Analyzing the Dynamics of Communication
in Online Social Networks

by

Munmun De Choudhury

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved February 2011 by the
Graduate Supervisory Committee:

Hari Sundaram, Chair
K. Selcuk Candan
Huan Liu
Duncan Watts
Doree Duncan Seligmann

ARIZONA STATE UNIVERSITY

May 2011

ABSTRACT

This dissertation deals with the analysis of interpersonal communication dynamics in online social networks and social media. My central hypothesis is that communication dynamics between individuals manifest themselves via three key aspects: the information that is the content of communication, the *social engagement* i.e. the sociological framework emergent of the communication process, and the *channel* i.e. the media via which communication takes place. My approach to developing a comprehensive understanding of these aspects in this dissertation is essentially computational as well as empirical: I present characterization techniques, develop computational models and finally discuss large-scale quantitative observational studies for each of these organizing ideas.

Communication dynamics have been of interest to researchers from multifaceted domains over the past several decades. However, today there are several modern capabilities encompassing a host of social media websites. These sites feature variegated interactional affordances, ranging from blogging, micro-blogging, sharing media elements as well as a rich set of social actions such as tagging, voting, commenting and so on. Consequently, these communication tools have begun to redefine the ways in which individuals exchange information, their modes of social engagement, and mechanisms of how the media characteristics impact their interactional behavior. Additionally, there are accompanying challenges regarding the nature of the communication itself. For example, it is often temporally volatile, allows an inexpensive expanse and reach to a broad audience and typically involves high reciprocity from a wide and demographically diverse population. This dissertation is geared towards developing methods, tools and frameworks towards understanding such key issues and at the same time addressing the associated challenges of making sense of the modern online communication processes.

The outcomes of this research are manifold. I present the contributions in three parts, corresponding to the three key organizing ideas. First, it is observed that user context is key to characterizing communication between a pair of individuals. However interestingly, the probability of future communication seems to be

more sensitive to the context compared to the delay, which appears to be rather habitual. Further, it is observed that diffusion of social actions in a network can be indicative of future information cascades; that might be attributed to social influence or homophily depending on the nature of the social action. Second, it is also observed that different modes of social engagement lead to evolution of groups that have considerable predictive capability in characterizing external-world temporal occurrences, such as stock market dynamics as well as collective political sentiments. Finally, characterization of communication on rich media sites have shown that conversations that are deemed “interesting” appear to have consequential impact on the properties of the social network: in terms of degree of participation of the individuals, thematic diffusion as well as emergent cohesiveness in activity among the participants in the network. Based on all these outcomes, I believe that this research can make significant contribution into a better understanding of how individuals communicate online and how it is redefining the collective sociological behavior.

ACKNOWLEDGEMENTS

I am heartily thankful to my advisor, Dr. Hari Sundaram, whose encouragement, guidance and support from the very initial to the final day has enabled me develop into a thoughtful and responsible researcher. The most valuable piece of training I gained from Hari is the ability to think of a research problem from a very fundamental sense—research which enables thinking “out-of-the-box”, instead of focusing on incremental innovation. This is extremely valuable in an area like computational social science, that is building on its ideas from constructs in other established disciplines. Hari was always very supportive of my interest in interdisciplinary research and encouraged me to *think big* and *aim high*, despite the numerous roadblocks that showed up time to time. Perhaps the most wonderful memory I will take away from my graduate studies is the freedom that Hari gave me to pursue research problems that interested me, often driven by my observations of our digital presence online or by my intuitions. I have constantly been rather impulsive about working on a variety of research problems on social networks at different points in time; and Hari was very patient and supportive to encourage me tread that path.

I have had an amazing group of coauthors and collaborators. I spent three summers (2006, 2009 and 2010) in internships in various research labs throughout the country, Avaya Labs Research, Yahoo! Research and Microsoft Research respectively, that gave me the opportunity to look at interdisciplinary research from a completely different perspective. It has been an honor for me to have the opportunity to work with Dr. Duncan Watts, Dr. Scott Counts, Dr. Mary Czerwinski, Dr. Sihem Amer-Yahia and Dr. Doree Duncan Seligmann during these three internships. The research I accomplished during these internships, enabled with the wonderful computing infrastructure and the rich availability of data, will remain very central to my academic accomplishment during the graduate school years. Additionally, encouraging discussions, fruitful guidance and generous help from Winter, Jake, Cong, Ronny, Ajita and Michael also helped me gather enthusiasm, have fun and remain on top of exciting research problems.

I would like to thank my thesis committee members, Dr. K. Selcuk Candan,

Dr. Huan Liu, Dr. Duncan Watts and Dr. Doree Duncan Seligmann for their advice and comments especially during the phase of my proposal defense. As a special mention, I offer my deepest gratitude to Dr. Aisling Kelliher, for always being a friend, philosopher and guide regarding issues of research or otherwise. Aisling always provided the alterative perspective on a research problem, which was valuable in defining why and how the same problem can be interesting to a person with a different academic background.

I am also indebted to my labmates in the Reflective Living group who have constantly provided a thriving and stimulating research environment and been a constant source of fun within the peripheries of graduate school research. It was wonderful working with a self-driven researcher like Yu-Ru and to be able to look up to bright and hard-working seniors like Yinpeng, Ankur, Adithya, Atit and Shree-harsh. Finally, feedback and discussion with current labmates Ryan, Shawn, Heng, Zhen and Silvan has always been fulfilling. It was also a learning opportunity to help and mentor Ke and Lu during the last stages of my PhD.

Life in Tempe would have been difficult without the wonderful ensemble of friends I have gathered over my five year stay. It has been a pleasure to be friends with Kartik, Sushovan, Aravind, Uday, Mithra and others who have time to time been a constant source of mental support during the easy and tough days.

I am grateful to my family members, especially my mother and my late father for supporting and encouraging me to pursue this degree when I was still very young and not so mature academically. I am fortunate to have their constant faith in me, no matter how easy or difficult the days have been. I am also very happy to have the love and support of a partner like Santanu who has consistently been very cooperative about my research career and has always encouraged me to remain on top of my ambitions. Santanu's parents, Maa and Baba also hold a special place for always being so supportive.

Lastly, I offer my regards to all of those in Tempe and otherwise who have supported me in any respect during the completion of my graduate school.

To My Father

TABLE OF CONTENTS

	Page
TABLE OF CONTENTS	vi
LIST OF FIGURES	xiv
LIST OF TABLES	xxvii
1 Introduction	1
1.1 Background and Motivation	4
Applications	6
Challenges	7
1.2 Overview and Contributions	8
1.3 Thesis Organization	14
2 Survey	16
2.1 Overview of Social Network Analysis	17
2.2 Social Networks: Concepts and Definitions	20
2.3 Communication Modes in Social Networks/Social Media	21
2.4 Communication Data Collection	25
2.5 Overview of Datasets	28
I Information Diffusion	34
3 Characterizing Communication: Topical Communication Flow	37
3.1 Introduction	38
Motivating Applications	41
Our Approach	42
3.2 Related Work	44
Information Diffusion	45
Analysis of Communication	47
Analysis of Information Roles of Users	48
Limitations	49
3.3 Problem Definition	51
The Role of Context	51
Data Model	52
Problem Statement	53

CHAPTER	Page
3.4 Communication Context	54
Neighborhood Context	54
Topic Context	57
Recipient Context	59
3.5 Social Context	61
Information Roles	61
Strength of Ties	63
3.6 Predicting Communication Flow	65
SVM Regression Framework	65
Prediction for Arbitrarily Distant Users	67
3.7 Evaluation Techniques	71
Frequency of Past Communication based Method	71
Trend Learning based Method	72
Epidemic Disease Model based Method	73
Neighborhood Effect based Method	74
3.8 MySpace Dataset	74
3.9 Experimental Results	77
Qualitative Case Study	77
Quantitative Evaluation	81
3.10 Extended Experiments on Social Network Datasets	88
3.11 Discussion	89
3.12 Conclusions	93
4 Diffusion of User Actions: Impact of Social Influence	102
4.1 Introduction	102
Related Work	104
Our Approach	104
4.2 What is Social Synchrony?	105
4.3 Problem Definition	107
4.4 Learning Framework	110
DBN Representation	110

CHAPTER	Page
User Context	111
Estimating User States	113
Estimating User Actions	114
4.5 Evolution Framework	115
Estimating Network Size	115
Evolving User Models	116
Predicting Synchrony	117
4.6 Experimental Results	117
Digg Dataset	118
Prediction of Periods of Synchrony	118
Properties of Seed Users	120
Comparative Evaluation	120
4.7 Extended Experiments on Social Media Datasets	122
4.8 Open Issues and Discussion	123
4.9 Conclusions	124
5 Diffusion of User Actions: Impact of Homophily	127
5.1 Introduction	128
Motivating Study	129
Main Contributions	130
5.2 Related Work	131
5.3 Problem Formulation	133
Preliminaries	133
Social Graph Model	134
Attribute Homophily	136
Topic Diffusion	136
Diffusion Series	137
5.4 Problem Statement	138
5.5 Characterizing Diffusion	139
5.6 Prediction Framework	141
5.7 Predicting Hidden States	143

CHAPTER	Page
5.8 Predicting Observed Action	144
5.9 Distortion Measurement	145
5.10 Experimental Results	146
5.11 Data Preparation	146
5.12 Quantitative Analysis	149
Temporal Analysis across Diffusion Metrics	149
Analysis across Themes	151
Comparative Study	151
5.13 Extended Experiments on Social Media Datasets	154
5.14 Discussion	156
5.15 Conclusions	157
II Evolution of Social Communication Networks	159
6 Community Dynamics and External Phenomena	163
6.1 Introduction	164
6.2 Modeling Information Roles	166
Roles due to Response Behavior	166
Roles due to Measure of Activity	168
6.3 Contextual Modeling of Communication Dynamics	169
6.4 Determining Correlation with Stock Market Movement	172
6.5 Experimental Results	173
Baseline Methods	173
Data Preparation	174
Results of SVR Regressor	174
6.6 Extended Experiments on Social Media Datasets	176
6.7 Conclusions	178
7 Prototypical Communication Groups	180
7.1 Introduction	181
Our Approach	184
7.2 Related Work	186
Social Network Characterization	187

CHAPTER	Page
Analysis of Communication Properties	190
Utility / Evaluation of Social Networks Dynamics	190
Limitations of Prior Work	191
7.3 Problem Definition	192
Preliminaries	192
Notion of a Prototypical Group	193
Data Model	194
Problem Statement	195
7.4 Group Extraction	197
Communication Features of Individuals	197
Clustering Individuals	200
7.5 Characterizing Prototypical Groups	202
Composition Entropy	202
Topic Divergence	203
Identifying Prototypical Groups	206
7.6 Utility of Prototypical Groups	207
Utility Variables	208
Utility Framework	209
7.7 Evaluation Method	210
Evaluation Metrics	210
Baseline Techniques	211
7.8 Experiments on Engadget Dataset	214
Description of Dataset	214
Analysis of Characteristics of Prototypical Groups	215
Analysis of Utility of Prototypical Groups	221
Evaluation of Prototypical Groups	223
7.9 Experiments on Huffington Post Dataset	226
Description of Dataset	226
Analysis of Characteristics of Prototypical Groups	227
Analysis of Utility of Prototypical Groups	230

CHAPTER	Page
Evaluation of Prototypical Groups	231
7.10 Experiments on Other Social Media Datasets	233
7.11 Discussion	237
7.12 Conclusions	239
III Rich Media Communication Patterns	247
8 Interestingness of Conversations	250
8.1 Introduction	251
Our Approach	255
Related Work	256
8.2 Problem Formulation	259
Definitions	259
Data Model	260
Problem Statement	262
8.3 Visual Themes from Media	263
8.4 Textual Themes from Media	265
Temporal Regularization	266
Co-participation based Regularization	268
8.5 Interestingness	270
Interestingness of Participants	271
Interestingness of Conversations	272
8.6 Consequences of Interestingness	275
8.7 Experimental Results	276
Analysis of Interestingness	278
Interestingness and Rich Media Attributes	283
8.8 Evaluation	286
Baseline Methods	286
Predictive Utility of Interestingness	293
Comparison of Visual and Textual Themes	295
8.9 Discussion	296
8.10 Conclusions	298

CHAPTER	Page
9 Quantifying Collective Participation in Social Media Conversations	299
9.1 Introduction	300
Our Approach	302
Main Results	303
9.2 Related Work	304
9.3 Social Media Conversations	306
Conversations	306
Datasets	307
9.4 Factors in Social Participation	309
Intrinsic Network Factors	309
Extrinsic Network Factors	313
9.5 A Prediction Framework	314
9.6 Validating Hypotheses	316
Prediction Performance	316
Statistical Significance	318
Summary of Findings	319
9.7 Combining Multiple Hypotheses	320
BIC Measure	321
Results	321
Open Issues	323
9.8 Conclusions	325
IV Conclusions	329
10 Summary of Contributions	330
11 Future Directions	334
11.1 Short-term Goals	334
11.2 Long-term Goals	337
V Appendix	340
12 Sampling Social Media Networks to Understand Social Processes	343
12.1 Introduction	343
12.2 Related Work	346

CHAPTER	Page
Graph Sampling	346
Social Diffusion Analysis	346
12.3 Problem Definition	347
12.4 Sampling Diffusion Data	350
12.5 Evaluation of Diffusion Samples	351
Diffusion Response Metrics	353
Results	356
12.6 Discussion	361
12.7 Conclusions and Future Work	362
13 Connecting Media to Community	364
13.1 Introduction	364
Related Work	365
Our Approach	366
13.2 Image Features	367
13.3 Tag Features	368
13.4 Communication Activity Features	369
13.5 Group Recommendation Framework	370
Main Idea	370
Model Learning	371
Folding-In Technique	371
13.6 Experimental Results	372
Flickr Dataset	372
Model Evaluation	372
Evaluation of Feature Types	374
Discussion	375
13.7 Conclusions	376
BIBLIOGRAPHY	377

LIST OF FIGURES

Figure		Page
1.1	Illustration of the three key organizing ideas that embody online interpersonal communication processes: namely, the information or concept that is the content of communication, the social engagement i.e. the sociological framework emergent of the communication process, and the channel or the media via which communication takes place.	9
2.1	The simplest social network.	17
3.1	Schematic representation of the data model used for predicting communication flow. A social network is represented as a weighted, multi-edge directed communication graph where an edge implies communication between the two connected users. In this figure, the thickness of an edge implies its weight, the color the topic and the directionality the flow of messages. Given a pair of users u and v and a topic Λ , we are interested in predicting two parameters—the probability of communication of u with v , $c_{uv}(\Lambda)$ and an association time delay in communication $d_{uv}(\Lambda)$	52
3.2	A Physics wavefront metaphor depicting primary and secondary wavefronts generated due to communication. Messages from u to v and w create primary wavefronts; while those which travel back to u , it is called backscatter. If v and w further propagate the messages, they constitute the secondary wavefronts.	55
3.3	Schematic representation explaining backscatter and susceptibility. . . .	56
3.4	Schematic representation explaining the information roles of users in a social network based on the dynamics of transmission and reception of messages.	62
3.5	Schematic representation explaining the notion of strength of ties between users in a social network. Strong ties usually form in a sub-network of densely connected users; while sparsely connected user sets are likely to have weak ties between them.	64

Figure	Page
3.6 Schematic representation of predicting communication flow between a pair of users who are arbitrarily n hops distant in the social network. The goal is to find a maximum probability $c_{uv}(\Lambda)$ and minimum delay $d_{uv}(\Lambda)$ path p between a pair of users u and v	68
3.7 Average Path Length Distribution of crawled MySpace data—we observe that the mean shortest path is ~ 5.9 which satisfies the criterion for a ‘small world’ network.	75
3.8 Logarithmic plot of the out-degree degree distribution of MySpace dataset. We observe that it follows a power-law, with exponent ~ 2.01 . This is in conformity with the observations in other small-world networks.	75
3.9 Distribution of messages over the different topics (synsets) extracted using WordNet.	77
3.10 Qualitative analysis of intent to communicate for a specific single node network. We show the predicted measures of intent over a set of four diverse topics for a user and each of her eight contacts. The visualization shows a radar plot of the measures of intent over a period of six weeks from Feb 19 to Apr 6, 2007.	80
3.11 Qualitative analysis of delay in communication for a specific single node network. We show the predicted measures of delay over a set of four diverse topics. The visualization shows a user and her local network comprising eight contacts (colored circles) over a period of six weeks from Feb 19 to Apr 6, 2007. The intensity of an edge between the center user and a contact reflects the measure of predicted delay when the center user intends to communication with the particular contact. Lighter shades of the edge reflect long delays; while darker shades imply the delay is short.	95
3.12 Error in prediction of intent to communicate over the MySpace dataset. The results are shown over a period of 40 weeks spanning from Jun 16, 2006 to Apr 24, 2007 and are averaged over all 25 topics. Comparison with four other baseline methods indicates that our method yields least error.	96

Figure	Page
3.13 Accuracy in prediction of delay in communication over the MySpace dataset. The results are shown over a period of 40 weeks spanning from Jun 16, 2006 to Apr 24, 2007 and are averaged over all 25 topics. Comparison with four other baseline methods indicates that our method yields highest accuracy.	96
3.14 Comparison of intent, (a) and delay, (b) predicted using our method and four other baseline methods against actual intent and delay. The values are shown over a set of different neighborhood sizes of MySpace users, varying from 1 to 1000. These results are averaged over the time span of 40 weeks of analysis, ranging from Jun 16, 2006 to Apr 24, 2007 and also averaged over all 25 topics.	97
3.15 Predicted intent to communicate and delay as a function of the number of hops between a pair of users, shown over the period of 40 weeks from Jun 16, 2006 to Apr 24, 2007 and averaged over all 25 topics.	98
3.16 Comparison of different feature selection types on the prediction performance of intent to communicate and delay from Jun 16, 2006 to Apr 24, 2007 and averaged over all 25 topics. We compare five different classes of feature selection: (1) use of our proposed dynamic feature selection method, (2) use of all features, (3) use of only communication context features, (4) use of only social context features, and (5) use of a random set of features. The colorbar represents the correspondence between the color of the blocks in the figure, and the measure of error. We observe that our method yields best performance.	99
3.17 The different features belonging to communication and social context that were dynamically selected over a period of 40 weeks from Jun 16, 2006 to Apr 24, 2007 and averaged over all 25 topics. A white box represents that the feature was selected at the corresponding time slice (shown on the X-axis) while a black one indicates that it was not. . . .	99

Figure	Page
3.18 Accuracy in prediction of intent to communicate and delay on three social network datasets—Twitter, YouTube and Flickr. The results are shown over a period of 50 days and are averaged over all topics in each dataset. Comparison with four other baseline methods indicates that our method yields highest accuracy.	100
3.19 Predicted intent to communicate and delay as a function of the number of hops between a pair of users, shown over the period of 50 days and averaged over all topics, corresponding to the three different datasets. .	101
4.1 Dynamics of the size of user set on two topics from Digg—‘Olympics’ and ‘Celebrity’ (September 1-30, 2008). Topic ‘Olympics’ is observed to exhibit synchrony where old users continue to be involved in the action of digging stories, as well as large number of new users join in the course of time (Sept 3-Sept 13).	106
4.2 Conceptual representation of synchrony in a social network over three time slices: t_1 , t_2 and t_3 . Synchrony is likely to involve (a) seed users, (b) sustained participation from old users (shaded circles) as well as attract new users (white circles) to perform a specific action over a period of time.	107
4.3 Representation of the two conditions (via two latent states) under which a user can perform a specific action—socially aware (e.g. digging on the Digg website), denoted by ‘1’ and socially unaware states (e.g. digging from an external site like the NY Times), denoted by ‘0’.	108
4.4 Illustration of the learning and evolution problems in prediction of synchrony.	110
4.5 Dynamic Bayesian Network representation of user actions using user context and states.	111
4.6 (a) Effect of a user’s neighborhood (i.e. contacts) on her actions, shown by the directed edges. (b) Coupling of a user’s actions with the seed users. The degree of coupling is represented using the thickness of an edge. Note the semantics of the two edges are different—a seed user in (b) may or may not be a contact of u in (a).	112

Figure	Page
4.7 Prediction of user actions based on our framework for six topics: ‘US Elections’, ‘World News’, ‘Olympics’, ‘Comedy’, ‘Celebrity’ and ‘Tennis’. Three variables—our predicted error, α and β (normalized) are shown on the left Y -axis, while the number of users is shown on the right Y -axis. Note, our model predicts the user actions with low mean error rate. . . .	125
4.8 Prediction of user actions (re-tweeting activity) based on our framework for four thematic categories on Twitter. Three variables—our predicted error, α and β (normalized) are shown on the left Y -axis, while the number of users is shown on the right Y -axis. Note, our model predicts the user actions with low mean error rate.	126
5.1 Examples of how attribute homophily impacts diffusion. We visualize diffusion on Twitter on the theme ‘Politics’ using four different attribute social graphs: location, activity behavior, information roles and content creation (ref. section 5.3). As noted, the discovered diffusion phenomenon is significantly different for the four attributes—e.g. the time span (size) of the ripple, the volume of users involved (width of the arcs) as well as the extent of tweeting activity (color intensity of each arc). . . .	130
5.2 (a) Distribution of Twitter users in different continents based on the location attribute; (b) Comparative examples of activity behavior distributions of three users over a period of 24 hours.	135
5.3 The three categories of user behavior corresponding to the attribute: information roles. Out-going links indicate post creation while incoming links indicate post responses from other users.	135
5.4 Example of a diffusion series from Twitter on the topic “global warming”. The nodes are users involved in diffusion while the edges represent “friend links” connecting two users.	137

Figure	Page
5.5 (a) Structure of the Dynamic Bayesian network used for modeling social action of a user u_i . The diagram shows the relationship between environmental features ($\mathbf{F}_{i,N}(\theta)$), hidden states ($S_{i,N}(\theta)$) and the observed action ($O_{i,N}(\theta)$). (b) State transition diagram showing the ‘vulnerable’ ($S_i = 1$) and ‘indifferent’ states ($S_i = 0$) of a user u_i	142
5.6 Performance analysis of saturation and utility measurement of predicted diffusion characteristics over time (the higher, the better). Results are shown across different categories of saturation measurement metrics: user-based (volume, participation, dissemination), topology-based (reach, spread, cascade instances, collection size), time-based (rate), utility measurement metrics: search and news trends.	148
5.7 Mean saturation and utility measurement of predicted diffusion characteristics shown across different themes. The themes are: A–Business-Finance, B–Politics, C–Entertainment-Culture, D–Sports, E–Technology-Internet, F–Human Interest, G–Social Issues, H–Hospitality-Recreation.	152
5.8 Comparison of mean saturation and utility measurement of predicted diffusion characteristics across different methods. Results are shown over the location attribute.	152
5.9 Comparison of mean saturation and utility measurement of predicted diffusion characteristics across different methods on Digg. Results are shown for the baseline graph and three attributes, baseline graph (a–b), content creation (c–d), information roles (e–f) and activity behavior (g–h).	158
6.1 Skewness in normalizing response time.	167
6.2 Activity distribution.	168
6.3 Visualization of stock movements with time on vertical scale. B_1 and B_2 are the two baseline techniques. Blue bubbles indicate positive movement and red bubbles negative movements. Sizes of the bubbles represent magnitude of movement. The SVR prediction is found to follow the movement trend very closely with an error of 13.41%.	175

Figure	Page
6.4 Prediction of stock market movement using two social media and one blog dataset: Digg, Twitter and Huffington Post.	177
7.1 Conceptual representation of prototypical groups in a communication set of individuals. Individuals (smaller bubbles) in a communication set form groups (larger bubbles) and prototypical groups (larger bubbles that are shaded) are subsets of groups whose content and context of communication is representative of the entire set.	193
7.2 Data model showing relationship between individuals, comments, blog posts and tags. Each blog post is written by an Editor who associates it with relevant tag(s). Individuals can write comments on such blog posts as well as can reply to existing comments by others.	195
7.3 Schematic representation of the key challenges involved in this article. They are (a) extraction of groups in a communication set, (b) character- izing the prototypical groups and finally (c) justifying the utility of the prototypical groups.	197
7.4 Schematic representation of extraction of groups in a communication set.	197
7.5 Graphical representation of leaders and followers. In (a), the individual u communicates following the four other individuals, hence a follower. In (b), the individual u 's communication leads that of the four others, hence a leader.	200
7.6 Schematic representation of characterizing key groups from groups in a communication set. Key groups are a subset of the groups at a certain time slice.	202

Figure	Page
7.7 Illustration of composition entropy of a group g at t_i with respect to other groups in the previous time slice t_{i-1} . The thickness of an edge connecting group g at t_i to the groups in the previous time slice t_{i-1} represents the number of individuals who moved to g from a particular group in the past. In (a) group g at time t_i comprises individuals who belonged to several different groups at t_{i-1} . In (b), individuals in group g come from only two groups in the past. Hence the composition entropy in (a) is higher than that in (b).	203
7.8 The evaluation method used in this paper to test various aspects of our framework. There are two major ideas to be evaluated—the group extraction method (several baselines are shown in the left column) and the method used for identifying the prototypical groups (several baselines are shown in the right column). Corresponding to all possible cases (shown through the edges—in all 12 cases), four different measures are computed at each time slice t_i : conductance (ρ_i), coverage (χ_i), predictivity (α_i) and resilience (β_i). Our approach has been shown in the red colored edge.	213
7.9 Visualization of groups (shown in circles) and the prototypical groups (shaded circles, intensity representing the measure of optimality) in the communication sets on Apple and Microsoft. Every row corresponds to the groups at each week, shown vertically downwards. We observe the dynamics of composition entropy and topic divergence of the prototypical groups during periods of significant related external events.	242
7.10 Visualization of groups (shown in circles) and the prototypical groups (shaded circles, intensity representing the measure of optimality) in the communication sets on Google and Nokia. Every row corresponds to the groups at each week, shown vertically downwards. We observe the dynamics of composition entropy and topic divergence of the prototypical groups during periods of significant related external events.	243

7.11	Analysis of the prototypical groups against the non-prototypical ones based on two metrics—conductance and coverage. The results are shown for four communication sets—Apple, Microsoft, Google and Nokia, in the time period from May 19 to Aug 31, 2008. Five different values are shown per week and per communication set. The details of the different columns are as follows: Column A: the number of prototypical groups; Column B: the mean conductance over all the prototypical groups; Column C: the mean coverage over all the prototypical groups; Column D: the mean conductance over all the non-prototypical groups, i.e. the groups in the communication set not identified as prototypical groups; and finally, Column E: the mean coverage over all the non-prototypical groups. Note, the conductances for the prototypical groups are considerably lower than that of the non-prototypical ones; while the coverage is higher. This implies that the identified prototypical groups are of high quality and hence meaningfully model communication existing among the individuals in the communication sets.	244
7.12	Visualization of groups (shown in circles) and the prototypical groups (shaded circles, intensity representing the measure of optimality) in the communication sets on Obama and McCain. Every row corresponds to the groups at each week, shown vertically downwards. We observe the dynamics of composition entropy and topic divergence of the prototypical groups during periods of significant related external events.	245
7.13	Dynamics of composition entropy and topic divergence (shown in left Y-axis) of the prototypical groups of four datasets during periods of significant related external events, extracted from NY Times (shown as the NYT mentions trends on the right Y-axis). For each dataset, we show the dynamics of prototypical groups defined over two topical communication sets.	246

Figure	Page
8.1 Example of a conversation from YouTube. A conversation is associated with a unique media object comprising several temporally ordered comments on different latent themes from different authors (participants of the conversation).	252
8.2 Timing diagrams of the random walk models for computing interestingness of participants ($\mathbf{I_P}^{(q-1)}$) and of conversations ($\mathbf{I_C}^{(q-1)}$). The relationships between different variables affecting the two kinds of interestingness are shown.	271
8.3 Ripple visualization of conversational themes over time—15 week period between Oct. to Dec. 2008.	278
8.4 Interestingness of 45 participants from YouTube, ordered by decreasing mean number of comments from left to right, is shown along with the corresponding number of comments over 15 weeks (rows). The Pearson correlation coefficient between the number of comments and interestingness is also shown; which implies that interestingness of participants is less affected by number of comments during periods of significant external events.	282
8.5 Mean and maximum interestingness of conversations on two topical categories from YouTube: “News and Politics” and “Sports” over a 15 week period between Oct. and Dec. 2008.	283
8.6 Best lag for correlation of interestingness measures to the three consequence-based metrics: activity, cohesiveness and thematic interestingness. Our method with temporal smoothing (I_1) is seen to be sharply correlated with the three metrics of consequences having the following lags—8 days for activity, 9 days for cohesiveness and 7 days for thematic interestingness.	288

Figure	Page
8.7 Mutual Information based response over the three consequence metrics. Results are shown for two rich media based datasets: YouTube and Flickr. We evaluate our computed interestingness I_1 and I_2 against baseline methods, B_1 (comment frequency), B_2 (novelty of participation), B_3 (co-participation based PageRank). Our method incorporating temporal smoothness (I_1) uses its corresponding best lags and maximizes the mutual information based response metric for the three consequence-based metrics (activity, cohesiveness and thematic diffusion).	290
8.8 Mutual Information based response over the three consequence metrics. Results are shown for two blog-based datasets: Engadget and Huffington Post. We evaluate our computed interestingness I_1 and I_2 against baseline methods, B_1 (comment frequency), B_2 (novelty of participation), B_3 (co-participation based PageRank). Our method incorporating temporal smoothness (I_1) uses its corresponding best lags and maximizes the mutual information based response metric for the three consequence-based metrics (activity, cohesiveness and thematic diffusion).	293
8.9 Prediction of different conversational features based on (1) interestingness and (2) auto-regression over historical values for the YouTube dataset. Note, that interestingness value of a conversation is a better predictor. .	294
8.10 Comparison of visual and textual themes of conversations over different topical categories on YouTube. The performance of each thematic type is evaluated based on the mutual information based response metric, that explains how much our computed interestingness measure can account for the three consequence metrics.	295

Figure	Page
9.1 Participation (percentage of users on the y -axis and number of posted comments per user, on the x -axis) on conversations from two types of social media sites: rich media and blog forums. We notice a marked difference in newcomer participation — rich media datasets (YouTube:72.37%, Flickr: 57.43%) attract more newcomers than blog forums (Engadget: 29.35%, Huffington Post: 21.35%). Notice that Flickr has a significant core following (20.76%)—sustained participation from individuals who have posted more than 100 comments.	301
9.2 Prediction accuracies (higher numbers are better) of collective participation in social media conversations over four different datasets; corresponding error bars are also shown to illustrate the deviations. Media context and conversational interestingness perform better for rich media datasets, while social awareness and community characteristics perform better for blogs. Creator reputation appears to perform relatively poorly over all datasets.	327
9.3 Performance of combining different feature categories (or hypotheses) in predicting collective participation. For each starting feature set, we show which feature sets were selected at each iteration, that minimizes the BIC. The plot at the bottom shows the actual BIC measures of the combinations at each step (lower BIC values are better). Here SA: Social Awareness, CC: Community Characteristics, CR: Creator Reputation, MC: Media Context and CI: Conversational Interestingness.	328
12.1 Example of a diffusion series from Twitter on the topic “global warming”. 349	
12.2 Distortion of different diffusion metrics for different choices of sampling ratio ρ , averaged over all the nine themes (i.e. ~ 125 trending topics from Twitter).	356
12.3 Distortion of diffusion metrics across themes, averaged over sampling ratio ρ . Metrics – user-based: volume, participation and dissemination; topology-based: reach, spread, cascade instances and collection size; time-based: rate.	358

Figure	Page
12.4 Response behavior (using the KS statistic) with respect to search and news trend for different choices of sampling ratio ρ , averaged over all themes.	359
12.5 Response behavior (using the KS statistic) with respect to search and news trend across different themes, averaged over sampling ratio ρ	359
12.6 Comparative distribution of (a) volume and (b) spread over different strategies for sampling ratio $\rho=0.3$	361
13.1 The overall system overview of our group recommendation framework. .	366
13.2 Evaluation of recommendations suing different methods: user-based and group-based.	374
13.3 Evaluation of different feature categories in the group recommendation process.	375

LIST OF TABLES

Table	Page
1.1 Some social media statistics.	8
1.2 Thesis overview of different research directions and problems.	15
2.1 Datasets and their descriptions used in this thesis.	31
2.2 Data collection methods used in the thesis.	32
2.3 Thesis overview of different research directions and problems.	33
3.1 Algorithm for SVR prediction.	68
3.2 Restricted Shortest Path Problem.	70
3.3 Average Network Statistics for the MySpace dataset.	76
3.4 Example messages for the topic abstractions, based on WordNet synsets used in the experiments.	78
3.5 Error in prediction of intent to communicate and delay, for our method ‘communication flow’ and the four baseline techniques.	82
4.1 Properties of the seed user set for six different topics. $ U_0 $ is the size of seed set, μ_0, σ_0 are the mean and variance of the degree distribution of U_0 , ρ_0 is the clustering coefficient of the seed users’ sub-network and ς_0 is the measure of vulnerability.	121
4.2 Error in predicting user actions based on our method against baseline techniques—trend learning method (B_1), user activity regression based framework (B_2), SIR based epidemiological model (B_3) and a simple threshold based model of global cascades (B_4). Our method performs the best among all baseline techniques.	121
4.3 Error in predicting user actions (re-tweeting activity) based on our method against baseline techniques—trend learning method (B_1), user activity regression based framework (B_2), SIR based epidemiological model (B_3) and a simple threshold based model of global cascades (B_4). Our method performs the best among all baseline techniques, in the case of the social media Twitter.	123
5.1 Summary of statistics of the data used for studying diffusion on Twitter.	147
5.2 Example theme and trending topic associations on Twitter dataset. . . .	147

Table	Page
5.3 Comparison of mean saturation and utility measurement of predicted diffusion (the higher, the better) across all methods and attribute and baseline social graphs.	154
7.1 Statistics of different communication sets in the Engadget dataset. . . .	215
7.2 Comparison of the content of communication of the prototypical groups with that of the entire communication set, for the four different sets in Engadget. The results are shown aggregated over the 15 weeks period from May 19 to Aug 31, 2008. We observe that the content of communication in the prototypical groups is representative of that in the entire communication set.	219
7.3 Analysis of the utility of prototypical groups of four communication sets in the Engadget dataset. The prototypical groups have been used to predict a related external variable, stock market movements of the corresponding sets in the time period from May 19 to Aug 31, 2008. The actual stock price at each Friday at each week is shown, along with the error in predicting the corresponding movement based on the prototypical groups. Considerably low errors in prediction are observed; that justifies that the characterized prototypical groups are meaningful. . . .	222
7.4 Evaluation of our framework against 11 different baseline scenarios, involving different group extraction and prototypical group identification methods. For each method we show four values averaged over all the prototypical groups, per communication set—conductance (ρ), coverage (χ), predictivity (α) and resilience (β). These values are averaged over the 15 weeks of analysis, May 19 to Aug 31, 2008 for the Engadget dataset.	223
7.5 Statistics of the two communication sets in the Huffington Post dataset.	226

Table	Page
7.6 Comparison of the content of communication of the prototypical groups with that of the entire communication set, for the two different sets in Huffington Post. The results are shown aggregated over the 15 weeks period from Jun 23 to Oct 3, 2008. We observe that the content of communication in the prototypical groups is representative of that in the entire communication set.	229
7.7 Analysis of the prototypical groups against the non-prototypical ones based on two metrics—conductance and coverage. The results are shown for two communication sets -Obama and McCain, in the time period from Jun 23 to Oct 2, 2008. Five different values are shown per week and per communication set. The details of the different columns are as follows: Column A: the number of prototypical groups; Column B: the mean conductance over all the prototypical groups; Column C: the mean coverage over all the prototypical groups; Column D: the mean conductance over all the non-prototypical groups, i.e. the groups in the communication set not identified as prototypical groups; and finally, Column E: the mean coverage over all the non-prototypical groups. Note, the conductances for the prototypical groups are considerably lower than that of the non-prototypical ones; while the coverage is higher. This implies that the identified prototypical groups are of high quality and hence meaningfully model communication existing among the individuals in the communication sets.	231
7.8 Analysis of the utility of prototypical groups of two communication sets in the Huffington Post dataset. The prototypical groups have been used to predict a related external variable, national polls of the corresponding political candidates in the time period from Jun 23 to Oct 3, 2008. The actual percentage of the poll at each week is shown, along with the error in predicting the corresponding movement based on the prototypical groups. Considerably low errors in prediction are observed; that justifies that the characterized prototypical groups are meaningful.	232

Table	Page
7.9 Evaluation of our framework against 11 different baseline scenarios, involving different group extraction and prototypical group identification methods. For each method we show four values averaged over all the prototypical groups, per communication set—conductance (ρ), coverage (χ), predictivity (α) and resilience (β). These values are averaged over the 15 weeks of analysis, Jun 23 to Oct 3, 2008 for the Huffington dataset.	234
7.10 Description of topics and categories over which prototypical groups were detected.	235
7.11 Statistics of the different communication sets over the four social media datasets.	236
7.12 Analysis of the prototypical groups against the non-prototypical ones based on two metrics—conductance and coverage. The results are shown for several topical communication sets belonging the four social media datasets. Note, p=Prototypical roudp, NP=Non-prototypical groups. .	237
8.1 Statistics (a total set of 272,810 videos involving 17,736,361 unique participants and 145,682,273 comments during 2008-2009) to demonstrate returning participants to videos on YouTube. We observe that the number of unique participants posting comments is much lesser compared to the number of comments, indicating that participants return to videos because of their engagement in the conversation (also note the comment-to-author ratio). The numbers are shown across different featured topical categories on YouTube.	253
8.2 Experimental Setup based on multiple rich media and blog datasets. . .	277
8.3 Details of the four datasets.	277
8.4 Word cloud of a representative theme belonging to three topical categories on YouTube.	279
8.5 Relevant media attributes corresponding to multiple rich media and blog datasets. These attributes were used to investigate how much useful they are for accounting for conversational interestingness.	284

Table	Page
8.6 Mean correlation coefficient between interestingness and media attributes over all four datasets. Note that media tenure implies the time elapsed since the video, image or blog post was uploaded or posted.	284
8.7 Correlation coefficient between interestingness and media attributes, shown for each topical category on YouTube.	286
8.8 Best lags in explaining the consequence metrics using our proposed method. The numbers are shown over multiple rich media and blog datasets. . .	287
8.9 Mean prediction error (%) of different conversational features based on (1) interestingness and (2) auto-regression over historical values over all four datasets. The errors are averaged over a period of ~ 15 weeks. . . .	294
9.1 Description of conversations on different rich media and blog datasets. .	308
9.2 Details of the four datasets.	309
9.3 Media context on multiple rich media and blog datasets.	311
9.4 Statistical significance (based on one-tail paired t -test; significance level of 0.05) of the two best performing factors for each dataset, compared to the other features, for same datasets. Here, SA: Social Awareness, CC: Community Characteristics, CR: Creator Reputation, MC: Media Context and CI: Conversational Interestingness. p -values below the significance level are shown in italics. We observe that in each dataset, the performances of the “best” factors are statistically significant compared to the other factors. Note that because we used a one-tail paired t -test, the significance results of X & Y is different from Y & X.	318
9.5 Summary of results in prediction of collective participation by newcomers and existing participants. Here SA: Social Awareness, CC: Community Characteristics, CR: Creator Reputation, MC: Media Context and CI: Conversational Interestingness.	320
9.6 Summary of results of combining hypotheses in prediction of collective participation.	322

Table	Page
9.7 Prediction accuracies using (I) just the best performing hypotheses (Figure 2), (II) optimal hypotheses combination (Figure 3), and (III) all five hypotheses.	323
10.1 Thesis structure and research problems.	331
12.1 Summary of statistics of the data used for studying diffusion on Twitter.	355
12.2 Summary of performance of different sampling techniques over the saturation and response metrics.	360
13.1 Evaluation of our method (M_1) against k -Nearest Neighbor (M_2) using Precision, Recall and F1-measure. Metrics are computed at $k=1, 3, 10$ and for three cases: precision, recall and F1-measure over all images, over images owned by each user and over images belonging to each group. Our method outperforms k -NN.	373
13.2 Evaluation of the three types of features (image content, tags and communication features) used for image characterization against our optimal method (all features).	375

Chapter 1

Introduction

“Man is by nature a social animal...”—Aristotle (384 BC–322 BC), ancient Greek philosopher.

I was still in college back in 2003, when I heard about MySpace, as something called a “social network”; a website that lets you stay connected with your friends and family online. This was soon followed by Orkut, that started in 2004; and by that time I was already ready to graduate and looking forward to joining graduate school and plunging on a period of intense research. Orkut caught my attention, because I realized it was a wonderful way to stay in touch with my high school and college friends alike, while I was away in a different country for graduate studies.

My interest in these social sites has ever been looming; and I have always been captivated by their *dynamics*: why and how we interact with our friends, family and acquaintances online. I was particularly amazed by how it reflected our individual behavior. However, starting around two to three years back, the online social networking aspect reached newer heights, with the advent of the *social media* sites that rendered the act of sharing information, opinions and news dissemination on timely happenings with completely new dimensions altogether.

In the summer of 2009, the world witnessed the “Twitter Revolution”, a demonstration of protests by people in Iran, post their Presidential elections. The street demonstrations, rioting and other expressions of discontent found their way to the world in real time through social-media networks and online video, including Twitter and YouTube. Not only were these tools used by the Iranians to interact with the rest of the world, but also we witnessed the emergence of citizen journalism via these widespread multi-faceted interactions. The immediacy of the reports was

gripping. Well-developed Twitter lists showed a constant stream of situation updates and links to photos and videos, all of which painted a portrait of the developing turmoil. The same trend followed during the crisis and the protests in Egypt early in 2011, as well as during the rehabilitation work post the earthquake in Haiti. Haiti earthquake Twitter updates helped the world stay informed about what was going on in the country as well as providing information on ways to make donations.

Such new paradigms of online communication had thereby transcended from the erstwhile boundaries of using social sites only to connect to friends and family. They were now a podium to voice one’s opinions, thoughts, interests and cater to exogenous events of political, economic and social flavor.

Motivated by the emergent spectrum of changes brought about by these communication paradigms, this thesis develops a comprehensive understanding of our communication process that typically happens on various online social platforms. In our research, there are three key organizing ideas, characterizing our online social interactions: the information that is the content of communication, the “social engagement” i.e. the social system or the network that embodies the communication process, and finally, the “channel”, i.e. the media artifact via which communication takes place. For each of these organizing ideas in this thesis, we present characterization techniques, develop computational models and finally discuss large-scale quantitative observational studies.

Communication is central to the evolution of social systems. It is the process by which participating individuals create and share information with one another in order to reach a mutual understanding (Berger and Calabrese, 1975). Over the years, numerous empirical studies (Bass, 1969; Granovetter, 1973; Hammer, 1980; Dacin et al., 1999; Barabasi, Jeong, Neda, Ravasz, Schubert, and Vicsek, Barabasi et al.) on online social communication processes have revealed that properties of the associated social system, including the network structure and dynamics can be useful pointers to determining the outcome of many important social and economic relationships (Adamic and Adar, 2005; Adar and Adamic, 2005).

Despite the fundamental importance laid on the understanding of these

structures and their temporal behavior in many social and economic settings (Burt, 2004; Coleman, 1998; Cialdini and Goldstein, 2004; Kumar et al., 2006; Leskovec et al., 2008), the development of characterization tools, foundational computational models as well as insightful observational studies on large-scale social communication datasets has begun to happen only recently. This is because communication patterns on online social platforms are significantly different from their physical world counterpart. Consequently the methods, tools and studies designed to cater to longitudinal ethnographic studies on observed physical world interactions are different from online social data. This distinction can be viewed on several aspects relating to the nature of the online communication process itself: such as inexpensive reach to a global audience, volatility of content and easy accessibility of publishing information content online. The outcome of these differences is that today there is an ardent necessity to develop robust computational frameworks to characterize, model and conduct observational studies on online communication processes prevalent, rather pervasively, on the online domain.

Our research is also motivated from the potential ability of online communication patterns in helping to address multi-faceted sociological, behavioral as well as societal problems (Doreian, 1980; Cortes et al., 2003; Bearman et al., 2004; Fiore and Donath, 2005; Liben-Nowell et al., 2005; Leskovec et al., 2009). For example, the patterns of social engagement, reflected via the network structure play a fundamental role in determining how concepts or information are exchanged. Such information may be as simple as an invitation to a party, or as consequential as information about job opportunities, literacy, consumer products, disease containment and so on. Additionally, understanding the evolution of groups and communities can provide meaningful insights into the ways in which concepts form and aggregate, opinions develop as well as ties are made and broken, or even how the decisions of groups of individuals contribute to impact on external temporal occurrences. Finally, studies of shared user-generated media content can enable us to reflect on the ways in which our communication patterns affect our social memberships or our observed behavior on online platforms.

In the light of the above observations, the following three parts summarize the key research investigations in this thesis:

1. **Information Diffusion.** This part instruments the characterization of the information involved in the social communication process and how it diffuses in a social network. Our primary questions of interest include: (i) how do we model communication characteristics between a pair of individuals that affects the diffusion of information between them? and (ii) how do social actions, impact information diffusion, from two perspectives—social influence and user homophily?
2. **Evolution of Social Communication Networks.** This part of the thesis deals with analysis of the ‘social engagement’ that embodies the communication process—groups and the communities, and their dynamics. We address the following questions: (i) what are the characteristics of individuals in online communities? What is their correlation with external social, economic and political phenomena? and (ii) how do we model an evolving community? What are prototypical temporal representations of such evolving communities? How do we validate the properties of such prototypical representations?
3. **Rich Media Communication Patterns.** The final part of this thesis investigates rich media communication patterns, i.e. the characteristics of the emergent communication, centered around the shared media object. There are two questions of interest: (i) what are the characteristics of conversations centered around shared rich media objects? and (ii) how does we quantify collective participation on rich media conversations?

1.1 Background and Motivation

About two decades ago when the Web first came into being at CERN, as a simple network of computers communicating over a protocol, little did we know what are going to be its implications to our day-to-day lives today. From a primarily information-consumption model of the Web of the nineties, we have traversed a

significantly long way to the Web that we see today—the Web that is predominantly characterized by extensive information creation *and* information consumption parallelly on the part of a very diverse set of users round the globe; a model of the Web also popularly known as “Web 2.0” (a set of technologies such as Ajax, that underpins these websites).

The advent of these technologies has provided considerable leeway to a rich rubric of platforms that promote widespread user interactions on shared spaces. These interactions often take place in the context of either a shared media including, an image, a video, a ‘blog’ / ‘microblog’; or are built across social ties that reflect physical world human relationships. Over the past few years, there has been a rapid proliferation of such social platforms that include Flickr, YouTube, Twitter, Digg, Facebook and blogs such as Slashdot, Engadget, Live Journal and Huffington Post (Kleinberg, 2002; Kumar et al., 2003; Kempe et al., 2003; Gruhl et al., 2004).

The resultant impact of these social websites has been widespread. Individuals can express their opinions on personal blogs (Song et al., 2007; Iwata et al., 2007) as well as can share media objects to engage themselves in discussion (Haythornthwaite, 1996; Gill et al., 2007; Kaltenbrunner et al., 2007). Right from shopping a new car, to getting suggestions on investment, searching for the next holiday destination or even planning their next meal out, people have started to rely heavily on opinions expressed online or social resources that can provide them with useful insights into the diversely available set of options (Goldenberg and Muller, 2001; Howe, 2006; Gómez et al., 2008). Moreover, personal experiences as well as thoughts and opinions on external events also manifest themselves through “memes” (Leskovec et al., 2009), “online chatter” or variegated “voting” mechanisms (Wu and Huberman, 2008) in several people’s blogs and social profiles.

These interactional affordances provided by the online social media and social network sites present to us a broad podium of opportunities and ample scope to conduct research on the analysis of communication in social networks. Instead of focusing on longitudinal studies of relatively small groups such as participant observation (Newcomb, 1961; Granovetter, 1973) and surveys (Burt, 2004), researchers

today can study social processes such as information diffusion or community emergence at *very large scales*. This is because electronic social data can be collected at comparatively low cost requiring little resource maintenance, can span over diverse populations and be acquired over extended time periods. The result is that the study of social processes on a scale of million individuals, that would have been barely possible a few years back, is now gaining a lot of interest currently (Kumar et al., 2006; Leskovec and Horvitz, 2008).

The research goal addressed in this thesis therefore spans the study of communication patterns associated with large-scale social processes online¹. Communication is central to these social processes and to the evolution of the corresponding social systems (Feld, 1981). We are interested in the broad question: how do online social communication processes affect our social dynamics and our collective behavior?

Applications

We demonstrate the utility of addressing this research question. The structural and temporal dynamics of social communication, manifested via emergent social roles of individuals, propagation of information, medical or technological innovations, or community dynamics can lend useful insights into understanding user behavior. They can also help drive targeted advertising, impact ranking information sources as well as can help reflect on the evolution and characteristics of social relationships at micro and macro scales.

From the perspective of business intelligence, social networks are attractive because consumers are connecting with other consumers and their mutual trust tends to be higher, because of the reflection on their physical world relationships. The power of people interacting with people in an online setting has thus driven the success or failure of many companies in the Internet space.

Secondly, there is a tremendous amount of buzz from the media for this newest form of marketing. Social networks are constantly changing the catalog of

¹An area that has popularly begun to be known as “computational social science”.

consumer sentiment, attitudes and information. Companies that want to maximize their presence on the social Web are therefore likely to benefit by taking advantage of social networks in all stages of the business models, from awareness to learning to buying to loyalty. For example, online retailing relating to the buying and selling of physical items as well as opinion formation platforms such as blogs owe much of their success to the power of social interactions taking place on the Web. These include for instance, the product reviews on Amazon, the reputation mechanism on Ebay or the voting patterns on Digg.

There are several additional reasons why understanding the dynamics of communication in social networks can be of value. Communication characterization within a network could suggest alternative ranking strategies for ‘impactful’ individuals (Domingos and Richardson, 2001), groups and communities in social networks. These include: *what* are the sources of reliable information and *who* are the people capable of suggesting answers to specific questions. It could also help user behavior prediction underpinning different sociological processes, as well as support tools for marketing analysis and operations management that are required to cater to user needs in response to temporally changing external impulses or in situations involving crisis or emergency. Besides, it can be of great value in opinion tracking in business intelligence (Vijay Mahajan and Bass, 1990). For example, corporations are often interested in knowing how they and their products are being received at large to customers. Companies also want to understand shareholder sentiment, as well as to maximize exposure on a certain product. Hence they are likely to be interested in knowing which communities are discussing them, including the dynamics of such communication activities.

Challenges

There are several challenges involved in tackling the research question of analyzing communication dynamics around online social networks and social media. The central challenge is the scale, growth and expanse of online communication data. We

Table 1.1: Some social media statistics.

SOCIAL MEDIA TYPE	KEY STATISTICS
YouTube	139M users; US\$200M
Flickr	3.6B images; 50M users
Facebook	500M active users; 1B pieces of content shared each week
MySpace	110M monthly active users; 14 B comments on the site
Digg	3M unique users; \$40M
Engadget	1,887,887 monthly visitors
Huffington Post	8.9M visitors
Live Journal	19,128,882 accounts
Twitter	7B+ tweets, 55–60M tweets posted per day

highlight some of the key statistics² of different social sites available on the Web today in Table 1.1. It is obvious from the statistics in the table that traditional tools to understand social interactions in physical spaces or over industrial media, or even prior work involving longitudinal studies of groups of individuals are often inadequate in characterizing, modeling and observing the modern online communication of today.

In this thesis, we identify three key components that are associated with the online social communication process today. These three components are as follows (also shown illustratively in Figure 1.1): (1) The information that is exchanged due to the communication process; (2) The social engagement (e.g. the community dynamics) that embodies the communication process; and (3) The channel or the media (e.g. text, video or image) around which the social engagement manifests itself and the communication takes place. We will discuss these three coupled research

²Refer to the following websites for the statistics:

<http://mashable.com/2006/07/17/youtube-hits-1-million-videos-per-day/>,
<http://www.flickr.com/help/forum/en-us/97258/>,
<http://www.nytimes.com/2010/07/22/technology/22facebook.html>,
<http://www.nytimes.com/2009/05/04/technology/companies/04myspace.html>,
<http://techcrunch.com/2008/01/29/digg-nearly-triples-registered-users-in-a-year-says-sleuth-programmer/>,
<http://normalkid.com/2008/05/15/engadget-vs-gizmodo-stats/>,
<http://mediamemo.allthingsd.com/20100304/huffington-post-still-growing-like-a-weed/>,
<http://www.livejournal.com/stats.bml>,
<http://themetricsystem.rjmetrics.com/2010/01/26/new-data-on-twitters-users-and-engagement/>

Figure 1.1: Illustration of the three key organizing ideas that embody online interpersonal communication processes: namely, the information or concept that is the content of communication, the social engagement i.e. the sociological framework emergent of the communication process, and the channel or the media via which communication takes place.

perspectives in three major parts in this thesis.

1.2 Overview and Contributions

Given, individuals communicating among themselves on an online social platform, (a) exchanging information, (b) via a channel or a specific interactional media, and (c) provided with the affordances of a social engagement i.e. an emergent social network, this thesis address the following technical challenges.

Part I. How do we analyze information diffusion among communicating individuals?

Today, the widespread use of online social media has made the cost involved in propagating a piece of information to a large audience negligible, providing extensive evidences of large-scale information diffusion. In this part of the thesis, we study how communication dynamics impact diffusion of information: including characterization

of communication associated with flow of information, as well as factors that impact this flow (i.e., social influence manifested via the network, and personal properties such as homophily).

Topical Communication Flow: First we develop a representation framework (De Choudhury et al., 2007, 2009a) for communication characteristics among a given pair of individuals that impacts the flow of information on a certain topic. Communication is represented via different types of contextual attributes: network topology based context (e.g. a node’s embeddedness in her local network), social context (e.g. strength of ties), topical context (e.g. content of communication) and user identity context (e.g. who are the people involved in communication). Such contextual representation is then used to predict the flow of information on a large dataset crawled from MySpace.

We have excellent results on several social network datasets. Our model yields excellent prediction results with error rates lower than $\sim 15\%$. Finally we observed qualitatively that intent is more affected due to contextual dynamics than delay. Delay seemed to be more dependent on other latent factors characterizing communication, e.g. ‘age’ of the information content under consideration or the evolution of user behavior over time.

Diffusion of User Actions: Second, we investigated two important facets enveloping communication, that impact the information diffusion process. The first facet relates to “social influence” in the light of the interactional process, e.g. temporal reciprocity in communication patterns between pairs of individuals in a network. In the second facet, we investigate the effect of “homophily” on the diffusion of information, e.g. similarity among individuals in terms of their communication engagement.

First, we develop a computational framework to study the process of diffusion in the light of network factors (De Choudhury et al., 2009b). The diffusion process is defined in the form of information cascades that reflect a temporal social network phenomenon. A cascade is characterized by a large number of users who mimic a certain action over a period of time with sustained participation from early

users. Our proposed computational framework comprises two parts: the learning framework and the evolution framework. In the learning framework, we develop a Dynamic Bayesian Network based representation that includes an understanding of user’s network context to predict the probability of user actions in the future. In the evolution framework, we simulate the social network and the user models over a set of future time slices to predict the information cascades responsible for diffusion.

Next we undertake large-scale observational studies on the impact of attribute homophily on information diffusion (De Choudhury et al., 2010). Homophily structures networks: people’s ego-centric social networks are often homogeneous with regard to diverse social, demographic, behavioral, and intra-personal characteristics. They can also revolve around social foci such as co-location or commonly situated activities. As a consequence of interaction in these homogeneous constructs, individuals tend to become interpersonally tied to each other. Hence, the existence of such *homophily* is likely to impact the information these individuals receive and propagate, the communication activities they engage in, and the social roles they form. In our work, we consider communication occurring via posts (also known as “tweets”) on the popular micro-blogging service Twitter and investigate the relationship between homophily among users and the social process of diffusion. In particular, we study four kind of contextual attributes on Twitter: location, activity behavior, social role and activity distribution. Thereafter, using a dynamic Bayesian network framework, we predict diffusion characteristics under homophily on each of these attributes at a certain time slice.

We conducted extensive experiments on large datasets crawled from popular social media sites, including Digg and Twitter. The results showed that our model yielded low error ($15.2 \pm 4.3\%$) in predicting user actions during time periods with and without observable information cascades—implying that network factors do indeed affect a user’s intent to participate in a cascade. We further demonstrate, based on experiments on social media datasets, that the choice of the homophilous attribute can impact the prediction of information diffusion as well, given a specific metric and a topic. In most cases, attribute homophily is able to explain the actual

diffusion and external trends by a margin of $\sim 15\text{-}25\%$ lower distortion compared to cases when homophily is not considered.

Part II. How does the social engagement (the emergent network) give rise to emergent collective behavior of groups and communities?

Communication process gives rise to commentary regarding shared interests, often on topics relating to events in the real world. In this part of the thesis, we study how we can understand and model the overall collective behavior of communities, particularly, communication sets on social websites. We are interested in investigating the utility of community behavior in predicting external phenomena, as well as methods to identify representative prototypical groups.

Community and External Phenomena: We characterize collective behavior of communities via the information roles of the communicating individuals—roles emerging out of their communication activity, such as early responders and late trailers (De Choudhury et al., 2008a). We further utilize the collective behavior of these communities to determine correlation with stock market movements of companies and political polls in favor of election candidates, based on a Support Vector Regression prediction framework (De Choudhury et al., 2008c,b).

We observe from our experiments that the communication activity on different social network datasets has correlation with stock market movement and political polls. These correlation measures are further cross-validated against two baseline methods. Our results are promising yielding about 78% accuracy in predicting the magnitude of stock market movement and political polls.

Prototypical Groups: We develop computational models to extract, characterize and study the utility of prototypical communication groups in the Blogosphere (De Choudhury et al., 2011). Our primary motivation in this regard is that although useful, tracking the communication dynamics of individuals over an entire network over time can be extremely expensive and inefficient. Moreover, the nature of the overall communication cannot be deciphered on the basis of a single individual or a large

set of individuals. We need a unifying notion that helps us characterize macroscopic interactions in a communicating set of individuals based on the microscopic communication properties of the individuals. Hence we analyze communication within a set of individuals to extract the representative prototypical groups and provides a novel framework to establish the utility of such groups. First, we extract groups by developing features representing communication dynamics of the individuals. Second, to characterize the overall communication set on a certain topic, we identify a subset of groups within the community as the prototypical groups. Third, we justify the utility of these prototypical groups by using them as predictors of related external phenomena; specifically, stock market movement of technology companies and political polls of Presidential candidates in the 2008 US elections.

We have conducted extensive experiments on different social network datasets. We observe that the prototypical groups can predict stock market movement / political polls satisfactorily with mean error rate of 20.32%. We evaluate the quality of the extracted groups based on their conductance and coverage measures and develop metrics—predictivity and resilience to evaluate their ability to predict a related external time series variable (stock market movement / political polls). This implies that communication dynamics of individuals are essential in extracting groups in a community, and the prototypical groups extracted by our method are meaningful in characterizing the overall communication sets.

Part III. How do we characterize the properties of the media objects around which communication takes place?

Online groups often engender significant dialogue based commentary (or “conversations”) centered around shared rich social media objects (text, image, video). This often leads to observable of the media objects itself, including its interestingness, popularity and rating. At the same time, we observe differential degree of participation on various conversations. Participation impacts the sustainability of the associated online communities as a whole. Hence in this part of the thesis, we inves-

tigate dynamics of conversations around social media objects, and the factors that drive individuals to participate in these conversations.

Conversational Interestingness: First, in our attempt towards understanding communication dynamics around rich media, we propose a computational framework to predict the “interestingness” property of conversations associated with media objects; e.g. videos on YouTube, images on Flickr (De Choudhury et al., 2009c). Interestingness is defined as a subjective property that depends on two properties: conversational themes and co-participation of individuals in commentary. Based on this operational definition, our proposed framework utilizes a one-dimensional Markov chain to determine the measure of interestingness of a media object dynamically.

In our experiments we determine conversational interestingness across several large social media datasets. In the absence of ground truth to validate our computed measure, we determine the consequence of a conversation by measuring how interestingness affects the following three variables—participation in related themes, participant cohesiveness and theme diffusion. Our results show that our method of measuring interestingness maximizes the mutual information, and is significantly better (twice as large) than three other baseline methods (number of comments, number of new participants and PageRank based assessment). This reflects that the interestingness of conversations associated with rich media often has consequential impact on the social network as a whole.

Collective Participation in Social Media Conversations: Second, we quantitatively study user behavior responsible for the widespread media participation predominant on social media conversations today. By media participation, we mean the social action of leaving a comment on a photo / video. We propose to seek answers to the following question: *how do we quantitatively characterize large-scale media participation on a topic and a given point in time?* In our framework, we propose several intrinsic and extrinsic network factors that are likely to drive participation on social media conversations.

We observed that conversations on rich media are more affected by factors

relating to the conversational characteristics, those on blog forums depend more on the participant’s awareness and the community’s behavior. Additionally, whether a newcomer will participate in a conversation seems to be more affected by his awareness of what kind of responses and feedback he receives from the community, whereas the existing participants tend to rely more on what is the behavior of the overall community, such as whether the community support cohesiveness in participation, or whether it sustains sufficient individuals as a consequence of prior participation. Using a Bayesian Information Criterion (BIC) metric, we also identify that there is usually an *optimal combination of factors* that drive collective participation; and it often depends on the associated *media type*.

Table 1.2: Thesis overview of different research directions and problems.

Information Diffusion	Topical Communication Flow (Chapter 3), Diffusion of User Actions (Chapters 4, 5)
Evolution of Social Communication Networks	Community Dynamics and External Phenomena (Chapter 6), Prototypical Communication Groups (Chapter 7)
Rich Media Communication Patterns	Interestingness of Conversations (Chapter 8), Quantifying Collective Participation on Social Media Conversations (Chapter 9)

1.3 Thesis Organization

We summarize the organizational structure of this thesis in Table 1.2. The thesis has three main threads, that coherently tie the three aspects along which we characterize the online communication process. Additionally, note that for each research problem addressed in this thesis, there are three parts: (1) characterization: where we develop techniques and frameworks to characterize properties and attributes of the social phenomenon being studied, (2) models: where we utilize the characterized social phenomena to develop computational models that are useful to predict future events, or to quantitatively study social network communication dynamics, and (3) empirical studies: where the computational models developed are used to instrument interesting observations made on large-scale data.

Chapter 2

Survey

Social network analysis is focused on uncovering the patterning of people's interaction. Network analysis is based on the intuitive notion that these patterns are important features of the lives of the individuals who display them. Network analysts believe that how an individual lives depends in large part on how that individual is tied into the larger web of social connections. Many believe, moreover, that the success or failure of societies and organizations often depends on the patterning of their internal structure. From the outset, the network approach to the study of behavior has involved two commitments: (1) it is guided by formal theory organized in mathematical terms, and (2) it is grounded in the systematic analysis of empirical data. It was not until the 1970s, therefore—when modern discrete combinatorics (particularly graph theory) experienced rapid development and relatively powerful computers became readily available—that the study of social networks really began to take off as an interdisciplinary specialty. Since then its growth has been rapid. It has found important applications in organizational behavior, inter-organizational relations, the spread of contagious diseases, mental health, social support, the diffusion of information and animal social organization.

In this chapter, we present a comprehensive discussion of the various concepts and definitions associated with network theory in general and social networks in particular. Next we present a discussion of interactional or communication modes prevalent in different social networks today and how they are embedded in the structural characteristics of the corresponding network. We present several methods of collection of such interactional social data, in the interest of this thesis. Finally we discuss a detailed overview of the different online datasets of relevance in this thesis.

2.1 Overview of Social Network Analysis

What is a Network?

Definition 1 *A network is a set of relationships. More formally, a network contains a set of objects (in mathematical terms, nodes) and a mapping or description of relations between the objects or nodes.*

In other words, the elements of a networked system are represented as nodes (also known as actors or vertices) and the connections among interacting elements are known as ties, edges, arcs, or links. The nodes might be neurons, individuals, groups, organizations, airports, or even countries, whereas ties can take the form of friendship, communication, collaboration, alliance, flow, or trade, to name a few.

The simplest network contains two objects, 1 and 2, and one relationship that links them (Figure 2.1). Nodes 1 and 2, for example, might be people, and the relationship that links them might be “are part of the same photography club at ASU.”

Figure 2.1: The simplest social network.

There are also directional relationships such as 1 likes 2. In this simple network, the relationship could be symmetrical or non-directional: 1 and 2 like one another, or their liking is mutual. A weighted network is a network where the ties among nodes have weights assigned to them. There also need not be just one relationship mapped between nodes 1 and 2. For example, 1 and 2 might be in the same room and might like one another. When there is more than one relationship, this is called a *multiplex* relationship.

Aside from their directionality, or lack of it, relationships might be more than the sharing of an attribute or being in the same place at the same time. There can be a flow between the objects or the nodes. Liking, for example, might lead to an exchange of gifts. Flows and exchanges can be very important in network theory.

Social network analysis is based on an assumption of the importance of relationships among interacting units in such a network, to answer questions that are of interest from the perspective of social theory (Wasserman and Faust, 1994). Social scientists have investigated four kinds of networks: ego-centric, socio-centric, open-system and snow-ball networks.

Ego-centric networks are those networks that are connected with a single node or individual, for example, my good friends, or, all the students in the Computer Science Department at ASU. To be considered networks these connections must not only be lists of people or organizations, but information must be available about the connections between these people or organizations. Otherwise, there is no network to analyze. In popular discourse, especially when social support is discussed, any list is considered to be a network. A person with a large number of good friends whom he or she can count on is said to have a large “network”. This network cannot be discussed in social network terms, however, unless we know whether and how these people are connected with one another. It is obviously one thing to have a supporting network in which most people know one another and a very different matter if the people are unknown to one another.

Socio-centric networks are networks in a box. Connection between children in a classroom, between executives or workers in an organization are closed system networks and the ones most often studied in terms of the fine points of network structure.

Open system networks are networks in which the boundaries are not necessarily clear, they are not in a box—for example, the elite of the United States, or connections between corporations, or the chain of influencers of a particular decision, or the adoption of new practices. In some ways these are the most interesting networks. They are also the most difficult to study.

A snowball network refers to the idea that the alters identified in an egocentric survey then become egos themselves and are able in turn to nominate additional alters. While there are severe logistic limits to conducting snowball network studies, a method for examining hybrid networks has recently been developed in which egos

in complete networks can nominate alters otherwise not listed who are then available for all subsequent egos to see. The hybrid network may be valuable for examining whole/complete networks that are expected to include important players beyond those who are formally identified. For example, employees of a company often work with non-company consultants who may be part of a network that cannot fully be defined prior to data collection.

The social network perspective thus encompasses theories, models, and applications that are expressed in terms of relational concepts or processes. Along with growing interest and increased use of network analysis has come a consensus about the central principles underlying the network perspective. In addition to the use of relational concepts, we note the following as being important:

1. Actors and their actions are viewed as interdependent rather than independent, autonomous units.
2. Relational ties (linkages) between actors are channels for transfer or "flow" of resources (either material or nonmaterial).
3. Network models focusing on individuals view the network structural environment as providing opportunities for or constraints on individual action.
4. Network models conceptualize structure (social, economic, political, and so forth) as lasting patterns of relations among actors.

We now discuss some of the characteristics of social networks. The shape of a social network helps determine a network's usefulness to its individuals. Smaller, tighter networks can be less useful to their members than networks with lots of loose connections (weak ties) (Granovetter, 1973) to individuals outside the main network. More open networks, with many weak ties and social connections, are more likely to introduce new ideas and opportunities to their members than closed networks with many redundant ties. In other words, a group of friends who only do things with each other already share the same knowledge and opportunities. A group of individuals with connections to other social worlds is likely to have

access to a wider range of information. It is better for individual success to have connections to a variety of networks rather than many connections within a single network. Similarly, individuals can exercise influence or act as brokers within their social networks by bridging two networks that are not directly linked (called filling structural holes) (Burt, 1982).

The power of social network analysis stems from its difference from traditional social scientific studies, which assume that it is the attributes of individual actors—whether they are friendly or unfriendly, smart or dumb, etc.—that matter. Social network analysis produces an alternate view, where the attributes of individuals are less important than their relationships and ties with other actors within the network. This approach has turned out to be useful for explaining many real-world phenomena, but leaves less room for individual agency, the ability for individuals to influence their success, because so much of it rests within the structure of their network.

Social networks have also been used to examine how organizations interact with each other, characterizing the many informal connections that link executives together, as well as associations and connections between individual employees at different organizations. For example, power within organizations often comes more from the degree to which an individual within a network is at the center of many relationships than actual job title. Social networks also play a key role in hiring, in business success, and in job performance. Networks provide ways for companies to gather information, deter competition, and collude in setting prices or policies (Wasserman and Faust, 1994).

2.2 Social Networks: Concepts and Definitions

A comprehensive list of different concepts and definitions associated with social network analysis is presented below¹:

Centrality. This measure gives a rough indication of the social power of a node

¹http://en.wikipedia.org/wiki/Social_network#Metrics_.28Measures.29_in_social_network_analysis

based on how well they “connect” the network. “Betweenness”, “Closeness”, and “Degree” are all measures of centrality.

Clustering coefficient. A measure of the likelihood that two associates of a node are associates themselves (Watts and Strogatz, 1998). A higher clustering coefficient indicates a greater ‘cliquishness’.

Degree. The count of the number of ties to other actors in the network. See also degree (graph theory).

Density. The degree a respondent’s ties know one another/ proportion of ties among an individual’s nominees. Network or global-level density is the proportion of ties in a network relative to the total number possible (sparse versus dense networks).

Path Length The distances between pairs of nodes in the network. Average path-length is the average of these distances between all pairs of nodes.

Structural hole. Static holes that can be strategically filled by connecting one or more links to link together other points (Burt, 1995). Linked to ideas of social capital: if you link to two people who are not linked you can control their communication.

Embeddedness. Embeddedness (Granovetter, 1973) is the degree to which individuals are enmeshed in a social network.

Triadic Closure. Triadic closure is the property among 3 nodes A, B, and C, such that if a strong tie exists between A-B and A-C, there is a weak or strong tie between B-C (Granovetter, 1973).

2.3 Communication Modes in Social Networks/Social Media

We discuss several different communication modes popularly existent in social networks and social media sites today. These diverse modalities of communication allow users to engage in interaction often spanning a commonly situated inter-

est, shared activities or artifacts, geographical, ethnic or gender-based co-location, or even dialogue on external news events. In this thesis, we have focused on the following forms of communication among users, that are likely to promote social interaction:

1. *Messages*. Social websites such as MySpace feature an ability to users to post short messages on their friends' profiles. A similar feature on Facebook allows users to post content on another user's "Wall". These messages are typically short and viewable publicly to the common set of friends to both the users; providing evidences of interaction via communication.
2. *Blog Comments / Replies*. Commenting and replying capability provided by different blogging websites, such as Engadget, Huffington Post, Slashdot, Mashable or MetaFilter provide substantial evidence of back and forth communication among sets of users, often relating to the topic of the blog post. Note, replies are usually shown as an indented block in response to the particular comment in question.
3. *Conversations around Shared Media Artifact*. Many social websites allow users to share media artifacts with their local network or set of contacts. For example, on Flickr a user can upload a photo that is viewable via a feed to her contacts; while YouTube allows users to upload videos encompassing different topical categories. Both these kinds of media sharing allow rich communication activity centered around the media elements via comments. These comments often take a conversational structure, involving considerable back and forth dialogue among users.
4. *Social Actions*. A different kind of a communication modality provided by certain social sites such as Digg or del.icio.us involves participation in a variety of social actions by users. For example, Digg allows users to vote (or rate) on shared articles, typically news, via a social action called "digging". Another example is the "like" feature provided by Facebook on user statuses, photos, videos and shared links. Such social action often acts as a proxy for

communication activity, because first, it is publicly observable, and second it allows social interaction among the users.

5. *Micro-blogging.* Finally, we define a communication modality based on micro-blogging activity of users, e.g. as provided by Twitter. The micro-blogging feature, specifically called “tweeting” on Twitter, often takes conversational form, since tweets can be directed to a particular user as well. Moreover, Twitter allows the “RT” or re-tweet feature, allowing users to propagate information from one user to another. Hence micro-blogging activity can be considered as an active interactional medium.

Now we will survey some prior work on the above presented communication modalities. In the interest of the discussion we organize the different modalities into two broad categories: conversations and social actions. Typically in the first category, we discuss work that have replied on direct communication events, such as messages, emails and comment-reply artifacts to understand social phenomena. The second category deals with more abstract forms of communication, such as voting mechanism on shared media, to promote social interaction among sets of individuals.

Conversations. Social networks evolve centered around communication artifacts. The conversational structure by dint of which several social processes unfold, such as diffusion of innovation and cultural bias, discovery of experts or evolution of groups, is valuable because it lends insights into the nature of the network at multi-grained temporal and topological levels and helps us understand networks as an emergent property of social interaction.

Comments and messaging structure in blogs and shared social spaces have been used to understand dialogue based conversational behavior among individuals (Schuth et al., 2007) as well as in the context of summarization of social activity on the online platform (Lu et al., 2009; Hu et al., 2007) or to understand the descriptive nature of web comments (Potthast, 2009). Some prior work have also deployed conversational nature of comments to understand social network structure as well

as in statistical analysis of networks (Gómez et al., 2008). There has also been considerable work on analyzing discussions or comments in blogs (Kaltenbrunner et al., 2007; Mishne and Glance, 2006) as well as utilizing such communication for prediction of its consequences like user behavior, sales, stock market activity etc (Adar et al., 2007; De Choudhury et al., 2008b; Gruhl et al., 2005; Liu et al., 2007).

Prior research has also discovered value in using social interactional data to understand and in certain cases predict external behavioral phenomena. There has been considerable work on analyzing social network characteristics in blogs (Kumar et al., 2006) as well as utilizing such communication for prediction of its consequences like user behavior, sales, stock market activity etc (Adar et al., 2007; Gruhl et al., 2005; Liu et al., 2007). In (Gruhl et al., 2005) Gruhl et al. attempt to determine if blog data exhibit any recognizable pattern prior to spikes in the ranking of the sales of books on Amazon.com. Adar et al. in (Adar et al., 2007) present a framework for modeling and predicting user behavior on the web. They created a model for several sets of user behavior and used it to automatically compare the reaction of a user population on one medium e.g. search engines, blogs etc to the reactions on another.

While these prior work lend insights into understanding the nature of discussions in online social spaces, we do not have a clear understanding of how the evolving characteristics of the communication patterns affect the propagation of information engraved in the interactions, how it impacts the social network structure in terms of evolution of groups and communications and in terms of relationship to external temporal phenomena, or how these patterns can reveal dynamic properties of shared media artifacts. This thesis analyzes conversations computationally in the context of large-scale online social data in the interest of these problems.

Social Actions. The participation of individual users in online social spaces is one of the most noted features in the recent explosive growth of popular online communities ranging from picture and video sharing (Flickr.com and YouTube.com) and collective music recommendation (Last.fm) to news voting (Digg. com) and

social bookmarking (del.icio.us). However in contrast to traditional communities, these sites do not feature direct communication or conversational mechanisms to its members. This has given rise to an interesting pattern of social action based interaction among users. The users' involvement and their contribution through non-message-based interactions, e.g. digging or social bookmarking have become a major force behind the success of these social spaces. Studying this new type of user interactional modality is crucial to understanding the dynamics of online social communities and community monetization.

Social actions performed on shared spaces often promote rich communication dynamics among individuals. In prior work, authors have discussed how the voting i.e. digging activity on Digg impacts the discovery of novel information (Wu and Huberman, 2008). Researchers (Viswanath et al., 2009) have also examined the evolution of activity between users in the Facebook social network to capture the notion of how social links can grow stronger or weaker over time. Their experiments reveal that links in the activity network on Facebook tend to come and go rapidly over time, and the strength of ties exhibits a general decreasing trend of activity as the social network link ages. Social actions revealed via third party applications as featured by Facebook have also lent interesting insights into the social characteristics of online user behavior. In (Gjoka et al., 2008), the authors characterize the popularity and user reach of Facebook applications. Their findings include that the popularity of Facebook applications is highly skewed across different users and that users with more applications installed are more likely to install new applications.

Social actions have also been analyzed in the context of information creation and dissipation. With content fast becoming a commodity, interest in using social networks to promote and find content has grown, both on the side of content producers (viral marketing) and consumers (recommendation). Lerman et al. in (Lerman and Galstyan, 2008) study the role of social networks in promoting content on Digg, a social news aggregator that allows users to submit links to and vote on news stories. Their empirical findings suggest that pattern of the spread of interest in a story on the network is indicative of how popular the story will become. Stories that

spread mainly outside of the submitter’s neighborhood go on to be very popular, while stories that spread mainly through submitter’s social neighborhood prove not to be very popular.

2.4 Communication Data Collection

Network studies focusing on different levels of analysis of interpersonal communication face different challenges, especially in terms of whether or not complete data is needed. The question that must be asked is whether all actors and relationships need to be identified and measured. Analysis at the macro level often calls for the complete enumeration of a closed population whereas analysis at the micro level is more concerned with complete enumeration of the local network surrounding “ego” or the local neighborhood of the actors or the nodes. Problems of boundary specification are evident at all levels of communication analysis, as omission of important social ties can lead to misleading results. Boundaries can be defined occupationally, geographically, or relationally (distance from the egocentric actor).

In the light of these challenges, we discuss several different data collection strategies for social networks, particularly focused on interaction or interpersonal communication. A detailed analysis of these methods, in the scope of prior work, as well in relevance to the thesis is discussed below:

1. *Questionnaires.* Questionnaires or surveys have been used by sociologists since several decades to understand behavioral characteristics of individuals as well as groups. Typically they contain questions, both objective and subjective in nature dealing with the nature of the problem under consideration. These self-report questions thus, often capture the presence or absence of a social tie. Specifically, name generators are used to discover who belong to the respondent’s network; additional information is then collected about these individuals.
2. *Direct Observation.* Anecdotal observations concerning case studies of interest to a problem domain have also been used extensively in the past to discover

network structure and associated interaction among the individuals.

3. *Written Records: archival or diary.* Also common is the use of archival data sources. Archives are useful in obtaining information about the past or about otherwise inaccessible networks. Archival methods can be used independently or can complement survey research.
4. *Self-reported Responses.* Several sociological studies have also relied on self-reported responses on the nature of social ties of individuals. However the problem with self-reported responses is that different kinds of relationships may exhibit different patterns and not be an indication of an underlying link. So while respondents might not know with whom they talk over any given period of time, they do know with whom they typically speak and what types of exchanges they engage in. Other methodological advances have also shown that rates of reciprocation among respondents are high enough to assume that self-reports contain more than mere perceptions.
5. *Derivation.* A method that has started to become increasingly popular to the social researchers of late is accumulation of electronic communication data, and their subsequent deployment in the derivation of network structures. Electronic social data can be collected at comparatively low cost of acquisition and resource maintenance, can span over diverse populations and be acquired over extended time periods. As a result, today we are able to study social processes involving interpersonal communication at unprecedented scales. Examples include websites such as Twitter, Facebook, Digg, email networks, citation and blog networks. Typically collection of electronic data is undertaken based on three different methods:

- *Application Programming Interfaces.* Many of the online social websites such as Twitter and Digg feature publicly accessible developer APIs (or Application Programming Interfaces) that allow collection of network data, including communication activity of the users. E.g. the Twitter API allows developers to collect data about the social graph of the users

i.e. “followers” and “followings”, the “tweets”, including their timestamp of posting, content and the owner of the tweet.

- *HTML Scraping / Parsing.* Publicly accessible social profiles or blog websites can often be crawled based on scraping or parsing the associated HTML, using one of the several publicly available open source parser libraries, e.g. Beautiful Soap with Python.
- *Gardenhose or firehose.* This method involves collection of complete social data involving interaction as well as social actions based on real-time data creation and storage as well as new data gets available. This technique is relatively less common.

2.5 Overview of Datasets

In this thesis we have primarily focused on data collection based on electronic data accumulated from different social websites. We present a description of the different data sources in Table 2.5.

Further, a summary of the data collection techniques for different datasets used in this thesis is presented in Table 2.5. The table also presents the different attributes that were collected for each dataset.

As pointed out earlier, wherever applicable, we apply all of these datasets in our empirical studies of each chapter. However each dataset does not exhibit the necessary characteristics or attributes required to analyze or evaluate our proposed methods. Hence in Table 2.3 we provide a list of the different problems where each dataset has been used. We discuss briefly the rationale behind use of different datasets in different problem contexts:

1. *Topical Communication Flow (Chapter 3).* In this problem we characterize communication between pairs of individuals via a variety of contextual factors, that are predictive of information flow. The notion of communication in this work is the exchange of messages between individuals in a social network, that are likely to be visible to others. Hence the datasets that are applicable

in this context are MySpace, Twitter, YouTube and Flickr, where we observe directed communication in the form of messages, tweets and comments among individuals, either independently, or around a shared media (e.g., an image or video). Social media sites Digg and blog networks like Engadget and Huffington Post are not usable in this case because they do not have an explicit notion of a “social network” due to which the communication of others can be observable.

2. *Diffusion of User Actions—Social Influence (Chapter 4)*. In this problem we develop a computational framework to model the impact of social influence (due to network) on diffusion of information. We consider the diffusion process to occur via a particular social action that is executed by the user. In this context, therefore, the datasets that are applicable are ones in which there is an observed action. Hence we use Digg (where the action is digging a URL) and Twitter (where the action is the act of re-tweeting a post made by another author). Note that because of the assumption of diffusion centered around a social action, datasets such as MySpace, blog networks like Engadget and rich media like YouTube are not directly usable—they do not have a notion of an observable social action that can be associated with the diffusion process.
3. *Diffusion of User Actions—Homophily (Chapter 5)*. This problem also investigates diffusion of information via user actions, but from the perspective of attribute homophily among users. For the same reason as above, we use Digg and Twitter as the datasets; other datasets do not seem to feature an observed social action around which we can study the diffusion process.
4. *Community Dynamics and External Phenomena (Chapter 6)*. This problem models various characteristics of online communities typically featured over blogs. We are interested to study the utility of the communication characteristics of individuals in these communities in correlating with exogenous temporal events. Hence we use blog datasets Engadget and Huffington Post, because they feature focused technological and political discourse, which can

be conveniently correlated with stock market variable or election related polls. The reason behind not using other datasets is because of the lack of availability of a correlating dataset which can demonstrate the utility of the extracted characteristics of the communities.

5. *Prototypical Communication Groups (Chapter 7)*. In this problem our goal is to identify representative prototypical groups in communication sets in the blogosphere. We again use external temporal time series variables to validate the utility of these prototypical groups of communication activity. Hence as above, we use the two blog datasets Engadget and Huffington Post.
6. *Interestingness of Conversations (Chapter 8)*. Our goal in this chapter is to model the dynamics properties of social media conversations, particularly its interestingness. Hence we are interested in datasets which feature commentary around a shared media object. For the purpose we use the two rich media datasets, YouTube and Flickr (commentary around videos, images) and the two blog datasets Engadget and Hunffington post (commentary around blog posts). Note that the other datasets do not explicitly feature communication activity around a shared media, hence are not directly applicable in this context.
7. *Quantifying Rich Media Collective Participation (Chapter 9)*. In this problem we study the factors behind participation of individuals in social media conversations. Based on the same reasons as above, we focus on the two rich media datasets, YouTube and Flickr (commentary around videos, images) and the two blog datasets Engadget and Hunffington post (commentary around blog posts).

Table 2.1: Datasets and their descriptions used in this thesis.

DATASET	DESCRIPTION
MySpace	A social networking site that allows users to add contacts share content, in the form of multimedia, blurbs and blogs, update moods in the form of emoticons as well as send each other messages, usually publicly viewable. Until 2008, it was the largest social networking website.
Digg	Digg is a social news website made for people to discover and share content from anywhere on the Internet, by submitting links and stories, and voting and commenting on submitted links and stories. Voting stories up and down is the site’s cornerstone function, respectively called digging and burying.
Twitter	Twitter is a free social networking and microblogging service that enables its users to send and read messages known as tweets. Tweets are text-based posts of up to 140 characters displayed on the author’s profile page and delivered to the author’s subscribers who are known as followers. Senders can restrict delivery to those in their circle of friends or, by default, allow open access. Users can also follow lists of authors instead of following individual authors.
Engadget	Engadget is a multilingual technology weblog and podcast about consumer electronics. Engadget currently has nine separate websites, all operating simultaneously with each having its own staff, which cover technology news in different parts of the world in their respective languages. The blog is updated multiple times a day with articles on gadgets and consumer electronics. To extend readership, the blog is available in several languages including Spanish, Japanese, German, Polish (until 1.04.2010), Korean, and Chinese (traditional and simplified). Engadget uses the Blogsmith CMS to publish its content.
Huff Post	The Huffington Post is an American news website and aggregated blog featuring various news sources and columnists. The site offers coverage of politics, media, business, entertainment, living, style, the green movement, world news, and comedy, and is a top destination for news, blogs, and original content. The Huffington Post has an active community, with over one million comments made on the site each month.
YouTube	YouTube is a video sharing website on which users can upload and share videos. Most of the content on YouTube has been uploaded by individuals, although media corporations including CBS, the BBC, UMG and other organizations offer some of their material via the site, as part of the YouTube partnership program. Unregistered users may watch videos, and registered users may upload an unlimited number of videos.
Flickr	Flickr is an image and video hosting website, web services suite, and online community. In addition to being a popular website for users to share and embed personal photographs, the service is widely used by bloggers to host images that they embed in blogs and social media.

Table 2.2: Data collection methods used in the thesis.

DATASET	DATA COLLECTION TECHNIQUE	ATTRIBUTES COLLECTED
MySpace	We constructed an HTML parser to snowball from the super-user Tom and crawl his contacts.	User information, messages, timestamps of message postings, contact list.
Digg	We used the publicly available Digg API to collect data relating to the digging and commenting activity of the users, incrementally over a period of time. The crawl was seeded from the set of news stories featured in the “Popular” category on Digg.	User information, social graph i.e. contact network, digging activity on news stories, comments, replies, timestamps of digging, commenting and reply activity, time of posting of news stories as well as their topical characterization given by Digg.
Twitter	The Twitter open-source API was used in a snowball crawler, that seeded from a focused set of authoritative or “celebrity” users. We did a breadth-first data collection based on the “following” links of the users.	User information such as location, time-zone, social graph i.e. “followers” and “followings” of a given user, content and timestamp of tweets of each user.
Engadget	We created an HTML parser based crawler that used the “search” feature on the website to collect information related to the blog posts featuring technology companies, particularly we focused on Apple, Microsoft, Google and Nokia.	Blog post information such as author, time of posting, tags, comments and replies, including their content, time of posting and author of the comment or reply.
Huffington Post	An HTML parser based crawler was used again in this case that collected blog posts and their associated information based on a search on the names of different political candidates, e.g. Clinton, Obama and McCain.	Blog post information such as author, time of posting, tags, comments and replies, including their content, time of posting and author of the comment or reply.
YouTube	We used the YouTube API to collect data regarding the videos and their associated conversations. We seeded the crawl from the popular videos in different topical categories.	Video information such as tags, time of upload, owner, comments and replies, including their content and authors.
Flickr	The Flickr API was used to collect information regarding the photos; while an HTML crawler was used to download the photos themselves as jpeg images. The crawling was seeded from the photos featured as “interesting” everyday on Flickr.	Photo-related information such as tags, notes, owner, time of upload, group information, set information, content, author and timestamp of comments.

Table 2.3: Thesis overview of different research directions and problems.

Dataset	Problems
MYSPACE	Topical Communication Flow (Chapter 3)
DIGG	Diffusion of User Actions (Chapter 4, 5), Community Dynamics and External Phenomena (Chapter 6), Prototypical Communication Groups (Chapter 7)
TWITTER	Topical Communication Flow (Chapter 3), Diffusion of User Actions (Chapter 4, 5), Community Dynamics and External Phenomena (Chapter 6), Prototypical Communication Groups (Chapter 7)
ENGADGET	Community Dynamics and External Phenomena (Chapter 6), Prototypical Communication Groups (Chapter 7), Interestingness of Conversations (Chapter 8), Quantifying Rich Media Collective Participation (Chapter 9)
HUFFINGTON POST	Community Dynamics and External Phenomena (Chapter 6), Prototypical Communication Groups (Chapter 7), Interestingness of Conversations (Chapter 8), Quantifying Rich Media Collective Participation (Chapter 9)
YOUTUBE	Topical Communication Flow (Chapter 3), Prototypical Communication Groups (Chapter 7), Interestingness of Conversations (Chapter 8), Quantifying Rich Media Collective Participation (Chapter 9)
FLICKR	Topical Communication Flow (Chapter 3), Prototypical Communication Groups (Chapter 7), Interestingness of Conversations (Chapter 8), Quantifying Rich Media Collective Participation (Chapter 9)

Part I

Information Diffusion

“At the same time, as social beings, we are moved in the relations with our fellow beings by such feelings as sympathy, pride, hate, need for power, pity, and so on.”—Albert Einstein (1879–1955), theoretical physicist, philosopher.

Human communication activity involves mutual exchange of information. Social interaction among a set of individuals can be modified, refined, shared or amplified based on the information that flows from one individual to another. Thus, over several decades, the structure of social groups, society in general and the relationships among individuals in these societies have been shaped to a great extent by the flow of information. Diffusion is the process by which a piece of information, an idea or an innovation flows through certain communication channels over time among the individuals in a social system.

Today, the pervasive use of online social media, including Facebook, Twitter and MySpace has made the cost in propagating a piece of information to a large audience extremely negligible. As a result there is extensive evidence of large-scale social contagion in these sites. To understand the extent to which ideas are adopted on different online social platforms, it is important to understand how the dynamics of adoption in the underlying social network. Understanding dynamics include: the extent to which people are likely to be affected by decisions of their friends, or the extent to which “word-of-mouth” effects communication.

Analyzing the flow of information can be useful in understanding diffusion of medical and technological innovations, the sudden and widespread adoption of various strategies in game-theoretic settings as well as the effects of viral marketing. My interest lies in the study of such large-scale information propagation via a certain mode of communication or an observable social action among individuals. There are two directions to this research perspective discussed in the following three chapters. First we develop some formal ways to characterize the communication process between a pair of individuals, that is often responsible for the diffusion of information over time in a network. Next we investigate developing computational frameworks that can model the interactional process on a social setting, given

a particular communication modality². The frameworks are used to understand how the characteristics of the network (i.e. social influence) or the characteristics of the individual (i.e. homophily) impacts the social phenomenon of diffusion. We discuss a motivating real-world application to illustrate the utility of this research.

Motivating Application. What are the interesting memes on different social media sites post the release of the iPad by Apple?

Let us suppose that Alice is interested in identifying and gathering comments from the public related to being conduits for the recent release of the Apple iPad tablet computer. Social media sites are a viable option for Alice is to track the flow or diffusion of the iPad meme on different popular content-publishing sites. Twitter is a platform where she could track the meme. However several technical challenges might unfold. For example we would first need to address how to characterize the communication from the individuals who are posting content about the iPad. Next, our goal would be to model the diffusion of the concept, iPad via the social actions of the individuals. Furthermore, we would need to identify the sets of individuals who are socially influenced by or homophilous to Alice, along particular sets of attributes, to ensure that the information dissipated aligns with her interest in the concept i.e. the release of iPad.

We summarize our goals in the following three chapters via the questions discussed below.

1. How does the communication process influence the flow of topical information between a pair of individuals in a network? (Chapter 3)
2. How do different communication modalities, such as social actions, influence information diffusion? How does social influence relate to the diffusion process? How does attribute homophily relate to the diffusion process? (Chapter 4, 5)

²Note that the modality we assume here does not need to be strictly textual interaction e.g. messages or comments, but can actually be some form of social action, usually public in nature. Examples include digging activity on Digg, tweeting on Twitter or favoriting a video or photo on YouTube or Flickr respectively.

Chapter 3

Characterizing Communication: Topical Communication Flow

“The Internet is the biggest psychological and social human experience; we make encouraging viral activity.”—Cynthia Gordon, vice president of new media marketing for Universal Orlando Resort.

This chapter presents a computational framework for predicting efficient communication paths in social networks, leveraging knowledge about a user’s communication and social context. Understanding communication flow is useful to understand how information, innovations, trends and ideas propagate in social networks; as well as to facilitate targeted advertising in marketing domains.

We identify two parameters to characterize communication flow between a given pair of users: the intent to communicate and the associated delay in communication. Our method has three parts. First we identify communication context as the attributes of current communication, such as the neighborhood of a user, the topic under consideration and the relationship of the recipient to the user. Second, we characterize social context of a user via two features: information roles in communication (e.g. generator, mediator or consumer of information), and the nature of the social tie shared between the particular pair of users. Third, we propose a temporally evolving Support Vector Regression framework to predict the intent and the delay for a pair of users directly connected in the network. We further propose a generalized method to compute the intent and delay for cases when the pair is arbitrarily n hops distant in the network, by using an approximation of the restricted shortest path problem.

We have excellent results ($\sim 10\text{-}15\%$ prediction error, and $\sim 40\text{-}50\%$ boost over baseline techniques) on four large datasets—MySpace, Twitter, YouTube and

Flickr. We observe interestingly that while context can reasonably predict intent, delay seems to be more dependent on evolving user behavior and other latent factors, e.g. ‘age’ and popularity of the information content under consideration. Finally we present how our results lend insights into the structural and temporal aspects of information diffusion; such as the effect of network connectivity, social distance and the rate of diffusion over time.

3.1 Introduction

In this chapter, we develop a computational model for dynamically predicting efficient paths for communication flow in large-scale social networks that leverages communication and social context. Modeling propagation of information in social systems subject to interpersonal communication, has interested social network theorists since several decades (Domingos and Richardson, 2001; Hammer, 1980; Newcomb, 1961; Newman, 2002; Janssen and Jager, 2001; Valente, 1996; Zachary, 1977). Historically such interest has drawn in observations from sociological and physical phenomena, like epidemic disease propagation, colliding gas molecules, percolation theories and so on; while more recently they have found widespread usage in facilitating targeted advertising in the marketing domains.

Note, however, most of these traditional approaches to modeling diffusion, have been focused on longitudinal studies of how a certain genre of information flows among sets of individuals. The studies were mostly based on surveys, questionnaires, analysis of citation networks, simulation of networks of agents or even via networks of movie actors or power-grid systems across the United States. However, over the past few years, the overwhelming quantity of data available via the different Web 2.0 social websites, including email clients, instant messengers, and social media, while on one hand has made such tedious empirical studies on focused communities unnecessary, at the same time, has presented several challenges in the studies related to information diffusion over short and extended periods of time.

To understand these challenges, let us discuss the different popular approaches today that are focused on understanding information diffusion. Over the

years, information diffusion has primarily been tackled with two different schools of approaches - one that interested the social scientists, and the other that interested the economists. The ‘social’ approach primarily dealt with exploiting the network topology between individuals to understand the diffusion of information, without any consideration of their individual habits or behavior (i.e. focus on structure with respect to an individual (Watts, 2003)). Whereas the ‘economist’ approach relied on the personal attributes of individuals solely to understand their adoption behavior of products, irrespective of their local social network topology (i.e. focus on agency with respect to an individual (Watts, 2003)).

Nevertheless, though reasonably successful in their domains, we conjecture that both these approaches presumably have limitations. For example, on the different online social media today, answering questions, such as: if Alice has a party at her house, who is the person she needs to contact to invite as many people; or a different question such as, if Alice wants to buy a car, who is the person best suitable for reliable suggestions - thus, often needs going beyond the scope of either of these approaches and incorporate both knowledge from both the structure and agency of the individual in question. Hence it appears that the increasing attention to massive data collection from the different social sites has undermined a seemingly grave issue, that is of analyzing information flow under factors that affect the context of the user at hand, i.e. her past communication activity, relationship with her contacts or her general social behavior of handling different kinds of information - factors that subsume the user’s structure and agency, both. We therefore propose that there is a need to revisit the state-of-the-art models of information flow or diffusion.

This necessity to revisit also spring from the observation that all these social networks have evolved into complex network structures rich in a wide array of interactions among people (e.g. often having “multiplex” ties) unlike traditional “one-to-one modes of communication”, that were typically private in nature (e.g. emails or instant messages). For example, today individuals can develop interactional relationship between them due to a friend nomination on Facebook¹, due to

¹<http://www.facebook.com>

co-participation on commentary on the same YouTube² video, due to membership on the same group on Flickr³ or due to re-tweeting the same piece of information on Twitter⁴. Moreover these social sites, like MySpace⁵, Facebook, YouTube, Flickr, Twitter, Digg⁶ etc, with their multi-faceted interactional media, further allow the generation and evolution of tremendous amount of information on the web - and via the ‘friend-of-friend’ nature of relationships between users, increase their vulnerability to acquire, modify and transmit them in various different ways, subject to their contextual attributes at hand.

The problem turns more challenging when we encounter very strong evidence that these online social networks also promote easy information sharing among the users, due to the presence of a rather public platform. For example, these social spaces let users add others as contacts which provide them with an opportunity to visit their profiles and leave messages. Often these messages acquire conversational forms—where to and fro information exchange goes on between the pair of individuals involved. The ability to view a friend’s public messages sometimes draws in others in the conversation—especially their common friends. As a result we observe artifacts of group conversation, where we notice that the information has diffused in the neighborhood of the originators of the communication. The common friends might participate in communication with the originators, as well as might choose to talk about a similar topic to their other friends. Clearly, we observe flow of information via these means of communication.

Hence, the collective outcome of such overwhelming information generation, sharing and associated public interaction is not trivial to decipher with the current state-of-the-art tools. In this chapter, therefore, we attempt to model the flow of communication itself, among a pair of users, under the context of the new forms interaction that have come to emerge in these online social networks. Specifically, we define communication flow (or generically information diffusion) as the process

²<http://www.youtube.com>

³<http://www.flickr.com>

⁴<http://www.twitter.com>

⁵<http://www.myspace.com>

⁶<http://www.digg.com>

by which messages are transmitted through certain channels (acted upon by a set of relationships among people) over time. The actors in the process are the members of the social system under consideration.

Motivating Applications

The prediction of communication flow can be deemed important in several domains; such as determining information propagation through social networks, or to help organizations determine their experts or knowledge points, monitor the dynamics of effective information paths, as well as to facilitate understanding community evolution in social networks. To take an example, corporations are often interested in determining the people, who act as information hubs, and who need to be targeted for advertising their products. They are also interested to understand the structure in which a particular genre of information flows in the local network of individuals via their online interactions. Besides, information flow has also found wide spread applications in the marketing domain, diffusion of innovations and evolution of cooperation among agents in social sciences.

In our attempt to address communication flow in this chapter, our main intuition comes from the observation that *context is the dynamic subset of knowledge that is in attention and it influences the exchange of messages between the entities in communication* (Mani and Sundaram, 2007), specifically here in transmitting messages from one user to another. To elaborate more concretely, communication on the above-stated different social sites is usually promoted in the form of a digital record, for example, a shared link on Facebook, a “re-tweeted” item (e.g. a bit.ly URL) on Twitter, messages on MySpace⁷ or a vote on a YouTube comment and is heavily and closely intertwined with the topic at hand, the past behavior in communication with respect to the topic, types of ties of a specific user or her communication habits, in the form of different information roles. This reflects that the increasing trend of fostering rich communication on these sites is often subject to a variety of contextual artifacts, and hence studying the flow of communication

⁷<http://www.myspace.com>

needs to be driven from an analysis of the user context itself.

Our Approach

Now, we present our approach to predict efficient paths for communication flow in social networks. We develop a temporally evolving representation framework for context that can efficiently predict communication flow in social networks between a given pair of users, who can be directly connected in the network, or can be connected through an arbitrarily long path. We describe communication flow by two parameters: the intent to communicate and communication delay. There are several key contributions made in our chapter as follows:

1. Use of a comprehensive set of features belonging to communication context and social context for determining communication flow,
2. A dynamic feature selection framework that can determine the optimal set of features at any given time,
3. A Support Vector Regression based online learning framework to predict the intent to communicate and delay between a pair of users, who are connected directly at a certain time interval, and using the set of dynamically selected optimal features,
4. An approximation algorithm based on the restricted shortest path problem (Lorenz and Raz, 1999), to predict the intent to communicate and delay between a pair of user when they are connected arbitrarily through a path in a social network having n number of users between them.

We are motivated by a Physics based wave front metaphor in our understanding of communication flow. This allows us to assume conditional independence in communication from a user's contacts given a topic. We identify four different aspects that affect communication on a specific topic: (a) neighborhood context, (b) topic context, (c) recipient context and (d) social context.

Neighborhood context refers to the effect of the user’s social network on her communication. It is affected by the number of messages by the user’s contacts on the topic and the communication in the local neighborhood on the topic. Topic context refers to the effect of the semantics of a user’s past communication on a topic on her future communication on the same topic. Recipient context refers to effect of the recipient identity on the user’s intent to communicate. Since the first three aspects relate to a user’s communication with her local neighborhood (i.e. the set of contacts of the user), we call them to be part of the communication context. Modeling communication context is motivated from prior work (De Choudhury et al., 2007) which emphasizes that flow of communication is dependent on the vulnerability of the users to new content. We also identify social context (De Choudhury et al., 2008a), which refers to the patterns of participation in communication (information roles) and the degree of overlap of friends between two people (strength of ties). Modeling social context is motivated by the observation that there is empirical evidence that the flow of information is often found to be characterized by temporal properties of propagation resulting from the interactions among a set of people, who can be considered connectors or influencers.

In order to determine communication flow between a pair of users based on the above features, a subset of optimal features is chosen at a given time instant based on ‘feature voting’ using different standardized feature selection strategies. For a pair of users connected directly in the network, the intent to communicate and communication delay are estimated using Support Vector Regression (SVR) based on the selected optimal features at a certain time interval. when the pair of users are connected through a path of length n hops, we transform our prediction problem into the restricted shortest path problem (Lorenz and Raz, 1999) with two parameters, the intent and the delay. Our problem is therefore to find an optimal path between the given pair of users for which we have maximum intent to communicate across the path as well as minimum delay in communication. This dual optimization problem is solved considering a graphical representation of the social network where each edge between two nodes comprises two parameters: the intent to communicate and

associated delay. The restricted shortest path problem being NP-complete, we use a polynomial time approximation scheme (PTAS) (Arora, 1998) which yields us an optimal communication path in polynomial time.

To evaluate the predicted intent and delay determined by this model for both the network topologies, it is compared against (a) a baseline model based on frequency counts of communication participation for the pair of users concerned, (b) a trend learning based method, (c) an information diffusion model based on spread of infectious diseases ((Gruhl et al., 2004)) and (d) effect of local neighborhood of the user based method.

We have excellent results on four large datasets—MySpace, Twitter, YouTube and Flickr. For the directly connected topology, first, we computed the intent to communicate and delay on several topics for a specific user and her network. Secondly, we determined the intent to communicate as well as the communication delay for the same set of topics, with varying network sizes, averaged over all contacts. Results show that SVR out performs the baseline techniques. Similarly, in the case of prediction for users connected by a path of length n , we have satisfactory results on our dataset (10-20% error in prediction) for the intent and delay. We also analyze the impact of the different features of communication and social context on communication flow and to observe the temporal dynamics of their significance in prediction. Finally we observed qualitatively that intent is more affected due to contextual dynamics than delay. Delay seemed to be more dependent on other latent factors characterizing communication, e.g. ‘age’ of the information content under consideration or the evolution of user behavior over time.

3.2 Related Work

In this section we discuss prior work. The analysis of communication dynamics and its role in information propagation in social networks has been of interest to researchers from various domains ranging from social sciences, epidemiology, disease propagation, physics and economics (Domingos and Richardson, 2001; Newman, 2002; Janssen and Jager, 2001; Valente, 1996). It has found wide-spread applications

in the diffusion of medical and technological innovations, cultural bias, patterns of adoption of strategies in game-theoretic settings, modeling market dynamics in consumer markets, understanding information roles of users, emergence of ‘chatter’ and ‘spikes’ in discussions, and disease and computer virus transmission among a set of people. We organize an elaborate review of the prior work in this chapter in the following sections, organized into three major complementary but inter-related ideas and relevant to the scope of work in this chapter—information diffusion, analysis of communication in social networks and analyzing user social behavior via their information roles. Finally we present the major limitations of prior work.

Information Diffusion

Understanding the diffusion of information in social networks is one of the oldest topics of interest to researchers. There has been prior work on modeling and predicting pathways of diffusion of information in social networks useful for several applications, ranging from recommendation systems, online advertising, user behavior prediction and disease containment (Gruhl et al., 2004; Kempe et al., 2003; Song et al., 2006; Stewart et al., 2007; Wan and Yang, 2007). In an early work (Kempe et al., 2003), Kempe et al propose solution to the optimization problem of selecting the most influential nodes in a social network which could trigger a large cascade of further adoptions. They use sub-modular functions and a greedy strategy to yield approximate solutions that out-perform node-selection heuristics based on the popular notions of degree centrality and distance centrality.

In (Gruhl et al., 2004) the authors focus on analyzing the text in blog posts and use an epidemic disease propagation model for determining information diffusion. The authors focus on the propagation of topics from one blog to the next in the blogosphere, based on the textual content instead of analysis of the hyperlinking structure. Using this information, they have characterized information diffusion along two dimensions: topics and users.

Adar and Adamic in (Adamic and Adar, 2005) utilize a novel inference scheme that takes advantage of data describing historical, repeating patterns of

'infection'. They present a visualization system that allows for the graphical tracking of such information flow. In (Song et al., 2006), the authors present an early adoption based information flow model useful for recommendation systems. The authors in (Newman, 2001) provide simple models for the onset of epidemic behavior in diseases. Their results derived exact analytic expressions for the percolation threshold on one-dimensional small-world graphs under both site and bond percolation. They have also looked briefly at the case of simultaneous site and bond percolation, in which both susceptibility and transmissibility can take arbitrary values.

In (Gonzalez and Herrmann, 2006), the authors propose a novel model for networks of complex interactions, based on a granular system of mobile agents whose collision dynamics is governed by an efficient event-driven algorithm and generate the links (contacts) between agents. There has also been some prior work (Delre and Janssen, 2007) where the authors explain how diffusion dynamics in small world networks are affected with having heterogeneous consumers. Stewart et al in (Stewart et al., 2007) propose an algorithm to discover information diffusion paths in the blogosphere for helping online advertising domains. They present a 'frequent pattern mining' based method in which they focus on analyzing blog content and topic extraction. Using sequences of blogs, they discover information diffusion paths which are useful for effective information flow in social networks.

In (Wan and Yang, 2007), Wan and Yang define information diffusion to be the phenomenon of document forwarding or transmission between various web sites on the Web. They propose a method for mining information diffusion processes for specific topics on the Web and develop a system called LIDPW to address this problem using matching learning techniques. Saito et al in (Saito and Kimura, 2008) utilize the independent cascade (IC) model to determine the likelihood for information diffusion episodes, where an episode is defined as a sequence of newly active nodes. Thereafter they present a method for predicting diffusion probabilities by using the popular EM algorithm.

In a recent work, Bakshy et al (Bakshy et al., 2009) study how "gestures" make their way through an online community - the social gaming environment called

Second Life. Gestures are code snippets that Second Life avatars must acquire in order to make motions such as dancing, waving or chanting. Their empirical studies indicate that individuals who have already declared each other as friends are more prone to getting influenced by each other and subsequently acquiring assets, rather than two individuals arbitrarily apart. In another recent work, Sun, Rosenn, Marlow et al (Sun et al., 2009) study the diffusion patterns on the Facebook “News Feed” and conclude that in online social media, diffusion dynamics are often triggered by the collision of short chains of information trigger, rather than the preconceived notion in the literature that diffusion occurs due to several long chains generated by a small number of “seeds”.

Analysis of Communication

Analyzing interpersonal communication has interested social scientists since several decades due to their potential to demonstrate social phenomena. Traditionally social network data has been collected from cross-sectional studies of relatively small groups such as participant observation (Hammer, 1980; Zachary, 1977) and surveys (Newcomb, 1961). However today with multiple Web based communication mechanisms, such as emails, instant messaging, online games and so on, collection of communication data and its empirical and computational analysis over extended periods of times has become easier than before (Anagnostopoulos et al., 2008; Bakshy et al., 2009; Leskovec and Horvitz, 2008; Sun et al., 2009; Eagle et al., 2009). The result is that network analysis on a scale that would have been inconceivable scarcely a decade ago—in the order of 100M nodes—is becoming routine (Kumar et al., 2006; Leskovec and Horvitz, 2008).

There has also been more recent work on modeling communication properties (Kumar et al., 2003; Li et al., 2007) in online social networks and their correlation with external events (Antweiler and Frank, 2004; Gruhl et al., 2005). The authors in (Agrawal et al., 2003) focus on exploiting the link information in computations over hypertext corpora and conjecture that links arising out of social interactions are more valuable than text based analysis. They test their computational model

in a setting by ignoring the text and relying solely on links.

Antweiler et al. in (Antweiler and Frank, 2004) determine correlations between communication activity in Internet message boards and stock volatility and trading volume. Gruhl et al in (Gruhl et al., 2005) correlate postings in blogs, media, and web pages with the sales ranks of books on Amazon.com. They devise carefully hand-crafted queries to find matching posts which can be indicators of future sales rank of books. They also provide a method which could automate the generation of such queries. Finally they present algorithmic predictors which can use online postings and successfully predict spikes in sales rank.

In (Bird et al., 2006), the authors analyze emails of software developers, to understand the relationship between the email activities and the software roles. In another work related to analysis of communication properties in social networks, Mei et al. in (Mei et al., 2007) present a Topic-Sentiment Mixture (TSM) model that can reveal the latent topical facets in a Weblog collection, the subtopics with respect to a certain given query, and the associated sentiments reference to the query. It is also able to provide general sentiment models that are applicable to any arbitrary topic.

In a previous work Gomez et al in (Gómez et al., 2008) analyze several social network properties emerging from the communication activity on the website Slashdot. They study the structure of discussion threads using a radial tree representation. The findings show that nesting of conversations on Slashdot exhibit strong heterogeneity and self-similarity. Benevenuto et al in (Benevenuto et al., 2008) present a method for understanding user behavior in a social network created essentially by video interactions by characterizing different properties of the popular social media YouTube. In (Zhou et al., 2008), Zhou et al attempt to predict the tendency of topic discussion on online social networks using a dynamic probability model. We analyze the process of topic discussion, using three main factors - individual interest, group behavior, and time lapse.

Characterizing individuals in social networks has been of interest in different domains. In an early work, Domingos and Richardson propose a method to determine potential customers suitable for marketing products (Domingos and Richardson, 2001). They develop a measure called ‘network value’ of a customer and model the social network as a Markov random field to compute this measure per user. There has also been some work on identifying important bloggers based on communication activity in (Nakajima et al., 2005). The authors attempt to determine ‘hot’ conversations in the blogosphere through agitators and summarizers by establishing discriminants. Soroka et al in (Soroka and Rafaeli, 2006) analyze participation of users in communication activities in a social networks to understand their lurking or de-lurking behavior. They define lurking as passive attention over ongoing communication activity. Such characterization of users shows that fostering receptive participation may be important and constructive to encourage active contributions in online communities.

In (Song et al., 2007) the authors identify opinion leaders who are responsible for disseminating important information to the blog network using a variation of the PageRank algorithm. Ehrlich et al in (Ehrlich et al., 2007) are motivated to characterize individuals in expertise locator systems which should reflect the social contexts in which people are embedded as well as facilitate the path to conversation. They therefore develop a system to find experts in enterprises based on textual information and social network characteristics.

In (Watts and Dodds, 2007), Watts and Dodds study the ‘influentials hypothesis’ in the context of network diffusion and formation of public opinion. Individuals are characterized to be influential if they comprise a small set who influence an exceptional number of their peers. Extensive simulation in their work shows that large-scale information cascades are formed not due to influential individuals, but due to a set of easily susceptible ones.

Goyal et al in (Goyal et al., 2008) develop a pattern mining approach to discover leaders and tribes in social networks. They consider a social network in

which users can perform a set of actions over shared items and can easily be motivated to mimic others. Under such context, the authors study the propagation of influence, via detecting the leaders in the social network. Saito and Kimura et al in (Saito and Kimura, 2008) seek to identify influential nodes involved in diffusion of information in social networks based on the independent cascade model. They explore what structural features of the underlying network are relevant to a greedy solution of the independent cascade model.

Limitations

We observe that prior research of detecting information flow has been focused mainly on two aspects: (a) implicit assumption of presence of a (virtual) social network through which people can exchange information, and (b) disregard to the local or neighborhood information, information content or the points of information generation and reception. There are several limitations of the prior work:

1. The model of propagation has been based on static knowledge about people's probability to transmit information. But the nature and degree of propagation are contingent upon the micro and macroscopic relationships between people engaged in the process of transmission. Contextual information such as the local network topology of the sender and receiver, relationship of the topic of communication with past communication, relationship shared between two people as well as identities of the sender and the recipient, has not incorporated.
2. In web based analysis, the flow is estimated from indirect evidence (e.g. a topic appears on a blog several days after it appeared on another blog), not from evidence of direct communication. Thus the effect of contextual factors on communication is not easy to determine.
3. A static set of factors (e.g. psychology of people in buying products or providing recommendations) have been considered in the prior work. However,

these factors are not static, rather change over time. Hence the factors that govern information flow are likely to be a function of time.

4. Communication can be considered as a medium for exchange of information. We describe the interactions between people by two orthogonal yet complementary aspects: media and action. Every interaction involving two people thus comprises a medium for the propagation of information (e.g. emails, messages, images etc) and an associated action that embodies that interaction (e.g. writing blog posts, adoption of consumer goods etc). The prior work on information diffusion has focused on the actions undertaken by people. But the action can be attributed differently under different context of the media. Hence a framework for prediction of diffusion should model the context in which the information is propagated from one person to another - analysis of actions and intrinsic behavior of people.

In this chapter we address the problem of predicting the communication flow between a pair of users based on a comprehensive analysis of the dynamic factors that impact communication—communication context and social context that subsume both the current context of the users at hand, as well as their long-term behavioral and structural characteristics.

3.3 Problem Definition

In this section we present the problem definition. First we discuss the role of different contextual factors in predicting communication flow between a pair of users. Next we present our data model. Finally we discuss the problem statement and the key challenges involved in this problem.

The Role of Context

In this section we discuss the types of context and how it is useful to predict communication flow. We define two categories of context - communication and social context (Mani and Sundaram, 2007). These are the set of attributes, dependent

on the local network topology of the two communicators, their behavioral roles as well as social relationship that affect communication between them. From psychological studies (Mani and Sundaram, 2007), we know that when a user u sends messages to v on a certain topic, the messages are influenced by u 's current short-term memory, e.g. attributes of the current context: e.g. the topic at hand, the messages exchanged as well as the response received from u 's local social network. Contextual attributes are application dependent and are dynamic. In the application considered in this chapter, we identify a twofold context representation that affect communication—communication context and social context. Communication context is further categorized as (a) neighborhood context, (b) topic context and (c) recipient context; while social context has two features: information roles of people, and (b) strength of ties.

Let us assume that u plans to send a message to her contacts about a topic (e.g. movies) at a certain point in certain time. Neighborhood context refers to the effect of u 's contact network on her planned communication. Topic context refers to the effect of u 's past communication about the topic 'movies' on her planned communication about the same topic. Recipient context refers to effect of the recipient identity (i.e. the specific persons with whom u communicates) on the intent to communicate. And social context incorporates the role acquired by u due to her communication, and the nature of the relationship that she shares with her contacts. Based on these ideas, we now present our data model.

Data Model

Our data model is based on the world's largest social networking site MySpace. This is a useful dataset because the users of MySpace involve in diverse communication activities over time, subject to a rich set of contextual attributes. We construct a communication graph based on these users and their contact lists in MySpace. The communication graph is described as follows.

Let $G(V, E)$ be a weighted directed graph with $|V| = n$ and $|E| = m$ (V is the set of users and E is the set of edges based on their contact list, and a real-

Figure 3.1: Schematic representation of the data model used for predicting communication flow. A social network is represented as a weighted, multi-edge directed communication graph where an edge implies communication between the two connected users. In this figure, the thickness of an edge implies its weight, the color the topic and the directionality the flow of messages. Given a pair of users u and v and a topic Λ , we are interested in predicting two parameters—the probability of communication of u with v , $c_{uv}(\Lambda)$ and an association time delay in communication $d_{uv}(\Lambda)$.

valued weight function $f: E \rightarrow \mathbb{R}$). Each edge $e_{uv}(\Lambda) \in E$ from a user u to another user v is defined over a topic Λ . It is associated with two parameters—a measure of the probability of communication of u with v , $c_{uv}(\Lambda)$ and an association time delay in communication $d_{uv}(\Lambda)$. Note since u could communicate with v on a wide range of topics, it is very likely that the weighted directed graph G is a multi-edge graph, i.e. there could exist several edges between u and v , each defined over a topic. Obviously, the probabilities of communication and associated delays would be different across different topics. A diagrammatic representation of the data model is shown in Figure 3.1.

Problem Statement

We now present the problem statement, technical challenges and key issues addressed on this chapter. Given u and v as two users in the social network (or the communication graph $G(V, E)$ as introduced in the previous sub-section), the technical problem addressed in this chapter is to predict communication flow between u and v given a certain topic Λ when (a) they are directly connected through an edge in G , and when (b) they might have arbitrary number of users in the paths $p \in P$ that links them in the social network G . The problem of communication flow is addressed by computing the likelihood that u will communicate with v (intent to

communicate, $c_{uv}(\Lambda)$) on the particular topic Λ , and additionally predict the delay in communication, $d_{uv}(\Lambda)$.

The solution to estimate both the likelihood of u 's communication or the intent to communicate with v and the delay lies in determining the topic of the message, understanding the contextual factors that affect the communication between the two users on this specific topic, their social roles and the nature of relationship between them. These factors affecting communication flow is categorized into two parts: communication context and social context. There are three contextual factors relating to the sender's communication context that are examined in this chapter—(a) effect of the local social network of the sender, (b) relationship of message topic to the sender's past communication, and (c) relationship of the sender to the recipient with respect to communication. Social context comprises contextual factors which refer to the patterns of participation in communication for the two people (information roles) and the degree of overlap of friends between them (strength of ties). Hence the key challenges are as follows:

1. How do we model the communication context, comprising neighborhood, topic and recipient context between a pair of users?
2. How do we model the social context between them?
3. How can we dynamically select an optimal set of features at a certain point in time for predicting communication flow? How does this prediction method extend to different network topologies - directly connected users and any two users who are arbitrarily distant in the social network?

In the following three sections, we present communication context, social context and the prediction framework for determining communication flow between a pair of users.

3.4 Communication Context

In this section we present the features of the communication context that affects a pair of users to communicate. There are three parts to communication context—neighborhood context, topic context and recipient context.

Neighborhood Context

Neighborhood context refers to the effect of the user’s social network on her communication. There are two network effects of interest—backscatter and susceptibility. We explain our motivation behind the choice of features of neighborhood context by a physics based wave metaphor. In the classical wave theory, the phase and magnitude of a wave at a certain point and time in space is the linear superposition of all waves from all sources, at the same point in time and space. This superposition can result in constructive (when the phases align) or destructive interference (when the waves are out of phase). Waves are additionally affected by the properties of the medium and exhibit phenomena such as reflection.

We present a simple example to illustrate the role of the metaphor. Let us consider a user v ’s social network. Further let users u and w are v ’s friends (Figure 3.2). We now apply the metaphor to our problem. User v ’s communication to his contacts creates a primary wave front. When her recipients create messages to further propagate the topic they generate secondary wave-fronts. Note v will also be the recipient of messages from her contacts—these are reflections or backscatter. The multiple communication wave fronts reaching u will have constructive / destructive effects on her intent based on her communication and social context.

We use the following example for the following three sections. We have two users u and v and v ’s contacts (w). Now assume that u wants to discuss topic Λ with v at a specific time slice t_i . We also denote the count of messages on Λ from u to v as $n_{uv}(\Lambda)$.

Backscatter. Backscatter refers to the fraction of the messages received by u from

Figure 3.2: A Physics wavefront metaphor depicting primary and secondary wavefronts generated due to communication. Messages from u to v and w create primary wavefronts; while those which travel back to u , it is called backscatter. If v and w further propagate the messages, they constitute the secondary wavefronts.

her contacts that are about a topic Λ . We can reasonably assume that a message sent by u at a certain point of time t_i will be affected by all the messages sent by u 's contacts on the same topic Λ that arrived before t_i . A diagrammatic representation explaining backscatter is shown in Figure 3.3. We now assume that the number of messages received from a specific contact v_1 is independent of the number of messages from any other contact v_2 , given Λ . We further assume that each contact v has the same importance with respect to u . Thus the backscatter due to one contact v in time slice t_i is given by,

$$\theta_{v \rightarrow u|u}(\Lambda, t_i) = \sum_{j=1}^{n_{v \rightarrow u|u}} \varphi(\Lambda, t_j, t_i) \quad (3.1)$$

where, t_j is the time-stamp of the j -th message on topic Λ from v to u and $\varphi(\Lambda, t_j, t_i)$ is an indicator function: 1 if t_j lies in time slice t_i and 0 otherwise. Thus, $\theta_{v \rightarrow u|u}$ gives the backscatter due to a specific contact v to u and due to an earlier communication from u on the topic Λ . Backscatter due to a specific contact v is proportional to the number of messages sent by v to u . The total backscatter $B_u(\Lambda)$ is given by the sum over all messages received from all L contacts, prior to t_i .

$$B_u(\Lambda) = \frac{1}{N \cdot L} \sum_{i=1}^N \sum_{v=1}^L \theta_{v \rightarrow u|u}(\Lambda, t_i) \quad (3.2)$$

Figure 3.3: Schematic representation explaining backscatter and susceptibility.

where N is the total number of time slices prior to t_i .

Susceptibility. Susceptibility measures whether the social network that u interacts with is interested in the topic that she plans to communicate on. Intuitively, if a network is susceptible to communication on a certain topic, then u is more likely to send a message on topic Λ to her network. Susceptibility is proportional to the number of messages sent by u 's contacts to their contacts. This is the key difference between susceptibility and backscatter. A diagrammatic representation explaining susceptibility is shown in Figure 3.3. Making identical assumptions as in the calculation of backscatter, the susceptibility is captured as follows. The susceptibility due to one contact v to her entire social network at time slice t_i is given by,

$$\theta_{v|u}(\Lambda, t_i) = \sum_w \sum_{j=1}^{n_{v \rightarrow u|u}} \varphi(\Lambda, t_j, t_i) \quad (3.3)$$

where, t_j is the time-stamp of the j -th message from v to u and $\varphi(\Lambda, t_j, t_i)$ is an indicator function: 1 if t_j lies in time slice t_i and 0 otherwise. Thus $\theta_{v \rightarrow w|u}$ gives the susceptibility due to a specific contact v (to her contacts w) due to an earlier communication from u to v on the topic Λ . Susceptibility for a specific topic Λ , for a user u , is just the sum of susceptibilities per contact v , over all past time slices prior to t_i .

$$S_u(\Lambda) = \frac{1}{N \cdot L} \sum_{j=1}^N \sum_{v=1}^L \theta_{v|u}(\Lambda, t_i) \quad (3.4)$$

where, N is the number of time slices prior to t_i .

Topic Context

Topic context refers to the effect of the semantics of a user's *past* communication on the topic Λ on her future communication. We are interested in four measures - (1) message coherence (2) temporal coherence (3) topic relevance and topic quantity.

Message Coherence. Message coherence refers to consistency in message semantics and the semantic relationships of the messages with the current topic Λ (e.g. 'movies').

We use the common sense reasoning toolkit ConceptNet (Ino et al., 2005) to compute the distance between messages (Shevade et al., 2007). ConceptNet has several desirable characteristics that distinguish it from the other popular knowledge network—WordNet (Miller et al., 1990). First, it expands on pure lexical terms to include higher order compound concepts ("buy food"). The repository represents semantic relations between concepts like "effect-of", "capable-of", "made-of", etc. Finally, ConceptNet is powerful because it contains practical knowledge - it will make the association that "students are found in a library" whereas WordNet cannot make such associations. Since our work is focused on communication in online social networks, which typically deals with casual conversations, ConceptNet is very useful. Each message comprises a set of words. This is obtained after stop word removal and word stemming. Let $d_c(w_1, w_2)$ denote the ConceptNet distance between two words (concepts) w_1 and w_2 . Then, the distance between a message m and a topic Λ is given as:

$$d(m, \Lambda) = \max_q \min_k d_c(w_q, w_k) \quad (3.5)$$

where, w_q is a word in message m and w_k is a word corresponding to Λ . Given a topic, Wordnet [Miller et al. 1993] is helpful in determining the synonym set for that topic—this helps us determine the set of words w_k for a topic Λ .

Message coherence $C(\Lambda)$ is then computed as the ratio of $\Omega(\Lambda)$ to $\Omega(\hat{\Lambda})$, where $\Omega(\Lambda)$ is the measure of the average similarity of all messages with respect to

topic Λ (computed using eqn. 3.5), and where, $\hat{\Lambda}$ is the set of antonyms (obtained using WordNet) corresponding to the topic Λ . This ratio has interesting properties: (1) if $C > 1$, then the intent to communicate will be high since messages on Λ are highly coherent; (2) if $C < 1$, then the intent to communicate will be low, since it implies that there is probably a topic in $C(\hat{\Lambda})$ which is more coherent than Λ ; and (3) if $C \approx 1$, then the effect on the intent to communicate might be considered neutral due to the presence of several topics.

Temporal Coherence. Temporal coherence is defined as the correlation of the time-stamps of the messages on a topic received by a person u . High coherence of messages in a recent past would increase a u 's intent to communicate and vice versa. It is determined by the mean and variance of the time-stamp differences of messages received by u in the past referenced from current time t_i . The mean μ_j over a time slice t_j and topic Λ is given by the mean difference of the time-stamps (T) of all the messages in time slice t_j referenced from current time t_i .

$$\mu_j(\Lambda, t_j, t_i) = \sum_{m \in t_j} (T(m, \Lambda, t_j) - t_i) / (n(\Lambda, t_j)) \quad (3.6)$$

where m is the index of a message of topic Λ in the time slice t_j and where $n(\Lambda, t_j)$ is the number of messages on topic Λ at time slice t_j . Similarly, the variance σ_j^2 over a time slice t_j and topic Λ is easily computed. Hence for each time slice t_j (duration of one week in our experiments), we can compute mean and variance (μ_j, σ_j^2) .

Topic Relevance and Quantity. Topic relevance for user u on a topic Λ refers to the relationship between topics in her past communication to the topic Λ . We can compute topic relevance $\psi_R(u, \Lambda)$ for u on topic Λ by the ratio of the number of messages $n_{u \rightarrow v}(\Lambda)$ on Λ sent by u to all contacts v to the total number of messages sent by u to all contacts v on all topics.

$$\psi_R(u, \Lambda) = \sum_v n_{u \rightarrow v}(\Lambda) / \sum_{\Lambda} \sum_v n_{u \rightarrow v}(\Lambda) \quad (3.7)$$

Topic quantity is the number of topics on which user u has received messages in the recent past. We determine the number of topics via spectral clustering. We consider messages on the same topic to belong to a specific topic cluster. The number of such topic clusters (k) is determined dynamically in our experiments. The effect of topic quantity $Q_u(\Lambda)$ for a topic Λ on the intent to communicate for user u is inversely related to the number of topic clusters k : $Q_u(\Lambda) = 1/k$. This implies that if u receives messages on many topics (large k), then the intent to communicate will decrease.

Recipient Context

Recipient context refers to effect of the recipient identity on u 's intent to communicate. There are three measures of interest—(a) reciprocity, (b) communication correlation and (c) communication significance.

Reciprocity. Reciprocity refers to the ratio of the messages received from the recipient to those sent to the recipient, on the intended communication topic. Reciprocity $r_{v \rightarrow u}$ of a user u with respect to v is given by the ratio of the number of messages $n_{v \rightarrow u}$ sent by v to u to the number of messages $n_{u \rightarrow v}(\Lambda)$ sent by u to v on topic Λ . Reciprocity is given as follows:

$$r_{v \rightarrow u}(\Lambda) = \frac{n_{v \rightarrow u}(\Lambda)}{n_{u \rightarrow v}(\Lambda)} \quad (3.8)$$

Communication Correlation. Communication correlation (ρ_{uv}) refers to the topical alignment between a user u and her contact v with whom she wants to communicate. It is computed as a histogram intersection distance over all time slices:

$$\rho_{uv}(\Lambda) = \frac{\sum_i \min(\beta_u(\Lambda, t_i), \beta_v(\Lambda, t_i))}{\sum_i \max(\beta_u(\Lambda, t_i), \beta_v(\Lambda, t_i))} \quad (3.9)$$

where, $\beta_u(\Lambda, t_i)$ refers to the number messages sent by user u at time t_i on topic Λ .

Communication Significance. Communication significance refers to the fraction of past messages to the specific contact v on the current communication topic. It is given by the ratio $s_{u \rightarrow v}(\Lambda)$ of the number of messages $n_{u \rightarrow v}(\Lambda)$ from u to v to the number of messages from u to all v on topic Λ .

$$s_{u \rightarrow v}(\Lambda) = \frac{n_{u \rightarrow v}(\Lambda)}{\sum_v n_{u \rightarrow v}(\Lambda)} \quad (3.10)$$

To summarize, we have considered the attributes that are part of current communication context between a pair of users in the previous two sections. Note, the communication patterns of people can also be affected by the habitual and network properties that are acquired over time; e.g. *who* is communicating with *whom* and what is the *strength* of relationship shared between them. In the following section we define these contextual factors.

3.5 Social Context

In this section, we discuss social context. *Social context* is the set of attributes that refers to *who* is communicating with *whom* and what is the *strength* of relationship shared between them. These are the habitual and network properties acquired by a person over time. There are two features of interest that we discuss in this section: information roles and strength of ties. Communication flow between users is affected by the communicative behavior of each. For example, user u might be a person who is very active in sending messages to her contacts. In that case the probability that u would communicate with v is high. Further, suppose u and v have several friends in common, which implies they share a strong bond. The strength of the tie between u and v would also affect the probability that they would communicate.

Information Roles

Information role is a contextual attribute acquired over time which impacts a person's communication behavior. We formally define three different categories of roles

Figure 3.4: Schematic representation explaining the information roles of users in a social network based on the dynamics of transmission and reception of messages.

of people: (a) *generators*, people who generate information by themselves or from other sources (e.g. external events like the American Idol or iPhone release), (b) *mediators*, people who act as transmitters of information between people, and (c) *receptors*, people who mostly receive messages. Drawing an analogy with the hyper-linking structure of the web, we notice that the generators and the receptors act like authorities and hubs respectively in the social network. The roles presented are exhaustive but not mutually exclusive; clearly a person can play different roles, depending on the context.

Let us again consider users u , v and w to be part of a social network. The roles of the three (as generators, mediators and receptors) emerge from three different communication structures shown in Figure 3.4. User u is a generator characterized by several out-going communication links; v is a mediator with comparable number of incoming and out-going links and w represents a receptor with large frequency of incoming links.

We define the information role of a person to depend on the net communication activity, in which she participates. This is given by the following message frequency ratio:

$$R(u, t_i) = \frac{\sum_v n_{u \rightarrow v}(t_i)}{\sum_v n_{v \rightarrow u}(t_i)} \quad (3.11)$$

where $n_{u \rightarrow v}$ is the number of messages sent by u to a contact v at time slice t_i and $n_{v \rightarrow u}$ is the number of messages received by u from a contact v at time slice t_i . In our experiments one time slice is assumed to be equivalent to one week. Now

we define the following conditions to define the roles: (a) If $R(u, t_i)$ is significantly greater than 1, then the person is a generator, (b) If $R(u, t_i)$ is approximately equal to 1, then the person is a mediator, and (c) If $R(u, t_i)$ is significantly less than 1, then the person is a receptor.

Intuitively, the information roles of a pair of communicators would change over time. Hence their probability of communication would also change due to change of roles. If the duration of past communication of the two communicating users is divided into i time slices, we therefore need to determine their roles as well as the probability that they would communicate after i time slices.

We define a role transition matrix $\mathbf{P}_1$ over the three roles s_1, s_2, s_3 such that $P(s_m, s_n)$ gives the probability that a person in role s_m would communicate with another person in role s_n at the *first* time slice. The values of the transition matrix are determined empirically from the past communication based on frequency of messages exchanged. Now given this initial role transition matrix, our goal is to determine their probability of communication after i transitions (or time slices). We use the Chapman-Kolmogorov equation which defines a technique to compute i -step transition probabilities by multiplying the initial transition probability matrix i -times. This gives us the role transition matrix $\mathbf{P}_i$ for the two people involved in communication at time slice i . We can now easily determine the two roles at time slice i using eqn. 3.11 and the probability that these two roles will communicate from $\mathbf{P}_i$.

Strength of Ties

This section discusses a second feature of social context—the strength of ties between a pair of communicating users. It is well known from prior work (Haythornthwaite, 1996) that the nature of relationship between two people affects communication. This argument is based on evidences in (Haythornthwaite, 1996) which suggest that the pattern of relationships between actors (people) reveals the likelihood that users will be exposed to particular kinds of information.

We categorize the relationships between two people using the strength of

Figure 3.5: Schematic representation explaining the notion of strength of ties between users in a social network. Strong ties usually form in a sub-network of densely connected users; while sparsely connected user sets are likely to have weak ties between them.

shared ties: strong ties and weak ties, similar to (Granovetter, 1973). The strength of a tie is proportional to the number of common friends between two persons (Figure 3.5). We emphasize the real world observation of evidence of strong friendship (or bond) between two persons when there is frequent communication between them. To take an example, once again suppose users u and v are friends; and u and w are also friends. We further consider the induction of a ‘psychological strain’ for the two pairs. This is because both v and w attempt to make their communication congruent with their common friend u . This depicts the introduction of a positive tie between v and w .

Therefore, we consider the strength of a tie between two people to be dependent on the overlap of their friends’ circles. For example, as in the above example, if u and v have ten common friends then they share a *strong* tie, while if u and w have two common friends, then they share a *weak* tie. Prior work on the strong and weak ties (Granovetter, 1973) reveals two claims as follows.

The first claim is that the exchange of new information (an external event like London bombings in 2006) is higher along weak ties. This is explained by the intuition that the new information can traverse greater social distance along weak ties. Suppose user u sends a message on a recent movie review to her friends, and those friends send messages to their friends. Many of these people would form a small clique, sharing strong ties. As a result the information traveling through such ties will likely be limited to a small clique of friends. On the other hand, if the new

information is transmitted across weak ties, the less overlap of friends is likely to make the new information reach other cliques.

The second claim is that strong ties are better for transmission of existing information which is often characterized by execution of an action in the external world (e.g. referral for a job position). This claim works on the ground that weak ties work well when there is a lot of friction among the people (since they stay manageable and provide a fresh perspective). As this friction gets reduced in a social network with hundreds of contacts, weak ties become overwhelming and people ignore the information to cope with information overload. In this case strong ties are more reliable means of transmission. We therefore observe that the probability of information across strong or weak ties depends upon latent factors, e.g. the ‘age’ of the information. Determining the age of information in an online social network like MySpace is a challenging problem. Instead in this work, we *learn* the impact of strength of ties on communication by defining the overlap in the friends’ circles of the two people. Strength of tie is a symmetric measure, therefore given by the measure of ‘embeddedness’ (Granovetter, 1973) between users u and v :

$$\psi(u, v, t_i) = (L(u, t_i) \cap L(v, t_i)) / (L(u, t_i) \cup L(v, t_i)) \quad (3.12)$$

where $L(u, t_i)$ and $L(v, t_i)$ are respectively the friend lists (vector of contact names/ids) of u and v at time t_i .

We next discuss how our computed features for communication and social context are incorporated into a Support Vector Regression based technique. This is used to predict the intent to communicate as well as the average delay in communication based on a dynamically selected set of contextual features.

3.6 Predicting Communication Flow

In this section we present our method of predicting communication flow, via its two attributes—intent to communicate and associated delay. We discuss a Support Vector Regression framework for two scenarios—directly connected pair

of users, and second when the users are arbitrarily distant from each other in the network.

SVM Regression Framework

In this section we present our prediction framework. First, we discuss the method of dynamically selecting an optimal set of features of communication and social context. Second, we present a Support Vector Machine based incremental prediction algorithm.

Selecting Optimal Contextual Features. In order to model dynamic context, it is very important to eliminate some features of the communication and social context. This is because some features despite capturing the current communication context, worsen the process of learning sometimes by enormously increasing the learning time of the relationship between contextual features, and also producing more rules and less accurate prediction rates. Thus, we need to select the subset of the features such that the performance of the learning is optimal. The optimality can be a function of several environment constraints or combination of factors including the learning time, the size of the final knowledge, and the prediction rates of the intent and the delay, especially because we are modeling temporally varying data.

In this work, this problem of dynamically modeling context is solved by performing ‘feature voting’ using dynamic selection of contextual features at the beginning of each time slice. We use a set of popular feature selection techniques in the voting strategy to determine an optimal set of features at each time slice. They are: Mutual Information, Principal Component Analysis and k -NN based estimation. We provide brief descriptions of these techniques as follows.

Mutual information of two features is a quantity that measures the mutual dependence of the two. Thus the dependency between features F_i and F_j is measured by how much the uncertainty decreases after the value of the features gets known.

Principal Component Analysis (PCA) is a multivariate statistical technique which rotates the data such that maximum variabilities are projected onto its prin-

principal axes or eigen components. It thus involves the calculation of the Singular value decomposition of a data set and can be viewed as a rotation of the existing axes to new axes in the space. In this new rotation, there is no correlation between the new features defined by the rotation. Thus it facilitates feature selection.

k-NN estimation based technique uses feature-weighting over the popular k nearest neighbor based learning algorithm. It captures complex non-linear dependencies among the features and uses the leave-one-out error method as a natural regularization.

We now present the feature voting strategy used to determine an optimal set of dynamically selected features based on these three techniques. At each time slice, k contextual features are selected using each of the three feature selection algorithms separately. Thereafter voting of the features for selection is performed, and k features which have been voted for selection maximum number of times are selected as the final optimal features for that time slice. In this work k has been set to 5.

Using these dynamically selected k features, we would predict the two parameters of communication context—intent to communicate and delay in the next section.

Prediction. Now we present the prediction framework. The intent to communicate and delay can be modeled as a regression problem where the relationships between the different model parameters (based on the dynamically selected features of communication and social context) can be learnt over time and for specific pairs of users. Our regression model is based on a Support Vector Machine (Vapnik, 1995) based unsupervised learning. To avoid training the time series data every time when we get a new test pair, we use an incremental SVM regression (Zhang, 2000) method. The SVR prediction algorithm for intent to communicate $c_{uv}(N+1)$ for a pair of users u and v and at a future time slice $N+1$ is described in Table 3.1. Estimates of communication delay $d_{uv}(N+1)$ can be determined using the same procedure as in Table 3.1.

Table 3.1: Algorithm for SVR prediction.

<p>Input: $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $\mathcal{X} = \mathbb{R}^d$ pairs of training data.</p> <ol style="list-style-type: none"> 1. Predicted intent x_i, $i = 1, 2, \dots, N$ where N is the number of time slices over past communication for two <i>specific</i> users u and v on topic Λ. It consists of the contextual feature vectors, $x_i = \{S_u(\Lambda), B_u(\Lambda), C(\mu_b \sigma_i^2), \psi_R(u, \Lambda), Q_u(\Lambda), r_{v \rightarrow u}(\Lambda), \rho_{uv}(\Lambda), s_{v \rightarrow u}(\Lambda)\}$. 2. Actual communication (based on frequency count of messages sent by u to v) y_i, $i = 1, 2, \dots, N$ where N is the number of time slices over past communication for users u and v on a topic Λ. <p>Procedure:</p> <ol style="list-style-type: none"> 1. The SVM regression function $f(x)$ is trained on $\{(x_1, y_1), \dots, (x_N, y_N)\}$. It is tested on the incremental sample (x_{N+1}, y_{N+1}). 2. The training set is then augmented with the new samples (x_{N+1}, y_{N+1}) and a new regressor is learnt using this training set. 3. Repeat a-c until there are no more samples. <p>Output: Error in prediction, E.</p> <ol style="list-style-type: none"> 1. Use $f(x)$ with training set $\{(x_1, y_1), \dots, (x_N, y_N)\}$ and determine predicted $y_{N+1}^{\hat{}}$. 2. Determine the actual communication intent $I_{uv}(N+1) = y_{N+1}$. 3. Compute error as, $E = (y_{N+1} - y_{N+1}^{\hat{}}) / y_{N+1}$.
--

The use of the SVM based regression algorithm can be directly used to incrementally predict the intent to communicate and the communication delay for a user u with a specific contact v that is, when u and v are connected directly through an edge.

In the following section, we extend the prediction framework for two users u and v when they are separated along a path comprising arbitrary number of users (n , say) between them.

Prediction for Arbitrarily Distant Users

In this section, we present the prediction framework for determining the optimal path of communication flow between users who are distant from each other by arbitrary number of hops in the social network. In other words, the goal is to find the shortest communication path from one user to another in a social network. For the purpose, we first introduce an algorithm called the restricted shortest path al-

Figure 3.6: Schematic representation of predicting communication flow between a pair of users who are arbitrarily n hops distant in the social network. The goal is to find a maximum probability $c_{uv}(\Lambda)$ and minimum delay $d_{uv}(\Lambda)$ path p between a pair of users u and v .

gorithm (Lorenz and Raz, 1999) and then discuss its use in determining the optimal communication path between a pair of users.

Restricted Shortest Path Algorithm. Recall, in our case, the social network is a communication graph where each edge between two nodes can be defined by two parameters: intent to communicate and delay. Determining the most efficient path in this case can be assumed to be a dual-optimization problem where communication can flow from a user u to another user v (who can be arbitrarily n hops distant in the graph) with maximum probability and least delay. A diagrammatic representation is shown in Figure 3.6. This can be defined formally as follows.

Given the weighted directed communication graph $G(V, E)$ and further given a user u of V , find a path $p \in P$ from u to a different v of V so that, $f(p)$ is minimal among all paths P connecting u to v ; f being a function which yields maximizes probability and minimizes delay.

We reduce our problem to the restricted shortest path problem (Lorenz and Raz, 1999) (a variant of the shortest path problem involving two edge parameters) where the parameters of an edge in a graph are dependent on two values: for example, edge cost (logarithm of probability of communication) and edge weight (estimated delay). Using these semantics we attempt to find the shortest path of communication from u to v in the communication graph. The restricted shortest

Table 3.2: Restricted Shortest Path Problem.

Restricted Shortest Path Problem ($G(V, E)$, $\{c_e, d_e\} \in E$, c^* , L , U , ε), the graph $G(V, E)$ has edges m , c^* is a positive integer and L and U are the upper and lower bounds of the optimal cost respectively. The source and target nodes are denoted as u and v .

```

 $S \leftarrow \frac{L\varepsilon}{n+1}$ 
for each  $l \in E$ ,
    define  $\tilde{d}_l \equiv \lfloor d_l / S \rfloor + 1$ 
 $\tilde{U} \leftarrow \lfloor U / S \rfloor + n + 1$ 
for all  $w \neq u$ ,
     $c(w, 0) \leftarrow \infty$ 
     $c(u, 0) \leftarrow 0$ 
for  $i = 1, 2, \dots, \tilde{U}$ ,
    for  $w \in V$ ,
         $c(w, i) \leftarrow c(w, i-1)$ 
        for  $l \in \{(z, w) \mid \tilde{d}_l \leq i\}$ ,
             $c(w, i) \leftarrow \min \{c(w, i), (c_l + c(z, i - \tilde{d}_l))\}$ 
        if  $c(v, i) \leq c^*$ ,
            return the corresponding path as the communication path
            return the associated delay of cost  $d(v, i)$ 
return FAIL

```

path problem is to find a path p in G from u to v such that the probability along this path is no lesser than a positive integer c and the transition delay of p is minimal.

The above restricted shortest path problem is NP-complete but can be solved in polynomial time by allowing an approximation factor of ε and with given upper UB and lower bounds LB on the optimal cost (or probability of communication) for the shortest path. The complexity of the solution is given by $O((mn/\varepsilon) \log \log(UB/LB))$. Table 3.2 gives the algorithm.

Now we discuss how the restricted shortest path problem can be applied to our model for predicting the intent and delay between two users who are arbitrarily n hops away.

Predicting the Optimal Communication Path. We apply the restricted shortest path algorithm to our problem where we need to find the optimal communication

route between two users u and v who are n hops away. Our motivation behind using the restricted shortest path problem in our problem lies in the fact that it preserves the notion of use of the contextual features: neighborhood, topic, recipient and social context in determining the individual edge parameters: probability of communication and delay. With an allowed degree of approximation, it yields a sub-optimal communication path that can transmit certain information from one person to another who is n hops away in the social network.

The restricted path problem is used to compute the intent and delay between two users in the following manner:

1. First, given the directed social network graph $G(V, E)$, we have to determine the edge parameters $\{c_l, d_l\}$ or the values of intent to communicate c_l and delay d_l for each pair of adjacent users. We compute the logarithm of the intent to communicate (probability of communication) and the delay using our SVR algorithm in Table 3.1 for each pair of adjacent users in the graph.
2. Now we choose loose upper and lower bounds on the optimal cost (probability of communication) as $LB = 0.5$ and $UB = 0.99$ ⁸. The approximation factor ε (or the allowance for error in approximation) is tuned over a wide range of values and finally chosen to be 0.2 (or 20%). The restricted shortest path problem (Table 3.2) is now used on the directed graph $G(V, E)$ where each edge $l \in E$ has the parameters $\{c_l, d_l\}$. The algorithm converges given a cost value of c^* and determines the optimal path from user u to v which maximizes the logarithm of the intent (since logarithm is a concave function and is additive) and minimizes the delay.

In the next section, we discuss evaluation techniques that would be used in the experimental section to verify our prediction framework.

⁸We tuned over different ranges of LB and UB to figure out that values over a wide range in $[0,1]$ perform best in our case.

3.7 Evaluation Techniques

In this section, we discuss the techniques used for evaluating the prediction model of communication flow. First we introduce a simple baseline technique based on use of historical communication data between two users. The second baseline technique is based on learning the trend of past communication between the pair of users. We use a model of spread of infectious diseases (Gruhl et al., 2004) as the third validation technique. The fourth evaluation technique is based on the idea of how the local neighborhood of a user affects her decision to communicate.

Frequency of Past Communication based Method

The first baseline technique for predicting the intent to communicate is based on using the prior probability of communication on a given topic Λ . This probability is proportional to the frequency count of messages exchanged between the users u and v in the past (t_0 to t_k) on Λ . The predicted intent $c_{uv}(\Lambda, t_{k+1})$ at a future time slice t_{k+1} is given by the ratio of the number of messages $n_{uv}(\Lambda)$ sent by u to v on the topic Λ to the total number of messages on all Λ sent by u to v in the past, described as below:

$$c_{uv}(\Lambda, t_{k+1}) = n_{uv}(\Lambda, t_0 : t_k) / \sum_{\Lambda} n_{uv}(\Lambda, t_0 : t_k) \quad (3.13)$$

The prediction of delay $d_{uv}(\Lambda, t_{k+1})$ is based on determining the correspondences between two messages between the users and then computing the mean delay by examining the message time stamps. For determining the semantic correspondence of messages we use the ConceptNet distance—for each message, we assume that the nearest message (in terms of time) whose distance is below a threshold is the corresponding message. We acknowledge that the message correspondence issue will benefit through a linguistic analysis of messages—that is beyond the scope of this chapter. Then the mean delay between two users u and v on topic Λ is the mean delay between all pairs of corresponding messages on the same topic. The predicted delay $d_{uv}(\Lambda, t_{k+1})$ for the future time slice t_{k+1} is computed at time t_k as

the mean delay across all previous time slices till t_k :

$$d_{uv}(\Lambda, t_{k+1}) = \frac{1}{k} \sum_{i=1}^k d_{uv}(\Lambda, t_i) \quad (3.14)$$

Trend Learning based Method

Our second baseline technique deals with computing the intent to communicate $c_{uv}(\Lambda, t_{k+1})$ and the associated $d_{uv}(\Lambda, t_{k+1})$ delay at a future time slice t_{k+1} for users u and v based on learning the trend of each of these two parameters over the time period in the past, t_0 to t_k . The trend is learnt based on a logistic function that incorporates the slope of a straight line fit to the past frequency (and delay) of communication between u and v . We use the following equations to determine the intent to communicate and the delay at a future time slice t_{k+1} :

$$\begin{aligned} c_{uv}(\Lambda, t_{k+1}) &= \frac{1}{1 + \exp(-\sum_{i=1}^{k-1} n_{uv}(\Lambda, t_{i+1})/n_{uv}(\Lambda, t_i))}, \\ d_{uv}(\Lambda, t_{k+1}) &= \frac{1}{1 + \exp(-\sum_{i=1}^{k-1} d_{uv}(\Lambda, t_{i+1})/d_{uv}(\Lambda, t_i))}, \end{aligned} \quad (3.15)$$

where $d_{uv}(\Lambda, t_{k+1})$ is given according to the method described in the previous section, and the logistic form of the functions ensures that the values of the intent to communicate and delay are restricted in the range $[0,1]$.

Epidemic Disease Model based Method

In this section, we discuss the third baseline technique based on the idea of spread of epidemic diseases which has been used in prior work for predicting information diffusion in social networks (Briesemeister et al., 2003; Girvan et al., 2002). This framework is based on the popular Independent Cascade Model (Goldenberg and Muller, 2001). The steps involved in computing the intent to communicate and delay using this disease propagation model are as below:

1. Let us consider the social network to be a directed graph, where each edge (u, v) is labeled with a copy probability κ_{uv} . In our problem we define the copy

probability to be the chance that u would communicate with v on a certain topic and certain time.

2. We introduce the notion that a user v may respond to certain messages from u frequently and others infrequently. We capture this with an additional edge parameter r_{uv} denoting the probability that v responds to u 's message at a certain point of time.
3. The propagation in the model occurs as follows. If a topic Λ exists at vertex u on a given time slice, i.e., u has previously communicated about the topic from time slices t_0 to t_k —then we compute the probability that the topic Λ will propagate from u to a neighboring vertex v at t_{k+1} as follows. Node v reads the topic from node u on any given day with reading probability r_{uv} , so we choose a delay from an exponential distribution with parameter r_{uv} . Then, with probability κ_{uv} , v will choose to respond to it. Alternatively, this implies that once u is infected, node v will become infected with probability $\kappa_{uv} \cdot r_{uv}$ on any given day. This is the intent to communicate of v with respect to user u ($c_{uv}(\Lambda, t_{k+1})$). The value of $\kappa_{uv} \cdot r_{uv}$ is determined using an iterative algorithm that uses the mean delay of v 's response over all topics. The details of the iterative algorithm can be referred to in (Gruhl et al., 2004).

Note, this model does not incorporate methods to determine the associated delay, given the probability of communication. Hence it can only be used to evaluate the first parameter—intent to communicate.

Neighborhood Effect based Method

The fourth baseline technique is based on the idea that the local neighborhood⁹ (set of contacts) of a user u 's affects her decision to communicate with another user u 's message. We conjecture that a user v would respond to a particular topic Λ at time slice t_{k+1} with a probability (i.e. intent) proportional to the fraction of her contacts

⁹Note the messages crawled are the ones publicly visible on the profile page of a MySpace user. We did not collect any private messages between users.

Figure 3.7: Average Path Length Distribution of crawled MySpace data— we observe that the mean shortest path is ~ 5.9 which satisfies the criterion for a 'small world' network.

who have already communicated on the topic; and with delay that is the mean of the delays for the set of those contacts. The intent to communicate and associated delay are therefore given as:

$$\begin{aligned}
 c_{uv}(\Lambda, t_{k+1}) &= \frac{|L_{\Lambda}(u, t_k)|}{|L(u, t_k)|}, \\
 d_{uv}(\Lambda, t_{k+1}) &= \frac{1}{|L_{\Lambda}(u, t_k)|} \sum_v \sum_w d_{vw}(\Lambda, t_k),
 \end{aligned} \tag{3.16}$$

where $L_{\Lambda}(u, t_k)$ is the subset of contacts of u who have communicated on the topic Λ at time t_k , and w is a contact of v to whom she has sent a message at t_k .

3.8 MySpace Dataset

MySpace is one of the world's largest social networking site with over 108 million users. The dataset used for our experiments comprises approximately 200,000 users who have exchanged about 12,425,010 messages in the time snapshot from September 2005 to April 2007. The crawling process was seeded from one of Tom's (super-user of MySpace who is a contact of all the users) top eight friends. A depth first strategy was adopted to continually crawl the friends of a user who have sent messages to the user in the said time period (included the timestamps and the content of the messages). The process was continued till we reached the third level

Figure 3.8: Logarithmic plot of the out-degree degree distribution of MySpace dataset. We observe that it follows a power-law, with exponent ~ 2.01 . This is in conformity with the observations in other small-world networks.

friend of each user (friends-of-friends). This strategy was adopted to ensure that the data we are dealing with exhibits sufficient traces of communication among the crawled users as well as the network of users was sufficiently cohesive in relationship pertaining to communications.

We now describe the network topology of the crawled dataset using three standard metrics: average shortest path length, degree distribution and clustering coefficient. Our analysis shows that the average shortest path length is approximately 5.952 (Figure 3.7). The out-degree distribution (Figure ??) is long-tailed and follows a power law distribution $P(k) \propto k^{-\gamma}$ (γ is a network coefficient) with $\gamma = 2.01$. Finally the clustering coefficient, defined as the probability that friends of a person will mutually be friends too, was determined to be 0.49. These measures are consistent with statistics of other social network datasets (Watts and Strogatz, 1998) which follow a topology akin to scale-free networks and observe the ‘small-world phenomenon’. This validates the suitability of our dataset to study communication flow—because literature suggests that information is likely to propagate at a fast rate between users in small-world networks. A detailed list of several social network features that characterize our dataset have been shown in Table 3.3.

We further provide some empirical studies regarding the content associated

Table 3.3: Average Network Statistics for the MySpace dataset.

IN-DEGREE	OUT-DEGREE	CLUSTERING COEFFICIENT
43.4	35.6	0.49
AVG. PATH LENGTH	#COMPONENTS	SIZE OF LARGEST COMPONENT
5.9	58	138,573 (~90%)

Figure 3.9: Distribution of messages over the different topics (synsets) extracted using WordNet.

with the communication among users. Since the communication flow parameters—intent to communicate and delay are defined per topic, we now discuss how each of these message from the crawled MySpace data is assigned a ‘topic’. This is done using a simple aggregation algorithm that exploits the tree structure of using WordNet (Miller et al., 1990). In WordNet, each word belongs to a synonym set or a set of words that are interchangeable in some context (synset), representing a unique lexical concept. For each word in the message, we determine the synset to which belongs. Using WordNet, we also determine the third-level generalization for each synset. These typically are abstractions (‘entity, physical, object’ is the third level generalization for the concept ‘car’) and we refer to them as topics in this chapter. Two lower level synsets are similar, if they share the same topic. We assign the topic to a message that covers the largest number of message synsets. A distribution of the number of messages over the most frequent 140 topics (synsets) is shown in Figure 3.9.

In this research we examine the 25 most frequently occurring topics concern-

ing about 150,000 users and 9,140,000 messages exchanged between them. We notice that there are only a few topics which are highly discussed; the topic distribution in general is long-tailed.

3.9 Experimental Results

In this section we discuss the experimental results conducted using MySpace data. We start with a qualitative case study of a single node network in the prediction of intent and delay. Next, we discuss the quantitative results of the predicted intent to communicate and the delay for (a) the directly connected topology where two users, u and v are connected through an edge; and (b) where u and v are arbitrarily distant in the social network through n hops. Finally we discuss the evaluation of the different contextual features and impact of the communication context and social context on the results.

Qualitative Case Study

We present a qualitative case study of the predicted intent to communicate and associated delay for a particular MySpace user (anonymized, and denoted as u) and her network of eight other users (denoted as users " v_1 " through " v_8 "). This study is shown for four different topic abstractions separately, to understand the specific variations in intent and delay per topic, and also over a period of six weeks. Example messages corresponding to the four topics discussed in these qualitative experiments are shown in Table 4.

The experimental results of prediction of intent for the single node network are shown in a radar plot in Figure 3.10. There are several interesting observations in the results and for each of the four topics we observe distinct dynamics of how the intent to communicate changes over the user u 's set of contacts as time is increased. We categorize the four kinds of dynamics over the four topics respectively as: (a) periodic, (b) personal, (c) habitual and (d) circumstantial. We discuss each of these four types of dynamics as follows.

In Figure 3.10(a) we observe that u 's intent to communicate with her neigh-

Table 3.4: Example messages for the topic abstractions, based on Word-Net synsets used in the experiments.

TOPIC ID	ABSTRACTION/SYNSET	EXAMPLE MESSAGE
TOPIC I	‘event, occurrence’	“I had no idea that it was XX’s birthday tonight. You guys should let me know of such parties and events beforehand.”
TOPIC II	‘movie, music, screenplay’	“XX’s band makes their appearance on the TV today! Don’t miss it out! Their music rocks!”
TOPIC III	‘outing, travel, hike’	“Hi XX would like to invite you guys over to the camping plan! A trip outside the city and some hiking is going to be fun.”
TOPIC IV	‘person, entity, someone’	“The party went off fine. We were eagerly waiting for XX to come though. She would really make a difference!”

borhood i.e. the set of contacts periodically increase and decrease over the period of the six weeks of analysis. Considering the content of the topic, i.e. ‘event, happening, occurrence’, it is intuitive that u and her contacts engage in get-togethers or social events periodically over a period of time—that explains the *periodic* pattern of increase and decrease of intent to communicate.

In Figure 3.10(b) we observe that u has very high intent to communicate for a subset of her contacts, while for the rest it is low. If we analyze the content of the topic, it appears to involve ‘music, band, screenplay’. Hence we conjecture that u shares similar interest in this topic with some of her friends, yielding high intent consistently over the time period of our analysis. While for others who are not as much interested in the topic, she almost never communicates, in the same period—reflecting her *personal* taste.

Figure 3.10(c) shows an extremely regular and almost monotonically increasing intent over the set of contacts for u . We conjecture that u and her contacts were planning on an outing together in the time period of analysis, and hence over time we observe increasing intent to communicate; because it is likely that as the outing neared, it generated more communication. Thus this regularity in increase reflects

Figure 3.10: Qualitative analysis of intent to communicate for a specific single node network. We show the predicted measures of intent over a set of four diverse topics for a user and each of her eight contacts. The visualization shows a radar plot of the measures of intent over a period of six weeks from Feb 19 to Apr 6, 2007.

that u and her contacts have a *habitual* behavior of communicating consistently on planned outings together.

Finally from Figure 3.10(d) we observe that high measures of intent feature on the weeks of Mar 14-Mar 21 and Mar 30-Apr 6. On considering the actual comments, we observe that this artifact arises because u and her contacts suddenly start discussing about a person on these weeks—and that results in high frequency

communication among them. On the other hand, on other weeks we do not observe any pattern in the measures of intent across different contacts. Our conjecture about this kind of dynamics therefore is that it is *circumstantial*.

Next we show qualitative case study of delay involved in the prediction communication for the same user u and her set of eight contacts. The visualization in Figure 3.11 shows u and her contacts (colored circles) over the same period of six weeks from Feb 19 to Apr 6, 2007. The intensity of an edge between the center user and a contact reflects the measure of predicted delay when the center user intends to communication with the particular contact. Lighter shades of the edge reflect long delays; while darker shades imply the delay is short.

We arrive at similar observations as the intent to communicate in the case of the visualizations of delay in Figure 3.11. The first topic appears to have periodic low delays in communication, reflecting the occurrence of an event among the set of contacts and u . For the second topic as well, we observe that for a subset of the contacts the delays are consistently low, representing similarity in personal interests between u and these contacts. In the case of the third topic, we again observe gradual decrease in delay, presumably due to an upcoming outing tour on the part of u and her contacts. And finally for the fourth topic, we observe that the delays are low based on the circumstance at hand—whenever the network is involved in discussions about a particular person, it decreases and vice versa.

Quantitative Evaluation

This subsection presents extensive quantitative analysis on the predicted intent and delay, involving evaluation with respect to alternative baseline techniques discussed previously as well as the actual measures of intent and delay. We are interested in analyzing the predictions from the following angles: the impact of the topology, the impact of number of contacts and the role of the different features proposed in this chapter. Motivated by these ideas, first, we consider analysis in a scenario involving directly connected pair of users; second, we analyze the effect of different neighbor-

hood¹⁰ sizes on the predictions; third, we consider the general scenario when the pair of users involved in communication are arbitrarily distant in the network; and finally, we present an evaluation of the different features corresponding to communication and social context.

Directly Connected Topology. We first consider the case when the pair of users, between which we are interested to predict communication flow, are connected directly through an edge in the MySpace dataset. For this purpose, we present extensive quantitative analysis of predicted intent to communicate and associated delay involve incremental training and testing over all users in the network¹¹, and over all the 25 topics of interest, based on the Support Vector Regression (SVR) method. We present the results of prediction over a period of 40 weeks between Jun 16, 2006 and Apr 24, 2007; and also present evaluation with respect to the baseline techniques and the actual measures of intent and delay.

The mean errors in prediction of intent to communicate over the 40 week period is shown in Figure 3.12. Note that the errors in prediction for both our method and the baseline techniques gradually decrease over time; due to the fact that we are gradually training over larger samples as part of the incremental method of prediction followed for each.

We further observe that for our SVR prediction (i.e. ‘communication flow’) the mean error is between 10-20%; while the error for the epidemic disease model is about 20-35%; for closely followed by the neighborhood effect method—about 25-40%; about 32-47% for the trend-learning method; and finally for the frequency based technique between 33-49%. We show a summary of the mean errors over the 40 week period in Table 3.5.

Our explanation for these observations are as follows. The communication

¹⁰We define neighborhood of a user as the set of contacts she is directly connected to in the network.

¹¹Note that the dataset being sparse, for certain pairs of users the intent to communicate will be zero, while the delay infinity. We assume that communication flow is not defined for these sets of users, and are not considered in computing error in prediction.

Table 3.5: Error in prediction of intent to communicate and delay, for our method ‘communication flow’ and the four baseline techniques.

METHOD	INTENT	DELAY
COMMUNICATION FLOW	17.10%	20.91%
EPIDEMIC DISEASE METHOD	24.78%	26.05%
NEIGHBORHOOD EFFECT METHOD	27.86%	28.59%
TREND-LEARNING METHOD	38.28%	42.10%
FREQUENCY-BASED METHOD	41.71%	44.17%

intent depends on a wide variety of contextual factors (neighborhood, topic, recipient and social) and not just on prior probability of communication on that topic. Hence we see that the four baseline techniques do not work very well in prediction. We believe that our SVR approach captures the important contextual factors at each time slice (through feature selection), yielding effective results. Along the baseline techniques, we observe that the epidemic disease model and the neighborhood effect method have errors that closely follow each other. Our conjecture about this is that both of these methods are dependent upon the communication from the contacts of a user. While, the trend-learning method and the frequency based methods yield similar performance because both are based on the degree of communication of a user in the past.

We observe similar interesting dynamics in the predicted delay as well (Figure 3.13), also shown over the 40 week period. As before, the error gradually decreases over time, implying that incremental training improves prediction. We further observe that compared to the four baseline techniques, our method (i.e. ‘communication flow’) yields least error. A summary of the mean errors over the time period over all techniques is shown in Table 3.5.

However note that in this case, the two baseline techniques, epidemic disease method and the neighborhood effect method follow our method closely in terms of the error in prediction over time. Our explanation for this observation is that it appears that delay is less affected by the contextual features (i.e. communication and social context), and more by the neighborhood of the user; hence the similarity in performance. On the other hand, the trend-learning method and the frequency

based methods, similar to the intent, again perform poorly in prediction—implying that just considering the past degree of communication of the user is not a predictor of her future communication. Overall, it seems that the mean errors are higher for delay than for intent (Table 3.5), that might suggest that contextual feature impact the intent more than the delay. Delay is presumably associated with other factors intrinsic to the user, such as the quality or ‘age’ of interest or factors that are predominantly ‘habitual’ on the part of the user.

Neighborhood Scaling Properties. In this section, we now present the results of the predicted intent to communicate and delay averaged over all 25 topics and over the entire period of 40 weeks of analysis, but with varying neighborhood sizes. The goal is to understand the role of the size of the local neighborhood of user on the prediction results.

We created a set of networks by sampling the MySpace dataset. We used an exponential function: $f(M) = \exp(M/k)$, where $k = 4.6$ and $M = 1, 2, 3, 4, \dots$ to choose networks with node out-degree values $f(M)$. We selected the top five users corresponding to each $f(M)$ based on high message density (number of messages exchanged by the user with neighborhood) from the MySpace dataset. Next, the intent to communicate for each topic and for each of these five users with their individual neighborhoods is determined. The mean intent to communicate per neighborhood size is then reported as the mean over all the five users, the topics and the time period of 40 weeks from Jun 16, 2006 to Apr 24, 2007.

We observe from Figure 3.14(a) that, compared to the actual probability of communication, our prediction method (i.e. ‘communication flow’) outperforms the baseline techniques with a mean error of 18-20%. Interestingly, the predicted intent computed using our method as well as the baseline techniques follow a gradual decay as the out-degree increases. This overall monotonic decrease may be explained as follows. With an increase in neighborhood size, the user may be in regular correspondence with only a small fraction of the contacts. Since we calculate the average over all contacts, i.e. the size of the entire neighborhood of the user, this

leads to an overall decrease in the intent to communicate.

Again, the analysis of (normalized) delay prediction based on the variation of neighborhood size is revealing. We observe from Figure 3.14(b) that with increase in out-degree, the mean delay increases. We believe that this is reflective of the fact that users may only correspond regularly with a small fraction of their neighborhood—since we take the average over entire neighborhood, this is influenced by a majority fraction of the neighborhood with whom the user is not in active communication.

Interestingly, the baseline techniques for delay perform better than the case with intent, and are also able to follow closely the monotonic increase; indicating that the delay in communication may often be due to intrinsic factors (e.g. habitual) and less affected by the contextual factors. Note, that in this case, the errors in delay for our algorithm range between 12% and 20%. The errors for the baseline cases range between approximately 25% and 75%.

Topology with Arbitrarily Distant Users. Now we discuss a generalized scenario of our prediction of communication flow, where the pair of users under consideration can be arbitrarily distant from each other by n number of hops in the social network.

For the purpose of the experimental results, we consider the same sampling procedure of the optimal paths between users in the MySpace dataset. We again used an exponential function: $f(P) = \exp(P/c)$, where $c = 5$ and $P = 1, 2, 3, 4, \dots$ to choose pairs of users who are separated by the number $f(P)(= n)$ users between them.

In the set of experiments reported in Figure 3.15, we consider every user u (who is not an isolate) and for all the paths going out of that node; then we compute n -hop optimal paths for several values of n . Here each value of n corresponds to the prediction of the intent and delay for u with another person v . This is repeatedly conducted for the set of 25 topics and over each time slice in the 40 week period of analysis.

The surface plots in Figure 3.15(a-b) show that for a particular time slice, the prediction errors of the intent and delay increase over the number of hops n . Our explanation for this is intuitive and is presented as follows. Note that, the farther away the pair of users under consideration are, the less likely would they communicate. This is because, MySpace being a social network which most likely reflects *real social ties*, most of the communication on a topic presumably happen in short neighborhoods around a user; however in some rare cases communication also does occur across long distances. Therefore based on our features which mostly exploit the communication and social characteristics within the scope of the sender user and her neighborhood, we can reasonably predict well the communication flow when n is small. As n increases, we conjecture that reasons not quite accountable by our features impact the communication and hence our error shows an increase.

Interestingly though, note that the increase in error for intent is more exponential in nature, than that for delay, which appears more or less linear. It seems to be because of the fact that our predictions of intent are better for smaller values of n , but after n reaches certain large value, the error seems to increase rapidly. On the other hand, delay being more affected by intrinsic and habitual characteristics of the users, the increasing n seems to have only a linear effect on the deteriorating prediction performance.

Feature Evaluation. In this section, we perform evaluation of our set of contextual features used for prediction of the intent to communicate and the delay. We discuss some experiments performed to evaluate the significance of the features in the model.

First we consider several feature selection scenarios to understand the impact of our voting-based feature selection strategy on the predicted intent to communicate and delay. In this paper, we consider five different selection scenarios - (1) use of our proposed dynamic feature selection method, (2) use of all features, (3) use of only communication context features, (4) use of only social context features, and (5) use of a random set of features¹². Experiments were conducted over the 25 topics

¹²For the random feature selection case, we sample five features at random over

as well as the 40 year period for each of these selection techniques. The mean error in prediction of the intent and delay are reported in Figure 3.16.

In the case of prediction of intent, we observe from Figure 3.16(a) that our method of dynamic feature selection yields the best performance, in terms of the least error over the period of analysis. This is followed by scenarios (2), (3), (4) and (5) respectively. Our explanation for this observation is as follows. Scenario (2) yields poorer performance than (1) because it is likely to include features which are not useful in predicting communication between a given pair of users. For scenarios (3) and (4), it seems that the communication context performs better than the social context in the prediction of intent. We conjecture that this is because the probability that a user u would communicate with v are more likely to be affected by the topic at hand and the identity of the recipient than the intrinsic characteristics of u (i.e. information role). Finally we observe that the random selection gives the poorest performance; because the features selected are likely to be ones which play little role in affecting the intent to communicate.

The prediction of delay based on the five feature selection techniques yields slightly different observations as shown in Figure 3.16(b). Our method still yields the least error (scenario (1)), and followed by scenario (2). However in this case, the social context (scenario (4)) seems to perform better than both scenario (3) and (5). We conjecture that this implies that delay is more affected by the social features than the communication context features. Moreover, very interestingly we observe that the random selection of features (scenario (5)) performs better than the communication context alone (scenario (3)). We again conjecture that this happens because delay is likely to be very less affected by the communication context than the social context - and hence leaving out the social context altogether in scenario (3) gives higher measure of error.

As the second part of the feature evaluation, we investigate the different individual features belonging to communication and social context that were selected by our voting-strategy based feature selection technique. A visualization of the 100 independent iterations, and report the mean error.

feature selection over the 40 week period of prediction and averaged over the 25 topics is shown in Figure 3.17. The checkerboard representation shows time on the X-axis while the features on the Y-axis; and the white and black boxes respectively represent whether a particular feature at a particular time slice was selected or rejected by the feature selection strategy.

There are several interesting details in the visualization. First, we observe that certain features, such as reciprocity and information roles are selected consistently at majority of the time slices; reflecting the significance of these features on the prediction. It also implies that the response behavior of the end user and the intrinsic habitual characteristics of the sender user are extremely importance in prediction of communication flow.

Our second observation is that certain features such as backscatter, susceptibility and communication significance are selected most times in the beginning time slices of prediction, than the later. Our explanation for this temporal artifact in the visualization is that, initially when we begin with smaller samples of training data (i.e. training over shorter duration of time), the prediction seems to depend upon mostly the neighborhood based features, due to the lack of sufficient statistics to make a reasonable representation of either the information roles or reciprocity. That is, the features that take into account the communication of a particular user with her set of contacts such as, backscatter, susceptibility or communication significance seem to be good indicators of predicting communication flow, when we do not have a long history of communication between users available.

Third, we observe that as an opposite scenario, certain features like topic quantity and strength of ties appear to be selected more frequently in the later part of our analysis, than the initial time span. We conjecture that both of these features, being dependent upon the longer-term communication and habitual characteristics of the users, presumably turn to be significant in the prediction when we are able to perform training over comparatively longer periods of time.

And finally, as the fourth observation, there seem to be features such as message coherence, topic relevance and communication correlation which do not show

any consistency in their selection over the 40 week period. Intuitively we conjecture that all of these three features, being extremely circumstantial, i.e. depending upon the current communication e.g. what the message is about, how relevant it is to the pair of users' recent communication and how much aligned the two users have been in the recent past, reveal this observation about their inconsistency in getting selected over the time period of our analysis.

3.10 Extended Experiments on Social Network Datasets

In order to demonstrate the extensibility of our proposed framework that predicts communication flow, we attempt to undertake the prediction exercise of determining the intent to communicate and delay for several topics over three social network datasets—Twitter, YouTube and Flickr. Each of these datasets were collected based on a snowball crawl. The number of users in each of the three datasets are: 465,362, 17,736,361 and 4,304,525 respectively. The number of messages (i.e., tweets, comments, replies) that were used to study communication flow are: 29.5M, 145,682,273 and 26,557,446 respectively. In the case of Twitter, we used a set of 25 trending topics to determine communication flow (e.g., 'oil spill', 'iPhone', 'Lady Gaga'). For YouTube, we used the website defined topical taxonomy of videos (e.g., Politics, Music, Comedy) and finally for Flickr we used the group name assignments of photo comments as the corresponding topics (e.g., 'HDR', 'Black and White', 'Nature', 'Arizona Mountains').

First, for the directly connected contacts case, we present the prediction accuracy results of intent to communicate and delay for all three datasets, over a 50 day period (ref. Figure 3.18). We also present the respective prediction accuracies using the same set of topics and over the same time period, but computed using the four baseline techniques. The main observation from the results is that our proposed method (i.e., communication flow model) yields the highest prediction accuracy, whereas the baseline techniques perform significantly poorly by wide difference of accuracies over both intent to communicate and delay. Note that as with the MySpace dataset, in the case of these three datasets too we observe the

prediction accuracy of intent to communicate to be higher compared to delay. This is in conformity with our earlier observations.

Finally, we present the results of predicting intent to communicate and delay for these datasets, when the two individuals involved in communication are arbitrarily distant in the network. This is presented in Figure 3.19. We consider hop distances between 1 and 500 in the network, and predict the measures of intent to communicate and delay for each case over a rolling time window of 50 days. The reported results are for each dataset and averaged over all topics. Note that while the number of hops is smaller the prediction errors are lower; gradually it shows a monotonic increase. We conjecture that this reflects that individuals arbitrarily distant in the network are likely to be motivated to communicate due to factors which are not observable under communication context or social context. Fleshing out intrinsic factors behind communication flow in these situations is an interesting direction for future research.

3.11 Discussion

In this section, we now present a discussion of our method and the experimental results. Going back to our main goal of being able to model and predict communication flow between a given pair of users and a given topic, we have developed a user context based framework that measures two variables—intent to communicate and associated delay between users to determine efficient communication paths. While our method seems to yield promising results in prediction, the results shed light into several associated interesting artifacts of communication flow in online social networks, as discussed below.

Role of contextual features. From the extensive qualitative and quantitative results conducted in this chapter, our primary observation is that intent to communicate shows lower prediction error across time as compared to delay. This might reveal that factors such as external events, habit (e.g. activity schedule, time zone etc), intrinsic or unobservable factors (e.g. external sources of trigger on a cer-

tain piece of information, real social relationships etc) and also personal contextual changes presumably affect delay more than intent. While our proposed contextual features can reasonably predict intent, delay also seems to be dependent other latent factors, like ‘age’ of information. For example, a topic, if old enough in a user’s neighborhood, is likely to be transmitted less than something which is more current and new. Besides, this also indicates that the intent and delay are presumably orthogonal parameters in characterizing communication flow.

Our results also reveal that the topic under consideration has significant effect on communication flow - an issue which has not been addressed deeply in majority of the prior work. As is evident from our qualitative analysis, different kinds of topics, for the same set of users, can have dramatically different intent and delay. This leads us to infer that the content plays a key role in a person’s communication and whether or not she is likely to transmit it. More sophisticated methods to model multiple social and information dimensions of topics in the future can therefore be interesting to boost the understanding on communication flow.

Effect of network connectivity. We observe from our experimental results that the network connectivity of a user affects significantly her desire to propagate information to another user; as the connectivity increases, i.e. the neighborhood size increases, the overall intent to communicate of a person appears to decrease considerably. However, depending on the topic, the user presumably communicates with a high probability, that can result in high communication flow. On a different note, the increase in delay as a function of the connectivity reflects an intrinsic aspect of these online social networks; which is the increased response times of users due to the overwhelming availability of multi-faceted content. Being able to model the relationship with information availability and the response behavior of users can be worth investigating in the future.

Effect of “social distance”. Our results also indicate that as the number of hops, i.e. the social distance between the pair of users in consideration increases, the

prediction errors for intent and delay also increase. This is an intuitive artifact on social networks because with the increase in social distance, first, the information loses to be interesting to new users; and second, the delay in transmission usually increases to such a considerable extent along large social distances, that it might generate a new communication path altogether. It would therefore be interesting to further investigate branching of the communication paths as the information grows older in a social network.

Temporal evolution of user behavior. In this chapter, though we propose an incremental training based prediction framework for communication flow, it is possible and very likely that the contextual features themselves evolve over time, due to evolution of user behavior. Therefore, it would be interesting to focus on the temporal evolution of communication context, possibly using a partially observable Markov decision process. The generic model of temporal evolution thus presents the scenario where the users engaged in communication are assumed to exhibit communication behavior which are only partially observable to us e.g. a user who travels to different locations frequently, a user who have different schedules of communicating online or a change in the real social relationship between a pair of users etc. The goal of such a model is then, to be able to infer the hidden contextual state related to the evolved behavior, based on the observed communication.

Role of external triggers / content popularity. An important aspect of how communication flow occurs in a social network is attributed to the nature of the content being transmitted - i.e. whether it was a result of a trigger from an external happening, or it is related to a theme that is extremely popular. In this chapter, we have modeled predicting the communication paths as a function of a topic (which is a WordNet abstraction over messages discussed), however we have not account for its relationship to an external event or its intrinsic popularity that might impact its propagation immensely. Hence modeling the nature of the content is one of our interests for future research.

Topology of communication paths. Moreover, prior work and our empirical observations on large-scale online social networks e.g. Facebook, Twitter, Digg etc lend the insight that the typical topology of communication paths is presumably quite short—often triggered at multiple user ends, due to an external happening; and information flow happens along these paths due to the collision among several similar paths. Although we have not investigated our results in this direction, yet the decrease in prediction performance over the number of hops n , between two users, reveals that likely MySpace also features such short-length topology of the communication paths. However without sufficient empirical and computational backing to support this observation, we leave this investigation to future research directions.

Diffusion rates in relation to communication flow. Communication paths in social networks are often associated with a temporal factor - either how frequently a topic diffuses in a network via communication, or how easily a network gets "infected" with an information. We have not addressed communication flow in these directions, however we have addressed a related temporal artifact; which is how long it will take for an information to travel from one user to another, i.e. our measure of delay in predicting communication flow. As future research directions we would be interested to connect this measure of delay with models of network infection as well as the rate at which a user is responsible for diffusing a particular piece of information via her communication.

3.12 Conclusions

With consumers showing increasing resistance to traditional forms of advertising such as TV or newspaper advertisements, marketers have turned to alternate strategies, including viral marketing. Viral marketing exploits existing social networks by encouraging customers to share product information with their friends. Previously, a few in depth studies have shown that social networks affect the adop-

tion of individual innovations and products. But until recently it has been difficult to measure how influential person-to-person communication actually is over a wide range of topics. We also note that online social networks have become an important medium of communication and information on the web. Due to their accessible and timely nature, they are also an intuitive source for data involving the spread of information and ideas. By examining linking propagation patterns from one blog post to another, we can infer answers to some important questions about the way information spreads through a social network over the web.

In this light, we are motivated to study the characterization of communication between pairs of individuals in social networks. In this chapter, we have developed a novel computational framework to predict communication flow along efficient communication paths in a large-scale social network. Our key contribution lies in characterizing these communication paths via two measures—the intent to communicate and the associated delay.

Our method is based on modeling the user context; which we ascribe, communication and social context. To elaborate on these two context types, we first identified three aspects that affect communication on a specific topic: (a) neighborhood context, (b) topic context and (c) recipient context. We thereafter modeled features belonging to social context, that include information roles acquired by the users, as well as the overlap or strength of ties with their local neighborhood, i.e. the set of contacts. The intent to communicate and communication delay were then estimated using Support Vector Regression over a dynamically selected set of contextual features, for a pair of users directly connected in the network. We also proposed a method to compute the intent and delay for cases when the pair of users are arbitrarily n hops distant in the social network using an approximation of the restricted shortest path problem.

From the qualitative and quantitative experimental results conducted on four large datasets—MySpace, Twitter, YouTube and Flickr, we observed qualitatively and quantitatively that our proposed context modeling based method yields low errors in predicting the intent and the delay; with as much as $\sim 40\text{-}50\%$ boost

over other baseline techniques. We also conducted extensive empirical studies with promising results via varying the (a) size of the neighborhood of users, (b) number of hops (or social distance) between the pair of users under consideration, when they are not directly connected through an edge, and finally (c) choosing diverse feature selection scenarios. Our results lend the insight that the intent is observed to be more affected due to contextual dynamics than delay. Delay seemed to be more dependent on other latent factors characterizing communication, e.g. ‘age’ of information, or intrinsic user behavior and habits.

In closing, we have presented in this chapter a comprehensive discussion of how efficient communication paths can be predicted in large-scale social networks by leveraging the user context. Nevertheless, naturally our results indicate the scope of ample future work, many of which have been elaborately discussed here, and in summary include user behavior modeling, modeling rates of diffusion as well as investigating the correlation between communication flow and social distances or the topology of the communication paths.

(a) Topic – ‘event, happening, occurrence’

(b) Topic – ‘movie, music, band, screenplay’

(c) Topic – ‘outing, travel, hike’

(d) Topic – ‘person, entity, someone’

Figure 3.11: Qualitative analysis of delay in communication for a specific single node network. We show the predicted measures of delay over a set of four diverse topics. The visualization shows a user and her local network comprising eight contacts (colored circles) over a period of six weeks from Feb 19 to Apr 6, 2007. The intensity of an edge between the center user and a contact reflects the measure of predicted delay when the center user intends to communication with the particular contact. Lighter shades of the edge reflect long delays; while darker shades imply the delay is short.

Figure 3.12: Error in prediction of intent to communicate over the MySpace dataset. The results are shown over a period of 40 weeks spanning from Jun 16, 2006 to Apr 24, 2007 and are averaged over all 25 topics. Comparison with four other baseline methods indicates that our method yields least error.

Figure 3.13: Accuracy in prediction of delay in communication over the MySpace dataset. The results are shown over a period of 40 weeks spanning from Jun 16, 2006 to Apr 24, 2007 and are averaged over all 25 topics. Comparison with four other baseline methods indicates that our method yields highest accuracy.

(a)

(b)

Figure 3.14: Comparison of intent, (a) and delay, (b) predicted using our method and four other baseline methods against actual intent and delay. The values are shown over a set of different neighborhood sizes of MySpace users, varying from 1 to 1000. These results are averaged over the time span of 40 weeks of analysis, ranging from Jun 16, 2006 to Apr 24, 2007 and also averaged over all 25 topics.

Figure 3.15: Predicted intent to communicate and delay as a function of the number of hops between a pair of users, shown over the period of 40 weeks from Jun 16, 2006 to Apr 24, 2007 and averaged over all 25 topics.

Figure 3.16: Comparison of different feature selection types on the prediction performance of intent to communicate and delay from Jun 16, 2006 to Apr 24, 2007 and averaged over all 25 topics. We compare five different classes of feature selection: (1) use of our proposed dynamic feature selection method, (2) use of all features, (3) use of only communication context features, (4) use of only social context features, and (5) use of a random set of features. The colorbar represents the correspondence between the color of the blocks in the figure, and the measure of error. We observe that our method yields best performance.

Figure 3.17: The different features belonging to communication and social context that were dynamically selected over a period of 40 weeks from Jun 16, 2006 to Apr 24, 2007 and averaged over all 25 topics. A white box represents that the feature was selected at the corresponding time slice (shown on the X-axis) while a black one indicates that it was not.

Figure 3.18: Accuracy in prediction of intent to communicate and delay on three social network datasets—Twitter, YouTube and Flickr. The results are shown over a period of 50 days and are averaged over all topics in each dataset. Comparison with four other baseline methods indicates that our method yields highest accuracy.

Figure 3.19: Predicted intent to communicate and delay as a function of the number of hops between a pair of users, shown over the period of 50 days and averaged over all topics, corresponding to the three different datasets.

Chapter 4

Diffusion of User Actions: Impact of Social Influence

“The way of the world is meeting people through other people.”—Anonymous.

We propose a computational framework to predict diffusion of user actions, referred to as “synchrony” of actions in online social media. Synchrony is a temporal social network phenomenon in which a large number of users are observed to mimic a certain action over a period of time with sustained participation from early users. Understanding social synchrony can be helpful in identifying suitable time periods of viral marketing. Our method consists of two parts—the learning framework and the evolution framework. In the learning framework, we develop a DBN based representation that includes an understanding of user context to predict the probability of user actions over a set of time slices into the future. In the evolution framework, we evolve the social network and the user models over a set of future time slices to predict social synchrony. Extensive experiments on Twitter and Digg datasets (comprising $\sim 7M$ diggs and $\sim 29.5M$ tweets) show that our model yields low error ($15.2+4.3\%$) in predicting user actions during periods with and without synchrony. Comparison with baseline methods indicates that our method shows significant improvement in predicting user actions.

4.1 Introduction

This chapter presents a framework for predicting social synchrony in online social media. By synchrony, we mean the tendency of a large group of people to perform similar actions in unison, in response to a contextual trigger. Consider the familiar observation from a performance in an auditorium. When a small set of individuals starts to clap, the rest of the audience follows. This lasts for a short period of time, till the claps die. In Nature, biological oscillators, including fireflies

are often observed to fire light at the same time, triggered by a few of them. This phenomenon of oscillating light continues for a certain period of time. Note in both examples, the population moves to a state of sync with respect to a certain action (clapping and light oscillation), in response to trigger by a small set of participants.

Do social media websites including Facebook¹ and Digg², exhibit this temporal property of sync? These sites provide a variety of user affordances including observing their friend behavior, commenting, and posting / sharing media content. For example on Facebook, certain types of behavior can be seen to be performed in near unison—e.g. taking of online quizzes, using certain applications to send message to one’s contacts; while on Digg, this can include liking or sharing stories around a particular topic. There is typically a contextual trigger (typically one’s contacts or a set of users the user is familiar with), that prompts similar behavior from the user. An important motivation for this chapter is to identify the conditions under which large number of users in a social network exhibit synchrony in mimicking a certain action.

The synchrony problem is related to, but distinct from, the problem of network cascades (Leskovec et al., 2007; Watts, 2002). Unlike a cascade, in the examples discussed above, the original set of seed users who trigger the behavior continue to participate in performing the same set of actions. Cascades do not include the idea of performance in unison, or the idea of sustained participation. Like cascades, synchrony must attract new users to act along with the existing users.

Understanding social synchrony can be of utility in several diverse domains. Corporations might be interested to know what could be the suitable time periods to market a particular product. Predicting time periods involving high user participation in a network can also be useful in resource allocation and management. It can also help us understand the responses of user groups to certain types of events.

¹<http://www.facebook.com>

²<http://www.digg.com>

Related Work

Although social synchrony has not been studied in prior work, there has been work on related ideas in peer disciplines which have dealt with the response of a social network to dynamic information—epidemiology and information flow (Adar and Adamic, 2005; Gruhl et al., 2004; Leskovec et al., 2007), social information cascades (Cha et al., 2008; Leskovec et al., 2007; Watts, 2002), social correlation (Anagnostopoulos et al., 2008) and social recommendations (Fernandess and Malkhi, 2008; Leskovec et al., 2007).

Gruhl et al (Gruhl et al., 2004) develop a topic propagation framework and use a model of infectious diseases to study information diffusion. Adar and Adamic in (Adar and Adamic, 2005) study the phenomenon of information epidemics in the blogosphere based on the idea of propagation of memes. Leskovec et al in (Leskovec et al., 2007) use the property of sub-modularity to develop an optimization algorithm that detects points of origin of outbreaks in networks. In (Watts, 2002) Watts analyzes the conditions under which global information cascades occur, based on a simple threshold idea of changing user states. Leskovec et al in (Leskovec et al., 2007) attempt to understand how realistic cascades emerge in blogs and study information propagation in the blogosphere.

The prior work address issues such as information flow or emergence of social cascades well. However, they assume that user participation in an action occurs at a single point in time without continued participation, and without the idea that addition of new users affects the dynamics of the phenomena over time. In online social media, these assumptions are not always true - for example, a user on Digg can continually ‘digg’ news stories on a topic over a period of time, and this continued participation can impact other new users in the network to participate as well. To the best of our knowledge, this chapter is one of the first attempts to study such temporal properties characterizing synchrony in social networks.

Our Approach

There are two key contributions in this chapter:

1. First, we propose an operational definition of synchrony of user actions in online social media. Our definition incorporates the following ideas—presence of a specific topic, an agreed upon action, a seed set who triggers an action and high frequencies of continuing old users as well as new users over a period of time.
2. Second, we develop a computational model to predict emergence of synchrony over a period of time. It uses a learning framework and an evolution framework. In the former, a dynamic Bayesian representation of a user based on latent states and user context is developed to predict her probability of a specific action at a certain point in time. In the latter the social network of users is evolved to account for varying network sizes and user models over time.

We have conducted extensive experiments on large datasets crawled from Twitter and Digg. The results show that our model yields low error ($15.2+4.3\%$) in predicting user actions (diggings/tweeting activity) during time periods with and without synchrony. Comparison with baseline techniques indicates that our method shows significant improvement in prediction of user actions in the range 18-56%.

4.2 What is Social Synchrony?

We now examine a popular real world social network Digg. Users on Digg can vote on news stories posted on an external website (popularly known as ‘digging’). They can also express their disapproval over items by ‘burying’ them.

As an example, consider the pool of stories on two different topics on Digg—‘Olympics’ and ‘Celebrity’. Figure 4.1 shows the number of new users (i.e. the users who digg a story on the particular topic for the first time) on each day over the month of September 2008 and continuing old users (i.e. the users who have *dugg* a story on the topic before) across each day for each of the two topics.

We observe that the topic ‘Olympics’ is characterized by a time period of large number of new users (Sept 3-Sept 13). During this period, we also observe considerable numbers of continuing old users. On the other hand, in case of the topic

Figure 4.1: Dynamics of the size of user set on two topics from Digg—‘Olympics’ and ‘Celebrity’ (September 1-30, 2008). Topic ‘Olympics’ is observed to exhibit synchrony where old users continue to be involved in the action of digging stories, as well as large number of new users join in the course of time (Sept 3-Sept 13).

‘Celebrity’ we observe that the number of continuing users is rapidly decreasing over time and the rate of new users joining the existing user set is nearly constant. Hence in the case of ‘Olympics’, the rate of new users getting involved in the action of digging is high, as well as several old users continue to perform the action of digging stories on the topic. We conjecture that the topic ‘Olympics’, unlike ‘Celebrity’ exhibits social synchrony (over the action of digging) between Sept 3 and Sept 13, in a manner similar to the earlier examples of clapping and light oscillation in the previous section.

These examples indicate that synchrony given a topic would involve two properties: (a) sustained participation from old users over time in perform a specific action—we call this property *sustainability* α , and (b) attraction of large number of new users over time to perform the same action—we call this property as the *rate of attraction* β . The measure of sustainability α is therefore given by the ratio of the number of common users across consecutive time slices, to the total number of users in the previous time slice. Similarly, the rate of attraction β is further given by the ratio of the number of new users at a time slice, to the total number of unique users across all time slices.

Figure 4.2: Conceptual representation of synchrony in a social network over three time slices: t_1 , t_2 and t_3 . Synchrony is likely to involve (a) seed users, (b) sustained participation from old users (shaded circles) as well as attract new users (white circles) to perform a specific action over a period of time.

Definition. Social synchrony is therefore a temporal phenomenon occurring in social networks which is characterized by (a) a certain topic (including a meaningful theme), (b) an agreed upon action(s) that the users in the network can perform with respect to the topic, (c) a set of seed users who are involved in performing the action at a certain point in time, and (d) large numbers of continuing old users as well as new users getting involved in the action over a period of time in the future, following the actions of the seed set (i.e. sustainability $\alpha \sim 1$ and rate of attraction $\beta \gg 1$).

We explain the idea of social synchrony in the conceptual representation in Figure 4.2. Starting with a small set of seed users, the figure shows how large numbers of new users join the initial set in performing a certain action (white circles). It further shows sustained participation from the old users (shaded circles) over a period of three time intervals.

4.3 Problem Definition

Now we present our data model, followed by the problem statement and finally, the key challenges.

Data Model. Our data model is based the popular news-sharing social media,

Figure 4.3: Representation of the two conditions (via two latent states) under which a user can perform a specific action—socially aware (e.g. digging on the Digg website), denoted by ‘1’ and socially unaware states (e.g. digging from an external site like the NY Times), denoted by ‘0’.

Digg. News stories on Digg are organized into a two-level Digg-defined content taxonomy—the higher level theme being called a ‘container’; while each container further comprising a number of ‘topics’. To study social synchrony on Digg, we focus on stories at the granularity of a topic. Each user on Digg can have a set of contacts, whose digging actions on stories of different topics can be viewed in the user’s Digg profile through feeds.

It is important to note that a user can digg a story under two conditions. First, the user can be present on the Digg website while digging an item. This can be described to be a socially aware state because her action of digging a story in this case is likely to be impacted by the actions of other Digg users. Second, she can digg it from the source website of the news story itself (e.g. digg a news story on the NY Times website), which we call the socially unaware state. In this chapter we will use the symbol ‘1’ to represent the socially aware state, and ‘0’ to represent the socially unaware state (ref. Figure ??). Note the state of the user is hidden—we have no knowledge about the condition under which a particular item was dugg by a user.

Besides the action of digging, users can further engage in communication by posting comments on an already dugg story (i.e. a story that has been voted upon earlier by other users), or involve in discussion with other users via replies to existing comments.

Although we focused on the social media Digg in this chapter, our framework can be easily extended to other social networks with evidence of user actions

and user communication. Several other social media sites like Flickr, del.icio.us, YouTube, Yahoo! Buzz, Reddit, StumbleUpon, and social networks like Facebook and MySpace also allow users to execute specific actions both on and off their website - hence the notion of the two conditions under which users can perform a specific action is also extensible.

Problem Statement. We now frame our problem statement as follows. Given:

1. a topic ζ and a social network $G(V, E)$ where V is the set of users and E is the set of edges between them. An edge $e_{uv} \in E$ exists if u and v are contacts of each other;
2. a set of actions $\Psi : a_1, a_2, \dots$ and a set of communication types $\tau : x_1, x_2, \dots$ that can be performed by each user $u \in V$ over the topic;
3. a seed set of users U_0 who perform actions, $a_i \in \Psi$ with respect to the topic ζ at a certain time slice t_0 ; and
4. a history of actions a_i and communication x_i (comments and replies among users in V) on the same topic ζ and over a period of T time slices (prior to t_0) corresponding to each user, $u \in V$,

we are interested in predicting whether the network G will exhibit social synchrony (operationally defined in the previous section) over M future time slices after t_0 , i.e. $t_1 - t_M$ on the same topic.

Key Challenges. Predicting an emergent social synchrony with respect to a topic and over a set of M future time slices involves the primary challenge of determining the temporal evolution of the social network G with respect to the predicted actions performed by its users (Figure 4.2). This, in turn, involves the following sub-challenges:

1. *Learning:* for each user in the social network, we need to predict her probability of actions at each future time slice (Figure 4.4). The probability estimate of

Figure 4.4: Illustration of the learning and evolution problems in prediction of synchrony.

user action should take into account the condition under which the action is performed (socially aware / unaware state) as well as how her context affects it. The user context should further incorporate the effect of the actions of other users in the network in the previous time slice, degree of coupling with the seed set with respect to the specific action, as well as the user’s own history of actions and communication.

2. *Evolution:* recall, synchrony in a social network (a) is likely to involve sustained participation from old users as well as attract new users to perform a certain action; and (b) persists over a period of time. Hence, we need to determine the network size at each of the M future time slices, choose the set of users in each of the evolved networks, and finally determine estimates of the probability of actions of all users in the evolved networks at each future time slice. These probability estimates would indicate the presence (or absence) of an emergent synchrony in the future for the particular topic (Figure 4.4).

In the following two sections, we discuss our framework that addresses each of the above challenges.

4.4 Learning Framework

In this section we present a framework for predicting the probability of actions of each user in the social network, given a topic, at each of the M future time slices. We need to solve four related sub-problems: (a) represent the user actions as a dynamic Bayesian network (DBN), (b) estimate the user context, (c) predict probabilities of the user states, and (d) predict probabilities of her actions.

Figure 4.5: Dynamic Bayesian Network representation of user actions using user context and states.

DBN Representation

A user's intent to perform an action (on a certain topic) can be triggered due to her being either socially aware or socially unaware of the actions in her local social network i.e. her immediate neighbors. Note each of the two states, in turn, would be affected by the user context (e.g. actions of the neighboring contacts, coupling with seed users and / or the user's communication over the topic).

We represent the temporal dependencies among all these variables—user actions, states and contextual attributes via a dynamic Bayesian network shown in Figure 4.5. Assuming first order Markov property, the DBN yields the following time-based inference equation for determining the probability of user action at a time slice $t_j (1 \leq j \leq M)$:

$$\begin{aligned}
 & P(A_{u,j} | \mathbf{A}_{u,j-1}, \mathbf{C}_{u,j-1}) \\
 &= \sum_{S_{u,j}} P(A_{u,j} | S_{u,j}, \mathbf{A}_{u,j-1}, \mathbf{C}_{u,j-1}) \cdot P(S_{u,j} | \mathbf{A}_{u,j-1}, \mathbf{C}_{u,j-1}) \\
 &= \sum_{S_{u,j}} P(A_{u,j} | S_{u,j}) \cdot P(S_{u,j} | S_{u,j-1}, \mathbf{C}_{u,j-1}),
 \end{aligned} \tag{4.1}$$

where $A_{u,j}$ is the action performed by user u at time t_j , $S_{u,j}$ is the state of u at time t_j , $C_{u,j-1}$ is the context of user u at the previous time slice, the vectors $\mathbf{A}_{u,j-1}$ and $\mathbf{C}_{u,j-1}$ represent $A_{u,1}, A_{u,2}, \dots, A_{u,j-1}$ and $C_{u,1}, C_{u,2}, \dots, C_{u,j-1}$ respectively. We show how each of the probabilities are estimated in the following subsections. First we present how we estimate user context.

User Context

We discuss the different attributes of the user context that are associated with the two latent user states. Let us denote the contextual attributes at time slice t_j corresponding to the socially aware state by $C_{u,j}^1$. The different attributes in $C_{u,j}^1$ are as follows:

1. A user is in the socially aware state would be affected by the actions of her neighborhood, i.e. the set of immediate contacts (Figure 4.6(a)). Hence the first contextual attribute is the mean probability of actions over user u 's neighborhood in the previous time slice. This is denoted as $\eta_{u,j-1}$.
2. A user u in the socially aware state would also be affected by the coupling $\omega_{u,j-1}$ with the seed set of users U_0 in the previous time slice (Figure 4.6(b)). It is given by the ratio of how many times u followed v 's actions on a topic in the previous time slice, to the total frequency of u 's actions at that time slice.
3. The prior probability of u 's own actions $p(A)_{u,j-1}$ in the previous time slice; and
4. The user u 's intrinsic interest on the particular topic in the previous time slice—her probability of communication. This is estimated via the frequency of commenting, $p(\kappa_{u,j-1})$ and replying, $p(R_{u,j-1})$.

The attributes in $C_{u,j}^0$ associated with the socially unaware state are also determined in a similar manner. However since the user does not have knowledge about her neighborhood or the seed users in this state, hence $\eta_{u,j-1}$ and $\omega_{u,j-1}$ are assumed to be equal to zero.

Estimating User States

We now present our method of estimating the probability of the user states.

Let us assume that the probability of a certain user state at t_j given the previous state and the user context at t_{j-1} ($P(S_{u,j}|S_{u,j-1}, C_{u,j-1})$) is given by a probability density function whose parameters are unknown to us. To estimate this

Figure 4.6: (a) Effect of a user's neighborhood (i.e. contacts) on her actions, shown by the directed edges. (b) Coupling of a user's actions with the seed users. The degree of coupling is represented using the thickness of an edge. Note the semantics of the two edges are different—a seed user in (b) may or may not be a contact of u in (a).

probability for a specific set of contextual attributes in $C_{u,j-1}$ our goal is therefore to determine the model parameters of the pdf that describes this conditional probability. In this chapter we assume that the user states $S_{u,j}$ at t_j have a multinomial density (Zhai and Lafferty, 2004) over the contextual attributes in $C_{u,j-1}$ (with parameter $\phi_{u,j-1}$), along with a conjugate Dirichlet prior (Zhai and Lafferty, 2004) over the previous state $S_{u,j-1}$ (with parameter $\gamma_{u,j}$). This is given by,

$$\begin{aligned}
 P(S_{u,j}|S_{u,j-1}, \mathbf{C}_{u,j-1}) &\propto P(\mathbf{C}_{u,j-1}|S_{u,j}) \cdot P(S_{u,j}|S_{u,j-1}), \\
 P(\mathbf{C}_{u,j-1}|S_{u,j}) &= \mathbf{multinom}(\text{vec}(\mathbf{C}_{u,j-1}); \phi_{u,j-1}), \\
 P(S_{u,j}|S_{u,j-1}) &= \mathbf{Dirichlet}(\text{vec}(S_{u,j}); \gamma_{u,j}),
 \end{aligned} \tag{4.2}$$

that is,

$$\begin{aligned}
 P(\mathbf{C}_{u,j-1}|S_{u,j}) &= \frac{\sum_{jk} \mathbf{C}_{u,j-1;jk}!}{\prod_{jk} \mathbf{C}_{u,j-1;jk}!}, \\
 P(S_{u,j}|S_{u,j-1}) &= \frac{1}{B(\gamma_{u,j})} \prod_{jl} S_{u,j}^{S_{u,j-1;jl}},
 \end{aligned} \tag{4.3}$$

and $B(\gamma_{u,j})$ is a Beta function with the parameter $\gamma_{u,j}$. Based on eqn. 4.2, let $\lambda_{u,j-1}$ denote the parameters corresponding to the pdf of $P(S_{u,j}|S_{u,j-1}, \mathbf{C}_{u,j-1})$. Our goal is to maximize the likelihood of $\lambda_{u,j-1}$ based on the temporal relationships between user states and context in the DBN in Figure 4.5:

$$\lambda^* = \arg \max_{\mathbf{C}_{u,j-1}} \log P(S_{u,j}|S_{u,j-1}, \mathbf{C}_{u,j-1}). \quad (4.4)$$

Assuming independence between the user state and the user context at the same time slice, and using first order Markov property in our DBN, it is sufficient to maximize the corresponding log-likelihood function in eqn. 4.4:

$$L(\lambda) = \log P(\mathbf{C}_{u,j-1}|S_{u,j}) + \log P(S_{u,j}|S_{u,j-1}). \quad (4.5)$$

We compute the above log likelihood using maximum a posteriori estimation (MAP) to get optimal estimates of the parameters $\lambda_{u,j-1}$. Substituting these optimal weights in eqn. 4.2 along with the actual values of $S_{u,j-1}$ and $\mathbf{C}_{u,j-1}$ gives the conditional probability $P(S_{u,j}|S_{u,j-1}, \mathbf{C}_{u,j-1})$ for the user u at time slice t_j .

Estimating User Actions

Now we discuss the method of computing the second conditional probability of eqn. 5.3, that is, the probability of user action given the user state, $P(A_{u,j}|S_{u,j})$.

To estimate this probability, it is sufficient to determine the emission probabilities of actions given the states, that is, $b_u(A|S)$ learnt for user u over the training time period of T time slice prior to t_0 . Since the user action distribution during the training period is known and the states are hidden, we can use a generative model (Jung et al., 2000) to decode the state sequence given the observed actions and thereby obtain optimal estimates of the emission probabilities b_u . Note, using a straight-forward Hidden Markov Model in this case to determine the emission probabilities of actions is not suitable because the state transition probabilities are affected by the user context. Finally we compute the conditional probability of user action given the user state, $p(A_{u,j}|S_{u,j})$ as follows:

$$P(A_{u,j}|S_{u,j}) = b_u(A|S). \quad (4.6)$$

To conclude this section, we substitute the probabilities from eqn. 4.2 and eqn. 4.6 into eqn. 5.3 to determine the probability of actions for each user u in the

social network at each time slice t_j , where $1 \leq j \leq M$.

4.5 Evolution Framework

Synchrony is a temporal phenomenon that persists over a certain time period. Hence online learning methods (e.g. incremental SVM Regression) that incrementally train and predict a value at each time slice, are not helpful in our case—we need estimates of the probability of actions of users in a social network for a set of M future time slices altogether.

To tackle this problem, we propose an evolution framework. First, we need to estimate the size of the social network for each of the M future time slices. This is because it is likely that certain users would leave and some would join over time in performing the specific action. Second, we need to determine which user models to use at the M time slices. Finally we need to determine a measure of the probability of actions for each of the evolved networks to predict an emergent synchrony.

Estimating Network Size

Synchrony is likely to involve sustained participation from old users, given by the parameter, sustainability α , as well as attract newer users to perform the action, given by the parameter called the rate of attraction β . However note that the values of α and β are not available to us over the set of M time slices wherein we intend to predict synchrony. Hence we have to learn the values of α and β in order to estimate the size of the social network at each future time slice. We make a simple conjecture that α and β in a small time window in the past (say, q time slices before t_1 , where t_1 is the time slice of the start of synchrony prediction) would reflect how the size of the user set is changing over time. Using the mean rates of α and β over the prior q time slices, we predict the number of users at $t_j (1 \leq j \leq M)$ to be approximately the sum of the number of continuing old users and the new users:

$$|U_j| \approx \alpha \cdot |U_{j+1}| + \beta \cdot |\cup_{i=1}^{j-i} U_i|. \quad (4.7)$$

Given the sizes of the evolved networks over the M future time slices, we discuss the evolution of user models in the next sub-section.

Evolving User Models

Now we need to determine the specific set of users $\{U_1, U_2, \dots, U_M\}$ corresponding to each of the evolved networks. Note, the size of network at a time slice t_j could be related to that at t_{j-1} by two relationships— $|U_j| \leq |U_{j-1}|$, or $|U_j| > |U_{j-1}|$. We construct the user set for these cases in the following manner:

1. In the first case, U_j comprises the subset of users from U_{j-1} who had the maximum mean probability of actions at t_{j-1} .
2. In the second case, we add those users to U_j who have the maximum probability of actions over all other topics at t_{j-1} .

Given the user sets over the M time slices, we now predict their probability of actions at each time slice—that is, evolve the user models. Note our learning framework gives estimates of the probability of actions of a user at t_j , given the previous actions and user context $C_{u,j-1}$ at t_{j-1} . In our set of evolved networks, since the actual values of the contextual attributes are not available for time slices after t_0 , we project estimates of the user context over each of the M time slices in order to predict a user's probability of actions.

From the previous section we know that the contextual attributes for the socially aware state is given by, $C_{u,j}^1 = (\eta_{u,j-1}, \omega_{u,j-1}, p(A_{u,j-1}), p(\kappa_{u,j-1}), p(R_{u,j-1}))$ at a certain time slice t_j . The attributes $\eta_{u,j-1}$ and $\omega_{u,j-1}$ can be updated based on the predicted actions of the user u 's contacts and those of the seed users in the previous time slice. Similarly $p(A_{u,j-1})$ can be updated based on the predicted value of u 's actions. The estimates of the frequency of comments and that of replies from u are not available to us; hence the two attributes $p(\kappa_{u,j-1})$ and $p(R_{u,j-1})$ are held constant over the period of the M time slices. Note, for the contextual attributes $C_{u,j}^0$ associated with the socially unaware state, we are only able to update $p(A_{u,j-1})$.

Based on the predicted actions and the predicted estimates of the user context, we use our learning framework to compute measures of $p(A_{u,j})$ at each time slice $t_j, 1 \leq j \leq M$ and thereby evolve the user models over the M time slices.

Predicting Synchrony

Based on the evolved networks sizes and evolved user models that comprise these networks, we now use the predicted probability of actions of the user to quantitatively predict an emergent synchrony in the set of M future time slices. Our main idea is that if the mean probability of actions of all users in $\{U_1, U_2, \dots, U_M\}$ over each of the M future time slices is very high, it implies (quantitatively) an emergent social synchrony in the evolved social networks. In particular, on Digg this implies a large digging probability.

However, note that social synchrony being a social network phenomenon, we are only interested in those users who are engaged in the actions given the socially aware state. Hence the quantitative measure of social synchrony for a certain topic over a set of M future time slices is given by the vector $(p_1, p_2, \dots, p_M)$ where p_j is the mean probability of actions for all users $u \in U_j, (1 \leq j \leq M)$, given u is in the socially aware state.

Now we briefly summarize our computational framework. We proposed a two-stage method (learning and evolution) to predict emergent synchrony in social networks. The learning framework yields estimates of user actions given her context and previous actions. In the evolution framework, we have developed a framework which can evolve network sizes over the set of M future time slices as well as evolve user models to predict the probability of actions of users. These predicted probabilities are then used to predict synchrony.

4.6 Experimental Results

In this section we present our experimental results comprising a description of the dataset, quantitative analysis of synchrony, comparative study of our method and finally some open issues.

Digg Dataset

The dataset used for the experiments is crawled from the social media Digg. We seeded our crawling from the stories in the featured category ‘Popular’ on the Digg website. We crawled all stories in this list and which submitted over August and September 2008. We identified the unique users from these set of stories and constructed their degree distribution (i.e. the number of contacts). From the degree distribution, we picked a set of 500 users with the highest degrees. We crawled all the stories, diggs, comments, replies submitted by them over the two months and collected their contacts. We used a snowballing technique, i.e. iteratively followed this procedure for a set of 21,919 users. In total, this dataset comprises 187,277 stories, 7,622,678 diggs, 687,616 comments and 477,320 replies over a set of 51 topics in this time range.

Now we present a quantitative analysis of our prediction results from two different perspectives—prediction of synchrony and relationship of certain properties of the seed set to synchrony.

Prediction of Periods of Synchrony

We discuss an analysis of the effectiveness of our model in predicting user actions over time, and relationship to presence / absence of synchrony over a set of topics. We focused on six different topics—‘US Elections’, ‘World News’, ‘Olympics’, ‘Comedy’, ‘Celebrity’ and ‘Tennis’ to demonstrate the performance our prediction. Each of these topics has a total number of 14,245, 23,935, 10,732, 8,356, 4,735 and 6,774 users respectively. The rationale behind choosing these topics was to test our model on a diverse range of topics—some of which were current at that point in time (e.g. ‘US Elections’), as well as some of which were consistently discussed topics (e.g. ‘Comedy’).

Experiments were conducted over the time period of the testing data taken to be September 2008, with training data over the month of August 2008. At each day in the test set, we predicted the user actions for future M days based on a time slice duration of t_j days. Note that choosing reasonable values of t_j and M are

extremely important because continual evolution of network and user models over several time slices (without retraining) could soon yield very high prediction error rates of the probabilities of user actions. We therefore chose an empirical threshold of prediction error equal to 20%, and our goal was to determine a value pair t_j , M such that the error was no more than 20%. Obviously we want M to be as high as possible, and error to be still less than 20%. To determine such a pair, for each topic, we considered different durations of time slices $t_j = 0.5, 1, 2, 3$ and 4 days and predicted the user actions for different values of $M = 1-20$ days. The optimal durations of time slices were found to be 2, 3, 2, 3, 1 and 2 days respectively for each of the six topics. And the corresponding optimal values of M were 6, 18, 8, 9, 7 and 5 days respectively.

The results of prediction based on these values are shown in Figure 4.7. Four different variables are shown in each plot in the figure—the actual rate of attraction β , sustainability α , the actual number of users and the error on predicting user actions. The first four topics indicate the presence of synchrony in certain time regions. This can be understood from the following observations—the number of users is high in certain time periods than others and the rate of attraction β and sustainability α are also reasonably large (twice as large as the previous value). For example, synchrony is observed from Sept 6 to Sept 13 for the topic ‘US Elections’, while from Sept 8 to Sept 11 for the topic ‘World News’.

We now observe the performance of our framework in predicting user actions during these periods of social synchrony. We notice that our model performs very well in prediction after the onset of synchrony ($\leq 15\%$ error), compared to the beginning and the end ($\sim 20\%$ error). This can be accounted for by the fact that our model learns the values of β and α based on a short time span in the past. At the onset of the synchrony, the total number of users and the values β and α exhibit rapid increase. Hence their estimates based on the past do not match well with the actual values, affecting our prediction performance.

Similarly towards the end of synchrony our model is trained on the rapidly rising values of β and α and it is not able to account for the decrease in the

two values. In the case of topics without synchrony, e.g. ‘Celebrity’ and ‘Tennis’, our framework is able to predict user actions with sufficiently low error rates.

To summarize, we observe that our model on average is consistently able to predict user actions with low error (15.2+4.3%) during time periods with and without synchrony over all the 51 topics present in the crawled dataset.

Properties of Seed Users

We now investigate if network properties of the seed user set, including size, degree distribution, clustering coefficient and measure of vulnerability (i.e. the weighted sum of the degrees of all individual users in the seed set) are correlated with the phenomenon of synchrony.

Table 4.1 shows the measures of the different properties of the seed set for six topics. For the first four topics where synchrony is present, the mean of the degree distribution is high, variance is low, clustering coefficient is low and the vulnerability measure is high. This is due to the fact that during synchrony, the network size evolves very rapidly and hence the seed set of users does not form a cohesive sub-network - hence the low clustering coefficient. High measure of vulnerability indicates high degrees of users, which further indicates that the action of digging is observable to a large set of users (through feeds over their contacts), triggering several new users to mimic the action, leading to an emergent synchrony. However the topics ‘Celebrity’ and ‘Tennis’ do not exhibit synchrony. In this research hence we have preliminary evidence that the properties of the seed set—high mean and low variance of degree distribution and low clustering coefficient are likely to be correlated with an emergent synchrony at a later point in time.

Comparative Evaluation

We now compare the performance of our model, i.e. the error predicting user actions based on four baseline methods. The techniques consist of two non-social network based methods and two social network based methods. The first is a simple temporal trend learning method of user actions. The second is a linear regressor based

Table 4.1: Properties of the seed user set for six different topics. $|U_0|$ is the size of seed set, μ_0 , σ_0 are the mean and variance of the degree distribution of U_0 , ρ_0 is the clustering coefficient of the seed users’ sub-network and ς_0 is the measure of vulnerability.

TOPICS	$ U_0 $	μ_0	σ_0	ρ_0	ς_0	SYNCHRONY
US ELECTIONS	334	42	0.13	0.21	0.72	Yes
WORLD NEWS	582	37	0.18	0.11	0.76	Yes
OLYMPICS	121	28	0.21	0.23	0.67	Yes
COMEDY	79	27	0.17	0.26	0.60	Yes
CELEBRITY	42	11	0.61	0.53	0.23	No
TENNIS	64	14	0.44	0.54	0.13	No

Table 4.2: Error in predicting user actions based on our method against baseline techniques—trend learning method (B_1), user activity regression based framework (B_2), SIR based epidemiological model (B_3) and a simple threshold based model of global cascades (B_4). Our method performs the best among all baseline techniques.

TOPICS	OUR METHOD	B_1	B_2	B_3	B_4
US ELECTIONS	0.19	0.67	0.52	0.38	0.35
WORLD NEWS	0.11	0.41	0.36	0.29	0.28
OLYMPICS	0.19	0.54	0.49	0.44	0.41
COMEDY	0.13	0.46	0.4	0.31	0.27
CELEBRITY	0.12	0.49	0.36	0.29	0.22
TENNIS	0.15	0.53	0.41	0.32	0.27

method which uses the coefficients associating a user’s activities like commenting and replying with her actions to predict actions in the future. The third baseline is the SIR (susceptible-infected-removed) epidemiological model (Cha et al., 2008) popularly used in determining information cascades in social networks. The last baseline is a threshold based model of global cascades (Watts, 2002) based on the idea that a user participates in an action by changing her state to ‘active’ only when a certain sufficiently large fraction of her contacts have already done so. Note that these baseline methods are used in an online learning setting to train and predict the probabilities of the actions over the future. We believe that this constructs a more rigorous comparative benchmark instead of extrapolating the probabilities over the set of future time slices.

The results of evaluation against the four baseline methods have been shown

in Table 4.2. The errors in prediction are shown for the same six different topics and over the time period of the testing data—September 2008. The evaluation results of our method against baseline techniques reveals that it performs the best among all with a mean improvement in prediction ranging 18-56% over all the 51 topics in the dataset. The trend based learning method (B_1) and the regressor based user activity measure (B_2), both perform worse than the other two social network based baseline methods. This explains that synchrony is a social network phenomenon and simple time series based prediction methods naturally yield poor results. The performances of the SIR epistemological model (B_3) and the global cascades model (B_4) are similar; however the cascades model yields slightly lower error. This indicates that the local network topology, that is, the contacts of a user affect her decision to perform an action. Our method has two key improvements over the baseline techniques—(a) the ability to predict user actions for a set of time slices into the future unlike online learning, and (b) ability to take into account the dynamically changing sizes of the user set via the rate of attraction and sustainability. They help boost our prediction performance.

4.7 Extended Experiments on Social Media Datasets

In order to test the generalizability of our prediction framework on other social media datasets, we conducted a similar set of experiments on the micro-blogging site Twitter. Our notion of a social action in this context is the re-tweet (RT) feature, wherein a user copies and posts the content from another user. The goal is to be able to predict whether there is going to be a information cascade in the network (via the re-tweeting activity) at a future point in time, typically spanning a broad thematic category. The Twitter dataset that is used for this set of experiments is a snowball crawl, with ~ 465 K users and ~ 29.5 M tweets. In order to conceive of the thematic categories, we assigned each re-tweet a thematic label using the natural language processing toolkit OpenCalais. Some of the example thematic categories are “Politics”, “Sports” and “Technology-Internet”.

As in the case of the Digg dataset, in this case again we show the performance

Table 4.3: Error in predicting user actions (re-tweeting activity) based on our method against baseline techniques—trend learning method (B_1), user activity regression based framework (B_2), SIR based epidemiological model (B_3) and a simple threshold based model of global cascades (B_4). Our method performs the best among all baseline techniques, in the case of the social media Twitter.

TOPICS	OUR METHOD	B_1	B_2	B_3	B_4
POLITICS	0.1030	0.1740	0.4292	0.4768	0.4809
SPORTS	0.1228	0.1395	0.1488	0.2589	0.3225
TECHNOLOGY-INTERNET	0.1050	0.1995	0.3252	0.4306	0.4457
ENTERTAINMENT-CULTURE	0.1036	0.1167	0.1806	0.3167	0.4930

of our proposed model in predicting user actions (i.e. re-tweeting) is emergent of synchrony, i.e. information cascades in the network. In Figure 4.8 we show the prediction results for four themes in which synchrony has been observed over a one month period in November 2009. Each plot shows four variables: our predicted error, sustainability α and rate of attraction β (normalized), shown on the left Y -axis, and the number of users shown on the right Y -axis. Note that each theme exhibits observable presence of a cascade; and our method is satisfactorily able to predict the associated re-tweeting activity leading to the cascade. The corresponding errors of prediction lie in the range of 11–19%.

Finally, we show comparative studies of our prediction model against other baseline techniques in Table 4.3. As before, we consider the following four baseline methods: trend learning method (B_1), user activity regression based framework (B_2), SIR based epidemiological model (B_3) and a simple threshold based model of global cascades (B_4). We observe that the performance of our method, in terms of prediction accuracy is much better compared to the others, by a margin of 8–35%. This implies that the DBN based model that we proposed in this chapter is useful and effective in quantifying observed synchrony in social media.

4.8 Open Issues and Discussion

Among the open issues in our approach, note that the notion of synchrony is based on prediction of user actions over a set of time slices in the future. However,

determining the exact size of the time horizon (i.e. the number of time slices into the future we can predict) is challenging. This is because typical periods of synchrony could differ for different topics as well as could depend upon contextual factors which are directly not observable in the social network. Moreover the socially aware and unaware states of a user could be affected by her intrinsic preferences—for example, Alice diggs political news because her friends blog about them in the Blogosphere. These factors are not captured by our representation of user context.

4.9 Conclusions

In this chapter, we developed a computational framework to predict emergent social synchrony in social media. Synchrony was defined to have the following characteristics—a given topic and an agreed upon user action, notion of a seed user set, and the rate of attracting new users as well as sustained participation of earlier users. Our method consisted of two parts—the learning framework and the evolution framework. In the learning framework, we developed a DBN based representation to predict the probability of user actions over a set of time slices into the future. In the evolution framework, we evolved the social network size and the user models over a set of future time slices to predict social synchrony. We conducted extensive experiments on social media datasets Twitter and Digg. The results showed that our model yielded low error (15.2+4.3%) in predicting user actions (or diggs) during time periods with and without synchrony. Comparison with different baseline techniques indicated that our method showed significant improvement in predicting user actions.

For future work, we intend to develop sophisticated theoretical models that can identify the conditions triggering social synchrony. The role of social synchrony in citation networks can be useful to identify time periods of paradigm shifts as well. We are also interested to explore how the action of burying a news article on Digg (apart from digging) affects the emergence of a synchrony in the future.

Figure 4.7: Prediction of user actions based on our framework for six topics: ‘US Elections’, ‘World News’, ‘Olympics’, ‘Comedy’, ‘Celebrity’ and ‘Tennis’. Three variables—our predicted error, α and β (normalized) are shown on the left Y-axis, while the number of users is shown on the right Y-axis. Note, our model predicts the user actions with low mean error rate.

Figure 4.8: Prediction of user actions (re-tweeting activity) based on our framework for four thematic categories on Twitter. Three variables—our predicted error, α and β (normalized) are shown on the left Y-axis, while the number of users is shown on the right Y-axis. Note, our model predicts the user actions with low mean error rate.

Chapter 5

Diffusion of User Actions: Impact of Homophily

“Birds of a feather flock together.”—Plato (428/427 BC–348/347 BC), a Classical Greek philosopher, mathematician.

Over several decades, social scientists have been interested in the idea that *similarity breeds connection*—precisely known as “homophily”. Homophily structures the ego-networks of individuals and impacts their communication behavior. It is therefore likely to effect the mechanisms in which information propagates among them. To this effect, we investigate the interplay between homophily along diverse user attributes and the information diffusion process on social media. Our approach has three steps. First we extract several diffusion characteristics along categories such as user-based (volume, number of seeds), topology-based (reach, spread) and time (rate)—corresponding to the baseline social graph as well as graphs filtered on different user attributes (e.g. location, activity behavior). Second, we propose a Dynamic Bayesian Network based framework to predict diffusion characteristics at a future time slice. Third, the impact of attribute homophily is quantified by the ability of the predicted characteristics in explaining actual diffusion, and external time series, e.g. search and news trends. Our experiments conducted on large Twitter and Digg datasets indicate that the particular attribute that can best explain diffusion depends upon the diffusion metric as well as the topic under consideration. In short, attribute homophily based prediction is able to quantify the actual diffusion characteristics and external trends by a significant $\sim 15 - 25\%$ lower distortion compared to the case when homophily is not considered.

5.1 Introduction

The central goal in this chapter is to investigate the relationship between homophily among users and the social process of information diffusion. By “homophily,” we refer to the idea that users in a social system tend to bond more with ones who are “similar” to them than ones who are dissimilar. The homophily principle has been extensively researched in the social sciences over the past few decades (Burt, 1982; McPherson and Smith-Lovin, 1987; Mcpherson et al., 2001). These studies were predominantly ethnographic and cross-sectional in nature and have revealed that homophily *structures* networks. That is, a person’s ego-centric social network is often homogeneous with regard to diverse social, demographic, behavioral, and intra-personal characteristics (Mcpherson et al., 2001) or revolves around social foci such as co-location or commonly situated activities (Feld, 1981). Consequently, in the context of physical networks, these works provide evidence that the existence of homophily is likely to impact the information individuals receive and propagate, the communication activities they engage in, and the social roles they form.

Homophilous relationships have also been observed on online media such as Facebook, Twitter, Digg and YouTube. These networks facilitate the sharing and propagation of information among members of their networks. In these networks, homophilous associations can have a significant impact on *very large scale* social phenomena, including group evolution and information diffusion. For example, the popular social networking site Facebook allows users to engage in community activities via homophilous relationships involving common organizational affiliations. Whereas on the fast-growing social media Twitter, several topics such as ‘#Elections2008’, ‘#MichaelJackson’, ‘Global Warming’ etc have historically featured extensive postings (also known as “tweets”) due to the common interests of large sets of users in politics, music and environmental issues respectively.

These networks, while diverse in terms of their affordances (i.e. what they allow users to do), share some common features. First, there exists a social action

(e.g. posting a tweet on Twitter) within a shared social space (i.e. the action can be observed by all members of the users’ contact network), that facilitates a social process (e.g. diffusion of information). Second, these networks expose attributes including location, time of activity and gender to other users. Finally, these networks also reveal these users attributes as well as the communication, to third party users (via the API tools); thus allowing us to study the impact of a specific attribute on information diffusion within these networks.

The study of the impact of homophily on information diffusion can be valuable in several contexts. Today, due to the plethora of diverse retail products available online to customers, advertising is moving from the traditional “word-of-mouth” model, to models that exploit interactions among individuals on social networks. To this effect, previously, some studies have provided useful insights that social relationships impact the adoption of innovations and products (Kempe et al., 2003). Moreover there has been theoretical and empirical evidence in prior work (Watts et al., 2002) that indicates that individuals have been able to transmit information through a network (via messages) in a sufficiently small number of steps, due to homophily along recognizable personal identities. Hence a viral marketer attempting to advertise a new product could benefit from considering specific sets of users on a social space who are *homophilous* with respect to their interest in similar products or features. Other contexts in which understanding the role of homophily in information diffusion can be important, include, disaster mitigation during crisis situations, understanding social roles of users and in leveraging distributed social search.

Motivating Study

We motivate our problem domain through a qualitative study on Twitter data. The study reveals how different attributes affect the diffusion process on a particular theme “Politics” (comprising topics such as ‘Obama’, ‘Tehran’ and ‘Afghanistan’) during Oct-Nov 2009. Figure 5.1 presents a “ripple” visualization of the diffusion process over a set of social graphs constructed using the attributes—location, infor-

Figure 5.1: Examples of how attribute homophily impacts diffusion. We visualize diffusion on Twitter on the theme ‘Politics’ using four different attribute social graphs: location, activity behavior, information roles and content creation (ref. section 5.3). As noted, the discovered diffusion phenomenon is significantly different for the four attributes—e.g. the time span (size) of the ripple, the volume of users involved (width of the arcs) as well as the extent of tweeting activity (color intensity of each arc).

mation roles, content creation and activity behavior. In order to explain diffusion, we qualitatively show sample news events (from Google News) for the different ripples. The visualization reveals that the choice of the attribute has a significant impact on the discovery of diffusion properties. For example, location seems to yield diffusion ripples over the longest period of time, while content creation the shortest. This implies that diffusion on “Politics” takes place extensively over the location attribute of users, i.e. there exists location-based homophily corresponding to the diffusion process over “Politics”. We conjecture that “Politics” being highly related to local happenings with respect to sets of users, it is able to quantify the diffusion process the best among all the chosen attributes.

Main Contributions

We propose a three step approach to investigate the role played by homophily in predicting diffusion characteristics on a given topic over time. First, we extract dif-

fusion characteristics along different categories, such as user-based (volume, number of seeds), topology-based (reach, spread) and time (rate), corresponding to social graphs defined on different user attributes (e.g. location, activity behavior). Second, we predict the users likely to get involved in the diffusion process at a future time slice based on a Dynamic Bayesian Network based probabilistic framework. Third, we utilize the predicted set of users to determine diffusion characteristics at the future time slice. We quantitatively define distortion metrics to study how the predicted characteristics corresponding to each attribute (i.e. presence of homophily along a certain attribute) can explain the actual characteristics as well as external time-series variables—search and news trends.

We demonstrate, based on large Twitter and Digg datasets, that the choice of the homophilous attribute can impact the prediction of information diffusion, given a specific metric and a topic. In most cases, attribute homophily is able to explain the actual diffusion and external trends by a margin of $\sim 15 - 25\%$ lower distortion compared to cases when homophily is not considered. Our method also outperforms baseline techniques in predicting diffusion characteristics subject to homophily, by $\sim 13 - 50\%$.

5.2 Related Work

We present prior work relevant to this work from three different perspectives—attribute homophily in the context of social networks, information diffusion subject to shared social actions, and role of homophily in social contagion.

Attribute Homophily. The role of homophily in the formation and sustenance of social ties and networks have been studied extensively by sociologists since several decades (Feld, 1981; Burt, 1982; McPherson et al., 2001; Kossinets and Watts, 2009). In the two seminal works (McPherson and Smith-Lovin, 1987; McPherson et al., 2001), McPherson et al study the inter-connectedness between homogeneous composition of groups and emergent homophily, and how user context such as geographic propinquity, kinship and isomorphic structural locations in a network allow

the formation of homophilous relations. In (Feld, 1981), the author presents theories to explain the origins of social circles, in terms of social foci that are social, psychological or physical entities around which joint activities of users are organized, e.g. workplaces, families and hangouts. Social foci are considered to be central to how ties are built—consequently how homophily emerges. Fiore et al in (Fiore and Donath, 2005) investigate the role of homophily in online dating choices made by users.

Information Diffusion. The analysis of social information diffusion has been of interest to researchers from various domains ranging from social sciences, epidemiology, physics and economics (Watts et al., 2002; Kempe et al., 2003; Gruhl et al., 2004; Bearman et al., 2004; Liben-Nowell and Kleiberg, 2008). There has been prior work on mining and predicting pathways of diffusion in social networks useful for several applications, ranging from recommendation systems, online advertising, user behavior prediction and disease containment (). Bass in (Bass, 1969) proposed a network independent method to determine the the rate of diffusion at a certain time, which was based on the rate at a previous time, a coefficient of adoption and a coefficient of incitation in the market, based on word-of-mouth. In (Kempe et al., 2003), Kempe et al propose solution to the optimization problem of selecting the most influential nodes in a social network which could trigger a large cascade of further adoptions.

In a recent work, Bakshy et al (Bakshy et al., 2009) study how “gestures” make their way through an online community—Second Life. In another work, Sun et al (Sun et al., 2009) study the diffusion patterns on the Facebook “News Feed” and conclude that in online social media, diffusion dynamics are often triggered by the collision of short chains of information trigger. Tang et al in (Tang et al., 2009) propose a topic affinity propagation (TAP) model for modeling social influence in large networks.

Homophily and Social Contagion. A body of prior work also studies the interplay between the homophily principle and social contagion; particularly in the

context of romantic and sexual networks (Bearman et al., 2004). Their primary observation, largely based on cross-sectional studies of populations, is that network structures inferred based on homophily in partner preferences e.g. race, religiosity are able to mimic the observed spread of sexually transmitted diseases. In their review paper (Cialdini and Goldstein, 2004), Cialdini et al discuss principal driving forces behind spread of social influence and therein the role played by conformity and compliance among individuals, e.g. affiliations, self-categorization. Crandall et al in (Crandall et al., 2008) study the interplay between similarity, emergence of social ties and subsequent social influence on the Wikipedia community. Finally, a recent work (Brzozowski et al., 2008) also explores the impact of “similarity networks” on the design of online social content aggregation services and recommender systems.

Although motivated by different research questions, the approaches taken in these studies do not provide any comprehensive computational analysis of the impact of attribute homophily on diffusion of content in large-scale social media datasets. Unlike problems involving disease containment in networks, information shared on social platforms such as Twitter are extremely content rich (i.e. diverse topics) and can often be correlated with external events (Leskovec et al., 2009). Moreover the diversity of users in terms of activity, demographics and roles is likely to induce various homophilous social relationships. Hence certain kinds of homophilous ties are likely to be more conducive to the flow of certain types of information. Additionally, specific online communities would presumably be interested in information content that is available to them along a homophilous social dimension, than that dissipated via traditional RSS feeds or retrieval techniques. To the best of our knowledge, this is the first time such interplay between homophily and diffusion is being investigated in depth in the context of online social media.

5.3 Problem Formulation

In this section we present our problem formulation. First, we present the key concepts involved in this paper, followed by the problem statement.

Preliminaries

We first present the social graph model used in this paper, and then definitions of attribute homophily and diffusion. Finally we introduce a structure called a diffusion series to quantify the diffusion in the social graph over time.

Social Graph Model

We define our social graph model as a directed graph $G(V, E)$ ¹, such that V is the set of users and $e_{ij} \in E$ if and only if user u_i and u_j are “friends” of each other (bi-directional contacts). Let us further suppose that each user $u_i \in V$ can perform a set of “social actions”, $\mathcal{O} = \{O_1, O_2, \dots\}$, e.g. posting a tweet, uploading a photo on Flickr or writing on somebody’s Facebook Wall. Let the users in V also be associated with a set of attributes $\mathcal{A} = \{a_k\}$ (e.g. location or organizational affiliation) that are responsible for homophily. Corresponding to each value v defined over an attribute $a_k \in \mathcal{A}$, we construct a social graph $G(a_k = v)$ such that it consists of the users in G with the particular value of the attribute, while an edge exists between two users in $G(a_k)$ if there is an edge between them in G .² E.g., for location, we can define sets of social graphs over users from Europe, Asia etc.

In this paper, our social graph model is based on the social media Twitter. Twitter features a micro-blogging service that allows users to post short content, known as “tweets”, often comprising URLs usually encoded via bit.ly, tinyurl, etc. The particular “social action” in this context is the posting of a tweet; also popularly called “tweeting”. Users can also “follow” other users; hence if user u_i follows u_j , Twitter allows u_i to subscribe to the tweets of u_j via feeds; u_i is then also called a

¹Henceforth referred to as the baseline social graph G .

²For simplicity, we omit specifying the attribute value v in the rest of the paper, and refer to $G(a_k = v)$ as the “attribute social graph” $G(a_k)$.

Figure 5.2: (a) Distribution of Twitter users in different continents based on the location attribute; (b) Comparative examples of activity behavior distributions of three users over a period of 24 hours.

“follower” of u_j . Two users are denoted as “friends” on Twitter if they “follow” each other. Note that, in the context of Twitter, using the bi-directional “friend” link is more useful compared to the uni-directional “follow” link because the former is more likely to be robust to spam—a normal user is less likely to follow a spam-like account. Further, for the particular dataset of Twitter, we have considered a set of four attributes associated with the users:

Location of users, extracted using the timezone attribute of Twitter users. Specifically, the values of location correspond to the different continents, e.g. Asia, Europe and North America (ref. Figure 5.2(a)).

Information roles of users, we consider three categories of roles: “generators”, “mediators” and “receptors” (Figure 5.3). Generators are users who create several posts (or tweets) but few users respond to them (via the @ tag on Twitter, which is typically used with the username to respond to a particular user, e.g. @BillGates). While receptors are those who create fewer posts but receive several posts as responses. Mediators are users who lie between these two categories.

Content creation of users, we use the two content creation roles: “meformer”

Figure 5.3: The three categories of user behavior corresponding to the attribute: information roles. Out-going links indicate post creation while incoming links indicate post responses from other users.

(users who primarily post content relating to self) and “informer” (users posting content about external happenings) as discussed in (Mor Naaman, 2010).

Activity behavior of users, i.e. the distribution of a particular social action over a certain time period. We consider the mean number of posts (tweets) per user over 24 hours (ref. Figure 5.2(b)) and compute similarities between pairs of users based on the Kullback-Leibler (KL) divergence measure of comparing across distributions.

Attribute Homophily

Attribute homophily (McPherson and Smith-Lovin, 1987; Mcpherson et al., 2001) is defined as the tendency of users in a social graph to associate and bond with others who are “similar” to them along a certain attribute or contextual dimension e.g. age, gender, race, political view or organizational affiliation. Specifically, a pair of users can be said to be “homophilous” if one of their attributes match in a proportion greater than that in the network of which they are a part. Hence in our context, for a particular value of $a_k \in \mathcal{A}$, the users in the social graph $G(a_k)$ corresponding to that value are homophilous to each other.

Topic Diffusion

Diffusion with respect to a particular topic at a certain time is given as the flow of information on the topic from one user to another via the social graph, and based

Figure 5.4: Example of a diffusion series from Twitter on the topic “global warming”. The nodes are users involved in diffusion while the edges represent “friend links” connecting two users.

on a particular social action. Specifically,

Definition 2 *Given two users u_i and u_j in the baseline social graph G such that $e_{ij} \in E$, there is diffusion of information on topic θ from u_j to u_i if u_j performs a particular social action O_r related to θ at a time slice t_{m-1} and is succeeded by u_i in performing the same action on θ at the next time slice t_m , where $t_{m-1} < t_m$.³*

Further, topic diffusion subject to homophily along the attribute a_k is defined as the diffusion over the attribute social graph $G(a_k)$.

In the context of Twitter, topic diffusion can manifest itself through three types of evidences: (1) users posting tweets using the same URL, (2) users tweeting with the same hashtag (e.g. #MichaelJackson) or a set of common keywords, and (3) users using the re-tweet (RT) symbol. We utilize all these three cases of topic diffusion in this work.

³Since we discuss our problem formulation and methodology for a specific social action, the dependence of different concepts on O_r is omitted in the rest of the paper for simplicity.

Diffusion Series

In order to characterize diffusion, we now define a topology called a *diffusion series*⁴ that summarizes diffusion in a social graph for a given topic over a period of time. Formally,

Definition 3 *A diffusion series $s_N(\theta)$ on topic θ and over time slices t_1 to t_N is defined as a directed acyclic graph where the nodes represent a subset of users in the baseline social graph G , who are involved in a specific social action O_r over θ at any time slice between t_1 and t_N .*

Note, in a diffusion series $s_N(\theta)$ a node represents an occurrence of a user u_i creating at least one instance of the social action O_r about θ at a certain time slice t_m such that $t_1 \leq t_m \leq t_N$. Nodes are organized into “slots”; where nodes associated with the same time slice t_m are arranged into the same slot l_m . Hence it is possible that the same user is present at multiple slots in the series if s/he tweets about the same topic θ at different time slices. Additionally, there are edges between nodes across two adjacent slots, indicating that user u_i in slot l_m performs the social action O_r on θ at t_m , after her friend u_j has performed action on the same topic θ at the previous time slice t_{m-1} (i.e. at slot l_{m-1}). There are no edges between nodes at the same slot l_m : a diffusion series $s_N(\theta)$ in this work captures diffusion on topic θ across time slices, and does not include possible flow occurring at the same time slice.

For the Twitter dataset, we have chosen the granularity of the time slice t_m to be sufficiently small, i.e. a day to capture the dynamics of diffusion. Thus all the users at slot l_m tweet about θ on the same day; and two consecutive slots have a time difference of one day. An example of a diffusion series on Twitter over the topic “global warming” has been shown in Figure 12.1, qualitatively annotated by the authors with significant relevant news events (<http://www.news.google.com/>).

⁴Note, a diffusion series is similar to a diffusion tree as in (Liben-Nowell and Kleiberg, 2008; Bakshy et al., 2009), however we call it a “series” since it is constructed progressively over a period of time and allows a node to have multiple sources of diffusion.

Since each topic θ can have multiple disconnected diffusion series $s_N(\theta)$ at any given time slice t_N , we call the family of all diffusion series a *diffusion collection* $\mathcal{S}_N(\theta) = \{s_N(\theta)\}$. Corresponding to each value of the attribute a_k , the diffusion collection over the attribute social graph $G(a_k)$ at t_N is similarly given as $\mathcal{S}_{N;a_k}(\theta) = \{s_{N;a_k}(\theta)\}$.

5.4 Problem Statement

Given, (1) a baseline social graph $G(V, E)$; (2) a set of social actions $\mathcal{O} = \{O_1, O_2, \dots\}$ that can be performed by users in V , and (3) a set of attributes $\mathcal{A} = \{a_k\}$ that are shared by users in V , we perform the following two preliminary steps. First, we construct the attribute social graphs $\{G(a_k)\}$, for all values of $a_k \in \mathcal{A}$. Second, we construct diffusion collections corresponding to G and $\{G(a_k)\}$ for a given topic θ (on which diffusion is to be estimated over time slices t_1 to t_N) and a particular social action O_r : these are given as $\mathcal{S}_N(\theta)$ and $\{\mathcal{S}_{N;a_k}(\theta)\}$ respectively. The technical problem addressed in this paper involves the following:

1. *Characterization*: Based on each of the diffusion collections $\mathcal{S}_N(\theta)$ and $\{\mathcal{S}_{N;a_k}(\theta)\}$, we extract diffusion characteristics on θ at time slice t_N given as: $\mathbf{d}_N(\theta)$ and $\{\mathbf{d}_{N;a_k}(\theta)\}$ respectively (section 5.5);
2. *Prediction*: We predict the set of users likely to perform the same social action at the next time slice t_{N+1} corresponding to each of the diffusion collections $\mathcal{S}_N(\theta)$ and $\{\mathcal{S}_{N;a_k}(\theta)\}$. This gives the diffusion collections at t_{N+1} : $\hat{\mathcal{S}}_{N+1}(\theta)$ and $\{\hat{\mathcal{S}}_{N+1;a_k}(\theta)\} \forall a_k \in \mathcal{A}$ (section 5.6);
3. *Distortion Measurement*: We extract diffusion characteristics at t_{N+1} over the (predicted) diffusion collections, $\hat{\mathcal{S}}_{N+1}(\theta)$ and $\{\hat{\mathcal{S}}_{N+1;a_k}(\theta)\}$, given as, $\hat{\mathbf{d}}_{N+1}(\theta)$ and $\{\hat{\mathbf{d}}_{N+1;a_k}(\theta)\}$ respectively. Now we quantify the impact of attribute homophily on diffusion based on two kinds of distortion measurements on $\hat{\mathbf{d}}_{N+1}(\theta)$ and $\{\hat{\mathbf{d}}_{N+1;a_k}(\theta)\}$. A particular attribute $a_k \in \mathcal{A}$ would have an impact on diffusion if $\hat{\mathbf{d}}_{N+1;a_k}(\theta)$, averaged over all possible values of a_k : (a) has lower distortion with respect to the actual (i.e. $\mathbf{d}_{N+1}(\theta)$); and (b) can

quantify external time series (search, news trends) better, compared to either $\hat{\mathbf{d}}_{N+1}(\theta)$ or $\{\hat{\mathbf{d}}_{N+1;a'_k}(\theta)\}$, where $k' \neq k$ (section 5.9).

5.5 Characterizing Diffusion

We describe eight different measures for quantifying diffusion characteristics given by the baseline and the attribute social graphs on a certain topic and via a particular social action. The measures are categorized through various aspects such as: properties of users involved in diffusion (volume, participation and dissemination), diffusion series topology (reach, spread, cascade instances and collection size) and temporal properties (rate). We discuss the measures for the diffusion collection corresponding to the baseline social graph (i.e. $\mathcal{S}_N(\theta)$); the computation of these measures on the attribute social graphs follow correspondingly over their respective diffusion collections (i.e. $\{\mathcal{S}_{N;a_k}(\theta)\}$).

Volume: Volume is a notion of the overall degree of contagion in the social graph. For the diffusion collection $\mathcal{S}_N(\theta)$ over the baseline social graph G , we formally define volume $v_N(\theta)$ with respect to θ and at time slice t_N as the ratio of $n_N(\theta)$ to $\eta_N(\theta)$, where $n_N(\theta)$ is the total number of users (nodes) in the diffusion collection $\mathcal{S}_N(\theta)$, and $\eta_N(\theta)$ is the number of users in the social graph G associated with θ .

Participation: Participation $p_N(\theta)$ at time slice t_N (Bakshy et al., 2009) is the ratio of the number of non-leaf nodes in the diffusion collection $\mathcal{S}_N(\theta)$, normalized by $\eta_N(\theta)$.

Dissemination: Dissemination $\delta_N(\theta)$ at time slice t_N is given by the ratio of the number of users in the diffusion collection $\mathcal{S}_N(\theta)$ who do not have a parent node, normalized by $\eta_N(\theta)$. In other words, they are the “seed users” or ones who get involved in the diffusion due to some unobservable external influence, e.g. a news event.

Reach: Reach $r_N(\theta)$ at time slice t_N (Liben-Nowell and Kleiberg, 2008) is defined as the ratio of the mean of the number of slots to the sum of the number of slots in all diffusion series belonging to $\mathcal{S}_N(\theta)$.

Spread: For the diffusion collection $\mathcal{S}_N(\theta)$, spread $s_N(\theta)$ at time slice t_N (Liben-Nowell and Kleiberg, 2008) is defined as the ratio of the maximum number of nodes at any slot in $s_N(\theta) \in \mathcal{S}_N(\theta)$ to $n_N(\theta)$.

Cascade Instances: Cascade instances $c_N(\theta)$ at time slice t_N is defined as the ratio of the number of slots in the diffusion series $s_N(\theta) \in \mathcal{S}_N(\theta)$ where the number of *new* users at a slot l_m (i.e. non-occurring at a previous slot) is greater than that at the previous slot l_{m-1} , to $L_N(\theta)$, the number of slots in $s_N(\theta) \in \mathcal{S}_N(\theta)$.

Collection Size: Collection size $\alpha_N(\theta)$ at time slice t_N is the ratio of the number of diffusion series $s_N(\theta)$ in $\mathcal{S}_N(\theta)$ over topic θ , to the total number of connected components in the social graph G .

Rate: We define rate $\gamma_N(\theta)$ at time slice t_N as the “speed” at which information on θ diffuses in the collection $\mathcal{S}_N(\theta)$. It depends on the difference between the median time of posting of tweets at all consecutive slots l_m and l_{m-1} in the diffusion series $s_N(\theta) \in \mathcal{S}_N(\theta)$. Hence it is given as:

$$\gamma_N(\theta) = 1 / (1 + \frac{1}{L_N(\theta)} \sum_{l_{m-1}, l_m \in \mathcal{S}_N(\theta)} (\bar{t}_m(\theta) - \bar{t}_{m-1}(\theta))), \quad (5.1)$$

where $\bar{t}_m(\theta)$ and $\bar{t}_{m-1}(\theta)$ are measured in seconds and $\bar{t}_m(\theta)$ corresponds to the median time of tweet at slot l_m in $s_N(\theta) \in \mathcal{S}_N(\theta)$.

These diffusion measures thus characterize diffusion at time slice t_N over $\mathcal{S}_N(\theta)$ as the vector: $\mathbf{d}_N(\theta) = [v_N(\theta), p_N(\theta), \delta_N(\theta), r_N(\theta), s_N(\theta), c_N(\theta), \alpha_N(\theta), \gamma_N(\theta)]$. Similarly, we compute the diffusion measures vector over $\{\mathcal{S}_{N;a_k}(\theta)\}$, given by: $\{\mathbf{d}_{N;a_k}(\theta)\}$, corresponding to each value of a_k .

Figure 5.5: (a) Structure of the Dynamic Bayesian network used for modeling social action of a user u_i . The diagram shows the relationship between environmental features ($F_{i,N}(\theta)$), hidden states ($S_{i,N}(\theta)$) and the observed action ($O_{i,N}(\theta)$). (b) State transition diagram showing the ‘vulnerable’ ($S_i = 1$) and ‘indifferent’ states ($S_i = 0$) of a user u_i .

5.6 Prediction Framework

In this section we present our method of predicting the users who would be part of the diffusion collections at a future time slice for the baseline and attribute social graphs. Our method comprises the following steps. (1) Given the observed diffusion collections until time slice t_N (i.e. $\mathcal{S}_N(\theta)$ and $\mathcal{S}_{N;a_k}(\theta)$), we first propose a probabilistic framework based on Dynamic Bayesian networks (Murphy, 2002) to predict the users likely to perform the social action O_r at the next time slice t_{N+1} . This would yield us users at slot l_{N+1} in the different diffusion series at t_{N+1} . (2) Next, these predicted users give the diffusion collections at t_{N+1} : $\hat{\mathcal{S}}_{N+1}(\theta)$ and $\{\hat{\mathcal{S}}_{N+1;a_k}(\theta)\}$.

We present a Dynamic Bayesian network (DBN) representation of a particular social action by a user over time, that helps us predict the set of users likely to perform the social action at a future time (Figure 5.5(a)). Specifically, at any time slice t_N , a given topic θ and a given social action, the DBN captures the relationship between three nodes:

Environmental Features. That is, the set of contextual variables that effect a user u_i ’s decision to perform the action on θ at a future time slice t_{N+1} (given by $F_{i,N}(\theta)$). It comprises three different measures: (1) u_i ’s degree of activity on θ in the past, given as the ratio of the number of posts (or tweets) by u_i on θ , to the

total number of posts between t_1 and t_N ; (2) mean degree of activity of u_i 's friends in the past, given as the ratio of the number of posts by u_i 's friends on θ , to the total number of posts by them between t_1 and t_N ; and (3) popularity of topic θ at the previous time slice t_N , given as the ratio of the number of posts by all users on θ , to the total number of posts at t_N .

States. That is, latent states ($S_{i,N}(\theta)$) of the user u_i responsible for her involvement in diffusion at t_{N+1} . Our motivation in conceiving the latent states comes from the observation that, in the context of Twitter, a user can tweet on a topic under two kinds of circumstances: first, when she observes her friend doing so already: making her *vulnerable* to diffusion; and second, when her tweeting is *indifferent* to the activities of her friends. Hence the state node at t_{N+1} that impacts u_i 's action can have two values as the vulnerable and the indifferent state (Figure 5.5(b)).

Observed Action. That is, evidence ($O_{i,N}(\theta)$) of the user u_i performing (or not performing) the action, corresponding values being: $\{1, 0\}$ respectively.

Now we show how to predict the probability of the observed action at t_{N+1} (i.e. $\hat{O}_{i,N+1}(\theta)$) using $\mathbf{F}_{i,N}(\theta)$ and $S_{i,N+1}(\theta)$, based on the DBN model. Our goal is to estimate the following expectation⁵:

$$\hat{O}_{i,N+1} = E(O_{i,N+1} | O_{i,N}, \mathbf{F}_{i,N}). \quad (5.2)$$

This involves computing $P(O_{i,N+1} | O_{i,N}, \mathbf{F}_{i,N})$. This conditional probability can be written as an inference equation using the temporal dependencies given by the DBN and assuming first order Markov property:

$$\begin{aligned} & P(O_{i,N+1} | O_{i,N}, \mathbf{F}_{i,N}) \\ &= \sum_{S_{i,N+1}} [P(O_{i,N+1} | S_{i,N+1}, O_{i,N}, \mathbf{F}_{i,N}) \cdot P(S_{i,N+1} | O_{i,N}, \mathbf{F}_{i,N})] \\ &= \sum_{S_{i,N+1}} P(O_{i,N+1} | S_{i,N+1}) \cdot P(S_{i,N+1} | S_{i,N}, \mathbf{F}_{i,N}). \end{aligned} \quad (5.3)$$

⁵Without loss of generalization, we omit the topic θ in the variables in this subsection for the sake of simplicity.

Our prediction task thus involves two parts: predicting the probability of the hidden states given the environmental features, $P(S_{i,N+1}|S_{i,N}, \mathbf{F}_{i,N})$; and predicting the probability density of the observation nodes given the hidden states, $P(O_{i,N+1}|S_{i,N+1})$, and thereby the expected value of observation nodes $\hat{O}_{i,N+1}$. These two steps are discussed in the following subsections.

5.7 Predicting Hidden States

Using Bayes rule, we apply conditional independence between the hidden states and the environmental features at the same time slice (ref. Figure 5.5(a)). The probability of the hidden states at t_{N+1} given the environmental features at t_N , i.e. $P(S_{i,N+1}|S_{i,N}, \mathbf{F}_{i,N})$ can be written as:

$$P(S_{i,N+1}|S_{i,N}, \mathbf{F}_{i,N}) \propto P(\mathbf{F}_{i,N}|S_{i,N}) \cdot P(S_{i,N+1}|S_{i,N}). \quad (5.4)$$

Now, to estimate the probability density of $P(S_{i,N+1}|S_{i,N}, \mathbf{F}_{i,N})$ using eqn. 5.4 we assume that the hidden states $S_{i,N+1}$ follows a multinomial distribution over the environmental features $\mathbf{F}_{i,N}$ with parameter $\phi_{i,N}$, and a conjugate Dirichlet prior over the previous state $S_{i,N}$ with parameter $\lambda_{i,N+1}$. The optimal parameters of the pdf of $P(S_{i,N+1}|S_{i,N}, \mathbf{F}_{i,N})$ can now be estimated using MAP:

$$\begin{aligned} & \mathcal{L}(P(S_{i,N+1}|S_{i,N}, \mathbf{F}_{i,N})) \\ &= \log(P(\mathbf{F}_{i,N}|S_{i,N})) + \log(P(S_{i,N+1}|S_{i,N})) \\ &= \log \mathbf{multinom}(\text{vec}(\mathbf{F}_{i,N}); \phi_{i,N}) \\ &+ \log \mathbf{Dirichlet}(\text{vec}(S_{i,N+1}); \lambda_{i,N+1}) \\ &= \log \frac{\sum_{jk} \mathbf{F}_{i,N;jk}!}{\prod_{jk} \mathbf{F}_{i,N;jk}!} \prod_{jk} \phi_{i,N;jk}^{\mathbf{F}_{i,N;jk}} + \log \frac{1}{B(\lambda_{i,N+1})} \prod_{jl} S_{i,N+1}^{S_{i,N;jl}} \\ &= \sum_{jk} \mathbf{F}_{i,N;jk} \cdot \log \phi_{i,N;jk} + \sum_{jl} S_{i,N;jl} \cdot \log S_{i,N+1;jl} + \text{const.} \end{aligned} \quad (5.5)$$

where $B(\lambda_{i,N+1})$ is a beta-function with the parameter $\lambda_{i,N+1}$. Maximizing the log likelihood in eqn 5.5 hence yields the optimal parameters for the pdf of $P(S_{i,N+1}|S_{i,N}, \mathbf{F}_{i,N})$.

The details of the convergence of the above estimation have been skipped due to page limit and can be found in (Murphy, 2002).

5.8 Predicting Observed Action

To estimate the probability density of the observation nodes given the hidden states, i.e. $P(O_{i,N+1}|S_{i,N+1})$ we adopt a generative model approach and train two discriminative Hidden Markov Models—one corresponding to the class when u_i performs the action, and the other when she does not. Based on observed actions from t_1 to t_N , we learn the parameters of the HMMs using the Baum-Welch algorithm. We then use the emission probability $P(O_{i,N+1}|S_{i,N+1})$ given by the observation-state transition matrix to determine the most likely sequence at t_{N+1} using the Viterbi algorithm. We finally substitute the emission probability $P(O_{i,N+1}|S_{i,N+1})$ from above and $P(S_{i,N+1}|S_{i,N}, \mathbf{F}_{i,N})$ from eqn. 5.5 into eqn. 5.3 to compute the expectation $E(O_{i,N+1}|O_{i,N}, \mathbf{F}_{i,N})$ and get the estimated observed action of u_i : $\hat{O}_{i,N+1}$ (eqn. 5.2). The details of this estimation can be found in (Rabiner, 1990).

We now use the estimated social actions $\hat{O}_{i,N+1}(\theta)$ of all users at time slice t_{N+1} to get a set of users who are likely to involve in the diffusion process at t_{N+1} for both the baseline and the attribute social graphs. Next we use G and $\{G(a_k)\}$ to associate edges between the predicted user set, and the users in each diffusion series corresponding to the diffusion collections at t_N . This gives the diffusion collection t_{N+1} , i.e. $\hat{S}_{N+1}(\theta)$ and $\{\hat{S}_{N+1;a_k}(\theta)\}$ (ref. section 5.3).

5.9 Distortion Measurement

We now compute the diffusion feature vectors $\hat{\mathbf{d}}_{N+1}(\theta)$ or $\{\hat{\mathbf{d}}_{N+1;a_k}(\theta)\}$ based on the predicted diffusion collections $\hat{S}_{N+1}(\theta)$ and $\{\hat{S}_{N+1;a_k}(\theta)\}$ from section 5.6. To quantify the impact of attribute homophily on diffusion at t_{N+1} corresponding to $a_k \in \mathcal{A}$, we define two kinds of distortion measures—(1) saturation measurement, and (2) utility measurement metrics.

Saturation Measurement. We compare distortion between the predicted and actual diffusion characteristics at t_{N+1} . The saturation measurement metric is thus given as $1 - D(\hat{\mathbf{d}}_{N+1}(\theta), \mathbf{d}_{N+1}(\theta))$ and $1 - D(\hat{\mathbf{d}}_{N+1;a_k}(\theta),$

$\mathbf{d}_{N+1}(\theta)$), averaged over all values of $\forall a_k \in \mathcal{A}$ respectively for the baseline and the attribute social graphs. $\mathbf{d}_{N+1}(\theta)$ gives the actual diffusion characteristics at t_{N+1} and $D(A, B)$ Kolmogorov-Smirnov (KS) statistic, defined as $\max(|A - B|)$.

Utility Measurement. We describe two utility measurement metrics for quantifying the relationship between the predicted diffusion characteristics $\hat{\mathbf{d}}_{N+1}(\theta)$ or $\{\hat{\mathbf{d}}_{N+1;a_k}(\theta)\}$ on topic θ , and the trends of same topic θ obtained from external time series. We collect two kinds of external trends: (1) *search trends*—the search volume of θ over t_1 to t_{N+1} ⁶; (2) *news trends*—the frequency of archived news articles about θ over same period⁷. The utility measurement metrics are defined as follows:

Search trend measurement: We first compute the cumulative distribution function (CDF) of diffusion volume as $E_{N+1}^D(\theta) = \sum_{m \leq (N+1)} |l_m(\hat{\mathcal{S}}_{N+1}(\theta))| / Q_D$, where $|l_m(\hat{\mathcal{S}}_{N+1}(\theta))|$ is the number of nodes at slot l_m in the collection $\hat{\mathcal{S}}_{N+1}(\theta)$. Q_D is the normalized term and is defined as $\sum_m |l_m(\hat{\mathcal{S}}_{N+1}(\theta))|$. Next, we compute the CDF of search volume as $E_{N+1}^S(\theta) = \sum_{m \leq (N+1)} f_m^S(\theta) / Q_S$, where $f_m^S(\theta)$ is the search volume at t_m , and Q_S is the normalization term. The search trend measurement is defined as $1 - D(E_{N+1}^D(\theta), E_{N+1}^S(\theta))$, where $D(A, B)$ is the KS statistic.

News trend measurement: Similarly, we compute the CDF of news volume as $E_{N+1}^{\mathcal{N}}(\theta) = \sum_{m \leq (N+1)} f_m^{\mathcal{N}}(\theta) / Q_{\mathcal{N}}$, where $f_m^{\mathcal{N}}(\theta)$ is the number of archived news articles available from Google News for t_m , and $Q_{\mathcal{N}}$ is the normalization term. The news trend measurement is similarly defined as $1 - D(E_{N+1}^D(\theta), E_{N+1}^{\mathcal{N}}(\theta))$.

Using the same method as above, we compute the search and news trend measurement metrics for the attribute social graphs—given as, $1 - D(E_{N+1;a_k}^D(\theta), E_{N+1}^S(\theta))$ and $1 - D(E_{N+1;a_k}^D(\theta), E_{N+1}^{\mathcal{N}}(\theta))$, averaged over all values of $\forall a_k \in \mathcal{A}$ respectively.

⁶<http://www.google.com/intl/en/trends/about.html>

⁷<http://news.google.com/>

Table 5.1: Summary of statistics of the data used for studying diffusion on Twitter.

ATTRIBUTE	VALUE
#nodes	465,107
#edges	836,541
#nodes with time-zone attribute	385,547
#tweets	25,378,846
Time span of tweets’ post times	Oct’06–Nov’09

5.10 Experimental Results

We present our experimental results in this section. First we discuss data preparation (section 5.11), followed by analysis of saturation and utility measurement of diffusion characteristics at the quantitative level: time-based, theme-based and comparative evaluation against baseline methods (section 5.12). We finally discuss some of the implications of this work in section 12.6.

5.11 Data Preparation

Twitter Dataset. We have undertaken a focused crawl⁸ on Twitter based on a snowballing technique, seeding from a set of authoritative users ($\sim 465K$) who post about a diverse range of topics. This seed set size is 500; and comprises politicians, musicians, environmentalists, techies and so on. These lists were collected from the popular social media blog, Mashable (<http://mashable.com/2008/10/20/25-celebrity-twitter-users/>). Next we expanded the social graph from the seed set based on their “friend” links. We finally executed a dedicated cron job that collected the tweets (and their associated timestamps) for users in the entire social graph every 24 hours. Table 12.5 gives some basic statistics of the crawled data spanning over a three year period⁹.

Experimental Setup. The crawled social graph, comprising the users and their

⁸<http://apiwiki.twitter.com/>

⁹The dataset is publicly available for download at <http://www.public.asu.edu/~mdechoud/datasets.html>

Table 5.2: Example theme and trending topic associations on Twitter dataset.

THEME	TRENDING TOPICS
Politics	Obama, Senate, Afghanistan, Tehran, Healthcare
Entertainment-Culture	Beyonce, Eagles, Michael Jackson, #britney3premiere
Sports	Chargers, Cliff Lee, Dodgers, Formula One, New York Yankees
Technology-Internet	Android 2, Bing, Google Wave, Windows 7, #Firefox5
Social Issues	Swine Flu, Unemployment, #BeatCancer, #Stoptheviolence

tweets is now deployed in the study of diffusion. Since we are interested in studying diffusion at the granularity of a topic, we first define how we conceive of the topics. For our experiments, we focus on a set of 125 randomly chosen “trending topics” that are featured on Twitter over a three month period between Sep to Nov 2009. For the ease of analysis, we organize the different trending topics into generalized themes based on the popular open source natural language processing toolkit called “OpenCalais” (<http://www.opencalais.com/>) e.g. Table 2.

Now our experimental goal is to utilize the crawled social graph to construct the baseline and the attribute social graphs, predict diffusion characteristics, and then study these characteristics over time, subject to homophily on each of their respective diffusion collections. For the purpose, we adopt a “batch” method of incremental training and testing. We begin with a base training set size comprising tweets posted during Sep 2009, and then incrementally train and predict the diffusion characteristics over Oct and Nov 2009.

5.12 Quantitative Analysis

Now we present quantitative analysis of impact of attribute homophily on diffusion. First we present the variations of predicted diffusion characteristics corresponding to different attributes, over diffusion metrics and time. Second, we discuss how prediction performance over different attributes varies across different themes.

Figure 5.6: Performance analysis of saturation and utility measurement of predicted diffusion characteristics over time (the higher, the better). Results are shown across different categories of saturation measurement metrics: user-based (volume, participation, dissemination), topology-based (reach, spread, cascade instances, collection size), time-based (rate), utility measurement metrics: search and news trends.

Finally we present a comparative study of the performance of our proposed method against several baseline techniques.

Temporal Analysis across Diffusion Metrics

Figure 5.6 presents the temporal variations of the performance of predicted diffusion characteristics based on saturation and utility measurements, averaged over all eight themes. We organize the results corresponding to different categories of metrics: user-based (volume, participation, dissemination), topology-based (reach, spread, cascade instances, collection size), time-based (rate), and external time-series variables such as search and news trends.

The observations from the results reveal interesting insights. Overall, firstly,

as is intuitive, we observe that as we increase the training data size (i.e. over time), the saturation and utility measure increase for the case of all attributes. The results in Figure 5.6 indicate that compared to the predictions over the baseline social graphs, several attribute social graphs yield higher saturation and utility measures in explaining the actual diffusion characteristics as well as the external trends respectively. However, note that the attributes corresponding to the best and the worst performance vary across the metrics:

- For the user-based metrics, we observe that the attribute, information roles (IRO) yields the highest saturation over time. This is because diffusion characteristics such as volume and participation are often related to the information generation and consumption behavior of users, and hence the higher performance. The worst performance in this case corresponds to the activity behavior attribute (ACT), revealing that the user-based metrics are less affected by the time of the day corresponding to the tweeting activity.
- In the case of the topology-based metrics, the attribute, content creation (CCR) yields the best performance over the month of prediction. This is because diffusion characteristics such as reach and spread are often affected by how much the information in the tweets are associated with external happenings. Since the content creation attribute characterizes users based on the type of content they share via tweets, it explains the high saturation measures observed in this case. Note, the location attribute (LOC) also yields good performance. This is because the topology metrics are often related to the social connectivity in the social graph; and location is likely to play a significant role in defining the social connectivities in the social graph under consideration.
- The attribute activity behavior (ACT) corresponds to the highest saturation measures in the case of the time-based metric rate. This is because the “speed” of information diffusing in the network often depends upon the temporal pattern of activity of the users, i.e. when are they tweeting. Interestingly, in this case the baseline social graph (BAS) corresponds to a better performance

compared to several other attributes. We conjecture that it reveals that rate is less affected by our chosen set of user attributes.

- Finally, in the case of search and news trends, the best performance again corresponds to the activity behavior attribute (ACT); implying that search behavior and response to news events are often affected highly by the temporal patterns of posting of tweets by users. Moreover, LOC attribute’s performance is relatively poor in this case; for which we conjecture that search or news trends are usually less affected by the location of the users.

To summarize, our primary insight from these experiments is that the particular attribute that yields the best prediction over the diffusion process often depends upon the metric under consideration.

Analysis across Themes

Now we discuss attribute homophily subject to variations across the different themes, and averaged over time (Oct-Nov 2009). Figure 5.7 shows that there is considerable variation in performance (in terms of saturation and utility measures) over the eight themes.

In the case of saturation measurement, we observe that the location attribute (LOC) yields high saturation measures over themes related to events that are often “local” in nature: e.g. (1) ‘Sports’ comprising topics such as ‘NBA’, ‘New York Yankees’, ‘Chargers’, ‘Sehwag’ and so on—each of them being of interest to users respectively from the US, NYC, San Diego and India; and (2) ‘Politics’ (that includes topics like ‘Obama’, ‘Tehran’ and ‘Afghanistan’)—all of which were associated with important, essentially local happenings during the period of our analysis. Whereas for themes that are of global importance, such as ‘Social Issues’, including topics like ‘#BeatCancer’, ‘Swine Flu’, ‘#Stoptheviolence’ and ‘Unemployment’, the results indicate that the attribute, information roles (IRO) yields the best performance—since it is able to capture user interests via their information generation and consumption patterns.

Figure 5.7: Mean saturation and utility measurement of predicted diffusion characteristics shown across different themes. The themes are: A–Business-Finance, B–Politics, C–Entertainment-Culture, D–Sports, E–Technology-Internet, F–Human Interest, G–Social Issues, H–Hospitality-Recreation.

From the results on utility measurement, we observe that for themes associated with current external events (e.g. ‘Business-Finance’, ‘Politics’, ‘Entertainment-Culture’ and ‘Sports’), the attribute, activity behavior (ACT) yields high utility measures. This is because information diffusing in the network on current happenings, are often dependent upon the temporal pattern of activity of the users, i.e. their time of tweeting. For ‘Human-Interest’, ‘Social Issues’ and ‘Hospitality-Recreation’, we observe that the content creation attribute (CCR) yields the best performance in prediction, because it reveals the habitual properties of users in dissipating information on current happenings that they are interested in.

Comparative Study

Now we present a comparative study to evaluate the goodness of our proposed method—i.e. the DBN based social action model that is used to predict the users involved in diffusion at a future time slice. For the purpose of comparison, we utilize five different baseline methods. These methods are independently applied on the

Figure 5.8: Comparison of mean saturation and utility measurement of predicted diffusion characteristics across different methods. Results are shown over the location attribute.

baseline and the attribute social graphs to predict the diffusion characteristics and the corresponding distortion measurements over time.

Our first baseline technique (**GenModel**) uses a Hidden Markov Model based generative framework of action prediction where tweeting (or not tweeting) are the observations, while the latent states correspond to the vulnerable and indifferent states (however unlike our method the environmental features i.e. prior action, topic or friends’ activity are not considered). The next method (**Cascade**) is a threshold based model of ‘global cascades’ (Watts, 2002) based on the idea that a user participates in an action by changing her state to ‘active’ only when a certain sufficiently large fraction of her contacts have already done so. In the third baseline framework (**LinRegress**), we predict the probability of observed action (i.e. tweeting) using a linear regression model that uses the environmental features for training. The fourth baseline framework (**DegAct**) chooses users likely to participate in diffusion at a future time slice based on the degree of tweeting activity. Finally in the fifth baseline technique (**Random**), we select a set of users at random for the next time slice as participants in the diffusion process.

In Figure 5.8, we present the results of the comparative study over all these methods for a particular attribute location (LOC), averaged over all themes as well as time. In the case of both saturation measurement (computed over all diffusion

Table 5.3: Comparison of mean saturation and utility measurement of predicted diffusion (the higher, the better) across all methods and attribute and baseline social graphs.

METRIC	SOCIAL GRAPH	METHOD					
		HomDiff	GenModel	Cascade	LinRgress	DegAct	Random
Saturation Measurement	BAS	0.63	0.42	0.44	0.35	0.25	0.17
	LOC	0.78	0.56	0.65	0.48	0.31	0.27
	IRO	0.68	0.47	0.51	0.39	0.26	0.18
	CCR	0.67	0.45	0.49	0.36	0.28	0.21
	ACT	0.62	0.37	0.41	0.33	0.21	0.15
Utility Measurement	BAS	0.59	0.58	0.48	0.41	0.32	0.21
	LOC	0.51	0.52	0.44	0.31	0.28	0.16
	IRO	0.64	0.61	0.53	0.45	0.37	0.26
	CCR	0.72	0.65	0.57	0.52	0.39	0.29
	ACT	0.84	0.68	0.61	0.55	0.42	0.33

characteristics, such as volume, reach, rate) and utility measurement (computed over search and news trends), we observe that our proposed method (**HomDiff**) yields the highest saturation in explaining the actual diffusion characteristics. Interestingly, note that **Cascade**, being able to utilize the social graph’s structural properties, performs better than **GenModel** in case of saturation measurement, that quantifies the user involvement, topological and temporal aspects of the diffusion process. While for utility measurements, correlation of diffusion with external trends benefits more from **GenModel** as it utilizes the users’ tweeting behavior over time.

Finally we summarize the performance of different methods across different attributes, as well as the baseline social graphs (averaged over all themes) in Table 5.3. We observe that our method (**HomDiff**) gives the highest saturation and utility measures compared to the baselines. Additionally, across the two measurement metrics, the attributes that yield best performance differ: being **LOC** and **IRO** for saturation measurement, while **ACT** and **CCR** for utility.

5.13 Extended Experiments on Social Media Datasets

We further validate our hypothesis of the impact of homophily on diffusion on alternate social media datasets. The dataset used for this set of experiments is crawled from the social media Digg. We seeded our crawling from the stories

in the featured category ‘Popular’ on the Digg website. We crawled all stories in this list and which submitted over August and September 2008. We identified the unique users from these set of stories and constructed their degree distribution (i.e. the number of contacts). From the degree distribution, we picked a set of 500 users with the highest degrees. We crawled all the stories, diggs, comments, replies submitted by them over the two months and collected their contacts. We used a snowballing technique, i.e. iteratively followed this procedure for a set of 21,919 users. In total, this dataset comprises 187,277 stories, 7,622,678 diggs, 687,616 comments and 477,320 replies over a set of 51 topics in this time range.

Note that in the context of Digg, the notion of the social action is the activity of digging a news item; and hence we study the diffusion process as a result of the digging activity of users on a certain topic. As in the case of the Twitter dataset, we consider the following three attributes along which we intend to identify homophily—content creation, information roles and activity behavior. Note that location is not an attribute here because of the unavailability of geotagging or timezone information of Digg users.

In the first set of experiments, we present comparison of our proposed method (**HomDiff**) to the five baseline techniques: **GenModel**, **Cascade**, **LinRegress**, **DegAct** and **Random**. In Figure 5.9 we show the mean measures of saturation and utility measurement respectively for the baseline graph (a–b) as well as the three other attributes along which we characterize homophily in this work (c–h). High values of saturation and utility measurement for our method **HomDiff** shows that it is better than the other baseline techniques in quantifying observed diffusion via homophily in each case.

There are interesting insights into the performance of the different baseline techniques:

1. It appears that across the different attributes, **GenModel** and **DegAct** perform better compared to other baseline techniques, in the case of the saturation measurement metric. We conjecture that in the context of Digg, diffusion processes that drive the saturation measurement metrics often depend on the

nature of the digging activity of the individuals. **GenModel** and **DegAct** being able to characterize this aspect of user activity, quantify the observed diffusion better than other baseline models.

2. In the context of the utility measurement metrics, it appears that the baseline techniques **Cascade** and **LinRegress** yield better performance on the Digg dataset. Our explanation for this observation is that utility measurement metrics being mostly associated with exogenous timely happenings in the real world, are likely to be associated with the environmental factors of the social network. Since **Cascade** and **LinRegress** both model an individual’s participation in diffusion activity based on the environmental features, it is intuitive that they perform better compared to other baseline techniques.

5.14 Discussion

Our extensive experimentation on the Twitter and Digg datasets have revealed that attribute homophily *indeed* impacts the diffusion process; however the particular attribute that can best explain the actual diffusion characteristics often depends upon: (1) the metric used to quantify diffusion, and the (2) topic under consideration. For example, on Twitter the location attribute based homophily among users seems to predict well the topological diffusion characteristics (e.g. reach, spread) well, because users’ local social neighborhoods are often clustered around commonality in their locations. While in the case of topics, we have observed that activity behavior based homophily that can capture the temporal patterns of tweeting behavior of users can predict diffusion characteristics better for themes that are related to current external events, such as ‘Politics’, ‘Technology-Internet’ and ‘Sports’. Note, these results certainly indicate that attribute homophily appears to impact the diffusion process characterization and we do provide empirical evidence of which kinds of themes and diffusion metrics seem to be more sensitive to specific attributes. Nevertheless, we have not provided a principled way of how to *learn* a particular attribute that would minimize the error in the predicted diffusion

discovery. This avenue is left for future work.

We also acknowledge that our proposed method and experimentation is not without limitations. Results have indicated that for certain diffusion characteristics such as rate, our chosen set of attributes do not perform significantly better in predicting the diffusion process, compared to the case when we do not consider attribute homophily. This leaves future opportunities for us to explore presence of homophily along alternate attributes or attribute combinations that incorporate time, e.g. bursty or consistent tweeting behavior.

Additionally, an important point of discussion is that in this paper, to test our hypothesis on the relationship between diffusion and homophily, we have focused on the social media Twitter. This is because Twitter features a diverse user population who vary widely across attributes such as location and content creation behavior (Mor Naaman, 2010); it also provides extensive evidence of information propagation (via tweets). However we acknowledge that our framework can be extended to other datasets with evidence of social actions. Also note that despite the observed interplay between homophily and diffusion, we refrain from making general claims about user A “socially influencing” B as a consequence of homophily: because attitudes can become homophilous even without observable evidence of direct influence.

5.15 Conclusions

We have investigated the role played by attribute homophily on the information diffusion process in online social media. To this effect, we have proposed a dynamic Bayesian network based framework to predict diffusion characteristics corresponding to different user attributes, such as location, activity behavior and information roles. We have also developed two kinds of metrics—saturation and utility measurement metrics that utilize the predicted characteristics to quantify the impact of attribute homophily in explaining the actual diffusion as well as external time series trends. Extensive experimentation on Twitter and Digg datasets have revealed insights in favor of our hypothesis. Overall, attribute homophily is

able to quantify the actual diffusion and external trends by a margin of $\sim 15 - 25\%$ lower distortion compared to cases when homophily is not considered. Comparison with baseline techniques have also indicated that our proposed method outperforms others in predicting diffusion characteristics subject to homophily, by $\sim 13 - 50\%$.

In future we are interested in investigating extensive sets of attributes on diverse social datasets. Additionally, instrumenting the interplay between homophily, emergent ‘sync’ of social actions and social influence in a network is also an exciting future direction.

Figure 5.9: Comparison of mean saturation and utility measurement of predicted diffusion characteristics across different methods on Digg. Results are shown for the baseline graph and three attributes, baseline graph (a–b), content creation (c–d), information roles (e–f) and activity behavior (g–h).

Part II

Evolution of Social Communication Networks

“A self does not amount to much, but no self is an island; each exists in a fabric of relations that is now more complex and mobile than ever before.”—Jean-Francois Lyotard (1924–1998), French philosopher and literary theorist.

A natural extension to information diffusion is to study how the impact of information propagated via a certain mode of communication on the network: including structural and temporal network dynamics. Individuals exhibit a tendency to come together and form ‘groups’ or ‘communities’ as an artifact of their communication process. Thus the manner in which such groups form and evolve is of significant interest to social science research over a long period of time (Coleman, 1998).

Studying the evolving characteristics of groups in particular and social communication networks in general has several applications. The evolution of informal close-knit groups within a large organization, for example, can provide insight into the organization’s global decision-making process. Additionally, the dynamics of certain sub-populations susceptible to a particular disease can be crucial in tracking the early stages of an epidemic. Finally, the discussions within an Internet-based forum can be used to understand the emergence and popularity of new ideas, opinions and technologies in a community.

We focus on blogs in this part of the dissertation. Weblogs, or “blogs,” are personal online diaries managed by easy-to-use software packages that allow single-click publishing of daily entries. The contents are observations and discussions ranging from the mainstream to the personal. There is an increasing trend of different moderated blogs, maintained by a group of editors, who are topical experts, and discuss media related news on different posts. These posts often involve considerable commentary from the public, often leading to opinion formation on several sub-themes.

However, although the structure and dynamics of social groups has been a natural domain of interest, we lack a clear understanding of how these evolving groups are embedded in the context of a larger social network structure. Further-

more, the implications of group communication related to real world external events is not evident. In this part of the dissertation, we conjecture is that some groups are likely to be “representative” of the broader characteristics of the evolving network as a whole. We call such groups to be “prototypical” of the larger network. Finally, we also observe that macroscopic evolution of the networks (in group formation) is closely related to the microscopic evolution at the individual level—it is the individual communication dynamics that influence the nature of ties between them.

Motivating Application. What were the different public opinions in different social networks on the Presidential candidates in the 2008 US elections?

Today journalists often rely on social media to analyze, track and gather public opinions on important political, social and economic events. For example consider that Alice is interested in analyzing the political alignment of individuals interacting on different online platforms. Political blogs are a rich source of emergent public opinion online. There are several reasons why tracking public opinion on social media websites can be useful. First, these online platforms bring together individuals with diverse demographic and socio-economic backgrounds. Second, owing to the low cost of publishing information, these websites enable dialogue among individuals leading to the evolution of opinions via comments, blog posts, tweets, votes and so on. We focus on three main ideas. To begin with, we need to identify characteristics of how groups evolve centered around the formation of collective opinions, and how these evolving groups correlate with real-world political events related to a specific political candidate. Next, we need to develop models to extract the prototypical groups that significantly reflect the emergent opinions.

We address these two central themes in this part of the dissertation:

1. What are the characteristics of individuals in online communities? What is their correlation with external social, economic and political phenomena? (Chapter 6)
2. How do we model an evolving community? What are prototypical tempo-

ral representations of such evolving communities? How do we validate the properties of such prototypical representations? (Chapter 7)

Chapter 6

Community Dynamics and External Phenomena

“Internet marketing strategists = helping businesses build relationships on the Internet = building your brand (marketing your products and services online) appropriately—one tweet at a time.”—Social Media Marketing Conference.

In this chapter, we develop a simple model to study and analyze communication dynamics in communities in the blogosphere and social media and use these dynamics to determine interesting correlations with an external temporal phenomenon: specifically, stock market movement. This work can drive targeted advertising on the web as well as facilitate understanding community evolution in the blogosphere and social websites at large. We describe the communication dynamics by several simple contextual properties of communication, e.g. the number of posts, the number of comments, the length and response time of comments, strength of comments and the different information roles that can be acquired by people (early responders / late trailers, loyals / outliers).

We first study a “technology-savvy” community called Engadget (<http://www.engadget.com>), followed by replication of the same exercise over three other datasets (Digg, Twitter, Huffington Post) to test for generalizability of our observations. There are two key contributions in this chapter: (a) we identify information roles and the contextual properties for four technology companies, and (b) we model them as a regression problem in a Support Vector Machine framework and train the model with stock movements of the companies. It is interestingly observed that the communication activity on the blogosphere and social media has considerable correlations with stock market movement. These correlation measures are further cross-validated against two baseline methods. Our results are promising yielding

about 78% accuracy in predicting the magnitude of movement and 87% for the direction of movement for the technology blog dataset Engadget. However in the case of the other three social media datasets it appears that the nature of the communication activity, the affordances of the website and the agreed upon community ethics (e.g. popularly discussed topics) affect the degree of which the communication can be utilized to predict an external dependent variable (i.e. stock market movement in this case).

6.1 Introduction

In this chapter, we develop a simple model to study and analyze communication dynamics in the blogosphere and social media and use these dynamics to mine interesting correlations with stock market movement. The problem is important because it provides insights into understanding communication patterns of people; for example, how context affects these patterns; how the information roles of people affect communication, temporal and topical dynamics of information roles etc. The communication dynamics further seem to yield correlations with certain external events as well, justifying their predictive power. Often, these dynamics are useful to corporate organizations who are interested in identifying the ‘moods’ of people on external communities in response to product releases and company related events. The dynamics can also drive targeted advertising on the web as well as enable understanding community evolution in the blogosphere.

There has been prior work on modeling communication dynamics (De Choudhury et al., 2007; Kumar et al., 2003; Li et al., 2007) and their correlation with external events (Antweiler and Frank, 2004; Gruhl et al., 2005). The authors in (Antweiler and Frank, 2004) determine correlations between activity in Internet message boards (through frequency counts of relevant messages) and stock volatility and trading volume. In (Gruhl et al., 2005) the authors attempt to determine if blog data exhibit any recognizable pattern prior to spikes in the ranking of the sales of books on Amazon.com. They present hand-crafted predicates to show that correlation measures indicate visible blog mentions ahead of any evidence in sales rank. There has also

been some work on identifying important bloggers based on communication activity in (Nakajima et al., 2005). The authors attempt to determine ‘hot’ conversations in the blogosphere through agitators and summarizers by establishing discriminants. In (Song et al., 2007) the authors identify opinion leaders who are responsible for disseminating important information to the blog network using a variation of the PageRank algorithm. However, modeling information roles and communication dynamics in the prior work have been done in a context-independent manner.

The main contribution of this chapter is a simple contextual framework to model the communication dynamics among people and understand how they can be correlated with events external to the blogosphere and social media. In this work, we have specifically looked at the impact on stocks of technology companies due to postings in a gadget-discussing blog as well as three other social media datasets; however, the framework can be extended easily to other scenarios.

We define stock movement of a company as the normalized difference between the returns on two consecutive days. We assume that the stock movement on a certain weekday can be correlated with the communication dynamics in the past week. This is reasonable because blog communication is often found to precede the occurrence of a real world event (Gruhl et al., 2005). Hence we characterize the communication dynamics in a blog through several contextual features for a particular company. These contextual features are: the number of posts, the number of comments, the length and response time of comments, strength of comments and the different information roles that can be acquired by people (early responders / late trailers, loyalists / outliers). We use these features and stock market movement of the company over N weeks for training an SVM (Support Vector Machine) regressor. The trained parameters are used to predict the movement at the $(N + 1)$ -th week. These results are validated using two baseline methods: firstly by comparing with a non-context aware case and secondly using a linear combination of the contextual features. Our technique supersedes both with error of 22% and 13% in predicting the magnitude and direction of movement respectively in the case of the Engadget dataset. The implication of relatively poorer performance in the context of the rest

three datasets is also discussed henceforth.

6.2 Modeling Information Roles

In this section, we describe a set of information roles that can be assumed by people involved in communication in the blogosphere. Roles of people affect their communication and are therefore a part of the social context. We categorize posters of blog comments into the following information roles: early responders / late trailers corresponding to their response time; and loyals / outliers corresponding to the measure of communication activity (frequency count of posts or comments authored).

Roles due to Response Behavior

In this section we describe the information roles of people with respect to the response times of their comments (on blog posts).

Intuitively, people who are regularly involved in communication in the blogosphere develop certain structures which characterizes their response behavior or when they would write comments on a certain post. Analysis of such patterns can define the information role of the person over a long period of time. To determine roles due to response time, we define a normalized response time frequency distribution for each poster in the following manner.

Normalized Response Time. Response time of a comment is defined as the time (in seconds) elapsed between the publishing of the original blog and the publishing of the comment. We define the normalized response time of a comment such that it depends on (a) the time at which it was published and (b) the rank of the comment, defined as a metric that depends on its relative position among the set of all comments. If a comment is posted very soon after the blog post, its rank is considered high and vice versa. Our motivation for this definition of normalized response time follows from Figure 6.1. We consider the two comments shown in green dots. We notice that the 38th comment (first green dot) has a short response time while

Figure 6.1: Skewness in normalizing response time.

the 43rd comment (second green dot) has a very long response time. However, we also notice that the difference in their ranks is very low. Hence the two comments have been posted to the blog in a comparatively short span of time. The effective normalized response time thus incorporates the rank metric in order to curb the skewness due to response times.

We define the normalized response time as follows. Let t_m be the time at which a comment was posted by a person Alice to a blog post p_x . Let us further assume that t_s and t_e are, respectively, the publishing time of the post and the last comment. Also, let κ be the rank of Alice's comment. The response time $r^c(x)$ of post p_x is defined as,

$$r^c(x) = \theta_1 \cdot \left(1 - \frac{t_m - t_s}{t_e - t_s}\right) + \theta_2 \cdot \left(1 - \frac{\kappa}{n^c(x)}\right), \quad (6.1)$$

where θ_1 and θ_2 are two chosen weights and $n^c(x)$ is the number of comments on post p_x . The equation suggests that the normalized response time of the comment is minimized when Alice has responded early and her comment has low rank κ .

Early Responders / Late Trailers. We now define two categories of behavior: early responders and late trailers. Early Responders are people who respond to messages or blog posts quickly. Late Trailers are people who catch up with an on-going discussion towards the end of communication on the topic. If the mean

Figure 6.2: Activity distribution.

response time r^c over all comments in a period of time is less than a threshold ρ , then the behavior in that time period is taken to be Early Responder. If it is greater than ρ , then the behavior is defined as a Late Trailer.

Roles due to Measure of Activity

In this section we describe two different information roles of people with respect to their overall past communication activity. They are: loyals and outliers. People who are noticed to author large numbers of comments or posts on a certain topic can be considered as loyals to that topic; while, outliers are all the people who are not characterized by any structure in their communication activity. For example, they are the people who appear to comment on blog posts sporadically.

Assume that a person has written a total number of C comments on all posts in a large time period (say 50 weeks) about a certain company. We construct an activity distribution (Figure 6.2) for all such people. In order to determine the information role of a particular person using this distribution, we define a suitable threshold θ over the maximum number of comments in the distribution. If C is greater than θ , then the person's role would be a loyal while it will be an outlier if C is less than or equal to θ . For example in the figure, Charles is a loyal while Brian is an outlier with suitably chosen θ .

6.3 Contextual Modeling of Communication Dynamics

In this section we develop several contextual features which characterize the communication dynamics of people. For the set of all features, we assume that the stock movement y_t of a company on a certain weekday depends on the communication activity in the past week (about the same company), in the time period $(t-6)$ to t . These features are then used to predict stock movement and determine its correlation with communication activity at a future time.

We now introduce some features of our dataset on the Engadget blog. The blog is characterized by two modes of communication: posts and comments written in response to posts. Every comment on a certain post is also marked for its significance by other users. There are five levels: *highest ranked*, *highly ranked*, *neutral*, *low ranked* and *lowest ranked*. It might be noted that these strength levels are different from the ‘rank of a comment’ described in the previous section. The level associated with a comment at any instant of time represents the composite significance indicated by all users. Let us now discuss the features.

Number of Posts. The higher the number of posts about a certain company, the more impact that particular day has on a future event. Hence the first contextual feature is the number of posts per day in the past week, $n_{t-6}^p, n_{t-5}^p, \dots, n_t^p$. Let at day $(t-i)$ where $0 \leq i \leq 6$, $p_1, p_2, \dots, p_{k_i}$ be the k_i posts on a particular company.

Number of Comments. The higher the number of comments a certain company, the more impact that particular day has on a future event. Hence the second contextual feature is the number of comments in all the posts per day in the past week, $n_{t-6}^c, n_{t-5}^c, \dots, n_t^c$.

$$n_{t-i}^c = \sum_{x=1}^{k_i} n_{t-i}^c(x), \quad (6.2)$$

where $n_{t-i}^c(x)$ is the number of comments on post p_x , $1 \leq x \leq k_i$ and $0 \leq i \leq 6$.

Length and Normalized Response Times of Comments. The mean and standard deviation of the length and response times (normalized between 0 and 1) of comments per day might also affect the stock movements. For example, high mean and high standard deviation might reveal that the interest in the posts that day was high, but was peaky or fluctuating. While high mean and low standard deviation would reveal that the interests were high but were mostly flat or consistent among the posters. Let at day $(t-i)$, $l_{t-i}^c(x)$ be the average length of comments on post p_x , where $1 \leq x \leq k_i$ and $0 \leq i \leq 6$. Therefore we define our third contextual feature as the tuple $(\mu_{t-i}^l, \sigma_{t-i}^l)$,

$$\begin{aligned}\mu_{t-i}^l &= \frac{1}{k_i} \sum_{x=1}^{k_i} l_{t-i}^c(x), \\ \sigma_{t-i}^l &= \frac{1}{k_i} \sqrt{\sum_{x=1}^{k_i} (l_{t-i}^c(x) - \bar{l}_{t-i}^c)^2}.\end{aligned}\tag{6.3}$$

Again let at day $(t-i)$, $r_{t-i}^c(x)$ be the average normalized response time of comments on post p_x , where $1 \leq x \leq k_i$ and $0 \leq i \leq 6$. Therefore we define our fourth contextual feature as the tuple $(\mu_{t-i}^l, \sigma_{t-i}^l)$ where the mean and the standard deviation can be computed similar to equation < 3 >.

Strength of Comments. The significance of comments on each day of the week is a useful indicator of impact of the comments on external events. If several people have highly ranked comments on a certain day, those comments are likely to impact the stock movements of the corresponding company more than less ranked comments. Let R_{t-i}^1 be the number of comments that are highest ranked in all the posts $p_1, p_2, \dots, p_{k_i}$ at day $(t-i)$ where $0 \leq i \leq 6$. Similarly, we can define the other four sets as $R_{t-i}^2, R_{t-i}^3, R_{t-i}^4, R_{t-i}^5$ respectively. Then the 5th feature is the 5-tuple $(R_{t-i}^1, R_{t-i}^2, R_{t-i}^3, R_{t-i}^4, R_{t-i}^5)$ corresponding to each day $(t-i)$ for all the seven days in the week.

Size of the Early Responder / Late Trailer Set. This features takes into account the habitual behavior of the people involved in communication on a certain

day $(t-i)$. The sizes of these information roles on a particular day are useful because it contains useful information. If the size of the early responders' set is large, it means that the corresponding posts are likely newer and so the impact on stock movement might be high; while if the size of late trailer set is large, it might reflect an older post whose impact might have already happened or less likely to happen in the future. The 6th feature is therefore the tuple (E, L) where E is the set of all people on day $(t-i)$ who are early responders while L is the set of all late trailers on day $(t-i)$.

Size of the Loyals / Outliers Set. We conjecture that the impact of a particular post on the stock movement of a company in the future also depends on who is posting comments to that post (based on extent of communication activity in the past). We use the set sizes of the information roles: loyals and outliers that we discussed in the previous section.

Let us consider the activity distribution for all the posts $p_1, p_2, \dots, p_{ki}$ at day $(t-i)$ where $0 \leq i \leq 6$. Let S_{t-i}^L and S_{t-i}^O be respectively the set of loyals and outliers at day $(t-i)$.

$$\begin{aligned} S_{t-i}^L &= \{x : C(x) > \theta\}, \\ S_{t-i}^O &= \{x : C(x) < \theta\}, \end{aligned} \tag{6.4}$$

where $C(x)$ is the total number of comments written by poster x at day $(t-i)$ where $0 \leq i \leq 6$ and θ is a suitably chosen threshold.

Let us further assume that S^L and S^O be the set of loyals and outliers over the whole training period. There are several interesting implications of these sets. If the cardinality of the set $S_{t-i}^L \cap S^L$ is large, it means that most of the posters who are otherwise loyal, have responded to posts on day $(t-i)$. It might indicate regularity in communication activity which might further mean that the posts on that day might have low impact on external events. On the other hand, if the cardinality of the set $S_{t-i}^O \cap [S^L \cup S^O]$ is large, it indicates several outliers posting comments. Such large attention focus from external users might imply the pre-occurrence of a big

event. Hence the 8th feature is the tuple $|S_{t-i}^L \cap S^L|, |S_{t-i}^O \cap [S^L \cup S^O]|$ corresponding to each day $(t-i)$.

6.4 Determining Correlation with Stock Market Movement

In this section, we present a Support Vector regression framework to predict the stock movements for a company which would reveal the extent of correlation with the communication dynamics.

First of all we discuss the method of computing stock market movement. In order to determine their correlation with communication dynamics, it is important to take into account the effect of the overall stock market sentiment as well. For example, a negative movement of the stock returns of a particular company may be attributed due to negative movements in the overall stock market index (e.g. NASDAQ, ISE, S&P500 etc).

We define the stock movement of a company c at a day t to be the change in stock return from the closing value of the past day, normalized by the return of the past day. Closing value for a company is the value of stock which exists at the end of the accounting period (one day). The movement is determined as follows,

$$y_t^c = \frac{(\varphi_t - \varphi_{t-1})}{\varphi_{t-1}}, \quad (6.5)$$

where φ_t is the stock return of the company at day t . Similarly, we determine the overall market movement as,

$$y_t^\eta = \frac{(\psi_t - \psi_{t-1})}{\psi_{t-1}}, \quad (6.6)$$

where ψ_t is the stock return of the NASDAQ index at day t (NASDAQ because we are focusing on technology companies). Hence the net stock movement for the company is,

$$y_t = y_t^c - y_t^\eta. \quad (6.7)$$

Now we present an SVM regression framework to predict stock movement. Let us represent the communication data (comprising the contextual feature vectors) as $x_t, t = 1, 2, \dots, N$ where N is the number of weeks over the past for a certain company. Also let us assume, the stock movements data be, $y_t, t = 1, 2, \dots, N$ for the corresponding N weeks for the same company. SVM regression function $f(x)$ is trained on $\{(x_1, y_1), \dots, (x_N, y_N)\}$. It is tested on the incremental sample (x_{N+1}, y_{N+1}) to get the predicted movement $\hat{y}_{N+1}$. The error in prediction is computed as, $E = (y_{N+1} - \hat{y}_{N+1})/y_{N+1}$.

6.5 Experimental Results

In this section we present the experimental results. First we present two baseline frameworks and then describe the nature of the Engadget dataset. This is followed by experimental results.

Baseline Methods

Comment Frequency. The first baseline method for determining correlation of stock market movements uses the frequency of comments per day. We assume again that the stock movement y_t on a certain weekday depends on the number of comments in the past week, in the time period $(t-6) - t$. We further use a linear regressor to learn the correlation coefficients incrementally based on stock movements and number of comments.

Linear Relationship among features. In this method, we assume a linear relationship between the contextual features and use a linear regressor to learn the correlation coefficients incrementally. The regression fit is given as, $\mathbf{Y} = \alpha \cdot \mathbf{X}$, where α is the vector of $\alpha_1, \alpha_2, \dots, \alpha_n$, the correlation coefficients and $\mathbf{Y}$ and $\mathbf{X}$ are the vectors of movements and communication features. Hence for a given training set of n weeks, the regressor predicts the stock movement on a particular weekday at the $(n+1)$ -th week (using the correlation coefficients learnt during training).

Data Preparation

Our dataset comprises the following data items crawled from the Engadget¹ site: a set of blog posts, corresponding comments, length and strength measures as well as who posted them and at what time. These data were collected for four different companies: Apple, Microsoft, Google and Nokia to capture diversity of patterns. There were a total number of 2,469 blog posts, 41,372 comments and 862 users in the dataset in a time period starting January 2007 to November 2007. The stock market returns for the companies (as per the NASDAQ index) were collected from Google Finance².

Results of SVR Regressor

In this section we discuss our dataset and the results of prediction using SVR.

We present the results of the experiments (Figure 6.3) performed using the two baseline methods and the SVM regression technique. We compare them against actual stock movement for the four companies. In Figure 6.3, the predicted and actual movements have been shown across time on a vertical scale with blue bubbles indicating positive movement and red bubbles negative movements. The correlations between the communication activity and stock movement have been shown through several representative events. Each of these events has been collected from the New York Times³ website. For ease of reading and constraints of space, the movements in the figure have been chosen to be representative days in which changes are significant. They span over a span of 50 days and are chosen using suitable thresholds for each company. Hence the same row in Figure 6.3 might not imply the same day across companies. However, the error has been predicted over all 11 months of training data.

The results of the experiments are revealing. We observe that the two baseline methods are not able to adequately capture the subtleties in variation of stock

¹<http://www.engadget.com/>

²<http://finance.google.com/finance>.

³<http://www.nytimes.com/>

Figure 6.3: Visualization of stock movements with time on vertical scale. B_1 and B_2 are the two baseline techniques. Blue bubbles indicate positive movement and red bubbles negative movements. Sizes of the bubbles represent magnitude of movement. The SVR prediction is found to follow the movement trend very closely with an error of 13.41%.

movements. It is only after the occurrence of a ‘big’ event that the baseline methods try to compensate for it by showing a large movement later. However, the SVR is able to capture the fluctuations much better. This is because certain discussions on Engadget often occur regarding future events. The SVR technique, being able to capture a wide array of contextual features and also being able to learn their relationships dynamically, follows the actual stock movement better. It is therefore observed that in majority of the cases the SVR method is closest to the magnitude of actual movement (error: 22%) compared to the two baseline methods (error in 1st baseline method: 48%; in 2nd baseline method: 33%). It is interesting to note that the SVR does a better job in following the movement trend (correspondence in color of bubbles) with an error of 13.41% (error in 1st baseline method: 36%; in 2nd baseline method: 29%). This might be attributed to the fact that our context aware model can predict the direction of movement very well; but the magnitude

of movement is often affected by unprecedented factors, e.g. how the event affects other companies, company statements etc which might not have traces in the past.

It might also be noted here that prediction of stock market movement is an extremely challenging problem which not only depends upon the discussions in a community (e.g. messages exchanged in forums, blogs etc) but are affected by several unforeseen factors which might not be captured in discussions at an earlier point of time. For example, the events “Google outbids Microsoft for Dell bundling deal (May 25)” and “Open Text Corporation extends alliance with Microsoft (Aug 20)” are not found to be present in Engadget discussions in the week long activity used for prediction; this explains the discrepancy and higher error in predicted movement for that day. We also emphasize that this work is attempted to mine interesting correlations of blog communication with stock market activity. The existence of any causal relationship between the two remains an open question and is beyond the scope of this chapter.

6.6 Extended Experiments on Social Media Datasets

So far we have observed that the communication dynamics on the technology blog dataset Engadget is a reasonable predictor of stock market movement in our context. Next we are interested in studying what kind of social network activity are better predictors of stock market activity; for example, are all types of social network communication equally predictive of stock market movement? Or instead, are there differences across the social website we study for the prediction purpose? To answer this question we perform extended experiments on two social media (Digg and Twitter) and one more blog dataset (Huffington Post). An interesting feature of all of these datasets is that although there is considerable conversation among users about different companies and their products, they also feature extensive communication activity on a variety of other topical themes, such as Politics, Social Issues, Technology and Entertainment.

We present the results of prediction of stock market movement using communication activity on Digg, Twitter and Huffington Post in Figure 6.4. The results

Figure 6.4: Prediction of stock market movement using two social media and one blog dataset: Digg, Twitter and Huffington Post.

for all of them are shown over a 30 day period, however there was little temporal overlap across the different prediction time intervals (prediction for Digg: Feb 2008; Twitter: Oct 2008 and Huffington Post: Nov 2007). Further, for the three datasets, there were respectively 247,944, 179,624 and 86,631 unique individuals and 1,358,262, 10,264,289 and 385,284 comments from these individuals involved in prediction.

The bar plot in the Figure shows the predicted measure of stock market movement (normalized between -1 and 1), while the actual normalized value is shown in the line plot. Note that the error in prediction in the three cases are: 43.35%, 34.77% and 28.92% respectively. A key observation of these error rates (and also correspondingly the two trends not following each other over the month-long period) is that the communication activity on these social media datasets is significantly worse compared to the case of Engadget, where we had errors as low as 13.41%.

These results indicate that while predicting stock market movement is a challenging problem, because of the presence of a variety of factors affecting it explicitly or implicitly, there seems to be some value in utilizing social network communication in addressing the problem. However not any nature of social network activity is suitable enough to yield reasonable prediction—communication activity engendering a certain class of topical theme(s) seems to perform better than others. This is experimentally validated by the observation that Engadget communication appears to be a better predictor than the other three social media datasets. We

conjecture that this is an artifact of the observation that Engadget mostly features communication focused on technology companies and/or their products. Hence there is less noise in the conversations—they are mostly geared towards the companies and individuals’ opinions on their policies and products. Hence it is likely that such communication is better correlated to stock market activity than the other datasets, that feature communication on a wide array of topics.

Note that the central take away in this chapter is not that social network communication is the best predictor of stock market activity. Nor have we conducted experimental studies to determine how other prediction models in the finance and business literature would perform compared to our techniques. Rather, our hypothesis in this research was to be able to quantify the utility of social network communication from a prediction perspective. For the purpose we chose stock market movement as a dependent variable: since we noted that a significant degree of communication on various social media exhibit discussion on technology. From our experiments, we have documented this utility in concrete terms—and interestingly, it seems that the nature of the communication activity, the affordances of the website and the agreed upon community ethics (e.g. popularly discussed topics) affect the degree of which the communication can be utilized to predict an external dependent variable. This conclusions in turn, thus sheds light on how we can evaluate social network communication models with the help of their correlation to an external time series, especially in situations when the ground truth of the corresponding underlying social process is not available.

6.7 Conclusions

In this chapter, we have developed a simple model to study and analyze communication dynamics in blogosphere and social media and use those dynamics to determine interesting correlations with stock market movement. We characterized the communication dynamics in a blog or a social media site through several contextual features for a particular (technology) company. These contextual features were: the number of posts, the number of comments, the length of comments, response

time of comments, strength of comments and the different information roles that can be acquired by people (early responders / late trailers, loyals / outliers). We used these features and stock market movement of the company over N weeks for training an SVM regressor. We predicted the stock movement using an incremental sample at $N + 1$. Our technique supersedes two baseline methods with a mean prediction error of 22% for magnitude and 13.41% for predicting the direction of movement over a dataset from Engadget. However for three other social media datasets Digg, Twitter and Huffington Post, we observed poor prediction performance, indicating that the nature of the communication activity, the affordances of the website and the agreed upon community ethics (e.g. popularly discussed topics) affect the degree of which the communication can be utilized to predict an external dependent variable.

There are several interesting directions to future work. We would like to improve our analysis of the information roles to identify people with variable consequences of their communication activity. The contextual model can also be refined by incorporating clustering of tags of companies, characterizing people by identifying response regions of their comments etc. It might also be interesting to see if there is an implicit macro property that underlies communication dynamics on the blogosphere and if that macro property is accounted for by a vocal minority or majority.

Chapter 7

Prototypical Communication Groups

“A World is not an ideology nor a scientific institution, nor is it even a system of ideologies; rather, it is a structure of unconscious relations and symbiotic processes.”—William Irwin Thompson (1938-), social philosopher and cultural critic.

This chapter analyzes communication within a set of individuals to extract the representative prototypical groups and provides a novel framework to establish the utility of such groups. Corporations may want to identify representative groups (which are indicative of the overall communication set) because it is easier to track the prototypical groups rather than the entire set. This can be useful for advertising, identifying “hot” spots of resource consumption as well as in mining representative moods or temperature of a community. Our framework has three parts—extraction, characterization and utility of prototypical groups. First, we extract groups by developing features representing communication dynamics of the individuals. Second, to characterize the overall communication set, we identify a subset of groups within the community as the prototypical groups. Third, we justify the utility of these prototypical groups by using them as predictors of related external phenomena; specifically, stock market movement of technology companies and political polls of Presidential candidates in the 2008 US elections.

We have conducted extensive experiments on two popular blogs, Engadget and Huffington Post as well as four social media datasets, Digg, Twitter, YouTube and Flickr. We observe that the prototypical groups can predict stock market movement / political polls satisfactorily with mean error rate of 20.32%. Further our method outperforms baseline methods based on alternative group extraction and prototypical group identification methods. We evaluate the quality of the extracted groups based on their conductance and coverage measures and develop metrics—

predictivity and resilience to evaluate their ability to predict a related external time series variable (stock market movement / political polls). This implies that communication dynamics of individuals are essential in extracting groups in a community, and the prototypical groups extracted by our method are meaningful in characterizing the overall communication sets.

7.1 Introduction

In this article, we characterize communicating sets of individuals in the blogosphere to extract the representative prototypical groups and provide a novel framework to establish the utility of such groups. A fine-grained, temporal analysis of the prototypical groups can help us understand the evolutionary aspects (organization, dynamics) of communities as an emergent property of interaction.

The advent of Web 2.0 technologies through a plethora of social media websites (e.g. Facebook, MySpace, LiveJournal, Flickr, YouTube) has enabled users of the Web to participate in a wide variety of online communications including expression of their opinions on blogs, sharing images and videos and engaging in discussions centered around media objects and news articles. For example, users on Engadget¹, a popular web magazine covering gadgets and consumer electronics, can communicate through comments and replies on blog posts about different technology products as well as related company happenings. Besides, the users on the Huffington Post², an internet newspaper featuring blog posts on current happenings, can engage in extensive communication on diverse events. Both these blogs feature extensive user participation, implying that the communication and opinions often cater to a large audience—Engadget, as of May 2009(<http://www.ebizmba.com/articles/blogs>), is the fourth most popular blog on the Web having 18,162,567 inbound links and 1,887,887 monthly visitors; while Huffington Post comprised a set of 8.9M visitors as of February 2009(<http://www.nielsen-online.com/>).

A common artifact that emerges out of such communication is that people

¹<http://www.engadget.com/>

²<http://www.huffingtonpost.com/>

often consider the stated opinions of others when online shopping, researching investment companies, planning their vacation or even their next meal out. Personal experiences as well as opinions on external events manifest themselves extensively through online chatter. By analyzing these communication dynamics we can arrive at a better understanding of user behavior, mine useful knowledge applicable to business domains as well as study and reflect on the growth, evolution and characteristics of online communities. Opinion tracking can help drive business intelligence, to, for example, proactively address negative commentary. Companies can benefit from learning which communities are discussing them and the dynamics of such interaction to better understand shareholder sentiment or maximize product exposure. Communication characterization within a network (via groups) can also suggest ranking strategies for impactful individuals, groups and communities in social networks—*what* are the sources of reliable information and *who* are the people capable of suggesting answers to specific questions. It could help user interest prediction, study underlying sociological phenomena in the Blogosphere as well as support tools useful for marketing analysis and operations management that are required to cater to user needs in response to temporally changing external phenomena.

However, despite extensive work in the past on analysis of such structural and temporal dynamics in the Blogosphere, tracking the communication dynamics of a large set of individuals over time can be extremely expensive and inefficient. Moreover, the nature of the overall communication cannot be deciphered on the basis of a single individual or a large set of individuals—we need a unifying notion that helps us characterize *macroscopic* interactions in a communicating set of individuals based on the *microscopic* communication properties of the individuals. Let us consider a few examples to examine these problems. First, consider a news agency interested in providing poll predictions in favor of different political candidates prior to the elections; and therefore wishes to focus on the on-going political discussions on the Web. Second, consider a corporation who intends to advertise a particular product and hence is interested to know which communities might be targeted. Now given the extensive communication dynamics observable on different social media

sites as well as the Blogosphere, the major challenge for the news agency or the corporation is—which are the important or interesting groups that need to be tracked over time?

To answer this question, the insights gained by observing the dynamics of *representative prototypical groups* can be extremely helpful. In this article, we will define prototypical groups in a communication set to be the representative subsets of individuals which characterize the overall communication in terms of the content (e.g. what is the topic of discussion) and context (e.g. how is the membership of the individuals to different groups changing over time). Thus they can aggregate together the salient characteristics of the individuals and suggest a more concrete picture of the nature of communication in a large communication set. In the first example above, the news agency can focus on the opinions of the individuals in the prototypical groups and use them as predictors of polls with respect to different political candidates. In the second example, the corporation can observe the membership dynamics in the prototypical groups and figure out how the sentiments of its customers towards the new product are changing. This would in turn help them further to identify the groups that could be targeted for advertising the product.

Although extracting such representative prototypical communication groups can be extremely interesting, the obvious question arises—how do we validate this? Typically, in several social network problems (e.g. community extraction), determining the utility of the extracted knowledge (in the sense: “is the mined knowledge useful?”) is difficult due to the absence of ground truth. Typically, we analyze the properties of the extracted knowledge (e.g. graph based metrics such as conductance / coverage, when we extract sub-graphs), as a measure of quality. However, these measures do not inform us of the utility of the mined data. In this article, we answer the utility question, by correlating the mined data (the prototypical groups) with a semantically related, but independent time series data (stock market movement / political poll).

Our Approach

Our goal, in this chapter, is to answer the following set of questions—how do we extract the prototypical communication groups in a large set of mutually communicating individuals in the Blogosphere? What are their characteristics? Are such extracted representative groups useful?

To answer these questions, we have characterized online communication sets in social networks to extract their representative prototypical groups and developed a novel utility framework to establish the utility of such prototypical groups. There are two major contributions in this article:

- Extract and characterize the prototypical groups in a community—the prototypical groups can simplify representation of a communication set by capturing its global communication properties.
- Establish the utility of such prototypical groups with their ability to predict related external phenomena—specifically in this work we consider stock market movements of companies and political polls of candidates in 2008 US Presidential elections.

To recall, we define prototypical communication groups to be the representative sets of individuals in a communication set which capture the overall context and content of the entire set. The prototypical groups should have the following two properties—(a) their constituent individuals would be derived from several other groups in the past and (b) their topic distribution would highly aligned to that of the overall communication set. In order to identify such prototypical groups and their utility, we adopt a three step approach—extraction of groups, characterization of prototypical groups and utility of the prototypical groups:

1. First, we develop features representing communication dynamics of the individuals in a communication set—response behavior, measure and impact of an individual’s communication; and her intrinsic roles in communication. These characteristics are then used to extract groups where each group comprises

a set of individuals having similar communication characteristics. The extraction algorithm is based on unsupervised clustering and it emphasizes the temporal relationships among groups.

2. Second, we identify a subset of all the groups as prototypical communication groups which capture and characterize the overall communication set. Two properties are used to identify these groups—composition entropy indicating if their constituent individuals are derived from several other groups in the past and topic divergence, indicating if their topic distribution is highly aligned to that of the overall communication set. We further present a novel optimization framework to identify these groups.
3. Third, we justify that extracting the prototypical groups is useful as they can enable communication sets to predict relevant time series data. Our conjecture is that online communication sets often manifest public opinion about finance, stocks, and technological products as well as about different political candidates. Hence we present a novel utility framework where we show that the prototypical groups can be useful predictors of stock market movement of technology companies as well as of political polls of elections.

To test our model, we have conducted extensive experiments on two popular blogs, Engadget and Huffington Post, as well as host of social media sites, Digg, Twitter, YouTube and Flickr. We observe that during periods of significant external events related to the communities associated with these blogs, the mean composition entropy and the mean topic divergence show significant changes. This implies that the prototypical groups are able to capture the dynamics in the communication set, for all the six datasets. We then justify the utility of the prototypical groups by using them to predict stock market movement / political polls, wherever applicable. The error in prediction are given by 23.42%, 22.58%, 19.67% and 20.73% respectively for communication sets discussing ‘Apple’, ‘Microsoft’, ‘Google’ and ‘Nokia’ and 21.43% and 19.75% for the sets discussing ‘Obama’ and ‘McCain’, for the datasets Engadget and Huffington Post.

Next, we evaluate the characteristics of the prototypical groups by their quality i.e. conductance and coverage, and two novel metrics, high predictivity of stock market movement / political polls and low resilience of the communication set in predicting these external variables when the prototypical groups are not considered. During the evaluation, we observe that our method yields high quality prototypical groups (evident via low conductance and high coverage) and also maximizes predictivity of stock market movement / polls (our method: 0.73, baselines: 0.48) and minimizes resilience with respect prediction (our method: 0.18, baselines: 0.46), implying that the prototypical groups extracted by our method are more meaningful in characterizing the overall communication sets. The mean improvement over the four sets in the Engadget dataset compared to the baseline techniques (based on alternative group extraction and prototypical group identification methods) in conductance is approximately 30%, for coverage is 27%, for the predictivity of the polls is 45% and for resilience is 42%. In case of the Huffington Post dataset, the mean improvement in conductance over the two sets ‘Obama’ and ‘McCain’ is 41.5%, that in coverage is 56%, that in predictivity of the polls is 48% and in resilience is 40%. Similar observations can be made from the measures of conductance and coverage in the case of the four social media datasets, i.e. Digg, Twitter, YouTube and Flickr.

7.2 Related Work

In this section, we introduce relevant prior work with respect to extraction, characterization and evaluation of dynamics in social networks. Properties of social phenomenon, including communication, with respect to individuals, groups and networks have been studied immensely in the past to guide a wide range of problems, ranging from information diffusion (Gruhl et al., 2004; Song et al., 2007; Stewart et al., 2007), trust propagation (Gyöngyi et al., 2004), community detection (Chin and Chignell, 2007; Du et al., 2007; Falkowski et al., 2006; Lin et al., 2007; Zhou et al., 2006), social capital quantification and community-centric search (Almeida and Almeida, 2004; Balfe and Smyth, 2004; Boydell and Smyth, 2006; Coyle and Smyth, 2007), prediction of social collaboration (Fisher and Dourish, 2004; McDon-

ald, 2003; Ohira et al., 2005), mining user behavior (Song et al., 2007), expertise modeling (D’Amore, 2004) and analyzing predictive power of social communication (Gruhl et al., 2005). In this section we provide a detailed overview of the realm of prior work corresponding to (a) social network characterization at different granularities—individuals, groups and networks, (b) analysis of communication properties in social networks and (c) the evaluation of the dynamics of these characterizations and properties.

Social Network Characterization

We present the prior work on characterizing social network properties from three different facets - mining individual characteristics, group characteristics and finally network based characteristics.

Individual Characteristics. There has been considerable work on modeling characteristics of bloggers at an individual level. Newman in (Newman, 2001) studied social networks of scientists in which the actors are authors of scientific papers. He found that the typical distances between authors of a wide range of scientific databases are small-the networks exhibit small world phenomenon (Milgram, 1967) and scale logarithmically with total number of authors in the entire network. He also presented a novel model of weighting the collaborations between the authors. In another work (Newman, 2003), Newman characterized individuals based on their assortative mixing patterns with each other. He considered mixing according to the degree of nodes in a network and investigated on whether or not individuals with similar degree interact with each other. He presents models for several kinds of assortatively mixed networks, and concludes that assortative mixing is a pervasive phenomenon.

Liljeros et al. (Liljeros et al., 2003) characterize individuals in different epidemiological states with respect to propagation of sexually transmitted infections in social networks. They concluded that social network analysis like assortative or disassortative mixing offers useful insight into how to conceptualize and model social

interactions of individuals for these infections and suggested the potential to greatly enhance the understanding of disease epidemics.

Nakajima et al. in (Nakajima et al., 2005) identify important bloggers based on communication activity. They determine 'hot' conversations in the blogosphere through agitators and summarizers by establishing discriminants. Song et al. (Song et al., 2007) identify opinion leaders, who are responsible for disseminating important information to the blog network, using a variation of the PageRank algorithm.

In another work (Huang et al., 2008), Huang et al. investigated coauthorship networks for modeling network evolution and predicting future collaboration between authors. They conduct analysis at three different levels: network, community and individual levels. At the network level, they use several standardized social network analysis metrics to show that coauthorship networks follow the small world phenomenon (Milgram, 1967). They characterized evolution at the community level and at the individual level and developed a stochastic model to predict future collaborations.

Group Characteristics. Prior work has also been done to quantify groups of individuals. Toivonen et al. (Toivonen et al., 2006) presented a model for an undirected growing network useful in studying socio-dynamic phenomena. They analyzed the structural properties of the network using k -clique method for quantifying the groups. Zhou et al. in (Zhou et al., 2006) analyzed the semantic information content in social networks, apart from observable communication to propose two generative Bayesian models for semantic subgroup discovery in social networks. In (Palla et al., 2007), Palla and Barabasi developed a model for analyzing social group evolution. They observed distinct characteristics governing the evolution of large and small groups and concluded that knowledge of membership stability and time commitment could be used to estimate the lifetime of these groups.

Chi et al. in (Chi et al., 2007) proposed a novel technique that captures the structure and temporal dynamics of blog groups. They represent groups by its structure and temporal dynamics—a community graph that indicates how often one

blog communicates with another, and an intensity that indicates the activity level of the group on a temporal scale. They extract such groups from the blogosphere and provide a model to best explain the observed interactions. In (Du et al., 2007), Du et al. presented an efficient algorithm ComTector (Community DeTector) for group detection and their characterization in large-scale social networks based on the nature of overlapping communities in the real world. Falkowski et al. in (Falkowski et al., 2006) analyzed the interaction behavior of members of subgroups over time. They proposed two approaches to analyze subgroup evolution. The first method comprised statistical analyzes and visualizations for an interactive analysis of group evolutions that exhibit stable membership structure. The second method comprised the detection of subgroups in which membership is extremely volatile over time.

Network Characteristics. Characterization of micro-laws of agent networks has also been of interest to the social network analysis researchers in the past. In (Barabasi, Jeong, Neda, Ravasz, Schubert, and Vicsek, Barabasi et al.), Barabasi et al. characterize the dynamic and the structural mechanisms that govern the temporal evolution and topology of scientific collaboration networks. They modeled the topological measures that characterize the network at a given instant and over a period of time. Their results indicated that the collaboration network is scale-free, and evolution is governed by the property of preferential attachment, affecting the degree distributions of people. Further, unlike most other social networks, the mean degrees of nodes were observed to increase over time, and the node separation to decrease. They also proposed a simple model that captures the network’s time evolution.

In (Choudhury and Pentland, 2003), Choudhury et al. develop computational models for automatically learning social network structure within a human group based on wearable sensors. Their primary goal was to detect the influencers in the group and to be able to quantify that influence to better facilitate information diffusion. Borgs et al. in (Borgs et al., 2004) studied the network formed by cross-posts, messages that are posted to two or more newsgroups simultaneously. They

also analyzed the network to show that it followed the small-world phenomenon with significant clustering or tendency of individuals to form cliques. They further use a spectral clustering algorithm to determine newsgroups based on the cross-post data.

Kumar et al. in (Kumar et al., 2006) characterize large scale online social networks based on their evolution of structure. They present a model to segment these networks into three regions: singletons, individuals who do not participate in the network; isolated communities which overwhelmingly display star relationships; and a giant component with a well-connected core region. They further presented a simple model of network growth which emphasizes these aspects of component structure.

Ching et al. in (Chen et al., 2007) present a learning approach that discover agent dynamics driving evolution of social groups in a community. They associate micro-laws to personal attributes like people could be ‘social’ or ‘outgoing’ and groups could be ‘central’ inviting several members.

Analysis of Communication Properties

There has also been prior work on modeling communication properties (Kumar et al., 2003; Li et al., 2007) in social networks and their correlation with external events (Antweiler and Frank, 2004; Gruhl et al., 2005). Antweiler et al. in (Antweiler and Frank, 2004) determine correlations between communication activity in Internet message boards and stock volatility and trading volume. In (Bird et al., 2006), the authors analyze emails of software developers, to understand the relationship between the email activities and the software roles. The authors in (Agrawal et al., 2003) focus on exploiting the link information in computations over hypertext corpora and conjecture that links arising out of social interactions are more valuable than text based analysis. They tested their computational model in a setting where they completely ignored the text and relied only on links.

In another work related to analysis of communication properties in social networks, Mei et al. in (Mei et al., 2007) present a Topic-Sentiment Mixture (TSM)

model that can reveal the latent topical facets in a Weblog collection, the subtopics with respect to a certain given query, and the associated sentiments reference to the query. It is also able to provide general sentiment models that are applicable to any arbitrary topic.

Utility / Evaluation of Social Networks Dynamics

There has been considerable work on analyzing social network characteristics in blogs (Gómez et al., 2008; Kumar et al., 2006) as well as utilizing such communication for prediction of its consequences like user behavior, sales, stock market activity etc (Adar et al., 2007; Gruhl et al., 2005; Liu et al., 2007). In (Gruhl et al., 2005) Gruhl et al. attempt to determine if blog data exhibit any recognizable pattern prior to spikes in the ranking of the sales of books on Amazon.com. Adar et al. in (Adar et al., 2007) present a framework for modeling and predicting user behavior on the web. They created a model for several sets of user behavior and used it to automatically compare the reaction of a user population on one medium e.g. search engines, blogs etc to the reactions on another.

Limitations of Prior Work

There are several limitations of prior work:

1. First, the prior work on group characterization has focused on individual role identification and mining structural network properties. However, our prototypical group extraction emphasizes the observable communication properties among the individuals.
2. Second, we do not have a unified understanding that characterizes interactions in a community (groups / network) at a macro-scale based on microscopic interactions (individuals). Characterization has been focused either on modeling communication or the structural properties of individuals within the social networks; lacking a holistic idea about how people communicate. People tend to have variable communication properties, but prior work lack techniques

to determine long-running moods of sentiments of representative prototypical communication groups of people on a particular topic. Moreover different people tend to have different interpretations and opinions of the same topic; hence there is need for a unified framework which can capture such variability and still describe the nature of the community they belong to. In order to understand the nature of these collective interactions, we therefore propose a means to extract prototypical groups in a community.

3. Third, there has been little work on determining utility of the extracted groups. There is a dearth of convincing validation techniques in social network characterization. Traditionally, research in this area has relied on use of graph based metrics to validate the extracted knowledge; e.g. in community extraction, conductance and coverage measures are used for estimating the quality of communities. We believe that such techniques of validation do not relate with the usefulness of the ‘mined’ knowledge and the extracted properties of the individuals, groups or communities. Prior work, moreover, has not focused on how emergent group dynamics, representative of a set of mutually communicating individuals, can be correlated with related external phenomena.

In this work, we therefore model the communication among individuals in the blogosphere and present a framework which enables us to extract and characterize representative groups of individuals temporally and on the basis of its communication dynamics. We also develop a novel validation framework which shows that the representative groups can be meaningfully correlated with external sources of data—thereby emphasizing the utility of the knowledge mined by our model.

7.3 Problem Definition

We present our problem definition in this section. First, we discuss some preliminaries to understand the terminology used throughout the article. Next, we present a formal definition of the prototypical communication groups. We present our data model next, and finally the problem statement.

Preliminaries

This article deals with three basic terms associated with communication in the Blogosphere.

Individuals. An individual is a person who is involved in communication centered around blog posts. She must either (a) write comments on blog posts or (b) write replies to other comments associated with a certain topic. Note, an individual is associated with different topics at the same time if she is communicating on blog posts related to each topic.

Communication Set. A communication set in this article is a set of individuals who are involved in communication on a topic over shared artifacts e.g. blog posts³. Note that alternative definitions of a communication set based on arbitrary topics / user interests / actions / interactions traditionally discussed in the literature are also possible; have not been discussed in this work in order to reduce noise in our analysis.

Groups. Groups are clusters of individuals in a communication set who share similar communication characteristics, e.g. similar response behavior among themselves.

Notion of a Prototypical Group

Based on these basic terms, we now present a formal understanding of prototypical groups in communication sets in the Blogosphere.

A conceptual representation of prototypical groups in a communication set is shown in Figure 7.1. We define a prototypical group in a communication set to be a subset of individuals whose characteristics are representative of the context and content of the entire communication set. We discuss this in detail as follows.

First, a prototypical group should be able to capture the overall context of the communication properties of the individuals. For the purpose, we conjecture that

³In this work, the names of different technology companies / political candidates have been used to construct the communication sets.

Figure 7.1: Conceptual representation of prototypical groups in a communication set of individuals. Individuals (smaller bubbles) in a communication set form groups (larger bubbles) and prototypical groups (larger bubbles that are shaded) are subsets of groups whose content and context of communication is representative of the entire set.

at a certain point in time, a group would be prototypical if it comprises individuals who had diverse communication characteristics in the past, and therefore, belonged to several different groups.

Second, a prototypical group should capture the content of communication in the overall communication set. In other words, the topics discussed in the prototypical group should be aligned with the topics discussed in the entire communication set at a certain point of time.

Note, we acknowledge that the understanding of a prototypical group in the Blogosphere can be subjective and therefore alternative definitions of such groups are possible. In this article, we have considered the above two characteristics to identify the prototypical groups. However, our framework can be easily generalized to incorporate additional characteristics as well.

Data Model

We now introduce the data model used in this article. We have chosen to investigate the technology blog Engadget and the popular political blog Huffington Post. These sites primarily discuss topics related to consumer electronics and politics, respectively. They are highly influential and are forums of active discussions. Only editors can post news stories, and only authorized readers can comment and reply, thus creating a high quality information source. Each editor can also tag the

Figure 7.2: Data model showing relationship between individuals, comments, blog posts and tags. Each blog post is written by an Editor who associates it with relevant tag(s). Individuals can write comments on such blog posts as well as can reply to existing comments by others.

post with relevant terms. We have focused on Engadget and Huffington Post; however our data model would generalize to any social network with evidence of direct communication.

The data model (Figure 7.2) consists of a set of blog posts, tags on the posts, comments on posts and replies on comments. Each blog post has a timestamp and each comment and reply in turn has a unique individual associated with it and a timestamp of writing.

We segment the posts to assign them to a set of T time slices. We now discuss how we construct the communication sets. At each time slice $t_i (1 \leq i \leq T)$, we filter each post posted within t_i using the tag associated with it to assign it to the corresponding communication set. This creates reliable filters as the tag was suggested by a human (the blog editor). The tag filter used to construct communication sets in this work is the name of a company / political candidate, e.g. Apple, Obama etc. Hence our notion of a communication set is associated with company names and names of Presidential candidates in the 2008 US elections; however our framework could easily be used for characterizing communication on any arbitrary topic.

Each set can now be represented by set of individuals, U_i , the comments and replies written by them along with their corresponding posting times, and comment content. Further, let G_i be the set of groups to be extracted from the communication set at time slice t_i and K_i the set of prototypical groups in it at time slice t_i . Based

on this data model, we now formalize our problem statement.

Problem Statement

We formally present our problem statement as follows. Given a communication set discussing a particular topic, our goal is to, for every time slice t_i , extract a set of groups (G_i) comprising all the individuals in U_i , extract the prototypical groups K_i , determine their characteristics, and finally justify the utility of such prototypical groups. That is, we propose to find answers to the following questions:

1. How do we extract the representative groups K_i in a certain communication set comprising individuals U_i at a certain time slice t_i ?
2. What are the characteristics of these K_i groups?
3. Are such extracted representative K_i groups useful?

We present the main challenges involved in solving the above problem:

1. Since groups are an emergent property of interaction, for (1), we need to develop features for individuals in U_i based on their communication dynamics. Thereafter, groups G_i need to be extracted based on the features of individuals in U_i .
2. Recall, prototypical groups are reflective of the context and the content of communication dynamics in the entire communication set at a particular time slice—they should be able to attract individuals in large numbers who previously belonged to other groups, and their communication should be representative of the communication in the overall communication set. Hence to *characterize* prototypical groups K_i at each time slice t_i (the second challenge (2)), first we need to determine entropy to quantify the composition of all the groups in G_i with respect to the previous time slice t_{i-1} . Second, we need to estimate the divergence of the topic distributions of groups G_i with the overall communication set U_i . We then need to identify prototypical groups K_i by optimizing over the two characteristics.

Figure 7.3: Schematic representation of the key challenges involved in this article. They are (a) extraction of groups in a communication set, (b) characterizing the prototypical groups and finally (c) justifying the utility of the prototypical groups.

3. Finally, for (3) we need to establish the *utility* of these prototypical groups K_i . For this purpose, apart from traditional graph-based goodness metrics of groups, we propose a novel prediction framework which uses the prototypical groups K_i at each time slice t_i to predict stock market movement of companies as well as political polls at the next time slice t_{i+1} .

A summarized schematic view of the three major challenges has been shown in Figure 7.3. In the following three sections, we discuss (a) extraction of groups in a communication set, (b) characterizing the prototypical groups and finally (c) justifying the utility of prototypical groups.

7.4 Group Extraction

In this section, we present our framework of extracting groups in a communication set of individuals. First we present a set of communication based features for individuals, and then use an unsupervised method of group extraction. A schematic representation of group extraction is shown in Figure 7.4.

Communication Features of Individuals

Groups are clusters of individuals in a communication set who share similar communication characteristics. To extract groups, hence we develop several communication based features of individuals. Such features can be attributed to either their communication activity in the past, like response behavior, participation in communication

Figure 7.4: Schematic representation of extraction of groups in a communication set.

and impact of their communication on other individuals, or can be driven by their intrinsic habits of communication, like instigating or following others in discussions. In the following two sub-sections, we would discuss communication activity based features, followed by intrinsic roles in communication.

Communication Activity Features. We describe each individual in a social network with respect to her communication activity. We consider a three dimensional communication activity characterization: responsivity to blog posts, participation and impact of communication.

1. *Responsivity.* We characterize each individual u in U_i at a certain time slice t_i with respect to the time taken to respond to blog posts associated with the communication set. We define a measure called responsivity of an individual as follows. Let $t_{u,j,i}$ be the time at which a comment (or the first comment when an individual posts multiple comments on the same blog post) was posted by individual u to a blog post j within time slice t_i . Let us further assume that $t_{p,j}$ and $t_{l,j}$ are, respectively, the publishing time of the post and the last comment on the post. Further, let $\kappa_{u,j}$ be the rank of u 's comment, which means that u 's comment is the κ -th comment on a blog post, ordered by time of posting. Then the normalized responsivity $r_{u,j,i}$ on post j at t_i is defined as an additive function which is minimized when individual u has responded

early and her comment has low rank,

$$r_{u,j,i} = \omega_1 \cdot \left[1 - \frac{(t_{u,j,i} - t_{p,j})}{(t_{l,j} - t_{p,j})} \right] + \omega_2 \cdot \left[1 - \frac{\kappa_{u,j}}{c_j} \right], \quad (7.1)$$

where ω_1 and ω_2 are two chosen weights and c_j is the total number of comments on post j . The mean responsivity $R_{u,i}$ of individual u at time slice t_i is therefore given as,

$$R_{u,i} = \frac{1}{n_{u,i}} \sum_j r_{u,j,i}, \quad (7.2)$$

where $n_{u,i}$ is the total number of comments made by individual u on posts written in time slice t_i .

2. *Participation.* We determine a measure of participation of an individual u related to a blog post j at a particular time slice t_i . Intuitively it implies how frequently u comments on the blog posts at t_i . The measure $P_{u,i}$ is therefore given by the ratio of the number of comments $n_{u,i}$ by u on these posts, to the total number of posts $m_{u,i}$ that u is involved in at that time slice:

$$P_{u,i} = \frac{n_{u,i}}{m_{u,i}}. \quad (7.3)$$

3. *Impact.* An individual's comment on a blog post associated with a community could impact the communication of others in two ways: (a) the individual's comment is followed by a large number of replies to this comment and / or (b) the mean response time of all comments after her comment is greatly reduced compared to that before her comment. We define the impact of communication for an individual u over a blog post j at time slice t_i as,

$$\nu_{u,j,i} = \omega_3 \cdot \left[\frac{q_{u,j,i}}{n_{j,i}} \right] + \omega_4 \cdot \left[\frac{b_{u,j,i}}{a_{u,j,i}} \right], \quad (7.4)$$

where ω_3 and ω_4 are two chosen weights, $q_{u,j,i}$ is the number of replies to u 's comments, $n_{j,i}$ is the number of comments on blog post j at t_i , $b_{u,j,i}$ is the mean response time of comments *before* u 's comment and $a_{u,j,i}$ is the mean response time *after* u 's comment at t_i . The measure of impact of communication $I_{u,i}$ for u at time slice t_i is therefore given as the average over all posts j where u commented at t_i .

Figure 7.5: Graphical representation of leaders and followers. In (a), the individual u communicates following the four other individuals, hence a follower. In (b), the individual u 's communication leads that of the four others, hence a leader.

Intrinsic Roles in Communication. The communication patterns of individuals are affected by their habitual properties as well, apart from their communication activity. We consider two categories of individuals with respect to their inherent communication characteristics—leaders, individuals who precede others in communication, and followers, individuals who tend to trail others in their communication response.

In Figure 7.5, we show a graphical representation of leaders and followers. In Figure 7.5(a), the individual u communicates following the four other individuals, denoted by inward edges; while in Figure 7.5(b), the individual u 's communication leads that of the four others, as shown by the outward edges. Hence we determine two scores per individual u in U_i —an ‘authority’ measure $F_{u,i}$ saying the measure of ‘followership’ for the individual u at time slice t_i , and a ‘hub’ measure $L_{u,i}$ describing the measure of ‘leadership’ for her at t_i . We use the HITS algorithm (Kleinberg, 1999) for the purpose. We maintain the invariant that the weights of each type are normalized so their squares sum (over all the people) to 1. Since this is a reinforcement relationship, we iterate over x times to modify the leader and follower measures. Convergence is obtained for $x = 5$.

Clustering Individuals

Now we present how we can extract groups based on the communication properties of individuals. Recall, each group consists of a set of individuals having similar individual characteristics.

We first define a similarity measure in order to cluster individuals into groups. The similarity between a pair of individuals is defined using a diffusion kernel (Kondor and Lafferty, 2002). We thereafter deploy an unsupervised clustering based group extraction algorithm (Lin et al., 2007) known as ‘mutual awareness expansion’ (MAE) to extract groups over each of the total T time slices.

We represent each individual u in U_i at time slice t_i as feature vector $\mathbf{f}_{u,i} = \{R_{u,i}, P_{u,i}, I_{u,i}, F_{u,i}, L_{u,i}\}$ where the features are the individual characteristics discussed in the previous section. We define an adjacency matrix Λ_i where each element (u, v) gives the distance between a pair of individuals u and v based on L^2 norm of their corresponding feature vectors at t_i :

$$\Lambda_i(u, v) = \|\mathbf{f}_{u,i} - \mathbf{f}_{v,i}\|^2. \quad (7.5)$$

Now we define the following generator for diffusion kernel with Λ_i as the adjacency matrix,

$$\mathbf{H}_i(u, v) = \begin{cases} \deg_{u,i} & \text{for } u = v, \\ -\Lambda_i(u, v) & \text{for } u \neq v, \\ 0 & \text{otherwise} \end{cases} \quad (7.6)$$

where $\deg_{u,i}$ is the degree of individual u in Λ_i . The diffusion kernel similarity matrix $\mathbf{S}_i$ (Kondor and Lafferty, 2002) is therefore given as,

$$\mathbf{S}_i = \exp(-\beta \cdot \mathbf{L}_i), \quad (7.7)$$

where $\mathbf{L}_i$ is Laplacian of $\mathbf{H}_i$ and β is the diffusion parameter.

Now we use $\mathbf{S}_i$ at each time slice t_i to extract G_i groups with the help of the MAE algorithm in (Lin et al., 2007). The number of clusters or the number of groups for each communication set of individuals was determined based on the most dominant eigen-vectors of the similarity matrix $\mathbf{S}_i$. The MAE algorithm is based on the idea that groups emerge due to observable actions (direct communication, in this chapter) among the member individuals. MAE uses a measure called ‘symmetric

social distance’ which is estimated using a random walk process, to capture mutual awareness expanding among the individuals involved in communication. It then extracts groups by maximizing the distance between two sets of individuals.

The algorithm uses an interaction space based representation to quantify group dynamics, where each dimension represents a pair-wise interaction between two individuals, and each group is a vector in the interaction space. In this work, the different dimensions of the interaction are the different individual characteristics. Two groups are thereafter considered ‘close’ if they are close in the interaction space. MAE further computes the interaction correlation of two groups by using histogram intersection of their vectors (number of overlapping individuals). Then, given a group at a certain time slice, group evolution is determined by maximizing the interaction correlation over communities in the previous time slice.

This form of group extraction is useful in our case as it incorporates metrics to evaluate the temporal evolution of groups. From this set of extracted groups G_i , we now present our model of identifying the prototypical groups $K_i \subseteq G_i$ in a given communication set of individuals.

7.5 Characterizing Prototypical Groups

To recall, we define prototypical groups in a community at a particular time slice need to capture the context and the content of communication in a communication set of individuals. Hence groups at a certain time interval need to be characterized by: (a) a measure of the context - subscription of the composing individuals with respect to their subscription to groups in the previous time interval, and (b) a measure of divergence of the content i.e. topics in a group with respect to that of the community. Finally we discuss an optimization framework of how to identify the prototypical communication groups based on these two characteristics. A schematic representation of group extraction is shown in Figure 7.6.

Figure 7.6: Schematic representation of characterizing key groups from groups in a communication set. Key groups are a subset of the groups at a certain time slice.

Figure 7.7: Illustration of composition entropy of a group g at t_i with respect to other groups in the previous time slice t_{i-1} . The thickness of an edge connecting group g at t_i to the groups in the previous time slice t_{i-1} represents the number of individuals who moved to g from a particular group in the past. In (a) group g at time t_i comprises individuals who belonged to several different groups at t_{i-1} . In (b), individuals in group g come from only two groups in the past. Hence the composition entropy in (a) is higher than that in (b).

Composition Entropy

We define a measure called composition entropy that gives a measure of how the subscription of different individuals in a group at a particular time slice is associated with their subscription to groups in the previous time slice.

A conceptual illustration of composition entropy in a group for two different cases has been shown in Figure 7.7. In the figure, the thickness of an edge connecting group g at t_i to the groups in the previous time slice t_{i-1} represents the number of individuals who moved to g from a particular group in the past. In (a) group g at time t_i comprises individuals who belonged to several different groups at t_{i-1} . In (b), individuals in group g come from only two groups in the past. Hence the composition entropy in (a) is higher than that in (b). That is, when the composition entropy of individuals in a particular group at time slice t_i is very high, with respect

to all groups at t_{i-1} , the group can be considered to be a prototypical group because several individuals having diverse characteristics in the past and belonging to several different groups have moved to this particular group in the current time slice. The composition entropy of a group g in G_i is given by,

$$e(g) = \frac{1}{|G_i|} \sum_{h \in G_{i-1}} (-p(g|h) \log p(g|h)) \quad (7.8)$$

where $p(g|h)$ gives the fraction of the number of individuals who are in g at t_i given they had been in group h at time slice t_{i-1} .

Topic Divergence

In order to characterize prototypical groups, next we need to characterize the topic distributions of groups and the topic distribution of the overall communication set at each time slice. If the topic distribution of a group closely follows the topic distribution of the overall communication set, it means that the group is prototypical to the communication set with respect to the content of communication. Hence we describe topic modeling for groups in the next subsection, followed by our method of estimating the parameters of the topic models. Thereafter we would discuss how we determine the topic divergence given a certain group.

Mixture Model of Topics. The communication activities of individuals in a group are highly dynamic over time. In order to identify the prototypical groups, we need to identify the topics discussed by individuals in a group at each time slice.

We represent the textual data (comments and replies of individuals) associated with a group g at each time slice t_i as a bag-of-words λ_g (stemmed and stop-word eliminated). Our goal is to associate this bag-of-words λ_g with a topic distribution at the corresponding time slice t_i . Hence let us assume that the words in λ_g are generated from N multinomial topic models $\theta_1, \theta_2, \dots, \theta_N$ whose distributions are hidden to us. Our goal is to determine the log likelihood of the topic distributions that can represent our data best. Thereafter we can maximize the log likelihood to compute the parameters of the N topic models.

However we note that the group g is associated with each time slice t_i and therefore it is necessary to regularize our topic mixture model with the time slice indicator. This is based on our conjecture that a word in λ_g can be attributed either to the textual context in the group g , or the context of the time slice t_i : for example, certain words can be highly popular on certain time slices. Thus we represent λ_g at time slice t_i using the following probabilistic mixture model of multinomials (Mei et al., 2006):

$$p(w : \lambda_g, t_i) = \sum_{j=1}^N p(w, \theta_j | \lambda_g, t_i), \quad (7.9)$$

where w is a word in λ_g and θ_j is the j -th topic. The joint probability on the right hand side can be decomposed as:

$$p(w, \theta_j | \lambda_g, t_i) = p(w | \theta_j) \cdot p(\theta_j | \lambda_g, t_i) = p(w | \theta_j) \cdot ((1 - \gamma_i) \cdot p(\theta_j | \lambda_g) + \gamma_i \cdot p(\theta_j | t_i)), \quad (7.10)$$

where γ_i is a parameter that regulates the probability of a topic θ_j given λ_g and the probability of a topic θ_j given the time slice t_i . For the entire collection of comments corresponding to G_i groups at t_i , the log likelihood is therefore given by,

$$L(G_i) = \log p(G_i) = \sum_{\lambda_g \in G_i} \sum_{w \in \lambda_g} \log \sum_{j=1}^N p(w, \theta_j | \lambda_g, t_i), \quad (7.11)$$

where $n(w, \lambda_g)$ is the count of the word w in λ_g and $p(w, \theta_j | \lambda_g, t_i)$ is given by eqn. 7.10.

Parameter Estimation. Now we discuss how we estimate the N parameters in the topic models. We use the standard Expectation-Maximization (EM) algorithm (Duda et al., 2000) to estimate all these parameters by maximizing the data likelihood. For the **E-step**, we define two hidden variables similar to (Mei et al., 2006):

$$\begin{aligned}
z(w, \lambda_g, j) &= \frac{p^{(m)}(w|\theta_j) \cdot ((1 - \gamma_i) \cdot p^{(m)}(\theta_j|\lambda_g) + \gamma_i \cdot p^{(m)}(\theta_j|t_i))}{\sum_{j'=1}^N p^{(m)}(w|\theta_{j'}) \cdot ((1 - \gamma_i) \cdot p^{(m)}(\theta_{j'}|\lambda_g) + \gamma_i \cdot p^{(m)}(\theta_{j'}|t_i))}, \\
y(w, \lambda_g, j) &= \frac{\gamma_i \cdot p^{(m)}(\theta_j|t_i)}{(1 - \gamma_i) \cdot p^{(m)}(\theta_{j'}|\lambda_g) + \gamma_i \cdot p^{(m)}(\theta_{j'}|t_i)}.
\end{aligned} \tag{7.12}$$

The **M-step** has a closed form solution:

$$\begin{aligned}
p^{(m+1)}(\theta_j|\lambda_g) &= \frac{\sum_w n(w, \lambda_g) \cdot z(w, \lambda_g, j) \cdot (1 - y(w, \lambda_g, j))}{\sum_{j'=1}^N \sum_w n(w, \lambda_g) \cdot z(w, \lambda_g, j') \cdot (1 - y(w, \lambda_g, j'))}, \\
p^{(m+1)}(\theta_j|t_i) &= \frac{\sum_{\lambda_g \in t_i} \sum_w n(w, \lambda_g) \cdot z(w, \lambda_g, j) \cdot (1 - y(w, \lambda_g, j))}{\sum_{\lambda_g \in t_i} \sum_{j'=1}^N \sum_w n(w, \lambda_g) \cdot z(w, \lambda_g, j') \cdot (1 - y(w, \lambda_g, j'))}, \\
p^{(m+1)}(w|\theta_j) &= \frac{\sum_{\lambda_g} n(w, \lambda_g) \cdot z(w, \lambda_g, j)}{\sum_{w'} \sum_{\lambda_g} n(w', \lambda_g) \cdot z(w', \lambda_g, j)}.
\end{aligned} \tag{7.13}$$

The EM algorithm terminates when it reaches a local maximum of the log likelihood in eqn. 7.11. With the learnt parameters of the topic models, we can now compute the probability:

$$\begin{aligned}
p(\theta_j|\lambda_g, t_i) &= \text{sum}_w p(\theta_j|w) \cdot ((1 - \gamma_i) \cdot p(w|\lambda_g) + \gamma_i \cdot p(w|t_i)) \\
&= \sum_w (p(w|\theta_j) \cdot p(\theta_j)/p(w)) \cdot ((1 - \gamma_i) \cdot p(w|\lambda_g) + \gamma_i \cdot p(w|t_i)).
\end{aligned} \tag{7.14}$$

All the parameters on the right hand side are known from the above parameter estimation. λ_g being the representation of a group g , the above equation would give us the topic distribution $\Theta_{g,i}$ of group g at every time slice $t_i, 1 \leq i \leq T$. Note, the topic distribution Θ_i of the overall communication set at time slice t_i can be similarly determined from the comments and replies of all its individuals at each time slice using the same method.

Estimating Topic Divergence. The divergence of topic distribution of group g from that of the overall communication set is given by the KL-divergence of $\Theta_{g,i}$ from Θ_i ,

$$d(g) = D(\Theta_i \parallel \Theta_{g,i}). \tag{7.15}$$

Identifying Prototypical Groups

Now we discuss how we can identify the prototypical communication groups based on their composition entropy and the topic distributions discussed in the previous two sub-sections. To recall, the prototypical groups are those whose composition entropy is high and the divergence from the topic distribution of the overall communication set is low. Composition entropy of group $g \in G_i$ at time slice t_i is given by eqn. 7.8, while the topic divergence of group g with respect to the overall communication set is given by eqn. 7.15.

To extract prototypical groups in a communication set of individuals, for each g at each time slice t_i , we define the following objective function:

$$\phi(\mathbf{x}; g) = x_1 \cdot e(g) + x_2 \cdot \exp(-(d(g))), \quad (7.16)$$

where $\mathbf{x} = x_1, x_2$ is the vector of weights which determine the impact of each of the two characteristics, composition entropy and topic divergence in determining the prototypical groups. Maximizing the linear optimization function $\phi(\mathbf{x}; g)$ (or minimizing $-\phi(\mathbf{x}; g)$) for optimal $\mathbf{x}^*$ gives an optimal empirical measure $\phi^*(\mathbf{x}^*; g)$ which determines whether the group g is a prototypical group at a certain time slice t_i . In this work, we consider if $\phi^*(\mathbf{x}^*; g) \geq \delta$, then the group g is a prototypical group, i.e. $g \in K_i$. We set δ empirically to be 0.8 in our work.

Let us now briefly summarize the main idea of our proposed framework. Our framework has two parts: extracting groups in a communication set, and thereafter characterizing prototypical communication groups in it. We develop features representing communication dynamics of the individuals in a communication set. Such characteristics are then used to extract groups where each group comprises a set of individuals having similar communication characteristics. The extraction algorithm is based on unsupervised clustering and it emphasizes the temporal relationships among groups. Next, we identify a subset of all the groups as prototypical groups which capture and characterize the overall community through high composition entropy and low topic divergence. We finally present an optimization framework to

identify such prototypical groups.

Now we discuss how we can establish the utility of these prototypical groups that characterize a communication set.

7.6 Utility of Prototypical Groups

In this section we discuss the utility of the extracted prototypical groups. First we define two utility variables—stock market movements of companies and political polls corresponding to political candidates. Thereafter we discuss a utility framework which can predict these variables based on the extracted prototypical groups.

Traditional validation techniques of social network characterization have primarily been based on network topology metrics like entropy and graph based properties like conductance, coverage etc, typically used to validate group extraction. However the major drawback of such approaches is that these network based measures do not consider the meaningfulness of the characterized groups. Hence we are motivated to use a validation framework which can verify the usefulness of the extracted knowledge (or the characteristics of the communities). In this work, we use the mined data (the prototypical communication groups) to predict a semantically related, but independent time series data (stock market movement, political polls).

Utility Variables

There are two utility variables of interest used in this article—stock market movements of companies and political polls corresponding to political candidates. These variables are relevant to our framework because there is extensive communication in different blogs relating to technology and politics. Moreover our definition of a communication set corresponds to a tag, which could be the name of a company or a political candidate; hence it is meaningful to use stock market data or political polls to justify the utility of the extracted prototypical groups.

Stock Market Movement. We define the stock movement of the company at a

particular week to be the difference in stock prices between the closing values of the current Friday and the previous Friday, normalized by the closing prices on the previous Friday. We further subtract the overall NASDAQ index from this return so as to capture the effect of the entire market movement.

Note, we consider Fridays as they reflect stock market activity over the entire week. In this work, the stock market returns of four different technology companies (as per the NASDAQ index) were collected from Google Finance⁴.

Political Polls. We define political polls for a certain candidate to be the electoral count of that candidate in 2008 US Presidential elections, based on a National Average. In particular, the polls used in this article represent the share of support with respect to the two Presidential nominees of 2008 US Elections. They were collected from the political news site Real Clear Politics (RCP)⁵. RCP reports the poll averages across several media sources like CNN, Newsweek, Pew Research, AP, Reuters / Zogby on a per day basis.

Utility Framework

In order to justify that the extracted prototypical groups are useful, we determine if these groups in a communication set can predict the stock movement of the corresponding company or polls related to a political candidate. The utility framework is presented as follows:

1. We train a Support Vector Regressor at each time slice t_i , over the initial k time slices and for each individual u in a prototypical group $g \in K_i$.
2. We use u 's communication features $\mathbf{f}_{u,i} = \{R_{u,i}, P_{u,i}, I_{u,i}, F_{u,i}, L_{u,i}\}$ and the actual stock market movements of the corresponding company or political poll (V_i) to train the regressor.

⁴<http://finance.google.com/finance>

⁵<http://www.realclearpolitics.com/>

3. Thereafter we use the learnt regression coefficients to predict the stock movement / poll for each individual in a prototypical group at each of the remaining time slices t_{k+1} to t_T .
4. The predicted stock movement / poll measure for all individuals in a prototypical group g at t_i ($\hat{V}_{g,i}$) is therefore the mean of the predicted values across all individuals belonging to that group.
5. For all K_i prototypical groups in the community at time slice q , the final predicted stock movement / poll measure is given by the maximum value across all prototypical groups.

In the following section, we would discuss how we can use the utility framework to evaluate our method of extraction and characterization of the prototypical groups against other baseline techniques.

7.7 Evaluation Method

In this section we discuss the method of evaluating our framework. First we would discuss several evaluation metrics. Next we would discuss our main idea of the evaluation method, followed by the baseline techniques used in the chapter.

Evaluation Metrics

In this work, we have used four different metrics to evaluate our framework. The first two metrics, conductance and coverage are traditional quality metrics of sub-graphs, while the rest two—predictivity and resilience justify the quality of the extracted prototypical groups.

- *Conductance.* Conductance (Bollobas, 1998) is a measure of how ‘well-knit’ a graph is. A group g with small conductance is of high quality because the similarities of its member individuals with those in other groups are small relative to the similarities of either the group members or members in other groups. At t_i , conductance of a communication over all its prototypical groups

K_i is defined as,

$$\rho_i = \frac{1}{K_i} \sum_{g \in K_i} \frac{\sum_{u,v \in g} \mathbf{S}_i(u,v)}{\min(\sum_{u \in g, v \in K_i} \mathbf{S}_i(u,v), \sum_{u \notin g, v \in K_i} \mathbf{S}_i(u,v))}, \quad (7.17)$$

where $\mathbf{S}_i(u,v)$ is the similarity between individuals u and v .

- *Coverage.* Coverage measures the fraction of similarity edges that are intra-group with respect to all edges in the social network of the communication set. A group with high coverage would have high cohesiveness in terms of communication characteristics. At t_i , coverage of a communication over all its prototypical groups K_i is defined as,

$$\chi_i = \frac{1}{K_i} \sum_{g \in K_i} \frac{\sum_{u,v \in g} \mathbf{S}_i(u,v)}{\sum_{u,v} \mathbf{S}_i(u,v)}. \quad (7.18)$$

- *Predictivity.* The extracted prototypical groups in a communication set are meaningful when they can predict the stock market movement / polls reasonably well. Hence we define a metric, called predictivity which gives a measure of how well a communication set, using its prototypical groups, can explain the actual stock market movement / polls at a certain point of time.

The predictivity α_i (of stock market movement / polls) of a communication set at a particular time slice t_i is given by an exponential function of the normalized absolute difference between the maximum predicted value determined by one of its prototypical groups g ($\hat{V}_{g,i}$) and the actual value at the same point of time (V_i)—the lesser the difference, the higher the predictivity. Predictivity of a communication set at t_i is given as:

$$\alpha_i \triangleq \arg \max_g \exp \left[- \left(\frac{|\hat{V}_{g,i} - V_i|}{|V_i|} \right) \right]. \quad (7.19)$$

- *Resilience.* The prototypical groups in a communication set are meaningful when their removal affects greatly the ability of the communication set to predict stock market movement / polls. That is, if the decrease in predictivity of a communication set due to removal of the prototypical groups is very high, then it would imply that the communication set is not resilient to removal of the prototypical groups.

In this work, resilience is defined to be the ability of a community to predict stock market movement / political poll at a particular time slice satisfactorily when the prototypical groups are not considered. It is given as:

$$\beta_i = \frac{\gamma_i}{\alpha_i}, \quad (7.20)$$

where γ_i is the predictivity of the communication set using the groups other than the prototypical ones.

Baseline Techniques

We use a comprehensive set of baseline techniques to evaluate different aspects of our framework. We need to evaluate two aspects of our model: (a) the method of extracting groups in a communication set, and (b) the method of identifying the prototypical groups at each time slice. Hence we devise two different baseline techniques for alternate group extraction and three different methods to alternatively identify the prototypical groups.

Baseline Techniques for Group Extraction. The two different baseline techniques for group extraction used in this article are: (a) group extraction based on communication co-occurrence of individuals in a communication set, and (b) group extraction based on a standard unsupervised clustering algorithm, kernel k -means.

In (a) we extract groups by clustering individuals in a communication set according to their communication co-occurrence with each other. Our main motivation is that we want to illustrate the usefulness of characterizing individuals based on a several different communication properties. Hence for this baseline technique, we characterize individuals only based on their communication co-occurrence with others. We first construct several graphs at each time slice t_i where the individuals are the nodes and a directed edge from node u to node v indicates that u preceded v in writing comments on blog posts at t_i relating to the topic of the communication set. Next we use such graphs to compute similarities (L^2 norm) between pairs of individuals at each time slice t_i , which we call co-occurrence in communication—

two individuals would have high similarity if they co-occur with each other in their communication several times. The similarity matrix is finally used to extract groups using the spectral clustering algorithm (Ng et al., 2001).

In (b) we extract communication characteristics of individuals based on our method and compute pair-wise similarities between individuals using the diffusion kernel. Our motivation is to evaluate the usefulness of our group extraction algorithm. Hence the group extraction algorithm used in this baseline technique is a standard clustering algorithm, kernel k -means which does not emphasize on the communication existing between pairs of individuals. The kernel chosen for the algorithm is the RBF kernel and k is taken to equal to the number of dominant eigenvalues of the similarity matrix.

Baseline Techniques for Identifying Prototypical Groups. We use three different baseline techniques to identify the prototypical groups in a communication set—(a) conductance based, (b) coverage and (c) communication activity based.

In (a) we compute the measure of conductance for all the extracted groups in a communication set at a certain time slice. Based on an empirically set threshold, we pick a set of groups as the prototypical groups which have the minimum conductance.

Similarly in (b) we pick a set of groups as the prototypical groups which have the maximum coverage.

In (c) we pick those groups as the prototypical groups which have the maximum degree of communication activity, in the form of the number of comments and replies. That is, if a group g at t_i has several comments and replies compared to others, it can be assumed to be a prototypical group.

Thereafter, using all the baseline techniques together with our method, we construct several evaluation cases= scenarios (12 in all) to evaluate our framework. In each of these scenarios, we compute four values corresponding to the four evaluation metrics—conductance, coverage, predictivity and resilience. A schematic representation of the evaluation method is shown in Figure 7.8.

Figure 7.8: The evaluation method used in this paper to test various aspects of our framework. There are two major ideas to be evaluated—the group extraction method (several baselines are shown in the left column) and the method used for identifying the prototypical groups (several baselines are shown in the right column). Corresponding to all possible cases (shown through the edges—in all 12 cases), four different measures are computed at each time slice t_i : conductance (ρ_i), coverage (χ_i), predictivity (α_i) and resilience (β_i). Our approach has been shown in the red colored edge.

Now we discuss the experimental results in the following section. There are two datasets which have been used in this chapter—a technology blog called Engadget, and a political blog called Huffington Post.

7.8 Experiments on Engadget Dataset

In this section, we discuss the experimental results based on the technology blog Engadget. First we present the statistics of the Engadget dataset used for analysis. Next, we present an analysis of the characteristics of extracted prototypical groups relating to the four topics i.e. names of technology companies: Apple, Microsoft, Google and Nokia. Third we analyze the utility of the prototypical groups by their ability to predict stock market movements of the corresponding companies. Finally we evaluate our framework by comparing it against several base-

Table 7.1: Statistics of different communication sets in the Engadget dataset.

COMMUNICATION SET	#INDIVIDUALS	#COMMENTS/REPLIES
Apple	52,296	2,941,358
Microsoft	46,882	1,908,959
Google	27,525	1,787,004
Nokia	14,158	82,935

line techniques based on the four metrics—conductance, coverage, predictivity and resilience.

Description of Dataset

We have executed a web crawler on the gadget-discussing blog called Engadget to collect data. We have chosen four different topics - Apple, Microsoft, Google and Nokia to capture diverse communication dynamics in the blogs. The duration of the crawl was from Apr 1, 2008 to Aug 31, 2008. We collected a set of 78,740 individuals who wrote 6,580,256 comments and replies with mean number of comments / replies per individual being 83.57. For analysis we have considered the top 10% individuals with the maximum number of authored comments corresponding to each topic.

We briefly discuss the statistics of each of the communication sets - Apple, Microsoft, Google and Nokia in Table 7.8. We observe that the Apple and Microsoft communication sets are considerably large in size as well rich in communication compared to Google or Nokia. Moreover there seems to be considerable overlap among the individuals communicating on different sets; however the comments are almost always associated with exactly one of the communication sets, with little overlap among sets.

Now we present the experimental results in the following sub-sections based on this dataset.

Analysis of Characteristics of Prototypical Groups

In this sub-section, we discuss the results of extraction and characterization of prototypical groups in four Engadget communication sets—Apple, Microsoft, Google

and Nokia. We extracted groups for each of these communication sets based on our method, and characterized the prototypical groups. The duration of a time slice used for the experiments was one week, while the total time period of analysis discussed here ranged from May 19 to Aug 31, 2008. The prototypical groups are analyzed from two perspectives—their composition entropy and topic divergence, as well as their conductance and coverage.

Analysis of Composition Entropy & Topic Divergence. The dynamics of the prototypical groups based on their mean composition entropy and topic divergence for the four communication sets are shown in Figure 7.9 and Figure 7.10. The figures show (1) the groups (rows) at each week, extracted using our method, and (2) the prototypical groups in each communication set (shown as shaded circles—intensity representing their measure of optimality) at each week. The visualizations are shown for a period of 15 weeks spanning from May 19 to August 31, 2008. The figures also show representative significant events over time associated with each communication set, along with the number of such events (shown using the numbers in the circles). These are useful in analyzing the dynamics of the mean composition entropy and the mean topic divergence of the prototypical groups which are also shown for the four cases.

We briefly discuss our method of extracting these representative events. We focused on the popular news site New York Times⁶ which features news stories into several categories like ‘Most Searched’, ‘Most Blogged’ and ‘Most Emailed’. For each week of our analysis, we identified the stories having the keywords ‘Apple’, ‘Microsoft’, ‘Google’ or ‘Nokia’ in their titles and corresponding to each of the featured categories, ‘most searched’, ‘most blogged’ and ‘most emailed’. We collated all the news stories at each week per communication set to get a representative set of events. From this set, we picked one event at random to illustrate our analysis in Figure 7.9 and Figure 7.10. The total numbers of representative events have also been shown in the figure.

⁶<http://www.nytimes.com/>

We gain interesting insights from Figure 7.9 and Figure 7.10. In the case of the communication set ‘Apple’, we observe that the mean composition entropy of the prototypical groups increases while the topic divergence decreases during time periods involving significant company related events.

We explain this as follows. In our dataset Engadget, there are large sets of individuals who consistently discuss Apple related events and products and are likely to exhibit widely different communication characteristics, like response behavior or frequency of communication. An artifact of this bias in our dataset could result in high movement of individuals across groups in the Apple communication set prior to times of important external events; yielding high composition entropy. We further explain the low topic divergence by the observation that several of these groups being interested in Apple related events, would be aligned to each other in their discussion about the current significant happening(s) over time.

Now we analyze the dynamics of the prototypical groups for the communication set ‘Microsoft’. This communication set is also characterized by high composition entropy and low topic divergence during periods of representative significant events. Our explanation lies firstly, in the observation that the set of individuals on Engadget who are interested in discussing Microsoft related events and products, is very large and diverse. Hence they are likely to have widely varying communication characteristics which change dynamically over time based on the different external events related to Microsoft. This feature of the dataset yields large movement of individual across groups during different time periods and therefore we observe the composition entropy to be high during the periods involving significant external events. Secondly, the topic divergence is low because several of the individuals in this communication set appear to be interested in discussions about the significant events at a particular point of time. For the communication set ‘Google’ we observe that both the composition entropy and the topic divergence are low during time periods of significant events. This is explained by the observation that there are several cohesive groups on Engadget who are interested in Google related events. There is little movement of individuals across the groups in this communication set,

and this explains the low composition entropy. Low topic divergence implies that the prototypical groups in this communication set involve in discussions related to the external events that is aligned with the entire communication set.

Finally in the case of the communication set ‘Nokia’, we can make similar observations and the results can be explained by the fact that this communication set is a small focused community whose movement across groups is low while their discussions relating to events on Nokia are reflective of the communication in the entire set.

Nevertheless it is important to note that the variations in the composition entropy and topic divergence for each of these communication sets are not necessarily correlated with presence of significant external events - there could be time periods where there are significant events, however we do not notice the characteristic changes in the two values. For example, weeks of Jul 14 and Aug 25 exhibit low composition entropy for the communication set ‘Apple’ despite the presence of significant events. Similar examples are weeks of Jul 21 and Aug 18 for ‘Microsoft’; Jun 2, Jun 9 and Jun 30 for ‘Google’ and Jun 2, Jun 16 and Aug 25 for ‘Nokia’.

We further make interesting observations about the sizes of the different prototypical groups over the period of 15 weeks from Figure 7.9 and Figure 7.10. We notice that the size of a group does not necessarily indicate that it would be a prototypical group of the communication set. For the sets ‘Apple’ and ‘Microsoft’ we note that there are several groups of small sizes which emerge as prototypical at different points in time. On the other hand, in the case of ‘Google’ and ‘Nokia’ groups of large sizes are the prototypical groups at several time periods. This justifies that using composition entropy and topic divergence as characteristics to identify the prototypical groups is meaningful.

We further bolster our method of characterizing the key groups by illustrating some representative communication content of the prototypical groups with respect to the entire communication set. In Table 7.8, we show representative keywords of such groups against that of the entire set over the period of 15 weeks of analysis. We observe that there is considerable overlap between the two, further

Table 7.2: Comparison of the content of communication of the prototypical groups with that of the entire communication set, for the four different sets in Engadget. The results are shown aggregated over the 15 weeks period from May 19 to Aug 31, 2008. We observe that the content of communication in the prototypical groups is representative of that in the entire communication set.

COMMUNICATION SET	REPRESENTATIVE KEYWORDS OF THE PROTOTYPICAL GROUPS	REPRESENTATIVE KEYWORDS OF THE ENTIRE COMMUNICATION SET
Apple	iPhone, 3G, AT&T, cell phones, data plan, Wi-Fi connectivity, quad-band GSM, release, Mac sales, fourth quarter, ipod sales, apple financial, profit	iPhone, release, ambient light, photo-sharing Bluetooth, AT&T, 8 GB flash drive, 3G, Wi-Fi, storage, iPods, Sales growth, IMac sales, ipod sales, Mac sales, stock price
Microsoft	Social networking, Facebook, advertising, deal, Yahoo, search, stake, Live Search, business, Stocks, bid, Google, competition, web, shareholder, negotiation, merger, advertising	Facebook, Yahoo, deal, Xbox, Windows, Vista, Google, stake, search, ads, Xbox 360 controller, advertising, Yahoo, online advertising, software giant, bid, Zune, release
Google	Ebooks, Verizon, Google Talk, GPhone, GoogleOS, Google Apps, Mobile WiMax, Sprint Nextel, internet, Blackberry, OS, iPhone, Google Talk Wifi phone	QuickTime, GoogleVideo, GoogleMaps, GPS, alliance, Google Phone, GoogleOS, internet, hardware, rumor, software, speculation, Orange, Tom Tom, Google Apps, Clearwire, LG
Nokia	Nokia Ovi, Vodafone, nseries, Twingo Nokia, Nokia N95, Nokia N800, Nokia iPhone, Wisair, Nokia E71, N-Gage, T-Mobile, patent payments	Nokia_NGage, cellPhone, navigator, Nokia N95, Vodafone, nseries, 3GSM, Qualcomm, Nokia E71, N-Gage, T-Mobile, CDMA

justifying that characterizing the prototypical groups with our method can indeed be concise representation of the entire communication set.

We explain this as follows. In our dataset Engadget, there are large sets of individuals who consistently discuss Apple related events and products and are likely to exhibit widely different communication characteristics, like response behavior or frequency of communication. An artifact of this bias in our dataset could result in high movement of individuals across groups in the Apple communication set prior to

times of important external events; yielding high composition entropy. We further explain the low topic divergence by the observation that several of these groups being interested in Apple related events, would be aligned to each other in their discussion about the current significant happening(s) over time.

Now we analyze the dynamics of the prototypical groups for the communication set ‘Microsoft’. This communication set is also characterized by high composition entropy and low topic divergence during periods of representative significant events. Our explanation lies firstly, in the observation that the set of individuals on Engadget who are interested in discussing Microsoft related events and products, is very large and diverse. Hence they are likely to have widely varying communication characteristics which change dynamically over time based on the different external events related to Microsoft. This feature of the dataset yields large movement of individual across groups during different time periods and therefore we observe the composition entropy to be high during the periods involving significant external events. Secondly, the topic divergence is low because several of the individuals in this communication set appear to be interested in discussions about the significant events at a particular point of time.

Analysis of Conductance & Coverage. Now we analyze the prototypical groups in these communication sets compared to the non-prototypical ones based on two measures - conductance and coverage. A comprehensive set of results for the four sets over the period of 15 weeks from May 19 to Aug 31 has been presented in Figure 7.11. Two values are shown: conductance and coverage for two cases—the prototypical groups only, and second the non-prototypical groups.

We can make three observations from the results. First, the values of conductance for the prototypical groups are significantly lower than those of the non-prototypical ones; while coverage is higher. This implies that the prototypical groups are meaningful and are better qualitative representatives of the communication set compared to the other groups. Our second observation is that during time periods of significant events, the measure of the conductance of the prototypical groups,

decreases, while the coverage increases. This is because during these particular time periods, the groups are cohesively involved in communication, implying that there are more intra-group edges compared to inter-group edges in the social network of the communication set. The non-prototypical groups do not capture the communication dynamics of the set as much well as the prototypical ones, hence there are significant inter-group edges implying considerable inter-group similarity in characteristics of the individuals.

Third, we observe that variation of the number of prototypical groups over the four communication sets. The number of such groups for ‘Apple’ and ‘Microsoft’ are observed to be higher than that of ‘Google’ and ‘Nokia’. We explain the reason as follows. In our Engadget dataset, the size of the first two communication sets is much larger than that of the rest two; and so is the degree of communication activity. These sets are also characterized by highly volatile group membership of individuals as observed from Figure 7.9 and Figure 7.10. As a result the number of groups reaching our optimality value in terms of high composition entropy and low topic divergence is high. This explains why they are noticed to have comparatively larger numbers of prototypical groups.

To summarize, from this sub-subsection, we are able to study comprehensively the characteristics of the prototypical groups as well as are able to distinguish their properties from the rest of the groups in the communication sets. We have observed that the dynamics of such properties are closely related to the presence of significant happenings in the external world and their dynamics intuitively explain the behavior of the entire communication set.

In the next sub-section we would analyze the utility of the prototypical groups in their ability to predict an external variable. In this article, the external variable used for the communication sets in Engadget is stock market movement of the corresponding companies.

Table 7.3: Analysis of the utility of prototypical groups of four communication sets in the Engadget dataset. The prototypical groups have been used to predict a related external variable, stock market movements of the corresponding sets in the time period from May 19 to Aug 31, 2008. The actual stock price at each Friday at each week is shown, along with the error in predicting the corresponding movement based on the prototypical groups. Considerably low errors in prediction are observed; that justifies that the characterized prototypical groups are meaningful.

WEEK	APPLE		MICROSOFT		GOOGLE		NOKIA	
	Stock Price	E (%)	Stock Price	E (%)	Stock Price	E (%)	Stock Price	E (%)
May 19	181.17	22.16	28.05	18.02	544.62	23.14	28.45	22.63
May 26	188.75	20.09	28.32	20.20	585.80	21.64	28.40	19.87
Jun 2	185.64	24.86	27.49	22.74	567.00	22.46	26.13	24.75
Jun 9	172.37	19.36	29.07	21.26	571.51	22.76	26.08	18.64
Jun 16	175.27	26.88	28.23	23.83	546.43	24.72	24.29	22.13
Jun 23	170.09	25.67	27.63	23.36	528.07	19.42	24.03	19.24
Jun 30	170.12	23.24	25.98	22.24	537.00	18.84	24.50	17.36
Jul 7	172.58	22.78	25.25	21.52	533.80	18.15	26.01	25.98
Jul 14	165.15	22.24	25.86	20.82	481.32	24.37	27.51	20.78
Jul 21	162.12	21.25	26.16	20.93	491.98	18.64	28.13	21.36
Jul 28	156.66	18.85	25.44	22.73	467.86	25.21	27.12	20.84
Aug 4	169.55	24.32	28.13	24.85	495.01	23.73	27.07	17.64
Aug 11	175.74	27.46	27.81	24.36	510.15	23.15	25.47	22.22
Aug 18	176.79	25.84	27.84	23.58	490.59	22.37	26.31	20.86
Aug 25	169.53	23.56	27.29	23.27	463.29	21.58	25.17	20.46

Analysis of Utility of Prototypical Groups

Now we justify the utility of the extracted prototypical groups of different communication sets, Apple, Microsoft, Google and Nokia with their ability to predict stock market movement of the corresponding companies.

To predict the stock movements for each communication set using the prototypical groups at each week, we first trained a Support Vector Regressor over the period from Apr 1, 2008 to May 16, 2008. The predictions of stock movement were thereafter made at each week from May 19, 2008 to Aug 31, 2008 (15 weeks, same period as Figure 7.9 and Figure 7.10). The results of prediction of stock market movements for the four technology companies are shown in Table 7.8. The table shows actual stock price at each Friday at each week, along with the error in predicting the corresponding movement based on the prototypical groups.

We observe that the error in predicting the stock movement is reasonable: 23.42%, 22.58%, 19.67% and 18.73% respectively for communication sets Apple, Microsoft, Google and Nokia. This justifies the utility of the extracted prototypical

groups in the four sets. We observe that the error rates for Google and Nokia are lower than that of Apple and Microsoft. We conjecture that the former two sets exhibit small changes in the communication dynamics of their individuals—also the sizes of these sets are comparatively smaller, resulting in less diversity of communication characteristics. Hence our Regressor is able to predict the stock market movement better based on these communication dynamics.

Note, we do not claim that prediction necessarily indicates consequential relationship between the prototypical group dynamics and stock market movement; rather it simply illustrates that such groups are meaningful representations of the corresponding communication sets.

It is also important to point it out here that prediction of a highly volatile variable like stock market movement is extremely difficult due to its dependency on a number of unforeseen factors. A technology community like Engadget can only reflect the consumer sentiment to a certain degree, and hence this justifies our prediction error rates. Also note, our goal is not to be able to predict the movement, rather to analyze if the prototypical groups can meaningfully represent a communication via its ability to predict an external variable to a considerable degree.

In the following sub-section, we present a comprehensive evaluation of our method using several baseline techniques.

Evaluation of Prototypical Groups

Now we present the results to evaluate our method of extracting and characterizing the prototypical groups using a set of baseline methods.

We have constructed several baseline comparison cases (12 in all, including our own method⁷) utilizing the two types of baseline techniques—first, baseline tech-

⁷The descriptions of the 11 different scenarios including our method are given as follows: B_1 —communication co-occurrence + minimum conductance; B_2 —communication co-occurrence + maximum coverage; B_3 —communication co-occurrence + maximum activity; B_4 —kernel k -means + minimum conductance; B_5 —kernel k -means + maximum coverage; B_6 —kernel k -means + maximum activity; B_7 —communication co-occurrence + composition entropy, topic divergence

Table 7.4: Evaluation of our framework against 11 different baseline scenarios, involving different group extraction and prototypical group identification methods. For each method we show four values averaged over all the prototypical groups, per communication set—conductance (ρ), coverage (χ), predictivity (α) and resilience (β). These values are averaged over the 15 weeks of analysis, May 19 to Aug 31, 2008 for the Engadget dataset.

	APPLE				MICROSOFT				GOOGLE				NOKIA			
	ρ	χ	α	β	ρ	χ	α	β	ρ	χ	α	β	ρ	χ	α	β
B₁	0.43	0.4	0.27	0.65	0.66	0.54	0.26	0.59	0.28	0.51	0.26	0.5	0.44	0.3	0.28	0.59
B₂	0.36	0.46	0.29	0.56	0.62	0.58	0.3	0.55	0.25	0.57	0.35	0.5	0.39	0.32	0.36	0.59
B₃	0.34	0.46	0.32	0.56	0.6	0.6	0.39	0.54	0.24	0.63	0.35	0.46	0.38	0.34	0.41	0.45
B₄	0.33	0.5	0.36	0.54	0.47	0.6	0.4	0.49	0.24	0.69	0.39	0.44	0.38	0.36	0.45	0.44
B₅	0.32	0.52	0.39	0.51	0.45	0.61	0.52	0.43	0.23	0.7	0.53	0.42	0.38	0.41	0.47	0.43
B₆	0.3	0.53	0.39	0.47	0.43	0.62	0.54	0.41	0.23	0.74	0.54	0.36	0.32	0.45	0.5	0.43
B₇	0.27	0.58	0.49	0.42	0.38	0.66	0.57	0.36	0.19	0.74	0.55	0.35	0.25	0.46	0.54	0.36
B₈	0.21	0.61	0.5	0.32	0.31	0.66	0.59	0.34	0.16	0.77	0.59	0.17	0.25	0.47	0.57	0.3
B₉	0.19	0.62	0.54	0.27	0.28	0.7	0.68	0.29	0.15	0.79	0.67	0.14	0.23	0.51	0.69	0.25
B₁₀	0.17	0.64	0.54	0.21	0.23	0.73	0.73	0.28	0.12	0.84	0.73	0.13	0.16	0.55	0.8	0.18
B₁₁	0.15	0.66	0.79	0.21	0.22	0.74	0.8	0.21	0.12	0.85	0.78	0.12	0.14	0.56	0.81	0.18
M	0.14	0.71	0.81	0.14	0.21	0.78	0.82	0.11	0.11	0.86	0.83	0.12	0.12	0.69	0.82	0.16

niques for evaluating group extraction and second, baseline techniques for evaluating the method of identifying the prototypical groups. We discuss the results of comparison of our method against these different baseline scenarios based on two graph based evaluation metrics—conductance and coverage and also the two evaluation metrics presented in this article—predictivity and resilience as discussed before.

Now we present the results of evaluation in Table 7.8. The table shows the measures of stock market movement predictivity and resilience for four communication sets: ‘Apple’, ‘Microsoft’, ‘Google’ and ‘Nokia’ based on our method (M) and the baseline scenarios ($B_1 - B_{11}$) averaged in the period from May 19, 2008 to Aug 31, 2008.

From the results, we are interested in the following observations: (a) whether extracting groups in communities based on communication dynamics of individuals is important, and (b) whether identifying prototypical groups based on composition (our method); B_8 —kernel k -means + composition entropy, topic divergence (our method); B_9 —MAE (our method) + minimum conductance; B_{10} —MAE (our method) + maximum coverage; B_{11} —MAE (our method) + maximum activity; M —MAE (our method) + composition entropy, topic divergence (our method).

entropy and topic divergence is useful. Let us analyze how the different scenarios can address each of these aspects:

- The first three baseline scenarios ($B_1 - B_3$) extract groups simply based on communication co-occurrence of individuals and hence ignore the dynamics of their communication properties. For identifying the prototypical groups, they use the groups with minimum conductance, maximum coverage and maximum activity respectively. Since these measures of identifying the prototypical groups do not consider the temporal dynamics of how they represent the overall communication set, they are observed to perform poorly. Note, B_3 is seen to perform better than the other two implying that the communication activity of the individuals is useful; however high activity in a group does not necessarily indicate that it would be prototypical.
- The baseline scenarios $B_4 - B_6$ use a standard unsupervised clustering algorithm kernel k -means to extract groups. We observe that these scenarios perform better as the features used by the clustering algorithm are based on the communication dynamics of the individuals. We also note that the conductance and coverage based scenarios B_4 and B_5 perform poorer than B_6 as they do not capture the communication activity in identifying the prototypical groups.
- The scenarios B_7 and B_8 extract groups based on communication co-occurrence of individuals and kernel k -means respectively. However, they use our method of identifying the prototypical groups. Hence we observe that they perform better than $B_1 - B_6$, since the prototypical groups capture the context and content of the entire communication set well, improving their ability to predict an external variable. However since the quality of the groups themselves does not incorporate the communication dynamics of the individuals based on their mutual awareness in communication, we notice that the measures of predictivity and resilience are therefore poorer than that by our method.

- In $B_9 - B_{11}$, we extract groups using our method that incorporates the idea of mutual awareness in extracting the groups, as well as uses the different communication features of individuals. However the prototypical groups for these scenarios are based on conductance, coverage and maximum activity based measures. Hence they are not able to develop useful prototypical representations of communication in the entire set. The measures of predictivity of an external variable and the associated resilience are therefore low in these scenarios.

Hence finally we observe that our method M yields maximum predictivity of stock market movement (average for our method: 0.76, for baselines: 0.43) as well as minimum resilience of the four communication sets (average for our method: 0.16, for baselines: 0.44). The mean improvement in conductance over the four sets compared to the baseline scenarios is approximately 30%, that in coverage is 27%, that in predictivity of the polls is 45% and in resilience is 42%.

Thus our method can effectively address the two aspects of evaluation as stated above better than baseline methods, both in terms of quality, that is given by the graph based metrics conductance and coverage, as well as the ability to explain a correlated external variable—in this case being the stock market movements of the companies. This justifies our group extraction framework, as well as our method of identifying the prototypical groups.

7.9 Experiments on Huffington Post Dataset

In this section we present experiments involving the political blog Huffington Post to establish the generalizability of our framework. We present a description of the dataset first; then discuss results of the experiments conducted on it—analysis of the characteristics of the extracted prototypical groups, and their utility. Finally we evaluate our method using the baseline techniques and the corresponding measures of the four evaluation metrics.

Table 7.5: Statistics of the two communication sets in the Huffington Post dataset.

COMMUNICATION SET	# INDIVIDUALS	#COMMENTS/REPLIES
Obama	39,692	3,522,5663
McCain	27,858	2,893,896

Description of Dataset

In order to test the generality of our multi-scale characterization framework, we have collected data on blog discussions involving the two Presidential candidates of 2008 US elections (Barack Obama and John McCain) from a highly popular and active political blog Huffington Post. We deployed a web based crawler as before, to collect data on the blog comments, their timestamps and authoring individuals. The dataset comprises a set of 59,282 individuals who wrote a set of 4,748,837 comments in the time span from May 15, 2008 to Oct 3, 2008. The mean number of comments per individual in this dataset is 80.11.

In Table 7.9, we provide an overview of the statistics of the two communication sets—‘Obama’ and ‘McCain’ in this dataset. Interestingly, we observe that the two sets have considerable overlap in terms of the individuals as well as the comments posted on either of the two electoral candidates. This is a property of the Huffington Post dataset where, during the crawled time period there was extensive discussion on the two Presidential candidates, Obama and McCain, hence several blog posts were tagged with both of them, and thereby the comments and replies of each of the two sets had considerable overlap.

In the following sub-sections we would present the results of the experiments conducted on this dataset.

Analysis of Characteristics of Prototypical Groups

In this sub-section, we present an analysis of the characteristics of the prototypical groups extracted from the two communication sets—‘Obama’ and ‘McCain’ in the Huffington Post dataset. We extracted groups based on our method and thereafter characterized the prototypical groups. We now analyze these prototypical groups

from two perspectives as with the Engadget dataset—their composition entropy and topic divergence; and their conductance and coverage.

Composition Entropy & Topic Divergence. We analyze the prototypical groups in each of the communication sets ‘Obama’ and ‘McCain’ based on the two characteristics—composition entropy and topic divergence. We present a visualization of the characterization of the prototypical groups for the two sets over a period of 15 weeks from Jun 23, 2008 to Oct 3, 2008 in Figure 7.12.

For the communication set ‘Obama’, we observe here that both the composition entropy and topic divergence decrease during periods of significant external happenings related to the Presidential candidate Obama. Similar observations can be made for the communication set, ‘McCain’.

We observe that these dynamics occur because there are specific groups of individuals who support Obama and McCain on the website Huffington Post—both being candidates of the two major political parties, The Republicans and The Democrats. As a result these individuals form cohesive groups where the individuals in a group share similar communication characteristics. These characteristics are not subject to much change over time; hence the movement of individuals across groups is low—yielding low composition entropy. Low topic divergence implies that the discussions in the prototypical groups are focused on the major external happenings and hence are aligned to the overall discussion trend in the communication set.

However there are some exceptions to these observations as well, which show that these are not necessary and sufficient conditions for observing the external happenings in relation to the dynamics of the prototypical groups. For example, in the week of Aug 18 we observe high topic divergence. We explain this by the observation that the corresponding event being about the nomination of Senator Joe Biden as the Democrat VP candidate, there are likely to be a lot of communication concerning him, rendering the communication content in the entire set to be rather diverse about various topics. This yields high topic divergence at this week.

In the case of the communication set ‘McCain’, we observe high topic diver-

Table 7.6: Comparison of the content of communication of the prototypical groups with that of the entire communication set, for the two different sets in Huffington Post. The results are shown aggregated over the 15 weeks period from Jun 23 to Oct 3, 2008. We observe that the content of communication in the prototypical groups is representative of that in the entire communication set.

COMMUNICATION SET	REPRESENTATIVE KEYWORDS OF THE PROTOTYPICAL GROUPS	REPRESENTATIVE KEYWORDS OF THE ENTIRE COMMUNICATION SET
Obama	barack-obama, humanitarian, iran, iraq, joe-lieberman, jordan, syria, zbigniew-brzezinski, gen-petreaus, harry-reid, john-edwards, move-onorg, senatedemocrats, senate-republicans, obama-senate-foreign-relations-committee, joe-biden, California	barack-obama, barack-obama-troop-withdrawal, obama-petreaus, obama-senate-foreign-relations-committee, california, camp-obama, off-the-bus, san-francisco-speech, , iran, iraq, harry-reid, joe-biden, Michelle Obama Proud, Tennessee GOP, Dalai-Lama, political-news, on the ground 2008
McCain	Iran, National Security, Fear Watch 2008, John McCain Iran, John McCain National Security, McCain Diplomacy, McCain National Security, McCain Iran, McCain Unfit, Politics News, Cindy McCain, Cindy McCain Campaign	John McCain Socialist, McCain Kansas City Star, McCain Obama Extremist, McCain Liberal, McCain Obama Socialist, Politics News, Iran, McCain Diplomacy, McCain National Security, McCain Iran, Cindy McCain

gence in the weeks of Aug 25, Sep 1 and Sep 22. On noticing the associated external events, we observe that they are related to the nomination of the Republican VP candidate Governor Sarah Palin and the start of the Financial Crisis. As a result it is likely that several individuals discussing Senator McCain would be discussing about these events on the weeks of Aug 25, Sep 1 and Sep 22. This is likely to be responsible for the high topic divergence of the prototypical groups during these weeks.

Therefore these results suggest that our results are generalizable and our method captures the temporal dynamics of the overall communication sets significantly well based on the prototypical groups.

Now we analyze qualitatively how the prototypical groups represent the overall communication in the entire set. Representative keywords of the prototypical groups as well as the entire communication set for ‘Obama’ and ‘McCain’ is shown in Table 7.9. The results are aggregated over the time period of analysis from Jun 23 to Oct 3, 2008. We observe that there is considerable overlap between the two, further justifying that characterizing the prototypical groups with our method can indeed be concise representation of the entire communication set.

Conductance & Coverage. We analyze the prototypical groups in the two Huffington Post communication sets compared to the non-prototypical ones based on two measures—conductance and coverage. The results for the sets over the period of 15 weeks from Jun 23 to Oct 3 have been presented in Table 7.9. We show two values—conductance and coverage for two cases—the prototypical groups only, and second the non-prototypical groups.

We make similar observations as Engadget from this dataset as well. The measures of conductance and coverage for the prototypical groups are lower and higher respectively than those of the non-prototypical groups. Moreover, the number of prototypical groups for the communication set ‘Obama’ is higher than that of ‘McCain’. This can be explained by the fact that a characteristic of the Huffington Post dataset is that it is quite liberal inclined. Hence there are specific groups of individuals who consistently discuss about Senator Obama in their communication, yielding low composition entropy as well as having discussions highly inclined to the current external events. Therefore the number of the prototypical groups in this case is larger than that for the other set ‘McCain’.

From this sub-section we have observed the different characteristics of the prototypical groups for a political dataset Huffington Post. These characteristics indicate that they meaningfully capture the dynamics in the overall communication set at different points in time.

Now we would analyze the utility of these prototypical groups in their ability to predict the political polls in favor of each of the two Presidential candidates.

Table 7.7: Analysis of the prototypical groups against the non-prototypical ones based on two metrics—conductance and coverage. The results are shown for two communication sets -Obama and McCain, in the time period from Jun 23 to Oct 2, 2008. Five different values are shown per week and per communication set. The details of the different columns are as follows: Column A: the number of prototypical groups; Column B: the mean conductance over all the prototypical groups; Column C: the mean coverage over all the prototypical groups; Column D: the mean conductance over all the non-prototypical groups, i.e. the groups in the communication set not identified as prototypical groups; and finally, Column E: the mean coverage over all the non-prototypical groups. Note, the conductances for the prototypical groups are considerably lower than that of the non-prototypical ones; while the coverage is higher. This implies that the identified prototypical groups are of high quality and hence meaningfully model communication existing among the individuals in the communication sets.

WEEK	OBAMA					MCCAIN				
	A	B	C	D	E	A	B	C	D	E
Jun 23	7	0.22	0.66	0.34	0.55	2	0.55	0.32	0.67	0.20
Jun 30	5	0.27	0.69	0.37	0.54	8	0.32	0.70	0.43	0.57
Jul 7	5	0.24	0.81	0.38	0.66	5	0.29	0.74	0.42	0.61
Jul 14	7	0.31	0.76	0.45	0.63	3	0.17	0.66	0.27	0.51
Jul 21	5	0.21	0.80	0.33	0.69	3	0.23	0.73	0.36	0.59
Jul 28	7	0.19	0.67	0.31	0.55	5	0.23	0.76	0.33	0.62
Aug 4	5	0.42	0.27	0.53	0.15	3	0.18	0.74	0.29	0.62
Aug 11	6	0.18	0.78	0.32	0.63	6	0.25	0.62	0.38	0.47
Aug 18	7	0.19	0.74	0.30	0.63	5	0.16	0.80	0.29	0.67
Aug 25	4	0.18	0.75	0.28	0.64	4	0.56	0.43	0.66	0.31
Sep 1	8	0.56	0.35	0.69	0.24	7	0.64	0.47	0.78	0.32
Sep 8	7	0.23	0.62	0.34	0.48	4	0.23	0.75	0.36	0.60
Sep 15	7	0.22	0.76	0.33	0.62	4	0.21	0.76	0.33	0.63
Sep 22	4	0.19	0.81	0.31	0.67	6	0.54	0.50	0.66	0.36
Sep 29	5	0.57	0.36	0.67	0.23	3	0.19	0.71	0.31	0.58

Analysis of Utility of Prototypical Groups

In this sub-section we attempt predict the polls in favor of Barack Obama and John McCain obtained from the popular political news site Real Clear Politics⁸. Similar to our analysis method for Engadget, we first trained a Support Vector Regressor as discussed in section 6 over the period from May 15, 2008 to Jun 22, 2008. The predictions of political polls were thereafter made at each week from Jun 23, 2008

⁸<http://www.realclearpolitics.com/>

Table 7.8: Analysis of the utility of prototypical groups of two communication sets in the Huffington Post dataset. The prototypical groups have been used to predict a related external variable, national polls of the corresponding political candidates in the time period from Jun 23 to Oct 3, 2008. The actual percentage of the poll at each week is shown, along with the error in predicting the corresponding movement based on the prototypical groups. Considerably low errors in prediction are observed; that justifies that the characterized prototypical groups are meaningful.

WEEK	OBAMA		MCCAIN	
	National Poll	E (%)	National Poll	E (%)
Jun 23	48.2	17.12	41.4	25.59
Jun 30	47.7	21.32	40.9	15.47
Jul 7	48.5	28.73	42.8	19.15
Jul 14	46.6	26.88	42.6	15.69
Jul 21	46.0	29.39	41.3	16.45
Jul 28	46.5	24.83	43.3	27.35
Aug 4	46.6	15.53	44.3	25.42
Aug 11	46.8	27.73	42.0	19.75
Aug 18	45.4	29.00	41.8	29.25
Aug 25	45.5	25.18	43.9	15.51
Sep 1	48.8	26.36	44.3	21.58
Sep 8	48.3	26.14	45.4	20.72
Sep 15	46.3	20.88	44.7	26.48
Sep 22	48.1	24.83	45.4	26.92
Sep 29	47.9	17.56	43.3	17.80

to Oct 3, 2008 (15 weeks, same period as Figure 7.12). The results of prediction of polls in favor of the two Presidential candidates are shown in Table 7.9. The table shows actual value of the National poll averaged over each week, along with the error in predicting the poll based on the prototypical groups.

We observe that the mean error in prediction is 21.43% and 19.75% respectively for the corresponding two communication sets—‘Obama’ and ‘McCain’. The errors seem to be high on weeks which have several external happenings present—implying that the individuals engaged themselves in diverse communication topics and this affected our prediction performance. However the results do suggest that the prototypical groups are meaningful because they can explain a related external variable (polls) quite well.

We now present an evaluation of our framework of identifying prototypical groups in this dataset.

Evaluation of Prototypical Groups

In this section we present the evaluation of our framework using different baseline scenarios. We use two different baseline group extraction methods and three different prototypical group identification methods to compare the quality of prototypical groups extracted by our framework. As before, we compute four measures per scenario—conductance, coverage, predictivity and resilience. These results are presented in Table 7.9⁹.

The results reveal that communication co-occurrence based group extraction yields poor quality of prototypical groups as well as low predictivity of the national electoral polls as they cannot capture the rich communication dynamics of the individuals in the communication set. Kernel k -means based group extraction scenarios yield comparatively better results; however cannot capture the dynamics emerging out of mutual awareness of communication among them individuals. Among the prototypical groups identification methods, minimum conductance and maximum coverage based methods perform the worst, while maximum activity based method yields better predictivity and resilience with respect to the extracted prototypical groups. However it cannot capture the temporal dynamics of change in context and content of the communication set with respect to the prototypical groups. Among all methods, our framework yields best results in respect of the quality of the prototypical groups (i.e. least conductance and highest coverage) as well as ability to correlate blog communication dynamics with an external variable (i.e. predictivity and resilience for the political polls). The mean improvement in conductance over the two sets ‘Obama’ and ‘McCain’ compared to the baseline scenarios is 41.5%,

⁹The descriptions of the 11 different scenarios including our method are given as follows: B_1 —communication co-occurrence + minimum conductance; B_2 —communication co-occurrence + maximum coverage; B_3 —communication co-occurrence + maximum activity; B_4 —kernel k -means + minimum conductance; B_5 —kernel k -means + maximum coverage; B_6 —kernel k -means + maximum activity; B_7 —communication co-occurrence + composition entropy, topic divergence (our method); B_8 —kernel k -means + composition entropy, topic divergence (our method); B_9 —MAE (our method) + minimum conductance; B_{10} —MAE (our method) + maximum coverage; B_{11} —MAE (our method) + maximum activity; M —MAE (our method) + composition entropy, topic divergence (our method).

Table 7.9: Evaluation of our framework against 11 different baseline scenarios, involving different group extraction and prototypical group identification methods. For each method we show four values averaged over all the prototypical groups, per communication set—conductance (ρ), coverage (χ), predictivity (α) and resilience (β). These values are averaged over the 15 weeks of analysis, Jun 23 to Oct 3, 2008 for the Huffington dataset.

	OBAMA				MCCAIN			
	ρ	χ	α	β	ρ	χ	α	β
B₁	0.48	0.22	0.27	0.54	0.59	0.2	0.28	0.54
B₂	0.48	0.31	0.3	0.47	0.55	0.25	0.29	0.49
B₃	0.47	0.39	0.31	0.45	0.51	0.26	0.29	0.45
B₄	0.46	0.42	0.37	0.43	0.47	0.35	0.31	0.45
B₅	0.43	0.46	0.37	0.41	0.44	0.43	0.37	0.41
B₆	0.42	0.46	0.45	0.38	0.36	0.45	0.42	0.35
B₇	0.41	0.49	0.5	0.31	0.32	0.66	0.53	0.3
B₈	0.36	0.58	0.53	0.3	0.32	0.68	0.55	0.24
B₉	0.29	0.61	0.55	0.28	0.21	0.69	0.57	0.21
B₁₀	0.26	0.65	0.67	0.23	0.17	0.72	0.6	0.15
B₁₁	0.15	0.67	0.69	0.23	0.14	0.74	0.75	0.15
M	0.11	0.77	0.75	0.15	0.13	0.77	0.76	0.13

that in coverage is 56%, that in predictivity of the polls is 48% and in resilience is 40%.

Hence to summarize, from the results we observe that (a) group extraction based on communication features and mutual awareness among individuals is useful, and (b) identifying the prototypical groups based on composition entropy and topic divergence captures the context and content of communication in the entire communication set.

7.10 Experiments on Other Social Media Datasets

We further extend and evaluate our observations made from the experiments on Engadget and Huffington Post, to social media datasets Digg, Twitter, YouTube and Flickr. Note that since these datasets were crawled at different points in time, as well as the nature of content posted on them are rather eclectic (i.e. span a

Table 7.10: Description of topics and categories over which prototypical groups were detected.

Dataset	Topics
DIGG	Ontology-defined categories like ‘US Elections’, ‘World News’, ‘Olympics’, ‘Comedy’, ‘Celebrity’, ‘Tennis’.
TWITTER	Trending topic clusters like ‘Politics’, ‘Entertainment-Culture’, ‘Sports’, ‘Technology-Internet’, ‘Social Issues’, extracted using the natural language toolkit OpenCalais.
YOUTUBE	Ontology-defined categories like ‘Comedy’, ‘Education’, ‘Entertainment’, ‘News’, ‘Sports’, ‘Music’.
FLICKR	Tags associated with shared photos like ‘Arizona’, ‘Grand Canyon’, ‘2008’, ‘Spring’, ‘Travel’, ‘Nature’.

variety of topics from Politics, to Sports to Entertainment) it was not possible for us to utilize the stock market movements or the political polls for correlation based evaluation purposes. Hence in this section we present experiments on these datasets in a quantitative sense that judges the extensibility of our proposed algorithm of detecting prototypical groups in communication sets. Since a common feature of all these four social media datasets is that they involve communication among the individuals via comments (such as Digg comments on a voted item, tweets on a trending topic or a hashtag, comments on a shared video or a shared photo), we can deploy our algorithm directly on a set of pre-defined topics or categories for the purpose of detecting the prototypical groups over time.

We begin by narrating the different topics and ontological categories in each dataset (ref. Table 7.10). The prototypical groups were extracted for each of these topics and categories. Some of the statistics of the over these topics and categories for all the datasets are also shown in Table 7.11.

We perform two sets of experiments on evaluating prototypical groups on these four datasets. In the first set, we study how the composition entropy and topic divergence of the groups extracted from two communication sets vary over time, and compare these time series to the series of major external happenings obtained from the NY Times. In the second set, we study two measures, conductance and coverage

Table 7.11: Statistics of the different communication sets over the four social media datasets.

Communication Set	#Individuals	#Comments/Replies/Posts
DIGG		
‘US Elections’	40,737	1,027,393
‘World News’	45,290	746,609
‘Olympics’	6,350	833,515
‘Comedy’	45,669	817,446
‘Celebrity’	31,618	431,450
‘Tennis’	4,878	721,026
TWITTER		
‘Politics’	13,925	188,306
‘Entertainment-Culture’	27,345	776,651
‘Sports’	47,876	35,017
‘Technology-Internet’	48,245	304,616
‘Social Issues’	7,881	50,789
YOUTUBE		
‘Comedy’	48,530	106,845
‘Education’	47,859	905,804
‘Entertainment’	24,269	764,312
‘News’	40,015	348,810
‘Sports’	7,095	1,045,245
‘Music’	21,089	37,891
FLICKR		
‘Arizona’	45,787	482,619
‘Grand Canyon’	39,611	419,715
‘2008’	47,975	842,069
‘Spring’	32,788	874,720
‘Travel’	1,786	205,560
‘Nature’	42,457	538,741

for the sets of prototypical and non-prototypical groups and present comparative analysis of the two.

In Figure 7.13 we show the dynamics of two properties of the extracted prototypical groups over two different communication sets in the four datasets—composition entropy and topic divergence. In order to see the correlation with external events, we also show a trend line of MY Times mentions of each topic as well. Note that during periods of several media mentions, the composition entropy and topic divergence appear to increase, which is distinct from our observations made on the blog datasets. Our conjecture is that, because most of these social media feature shared content on timely happenings, hence the composition of the groups undergo

Table 7.12: Analysis of the prototypical groups against the non-prototypical ones based on two metrics—conductance and coverage. The results are shown for several topical communication sets belonging the four social media datasets. Note, p=Prototypical roudp, NP=Non-prototypical groups.

Communication Set	Cond. (P)	Cov. (P)	Cond. (NP)	Cov. (NP)
DIGG				
‘US Elections’	0.3444	0.8335	0.6199	0.2974
‘World News’	0.3717	0.7697	0.4580	0.2604
‘Olympics’	0.1381	0.7894	0.7412	0.3010
‘Comedy’	0.3740	0.7858	0.6488	0.2241
‘Celebrity’	0.2897	0.6981	0.5404	0.2330
‘Tennis’	0.1293	0.7639	0.6053	0.4355
TWITTER				
‘Politics’	0.1835	0.6428	0.5607	0.4390
‘Entertainment-Culture’	0.2641	0.7765	0.4304	0.3438
‘Sports’	0.3873	0.6080	0.4960	0.2149
‘Technology-Internet’	0.3895	0.6692	0.4493	0.2587
‘Social Issues’	0.1473	0.6115	0.4736	0.2883
YOUTUBE				
‘Comedy’	0.3912	0.6243	0.4960	0.4053
‘Education’	0.3872	0.8059	0.5669	0.2039
‘Entertainment’	0.2456	0.7737	0.4199	0.2108
‘News’	0.3401	0.6793	0.7611	0.2422
‘Sports’	0.1426	0.8376	0.7779	0.3623
‘Music’	0.2265	0.6086	0.5963	0.3829
FLICKR				
‘Arizona’	0.3747	0.7097	0.5957	0.3619
‘Grand Canyon’	0.3377	0.6954	0.5351	0.3127
‘2008’	0.3878	0.7914	0.7600	0.3368
‘Spring’	0.2967	0.7988	0.5477	0.2741
‘Travel’	0.1107	0.6467	0.4445	0.3862
‘Nature’	0.3547	0.7224	0.7121	0.2472

significant change during periods of important happenings—since several newcomers tend to participate. Also due to the increased volume of participation during these times, the divergence of the discussed topics appears to increase, relative to the overall communication in the entire communication set. Overall it appears that the properties of the prototypical groups are affected by external happenings in the real world.

Next we study two structural measures for the prototypical and the non-prototypical groups. These are conductance and coverage. For all the communication sets, we unanimously observe that conductance is lower for prototypical groups and coverage higher compared to the non-prototypical ones. This indicates the effectiveness of our method in being able to characterize useful representative groups in a large community, in which the individuals seem to be cohesive in their communication patterns.

7.11 Discussion

In this section we present a discussion of our results and its implications on generalizability of our framework on extraction, characterization and utility of prototypical groups in the Blogosphere to other diverse domains.

Communication among individuals in online communities provides a plethora of information and facts that associate themselves with several external events. Apparently, while bulk of this information seems to be extremely noisy and difficult to be sifted through, some of it, if mined, could be put to tremendous use in our day-to-day practical applications, e.g. determining who the key people in a community are, which groups are interesting or prototypical, which blogs or social media artifacts are important or highly ranked, who the information authorities are or how information flows in a social network. In this context, our framework is an extremely useful way of extracting the useful facts (or knowledge) in social network communication. The usefulness of such mined information can be verified by its correlation with semantically related external information sources—in certain cases, it might be interesting to observe if the mined knowledge possesses predictability characteristics, acts as a chatter creator or trails in its dynamics with respect to other significant events in the external world. Our framework and results show that the usefulness of the knowledge mined from specific types of social network communication (especially technology-related and political blogs) can be interestingly verified by correlating it with external happenings (i.e. in this context, stock market movements or political polls).

We justify our choice of a technology blog as one of our experimental benchmark because it has the affordance of correlating its communication with stock market movements of the companies which are widely discussed. It is a valid conjecture that the bloggers and commenters on Engadget are also likely the stock market investors of the companies being discussed in the blogs. Furthermore, the use of a political blog discussing the two major Presidential candidates of 2008 US elections is also justified because several of the communicators are the people who had voted in the primaries in various states. Therefore correlating this communication with media polls on the two candidates is meaningful. Note, our framework can be generalized to any social network with directly observable communication dynamics. Our framework of evaluating the characterized groups through predictivity and resilience of external variables could also be easily applied if there is another external set of data which is orthogonal in its source and content, but related in its nature and semantics.

However there are some open issues in our framework. We have focused on temporal analysis of the prototypical groups, but mining evolutionary patterns from such temporal analysis remains an open issue. For example, can we detect how a communication set behaves over time: a set of individuals who are very volatile in communication in summer, compared to winter. Besides, the posts associated with the various companies and the political candidates have been collected from Engadget and Huffington Post based on the Editor assigned tags only. Though the results validate our framework via metrics like composition entropy, topic divergence as well as graph-based measures like conductance and coverage, it would be interesting to note how the content of the posts and comments around the same time period of analysis are associated with the external events.

Nevertheless, compared to substantial previous studies on the blogosphere or on the variety of social media datasets used here (i.e. Digg, Twitter, YouTube and Flickr), this article aims at understanding the evolutionary features of social networks through a fine-grained temporal analysis of its groups. We establish a significance measure of the different groups in a set and propose how such groups

can be representative of the entire communication set. We also present a novel evaluation framework based on measure of correlation of group characteristics with an external variable, in cases where such variables are available.

7.12 Conclusions

In this chapter, we characterized online communication sets in social networks to extract their representative prototypical groups and developed a novel utility framework to establish the utility of such prototypical groups. We defined prototypical communication groups to be ones whose (a) constituent individuals are derived from several other groups and (b) whose topic distribution is highly aligned to that of the overall communication set. In order to identify such prototypical groups and their utility, we adopted a three step approach—extraction of groups, characterization of prototypical groups and utility of the prototypical groups.

First, we developed features representing communication dynamics of the individuals in a communication set. Such characteristics are then used to extract groups where each group comprises a set of individuals having similar communication characteristics. The extraction algorithm is based on unsupervised clustering and it emphasizes the temporal relationships among groups. Second, we identified a subset of all the groups as prototypical communication groups which capture and characterize the overall communication set. We presented a novel optimization framework to identify these groups. Third, we justified that extracting the prototypical groups is useful as they can enable communities to predict relevant time series data. Our conjecture is that online communities often manifest public opinion about finance, stocks, technology products as well as about different political candidates. Hence we presented a novel utility framework where we show that the prototypical groups can be useful predictors of stock market movement of technology companies as well as of political polls of elections.

We conducted extensive experiments on two popular blogs, Engadget and Huffington Post and four social media datasets, Digg, Twitter, YouTube and Flickr. We observed that during periods of significant external events related to the commu-

nities, the mean composition entropy and the mean topic divergence show significant changes. This implies that the prototypical groups are able to capture the dynamics in the communication set. We then justified the utility of the prototypical groups by using them to predict stock market movement / political polls for the blog datasets. The error in prediction were 23.42%, 22.58%, 19.67% and 20.73% respectively for communication sets discussing ‘Apple’, ‘Microsoft’, ‘Google’ and ‘Nokia’ on Engadget and 21.43% and 19.75% for the sets discussing ‘Obama’ and ‘McCain’ on Huffington Post.

Next we evaluated the characteristics of the prototypical groups by their quality i.e. conductance and coverage (on all datasets), and two novel metrics, high predictivity of stock market movement / political polls and resilience of the communication set in predicting these external variables when the prototypical groups are not considered (only blog datasets). During evaluation we observed that our method yielded high quality prototypical groups (evident via low conductance and high coverage) and also maximized predictivity of stock market movement / polls (our method: 0.73, baselines: 0.48) and minimized resilience with respect prediction (our method: 0.18, baselines: 0.46), implying that the prototypical groups extracted by our method are more meaningful in characterizing the overall communities. The mean improvement in conductance over the four sets in the Engadget dataset compared to the baseline techniques (based on alternative group extraction and prototypical group identification methods) is approximately 30%, that in coverage is 27%, that in predictivity of the polls is 45% and in resilience is 42%. In case of the Huffington Post dataset, the mean improvement in conductance over the two sets ‘Obama’ and ‘McCain’ is 41.5%, that in coverage is 56%, that in predictivity of the polls is 48% and in resilience is 40%. Similar improvements of the measures of conductance and coverage over non-prototypical groups can be made for the four social media datasets (i.e.Digg, Twitter, YouTube and Flickr) as well.

Our results are promising and there are several interesting directions to future work. One direction could be sentiment detection in communication leveraging the properties of the prototypical groups. Among other directions, we are interested

to observe how the temporal dynamics of the prototypical groups affect the discussions and the individuals they are associated with; for example, do prototypical groups indicate presence of prototypical individuals too? Also, can the discussions associated with the prototypical groups be deemed prototypical or ‘interesting’ as well? We are also interested in modeling cultural phenomenon in a community or organizing reaction with respect to an impulse or event, e.g. how traffic, road and communication in a community could be impacted at a particular point of time in response to a snowstorm. Besides, developing models of network growth, e.g. how individuals associate themselves to others based on their communication activity, as well as understanding the prediction and evolution of mood or sentiment of a community in response to external events can yield interesting insights into online communication dynamics in the Blogosphere as well as social media in general.

(a) Communication set – Apple

(b) Communication set – Microsoft

Figure 7.9: Visualization of groups (shown in circles) and the prototypical groups (shaded circles, intensity representing the measure of optimality) in the communication sets on Apple and Microsoft. Every row corresponds to the groups at each week, shown vertically downwards. We observe the dynamics of composition entropy and topic divergence of the prototypical groups during periods of significant related external events.

(c) Communication set – Google

(d) Communication set – Nokia

Figure 7.10: Visualization of groups (shown in circles) and the prototypical groups (shaded circles, intensity representing the measure of optimality) in the communication sets on Google and Nokia. Every row corresponds to the groups at each week, shown vertically downwards. We observe the dynamics of composition entropy and topic divergence of the prototypical groups during periods of significant related external events.

WEEK	APPLE					MICROSOFT					GOOGLE					NOKIA				
	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
May 19	8	0.28	0.76	0.41	0.63	7	0.55	0.32	0.68	0.20	6	0.16	0.59	0.26	0.45	3	0.35	0.69	0.45	0.58
May 26	8	0.19	0.69	0.29	0.56	6	0.33	0.69	0.44	0.58	7	0.50	0.44	0.60	0.32	7	0.49	0.32	0.63	0.17
Jun 2	10	0.43	0.46	0.54	0.32	8	0.32	0.71	0.44	0.59	5	0.25	0.66	0.37	0.52	6	0.18	0.71	0.28	0.58
Jun 9	6	0.17	0.71	0.27	0.57	5	0.36	0.67	0.49	0.53	5	0.21	0.71	0.34	0.57	5	0.58	0.36	0.71	0.23
Jun 16	7	0.48	0.38	0.58	0.24	5	0.43	0.38	0.57	0.25	5	0.53	0.39	0.67	0.25	5	0.19	0.73	0.33	0.62
Jun 23	9	0.29	0.64	0.43	0.52	8	0.23	0.63	0.37	0.51	3	0.30	0.72	0.40	0.59	5	0.53	0.44	0.67	0.29
Jun 30	6	0.22	0.74	0.35	0.60	4	0.18	0.74	0.30	0.59	5	0.19	0.80	0.31	0.69	4	0.20	0.66	0.30	0.52
Jul 7	6	0.19	0.73	0.30	0.59	7	0.65	0.40	0.75	0.27	7	0.51	0.35	0.63	0.23	4	0.50	0.48	0.62	0.36
Jul 14	6	0.55	0.44	0.69	0.33	6	0.66	0.35	0.76	0.22	3	0.23	0.77	0.33	0.62	6	0.21	0.78	0.32	0.65
Jul 21	7	0.42	0.45	0.52	0.34	5	0.19	0.73	0.30	0.58	3	0.58	0.41	0.69	0.30	3	0.18	0.68	0.32	0.55
Jul 28	8	0.26	0.81	0.38	0.68	7	0.22	0.70	0.36	0.58	5	0.16	0.81	0.26	0.66	7	0.52	0.38	0.64	0.27
Aug 4	6	0.18	0.72	0.29	0.57	6	0.19	0.75	0.30	0.61	5	0.61	0.27	0.74	0.14	3	0.17	0.69	0.31	0.57
Aug 11	6	0.38	0.46	0.51	0.34	8	0.21	0.76	0.35	0.62	6	0.17	0.65	0.28	0.50	2	0.48	0.29	0.58	0.18
Aug 18	10	0.29	0.78	0.42	0.65	4	0.26	0.80	0.37	0.68	4	0.24	0.67	0.36	0.56	5	0.60	0.36	0.71	0.25
Aug 25	11	0.23	0.79	0.33	0.67	7	0.59	0.30	0.73	0.17	4	0.29	0.68	0.39	0.55	4	0.21	0.66	0.31	0.54

Figure 7.11: Analysis of the prototypical groups against the non-prototypical ones based on two metrics—conductance and coverage. The results are shown for four communication sets—Apple, Microsoft, Google and Nokia, in the time period from May 19 to Aug 31, 2008. Five different values are shown per week and per communication set. The details of the different columns are as follows: Column A: the number of prototypical groups; Column B: the mean conductance over all the prototypical groups; Column C: the mean coverage over all the prototypical groups; Column D: the mean conductance over all the non-prototypical groups, i.e. the groups in the communication set not identified as prototypical groups; and finally, Column E: the mean coverage over all the non-prototypical groups. Note, the conductances for the prototypical groups are considerably lower than that of the non-prototypical ones; while the coverage is higher. This implies that the identified prototypical groups are of high quality and hence meaningfully model communication existing among the individuals in the communication sets.

(a) Communication set – Obama

(b) Communication set – McCain

Figure 7.12: Visualization of groups (shown in circles) and the prototypical groups (shaded circles, intensity representing the measure of optimality) in the communication sets on Obama and McCain. Every row corresponds to the groups at each week, shown vertically downwards. We observe the dynamics of composition entropy and topic divergence of the prototypical groups during periods of significant related external events.

Figure 7.13: Dynamics of composition entropy and topic divergence (shown in left Y-axis) of the prototypical groups of four datasets during periods of significant related external events, extracted from NY Times (shown as the NYT mentions trends on the right Y-axis). For each dataset, we show the dynamics of prototypical groups defined over two topical communication sets.

Part III

Rich Media Communication Patterns

“*New marketing is about the relationships, not the media.*”—Ben Grossman, founder and chief strategist for BiGMarK.

My third and final area of focus in this thesis deals with the relationship between online communication and properties of the shared rich media. In recent years we have seen a transformation in the types of content available on the web. During the first decade of the web’s prominence—from the early 1990’s onwards—most online content resembled traditional published material: the majority of web users were *consumers* of content, created by a relatively small amount of publishers. From the early 2000’s, more and more users participate in content creation, rather than just consumption. Popular user-generated content (or otherwise known as social media) domains include blogs and web forums, social bookmarking sites, photo and video sharing communities, as well as social networking platforms such as Facebook and MySpace, which offers a combination of all of these with an emphasis on the relationships among the users of the community.

An interesting property of large-scale user-generated content on social media sites is that shared media content seem to generate significant communication centered around shared media objects, e.g. videos in YouTube, images in Flickr etc. Hence apart from impact of communication on the dynamics of the individuals’ actions, roles and the community in general, there are additional challenges on how to characterize such “conversations”. These include understanding the relationship of the conversations to the community’s dynamics, as well as studying the observed user behavior responsible for participation around the shared content.

Motivating Application. Which is the best news source to read about the political developments centered around the elections in Iran?

Today, there is significant online chatter, discussion and opinions expressed via shared rich media artifacts, e.g. photos, videos etc, reflecting sentiment on political events. While different media sites can provide coverage over the same event with variable degrees of chatter, it becomes imperative to determine suitable methods

and techniques to identify media sources likely to provide “interesting” information to a certain user. Suppose Alice is interested in identifying “interesting” media sources dissipating information on public sentiments regarding the 2009 elections in Iran. To address Alice’s needs, we need to address three non-trivial problems. First, we need to be able to characterize useful chatter or conversations that emerge centered around rich media artifacts. Next, there needs to be a way to be able to model Alice’s observed participatory behavior, so that we can identify the factors impacting her decision to converse on social media. This knowledge can enable us suggest her the most suitable media channel dissipating the information on the elections in Iran.

Thus we are interested in analyzing rich media communication patterns from two different perspectives:

1. What are the characteristics of conversations centered around shared rich media artifacts? (Chapter 8)
2. How do we quantify collective participation on rich media conversations? (Chapter 9)

Chapter 8

Interestingness of Conversations

“Informal conversation is probably the oldest mechanism by which opinions on products and brands are developed, expressed, and spread.”—Johan Arndt (1555–1621), German Lutheran theologian.

The widespread proliferation of social media websites today have enabled web users to create and share rich multimedia objects exceedingly easily. This has made diverse media objects available to large audiences providing conducive platforms for critical feedback on the shared item. Note that a key property of these rich media websites is that they support rich interactions among individuals centered around the item, in the form of comments and replies. These interactions often take conversational forms, supporting considerable back and forth dialogue among the participants. What causes such a conversation to be interesting, that prompts a user to participate in the discussion on a posted video? Our central intuition in this chapter is that a conversation is perceived as interesting when its themes are engaging (including visual content of the media as well as the textual content of the comments), or when its participants are interesting in their manner of communication discourse. Importantly, a conversation that is interesting must be consequential—i.e. it must impact the social network itself.

We propose a computational framework to determine objectively a measure of conversational interestingness. We characterize conversational themes via two aspects: visual features of the associated media object, and thereafter, the textual content of the comments and replies subsuming the conversation at any point in time. Next we characterize the communication properties of the participants and the conversations via a random walk model. The interestingness measure of the

conversations is then determined via a novel joint optimization framework, that utilizes the themes, the interestingness of the participants and the past interestingness measure of the conversations. As a quantitative evaluation methodology, we measure the consequence of a conversation by determining how interestingness affects the following three variables—participation in related themes, participant cohesiveness and theme diffusion. On evaluating our proposed method on four different datasets, YouTube, Flickr, Engadget and Huffington Post, featuring conversations over multiple topical categories, we observe that our method can satisfactorily explain consequence metrics significantly better (two times) than three other baseline methods. Utilizing the interestingness measure as predictive indicators of rich media conversational characteristics (e.g. number of comments, rate of new participation) is also investigated through empirical studies, and its implications discussed henceforth.

8.1 Introduction

The emergence and growth of the social media websites, including YouTube and Flickr have created new opportunities for individuals to share rich multimedia objects online—a phenomenon that was scarcely feasible less than a decade ago. Not only does this social practice of media sharing open up multimedia browsing and content consumption to wide and diverse audiences, a key feature of such sharing process is that it facilitates extensive interaction among the individuals involved in the sharing and media consumption process. This communication process manifests itself via comments and replies centered around the media object, revealing a communication structure between users (user Alice comments on the upload, user Bob comments on the upload, Alice comments in response to Bob’s comment, Bob responds to Alice’s comment, and so forth), akin to a typical *conversational thread*. Example of a conversation from YouTube (<http://www.youtube.com/>) is shown in Fig. 8.1. In this chapter, a conversation is formally defined as the sequence of comments on a media object. Note that in the context of YouTube, the theme of a conversation is latent and depends on its content, that might be the visual features

Figure 8.1: Example of a conversation from YouTube. A conversation is associated with a unique media object comprising several temporally ordered comments on different latent themes from different authors (participants of the conversation).

of the video content, or the textual attributes in the comments.

There are two key characteristics of these “conversations”. First, note that over time, a conversational thread can often digress to themes that are marginally related, or unrelated to the original video. Such dynamics of the conversation are therefore likely to affect the characteristics of the media object over time, and even possibly, the associated social network of the interacting individuals.

Second, the communication activity in the conversations is likely to impact the semantic evolution¹ of the associated shared multimedia artifact. Hence the analysis of communication activity is crucial to understanding repeated visits of users to a rich media social networking site. Users return to a video post that they have already seen not only because the visual content of the video is captivating, but also because they intend to post and / or reply to the engaging conversation that embodies the video. To justify this claim empirically, we present some statistics in Table 8.1, that utilize the communication activity surrounding YouTube videos. We observe that on different topical categories, the mean number of unique users posting comments is much lesser in comparison to the mean number of comments

¹By semantic evolution, we imply the change in the meaning of a shared media object over time. This change is affected by user participation, as well as the dynamics of the thematic content they discuss.

Table 8.1: Statistics (a total set of 272,810 videos involving 17,736,361 unique participants and 145,682,273 comments during 2008-2009) to demonstrate returning participants to videos on YouTube. We observe that the number of unique participants posting comments is much lesser compared to the number of comments, indicating that participants return to videos because of their engagement in the conversation (also note the comment-to-author ratio). The numbers are shown across different featured topical categories on YouTube.

Topic Category	Mean #comments	Mean #authors	Ratio
“COMEDY”	288	124	2.32
“EDUCATION”	193	81	10.72
“ENTERTAINMENT”	372	105	3.54
“NEWS & POLITICS”	682	428	1.59
“SPORTS”	504	258	1.95
“MUSIC”	126	62	2.03

on the videos. Note that for categories such as “Music” the returning visitors might be the consequence of the music video being popular or preferred as shown by the higher number of unique comments authored. However for categories as “News & Politics” there is clear evidence of the conversation being central in increasing the engagement of participants to return to the videos. Thus it is the content of the communication activity itself that the people want to read (or see, if the response to a video post is another video, as is possible in the case of YouTube). Furthermore, these rich media sites as YouTube have notification mechanisms that alert users of new comments on a video post / image upload promoting this communication activity.

We denote the communication property that causes thematic evolution of discourse as well as invites people to further participate in a conversation as its “interestingness.” While the meaning of the term “interestingness” is subjective, we decided to use it to express an intuitive property of the communication phenomena that we frequently observe on rich media networks. Our goal is to determine a real scalar value corresponding to each conversation in an objective manner that serves as a measure of interestingness².

What causes a conversation to be interesting to prompt a user to participate?

²Modeling the user subjectivity is beyond the scope of our work; refer to section 8.4 for further elaboration.

We conjecture that people will participate in conversations when (a) they find the conversation theme interesting (what the previous users are talking about) (b) see comments by people that are well known in the community, or people that they know directly comment (these people are interesting to the user) or (c) observe an engaging dialogue between two or more people (an absorbing back and forth between two people). Intuitively, interesting conversations have an engaging theme, with interesting people.

However note that the content of the media itself (e.g. the visual attributes of a YouTube video) might also play an important role in triggering repeated user participation over time. Hence in this chapter, our notion of a conversation themes subsumes two aspects—not only the textual content of the comments in the conversation, but the visual content of the video as well.

A major challenge in this work is that there is no available ground truth to evaluate the interestingness measure of conversations. Hence we propose to study the impact of interestingness on future temporal events. We note that a conversation that is deemed interesting must be consequential—i.e. it must impact the social network itself. Intuitively, there should be three consequences (a) the people who find themselves in an interesting conversation, should tend to co-participate in future conversations (i.e. they will seek out other interesting people that they’ve engaged with) (b) people who participated in the current interesting conversation are likely to seek out other conversations with themes similar to the current conversation and (c) the conversation theme, if engaging, should slowly proliferate to other conversations.

There are several reasons why measuring interestingness of a conversation is of value. Besides understanding the semantic evolution of media artifacts, the interestingness measure can be useful in a variety of practical applications. First, it can be used to rank and filter both blog posts and rich media, particularly when there are multiple sites on which the same media content is posted, guiding users to the most interesting conversation. For example, the same news story may be posted on several blogs, our measures can be used to identify those sites where the postings and commentary is of greatest interest. It can also be used to increase efficiency.

Rich media sites, can manage resources based on changing interestingness measures (e.g. and cache those videos that are becoming more interesting), and optimize retrieval for the dominant themes of the conversations. Additionally, capturing the time periods during which conversations become interesting can help predict spikes and chatter in online communities. Besides, differentiated advertising prices for ads placed alongside videos can be based on their associated conversational interestingness.

It is important to note that frequency based measures of a video (e.g. number of views, number of comments and number of times it has been marked by a user as a favorite) do not adequately capture interestingness because these measures are properties of the video (content, video quality), not the communication. Furthermore, the textual analysis of comments alone are not adequate to capture conversational interestingness because it does not consider the dialogue structure amongst users in the conversation, or the dynamics of the themes that embody the conversational process over time.

Our Approach

In this chapter, we determine the interestingness measure of conversations around media objects operationally. It involved characterizing conversational themes of the media object, characterizing the communication of the individuals associated with the conversation and finally evaluating the consequence of the interestingness measure on the larger social network. There are three key contributions in this chapter:

1. We characterize conversational themes for determining the “interestingness” of online conversations. These themes could be visual themes (content features of the associated media object), or could be textual themes (topical assignment based on the comment contents). We detect visual themes via content features of the associated media object, followed by detection of conversational textual themes using a sophisticated mixture model approach.

2. We characterize the communication properties of participants of conversations, based on a recurrence temporal relation. Thereafter we also propose a novel joint optimization framework with temporal smoothness constraints, that utilizes the themes and the communication properties of participants to effectively compute interestingness.
3. Finally, we measure the consequence of conversational interestingness using a mutual information based metric. We compute the mutual information between the interestingness with consequence-based measures: activity, cohesiveness and thematic interestingness.

To evaluate our proposed framework, we have conducted extensive experiments using datasets from two rich media websites, YouTube and Flickr, and two blog forums, Engadget and Huffington Post³. We observe from the dynamics of conversational themes, interestingness of participants and of conversations that (a) conversational themes associated with significant external happenings become popular (or strong), (b) the mean interestingness of conversations increase due to the chatter about important external events. During evaluation, we observe that our method of interestingness maximizes the mutual information by explaining the consequences significantly better than three other baseline methods (our method 0.83, baselines 0.41). Finally, a set of predictive experiments to regress variables such as the number of comments or rate of new participants arriving at a conversation over time, based on the conversational interestingness variable, indicate that the interestingness measure of conversations can actually be an effective indicator of these variables, in comparison to more obvious methods as their respective auto-regression models.

Related Work

Analyzing dynamic properties of rich multimedia, such as conversational interestingness associated with media objects involves understanding of constructs from a

³The datasets are available for download at: <http://www.public.asu.edu/~mdechoud/datasets.html>

variety of domains. Hence in the following paragraphs we discuss prior work from the following three facets useful in solving this problem.

Analysis of Media Properties. There has been considerable work conducted in analyzing dynamic media properties, (e.g. associated tags on a media object). In (Dubinko et al., 2006) Dubinko et al. visualized the evolution of tags within Flickr and presented a novel approach based on a characterization of the most salient tags associated with a sliding interval of time. Kennedy et al. in (Kennedy and Naaman, 2008) leveraged the community-contributed collections of rich media (Flickr) to automatically generate representative views of landmarks. Their work suggests that community-generated media and tags can improve access to multimedia resources. In (Smith et al., 2008), Smith et al. explored the search methodologies of rich social media content by utilizing the social context of users—that includes both their personal social context (their friends and the communities to which they belong) and their community social context (their role and identity in different communities). Finally Singh et al. (Singh et al., 2009) analyzed and modeled user contributions in social media sites to study its associated dynamics. Their model was based on the idea of users as rational selfish agents, and considered domain attributes like voluntary participation, virtual reward structure and public-sharing to model the dynamics of this interaction.

Theme Extraction. Extraction of themes or topics from dynamic web collections via topic modeling is a common problem of interest to researchers (Mei and Zhai, 2005), (Mei et al., 2006), (Zhou et al., 2008), (Ling et al., 2008). “Topic models” (e.g. (Blei et al., 2003)) attempt to discover patterns in text corpus by representing the underlying topics with word distributions and the topics are combined to form documents. In addition, dynamic topic models (Blei and Lafferty, 2006) have been developed to capture the evolution of topics in a sequentially organized corpus of documents. In (Mei and Zhai, 2005) the authors study the problem of discovering and summarizing evolutionary theme patterns in a dynamic text stream. The authors modify temporal theme extraction in (Mei et al., 2006) by regularizing their

theme model with timestamp and location information.

In other work, authors in (Zhou et al., 2008) propose a dynamic probability model, which can predict the tendency of topic discussions on online social networks. Nallapati et al. (Nallapati et al., 2008) addressed the problem of joint modeling of text and citations in the topic modeling framework; with the help of two different models that were variants of the LDA model. Recently, Iwata et al. (Iwata et al., 2010) developed an online topic model for sequentially analyzing the time evolution of topics in document collections.

Social Media Communication Analysis. A significant amount of work on social networks and media so far have dealt with analyzing discussions or comments in blogs (Kaltenbrunner et al., 2007), (Gómez et al., 2008), (Mishne and Glance, 2006) as well as utilizing such communication for prediction of its consequences like user behavior, sales, stock market activity etc (Adar et al., 2007), (De Choudhury et al., 2008b), (Gruhl et al., 2005), (Liu et al., 2007). Antweiler et al. in (Antweiler and Frank, 2004) determine correlations between communication activity in Internet message boards and stock volatility and trading volume. Gruhl et al in (Gruhl et al., 2005) correlate postings in blogs, media, and web pages with the sales ranks of books on Amazon.com. They devise carefully hand-crafted queries to find matching posts which can be indicators of future sales rank of books. Similarly, in (De Choudhury et al., 2008b), we analyzed the communication dynamics (of conversations) in a technology blog and used it to predict stock market movement.

Another line of work in this domain deals with topical and structural analysis of commentary on social websites. Gomez et al in (Gómez et al., 2008) analyze several social network properties emerging from the communication activity on the website Slashdot. They study the structure of discussion threads using a radial tree representation. The findings show that nesting of conversations on Slashdot exhibit strong heterogeneity and self-similarity. Benevenuto et al in (Benevenuto et al., 2008) present a method for understanding user behavior in a social network created essentially by video interactions by characterizing different properties of the

popular social media YouTube. More recently, Shi et al. (Shi et al., 2009) studied the patterns of user participation behavior manifested through various interactions, and the feature factors that influence such behavior on different forum datasets.

Besides, a line of research also focuses on computed-mediated communication, as manifested via dialogue based discourse on rich media social networks. Honeycutt et al. in (Honeycutt and Herring, 2009) investigated the conversationality of Twitter, with special attention to the role played by the @ sign. Further, Naaman et al. (Naaman et al., 2010) performed an analysis of the content of messages posted by individuals on Twitter, to categorize individuals as those who focus on the “self”, while others who are driven more by sharing information.

Our unique contribution. Note that neither of these work captures the dynamics of the associated conversations (or the broader communication activity in general) on a media object. Additionally, the relationship of theme extraction with the co-participation behavior of the authors of comments (participants) has also not been analyzed; nor has the relationship or impact of a certain conversation property with respect to other attributes of the media object been considered. In this work, we characterize the consequences of conversations based on the impact of the themes and the communication properties of the participants, where we consider both static visual feature-based themes, as well as textual themes based on the content of the conversations.

8.2 Problem Formulation

We begin by discussing our problem formulation—definitions, data model and the problem statement.

Definitions

The main definitions used in the rest of this chapter are introduced as below:

1. *Conversation.* We define a conversation in online social media (e.g., an image, a video or a blog post) as a temporally ordered sequence of comments posted

by individuals whom we call “participants”. In this chapter, the content of the conversations are represented as a stemmed and stop-word eliminated bag-of-words.

2. *Conversational Themes.* A conversational theme is a set of salient topics associated with a conversation at different points in time. Note there are two kinds of themes defined in this chapter, that are assumed to characterize a conversation: visual themes and textual themes. Visual themes comprise of different video features that describe the content of the media object associated with the conversation (here a video is represented through image frames). Textual themes are topics based on the content of the comments appearing in a conversation over time.
3. *Interestingness of Participants.* Interestingness of a participant is a property of her communication activity over different conversations. We propose that an interesting participant can often be characterized by (a) several other participants writing comments after her, (b) participation in a conversation involving other interesting participants, and (c) active participation in “hot” conversational themes.
4. *Interestingness of Conversations.* We define “interestingness” as a dynamic property of user-user communication which is represented as a real non-negative scalar. The interestingness measure is dependent on the following aspects: the evolutionary conversational themes at a particular point of time, and the communication properties of its participants. It is important to note here that “interestingness” of a conversation is necessarily subjective and often depends upon context of the participant⁴.

Conversations used in this chapter are the temporal sequence of comments associated with media elements (videos) in the highly popular media sharing site

⁴We acknowledge that alternate definitions of interestingness are also possible. For example, interestingness being a subjective notion, an interesting conversation to an individual might imply communication from people who s/he is familiar with.

YouTube. However our model can be generalized to any domain with observable threaded communication around a visual media object.

Data Model

Our data model comprises the tuple $\langle m, C, P \rangle$ having the following three inter-related entities: a set of media objects, a set of conversations, C on shared media elements; and a set of participants P in these conversations. Each conversation is represented with a set of comments, such that each comment that belongs to a conversation is associated with a unique participant, a timestamp and some textual content (bag-of-words). A conversation is also associated with a media object, e.g. a video in the context of YouTube, that can be described by a set of visual content-based features, that comprise visual themes.

We now discuss the notations. We assume that there are N participants, M conversations, L visual themes, K conversational textual themes and Q time slices. Using the relationship between the entities in the tuple C, P from the above data model, we construct the following matrices for every time slice $q, 1 \leq q \leq Q$:

- $\mathbf{P_F}^{(q)} \in \mathbb{R}^{N \times N}$: Participant-follower⁵ matrix, where $\mathbf{P_F}^{(q)}(i, j)$ is the probability that at time slice q , participant j comments following participant i on the conversations in which i had commented at any time slice from 1 to $(q - 1)$. Note that this is a computed over all conversations.
- $\mathbf{P_L}^{(q)} \in \mathbb{R}^{N \times N}$: Participant-leader⁶ matrix, where $\mathbf{P_L}^{(q)}(i, j)$ is the probability that in time slice q , participant i comments following participant j on the conversations in which j had commented in any time slice from 1 to $(q - 1)$. Note, both $\mathbf{P_F}^{(q)}$ and $\mathbf{P_L}^{(q)}$ are asymmetric, since communication between participants is directional. Like before, it is a computed over all conversations.

⁵An individual is a follower in a conversation with respect to another individual, if his/her comment appears *after* the latter's comment.

⁶An individual is a leader in a conversation with respect to another individual, if his/her comment appears *prior to* the latter's comment.

- $\mathbf{P_C}^{(q)} \in \mathbb{R}^{N \times M}$: Participant-conversation matrix, where $\mathbf{P_C}^{(q)}(i, j)$ is the probability that participant i comments on conversation j in time slice q .
- $\mathbf{C_V}^{(q)} \in \mathbb{R}^{M \times L}$: Conversation-visual theme matrix, where $\mathbf{C_V}^{(q)}(i, j)$ is the probability that conversation i has the visual feature⁷ j in time slice q .
- $\mathbf{C_T}^{(q)} \in \mathbb{R}^{M \times K}$: Conversation-textual theme matrix, where $\mathbf{C_T}^{(q)}(i, j)$ is the probability that conversation i belongs to theme j in time slice q .
- $\mathbf{V_S}^{(q)} \in \mathbb{R}^{L \times 1}$: Visual theme-strength vector, where $\mathbf{V_S}^{(q)}(i)$ is the strength of visual theme i in time slice q . Note, $\mathbf{V_S}^{(q)}$ is simply the normalized column sum of $\mathbf{C_V}^{(q)}$.
- $\mathbf{T_S}^{(q)} \in \mathbb{R}^{K \times 1}$: Textual theme-strength vector, where $\mathbf{T_S}^{(q)}(i)$ is the strength of textual theme i in time slice q . Note, $\mathbf{T_S}^{(q)}$ is simply the normalized column sum of $\mathbf{C_T}^{(q)}$.
- $\mathbf{P_V}^{(q)} \in \mathbb{R}^{N \times L}$: Participant-visual theme matrix, where $\mathbf{P_V}^{(q)}(i, j)$ is the probability that participant i 's interest in visual theme j in time slice q . Note, $\mathbf{P_V}^{(q)} = \mathbf{P_C}^{(q)} \cdot \mathbf{C_V}^{(q)}$.
- $\mathbf{P_T}^{(q)} \in \mathbb{R}^{N \times K}$: Participant-textual theme matrix, where $\mathbf{P_T}^{(q)}(i, j)$ is the probability that participant i communicates on textual theme j in time slice q . Note, $\mathbf{P_T}^{(q)} = \mathbf{P_C}^{(q)} \cdot \mathbf{C_T}^{(q)}$.
- $\mathbf{I_P}^{(q)} \in \mathbb{R}^{N \times 1}$: Interestingness of participants vector, where $\mathbf{I_P}^{(q)}(i)$ is the interestingness of participant i in time slice q .
- $\mathbf{I_C}^{(q)} \in \mathbb{R}^{M \times 1}$: Interestingness of conversations vector, where $\mathbf{I_C}^{(q)}(i)$ is the interestingness of conversation i in time slice q .

For simplicity of notation, we denote the i -th row of the above 2-dimensional matrices as $\mathbf{X}(i, :)$.

⁷The visual feature extraction method is discussed in the following section.

Problem Statement

Now we formally present our problem statement: given a dataset C, P and associated meta-data, we intend to determine the interestingness of the conversations in C , defined as $\mathbf{I}_C^{(q)}$ (a non-negative scalar measure for a conversation) for every time slice $q, 1 \leq q \leq Q$. Determining interestingness of conversations involves three key challenges:

1. How to extract the evolutionary conversational themes—visual and textual?
2. How to model the communication properties of the participants?
3. Further in order to justify interestingness of conversations, we need to address the following challenge: what are the consequences of an interesting conversation?

In the following four sections, we discuss how we address these challenges through: (a) define visual themes of the media object associated with a conversation via different standard image features (section 4); (b) detecting conversational textual themes based on a mixture model that incorporates regularization with time indicator, regularization for temporal smoothness and for co-participation (section 5); (c) modeling interestingness of participants; and thereafter modeling interestingness of conversations; along with using a novel joint optimization framework of interestingness that incorporates temporal smoothness constraints (section 6); and finally (d) defining several consequence based metrics to evaluate the impact of an interesting conversation (section 7).

8.3 Visual Themes from Media

A key motivation for large-scale participation of users on social media conversations arises due to their interest in the content of the associated media object. Hence in this section we discuss a number of features that characterize the visual content of the media. We have chosen features that have been found effective in

image content analysis⁸. These include features on color (color histogram and color moments), texture (GLCM and phase symmetry), shape (radial symmetry and phase congruency) and interest points (SIFT).

We briefly discuss different content based features that have been used to characterize the key frames associated with videos of different conversations.

Color. There are two color-based features of interest—color histogram and color moments, which capture the distribution of different colors in the images.

Texture. YouTube videos show diversity of texture. We use two texture features: gray level co-occurrence matrix, GLCM, and a texture detector for arbitrary “blobs” in images—called phase symmetry (Xiao et al., 2005). Phase symmetry is based on determining local symmetry and asymmetry across an image from phase information. Given an image, phase-based symmetry detector (PSD) maps a pixel, p , an orientation, o , and a scale, n , to a phase congruency value $PC_{n,o}(p)$ and a special phase, $\phi_{n,o}(p)$:

$$PSD(p, n, o) = (PC_{n,o}(p) \cdot \phi_{n,o}(p)). \quad (8.1)$$

Here,

$$PC_{n,o}(p) = \frac{sumE(p)}{sumA(p) + \varepsilon} = \frac{\sum_{k,q} E_{n_k,o_q}(p)}{\sum_{k,q} A_{n_k,o_q}(p) + \varepsilon}, \quad (8.2)$$

where $sumE(p)$ is the total energy when phases are congruent under all scales and orientations and phases are zero; $sumA(p)$ is the total amplitude when phases are congruent under all scales and orientations and phases are zero; ε is a positive con-

⁸In order to extract visual features from videos in our context i.e. YouTube, we have chosen the first frame of the video as the ‘key’ frame and then performed image analysis on it. Hence for a conversation c_i , that has an associated media element, we index its key frame as the image κ_i . This assumption of the choice of the key frame is reasonable, because the dataset of YouTube videos utilized in this work are typical of short duration (~ 5 – 10 minutes). However we acknowledge that a limitation of this strategy is that we cannot capture the movement semantics of the video.

stant.

Shape. Frames in a video are often characterized by central themes or concepts. To extract such features, we use two shape features—radial symmetry (Loy and Zelinsky, 2003) and phase congruency (Liu and Laganière, 2007). The radial symmetry feature is based on the idea of detecting points of interest in an image. Phase congruency is an illumination and contrast invariant measure of feature significance. For a given image, phase congruency $PC(x)$ at some location x is expressed as the summation over orientation o and scale n :

$$PC(x) = \frac{\sum_o \sum_n W_o(x) [A_{n,o}(x) \Delta \Phi_{n,o}(x) - T_o]}{\sum_o \sum_n A_{n,o}(x) + \varepsilon}, \quad (8.3)$$

where A_n represents the amplitude of the n -th component (of the image) in the Fourier series expansion, $W_o(x)$ is the convolution of the given image with an even / odd filter, ε is added for cases of small Fourier amplitudes, T_o is a compensating measure for the influence of noise and $\Delta \Phi_n(x)$ is a sensitive phase deviation.

SIFT. SIFT or Scale Invariant Feature Transform (Lowe, 2004) is a content-based image feature that detects stable keypoint locations in scale space of an image (in our case, a key frame in a video). It computes the following function from the difference of two nearby scales separated by a constant multiplicative factor κ :

$$D(x, y, \sigma) = (G(x, y, \kappa\sigma) - G(x, y, \sigma)) \cdot I(x, y) = L(x, y, \kappa\sigma) - L(x, y, \sigma), \quad (8.4)$$

where $L(x, y, \sigma)$ is the scale space of image $I(x, y)$ and is produced from the convolution of a variable-scale Gaussian having scale (σ) with $I(x, y)$.

The details of these features may be referred to in (Lowe, 2004), (Xiao et al., 2005), (Liu and Laganière, 2007), (Loy and Zelinsky, 2003).

After extracting these features, we now construct a L -dimensional feature vector for each key frame associated with a conversation c_i , where $L = 1064$ in this work. This gives us the matrix $\mathbf{C}_V^{(q)}$ associated with conversations at time slice q .

Note from $\mathbf{C}_V^{(q)}$, we can easily construct the visual theme-strength vector, given as, $\mathbf{V}_S^{(q)}$.

8.4 Textual Themes from Media

In this section, we discuss the method of detecting textual themes in conversations. We elaborate on our theme model in the following two parts—first a sophisticated mixture model for textual theme detection incorporating time indicator based, temporal and co-participation based regularization is presented. Next, we discuss parameter estimation of this theme model.

Segmentation-based Mixture Model of Themes. Conversations are dynamically growing collections of comments from different participants. Hence, static keyword or tag based assignment of themes to conversations independent of time is not useful. Our model of detecting themes is therefore based on segmentation of conversations into ‘chunks’ per time slice. A chunk is a representation of a conversation at a particular time slice and it comprises a (stemmed and stop-word eliminated) set of comments (bag-of-words) whose posting timestamps lie within the same time slice. Our goal is to associate each chunk (and hence the conversation at that time slice) with a theme distribution. We develop a sophisticated multinomial mixture model representation of chunks over different themes (a regularized pLSA (Hofmann, 1999)) where the theme distributions are (a) regularized with time indicator, (b) smoothed across consecutive time slices, and (c) take into account the prior knowledge of co-participation of individuals in the associated conversations.

Let us assume that there are N conversations. A conversation c_i is segmented into Q non-overlapping chunks (or bag-of-words) corresponding to the Q different time slices. Let us represent the chunk corresponding to the i -th conversation at time slice q ($1 \leq q \leq Q$) as $\lambda_{i,q}$. We further assume that the words in $\lambda_{i,q}$ are generated from K multinomial theme models $\theta_1, \theta_2, \dots, \theta_K$ whose distributions are hidden to us. Our goal is to determine the log likelihood over the collection, while incorporating the three regularization techniques mentioned above. Thereafter we can maximize the

log likelihood to compute the parameters of the K theme models.

However, before we estimate the parameter of the theme models, we refine our framework by regularizing the themes temporally (time regularization of the theme models) as well as due to co-participation of participants. This is discussed in the following two sub-sections.

Temporal Regularization

We incorporate temporal characterization of themes in our theme model (Mei et al., 2006). We conjecture that a word in the chunk can be attributed either to the textual context of the chunk $\lambda_{i,q}$ (i.e. the term frequency distribution of the comments in the conversation), or the time slice q —for example, certain words can be highly popular on certain time slices due to related external events. Hence the theme associated with words in a chunk $\lambda_{i,q}$ needs to be regularized with respect to the time slice q . We represent the chunk $\lambda_{i,q}$ at time slice q with the probabilistic mixture model:

$$p(w : \lambda_{i,q}, q) = \sum_{j=1}^K p(w, \theta_j | \lambda_{i,q}, q) \quad (8.5)$$

where w is a word in the chunk $\lambda_{i,q}$ and θ_j is the j -th theme. The joint probability on the right hand side can be decomposed as:

$$p(w, \theta_j | \lambda_{i,q}, q) = p(w | \theta_j) \cdot p(\theta_j | \lambda_{i,q}, q) = p(w | \theta_j) \cdot ((1 - \gamma_q) \cdot p(\theta_j | \lambda_{i,q}) + \gamma_q \cdot p(\theta_j | q)), \quad (8.6)$$

where γ_q is a parameter that regulates the probability of a theme θ_j given the chunk $\lambda_{i,q}$ and the probability of a theme θ_j given the time slice q . Note that since a conversation can alternatively be represented as a set of chunks, the collection of all chunks over all conversations is simply the set of conversations C . Hence the log likelihood of the entire collection of chunks is equivalent to the likelihood of the M conversations in the set C , given the theme model. Weighting the log likelihood of the model parameters with the occurrence of different words in a chunk, we get the following equation:

$$L(C) = \log p(C) = \sum_{\lambda_{i,q} \in C} \sum_{w \in \lambda_{i,q}} n(w, \lambda_{i,q}) \cdot \log \sum_{j=1}^K p(w, \theta_j | \lambda_{i,q}, q), \quad (8.7)$$

where $n(w, \lambda_{i,q})$ is the count of the word w in the chunk $\lambda_{i,q}$ and $p(w, \theta_j | \lambda_{i,q}, q)$ is given by eqn. 8.6. However, the theme distributions of two chunks within a conversation across two consecutive time slices should not too different from each other. That is, they need to exhibit temporal smoothness. For a particular topic θ_j this smoothness is thus based on minimization of the following $L-2$ -norm distance between its probabilities across every two consecutive time slices:

$$d_T(j) = \sum_{q=2}^Q (p(\theta_j | q) - p(\theta_j | q-1))^2. \quad (8.8)$$

Incorporating this distance in eqn. 8.7 we get a new log likelihood function which smoothes all the K theme distributions across consecutive time slices:

$$L_1(C) = \sum_{\lambda_{i,q} \in C} \sum_{w \in \lambda_{i,q}} n(w, \lambda_{i,q}) \cdot \log \sum_{j=1}^K (p(w, \theta_j | \lambda_{i,q}, q) + \exp(-d_T(j))). \quad (8.9)$$

Now we discuss how this theme model is further regularized to incorporate prior knowledge about co-participation of individuals in the conversations.

Co-participation based Regularization

Our intuition behind co-participation regularization is based on the idea that if several participants comment on a pair of chunks, then the theme distributions of the chunks are likely to be closer to each other⁹.

To recall, chunks are representations of conversations at a particular time slice. We define a participant co-occurrence weighted graph $G(C, E; \omega)$ where each vertex in C is a conversation c_i and an undirected edge $e_{i,m}$ exists between two conversations c_i and c_m if they share at least one common participant. The edges are also associated with weights $\omega_{i,m}$ which is the fraction of common participants between two conversations. We incorporate participant-based regularization based

⁹If two individuals tend to co-comment around the same time on two conversations, its likely that they are talking about similar themes.

on this graph by minimizing the distance between the edge weights of two adjacent conversations with respect to their corresponding theme distributions.

The following regularization function ensures that the theme distribution functions of conversations are very close to each other if the edge between them in the participant co-occurrence graph G has a high weight:

$$R(C) = \sum_{c_i, c_m \in C} \sum_{j=1}^K (\omega_{i,m} - (1 - (f(\theta_j|c_i) - f(\theta_j|c_m))^2))^2, \quad (8.10)$$

where $f(\theta_j|c_i)$ is defined as a function of the theme θ_j given the conversation c_i and the L^2 distance between $f(\theta_j|c_i)$ and $f(\theta_j|c_m)$ ensures that the theme distributions of adjacent conversations are similar. Since a conversation is associated with multiple chunks, thus $f(\theta_j|c_i)$ is given as in (Mei et al., 2008):

$$f(\theta_j|c_i) = p(\theta_j|c_i) = \sum_{\lambda_{i,q} \in c_i} p(\theta_j|\lambda_{i,q}) \cdot p(\lambda_{i,q}|c_i). \quad (8.11)$$

Now, using eqn. 8.9 and eqn. 8.10, we define the final combined optimization function which minimizes the negative of the log likelihood and also minimizes the distance between theme distributions with respect to the edge weights in the participant co-occurrence graph:

$$O(C) = -(1 - \varsigma) \cdot L_1(C) + \varsigma \cdot R(C), \quad (8.12)$$

where the parameter ς controls the balance between the likelihood using the multinomial theme model and the smoothness of theme distributions over the participant graph. It is easy to note that when $\varsigma = 0$, then the objective function is the temporally regularized log likelihood as in eqn. 8.9. When $\varsigma = 1$, then the objective function yields themes which are smoothed over the participant co-occurrence graph.

Parameter Estimation. Minimizing $O(C)$ in eqn. 8.12 for $0 \leq \varsigma \leq 1$ would give us the theme models that best fit the collection. We discuss the parameter estimation in the following paragraphs.

We discuss how we can learn the hidden parameters of the theme model in eqn. 8.12. Note, the use of the more common technique of parameter estimation with

the EM algorithm in our case involves multiple computationally intensive iterations due to the existence of the regularization function in eqn. 8.12. Hence we use a different technique of parameter estimation based on the Generalized Expectation Maximization algorithm (GEM (Mei et al., 2008)).

The update equations for the E and M steps in estimation of the theme model parameters are illustrated as follows. Specifically, in the E-step, we first compute the expectation of the complete likelihood $\Theta(\Psi; \Psi^{(m)})$, where Ψ denotes all the unknown parameters and $\Psi^{(m)}$ denotes the value of Ψ estimated in the m -th EM iteration. In the M-step, the algorithm finds a better value of Ψ to ensure that $\Theta(\Psi^{(m+1)}; \Psi^{(m)}) \geq \Theta(\Psi^{(m)}; \Psi^{(m)})$. First we empirically fix the free transition parameters involved in the log likelihood in eqn. 8.12: γ_q to be 0.5 for all q and ς as well to be 0.5. For the E-step, we define a hidden variable $z(w, \gamma_{i,q}, j)$. Formally we have the **E-step**:

$$z(w, \gamma_{i,q}, j) = \frac{p^{(m)}(w|\theta_j)((1-\gamma_q)p^{(m)}(\theta_j|\lambda_{i,q}) + \gamma_q p^{(m)}(\theta_j|q))}{\sum_{j'=1}^K p^{(m)}(w|\theta_{j'})((1-\gamma_q)p^{(m)}(\theta_{j'}|\lambda_{i,q}) + \gamma_q p^{(m)}(\theta_{j'}|q))}. \quad (8.13)$$

Now we discuss the **M-step**:

$$\begin{aligned} \Theta(\Psi; \Psi^{(m)}) = & (1-\varsigma) \cdot \sum_{\lambda_{i,q} \in C} \sum_{w \in \lambda_{i,q}} n(w, \lambda_{i,q}) \cdot \sum_{j=1}^K z(w, \gamma_{i,q}, j) \cdot a_j + L_\lambda + L_q + L_j \\ & - \varsigma \cdot \sum_{c_i, c_m \in C} \sum_{j=1}^K (\omega_{i,m} - (1 - (f(\theta_j|c_i) - f(\theta_j|c_m))^2))^2, \end{aligned} \quad (8.14)$$

where $L_\lambda = \alpha_\lambda(\sum_j p(\theta_j|\lambda_{i,q}) - 1)$, $L_q = \alpha_q(\sum_j p(\theta_j|q) - 1)$ and $L_j = \alpha_j(\sum_w p(w|\theta_j) - 1)$ are the Lagrange multipliers corresponding to the constraints that $\sum_j p(\theta_j|\lambda_{i,q}) = 1$, $\sum_j p(\theta_j|q) = 1$ and $\sum_w p(w|\theta_j) = 1$. Based on several iterations of E and M-steps, GEM estimates locally optimum parameters of the K theme models. Details of convergence of this algorithm can be referred in (Mei et al., 2008).

With the learnt parameters of the theme models, we can now compute the probability that a chunk $\lambda_{i,q}$ belongs to a theme θ_j :

$$\begin{aligned}
p(\theta_j|\lambda_{i,q},q) &= \sum_w p(\theta_j|w)((1-\gamma_q) \cdot p(w|\lambda_{i,q}) + \gamma_q \cdot p(w|q)) \\
&= \sum_w \frac{p(w|\theta_j) \cdot p(\theta_j)}{p(w)} \cdot (1-\gamma_q) \cdot p(w|\lambda_{i,q}) + \gamma_q \cdot p(w|q).
\end{aligned} \tag{8.15}$$

All the parameters on the right hand side are known from parameter estimation. A chunk $\lambda_{i,q}$ being the representation of a conversation c_i at a time slice q , the above equation would give us the conversation-theme matrix $\mathbf{C_T}$ at every time slice $q, 1 \leq q \leq Q$.

Now, we discuss how the evolutionary conversational themes can be used to determine interestingness measures of participants and conversations.

8.5 Interestingness

In this section we describe our interestingness models and then discuss a method that jointly optimizes the two types of interestingness incorporating temporal smoothness.

Interestingness of Participants

We pose the problem of determining the interestingness of a participant at a certain time slice as a simple one-dimensional random walk model where she communicates either based on her past history of communication behavior in the previous time slice, or relies on her independent desire of preference over different themes (random jump). We describe these two states of the random walk through a set of variables as follows.

We conjecture that the state signifying the past history of communication behavior of a participant i at a certain time slice q , denoted as $\mathbf{A}(q-1)$ comprises the variables: (a) whether she was interesting in the previous time slice, $\mathbf{I_P}^{(q-1)}(i)$, (b) whether her comments in the past impacted other participants to communicate and their interestingness measures, $\mathbf{P_F}^{(q-1)}(i,:) \cdot \mathbf{I_P}^{(q-1)10}$, (c) whether she followed several interesting people in conversations at the previous time slice $q-1$, $\mathbf{P_L}^{(q-1)}(i,:)$

¹⁰To recall, $\mathbf{X}(i,:)$ is the i -th row of the 2-dimensional matrix $\mathbf{X}$.

Figure 8.2: Timing diagrams of the random walk models for computing interestingness of participants ($I_P^{(q-1)}$) and of conversations ($I_C^{(q-1)}$). The relationships between different variables affecting the two kinds of interestingness are shown.

) $\cdot I_P^{(q-1)}$, and (d) whether the conversations in which she participated became interesting in the previous time slice $q-1$, $P_C^{(q-1)}(i;:) \cdot I_C^{(q-1)}$. The independent desire of a participant i to communicate is dependent on her theme distribution and the strength of the themes at the previous time slice $q-1$: either due to the visual content of the themes, given as, $P_V^{(q-1)} \cdot V_S^{(q-1)}$, or due to the textual content of the themes, as given by $P_T^{(q-1)}(i, :) \cdot T_S^{(q-1)}$. The relationships between all these different variables involving the two states are shown in Figure 8.2.

Thus the recurrence relation for the random walk model to determine the interestingness of all participants at time slice q is given as:

$$I_P^{(q)} = (1 - \beta) \cdot A^{(q-1)} + \beta \cdot (\gamma_1 \cdot P_V^{(q-1)} \cdot V_S^{(q-1)} + \gamma_2 \cdot P_T^{(q-1)} \cdot T_S^{(q-1)}), \quad (8.16)$$

where,

$$A^{(q-1)} = \alpha_1 \cdot P_L^{(q-1)} \cdot I_P^{(q-1)} + \alpha_2 \cdot P_F^{(q-1)} \cdot I_P^{(q-1)} + \alpha_3 \cdot P_C^{(q-1)} \cdot I_C^{(q-1)}. \quad (8.17)$$

Here α_1 , α_2 and α_3 are weights that determine mutual relationship between the variables of the past history of communication state $A^{(q-1)}$, similarly, γ_1 and γ_2 are the weights determining the impact of the visual and the textual content of

the themes associated with the conversations, and β the transition parameter of the random walk that balances the impact of past history and the random jump state involving participant's independent desire to communicate. The value of β can be tuned to determine the impact of the two states on the interestingness measure of a participant.

Interestingness of Conversations

Similar to interestingness of participants, we pose the problem of determining the interestingness of a conversation as a random walk where a conversation can become interesting based on two states: the first state is when participants make the conversation interesting, and the second state is when themes make a conversation interesting (random jump). Hence to determine the interestingness of a conversation i at time slice q , we conjecture that it depends on whether the participants in conversation i became interesting at $q-1$, given as, $\mathbf{P_C}^{(q-1)}(i, :)^t \cdot \mathbf{I_P}^{(q-1)}$, or whether the conversations belonging to the strong themes in $q-1$ became interesting, which is given as, $\text{diag}(\mathbf{C_V}^{(q-1)}(i, :) \cdot \mathbf{V_S}^{(q-1)}) \cdot \mathbf{I_C}^{(q-1)}$ corresponding to the visual content of the themes associated with the conversations, or, $\text{diag}(\mathbf{C_T}^{(q-1)}(i, :) \cdot \mathbf{T_S}^{(q-1)}) \cdot \mathbf{I_C}^{(q-1)}$, which are due to the textual content of the themes. Thus the recurrence relation of interestingness of all conversations at time slice q is:

$$\begin{aligned} \mathbf{I_C}^{(q)} = & \psi \cdot \mathbf{P_C}^{(q-1)t} \cdot \mathbf{I_P}^{(q-1)} + (1 - \psi) \cdot (\varphi_1 \cdot \text{diag}(\mathbf{C_V}^{(q-1)}(i, :) \cdot \mathbf{V_S}^{(q-1)}) \cdot \mathbf{I_C}^{(q-1)} \\ & + \varphi_2 \cdot \text{diag}(\mathbf{C_T}^{(q-1)} \cdot \mathbf{T_S}^{(q-1)}) \cdot \mathbf{I_C}^{(q-1)}), \end{aligned} \quad (8.18)$$

where φ_1 and φ_2 are the weights determining the impact of the visual and the textual content of the themes and ψ is the transition parameter of the random walk that balances the impact of interestingness due to participants and due to themes. Clearly, when $\psi = 1$, the interestingness of conversation depends solely on the interestingness of the participants at $q-1$; and when $\psi = 0$, the interestingness depends on the theme strengths in the previous time slice $q-1$.

We observe that the measures of interestingness of participants and of conversations described in previous sections involve several free (unknown) parameters. In order to determine optimal values of interestingness, we need to learn the weights $\alpha_1, \alpha_2, \alpha_3, \gamma_1$ and γ_2 in eqn. 8.17 and the theme impact weights φ_1 and φ_2 for the conversations in eqn. 8.18. Moreover, the optimal measures of interestingness should ensure that the variations in their values are smooth over time. Hence we present a novel joint optimization framework, which maximizes the two interestingness measures for optimal $(\alpha_1, \alpha_2, \alpha_3, \gamma_1, \gamma_2, \varphi_1, \varphi_2)$ and also incorporates temporal smoothness. We elaborately discuss the joint optimization framework below.

The joint optimization framework of interestingness is based on the idea that the optimal parameters in the two interestingness equations are those which maximize the interestingness of participants and of conversations jointly. Let us denote the set of the parameters to be optimized as the vector, $\mathbf{X} = [\alpha_1, \alpha_2, \alpha_3, \gamma_1, \gamma_2, \varphi_1, \varphi_2]$. We can therefore represent $\mathbf{I_P}$ and $\mathbf{I_C}$ as functions of $\mathbf{X}$. We define the following objective function $g(\mathbf{X})$ to estimate $\mathbf{X}$ by maximizing $g(\mathbf{X})$:

$$g(\mathbf{X}) = \rho \cdot \|\mathbf{I_P}(\mathbf{X})\|^2 + (1 - \rho) \cdot \|\mathbf{I_C}(\mathbf{X})\|^2, \quad (8.19)$$

$$\text{s.t. } \alpha_1, \alpha_2, \alpha_3 \geq 0, \gamma_1, \gamma_2 \geq 0, \varphi_1, \varphi_2 \geq 0, \mathbf{I_P} \geq 0, \mathbf{I_C} \geq 0, \alpha_1 + \alpha_2 + \alpha_3 = 1, \gamma_1 + \gamma_2 = 1, \varphi_1 + \varphi_2 = 1.$$

In the above function, ρ is an empirically set parameter to balance the impact of each interestingness measure in the joint optimization. Now to incorporate temporal smoothness of interestingness in the above objective function, we define a L^2 norm distance between the two interestingness measures across all consecutive time slices q and $q - 1$:

$$\begin{aligned} d_P &= \sum_{q=2}^Q (\|\mathbf{I_P}^{(q)}(\mathbf{X})\|^2 - \|\mathbf{I_P}^{(q-1)}(\mathbf{X})\|^2), \\ d_C &= \sum_{q=2}^Q (\|\mathbf{I_C}^{(q)}(\mathbf{X})\|^2 - \|\mathbf{I_C}^{(q-1)}(\mathbf{X})\|^2). \end{aligned} \quad (8.20)$$

We need to minimize these two distance functions to incorporate temporal smoothness. Hence we modify our objective function,

$$g_1(\mathbf{X}) = \rho \cdot \|\mathbf{I_P}(\mathbf{X})\|^2 + (1 - \rho) \cdot \|\mathbf{I_C}(\mathbf{X})\|^2 + \exp(-d_P) + \exp(d_C), \quad (8.21)$$

where $\alpha_1, \alpha_2, \alpha_3 \geq 0, \gamma_1, \gamma_2 \geq 0, \varphi_1, \varphi_2 \geq 0, \mathbf{I_P} \geq 0, \mathbf{I_C} \geq 0, \alpha_1 + \alpha_2 + \alpha_3 = 1, \gamma_1 + \gamma_2 = 1, \varphi_1 + \varphi_2 = 1$.

Maximizing the above function $g_1(\mathbf{X})$ for optimal $\mathbf{X}$ is equivalent to minimizing $-g_1(\mathbf{X})$. Thus this minimization problem can be reduced to a convex optimization form because (a) the inequality constraint functions are also convex, and (b) the equality constraint is affine. The convergence of this optimization function is skipped due to space limit.

Now, the minimum value of $-g_1(\mathbf{X})$ corresponds to an optimal $\mathbf{X}^*$ and hence we can easily compute the optimal interestingness measures $\mathbf{I_P}^*$ and $\mathbf{I_C}^*$ for the optimal $\mathbf{X}^*$.

To summarize, we have developed a method to characterize interestingness of conversations based on the themes, and the interestingness property of the participants. We have jointly optimized the two types of interestingness to get optimal measure of interestingness of conversations. Now we discuss an evaluation method of this proposed measure.

8.6 Consequences of Interestingness

An interesting conversation is likely to have consequences. These include the (commenting) activity of the participants, their cohesiveness in communication and an effect on the interestingness of the themes. It is important to note here that the consequence is generally felt at a future point of time; that is, it is associated with a certain time lag (say, δ days) with respect to the time slice a conversation becomes interesting (say, q). Hence we ask the following three questions related to the future consequences of an interesting conversation:

Activity. Do the participants in an interesting conversation i at time q take part in other conversations relating to similar themes at a future time, $q + \delta$ We define

this as follows,

$$Act^{q+\delta}(i) = \frac{1}{|\varphi_{i,q+\delta}|} \sum_{k=1}^{|\varphi_{i,q+\delta}|} \sum_{j=1}^{|P_{i,q}|} \mathbf{P}_C^{(q+\delta)}(j,k), \quad (8.22)$$

where $P_{i,q}$ is the set of participants on conversation i at time slice q , and $\varphi_{i,q+\delta}$ is the set of conversations m such that, $m \in \varphi_{i,q+\delta}$ if the KL-divergence of the theme distribution of m at time $q+\delta$ from that of i at q is less than an empirically set threshold: $D(C_T^{(q)}(i, :)||C_T^{(q+\delta)}(m, :)) \leq \varepsilon$.

Cohesiveness. Do the participants in an interesting conversation i at time q exhibit cohesiveness in communication (co-participate) in other conversations at a future time slice, $q+\delta$ In order to define cohesiveness, we first define co-participation of two participants, j and k as,

$$O^{(q+\delta)}(j;k) = \frac{\mathbf{P}_P^{(q+\delta)}(j,k)}{\mathbf{P}_C^{(q+\delta)}(j,k)}, \quad (8.23)$$

where $\mathbf{P}_P^{(q+\delta)}(j,k)$ is defined as the participant-participant matrix of co-participation constructed as, $P_C^{(q+\delta)} \cdot (P_C^{(q+\delta)})^t$. Hence the cohesiveness in communication at time $q+\delta$ between participants in a conversation i is defined as,

$$Co^{(q+\delta)}(i) = \frac{1}{|P_{i,q}|} \sum_{j=1}^{|P_{i,q}|} \sum_{k=1}^{|P_{i,q}|} O^{(q+\delta)}(j;k). \quad (8.24)$$

Thematic Interestingness. Do other conversations having similar theme distribution as the interesting conversation c_i (at time q), also become interesting at a future time slice $q+\delta$ We define this consequence as thematic interestingness and it is given by,

$$TInt^{(q+\delta)}(i) = \frac{1}{|\varphi_{i,q+\delta}|} \sum_{j=1}^{|\varphi_{i,q+\delta}|} I_C^{(q+\delta)}(j). \quad (8.25)$$

To summarize, we have developed a method to characterize interestingness of conversations based on the themes, and the interestingness property of the participants. We have jointly optimized the two types of interestingness to get optimal

Table 8.2: Experimental Setup based on multiple rich media and blog datasets.

DATASET	MEDIA	CONVERSATION	CONTENT THEMES	TEXTUAL THEMES
YouTube	Video	Temporal sequence of comments associated to a shared video	The visual features of the video (features are computed over a key frame in each video, where the key frame corresponds to the median duration of the video)	The textual content of the comments and replies
Flickr	Photo	Temporal sequence of comments associated to a shared photo	The visual features of the photo	The textual content of the comments and replies
Engadget	Blog post	Temporal sequence of comments associated to a blog post (typically made by an Editor)	TF-IDF based features of the blog content, its timestamp	The textual content of the comments and replies
Huff Post	Blog post	Temporal sequence of comments associated to a blog post (typically made by an Editor)	TF-IDF based features of the blog content, its timestamp	The textual content of the comments and replies

interestingness of conversations. And finally we have discussed three metrics which account for the consequential impact of interesting conversations. Now we would discuss the experimental results on this model.

8.7 Experimental Results

We present our experiments in the context of four different datasets that have been found to be relevant to our problem of measuring interestingness property of conversations. We provide a summary of the definition of a conversation in each dataset, what we refer to by content themes as well as what we mean by textual themes of conversations in Table 8.2. Additionally some of the details of the four datasets used in the experiments here are shown in Table 9.2.

We discuss the data collection method for the YouTube dataset in particu-

Table 8.3: Details of the four datasets.

Dataset	#Participants	#Conversations	#Comments
YOUTUBE	17,736,361	272,810	145,682,273
FLICKR	4,304,525	305,258	26,557,446
ENGADGET	78,740	45,073	6,580,256
HUFFINGTON POST	59,282	24,479	4,748,837

lar. We executed a web crawler to collect conversations (set of comments) associated with videos in different featured categories on the YouTube website: such as “Comedy”, “Education”, “Entertainment”, “News & Politics”, “Sports”, “Music”. We seeded the crawling process from different YouTube classified genres, e.g. “Most Responded”, “Top Favorited”, “Top Rated” and “Most Discussed”; the motivation being, we snowball from a set of videos that have rich conversations around them, as well as are linked to several other videos. We continued the crawl from the seeds by spidering their linked related videos. For each video, apart from downloading its content i.e. the video itself, we collected its timestamp, tags, its associated set of comments, their timestamps, authors and content. We crawled a total set of 272,810 videos involving 17,736,361 unique participants and 145,682,273 comments during 2008-2009.

Now there are two parts to our experiments. First we present the quantitative analysis of the dynamics of interestingness. Next we present elaborate evaluation of our method against several baseline techniques.

Analysis of Interestingness

We analyze two components of the interestingness measure of the conversations—the temporal dynamics of the textual themes and the computed interestingness. Both of these analyzes are evaluated over time with respect to a set of external happenings (i.e. news items) obtained from the Google News website. We judged on the significance of the events based on their number of views. The goal is to be able to see whether the textual themes and the interestingness measure in general are meaningful with respect to the actual events in the physical world that they

Figure 8.3: Ripple visualization of conversational themes over time—15 week period between Oct. to Dec. 2008.

attempt to reflect upon.

To analyze the conversational textual themes, we developed a “ripple visualization” to show the following characteristics of each theme—(a) the strength/intensity (or degree of “hotness”) of the theme, (b) the degree of participation, given by the number of unique participants involved with the theme, and (c) the temporal spread of the theme in the period of analysis.

The ripple visualization of three themes¹¹ corresponding to two topical categories from YouTube—“News & Politics” and “Sports” are shown in Fig. 8.3.¹² Each ripple (i.e. set of concentric circles) represents a theme. Each concentric circle in the ripple implies a time slice, which is a week in our experiments; with the inner-most circle being the first time slice of analysis, while the outer-most one the last one (time span of 15 weeks between Oct.–Dec. 2008). The color intensity of

¹¹Note that in our proposed theme model, we need to specify the number of theme clusters, k . In this work, k is decided on the basis of the topmost k eigen components of the conversational chunk matrix.

¹²An elaborate description of the contents of representative themes belonging to these categories is given in Table 8.7. We present word clouds of the content of a representative theme belonging to three of the topical categories in our YouTube dataset in Table 8.7. Note that for all the three categories, the theme, extracted using our proposed method, captures the dynamics of both the interest of the YouTube community in general, as well as matches with the events during the time period of the data crawl.

Table 8.4: Word cloud of a representative theme belonging to three topical categories on YouTube.

NEWS & POLITICS	ahead alive alone america american benefits best better bill bush businesses care cars century challenges chance change child children college companies could've country create democrats deserve different doesn't dreams economy education eight election finish five fundamental future generation george government hands money moral nation night nuclear oil past pay people plan policies progress promise protect proud provide purpose pursue renewable republicans responsibility restore rise safe security seen senator sense serve sick someone something stand states stood strength woman workers years
SPORTS	Alex Rodriguez Bill Belichick Boston Bruins Boston Celtics Brett Favre Brian Scalabrine Brown Brown University CC Sabathia Celtics Chicago Bulls China Cleveland Cavaliers college basketball college football college sports courts Danny Ainge David Ortiz Doc Rivers Dustin Pedroia economy Eddie House Elizabeth Beisel ESPN Field Hockey Golf Paul Pierce Pawtucket Red Sox Providence Bruins Providence College Red Sox Rick Pitino Rob Lee Robert Lee SoxTalk sports cover SportsTalk Stephon Marbury steroids Tampa Bay Rays Tiger Woods Tom Brady Tony Allen UConn University of Rhode Island URI URI basketball Yankee Stadium
MUSIC	alternative amarok analysis artists Audacity audio audioscrobber backup band bands beats Bittorrent blog blogging blogs bpm brasil brazil Classical code codec collaboration Commons digitalstorytelling directory distribution distro DIY dj download Downloads drm drum drummachine drums electronics Flash foss foxytunes free free_music JavaScript jazz metal microphone midi mixes mixing mobile motif movie movies mozilla mp3 mpb multimedia music pandora pc PD playlist project remix remote research resource synthesis synthesizer techno technology Technology-News tempo video videos visualization wishlist YouTube

each circle implies the strength or ‘hotness’ of the theme¹³. Further, the thickness

¹³We determine the strength of a theme by the number of conversations that are clustered under that theme at any time slice. Note that although our proposed textual theme model generates a theme distribution corresponding to each conversation (soft clustering), we enforce a hard assignment of a single theme to a conversation by picking the theme with the highest probability in the distribution.

of each circle represents the degree of participation, in terms of the frequency of unique participants involved in the corresponding conversations at that time slice. The different circles in each ripple are also overlaid with a number of news items (from Google News), which have been observed to be of high significance during that time period. The correspondence between news items and the themes is obtained using cosine similarity measure between the contents of the news items and the content of the conversations associated with the themes. More details on an alternative visualization for studying the textual themes can be found in (De Choudhury et al., 2009c).

From the ripple visualization in Fig. 8.3, we observe that the themes exhibit high strength, shown in high color intensity, during the time slices where the news items appear. That is, there is high correlation between the time periods when our theme model shows themes to be hot, and the time periods when significant news events appear, that are related to the theme content. This shows that our proposed textual theme model, is able to capture the dynamics of the external happenings around which conversations emerge on rich media websites. However, an interesting point to note here is that, the same time slices when the theme is hot, the degree of participation during the same time, is not always necessarily high. This implies that themes can become strong or hot even with limited participation from users—likely because of high communication activity between participants who are highly interested in the particular news event happening.

The results of interestingness of the participants are shown in Figure 8.4. We have shown a set of 45 participants over the period of 15 weeks (June 20, 2008 to September 26, 2008) by pooling the top three most interesting participants over all conversations from each week. From left to right, the participants are shown with respect to decreasing mean number of comments over all 15 weeks. The figure shows plots of the comment distribution and the interestingness distribution for the participants at each time slice along with the Pearson correlation coefficient between the two distributions. From the results, we observe that on the last three weeks (13, 14, 15) with several political happenings, the interestingness distribution of participants

does not seem to follow the comment distribution well (we observe low correlation). Hence we conclude that during periods of significant external events, participants can become interesting despite writing fewer comments—high interestingness can instead be explained due to their preference for the conversational theme which reflects the external event.

Figure 8.4: Interestingness of 45 participants from YouTube, ordered by decreasing mean number of comments from left to right, is shown along with the corresponding number of comments over 15 weeks (rows). The Pearson correlation coefficient between the number of comments and interestingness is also shown; which implies that interestingness of participants is less affected by number of comments during periods of significant external events.

The results of the dynamics of interestingness of conversations are shown in Fig. 8.5. We show a temporal plot of the mean and maximum interestingness per week in order to understand the relationship of interestingness to external happenings¹⁴. From the plots over two YouTube topical categories “News & Politics”

¹⁴These news items were obtained from <http://www.news.google.com/> in the said

Figure 8.5: Mean and maximum interestingness of conversations on two topical categories from YouTube: “News and Politics” and “Sports” over a 15 week period between Oct. and Dec. 2008.

and “Sports”, we observe that the mean and maximum interestingness of conversations have significantly different patterns. The prime observation is that while it seems the mean interestingness shows hikes during periods of external happenings, the maximum interestingness over time exhibits rather stable characteristics. For example, in the case of “News & Politics”, the plot of mean interestingness shows increase between weeks 6 and 11, which is also the time period during which the outcomes of the 44th Presidential elections in the United States were out. Similarly, for “Sports”, the initial time period i.e. before week 8 seems to overlap with the football season in the US (Oct/Nov)—hence the higher values of mean interestingness. However there is much less volatility for the plots of maximum interestingness with respect to these news items.

Hence it seems that more conversations in general become highly interesting period and their significance were determined based on the number of views on each item.

when there are significant events in the external world—an artifact that online conversations are reflective of chatter about external events. However, certain highly interesting conversations always occur at different weeks irrespective of events. This implies that conversations could become interesting even if the themes they discuss are not very popular at that point of time—rather, the interestingness in such cases could be attributed to the communication activity of the participants.

Interestingness and Rich Media Attributes

For the interestingness measure corresponding to each dataset, we have explored the relationships between our computed interestingness of conversations and the attributes of their associated media objects. We consider correlation (using the Pearson correlation coefficient) between interestingness (averaged over all time slices) and a host of media attributes. Note that these media attributes are going to differ across different datasets; hence we provide a list of the relevant attributes that were used for this set of experiments, in Table 8.7. Note that there are significant differences in the set of attributes between the two rich media websites: YouTube and Flickr, and the blog forum based social media sites: Engadget and Huffington Post, indicating that the nature of the shared media object is significantly different in the two kinds of datasets considered in this chapter.

We now discuss the results of the correlation coefficient between our computed conversational interestingness and the media attributes corresponding to each of these datasets (Table 8.7). Overall, our observations reinstate that media object popularity, as accounted for by these attributes has little correlation or impact for that matter, on the interestingness of the associated conversations to participants. Conversational interestingness therefore does not imply media popularity or vice versa. We conjecture that this bolsters our claims regarding our proposed model of determining interestingness; where we consider the dynamic properties of the conversation such as themes and participants’ communication behavior, rather than attributes of the media.

Now we explore the relationships between our computed interestingness of

Table 8.5: Relevant media attributes corresponding to multiple rich media and blog datasets. These attributes were used to investigate how much useful they are for accounting for conversational interestingness.

DATASET	RELEVANT MEDIA ATTRIBUTES
YouTube	number of views, number of favorites, ratings, number of linked sites, time elapsed since video upload and video duration
Flickr	number of tags, number of notes, number of views, number of favorites, number of associated groups, time elapsed since photo upload
Engadget	number of tags, time elapsed since blog was posted, number of Facebook “likes”, length of the post
Huffington Post	number of tags, time elapsed since blog was posted, number of Facebook “likes”, length of the post

Table 8.6: Mean correlation coefficient between interestingness and media attributes over all four datasets. Note that media tenure implies the time elapsed since the video, image or blog post was uploaded or posted.

Dataset	#views	#tags	Media tenure
YOUTUBE	0.63	0.42	-0.22
FLICKR	0.42	0.37	-0.37
ENGADGET	0.57	0.34	-0.03
HUFFINGTON POST	0.67	0.15	-0.13

conversations and the attributes of their associated media objects. We consider correlation (using the Pearson correlation coefficient) between interestingness (averaged over 15 weeks) and number of views, number of favorites, ratings, number of linked sites, time elapsed since video upload and video duration which are media attributes associated with YouTube videos. Table 8.7 gives the correlation measures for these attributes over different topical categories featured on the YouTube website. The results reveal interestingly that these media attributes are in most cases not correlated (e.g. ratings, number of linked sites), and even negatively correlated (e.g. time elapsed from video upload, duration) with our computed interestingness measure of the conversations. This is intuitive because videos which are recently uploaded and generate lot of attention quickly are likely to be highly interesting; also, most interesting conversations have been observed to be those which are short in duration.

Table 8.7: Correlation coefficient between interestingness and media attributes, shown for each topical category on YouTube.

Categories	#Views	#Favorites	Ratings	#Linked Sites	Tenure	Duration
COM.	0.343	0.425	0.112	0.072	-0.277	-0.432
EDN.	0.662	0.692	0.096	0.153	-0.225	-0.142
ENT.	0.452	0.394	0.186	0.221	-0.248	-0.385
NEWS	0.693	0.732	0.196	0.152	-0.297	-0.236
SPT.	0.715	0.755	0.136	0.246	-0.312	-0.077
MUSIC	0.282	0.229	0.154	0.096	-0.332	-0.267

Additionally, across the different thematic categories, we observe that the correlation measures are considerably low for ‘Comedy’, ‘Entertainment’ and ‘Music’, compared to others such as ‘Sports’. We conjecture that the reason behind this result is that participants are more likely to return to videos belonging to ‘Comedy’ or ‘Music’ just to re-watch or re-listen the video itself, than to engage in conversations. On the other hand, the category ‘Sports’ often featuring real-world games, are likely to engage participants into long dialogues with each other due to either fan-following, sports gatherings / events and so on. Hence participants on such videos are likely to be more vocal and opinionated in expressing themselves via the conversations, instead of repeatedly watching the same game (or gaming clip) time and again across multiple visits.

Therefore overall, our observations reinstate that media object popularity, as accounted for by these attributes has little correlation or impact for that matter, on the interestingness of the associated conversations to participants. Conversational interestingness therefore does not imply media popularity or vice versa. We conjecture that this bolsters our claims regarding our proposed model of determining interestingness; where we consider the dynamic properties of the conversation such as themes and participants’ communication behavior, rather than attributes of the media.

8.8 Evaluation

We reiterate that a prime challenge in this research is the evaluation of a subjective measure as interestingness, in the absence of any ground truth data availability. In this light, we now discuss some elaborate experiments involving the consequence metrics, as well as establish comparative validation of our proposed measure against other alternative definitions of interestingness.

Baseline Methods

For the purpose of evaluation using the three consequence measures, we now discuss three baseline methods of determining alternative measures of interestingness. We define the first baseline interestingness measure of a conversation based on the number of comments in a particular time slice so that it satisfies the following two constraints as in (Dubinko et al., 2006): (a) a conversation is interesting at a time slice when it has several comments in that time slice, and (b) a conversation should not be considered interesting if all its comments are in a particular time slice and no comments occur in other time slices. The second baseline is based on the idea of novelty in participation: if several new participants join in a conversation at time q who did not appear at any time slice before q , then it implies the conversation is interesting. The third baseline is based on ranking conversations using the PageRank algorithm on the participant-co-occurrence graph $G(C, E)$ discussed previously. This is based on the motivation that if the participants of several conversations co-communicate on another conversation, it makes the latter interesting as it appeals to a large number of individuals.

In order to evaluate our proposed interesting measure, we utilize the notion of how well, our method can explain future consequences, compared to simpler baseline techniques (ref. Appendix). In order to understand this phenomenon, we conjecture that there will be a certain temporal lag (positive or negative) between the interestingness distribution of a conversation, and its associated consequence metric—such as activity, cohesiveness and thematic diffusion. Hence we compute

Table 8.8: Best lags in explaining the consequence metrics using our proposed method. The numbers are shown over multiple rich media and blog datasets.

DATASET	ACTIVITY	COHESIVENESS	THEMATIC DIFFUSION
YouTube	8 days	9 days	7 days
Flickr	14 days	11 days	10 days
Engadget	7 days	10 days	11 days
Huffington Post	3 days	4 days	2 days

Pearson correlation coefficient between the interestingness distribution and the consequence metric distribution, considering a range of lags between -40 days to 40 days. We report the ‘best lag’ for each metric, as found by our proposed method in the case of the four different datasets (two rich media websites, two blog-based forums) in Table 8.8.

From the best lags for each consequence metric, we observe significant differences across datasets. While the numbers seem to be more or less consistent across the different metrics, we observe that for Flickr, the lags are higher, and for Huffington Post, they are much smaller. We conjecture that the reason behind this is that Flickr is more of a interest-driven community—that is the consequences, activity, cohesiveness and thematic diffusion are driven more by the common interests of individuals or their social cliques indicating homophily along certain photographic attributes. Hence the ability of conversational interestingness to explain the consequences in a smaller temporal time scale is very low. On the other hand, Huffington Post appears to be a highly dynamic community; with consistent participants in conversations, as well as new participation, often triggered by different political news events. As a result the dynamics of the conversations are very volatile over time. This causes the lag involved in explaining the consequence metrics to be relatively low as observed in our results.

We now conduct a series of experiments on four different datasets to evaluate our proposed interestingness measure. Evaluation is performed based on a mutual information based “response metric”—i.e. how well our computed interestingness measure can explain future consequences, such as activity, cohesiveness of partici-

Figure 8.6: Best lag for correlation of interestingness measures to the three consequence-based metrics: activity, cohesiveness and thematic interestingness. Our method with temporal smoothing (I_1) is seen to be sharply correlated with the three metrics of consequences having the following lags—8 days for activity, 9 days for cohesiveness and 7 days for thematic interestingness.

pants, as well as thematic diffusion. We also compare our method with a series of baseline methods.

We present the results of measuring consequence of interestingness on the YouTube dataset captured by the three metrics discussed previously—activity, cohesiveness and thematic interestingness. To observe the consequential impact of interestingness, we determine its correlation to activity, cohesiveness and thematic interestingness using five methods—our interestingness measure with temporal smoothing (I_1), our interestingness measure without temporal smoothing (I_2), and the three baseline methods $B_1 - B_3$. The three consequence metrics would be felt after a certain time lag with respect to the point at which a conversation became interesting. Hence for each metric and method pair, we need to determine by what time lag the metric trails the interestingness with maximum correlation. Since interestingness of a conversation and its associated activity, cohesiveness or thematic interestingness computed over different time slices (weeks) can be considered to be time-series, we determine the cross-correlation between interestingness and each of the consequence-based metrics for various values of lags (-40 to 40 days for leading and trailing consequences). The lag corresponding to which the correlation is maximum, is taken as the ‘best lag’.

Fig. 8.6 shows the correlation between the consequence-based metrics and interestingness of conversations computed using various methods for various lags,

averaged over the entire period of 15 weeks. We observe that incorporating temporal smoothing significantly improves correlation (I_1 over I_2) for our method and this is explained by the fact that interestingness of conversations exhibits considerable relationship across time slices. We finally conclude from these results that our computed interestingness appears to have significant consequential impact on the three metrics due to high correlation compared to all baseline methods (mean correlation of 0.71 over all three metrics)—all the three baseline methods appear to have more or less flat correlation plots (mean correlation of 0.35 over all three metrics). Hence interestingness of conversations determined through our method could be predictors of communication dynamics in social media.

Next, the results of the MI-based response for the rich media datasets and the blog datasets are shown in Figure 8.7 and Figure 8.8 respectively. The following are the main insights gained from these experiments results:

1. *YouTube*: The results conducted on the extended dataset for YouTube (i.e. averaged over six different topical categories) are found to be consistent with our earlier observations on just the “Politics” category. We observe that our method I_1 maximizes mutual information for all three metrics (mean 0.83)—implying that our computed interestingness can successfully explain the three consequences compared to the baseline methods (mean 0.41). The baseline methods perform poorly because they have relatively flat correlation with the three consequences. This implies that our methods are effective in explaining the consequences reasonably.
2. *Flickr*: Note that a major issue with the Flickr dataset is that Flickr conversations are not inherently rich. That is, most of the conversations, though involve a lot of comment, but are essentially critical feedback on the media object, i.e. the photo shared. On some other occasions, comments are often invites to the uploader of the photo to associate the image to some relevant group. Hence our proposed measure of interestingness of conversations performs relatively poorly in accounting for the three consequence metrics. Nev-

Figure 8.7: Mutual Information based response over the three consequence metrics. Results are shown for two rich media based datasets: YouTube and Flickr. We evaluate our computed interestingness I_1 and I_2 against baseline methods, B_1 (comment frequency), B_2 (novelty of participation), B_3 (co-participation based PageRank). Our method incorporating temporal smoothness (I_1) uses its corresponding best lags and maximizes the mutual information based response metric for the three consequence-based metrics (activity, cohesiveness and thematic diffusion).

ertheless our proposed method incorporating temporal smoothness (I_1) yields the highest response for all the three cases. Among the three baseline cases, however, we do observe some variations. (a) We observe that the PageRank based measure i.e. B_3 performs the best among the baselines for cohesiveness. This is reasonable because on Flickr, participation is often driven by the social graph structure than the content of the media itself. A likely reason behind this could be that on friending someone on Flickr, the photos shared by the other individuals become visible to the friends via RSS feeds. This facilitates future participation, increasing cohesiveness. Hence graph structure based methods as B_3 perform better than ones that reply purely on comment attributes, as B_1 . (b) We further observe that thematic diffusion, as accounted for by the different methods is relatively low. This is because of the media

sharing culture on Flickr. Flickr is primarily used as a pro photographers social site, where the main goal is to obtain feedback on the shared item. Hence there aren't enough dynamic themes evolving around the conversations. As a consequence, thematic diffusion is barely affected by the characteristics of the conversation itself. In fact it is often based on the personal interests of users on different photography relate constructs.

3. *Engadget*: The results on the Engadget dataset are revealing. While we observe that our method yields the best performance (i.e. I_1), the main takeaway from the variations across the different baseline techniques is that Engadget is a close-knit technology-savvy community—often involving sets of users who are passionate about a particular type of technology / gadget / software. Hence we observe high cohesiveness response for all the methods—indicating that the characteristics of the conversations do indeed impact future co-participation of individuals on relevant blog posts in the future. Note that for all the three metrics, the worst performance seems to be given by B_2 , which is the novelty of participation based interestingness measure. Given Engadget behaves a close knit community, it is understandable that new participation does not adversely affect conversational interestingness. Hence B_2 is not able to explain the future consequences in a satisfactory manner. For the same reason of the community being close knit, the participation based metrics, activity and cohesiveness are better accounted for by B_1 , which is based on the properties of the comments exclusively.
4. *Huffington Post*: Although a similar blog forum as Engadget, Huffington Post appears to be a more open community. We conjecture that this is because it invites political opinions from a variety of individuals with different backgrounds and communication behavior. Often these individuals are brought together, not merely because they are interested in a certain political topic, but because of a certain temporal event. This explains some of the variations in the performance of different baseline techniques. Again, in summary,

Figure 8.8: Mutual Information based response over the three consequence metrics. Results are shown for two blog-based datasets: Engadget and Huffington Post. We evaluate our computed interestingness I_1 and I_2 against baseline methods, B_1 (comment frequency), B_2 (novelty of participation), B_3 (co-participation based PageRank). Our method incorporating temporal smoothness (I_1) uses its corresponding best lags and maximizes the mutual information based response metric for the three consequence-based metrics (activity, cohesiveness and thematic diffusion).

our method I_1 performs the best, but the variations across the baselines are insightful. Thematic diffusion response seems to be considerably low; we conjecture this is explained by the fact that individuals are brought together into participation in a conversation not because of the thematic spread, but likely because of a related external happening. Hence the interestingness of the conversations do not necessarily spread to other conversations at a future time slice. The baseline B_1 is also observed to perform poorly for this dataset, indicating that conversational interestingness is affected by factors external to the conversation, such as news events, instead of purely participation based factors (such as comment frequency rate as used by B_1).

The major observation from the experiments across these datasets is that

our proposed method of measuring conversational interestingness on different social communication based websites performs the best over alternative methods. However we do observe significant differences in the performance of the different baseline techniques over the datasets. It appears that interestingness of conversations is strongly affected by some of the accepted “cultures” of communication on these different websites. For example, there are different motivations on these websites that cause users to engage in communication activity with another user. Additionally, the role played by the social graph in each of these datasets, in triggering participation is also widely different. Due to all these reasons, we observe the differences. However our method, since it accounts for the thematic popularity as well as the dynamics of the communication process among the participants, seems to balance out the impact of each aspect on the datasets, leading to promising results.

Predictive Utility of Interestingness

As a complementary direction to understand the utility of conversational interestingness empirically, we conducted a set of predictive experiments. The goal of these experiments was to observe if the measure of interestingness of a conversation can be indicative of the measure of another orthogonal variable in the future, that is also a property of the conversation in general. We considered three such variables in this chapter: (a) the number of comments in a conversation at any point in time, (b) the rate of comments posted per unit time, and (c) the rate of arrival of new participants to the conversation over time. The granularity of time considered here is one week as before.

For the exact set of predictive experiments, we performed linear regression with each of the three variables above as the control variables at each time slice i.e. week $t + 1$, while we learned the associated weights given the interestingness measure of each conversation at time t . As baseline comparison, we also considered the case of auto-regression of each variable. The results of these experiments are shown in Fig. 8.9. These plots interestingly reveal that the predicted distribution of the variables over the 15 week period is much closer to the actual distribution (mean

Figure 8.9: Prediction of different conversational features based on (1) interestingness and (2) auto-regression over historical values for the YouTube dataset. Note, that interestingness value of a conversation is a better predictor.

Table 8.9: Mean prediction error (%) of different conversational features based on (1) interestingness and (2) auto-regression over historical values over all four datasets. The errors are averaged over a period of ~ 15 weeks.

Dataset	#comments	Rate: comments	Rate: newcomer arrival
YOUTUBE	18.36%	17.83%	21.13%
FLICKR	15.73%	18.46%	15.21%
ENGADGET	22.83%	20.25%	16.94%
HUFF POST	19.58%	17.36%	18.83%

error is $< 19\%$), in comparison to the baseline case of auto-regression (mean error is $> 80\%$). These experiments shed useful insights into how our proposed measure of interestingness can be useful in real world scenarios to understand future trends variables associated with the conversations or the media object in consideration. Similar observations can be made from the prediction tasks on the other three datasets: Flickr, Engadget and Huffington Post. This is shown in Table 8.8.

Comparison of Visual and Textual Themes

An important extension in the work presented in this chapter, over our previous publication (De Choudhury et al., 2009c) is that the conversational themes are now defined based on both the characteristics of the media object (i.e. video on YouTube) as well as the textual content of the comments in the conversation. Hence in this section we evaluate the performance of each of these theme types in computing the measure of interestingness. We compute the mutual information based response metric for each thematic type: (1) visual themes only, (2) textual themes only and (3) both (which is our proposed method), corresponding to each of the three

Figure 8.10: Comparison of visual and textual themes of conversations over different topical categories on YouTube. The performance of each thematic type is evaluated based on the mutual information based response metric, that explains how much our computed interestingness measure can account for the three consequence metrics.

consequence metrics.

The results over the six YouTube topical categories is shown in Fig. 8.10. In all the cases, we observe that using just the visual features yield low response for the consequence metrics; indicating that visual attributes of the shared media object cannot always explain the future consequences associated with a conversation. On the other hand, the textual themes perform relatively better; while the best performance is given by the case when we use both thematic types. Overall, compared to visual themes, we have about a 15% improvement on using both thematic types, which is statistically significant across all topical categories on YouTube.

Nevertheless, there are some interesting observations in Fig. 8.10. Note that for the categories “Sports” and “Music” the visual themes yield very good MI-based response for two of the three consequence metrics. We conjecture that this is because the two categories both have a strong emphasis on the content of the video: e.g. a recording of a game that people like or a music video from a favorite band—leading individuals to return to the video often to watch it again, than to engage in conversations. Hence the visual features yield very good performance in these cases.

Another interesting observation in the results is that the textual themes per-

form relatively well for categories ‘Comedy’ and ‘News & Politics’. This is because, the conversations in these categories generate substantial commentary among the visitors, and often times, the users come back to the videos due to the conversation itself, than to watch the comedy video or the politics related event. This explains why the textual themes yield good performance in this case.

8.9 Discussion

Rich media sites such as YouTube and Flickr were primarily developed to support active sharing of multimedia artifacts to large audiences. However, over time, they have begun to provide opportunities for complex but rich and flexible processes of interaction between participants, essentially via their comment feature. Such conversational nature of communication activity not only structures their social relations, it also provides the space within which social relations emerge (e.g. evolution of an online community) and the tool that they use to define that space, in terms of media creation and consumption. Consequently, understanding the dynamic nature of such conversations is very important to reveal participatory engagement in collaborative scenarios, because conversations have the potential to connect individuals in a variety of situations. In this chapter we have focused on tackling a narrow version of this larger problem by characterizing rich multimedia conversations via their interestingness.

Note that although YouTube is one of the most popular media sharing websites today, there are some distinct characteristics of this website which make it different from other media sharing tools used here, such as Flickr, Engadget or Huffington Post. Because of its diverse thematic span of videos, that includes videos on controversial topics, religion, politics as well as personal opinions, conversations on certain videos seem to feature a fair degree of “flame war” among its participants. It is possible that the presence of such flame wars can either make a conversation provocative to a visitor of the video, or can also make it uninteresting. In this chapter, we have not demonstrated any methodology to incorporate properties of such controversies in a separate manner. While it is often a decision with the

user to decide if these flame wars are of any interest to her, it remains to be seen in future research how modeling such distinct properties of conversations can affect the proposed measure.

Beyond our method and results, an important point to note in this work is that conversational interestingness is essentially a subjective construct; i.e. different sets of attributes might be responsible for how participants perceive of the interestingness property. The work presented in this chapter is essentially quantitative, hence we define interestingness in an objective manner without explicitly incorporating feedback from participants about their cognitive perception of interestingness. While we have demonstrated promising results via elaborate experiments on four different datasets, evaluation of our proposed method can also possibly benefit from ethnographic experimentation. Ethnographic studies involving geographically and demographically diverse populations will likely be able to account for the user subjectivity. Addressing subjectivity in this manner also has the potential to open avenues for alternate definitions of interestingness of rich media conversations. This provides exciting directions for future research.

Another point we would like to mention here is that the perception of conversational interestingness can vary widely across individuals. For example, will a “lurker” (or a passive participant) in a conversation find the same set of conversations to be interesting, compared to someone more active? How will gender, geographic location and other demographic attributes affect the notion of interestingness? In this chapter, we do not account for personalization to cater to different audiences, however it can be a viable extension in the future. For example, interestingness can be incorporated in a recommendation framework, which suggests interesting conversations based on the history of user actions on rich media websites.

8.10 Conclusions

In this chapter, we have proposed a computational framework to characterize the conversations in online social networks through their “interestingness”. Our central intuition is that participants return to the same video time and again

because they find not only the media object to be interesting, but the conversation associated with it as well. Hence we modeled interestingness via two key aspects: the conversational themes and the communication properties of the participants. To begin with, we first computed visual features of the media object associated with a conversation, and thereafter textual themes using a mixture model approach. Second we determined interestingness of participants and interestingness of conversations based on a random walk model. Third, we established the consequential impact of interestingness via metrics: activity, cohesiveness and thematic interestingness. We evaluated our method on four different datasets to observe that our proposed method maximizes the mutual information by explaining the consequences (activity, cohesiveness and thematic interestingness) significantly better than three other baseline methods. Interestingness also seemed to be a predictive indicator of a number of time series variables associated with the media, including the number of comments, rate of posting new comments and rate of new participation. We believe this work can enable a better understanding of the dynamics associated with rich media objects shared online, as well as can lend valuable insights into studying interactional participation and collaboration of users on the web.

Chapter 9

Quantifying Collective Participation in Social Media Conversations

We are motivated in our work by the following question: *what factors influence individual participation in social media conversations?* This question is important due to several reasons. First, conversations around user posted content, is central to the user experience in social media sites, including Facebook, YouTube and Flickr. Second, understanding *why* people participate, can have significant bearing on the following fundamental research questions: social network evolution including changes to the network structure, and information flow.

Our approach is as follows. We first identify several key aspects of social media conversations, distinct from both online forum discussions and other social networks. These aspects include intrinsic and extrinsic network factors. There are three factors *intrinsic* to the network : social awareness, community characteristics and creator reputation. The factors *extrinsic* to the network include: media context and conversational interestingness. We develop one hypothesis for each factor to test the influence of the factor on individual participation. There are two technical contributions of this paper, both related to testing of hypothesis: a Support Vector Regression based prediction framework to evaluate each hypothesis, and a Bayesian Information Criterion (BIC) metric to identify the optimal factor combination. We have interesting findings. First, we show that the factors that influence participation depend on the *media type*: YouTube participation is different from a weblog such as Engadget. Second, different sets of factors influence newcomer and existing participants. Finally, we show that an optimal factor combination improves prediction accuracy of observed participation, by $\sim 9\text{--}13\%$ and $\sim 8\text{--}11\%$ over using just the best hypothesis and all hypotheses respectively. This reveals that there is likely

to be a complex set of factors responsible for the nature of participation observed on different social media conversations today.

9.1 Introduction

Today, rich media sites including Flickr and YouTube as well as weblogs including Engadget and Huffington Post have emerged as popular channels for the expression of individual interests, ideas and opinions¹. These rich media sites allow users to share content, including uploading images, text and videos. Importantly, the shared content allows users to communicate with other users, through comments on the shared media object. We define a sequence of temporally-ordered comments on the shared media object, as a “conversation.”

Conversations are important to understand the nature of the underlying social network (?). In particular, conversations can be used to study the following: user behavior (Qiu and Cui, 2010) and information roles, including content dissipators (Biel and Gatica-Perez, 2009), impact on information cascades (Cha et al., 2009), and influence propagation (Anagnostopoulos et al., 2008). Hence, it is important to understand user participation in the context of social media conversations. For example, why do certain conversations exhibit continued and increasing participation from individuals? In this light, our work in this paper is motivated by the following question: *what are the factors that influence individual participation in social media conversations?* Notice that by “participation,” we mean that a user has posted comments on a conversation.

Understanding the motivations behind participation of individuals in social media conversations involves several challenges. These challenges are related to key aspects of the social network: the inherent culture of interaction within the greater community, the affinity of the community to invite new individuals, the standard practices of social actions and the goal and purpose of the community-wide inter-

¹As of May 2010, YouTube features more than 2 billion views a day and 24 hours of video uploaded per minute: <http://www.digitalbuzzblog.com/infographic-youtube-statistics-facts-figures/>

Figure 9.1: Participation (percentage of users on the y -axis and number of posted comments per user, on the x -axis) on conversations from two types of social media sites: rich media and blog forums. We notice a marked difference in newcomer participation — rich media datasets (YouTube:72.37%, Flickr: 57.43%) attract more newcomers than blog forums (Engadget: 29.35%, Huffington Post: 21.35%). Notice that Flickr has a significant core following (20.76%)—sustained participation from individuals who have posted more than 100 comments.

actions. Contemporary online communities support different types of social interaction, and cater to different kinds of audiences. Rich media sites, for example, including YouTube and Flickr, primarily cater to sharing of media objects. On the other hand, blog forums such as Engadget or Huffington Post are directed towards technology-savvy or liberal political audiences who intend to remain engaged in interactions around news events. Therefore, it is likely that different social media sites will have different factors driving conversational participation within their sites. Furthermore, it is likely that there are differences between the motivations of newcomers to participate, compared to the existing members.

To establish these differences empirically, we show the distribution of participation (percent) over two types of social media sites: two rich media sites and two blog forum sites in (Figure 9.1). We notice a marked difference in the nature of user participation — in contrast to blogs, rich media features a large percentage of

newcomers. Hence identifying factors influencing participation in each of these sites, and how they vary across the types of sites and participants, is critical. In particular, a careful analysis of participation can help contextualize network phenomena (e.g. distribution of information roles, or network dynamics including changes to the structure and information flow) within these sites. An application of our work includes better design of social media websites — in particular, sites where individuals interact with a shared media object (videos, photos, blogs).

Our Approach

We define the participation of individuals on a social media conversation as “collective participation.” There are two aspects to it: newcomers and existing participants. The former, includes individuals who have not posted a comment or reply on the particular conversation thus far. The latter includes participants who have posted at least one comment or reply at an earlier point in time.

We identify *intrinsic* as well as *extrinsic* network factors influencing collective participation from newcomers and existing participants. The nature of the social network in which the conversation is embedded influences *intrinsic* network factors. Intrinsic network factors include: an individual’s ‘social awareness,’ including peer feedback, ‘community characteristics’ including the ability to sustain users, and ‘reputation’ of the media creator. Participants also receive external ‘information signals’ through *extrinsic* network factors, that may be due to an image/video posted in response to an external event, or associated with emergent themes due to conflicting opinions. The extrinsic factors therefore include the ‘media context,’ including visual/textual content, tags, and ‘conversational interestingness.’

Following these two categories of factors, we develop one hypothesis for each factor to test the influence of the factor on participation. In order to examine how well each hypothesis can be attributed to the observed participation of individuals, we adopt a prediction approach. Our goal is to utilize each hypothesis as a feature in a prediction framework — we perform regression to determine a predicted measure of participation. The better the predicted measure, the more likely the hypothesis

is influencing participation.

There can be several ways to qualitatively validate the proposed hypotheses including via ethnographic studies. In this work, however, we adopt a quantitative prediction approach. Prediction of observed participation based on the different hypotheses helps us understand the motivation for an individual to come back to the different types of social media sites. There are two technical contributions of this paper, both related to testing of hypothesis via prediction of observed participation:

- We use a Support Vector Regression based prediction framework to evaluate each hypothesis. Specifically, we test the ability of a factor, including intrinsic and extrinsic factors, to explain observed participation, including newcomers and existing participants.
- We propose a Bayesian Information Criterion (BIC) metric to identify the optimal factor combination. A combination of factors may better explain collective participation.

Main Results

We tested our hypotheses on two dataset classes — two rich media datasets, Flickr and YouTube, and two blog forum datasets, Engadget and Huffington Post. Our results indicate that different factors influence conversations from the two data classes differently. On one hand, extrinsic network factors, including media context and conversational interestingness, explain participation on rich media conversations. On the other hand, intrinsic network factors, including social awareness and community characteristics seem to explain participation on blog forums. We show that shared media contextual attributes, including visual and textual content, tags etc., influence newcomer participation. In contrast, existing participants seem to rely more on the characteristics of the community for continued participation.

Testing of hypothesis combination also yields insights. Interestingly, we find that including *all* of the intrinsic and extrinsic network factors does not yield the best prediction accuracy. Instead, we note that the optimal combination of factors

improves prediction accuracy significantly by $\sim 9\text{--}13\%$ and $\sim 8\text{--}11\%$ respectively over using just the best factor and all factors. Specifically, we observe that: (a) for rich media conversations, a combination of the extrinsic network factors quantify participation better; and (b) in the case of blog forums, a combination of the intrinsic factors perform the best. This reveals that there is likely to be a complex set of factors responsible for the nature of participation observed on different social media conversations today.

The rest of this paper is organized as follows. In section 2 we discuss prior work. Sections 3, 4 and 5 present the datasets, the different factors behind participation and the framework used to predict it. In section 6 we present our experiments involving the validation of the different hypotheses or factors relating to participation. We test the impact of combining multiple hypotheses in section 7. We conclude with a discussion of some open issues and our contributions in sections 8 and 9 respectively.

9.2 Related Work

Over several years, sociologists have been interested in understanding individual participation that underpins social movements. Dixon et al. (Dixon and Roscigno, 2003) considered aggregate network processes that may condition the costs and benefits of participation in social movements. Recent work on understanding participation over the Internet has focused on factors associated with continued contribution of individuals on newsgroups, discussion forums, and online communities and networks (Beenen et al., 2004; Koh et al., 2007; Nov et al., 2010,?). In the following two paragraphs, we organize the related work towards understanding participation in communities and rich media repositories. This aligns with the organization of the analysis presented in this paper.

Participation in Communities. Lampe et al. (Lampe and Johnston, 2005) examined the participation of users on the technical community Slashdot and substantiated three explanations for participation. They were: learning transfer from previous

experiences, observation of, as well as feedback from other participants. Joyce and Kraut (Joyce and Kraut, 2006) studied the factors behind participation in news-groups. Specifically they considered whether the response received on the first post of an newcomer motivated them to further participate in the communities and observed that the emotional tone of the feedback received had little affect on the individual's motivation to post again. In the context of social networks and social media, Burke et al. (Burke et al., 2009) studied content contribution on Facebook. They predicted how mechanisms such as social learning, singling out, feedback, and distribution were able to quantify long-term sharing of content (e.g. photos) based on their experiences in the first two weeks.

Participation in Rich Media. In the context of the rich media site Flickr, Nov et al. (Nov et al., 2010) studied how the tenure in a community affects a variety of participation types for individuals. Negoescu et al (Negoescu et al., 2010) adopted a human-centered approach to study two photo-sharing communities: Flickr and Kodak Gallery. They identify differences and similarities in the usage of photo tags and thereafter propose a topic model to provide further insight into the differences between the two communities. Finally, Miller et al (Miller and Edwards, 2007) studied photography practices, privacy perspectives and socialization styles on Flickr.

Limitations of Prior Work: The state-of-the-art has made significant contributions to understanding factors behind voluntary participation in physical and online communities (e.g. open source forums, Wikipedia, Facebook). A key property of these online communities is the following: there are clear incentives behind an individual's participation in the discussion forum, in editing a Wikipedia article or posting/tagging a photo on Facebook. The incentives could range from contribution to an open source project (in open source forums), generating knowledge (in Wikipedia) or merely the desire to remain posted with one's real world social ties (in Facebook). Therefore, from the prior literature we gain the insight that participation can in these contexts be explained by considering intrinsic factors within the social network. Such factors include, the awareness of a participant to feedback/responses from her peers or her familiarity with the peers in the past.

However, prior research has not investigated participation in the context of the conversations in rich media around which a social network evolves. It is natural to conjecture that *a combination of factors*, such as awareness of the individual to feedback from peers, community behavior as well as conversational interestingness is likely to impact participation. Understanding such factors that influence participation has not received sufficient attention in prior work. Addressing these concerns is a major focus in this work.

9.3 Social Media Conversations

In this section, we first describe several key social media conversation features, and then provide an overview of the datasets used in this paper.

Conversations

Social media conversations possess unique characteristics. These features of social media conversations are different from online forum discussions, where user participation has been typically investigated. The key features of conversations include: community, presence of shared media and conversational interestingness.

Community. Shared media conversations can promote cohesive interaction amongst community members. Members of the community can join a specific conversation due to several reasons. First, individuals can come together because they share a common interest in the topic. Second, individuals may be interested in expressing their opinion on a media object related to a recent event. Finally, they may be interested in exchanging ideas with familiar community members, whom they observe participating in the conversation. Conversations also provide a unique framework for individuals to observe peer activity around a specific topic of interest. Such observations can thus be used by a participant to infer characteristics of the larger community around the conversation, that is interested in discussing the particular topic (?). These community characteristics may include the following: community size, community cohesiveness, including dense cliques, whether there is sustained

participation, and if the community can attract and retain new participants. Thus, an individual’s observations of the larger community is likely to influence her participation in a conversation.

Shared Media. Social media conversations take place in the context of a shared media object, including a video on YouTube, or a post on the technology blog, Engadget. Naturally, the content of the media object—e.g., visual features of an image/video, textual content of a blog post is likely to impact an individual’s desire to participate in the associated conversation. Additionally, each media object typically engenders several contextual attributes, which can attract an individual’s attention as well (Kennedy et al., 2007), prompting her to post a comment on the conversation. These contextual attributes includes tags, ratings, recency of the media object etc. Hence, analysis of factors behind voluntary participation in these conversations needs to consider the shared media context, including its content.

Conversational Interestingness. Temporal theme evolution is a key characteristic of social media conversations. New themes slowly emerge due to new user comments, and over time, the conversation topic can bear little resemblance to the original conversation topic (De Choudhury et al., 2009c). In this way, certain themes can emerge to be highly popular. The theme popularity affects the participants who comment in such themes: the participants become important in the context of the conversation. In (De Choudhury et al., 2009c), the authors operationalize temporal evolution of a conversation by the “interestingness” measure of the conversation. We conjecture that the degree of interestingness of a conversation, influences individual participation.

We also consider additional factors: an individual’s awareness of peer activities and the reputation and tenure of the media creator in the social network. Such factors are motivated by observations from prior research regarding online participation—self-development, enjoyment (Nov et al., 2010), reputation (Nov et al., 2010), feedback and attention (Lampe and Johnston, 2005; Koh et al., 2007; Burke et al., 2009).

Table 9.1: Description of conversations on different rich media and blog datasets.

DATASET	MEDIA	CONVERSATION
Rich Media Datasets		
YouTube	Video	Comments on the video
Flickr	Photo	Comments on the photo
Blog Forum Datasets		
Engadget	Blog post	Comments on the blog post
Huffington Post	Blog post	Comments on the blog post

Datasets

A key goal in this work is to understand the factors affecting collective participation in different types of social media conversations. Social media conversations take place under a variety of contextual conditions. We identify two different conversational contexts: conversations centered around a shared rich media object (image, video) and conversations centered around shared textual content, including blogs. We utilize two datasets from each of the two categories—two rich media websites, Flickr (<http://www.flickr.com/>) and YouTube (<http://www.youtube.com/>), and two blog forums, Engadget (<http://www.engadget.com/>)² and Huffington Post (<http://www.huffingtonpost.com/>)³.

In Table 9.1, we provide a summary definition of a conversation in each of the above datasets, along with the shared media object around which we intend to quantify the collective participation.

We describe the details of each dataset in Table 9.2. All of the datasets were crawled for research purposes using their respective APIs. Additionally,

²Engadget is a technology weblog and podcast about consumer electronics. The blog is usually updated multiple times a day with articles on gadgets and consumer electronics, typically by an Editor. It also posts rumors about the technological world, frequently offers opinion within its stories, and produces profuse commentary from registered users centered around the stories.

³Huffington Post is an American news website and aggregated blog featuring various news sources and columnists. The site offers coverage of politics, media, business, entertainment, living, style, the green movement, world news, and comedy, and is a top destination for news, blogs, and original content. Huffington Post has an active community, with over one million comments made around the posted blog stories.

the time spans of these crawled datasets have been chosen to be of approximately the same length (~ 147 days) and are given as follows. YouTube: Sep 1, 2008–Jan 31, 2009; Flickr: Feb 1–Jun 30, 2008; Engadget: Apr 1, 2008–Aug 31, 2008; and Huffington Post: May 15, 2008–Oct 10, 2008.

Table 9.2: Details of the four datasets.

Dataset	#Participants	#Conversations	#Comments
Rich Media Datasets			
YouTube	17,736,361	272,810	145,682,273
Flickr	4,304,525	305,258	26,557,446
Blog Forum Datasets			
Engadget	78,740	45,073	6,580,256
Huff Post	59,282	24,479	4,748,837

9.4 Factors in Social Participation

There are several factors that can affect the degree of participation in social media conversations — both for the newcomers and existing participants. We categorize them as intrinsic and extrinsic network factors. Factors *intrinsic to the network* include social awareness, community characteristics and creator reputation. Factors *extrinsic to the network* are features related to media context and conversational interestingness. In the rest of this section, we propose several features for intrinsic and extrinsic factors.

Intrinsic Network Factors

Social Awareness. Participation of individuals in social media conversations is dependent upon factors that induce social awareness in an individual. The factors that influence social awareness in a conversational setting in digital communities have been studied in prior literature (Lampe and Johnston, 2005; Joyce and Kraut, 2006; Koh et al., 2007; Burke et al., 2009; Nov et al., 2010). We utilize three measures of social awareness:

- *Familiarity:* We quantify the degree of *familiarity* with participants with re-

spect to an individual associated with a conversation to be a function of the number of times they co-participated in any prior conversation on the same topical category. For the n -th conversation, familiarity $F_K^{(n)}$ at time slice t_K is thus given by the ratio of the mean frequency of co-participation of every participant u with every other participant v on prior conversations, to the mean frequency of participation of all participants in all prior conversations between t_1 and t_K .

- *Feedback*: Next we quantify the degree of *feedback* with participants with respect to an individual associated with a conversation to be a function of the number of replies she receives from other participants. For the n -th conversation, feedback $D_K^{(n)}$ at time slice t_K is thus given by the mean number of replies each participant in the conversation receives, to the number of comments / replies s/he has had posted until t_K .
- *Dialogue*: Presence of *dialogue* $L_K^{(n)}$ among the participants in the n -th conversation is given by the ratio of the frequency of all the replies to frequency of all the comments until t_K . It is therefore a measure of the overall back and forth communication (comment/reply) has happened between the participants in the past.

Hypothesis 1 *Collective participation on a social media conversation is affected by the degree of social awareness of the participating individuals, including their familiarity with other participants in the conversation, feedback from others and dialogue among others.*

Community Characteristics. Properties of the overall community also influence collective participation in conversations. We consider a community to be a set of individuals who engage in commentary centered around a broad topic. A typical community in our dataset, for example, on YouTube: a set of individuals who write comments or replies around shared videos on the topic of “News & Politics”.

Table 9.3: Media context on multiple rich media and blog datasets.

DATASET	MEDIA CONTENT FEATURES	MEDIA META-DATA
Rich Media Datasets		
YouTube	Visual features of the video— color (color histogram, color moments), texture (GLCM, phase symmetry), shape (radial symmetry, phase congruency) and keypoint location features (SIFT) (De Choudhury et al., 2009). These features are computed over a key frame in each video, where the key frame corresponds to the one at the median time of the duration of the video	Number of views ^a , number of ‘favorites’ ^a , ratings, number of linked sites, time elapsed since video upload (recency), video duration
Flickr	Visual features of the photo— color (color histogram, color moments), texture (GLCM, phase symmetry), shape (radial symmetry, phase congruency) and keypoint location features (SIFT) (De Choudhury et al., 2009).	Number of tags ^a , number of notes, number of views ^a , number of ‘favorites’ ^a , number of associated groups, time elapsed since photo upload (recency)
Blog Forum Datasets		
Engadget	tf-idf (term frequency-inverse document frequency) based features of the blog content; where the content is represented as a stemmed and stop-word eliminated bag-of-words	Number of tags ^a , time elapsed since blog was posted (recency), number of Facebook “likes” ^a , length of the post
Huff Post	tf-idf (term frequency-inverse document frequency) based features of the blog content; where the content is represented as a stemmed and stop-word eliminated bag-of-words	Number of tags ^a , time elapsed since blog was posted (recency), number of Facebook “likes” ^a , length of the post

^a Variable is log-transformed to correct for skew.

We consider different properties, structural and temporal, to characterize online communities: community size, community activity, community cohesiveness and community sustenance. We describe each of these characteristics below:

- *Community size* S_K at a certain time slice t_K is defined as the number of

unique individuals who have posted a comment or a reply at least once on all conversations associated with media objects belonging to a certain topical category.

- *Community activity* A_K at a certain time slice t_K is the mean degree of activity of the individuals in a community. It is given by the mean number of postings of comments and replies across all the individuals in the community.
- *Community cohesiveness* H_K at a certain time slice t_K is defined as the mean clustering coefficient of the communication graph. The graph is induced by the co-participation of individuals commenting or replying to all conversations associated with media objects belonging to a certain topic. Notice that the communication graph is an undirected weighted graph $G(V, E; t_K)$ where the nodes V are the individuals who have posted a comment or a reply at least once on all conversations associated with media objects belonging to a certain topical category until t_K . An edge $e \in E$ exists between two individuals in V if they have commented/replied together (i.e. co-participated) on the same conversation belonging to the topical category at least once until t_K . The weight on the edge is proportional to the mean frequency the co-participation between t_1 and t_K .
- *Community sustenance* U_K at a certain time slice t_K is defined as the mean degree of retention of communicating individuals over time. Sustenance is a function of the number of individuals who repeatedly return to the community over time to post comments / replies on conversations belonging to the particular topic. For a community C_K at a certain time slice t_K sustenance is defined as follows:

$$U_K = \frac{1}{K-1} \sum_{m=1}^{K-1} \frac{|C_K \cap C_m|}{|C_K|}, \quad (9.1)$$

where C_m is the community at time slice t_m .

Hypothesis 2 *Collective participation on a social media conversation is affected by the characteristics of the larger community, including its size, how active and*

cohesive its members are, and to what degree it is able to sustain its members over time.

Creator⁴ Reputation. Since social media conversations are typically centered around a media object, the identity or characteristics of the creator is likely to play an important role in the communication. Let the reputation be defined as the one-dimensional vector $\mathbf{R}_K^{(n)}$ corresponding to the n -th conversation at time slice t_K . The vector contains measures of the following attributes: number of media objects uploaded / posted by the individual until t_K , his or her number of (social) contacts in the community until t_K (if applicable), i.e. his or her authority measure in the network⁵, and the duration of his or her ‘tenure’ i.e. the time elapsed until t_K , since the date s/he joined the website (Nov et al., 2010).

Hypothesis 3 *Collective participation on a social media conversation is affected by the reputation of the creator of the associated media artifact, including his or her activity in media creation, his network authority score and tenure in the larger community.*

Extrinsic Network Factors

Media Context. As mentioned earlier, a distinct feature of participation on social media conversations is that it takes place around a shared media object. Hence the media context is also useful in analyzing the degree of collective participation over time. We consider two kinds of collective participation media contexts: the visual/textual content (features) of the media object, and media meta-data. A detailed description of the two different aspects of the media context is described in Table 9.3. Corresponding to the n -th conversation and at time slice t_K , we therefore assume that the visual / content features are denoted as $\mathbf{V}_K^{(n)}$ and the meta-data features as $\mathbf{M}_K^{(n)}$ — note that both of these aspects are one-dimensional feature vectors.

⁴A creator is simply the individual who uploads a video on YouTube, shares a photo on Flickr or write blog posts on Engadget or Huffington Post.

⁵Variable is log-transformed to correct for skew.

Hypothesis 4 *Collective participation on a social media conversation is affected by the context associated with the media artifact, including its visual or textual content as well as media meta-data, including its ratings, views, tags and recency of upload.*

Conversational Interestingness. A typical aspect of social media conversations is that they engender communication around the shared media spanning a variety of external events. As a result, *we conjecture that collective participation will be significantly affected by the evolving nature of the conversation itself.* We consider a subjective temporal property of the conversations: known as “interestingness”. We utilize the interestingness model proposed in (De Choudhury et al., 2009c) to compute this measure as a real scalar value in the range $[0,1]$. Interestingness of a conversation at any given time depends on its themes (popular themes featured in a conversation are likely to make it interesting to individuals and facilitate participation); and also the prior communication activity of its participants.

Since conversational interestingness is a new feature, we briefly review its calculation. Specifically, the authors (De Choudhury et al., 2009c) propose the following steps. First, conversational themes are detected using a temporally regularized mixture of multinomials model. Second interestingness of participants and interestingness of conversations are determined based on an one-dimensional random walk model. Finally, a joint optimization framework of interestingness is used to effectively compute interestingness, that incorporates temporal smoothness constraints. In this work, we denote this interestingness measure of the n -th conversation at time slice t_K as $I_K^{(n)}$.

Hypothesis 5 *Collective participation on a social media conversation is affected by the characteristics of the conversation itself, such as its interestingness over time, where interestingness is characterized by the popularity of the conversational themes and the communication properties of the participants around those themes.*

We now provide a brief summary of the various factors behind participation that have been proposed in this section. We proposed a number of intrinsic and extrinsic network factors that are likely to impact participation of newcomers and

existing participants in social media conversations. These include three intrinsic factors: social awareness, community characteristics and creator reputation. While the two extrinsic factors are: media context and conversational interestingness.

9.5 A Prediction Framework

In this section, we propose a prediction approach to evaluate each hypothesis in explaining observed participation. In particular, we are interested in explaining participation for newcomers and existing users, in different types of social media conversations (see Figure 9.1). The goal is to identify the sets of factors which are more effective (i.e. higher accuracy in prediction) in accounting for the observed participation in social media conversations.

Our prediction framework uses a learning framework, that regresses over past degrees of participation using the various factors impacting participation (by treating them as features). Then it predicts the measure of participation at a future point in time by using the best fit coefficients. In this work, we utilize an incremental Support Vector Regression model (Burgess, 1998) to predict the degree of observed participation, that can be attributed to each of the five different types of factors.

We begin with by constructing our “ground truth” for quantifying the influence of each type of factor towards newcomer and existing user participation. Let $N_K^{(n)}$ be the number of comments that are generated on the n -th conversation at time slice t_K , by individuals who had not posted any comments (or replies) on the same conversation between t_1 and t_{K-1} . Furthermore, let $E_K^{(n)}$ be the number of comments that are generated on the same n -th conversation at time slice t_K , by individuals who had posted at least one comment (or reply) on the same conversation between t_1 and t_{K-1} . In the same way, we determine the degrees of participation over all time slices between t_1 and t_K . Without loss of generality, let us denote a participation vector to be $\mathbf{Y}_{1:K}^{(n)} \in \mathbb{R}^{K \times 1}$.

Next we define five different feature sets, corresponding to the five categories of factors discussed in the previous sub-section. Let, for the n -th conversation, $\mathbf{f}_K^{(n)} \in \mathbb{R}^{1 \times d}$ denote the feature vector corresponding to any of these five categories

at time slice t_K ; d being the number of features (or dimensionality) within the chosen category. The feature vectors can be similarly constructed for all time slices from t_1 to t_{K-1} . Let us represent the matrix of the feature vectors for all time slices between t_1 and t_K as $\mathbf{X}_{1:K}^{(n)} \in \mathbb{R}^{K \times d}$.

We use the data over the first p time slices (where $p < K$) to predict the number of comments from newcomer and existing participants. We split the ground truth vector (or the dependent variable) $\mathbf{Y}_{1:K}^{(n)}$, as well as the feature matrix (or the independent variables) $\mathbf{X}_{1:K}^{(n)}$ into training and testing sets. The first p slices form the training set, while the remaining $p+1$ to K time slices are the test set. The training phase of the SV Regression model (based on a Gaussian RBF kernel), gives us the support vectors, and the best-fit regression coefficients. These coefficients are thereafter applied on the test set over time slices $p+1$ to K to get the predicted measures of participation over time slices $p+1$ through K .

The effectiveness of the chosen feature set category is therefore given by the mean percentage accuracy in predicting the value $\hat{N}_i^{(n)}$ and $\hat{E}_i^{(n)}$ against the actual values $N_i^{(n)}$ and $E_i^{(n)}$ for all time slices $p+1 \leq i \leq K$. The accuracy measure is given as the ratio of the absolute difference between predicted and actual values to the actual value of each type of participation (the number of comments from newcomer and existing participants).

9.6 Validating Hypotheses

We conduct elaborate experimental studies on all the four datasets introduced in section 3, in order to find empirical grounding on the five different hypotheses behind collective participation proposed in this paper. For all the four datasets, we choose the first 97 days ($\sim 65\%$) as the training phase and the next 50 days ($\sim 35\%$) as test set in each case. We avoid using larger training set sizes to prevent overfitting in the prediction task.

Prediction Performance

We begin by presenting prediction performance of using different feature set categories in accordance with the different hypotheses framed in section 4. The performance is evaluated based on the corresponding percent accuracy metric (discussed in section 5.6) and we present the results for both dataset types, as well as for newcomer participation as well as that from existing participants (Figure 9.2).

Rich media vs. Blog Forums. We observe differences in the feature sets that yield the best prediction performance across the two dataset types. For rich media data, extrinsic network factors (media context and conversational interestingness; mean accuracy $\sim 80\%$) seem to better predictors of participation compared to social awareness and community characteristics. This is because the nature of the shared media is central to triggering users to participate in conversations. For blog forums data, intrinsic network factors (social awareness and community characteristics; mean accuracy $\sim 78\%$) seem to better predictors of participation compared to the others. This is because participation on these websites are often driven by personal opinions on technology or political happenings. Hence the overall community’s response and behavior to a certain event are likely to be important factors behind participation.

Newcomers vs. Existing Participants. There are also significant differences across the factors that affect participation in newcomers and existing participants. Conversational interestingness and community characteristics perform relatively better for all datasets in the case of existing participants. This is because over time they are able to ‘learn’ a community’s dynamics: its nature of activity as well as can judge better (via comparison) the interestingness of the on-going conversations. Newcomers seem to rely more on media context and social awareness. This is because their participation is likely to be triggered by the popularity of the media shared, or by how the rest of the participants are reacting to their comments.

Creator Reputation. The creator reputation feature does not explain collective participation well for any of the datasets (accuracy $\sim 49\%$). We believe that there

are two explanations: large number of authors in rich media sites and restrictive media authorship in blogs. Anyone can upload a media object in rich media websites. Since there is no restriction on *who* can upload—the number of creators on rich-media sites is very large. This overabundance of creator choice, in rich media sites, makes the creator a less likely candidate as the *sole* attribute on which to filter media. On the two blog forums analyzed in this work, only a fixed number of Editors can create content. Since all the content on these two blogs are created by editors, the reputation of the editor makes little difference to user participation.

Statistical Significance

From the results in Figure 9.2 we observe that there are differences in prediction performances for the different feature set categories, or hypotheses and the datasets. In order to substantiate the differences, we perform some tests of statistical significance (one-tailed paired Student’s *t*-test) on the prediction accuracy measure of each hypothesis, including both for newcomers and existing participants (ref. Figure 9.2). In particular, we are interested to investigate if the performances of the “best” hypotheses are statistically significant compared to that of the others.

We consider all the conversations in each dataset. Our experimental setup consists of a one-tail paired *t*-test that compares the prediction accuracies of the two best performing methods/hypotheses for each dataset (obtained from Figure 9.2), with that using each of the other methods. Our null hypothesis is that the accuracy measures are sampled from the same distribution and hence the distribution of accuracies over the two methods under consideration would have similar means and therefore account for little differences. We predict that for the two best performing methods in the case of each dataset, the null hypothesis will be false, because the differences in performances of these two methods against others are significantly better.

The measures of the *t*-statistic and the corresponding *p*-values for each of the comparisons is given in Table 9.4. We use the following abbreviations here — SA: Social Awareness, CC: Community Characteristics, CR: Creator Reputation,

Table 9.4: Statistical significance (based on one-tail paired t -test; significance level of 0.05) of the two best performing factors for each dataset, compared to the other features, for same datasets. Here, SA: Social Awareness, CC: Community Characteristics, CR: Creator Reputation, MC: Media Context and CI: Conversational Interestingness. p -values below the significance level are shown in italics. We observe that in each dataset, the performances of the “best” factors are statistically significant compared to the other factors. Note that because we used a one-tail paired t -test, the significance results of X & Y is different from Y & X.

	t	p		t	p
RICH MEDIA DATASETS					
YouTube: df= 272,810; best performing factors: CI, MC					
SA & CI	-926.3	<i>0.013</i>	SA & MC	-909.1	<i>0.028</i>
CC & CI	-937.9	<i>0.007</i>	CC & MC	-924.6	<i>0.016</i>
CR & CI	-959.7	<i>0.001</i>	CR & MC	-951.7	<i>0.002</i>
MC & CI	-892.3	0.116	CI & MC	-884.4	0.204
Flickr: df= 305,258; best performing factors: CI, MC					
SA & CI	-981.5	<i>0.011</i>	SA & MC	-914.5	<i>0.031</i>
CC & CI	-1035.1	<i>0.006</i>	CC & MC	-1052.6	<i>0.014</i>
CR & CI	-1263.3	<i>0.001</i>	CR & MC	-1352.2	<i>0.003</i>
MC & CI	-835.4	0.121	CI & MC	-862.7	0.193
BLOG FORUM DATASETS					
Engadget: df= 45,073; best performing factors: SA, CC					
CC & SA	-318.4	0.147	SA & CC	-298.6	0.192
CR & SA	-451.3	<i>0.001</i>	CR & CC	-431.4	<i>0.002</i>
MC & SA	-405.7	<i>0.008</i>	MC & CC	-424.2	<i>0.007</i>
CI & SA	-362.8	<i>0.009</i>	CI & CC	-379.1	<i>0.008</i>
Huffington Post: df= 24,479; best performing factors: SA, CC					
CC & SA	-224.3	0.091	SA & CC	-183.6	0.075
CR & SA	-373.5	<i>0.002</i>	CR & CC	-278.3	<i>0.001</i>
MC & SA	-285.8	<i>0.006</i>	MC & CC	-237.3	<i>0.003</i>
CI & SA	-194.7	<i>0.003</i>	CI & CC	-207.5	<i>0.002</i>

MC: Media Context and CI: Conversational Interestingness. Notice that the since the test is *one-tail*, the results are not symmetric. Consider, as an example, the attributes CI and MC. In the one-tailed case, comparing MC to CI will be different from the result from comparing CI to MC.

We show the results for the two rich media and two blog forum datasets. In the case of YouTube and Flickr, the results reveal that the best performing methods,

Table 9.5: Summary of results in prediction of collective participation by newcomers and existing participants. Here SA: Social Awareness, CC: Community Characteristics, CR: Creator Reputation, MC: Media Context and CI: Conversational Interestingness.

	Support–Rich Media	Support–Blog Forums
<i>Newcomers</i>		
SA	Less (-33%)	<i>High</i> (-11%)
CC	Less (-37%)	Moderate (-20%)
CR	Minimal (-49%)	Minimal (-47%)
MC	<i>High</i> (-19%)	Less (-34%)
CI	Moderate (-28%)	Less (-31%)
<i>Existing participants</i>		
SA	Less (-35%)	Moderate (-24%)
CC	Less (32%)	<i>High</i> (-15%)
CR	Minimal (-48%)	Minimal (-51%)
MC	Moderate (-26%)	Less (-36%)
CI	<i>High</i> (-18%)	Less (-33%)

i.e. CI and MC yield p -values below the significance level of 0.05, with respect to SA, CC and CR. The same is true for the two best performing methods SA and CC for Engadget and Huffington Post. Consequently, we reject the null hypothesis that there is no difference in the different factors.

Summary of Findings

We conclude that *not* all the stated hypotheses are able to quantify the observed participation equally well. There are differences across the two classes of dataset, rich media and blog forums, as well as between the participation from newcomers and existing participants. We summarize these findings in Table 9.5. For the purpose of easy comprehensibility, we indicate how much *support* each hypothesis provides towards quantifying the participation from newcomers and existing participants separately. Support is defined as the negative of the error—the difference between the predicted accuracy and the ground truth observed participation. At zero error, we have the highest support. We define following terms for the support (S): High ($-20 \leq S \leq 0$), Moderate ($-30 \leq S \leq -20$), Less ($-40 \leq S \leq -30$), and Minimal ($S \leq -40$).

9.7 Combining Multiple Hypotheses

Collective participation in online media will typically be manifested due to a collection of factors, rather than a single factor, as discussed in the previous section. In this section, we therefore use a Bayesian Information Criterion (BIC) based measure to determine the optimal set of factors to explain collective participation.

BIC Measure

Our goal is to determine an “optimal” number of factors, typically smaller than the set of all factors, that can best explain the observed participation. This problem can be reduced to a model selection problem, where we fit a model in a learning task, with a number of parameters. In our case, the different factors can be considered as the model parameters, and we are interested in identifying the optimal parameter combination that helps explain the observed participation. We utilize a measure frequently used in model selection—known as Bayesian Information Criterion (BIC) (Chen and Gopalakrishnan, 1998) to find the optimal factor subset.

We develop an iterative approach to determine the optimal hypothesis combination using the BIC measure. We start with a random hypothesis, and sequentially add hypotheses to it. The feature vector corresponding to the chosen starting hypothesis is used to predict the collective participation (of newcomers and existing participants) using the Support Vector Regression technique discussed in Section 5.6. Using the prediction error, we then compute the BIC measure of the combination at the current iteration. Then at the next step we add a hypothesis. The optimal hypothesis added at each iteration, is the one that *minimizes the Bayesian Information Criterion (BIC)* measure (Chen and Gopalakrishnan, 1998) in the combination, for predicting participation. This procedure terminates when there are no more hypothesis are left to be added.

This procedure is repeated for different starting seeds. The particular combination that yields the minimum BIC value overall is chosen as the optimal hypothesis combination.

Results

We present the results of combining hypotheses to predict collective participation in Figure 9.3. The figure has two parts. In the top part, we show a visual representation of which hypotheses were chosen at each iteration for each starting seed hypothesis. This is shown using linear paths between the hypotheses (each path has a different starting seed hypothesis). That is, in the figure, in the prediction of newcomer participation for YouTube, at iteration 3, for the starting seed hypothesis CI, we have the optimal combination as {CI, MC, SA}, shown with the green dotted path. Next in the bottom part of the figure, we show the BIC value of each hypothesis combination at each iteration (shown in a line plot with the same color and style as the corresponding path in the top part).

The results indicate that combining hypotheses does *indeed* appear to improve the prediction of collective participation for both newcomers and existing participants. It appears that the combinations that perform the best are the ones which have the starting seed as the best performing hypothesis in Figure 9.2. However, surprisingly enough, using all information in terms of all five hypotheses *does not* yield the best prediction. In fact the best performance, as seen in the BIC curves in Figure 9.3 are given by hypotheses combinations in the middle of the curve—that is, a selective few hypotheses, on combination, quantify collective participation in the best manner.

Table 9.6: Summary of results of combining hypotheses in prediction of collective participation.

Dataset	Newcomers	Existing Participants
YouTube	{MC, CI, SA}	{MC, CI, CC}
Flickr	{MC, CI, SA}	{MC, CI, CC}
Engadget	{SA, CC, MC}	{SA, CC, CI}
Huff Post	{SA, CC, MC}	{SA, CC, CI}

A summary of the best performing combinations is shown in Table 9.6. We also present in Table 9.7. the prediction accuracies for these best performing combinations and compare them to those of using just the best performing hypothesis (ref. Figure 9.2) and the combination of all five hypotheses. The best combination

improves prediction accuracy significantly by $\sim 9\text{--}13\%$ and $\sim 8\text{--}11\%$ respectively over using just the best hypothesis and all hypotheses.

Table 9.7: Prediction accuracies using (I) just the best performing hypotheses (Figure 2), (II) optimal hypotheses combination (Figure 3), and (III) all five hypotheses.

Dataset	Newcomers			Existing participants		
	I	II	III	I	II	III
YouTube	79%	88%	80%	80%	92%	81%
Flickr	82%	92%	82%	80%	91%	82%
Engadget	76%	89%	78%	76%	87%	76%
Huff Post	83%	93%	82%	81%	90%	80%

The combinations that work best (see Table 9.6), for example, for rich media are ones which utilize extrinsic network factors: MC and CI. For blogs, intrinsic network factors SA and CC play a key role. For newcomers MC and SA are important across both blogs and rich media, while for existing participants, CC and CI are key across all datasets. As before, the creator reputation least affects the participation measures. Hence its inclusion make the prediction worse by increasing the BIC of the combination.

It is reasonable to conclude from these experiments that collective participation on social media conversations are guided by *a complex set of factors*. However, different factors dominate depending on the type of the site: rich media site participation depends more on the properties of the conversation itself, while the blog forums participation are guided by the social attributes including awareness and community behavior.

Open Issues

We now discuss some of the open issues in this paper. These include the puzzling lack of influence of creator reputation, incompleteness of factors, and the variety of participatory mechanisms.

Social media conversations featuring on the social Web today exhibit a variety of underlying norms and principles on the part on its associated individuals who are involved in communication. These aspects drive the larger community's sus-

tained growth and participation over time. How do we characterize these underlying factors that motivate individuals to participate in community-centric conversations? How can the design of new social media sites, facilitating user communication, benefit from a consideration of these factors? How can the experience of a new member in a social media community be made more conducive from the standpoint of encouraging her to converse? And finally, does the type of media shared on a social media site affect which factors are more likely to impact greater participation in conversations? Our work in this paper has been motivated from the above questions. By dint of characterizing a number of factors behind participation in conversations, we have been able to utilize a prediction framework and determine what type of social media sites benefit from what kind of participatory factors.

We observed with some surprise that the factor relating to creator reputation barely influenced collective participation of individuals. There may be several reasons why this was so. First, since these communities are very large, it is likely that there is little awareness of the content creator identity. This is in contrast to smaller well-knit communities, where individuals are aware of media creation activities of fellow community members. Peer awareness, can become one important variable in development of creator reputation. Finally, while we had used a measure of reputation motivated from prior work (Nov et al., 2010), it may be worthwhile investigating the measure carefully, and developing new measure(s), that can contain additional factors.

It is possible that we have not exhaustively examined the set of factors influencing participation. Unobserved variables, including participant demographics, gender, age, location and cultural norms may also affect participant behavior. Additionally, the participation behavior might evolve over time, in the case of the existing participants. Sentiment and writing style may also influence participation. Fleshing out more extensive factors driving participation remains a ripe area for future research.

Additionally, we have considered only one kind of participation on social media sites: posting comments on conversations. An individual may also participate

in other ways: individuals can participate by rating comments, sharing posts and comments of interest. We would be interested to see in future work, if our intrinsic and extrinsic factors can explain these other forms of participation.

Finally, participation on a social media website may also be affected by an individual’s intrinsic or idiosyncratic behavior. As an auxiliary method of validation of the factors considered here, ethnographic studies may also be conducted in the future to capture individual behavior that lie beyond the scope of these factors.

9.8 Conclusions

In this paper, we investigated several factors to explain participation in social media conversations. Investigating the factors allows us to understand the nature of the underlying social network, including network structure and evolution, and information roles, and influence propagation. Efficient design of social media sites is one potential application of our work.

Our approach was as follows. We first identified several key intrinsic and extrinsic network factors influencing social media conversations, distinct from both online forum discussions and other social networks. We identified three intrinsic factors: social awareness, community characteristics and creator reputation. Furthermore, we identified two extrinsic factors: media context and conversational interestingness. We developed one hypothesis for each factor to test the influence of the factor on individual participation. We developed a Support Vector Regression based prediction framework to evaluate each hypothesis, and a Bayesian Information Criterion (BIC) metric to identify the optimal factor combination.

Our approach addressed two limitations of prior work. First, prior work typically looked at intrinsic network factors affecting the awareness of the participant, including peer familiarity and peer feedback. While these factors are of value and explored in this work, prior work paid little attention to extrinsic network factors, including conversational dynamics and content. We incorporated both intrinsic and extrinsic factors in our work. A second difference is that that we investigated how a combination of factors influence participation on social media conversations

We presented three interesting findings. First, we showed that extrinsic network factors significantly affected conversations on rich media, while intrinsic network factors were a significant factor for blog forums. Second, awareness of responses and community feedback affected newcomer participation in a conversation. In contrast, the behavior of the overall community, including community support for cohesive participation, or ability of the community to sustain participation, could better explain existing participant behavior. Finally, interestingly enough, we showed that an optimal factor combination improved prediction accuracy of observed participation by $\sim 9\text{--}13\%$ and $\sim 8\text{--}11\%$ over using just the best hypothesis and all hypotheses respectively.

We plan to investigate several research directions in the future, including a careful analysis of creator reputation, increasing the number of factors that may influence participation, and examining other forms of participation in social media conversations.

Figure 9.2: Prediction accuracies (higher numbers are better) of collective participation in social media conversations over four different datasets; corresponding error bars are also shown to illustrate the deviations. Media context and conversational interestingness perform better for rich media datasets, while social awareness and community characteristics perform better for blogs. Creator reputation appears to perform relatively poorly over all datasets.

Figure 9.3: Performance of combining different feature categories (or hypotheses) in predicting collective participation. For each starting feature set, we show which feature sets were selected at each iteration, that minimizes the BIC. The plot at the bottom shows the actual BIC measures of the combinations at each step (lower BIC values are better). Here SA: Social Awareness, CC: Community Characteristics, CR: Creator Reputation, MC: Media Context and CI: Conversational Interestingness.

Part IV

Conclusions

Chapter 10

Summary of Contributions

We present a summary of the major contributions accomplished in this thesis. Our central hypothesis in this thesis has been that communication dynamics between individuals manifest themselves via three key aspects: the information that is the content of communication, the *social engagement* i.e. the sociological framework emergent of the communication process, and the *channel* i.e. the media via which communication takes place. Our approach to developing a comprehensive understanding of these aspects in this thesis has been essentially computational as well as empirical: we presented characterization techniques, developed computational models and finally discussed large-scale quantitative observational studies for each of these organizing ideas. This naturally closes the loop between design and engineering on one hand, and empirical measurement and modeling on the other hand.

We organize our contributions, impact and the major outcomes along the three primary parts presented in the thesis, i.e. information diffusion, evolution of social communication networks and finally, rich media communication patterns (ref. Table 10.1).

- *Information Diffusion.* We have developed characterization frameworks of interpersonal communication processes that results in the flow of information between a pair of individuals. Our experiments on large datasets indicated that while the probability of communication between two individuals is dependent significantly on the user context at hand, the delay in the propagation of information appears to be more of a habitual property of the users themselves. Next we have attempted to develop computational models to predict the diffusion of actions on social media, that are capable of generating large-scale information cascades. We considered two facets that are likely to impact

Table 10.1: Thesis structure and research problems.

Information Diffusion	Topical Communication Flow (Chapter 3), Diffusion of User Actions (Chapters 4, 5)
Evolution of Social Communication Networks	Community Dynamics and External Phenomena (Chapter 6), Prototypical Communication Groups (Chapter 7)
Rich Media Communication Patterns	Interestingness of Conversations (Chapter 8), Quantifying Collective Participation on Social Media Conversations (Chapter 9)

diffusion of user actions—network factors like social influence, and individualistic factors like attribute homophily. Along the first direction, we have had excellent results on large social media datasets, that indicated that our proposed model can identify in advance emergent cascades due to diffusive nature of social actions with low error rates ($\sim \leq 15\%$). In the second direction, we conducted large-scale quantitative observational studies on the impact of individual attributes (such as location, information roles and activity behavior) on information diffusion process. Experiments indicated that similarity among individuals, that is, the homophily principle can explain diffusion process significantly well for a certain genre of topics, such as Politics and Sports, communication on which are highly affected by external world events.

- *Evolution of Social Communication Networks.* In this part of the thesis, we have studied the relationship of dynamics of the social engagement i.e. the groups and networks that individuals belong to, with external temporal happenings, like stock market movement dynamics as well as political sentiments on different candidates expressed via polls. Our experiments on large blog and social media datasets have indicated that suitably chosen communication features of individuals (e.g. responsivity and strength of ties) can characterize groups that exhibit considerable predictive power of future occurrences i.e. stock market dynamics and political polls. We have also developed a compre-

hensive model to understand the evolution of such technological and political groups via communication properties of individuals and have justified their utility via a novel evaluation framework, i.e. their correlation to stock movements of technology companies as well polls on the Presidential candidates in 2008 US Elections.

- *Rich Media Communication Patterns.* In the last part of this thesis, our goal has been to understand the communication patterns that happen centered around shared rich media artifacts, such as photos (Flickr) and videos (YouTube). First, we characterized the “interestingness” property of conversations that take place via comments on YouTube videos using a probabilistic framework that utilizes the properties of the conversational themes and the intrinsic interests of the participants over time. We tested our proposed framework on a large corpus of YouTube comments and videos to find that interesting conversations are often consequential i.e. they affect the social network that are associated with, in terms of degree of participation, thematic propagation of interestingness as well as cohesiveness of participants in mutual dialogue in the future. Secondly, we developed a probabilistic recommendation framework to understand the factors that impact individual participation on social media conversations. Our experiments indicated that the optimal factors often depend on the nature of the dataset; and also there is usually a complex set of factors of various genres that impact participation.

Based on all these outcomes, we believe that this research can make significant contribution into a better understanding of how we communicate online and how it is redefining our collective sociological behavior. Beyond exploring new sociological questions, the collective modeling of automatically measurable interactional data will also enable new applications that can take advantage of knowledge of a person’s social context or provide feedback about her social behavior. Communication modeling may also improve the automated prediction and recognition of human behavior in diverse social, economic and organizational settings. For collec-

tive behavior modeling, the social network can define dependencies between people's behavior with respect to their communication patterns, and features of the social network may be used to improve prediction and recognition. Additionally, some of the statistical techniques developed in this thesis for analyzing interpersonal communication may find new application to behavior modeling and machine learning.

Chapter 11

Future Directions

In the future, I believe the work presented in this dissertation will have impact on real world applications in a number of ways; and in turn will open up opportunities towards other interesting new directions in the broader domain of computational social science. My future research goal is therefore to harness large-scale networks to understand, predict, and ultimately, enhance social and technological systems. In this chapter, I organize my future research into short and long-term goals, as presented below.

11.1 Short-term Goals

My short-term goals are discussed as follows:

Multiplex Ties: Evolution and Empirical Thresholds on Ego-network Size. Today, online social media sites, such as Digg and Facebook provide multi-faceted interaction capabilities among users. For example, on Facebook, users can post content on each others' "Wall", write comments on a shared photo or video, join the same celebrity fan page or "like" each other's status update. Similarly on Digg, users can "digg" i.e. vote on a news article, comment and reply in the associated discussion on it, or subscribe to each other's activity feed via a contact relationship. All these modalities of interaction let users communication with each other in diverse ways, often depending upon the context at hand. When a user A intends to invite another user B to a weekend party, A might create an event page and invite B to RSVP on it; while on being back from the party, B might tag A on a photo that was taken at the party. These variegated interactional mechanisms reveal that the nature of relationship between two individuals is often a function of the context, and that it is likely to change over time due to the dynamics of context.

The most intuitive way to describe such time-evolving temporal and multi-relational ties among individuals is therefore, through a “multiplex” network. Multiplex networks are defined as the set of nodes that link to other nodes on the basis of more than one relation (Wasserman and Faust, 1994). Each type of relation can be treated as a separate network defined on the same set of nodes. The results networks can then be studied separately as a set of uniplex networks or collectively as a single multiplex network. The proposed work in this chapter studies the evolution of ties in these multiplex networks. We consider the ego networks of individuals, and our interest lies in empirically studying how these ego-networks change over time across multiple relations.

A consequential but orthogonal aspect related to evolution of ego-networks, is to understand the existence of thresholds on the sizes of the multiplex ego-networks. Note, that since ego-networks are likely to evolve and change over time, due to variations in the degrees of interactions of an individual with her contacts, at any given point in time, it is likely that she maintains consistent relationship with only a limited set of contacts. Hence the ego-networks of individuals for a given relation would exhibit evidences of a thresholded behavior. The ego-network size identified by this threshold thus refers to the maximum number of individuals with whom an individual can maintain social relationships by consistent social interaction over a period of time. In this chapter, we, therefore, further propose to investigate what are the expected sizes of the ego-networks for a given relation, and across relations, what are the inter-dependencies among the corresponding thresholds. A central proposition about multiplexity is that multiple types of network ties are interdependent, and ties in one network influence the formation or dissolution of ties in other networks (Lomi and Pattison, 2006; Robins et al., 2007; Galaskiewicz and Bielefeld, 1998). In other words, one network of interaction becomes the context for a set of other networks. Our goal in this regard is that, such investigation would lend interesting insights into the co-evolution of networks along multiples social ties over time.

We propose to seek answers to the following questions: (1) What are appro-

appropriate representations of multiplex networks based on multi-relational interactions on online social media? How do multiplex networks evolve over time, from an ego-centric perspective? (2) Are there empirical thresholds on the sizes of ego-networks corresponding to a given relation / modality? Do dependencies exist among the thresholds across these different relations? That is, does a certain relation affect the threshold for a different relation for a given ego-centric network of an individual? Can we model these inter-dependencies among the thresholds of multiple relations to develop a generalized model of evolution of ego-networks?

Large-scale Data and Evaluation Standards. Despite the immense potential today bolstered by the realm of available social data, and the range of computational tools, the field of computational social sciences is relatively new and there lie a lot of challenges ahead in the next few years. A primary challenge, for example, is the lack of suitable standards across researchers of various domains to share and exchange their findings on these datasets and tools. I would therefore like to develop suitable metadata standards that can promote efficient exchange of data and findings across researchers. To cater to this issue, I have been actively involved in public release of large-scale social datasets (Twitter/Digg etc.) that I have had collected or crawled. These avenues have already benefitted a host of scientists in their respective research problems, spanning both academia and the industry research labs¹. Transparency of data and methods in this manner is sure to help us develop a set of standards, metrics and evaluation methodologies in the coming years.

Understanding Emergent Order in Large Human Networks. A fundamental aspect of today's large-scale human networks evolving continuously around online interactions is that, they essentially exhibit substantial noise, or structural or semantic anomalies. However, in spite of such irregular perturbations in the network, these networks strongly feature the emergence of "order". Order can be described in terms of how the network allows consistent communities to form (e.g. the presence of consistent left and right wing communities on blog forums), semantic constructs to emerge

¹The released version of the Twitter dataset has had 100+ downloads already, since March 2010.

(e.g. “pwn” a new word extensively used in gaming communities that means “to own”), or how large groups of individuals reach a commonality in their culture of information exchange (e.g. ‘RT’ or re-tweet tag has culturally evolved over Twitter to indicate propagation of the same piece of information across individuals). How do we quantify such order amidst irregularity?

I believe answer to this question can be motivated from literature in ethnomethodology. This literature suggests that the fundamental question in sociology should not be a top-down study of social order (capitalism, socialism etc.), but instead how these social systems arise out of ordinary interactions between people—i.e. he suggested that the social systems were emergent artifacts of mutually observable human interaction. My key insight therefore is that, the order in these large online networks emerge through observable microscopic actions and interactions of the individuals. The network behaves as a closed system and eventually reaches a state of equilibrium wherein individuals develop a common sense understanding of structure, constructs and culture through the reciprocal actions. In the near future, I intend to formalize these concepts to model emergent order in large online networks.

In other words, I would like to create explanatory and predictive models of actions of large groups of people and societies, and biological and technological systems. Although the actions of a particular individual or component may be too difficult to model, computational methods can be applied to large groups or ensembles, which can yield effective models with the ability to predict the flow of future events. Based on our recent results and research experience, we believe that the study of large networks is a promising approach to developing such understandings, as graphs capture local dependencies, and also reveal large-scale structure and phenomena arising from the multitude of local interactions. Seemingly “random” local behavior can thus propagate to the macro scale where emergent global regularities and patterns emerge.

11.2 Long-term Goals

The interest of scientists to infer, understand and predict collective interactional behavior and deploy it in the design of systems or other practical applications is not new. What is new is the scale of opportunities brought to us today by the availability of massive amounts of online social data, spanning over millions of interacting individuals and often available upto the granularity of milli-seconds. Over several years, several theories of communication between individuals have been developed to help us understand how people behave and communicate in different ways to construct and negotiate a social reality. The same is yet to happen for our new forms of communications emerging online. My long-term vision encompasses devising generalized models and theories to account for our online interactional patterns. My prime motivation behind this goal is that, we have slowly but significantly deviated from our conventional modes of communication in the physical world. Contrasting and being able to quantify this distinction between physical and online interactions is important. First, it is because these interactions affect our decisions, modes of information consumption, search behavior and so on. Hence it can improve the automated prediction and recognition of human behavior in diverse social, economic and organizational settings. Second, the available genre of tools and methods, being mostly crafted for small-scale “snapshot”-based studies of a small number of individuals, do not naturally scale up to the domain of online interactional data. As a consequence, it is important to devise novel tools, methods and theories that can computationally model the online form of human interaction.

Note that one of my primary research agendas focuses on large scale data and computing architectures for massive social data manipulation and analysis. In the longer term, I also plan to explore map-reduce type programming abstractions for large scale computing and extend our methods to distributed architectures. The question here is what kinds of analyses are suitable for such architectures, and how to parallelize these computational methods to scale to thousands of machines. This line of research will allow us to perform near real-time analysis of planetary and

internet scale data and find patterns that are practically unobservable at smaller scales.

My future goals therefore encompass the broad vision of utilizing such profuse amount of Internet information to answer questions in social sciences and psychology in a more precise and scalable manner. There are two broad directions. (a) Deploying the knowledge of social processes in the context of understanding business, advertising and financial transactions occurring in social, collaborative as well as game-theoretic settings; and (b) Developing efficient scalable mining algorithms that can enable the discovery of legitimate semantic knowledge to address (a).

In closing, understanding the characteristics and dynamics of interactions over social networks and media is key to the design better usable interfaces, as well as build, maintain and allocate resources efficiently over large socio-technological systems. My academic training as well as the industrial research experience and collaboration in a trans-disciplinary setting have given me significant exposure to develop sound computational and empirical solutions to address research in this direction. At the same time they have enabled me to identify their practical applications in the real world via ethnographic experimentation and prototype development; parts of which have also found their way through to general audiences via press coverage. I strongly believe that summation of these skills and endeavors would prove to be immensely beneficial in accomplishing my future vision in the domain of computational social science.

Part V

Appendix

Notice that a key aspect of our research has been our empirical evaluations on large-scale online data, that has been collected (crawled) over time. Typically the datasets we have focused on are generalized enough, in the sense that the crawling process attempted to collect those individuals, their activity and their network who are representative of the network as a whole. However, these datasets are “samples”, and hence it is not well understood how the choice of a different sample would have impacted our results, or more precisely, to what extent; essentially raising the question whether our datasets are *biased*.

Sampling is thus a natural approach for learning about the various social systems using light-weight data collection, and had been relied on by most previous studies (for example, see (Adamic and Adar, 2005; Kennedy et al., 2007; Cha et al., 2008; Gjoka et al., 2008)). It is defined as a statistical practice concerned with the selection of a subset of individual observations within a population of individuals intended to yield some knowledge about the population of concern, especially for the purposes of making predictions based on statistical inference. The sampling process comprises several stages:

1. Defining the population of concern.
2. Specifying a sampling frame, a set of items or events possible to measure.
3. Specifying a sampling method for selecting items or events from the frame.
4. Determining the sample size.
5. Implementing the sampling plan.
6. Sampling and data collecting.

Prior research has focused on various aspects of the above stages in the context of sampling of graphs. Sampling from a class of graphs has been well studied in the graph theory literature (Jerrum and Sinclair, 1990; Hoppen and Wormald, 2008), where they define a class of graphs sharing some property (e.g., degree distribution) and prove that a particular random algorithm can generate all graphs in

the class. Cooper et al. (Cooper and Frieze, 2003) use this approach to show their algorithm for overlay construction generates graphs with good properties. Others use sampling to extract information from graphs, e.g., sampling a representative subgraph from a large, intractable graph, while maintaining properties of the original (Airoldi and Carley, 2005; Stumpf et al., 2005; Leskovec and Faloutsos, 2006). Others use sampling as a component of efficient, randomized algorithms (Karger, 1994).

However as the focus of this thesis has been, to recall, the understanding of the social processes that are emergent of our communication process, the prior literature does not lend insights into how our understanding (such as predictive modeling, introduced in chapters 3–9) of social processes will be affected by the sampling methodology. In the appendix section of the thesis, we probe into this observation and see how different ways to sample social media networks, such as Twitter, will affect a particular social process, information diffusion.

In this final part of the thesis, we also present an orthogonal problem that investigated the interplay and the role of considering content features of shared rich media objects in understanding network properties on Flickr.

Chapter 12

Sampling Social Media Networks to Understand Social Processes

Platforms such as Twitter have provided researchers with ample opportunities to analytically study social phenomena. There are however, significant computational challenges due to the enormous rate of production of new information: researchers are therefore, often forced to analyze a judiciously selected “sample” of the data. Like other social media phenomena, information diffusion is a social process—it is affected by user context, and topic, in addition to the graph topology. This chapter studies the impact of different attribute and topology based sampling strategies on the discovery of an important social media phenomena—information diffusion.

We examine several widely-adopted sampling methods that select nodes based on attribute (random, location, and activity) and topology (forest fire) as well as study the impact of attribute based seed selection on topology based sampling. Then we develop a series of metrics for evaluating the quality of the sample, based on user activity (e.g. volume, number of seeds), topological (e.g. reach, spread) and temporal characteristics (e.g. rate). We additionally correlate the diffusion volume metric with two external variables—search and news trends. Our experiments reveal that for small sample sizes (30%), a sample that incorporates both topology and user-context (e.g. location, activity) can improve on naïve methods by a significant margin of $\sim 15\text{-}20\%$.

12.1 Introduction

Over the past forty years, traditional methods of studying social processes such as information diffusion, expert identification or community detection have been focused on longitudinal studies of relatively small groups. However, the widespread

proliferation of several social websites such as Facebook, Twitter, Digg, Flickr and YouTube has provided ample avenues to researchers to study such processes at *very large scales*. This is because electronic social data can be acquired and stored over extended time intervals, and for very large populations. The result is that study of social processes on a scale of million nodes, that would have been inconceivable a decade ago, is becoming routine.

Consider the particular social process of information diffusion. The pervasive use of social media has made the cost involved in propagating a piece of information to a large audience extremely negligible, providing extensive evidences of large-scale social contagion. As a result researchers, today, are able to conduct massive empirical studies on diffusion, such as involving blog postings (?), Internet chain-letter data (Liben-Nowell and Kleiberg, 2008), social tagging (Anagnostopoulos et al., 2008), Facebook news feed (Sun et al., 2009), online games (Bakshy et al., 2009) and so on.

The attention paid to data volume, however, has overshadowed seemingly less obvious but two equally important challenges: namely, *data acquisition bottleneck* and *information analysis complexity*. For example, the social network Facebook currently features more than 350M users, while the social media Twitter has a rate of approximately 17,000 posts (tweets) per minute. Under such circumstances, firstly, typical data acquisition tools as provided by the publicly available APIs (Application Programming Interfaces) are often not sufficient to track all the data that is being generated—note, the rate limit of API calls for Twitter is only 20,000 per hour¹—hence creating an acquisition bottleneck. Second, there is extensive resource cost involved in storage of data of this scale, and also, thereby considerably high complexity in analyzing the data itself.

These challenges necessitate the need for collecting a *sample* of the social data that spans over a diverse set of users. Typically, researchers rely on some judiciously chosen sampling practice (e.g. random sampling or snowballing (Frank,

¹The default rate limit for API calls is 150 requests per hour; a whitelisted account or IP is allowed 20,000 requests per hour.

1978)) that can recover the topological characteristics of the particular social graph independent of the particular application in question (Leskovec et al., 2005), (Leskovec and Faloutsos, 2006). However, in order to study complex dynamic social processes such as diffusion, apart from topology of the social graph, there is a need to consider the nature of the shared information content as well as the rich social context—users (nodes) in a social network are associated with various attributes (e.g. location, age, profession, etc.) and the relationship (edges) may have various properties (e.g. friendship may have duration, or may be asymmetric).

In this chapter, we formally study how the choice of different sampling strategies impacts discovery of the particular social phenomenon, diffusion. Diffusion has found extensive potential in addressing the propagation of medical and technological innovations (Newman, 2002), cultural bias (Zachary, 1977), (Bakshy et al., 2009) and understanding information roles of users (Kempe et al., 2003), (Watts and Dodds, 2007).

Our approach comprises two steps. First, we utilize several popularly used sampling techniques such as random sampling, degree of user activity based sampling, forest-fire and location-attribute based sampling to extract subgraphs from a social graph of users engaged in a social activity. Second, these subgraphs are used to study diffusion characteristics with respect to the properties of the users (e.g. participation), structural (e.g. reach, spread) and temporal characteristics (e.g. rate) as well as relationship to events in the external world (e.g. search and news trends).

We have conducted extensive experiments on a large-scale dataset collected from Twitter to understand, up to what extent the results of diffusion analysis obtained from different types of samples are affected by the corresponding sampling methods. Our experiments reveal that methods that incorporate both network topology and user-context such as activity, or attributes related to “homophily” (e.g. location) are able to explain diffusion characteristics better compared to naïve methods (e.g. random or activity based sampling) by a large margin of $\sim 15\text{-}20\%$. Besides, for moderately small sample sizes (30%), these methods can explain the

metrics computed on unsampled graph better than pure attribute or topology based strategies.

The rest of the chapter is organized as follows. In the next section we present a discussion on related work. We present our problem definition in the following section, and then discuss different sampling techniques. The following section deals with our evaluation metrics, used to estimate diffusion bias under sampling. Thereafter we present experimental studies over Twitter, present a discussion of our work and finally conclude with our major contributions.

12.2 Related Work

We discuss related prior work from two different perspectives: first, sampling of large-scale graphs, and second, information diffusion in social media and networks.

Graph Sampling

Our work deals with extracting information from large-scale social networks, which is closely related to the problem of “subgraph sampling.” A subgraph sampling method commonly used in sociology studies is *snowball sampling* (Frank, 1978); another well-known method is *random walk sampling* (Klov Dahl et al., 1977). Recent work has investigated sampling of large-scale graphs, with a focus on recovering topological characteristics such as degree distribution, path length etc. (Rusmevichientong et al., 2001) as well as analyzing the impact of missing data on social network properties (?). For example, Leskovec et al. in (Leskovec et al., 2005; Leskovec and Faloutsos, 2006) focused on empirically observed static and dynamic graph properties such as densification and shrinking diameter. They studied different sampling methods, including random node/edge selection, random walk etc for recovering solely these topological properties. They also introduced *forest fire sampling*, which randomly selected a subset of neighbors of current traversed node to form a sample according to a forwarding probability.

The analysis of social information diffusion has been of interest to researchers from various domains ranging from social sciences, epidemiology, disease propagation, physics and economics (Zachary, 1977; Newman, 2002; Watts and Dodds, 2007). There has been prior work on modeling and predicting pathways of diffusion of information in social networks useful for several applications, ranging from recommendation systems, online advertising, user behavior prediction and disease containment (Kempe et al., 2003; Gruhl et al., 2004; Song et al., 2006; Anagnostopoulos et al., 2008; Kossinets et al., 2008; Liben-Nowell and Kleinberg, 2008).

In an early work (Kempe et al., 2003), the authors propose solution to the optimization problem of selecting the most influential nodes in a social network which could trigger a large cascade of further adoptions. In a recent work, Bakshy et al (Bakshy et al., 2009) study how “gestures” make their way through an online community—Second Life. In another work, Sun et al (Sun et al., 2009) study the diffusion patterns on the Facebook “News Feed” and conclude that in online social media, diffusion dynamics are often triggered by the collision of short chains of information trigger.

Although these prior work provide useful insights into the subgraph sampling problem as well as into characteristics of diffusion in social media separately, they suffer from limitations in the context of this work:

- *Information content:* Today’s online social media feature extensive activity that is dependent on the information content being shared (i.e. the topic) as well as have been historically observed to exhibit correlation with external events (Gruhl et al., 2005). Hence, pure topology-based sampling might not be suitable to study social processes that depend on the relationship between the shared content and external user actions and events.
- *Social context:* Most of the prior research does not consider the *contextual information* of the users in the social graph, such as the geographical location, or how quickly the user changes her status (e.g. the rate of social activity). Such

contextual information is crucial in studying social phenomena like diffusion, whose impact will be investigated in this chapter.

12.3 Problem Definition

We now introduce the social graph model, the key concepts and then define our research problem.

Social Graph Model. Our social graph model is based on the social media Twitter. Twitter features a micro-blogging service that allows users to post short content, known as “tweets”, often comprising URLs usually encoded via bit.ly, tinyurl, etc. The particular social action of posting a tweet is popularly called “tweeting”. Users can also “follow” other users; hence if user u follows v , Twitter allows u to subscribe to the tweets of v via feeds. Two users are denoted as “friends” on Twitter if they “follow” each other.

We now define a social graph $G(V, E)$ that is directed and where V is the set of users and $e_{uv} \in E$ if and only if users u and v are “friends” of each other. Note that, using the bi-directional link is more useful in the context of Twitter compared to the uni-directional “follow” link because the former is more likely to be robust to spam—a normal user is less likely to follow a spam-like account.

Topic Diffusion. We define diffusion with respect to a particular topic as the flow of information from one user to another via the social graph $G(V, E)$, also called “social contagion”. Given two users v and u sharing a “friend” link, topic diffusion on Twitter can manifest itself through three types of evidences: (1) users posting tweets using the same URL, (2) users tweeting with the same hashtag (e.g. #Election2008) or a set of common keywords, and (3) users using the re-tweet (RT) tag. We utilize all these three cases of diffusion in this chapter.

Diffusion Series. In order to study diffusion characteristics, we now define a topology called a *diffusion series* that summarizes diffusion via social contagion in the social graph G for a given topic θ . Note, a diffusion series is similar to a diffusion tree as in (Liben-Nowell and Kleinberg, 2008), (Bakshy et al., 2009), however we call it a “series” since it is constructed progressively over a period of time and allows a

node to have multiple sources of diffusion from more than one contact and at more than one time interval.

A diffusion series δ_θ on topic θ is a directed graph where the nodes are users tweeting on a topic over a period of time. Specifically, a node represents an occurrence of a user creating at least one tweet about a topic at a particular time, and nodes associated with the same period of time are arranged into the same level². Each level l_m in the diffusion series δ_θ is defined over a day in this chapter, i.e. in the context of Twitter, all the users in a particular level tweet about the information on the same day; and two consecutive levels have a time difference of one day. The edges across nodes between two adjacent levels indicate that user u in level l_m tweets about the information on the m -th day, after (via the subscribed feed) her contact (i.e. friend) v has tweeted about the same information on the previous day (at level l_{m-1}). Note there are also some other approximations on the diffusion process captures by this topology. Social diffusion is a continuous-time process, but the diffusion series are constructed over discrete time, at 24 hr increments. That is, there are no edges between nodes at the same level—a diffusion series in this work captures flow of information *across* days, and does not include possible flow occurring at the same day. However, our definition is generic enough to construct diffusion series over any arbitrary discrete temporal granularity. An example of a diffusion series on Twitter over the topic “global warming” has been shown in Figure 12.1. Significant news events associated with diffusion have also been annotated in the series in the figure.

Since each topic θ can have multiple disconnected diffusion series δ_θ at any given time, we call the set of all diffusion series a *diffusion collection* $D_\theta = \{\delta_\theta\}$.

Problem Statement. Given a topic θ and observable social actions of users in a social graph G , our goal is to: (1) utilize a set of sampling techniques $S \in \mathcal{S}$ to extract samples from the original social graph G , as given by $\hat{G}(S)$ where $\hat{G} \subset G$; and (2) empirically study the diffusion characteristics in the original social graph

²Hence, the same user may be present multiple times at different levels in a series if s/he tweets about the same topic at different points in time (different days).

Figure 12.1: Example of a diffusion series from Twitter on the topic “global warming”.

G (given by the diffusion collection D_θ) as well as in the samples $\hat{G}(S)$ (given by $\hat{D}_\theta(S)$); thereby empirically estimating the robustness and effectiveness of various sampling techniques $S \in \mathcal{S}$ useful for selectively retrieving information on the topic θ .

12.4 Sampling Diffusion Data

We define a sampling strategy $S \in \mathcal{S}$ as a technique that selectively chooses nodes from the social graph G based on a certain attribute or technique. Typically the sampled graph $\hat{G}(S)$ is considerably smaller in size in terms of number of nodes, compared to G ; the size being determined based on a ratio (called the sampling ratio): $\rho = \text{size}(\hat{G}(S)) / \text{size}(G)$. In this chapter for each strategy S , we have defined samples based on ρ linearly ranging between 0.1 and 1.0.

We now examine six different sampling strategies, that include three attribute-based and three topology-based techniques. For attribute based techniques, we form the sampled graph $\hat{G}(S)$ as follows. Using the nodes obtained from the sampling method S , we examine G for existence of edges between all pairs of the selected nodes. Thereafter we associate an edge between nodes u and v in $\hat{G}(S)$ if u and v are connected in G . For the topology-based techniques, we select a seed user set

based on an attribute, and then use the topology of G to expand the sampled graph $\hat{G}(S)$.

Attribute-based Techniques

1. *Random Sampling*: We focus on a random sampling strategy where we select users based on a uniform distribution.
2. *Activity-based Sampling*: This sampling strategy involves choosing a subgraph $\hat{G}(S)$ that comprises the most active users (in terms of their number of tweets).
3. *Location-based Sampling*: For this method, we divide the users in G into different categories based on their location attribute³. The different locations considered in this chapter are the different continents, e.g. ‘North America’, ‘Asia’ and ‘Europe’. For ℓ different locations and a given ρ , we randomly select ρ/ℓ users corresponding to each location to construct the subgraph sample $\hat{G}(S)$.

Topology-based Techniques

We adopt the forest-fire sampling method as described in (Leskovec and Faloutsos, 2006). We have used three different ways to choose the seed user set in this method. In the first case, the seed user set is selected at random. We expand from this seed set based on their contacts (i.e. friends), where the contacts are chosen based on a forwarding probability p_f . This process is repeated in turn for each of the contacts and so on, until we “burn” sufficient number of users for each sample $\hat{G}(S)$. In the second case, the seed user set is chosen based on a particular location, e.g. ‘Asia’, ‘Europe’ etc; while in the third, it is chosen in terms of measure of user activity.

³We use the timezone attribute of the users on Twitter to extract locations, e.g. Eastern Time would indicate users from New York, Massachusetts, Pennsylvania etc.

12.5 Evaluation of Diffusion Samples

Given the sampled social graphs $\hat{G}(S)$ constructed through sampling strategies $S \in \mathcal{S}$, we construct the diffusion collections $\hat{D}_\theta(S)$. To evaluate the quality of these diffusion collections, we now propose two types of metrics: (a) *saturation metrics*: characterizing the quality of $\hat{D}_\theta(S)$ with respect to the diffusion collections D_θ constructed from unsampled graph G ; (b) *response metrics*: quantifying how the diffusion results obtained from $\hat{D}_\theta(S)$ corresponding to popular external activities e.g. search and news trends.

Diffusion Saturation Metrics

We describe eight different metrics for quantifying diffusion on a certain topic that are discovered via a variety of sampling techniques. The metrics are categorized through various aspects such as: properties of users involved in diffusion (volume, participation and dissemination), diffusion series topology (reach, spread, cascade instances and collection size) and temporal properties (rate):

1. *Volume*: Volume is a notion of the overall degree of contagion in the social graph. For a sampling technique $S \in \mathcal{S}$, we formally define volume $v_\theta(S)$ with respect to θ as the ratio of $n_\theta(S)$ to $N_\theta(S)$, where $n_\theta(S)$ is the total number of users (nodes) in the diffusion collection $\hat{D}_\theta(S)$, and $N_\theta(S)$ is the number of users in the sampled social graph $\hat{G}(S)$ associated with topic θ . Note, $N_\theta(S)$ would include users who are not part of the diffusion collection, but nevertheless have tweeted about θ .
2. *Participation*: Participation $p_\theta(S)$ ((Bakshy et al., 2009)) is fraction of users involved in the diffusion of information on a particular topic who further trigger other users in the social graph to get involved in the diffusion. It is the number of non-leaf nodes in the diffusion collection $\hat{D}_\theta(S)$, normalized by $N_\theta(S)$.
3. *Dissemination*: Dissemination $d_\theta(S)$ is given by the ratio of the number of users in the diffusion collection $\hat{D}_\theta(S)$ who do not have a parent node, nor-

malized by $N_\theta(S)$. In other words, they are the “seed users” or ones who get involved in the diffusion due to some unobservable external influence, e.g. a news event.

4. *Reach*: Reach $r_\theta(S)$ ((Liben-Nowell and Kleiberg, 2008)) is conceptually defined as the extent in the social graph, to which information on a particular topic θ reaches to users. We define it formally as the mean of the number of levels in all the diffusion series $\hat{\delta}_\theta(S) \in \hat{D}_\theta(S)$.
5. *Spread*: For the diffusion collection $\hat{D}_\theta(S)$, spread $s_\theta(S)$ ((Liben-Nowell and Kleiberg, 2008)) is defined as the ratio of the maximum number of nodes at any level in $\hat{\delta}_\theta(S) \in \hat{D}_\theta(S)$ to $n_\theta(S)$.
6. *Cascade Instances*: Cascade instances $c_\theta(S)$ is defined as the ratio of the number of levels in the diffusion series $\hat{\delta}_\theta(S) \in \hat{D}_\theta(S)$ where the number of *new* users at a level l_m (i.e. non-occurring at a previous level) is greater than that at the previous level l_{m-1} , to L_δ , the number of levels in $\hat{\delta}_\theta(S) \in \hat{D}_\theta(S)$.
7. *Collection Size*: Collection size $\alpha_\theta(S)$ is the number of diffusion series $\hat{\delta}_\theta(S)$ in $\hat{D}_\theta(S)$ over a certain topic θ .
8. *Rate*: We define rate $\gamma_\theta(S)$ as the “speed” at which information on θ diffuses in the collection $\hat{D}_\theta(S)$. It depends on the difference between the median time of posting of tweets at all consecutive levels l_m and l_{m-1} in the diffusion series $\hat{\delta}_\theta(S) \in \hat{D}_\theta(S)$. Hence it is given as:

$$\gamma_\theta(S) = 1 / (1 + \frac{1}{L_\delta} \sum_{l_{m-1}, l_m \in \hat{\delta}_\theta(S)} (t_\theta^m(S) - t_\theta^{(m-1)}(S))), \quad (12.1)$$

where $t_\theta^m(S)$ and $t_\theta^{(m-1)}(S)$ are measured in seconds, $t_\theta^m(S)$ corresponds to the median time of tweet at level l_m in $\hat{\delta}_\theta(S) \in \hat{D}_\theta(S)$.

We now define a distortion metric in order to evaluate quantitatively the performance of each of these diffusion saturation metrics for the different sampling strategies, $S \in \mathcal{S}$. The distortion metric is defined as:

$$F_\theta(m; S) = \frac{|m_\theta(S) - \hat{m}_\theta(S)|}{m_\theta(S)}, \quad (12.2)$$

where m is the particular metric under consideration. $\hat{m}_\theta(S)$ is the measure of metric m under S and computed over the diffusion collection $\hat{D}_\theta(S)$, while $m_\theta(S)$ is the metric over D_θ , corresponding to the unsampled social graph.

Diffusion Response Metrics

We now describe metrics for quantifying the relationship between diffusion characteristics obtained from samples within Twitter, and the trends of the same given topic obtained from external world. We collect two kinds of external-world trends: (1) *search trends*—the search volume of a particular topical keyword over a period of time⁴; (2) *news trends*—the frequency of archived news articles about a particular topical keyword over a period of time⁵. Based on these trends, we define two diffusion response metrics:

1. *Search response*: We first compute the cumulative distribution function (CDF) of diffusion volume as $E_D(x) = \sum_{i \leq x} |l_i(\hat{D}_\theta(S))| / Q_D$, where $|l_i(\hat{D}_\theta(S))|$ is the number of nodes at the i th level in the collection $\hat{D}_\theta(S)$ obtained via sampling technique S . Q_D is the normalized term and is defined as $\sum_i |l_i(\hat{D}_\theta(S))|$. Next, we compute the CDF of search volume as $E_S(x) = \sum_{i \leq x} f_S(i) / Q_S$, where $f_S(i)$ is the search volume returned by the Google Trends API for the given time i , and Q_S is the normalization term. The search response is defined as $1 - D(E_D, E_S)$, where $D(A, B)$ is the Kolmogorov-Smirnov (KS) statistic and is defined as $\max(|A(x) - B(x)|)$.
2. *News response*: Similarly, we compute the CDF of news volume as $E_N(x) = \sum_{i \leq x} f_N(i) / Q_N$, where $f_N(i)$ is the number of archived news articles available from Google News for the given time i , and Q_N is the normalization term. The news response is similarly defined as $1 - D(E_D, E_N)$, where $D(A, B)$ is the KS statistic.

⁴Google Trends: (<http://www.google.com/intl/en/trends/about.html>).

⁵Google News: (<http://news.google.com/>).

Table 12.1: Summary of statistics of the data used for studying diffusion on Twitter.

#nodes	465,107
#edges	836,541
#nodes with time-zone attribute	385,547
#tweets	25,378,846
Time span of tweets posting times	Oct 2006–Nov 2009

Experimental Study

We have focused on a large dataset crawled from Twitter. We have undertaken a focused crawl⁶ based on a snowballing technique, over a set of quality users ($\sim 465K$), who mutually form a reasonably large connected component. First, we seeded the crawl from a set of genuine (or authoritative) users, who post about a diverse range of topics and reasonably frequently. Our seed set size is 500; and comprises politicians, musicians, environmentalists, techies and so on. These lists were collected from the popular social media blog, Mashable (<http://mashable.com/2008/10/20/25-celebrity-twitter-users/>). Next we expand the social graph from the seed set based on their “friend” links⁷. We finally executed a dedicated cron job that collected the tweets (and their associated timestamps) for users in the entire social graph every 24 hours. Table 12.5 gives some basic statistics of the crawled data that were used for studying diffusion.

Experimental Procedure

The crawled social graph, comprising the users and their tweets are now deployed in the study of diffusion. Since we are interested in studying diffusion at the granularity of a topic, we first define how we conceive of the topics. For our experiments, we focus on the “trending topics”⁸ that are featured on Twitter over a two month period

⁶<http://apiwiki.twitter.com/>

⁷Note that the social graph crawled in this work is a static snapshot made only once at the time of the crawl.

⁸Trending topics are Twitter-generated list of popular topics. Note they can either be hashtags (i.e. words or phrases preceded by the # symbol), or could be groups of words.

between Oct and Nov 2009. From the ensemble of these trending topics, a set of ~ 125 topics are selected at random; of which there are 25 hashtags and the rest, phrases or groups of words.

For the ease of analysis, we organize the different trending topics into generalized themes. For automatically assigning theme to trending topic associations, we use the popular open source natural language processing toolkit called “OpenCalais”⁹. In the context of Twitter, we filter tweets give a trending topic, and then use OpenCalais to return theme labels over those tweets. Based on this process, we associated the 125 trending topics with a total of nine themes, such as ‘Business Finance’, ‘Sports’ etc’.

Now our experimental goal is to utilize the crawled social graph to construct diffusion samples *per topic* and thereby study the impact of sampling on diffusion.

Results

We present our results from two different perspectives:

1. *What is a good sampling strategy?* i.e. how the choice of a particular sampling technique affects diffusion characteristics and their relationship to external trends.
2. *What is a good sampling ratio?* i.e. how to choose the sample size for different sampling strategies, such that it can explain well the distributions over diffusion metrics on the unsampled graph.

Analysis of Sampling Strategies

Saturation Metrics. Figure 12.2(a-h) gives the results of how the six different sampling methods impact the saturation metrics over the sampling ratio ρ . The results are averaged over all trending topics. Note, for the forest-fire based methods, the results are averaged over different values of the forwarding probability p_f and for all methods, 50 iterations were undertaken to ensure statistical significance.

⁹<http://www.opencalais.com/>

Figure 12.2: Distortion of different diffusion metrics for different choices of sampling ratio ρ , averaged over all the nine themes (i.e. ~ 125 trending topics from Twitter).

We observe that there is significant variation in the distortion measures across the six different strategies. Primarily, for the user-based metrics—volume, participation and dissemination, we observe that the values of distortion lie in a range of 20-25% across the different methods. Also the relatively narrower range of distortion across strategies shows that these three metrics seem to be less sensitive to different sampling methods. Nevertheless, note that the forest-fire with activity based seeds sampling outperforms others.

In the case of the diffusion series based topology metrics—reach, spread, cascade instances and collection size, the distortion measures across the different strategies seem to be comparatively more widely-ranged, i.e. the variations range between 30-35%. For example, the three forest-fire based sampling techniques perform significantly better compared to just the activity based, random or location based methods—reinforcing the fact that incorporating topology in the sampling process gives less distortion in terms of the diffusion series topology.

The last metric, rate (that is time-based) also exhibits monotonic increase in distortion for decreasing ρ . Interestingly, the range of variation for different methods is narrowly around 10% in this case; indicating that the choice of the particular sampling strategy does not seem to impact much the measurement of the rate.

Now we present some results that illustrate how diffusion on each topic is affected by the choice of the sampling strategies, averaged across all values of the sampling ratio ρ . Figure 12.3(a-c) gives the mean measures of distortion for the nine themes and for different sampling techniques, shown for the three categories of diffusion metrics. Across these three metric categories, we observe the following variations over different themes:

- Context such as demographics (location) seem to perform well in yielding quality samples for themes that are ‘local’ in nature, e.g. ‘Sports’ comprising topics such as ‘NBA’, ‘New York Yankees’, ‘Chargers’, ‘Sehwag’ and so on—each of them being of interest to users respectively from the US, NYC, San Diego and India.
- Pure topology based sampling (i.e. forest-fire with random seeds) seem to perform well for themes that are of global importance, such as ‘Social Issues’ that subsumes topics like ‘#BeatCancer’, ‘Swine Flu’, ‘#Stoptheviolence’ and ‘Unemployment’.
- Incorporating both context (i.e. activity) and topology (i.e. forest-fire) in sampling seems to perform well for themes that relate to external events or issues, e.g. ‘Technology-Internet’ comprising topics like ‘Android 2’, ‘Google Wave’ and ‘Windows 7’. Similar observations can be made for ‘Politics’ that subsumes topics like ‘Tiger Woods’, ‘Healthcare’ and ‘Afghanistan’—all of which were associated with important external happenings during the period of our analysis.

Response Metrics. Now we present analysis of diffusion in terms of its response to external variables: search and news trends. Figure 12.4 shows the impact of sampling ratio ρ on response. We observe that for search response, the range of variation of response (based on the KS statistic) lying between $\rho = 0.1$ and 1.0 is larger compared to that of news. This indicates that for smaller values of ρ , the news trends are more responsive to diffusion characteristics compared to search trends.

Figure 12.3: Distortion of diffusion metrics across themes, averaged over sampling ratio ρ . Metrics – user-based: volume, participation and dissemination; topology-based: reach, spread, cascade instances and collection size; time-based: rate.

We conjecture that it is because diffusion processes on Twitter are heavily related to external news-related events.

Figure 12.4: Response behavior (using the KS statistic) with respect to search and news trend for different choices of sampling ratio ρ , averaged over all themes.

Additionally, we also observe that the sampling technique that yields maximum response in case of search trends is the forest-fire with location based seeding technique. Moreover, pure location based sampling appears to perform better compared to attributes such as activity based (that performs quite poorly). This implies that since search behavior often heavily relies on user demographics, diffusion samples drawn based on attributes like location yield good response measures. While for news trends, best performance is given by the forest-fire technique seeded based

Figure 12.5: Response behavior (using the KS statistic) with respect to search and news trend across different themes, averaged over sampling ratio ρ .

on activity.

Finally, in Figure 12.5, we present the results of response for search and news over the nine different themes averaged across ρ . In the case of search trends, the results indicate that location based sampling and forest-fire seeded based on location perform considerably well in comparison to other techniques; especially for themes that heavily reflect user interest and are aligned along certain demographic attributes, e.g. ‘Entertainment-Culture’ (example topics are ‘Chris Brown’, ‘Eagles’), ‘Sports’ and ‘Technology-Internet’. While for news trend, we again observe that for several themes, activity based sampling yields good performance. For example, themes such as ‘Politics’ (subsuming topics like ‘Healthcare’ and ‘Afghanistan’), ‘Sports’ and ‘Technology-Internet’ being associated with external events, diffusion samples for these themes drawn using user activity seem to be highly responsive to news trends.

We summarize performance of the sampling techniques over the saturation and the response metrics in Table 12.5. Note, the results have been found to be statistically significant based on the student t-test statistic and using 50 independent runs of each sampling method.

Table 12.2: Summary of performance of different sampling techniques over the saturation and response metrics.

Method	Saturation Distortion	Response
Random	0.41	0.68
Activity	0.44	0.64
Location	0.35	0.66
Forest-fire+Random	0.28	0.73
Forest-fire+Location	0.27	0.76
Forest-fire+Activity	0.22	0.78

Analysis of Sampling Ratio

In Figure 12.2 we had observed that a moderate sampling ratio of $\rho=0.3$ yields low distortion for user-based and topology-based metrics. To account for choice of ρ more concretely, we present the performance of different strategies for $\rho=0.3$. Figure 12.6 shows the topic distribution for diffusion volume and spread—X-axis representing values of the metric while Y-axis, the number of topics having a particular value of the metric. We compare three sampling methods against the distributions on unsampled graph. It appears that, for a moderately small $\rho=0.3$, forest-fire method seeded on location outperforms pure attribute or topology-based techniques in explaining these distributions, with mean error of $\sim 8.5\%$.

Figure 12.6: Comparative distribution of (a) volume and (b) spread over different strategies for sampling ratio $\rho=0.3$.

12.6 Discussion

Our primary observation from the results is that sampling impacts the discovery of dynamic social processes, such as diffusion, in a non-trivial manner. Contrary to prior empirical observations that topology alone can yield good subgraph samples (Leskovec and Faloutsos, 2006), we have found evidence that sampling techniques that incorporate user context, e.g. activity or location along with the graph topology seem to perform better in discovery of diffusion. Interestingly also note, pure context based techniques, such as location appear to perform reasonably well—better than activity based or random sampling. We conjecture that it is related to the concept of user “homophily” (Mcpherson et al., 2001); that explains that users engaged in a social activity seem to be associated more closely with ones who are “similar” to them along a certain (contextual) dimension, such as location, age, political view or organizational affiliation, compared to ones who are “dissimilar”.

We also observe that diffusion characteristics are widely varied across the different themes; hence content has great impact on the quality of the sample. For example, studies of diffusion related to a political event of importance in the US would benefit more from samples chosen based on location than on pure graph topology. Or if the interest is related to a recent technological event, such as release of an electronic gadget, one can benefit more from sampling techniques based on both topology and activity.

Our results are promising, however are limited by the scope of our dataset, which itself is based on a crawl. Hence the observations on diffusion are likely to be only approximate, because we do not quantify the inherent bias in our initial snowball sample of the Twitter population. Moreover, the non-uniformities within each sampling process also have not been considered, nor have we evaluated the bias in each strategy using any form of sampling bias estimators (?). Additionally note that in this chapter we have focused on only one social process: diffusion. Nevertheless, we believe that our empirical observations are extensible to other phenomena as well, e.g. community discovery; because most social processes are

affected by both topology and context. Finally, we acknowledge that alternative sampling techniques are also possible. For example, a viral marketer intending to maximize the flow of information in a very short span of time might be interested in sampling that chooses nodes based on time-varying properties of the edges (Kossinets et al., 2008).

12.7 Conclusions and Future Work

We have empirically studied the impact of attribute and topology based sampling methods on discovery of information diffusion in data from Twitter. Our main conclusion is that methods that incorporate both network topology and user-contextual attributes such as activity estimate information diffusion with lower error, for the same sample size, when compared to naïve methods (e.g. random or activity based sampling). The improvements are significant: $\sim 15\text{-}20\%$. Our results also show that for a reasonably small sample size ($\sim 30\%$) these methods can explain well the topic distributions for diffusion metrics on the unsampled graph.

There are several promising future research directions. We plan to extend this study to different social media datasets (e.g. Digg, Flickr), and social phenomena (e.g. community discovery) to see the effects of sampling. Second, we propose to improve on the topology based methods (e.g. forest fire) to incorporate the effect of context, and content on how the forest fire grows—at present, only the seed nodes are affected by context, i.e. location and activity.

Chapter 13

Connecting Media to Community

“Interdependence is and ought to be as much the ideal of man as self-sufficiency. Man is a social being.”—Mahatma Gandhi (1869–1948), political and spiritual leader of India.

In this chapter we develop a recommendation framework to connect image content with communities in online social media. The problem is important because users are looking for useful feedback on their uploaded content, but finding the right community for feedback is challenging for the end user. Social media are characterized by both content and community. Hence, in our approach, we characterize images through three types of features: visual features, user generated text tags, and social interaction (user communication history in the form of comments). A recommendation framework based on learning a latent space representation of the groups is developed to recommend the most likely groups for a given image. The model was tested on a large corpus of Flickr images comprising 15,689 images. Our method outperforms the baseline method, with a mean precision 0.62 and mean recall 0.69. Importantly, we show that fusing image content, text tags with social interaction features outperforms the case of only using image content or tags.

13.1 Introduction

There has been an unprecedented increase in the number of social media websites (e.g. Flickr¹, YouTube, Slashdot, Digg, del.icio.us) in the past few years which have allowed users to create, share and consume rich media very easily. Social media sites are popular not just for the content, but also due to the accompanying social interaction. In popular image sharing sites such as Flickr, enthusiastic

¹<http://www.flickr.com/>

photographers are interested in receiving critical comments on their photos. Note that simply uploading an image onto Flickr does not ensure rich social interaction or reachability to other users for critical feedback.

Flickr allows people to connect their images to communities, through the mechanism of image ‘groups’ (also known as image pools). A Flickr group is a repository of images shared by a set of users and is usually organized under a certain coherent theme (e.g. the group “The Magic of Nature”). However, finding the right community that will give useful comments is not easy. Simple text based search for a group will reveal a large number of similar communities (also known as image groups / pools on Flickr) e.g. “Travel / Travel Photography / Travel in Asia” etc. The fundamental challenge addressed in this chapter is to connect user content to the correct community—i.e. given an image, recommend the relevant group(s) that would enable social interaction and enhance the reachability to other users.

Related Work

There has been considerable work in recommendation of items (e.g. books, movies) (Schein et al., 2002) to users as well as on recommending tags to media objects (Garg and Weber, 2008; Sigurbjörnsson and van Zwol, 2008). A fundamental distinction between prior work and our own, is that prior work has tended to pay close attention to the content (e.g. automated tag recommendation systems for images), while paying less attention to the social interaction, a key component of social media. Since we are interested in ‘connecting’ content (image) with community (groups, where significant social interaction occurs), we pay close attention to social interaction (user-user comments), in addition to content based visual features and text tags.

In a recent work on group and tag recommendation (Chen et al., 2008), authors use appearance based image concepts, but they do not incorporate social interaction in their analysis. In another work on analysis of Flickr groups (Negoescu and Gatica-Perez, 2008), the authors have analyzed user behavior in these groups. While the group representation framework is rich, it does not incorporate social interaction, and has not been used for connecting content with community. To the

best of our knowledge, this is one of the first works where image content is connected to social media communities, through the use of appearance based features, text tags and social interaction history.

Our Approach

A system overview of our group recommendation framework is shown in Figure 13.1. The framework is based on three ideas: (a) users can associate their images with groups whose themes they consider fit to the image, or (b) are interested in the concepts (tags) or (c) the on-going communication (comments) among the group members. Hence at the first step (feature extraction), three types of features are extracted for all images in the dataset: image content, tag and communication features. Tag features are given by a vector of the frequency counts of the tags the owner of the image has used over the past. Communication features are given by the frequency of comments written by the owner of the image on different groups. In the second step (model learning), bag-of-features based representations of the groups are generated and a model is learnt to represent the groups in a latent space. Finally, we use the learnt model parameters to recommend k groups for each image.

Figure 13.1: The overall system overview of our group recommendation framework.

We have performed extensive experimentation on a dataset (15,689 images) crawled from Flickr. Our method yields good results in recommending groups to

images compared to a k -Nearest Neighbor based baseline framework. In our dataset, each image on an average is associated with three groups, and we observe that our results yield high precision and recall of 0.65 and 0.69 respectively for $k = 3$.

13.2 Image Features

In this section, we briefly discuss different content based features that have been used to characterize the images.

Color. There are two color-based features of interest—color histogram and color moments, which capture the distribution of different colors in the images.

Texture. Flickr images show diversity of texture. We use two texture features: gray level co-occurrence matrix, GLCM, and a texture detector for arbitrary “blobs” in images—called phase symmetry (Xiao et al., 2005). Phase symmetry is based on determining local symmetry and asymmetry across an image from phase information. Given an image, phase-based symmetry detector (PSD) maps a pixel, p , an orientation, o , and a scale, n , to a phase congruency value $PC_{n,o}(p)$ and a special phase, $\phi_{n,o}(p)$:

$$PSD(p, n, o) = (PC_{n,o}(p) \cdot \phi_{n,o}(p)). \quad (13.1)$$

Here,

$$PC_{n,o}(p) = \frac{sumE(p)}{sumA(p) + \varepsilon} = \frac{\sum_{k,q} E_{n_k,o_q}(p)}{\sum_{k,q} A_{n_k,o_q}(p) + \varepsilon}, \quad (13.2)$$

where $sumE(p)$ is the total energy when phases are congruent under all scales and orientations and phases are zero; $sumA(p)$ is the total amplitude when phases are congruent under all scales and orientations and phases are zero; ε is a positive constant.

Shape. Images are often characterized by central themes or concepts. To extract

such features, we use two shape features—radial symmetry (Loy and Zelinsky, 2003) and phase congruency (Liu and Laganière, 2007). The radial symmetry feature is based on the idea of detecting points of interest in an image. Phase congruency is an illumination and contrast invariant measure of feature significance. For a given image, phase congruency $PC(x)$ at some location x is expressed as the summation over orientation o and scale n :

$$PC(x) = \frac{\sum_o \sum_n W_o(x) [A_{n,o}(x) \Delta \Phi_{n,o}(x) - T_o]}{\sum_o \sum_n A_{n,o}(x) + \varepsilon}, \quad (13.3)$$

where A_n represents the amplitude of the n -th component (of the image) in the Fourier series expansion, $W_o(x)$ is the convolution of the given image with an even / odd filter, ε is added for cases of small Fourier amplitudes, T_o is a compensating measure for the influence of noise and $\Delta \Phi_n(x)$ is a sensitive phase deviation.

SIFT. SIFT or Scale Invariant Feature Transform (Lowe, 2004) is a content-based image feature that detects stable keypoint locations in scale space of an image. It computes the following function from the difference of two nearby scales separated by a constant multiplicative factor κ :

$$D(x, y, \sigma) = (G(x, y, \kappa\sigma) - G(x, y, \sigma)) \cdot I(x, y) = L(x, y, \kappa\sigma) - L(x, y, \sigma), \quad (13.4)$$

where $L(x, y, \sigma)$ is the scale space of image $I(x, y)$ and is produced from the convolution of a variable-scale Gaussian having scale (σ) with $I(x, y)$.

The details of these features may be referred to in (Lowe, 2004; Xiao et al., 2005; Liu and Laganière, 2007; Loy and Zelinsky, 2003). We discuss tag features to characterize images in the next section.

13.3 Tag Features

We develop a set of tag features for each image based on the tagging activity of the owner of the image. Tag feature extraction is useful because: (1) users develop a set of ‘favorite’ concept spaces (in the form of tags) over time to describe their

images, and (2) an image is likely to be associated by the user to a group whose concepts are similar to the high frequency tags she has used in the past.

Hence, if the tag distribution given by the prior tagging activity of the owner of an image is close to the distribution of tags in a group, this group should be assigned a high probability of recommendation. We construct a vector of the frequency counts of the tags the owner of the image has used in the past (on other images), prior to the date of upload of the image. Let for image i the timestamp of its upload be t_i and τ_u be the set of all unique tags that user u (owner of i) has used for her other images from time 0 to t_i . The frequency count $n_{u,j}$ of usage of each tag j in τ_u gives the tag feature vector $\mathbf{T}_i$ for image i :

$$\mathbf{T}_i = [n_{u,1}, n_{u,2}, \dots, n_{u,L}], \text{ s.t. } L = |\tau_u|. \quad (13.5)$$

We now discuss communication features to characterize images in the next section.

13.4 Communication Activity Features

We develop communication based features for each image based on the frequency of comments written by the owner of the image on different groups. A user participating in the communication in a certain set of groups through comments is likely to be interested in those groups. Hence recommending them to the user is useful. Let Alice be interested in groups related to travel and frequently leave comments on the images in such groups. Further suppose Alice uploads a new image on Grand Canyon for which she intends to find suitable groups. Since Alice has earlier expressed interest in the travel related groups through her communication, our framework should recommend her such groups. Let $n_{u,j}$ be the number of comments written by user u on group j , ($1 \leq j \leq M$) in the time period from 0 to t_i where t_i is the timestamp of posting image i by u . Then the communication activity feature $\mathbf{O}_i$ for image i is given by the vector of $n_{u,j}$ over all groups j :

$$\mathbf{O}_i = [n_{u,1}, n_{u,2}, \dots, n_{u,M}]. \quad (13.6)$$

Based on the three types of features, we now characterize all the N images in the dataset. Suppose D is the dimensionality of the feature vector of each image i , then, $\mathbf{f}_i \in \mathbb{R}^{1 \times D}$, $\mathbf{f}_i = [\alpha_1 \cdot \mathbf{C}_i, \alpha_2 \cdot \mathbf{T}_i, \alpha_3 \cdot \mathbf{O}_i]$, where $\mathbf{C}_i$, $\mathbf{T}_i$ and $\mathbf{O}_i$ are the image content, tag and communication feature vectors; and α_1 , α_2 and α_3 are the weights determining the impact of each kind of feature. Based on these features, we now discuss the group recommendation framework.

13.5 Group Recommendation Framework

We now present our group recommendation framework. First, we present the main idea. Second, we discuss the learning of the model parameters. Finally, we discuss the folding-in technique where we determine top k recommended groups.

Main Idea

The goal of the recommendation framework is to determine the following probability over all M groups G_j for a given image i :

$$P(G_j|i) \propto P(G_j, i). \quad (13.7)$$

In order to compute the above joint probability, we develop a mixture model representation of each image over a set of latent states, motivated from pLSA (probabilistic latent semantic analysis) (Hofmann, 1999). This latent space could be impacted by different factors—the content of the image, owner’s prior tagging activity or her prior commenting activity over different groups. The joint probability $P(G_j, i)$ above can thus be represented as:

$$P(G_j, i) = \sum_z P(z) \cdot P(i|z) \cdot P(G_j|z) \propto \sum_z P(z|i) \cdot P(G_j|z), \quad (13.8)$$

where z are the latent states. Hence determining the group recommendation probabilities given an image can be reduced to computing the two conditional proba-

bilities: $P(G_j|z)$ and $P(z|i)$. Computing $P(G_j|z)$ can be considered as training or model learning, as it is independent of the image, given the latent states. While, computing $P(z|i)$ directly depends on the image whose recommendations we are seeking, and hence can be considered as testing or folding-in technique.

Model Learning

We first discuss the construction of the training set based on computing feature vector representations for each group; and second, learning the model. We assume that each group is a ‘bag-of-features’, comprising its constituent image content, tags and user comments. Let Q images be used for the training set and $N - Q$ images for the test set. Using these Q images, the training set, $\mathbf{Y} \in \mathbb{R}^{D \times M}$ is defined over M groups where each group G_j is represented by a feature space of its associated images (centroid). The feature vector for the j -th group in $\mathbf{Y}$ is thus computed as:

$$\mathbf{Y}_j = \frac{1}{|G_j|} \sum_{i \in G_j} \mathbf{f}_i, \quad (13.9)$$

where $\mathbf{f}_i$ is the i -th image in the group G_j . Using this training set, we now discuss learning the conditional probability $P(G_j|z)$ based on the EM-algorithm. If F_m is the m -th feature attribute in $\mathbf{Y}$, the update equations of EM (Hofmann, 1999) are given as:

$$\begin{aligned} \text{E-step: } P(z|F_m, G_j) &= \frac{P(z) \cdot P(F_m|z) \cdot P(G_j|z)}{\sum_{z'} P(z') \cdot P(F_m|z') \cdot P(G_j|z')}. \\ \text{M-step: } P(G_j|z) &\propto \sum_m \mathbf{Y}_{m,j} \cdot P(z|F_m, G_j), \\ P(F_m|z) &\propto \sum_j \mathbf{Y}_{m,j} \cdot P(z|F_m, G_j), \\ P(z) &\propto \sum_m \sum_j \mathbf{Y}_{m,j} \cdot P(z|F_m, G_j). \end{aligned} \quad (13.10)$$

Now we discuss the folding-in technique for determining the groups to be recommended to a given new image (test image).

Folding-In Technique

We discuss learning the conditional probability $P(z|i)$ for a given new image (test image) in $\mathbf{Z} \in \mathbb{R}^{D \times (N-Q)}$ and how based on the learnt parameters we can determine the top k recommended groups for the images. The basic idea of computing the probability $P(z|i)$ is based on how we can "fold-in" (Schein et al., 2002) a new image in our existing data to predict its probability of being recommended to different groups. We again use the following EM update rules to determine the probability:

$$\begin{aligned} \text{E-step: } P(z|i, F_m) &\propto P(F_m|z) \cdot P(z|i). \\ \text{M-step: } P(z|i) &\propto \sum_z \mathbf{Z}_{m,i} \cdot P(z|i, F_m), \end{aligned} \tag{13.11}$$

where F_m is the m -th feature attribute of test image i . Substituting the learnt probabilities $P(G_j|z)$ and $P(z|i)$ from eqns. 13.10 and 13.11 in eqn. 13.8 and finally in eqn. 13.7, we can determine the probability of recommending a group G_j to test image i . Hence the top k recommendations $g_i \in \mathbb{R}^{1 \times k}$ for a test image i is given by those k groups among M for which $P(G_j|i)$ is maximum. Let us now discuss the experimental results.

13.6 Experimental Results

In this section we discuss the experimental results. First we present a brief overview of the Flickr dataset. Second we discuss the results and finally we present a brief discussion of the limitations of our results.

Flickr Dataset

We have tested our group recommendation framework on a dataset crawled from the popular media-sharing site, Flickr. We downloaded images ranked by Flickr's proprietary "interestingness" criterion. The dataset comprises 15,689 images which belong to 925 groups; each group on an average consisting of 17 images. The mean number of tags per image is six, that of groups per image is three and comments per image is 14. The upload time period of these images ranged from March 21, 2008 to

Table 13.1: Evaluation of our method (M_1) against k -Nearest Neighbor (M_2) using Precision, Recall and F1-measure. Metrics are computed at $k=1, 3, 10$ and for three cases: precision, recall and F1-measure over all images, over images owned by each user and over images belonging to each group. Our method outperforms k -NN.

		Precision		Recall		F1-measure	
		M_1	M_2	M_1	M_2	M_1	M_2
Image-based	$k=1$	0.68	0.50	0.63	0.47	0.66	0.49
	$k=3$	0.63	0.49	0.69	0.52	0.64	0.50
	$k=10$	0.52	0.47	0.71	0.55	0.59	0.51
User-based	$k=1$	0.64	0.48	0.59	0.46	0.61	0.48
	$k=3$	0.59	0.46	0.62	0.48	0.60	0.47
	$k=10$	0.51	0.43	0.68	0.52	0.58	0.47
Group-based	$k=1$	0.74	0.52	0.72	0.52	0.72	0.53
	$k=3$	0.69	0.50	0.71	0.54	0.70	0.52
	$k=10$	0.61	0.48	0.81	0.57	0.69	0.55

August 20, 2008. About 80% of the images (randomly selected) were used to construct the training set (12,551 images) and the rest of the 3,138 images constituted the test set.

Model Evaluation

The results of validation of our group recommendation framework (M_1) against a baseline method k -Nearest Neighbor (M_2) is shown in Table 13.1. The performance of our method is evaluated using three metrics: precision, recall and F1-measure. We compute these metrics for different values of the number of recommended groups ($k=1, 3$ and 10). Further, precision, recall and F1-measure are computed for three cases: (a) over all images, (b) over images from each user and (c) over images in each group. The motivation for (b) lies in the fact that since our method uses communication features, the recommendation performance is going to be different for different users. Similarly for (c), we want to be able to analyze our method at the group level, since for some groups we might be able to yield recommendations better than others. Note, all features are used in both the methods.

The results yield interesting insights (Figure 13.2). First, we observe that precision gradually decreases, while recall increases with increase in k . This is because with increase in the number of recommended groups, more false positives are

likely to be returned (because recall, the average number of groups per image is three), resulting in low precision. While, more ground truth groups are likely to be returned for larger values of k , yielding high recall. Second, maximum precision and recall occur in the group-based case, while minimum for the user-based case. This is explained by the fact that groups are likely to comprise images which are consistent content-wise or tag feature-wise. Whereas, users often associate their images to different groups due to personal preferences, apart from image content, tags or commenting behavior, resulting in diversity of image-group association. Third, comparing with the baseline method k -NN our method seems to yield higher precision and recall; mean precision is 0.62 against 0.49 for k -NN; and mean recall is 0.69 against 0.59 for k -NN. Also, overall, we observe that the mean F1-measure (that combines precision and recall together) for our method is approx. 0.65, while for k -NN is 0.51. Thus our method outperforms the baseline method.

Figure 13.2: Evaluation of recommendations using different methods: user-based and group-based.

Evaluation of Feature Types

The results of evaluation of the different types of features extracted in this chapter have been shown in Table 13.2 and Figure 13.3. We observe that the combination of

Table 13.2: Evaluation of the three types of features (image content, tags and communication features) used for image characterization against our optimal method (all features).

FEATURE TYPE	PRECISION	RECALL	F1-MEASURE
Image Content features	0.43	0.48	0.46
Tag features	0.61	0.66	0.63
Communication features	0.57	0.64	0.61
Optimal method (all features)	0.62	0.69	0.65

all features yields the highest values of precision, recall and F1-measure. Among the three types of features, we observe that content features perform the worst while the tag features perform the best. This is explained by the nature of groups on Flickr. Several groups are organized along certain concept spaces / themes (e.g. “The South-west of United States”); as a result they consist of images which are visually quite diverse; however are likely to contain consistent tags like “Grand Canyon” or “Arizona” which might be reflected in the corresponding user activity over the past. Interestingly, communication features perform well; implying that comments do indeed impact users’ intent to associate groups to images.

Figure 13.3: Evaluation of different feature categories in the group recommendation process.

Discussion

Online social media are not mere repositories of diverse content. Hence group recommendation to images on social media is an extremely challenging problem unlike

traditional classification; because it needs to account for the inter-user interactions in the groups. However such interactions might always not be directly observable from the image content, tagging or communication activity of the users. Users could have intrinsic motivations affecting these interactions, which in turn might be responsible for associating an image to a group. Despite this, this chapter gives a novel approach to characterize images along several feature types and yields promising results against methods incorporating image content or tags alone.

13.7 Conclusions

In this chapter, we developed three kinds of features to characterize images in online social media: image content, user tagging activity and user communication activity. A group recommendation framework based on learning a latent space for the groups was developed which recommended k most likely groups to a given image. Experiments on the Flickr dataset indicated satisfactory results in recommending groups to images with a mean precision of 0.62 and a mean recall of 0.69, compared to 0.49 and 0.59 respectively for a k -NN based baseline framework. We conclude that user tagging and communication based characterization of images helps improve recommendation performance significantly against image content alone. Our recommendation framework also captures social interactions among users through user communication history which is central to online social media.

As future work, elaborate understanding of communication among users can help provide better recommendations. Moreover exploiting the social network of users to understand their mutual coupling can also improve recommendation performance.

BIBLIOGRAPHY

- Adamic, L. and E. Adar (2005, July). How to search a social network. *Social Networks* 27(3), 187–203.
- Adar, E. and L. A. Adamic (2005). Tracking information epidemics in blogspace. In *WI '05: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, Washington, DC, USA, pp. 207–214. IEEE Computer Society.
- Adar, E., D. S. Weld, B. N. Bershad, and S. S. Gribble (2007). Why we search: visualizing and predicting user behavior. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, New York, NY, USA, pp. 161–170. ACM.
- Agrawal, R., S. Rajagopalan, R. Srikant, and Y. Xu (2003). Mining newsgroups using networks arising from social behavior. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, New York, NY, USA, pp. 529–535. ACM.
- Airoldi, E. M. and K. M. Carley (2005, December). Sampling algorithms for pure network topologies: a study on the stability and the separability of metric embeddings. *SIGKDD Explor. Newsl.* 7, 13–22.
- Almeida, R. B. and V. A. F. Almeida (2004). A community-aware search engine. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, New York, NY, USA, pp. 413–421. ACM.
- Anagnostopoulos, A., R. Kumar, and M. Mahdian (2008). Influence and correlation in social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 7–15. ACM.
- Antweiler, W. and M. Z. Frank (2004). Is all that talk just noise? the information content of internet stock message boards.
- Arora, S. (1998). Polynomial time approximation schemes for euclidean traveling salesman and other geometric problems. *J. ACM* 45(5), 753–782.
- Bakshy, E., B. Karrer, and L. A. Adamic (2009). Social influence and the diffusion of user-created content. In *EC '09: Proceedings of the tenth ACM conference on Electronic commerce*, New York, NY, USA, pp. 325–334. ACM.
- Balfe, E. and B. Smyth (2004). Query mining for community based web search. In *WI '04: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, Washington, DC, USA, pp. 594–598. IEEE Computer Society.
- Barabasi, A. L., H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A*.
- Bass, F. M. (1969). A new product growth model for consumer durables. *Management Science* 15, 215–227.
- Bearman, P. S., J. Moody, and K. Stovel (2004, July). Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology* 110(1), 44–91.

- Beenen, G., K. Ling, X. Wang, K. Chang, D. Frankowski, P. Resnick, and R. E. Kraut (2004). Using social psychology to motivate contributions to online communities. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, CSCW '04, New York, NY, USA, pp. 212–221. ACM.
- Benevenuto, F., F. Duarte, T. Rodrigues, V. A. Almeida, J. M. Almeida, and K. W. Ross (2008). Understanding video interactions in youtube. In *Proceeding of the 16th ACM international conference on Multimedia*, MM '08, New York, NY, USA, pp. 761–764. ACM.
- Berger, C. R. and R. J. Calabrese (1975). Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Human Communication Research* 1(2), 99–112.
- Biel, J.-I. and D. Gatica-Perez (2009). Wearing a youtube hat: directors, comedians, gurus, and user aggregated behavior. In *Proceedings of the seventeen ACM international conference on Multimedia*, MM '09, New York, NY, USA, pp. 833–836. ACM.
- Bird, C., A. Gourley, P. Devanbu, M. Gertz, and A. Swaminathan (2006). Mining email social networks. In *MSR '06: Proceedings of the 2006 international workshop on Mining software repositories*, New York, NY, USA, pp. 137–143. ACM.
- Blei, D. M. and J. D. Lafferty (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, New York, NY, USA, pp. 113–120. ACM.
- Blei, D. M., A. Y. Ng, and M. I. Jordan (2003, March). Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022.
- Bollobas, B. (1998). *Modern graph theory*. New York: Springer-Verlag.
- Borgs, C., J. Chayes, M. Mahdian, and A. Saberi (2004). Exploring the community structure of newsgroups. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 783–787. ACM.
- Boydell, O. and B. Smyth (2006). Capturing community search expertise for personalized web search using snippet-indexes. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, New York, NY, USA, pp. 277–286. ACM.
- Briesemeister, L., P. Lincoln, and P. Porras (2003). Epidemic profiles and defense of scale-free networks. In *WORM '03: Proceedings of the 2003 ACM workshop on Rapid malware*, New York, NY, USA, pp. 67–75. ACM.
- Brzozowski, M. J., T. Hogg, and G. Szabo (2008). Friends and foes: ideological social networking. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, New York, NY, USA, pp. 817–820. ACM.
- Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* 2(2), 121–167.

- Burke, M., C. Marlow, and T. Lento (2009). Feed me: motivating newcomer contribution in social network sites. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, New York, NY, USA, pp. 945–954. ACM.
- Burt, R. S. (1982). Toward a structural theory of action: Network models of social structure, perception and action. *The American Journal of Sociology* 90(6), 1336–1338.
- Burt, R. S. (1995). *Structural Holes : The Social Structure of Competition*. Cambridge, Massachusetts: Harvard University Press.
- Burt, R. S. (2004). Structural holes and good ideas. *The American Journal of Sociology* 110(2), 349–399.
- Cha, M., A. Mislove, B. Adams, and K. P. Gummadi (2008). Characterizing social cascades in flickr. In *WOSP '08: Proceedings of the first workshop on Online social networks*, New York, NY, USA, pp. 13–18. ACM.
- Cha, M., A. Mislove, and K. P. Gummadi (2009). A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, WWW '09, New York, NY, USA, pp. 721–730. ACM.
- Chen, H.-C. J., M. Magdon-Ismael, M. Goldberg, and W. A. Wallace (2007). Inferring agent dynamics from social communication network. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, New York, NY, USA, pp. 36–45. ACM.
- Chen, H.-M., M.-H. Chang, P.-C. Chang, M.-C. Tien, W. H. Hsu, and J.-L. Wu (2008). Sheepdog: group and tag recommendation for flickr photos by automatic search-based learning. In *MM '08: Proceeding of the 16th ACM international conference on Multimedia*, New York, NY, USA, pp. 737–740. ACM.
- Chen, S. S. and P. S. Gopalakrishnan (1998). Clustering via the bayesian information criterion with applications in speech recognition. In *ICASSP '98: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Volume 2, pp. 645–648. IEEE.
- Chi, Y., S. Zhu, X. Song, J. Tatemura, and B. L. Tseng (2007). Structural and temporal analysis of the blogosphere through community factorization. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 163–172. ACM.
- Chin, A. and M. Chignell (2007). Identifying active subgroups in online communities. In *CASCON '07: Proceedings of the 2007 conference of the center for advanced studies on Collaborative research*, New York, NY, USA, pp. 280–283. ACM.
- Choudhury, T., C. B. B.-S. and A. Pentland (2003). Learning communities: Connectivity and dynamics of interacting agents. In *Proceedings of the International Joint Conference on Neural Networks*, pp. 2797 – 2802.

- Cialdini, R. B. and N. J. Goldstein (2004, February). Social influence: Compliance and conformity. *Annual Review of Psychology* 55, 591–621.
- Coleman, J. (1998, August). *Foundations of Social Theory*. Belknap Press of Harvard University Press.
- Cooper, C. and A. Frieze (2003). A general model of web graphs. *Random Struct. Algorithms* 22(3), 311–335.
- Cortes, C., D. Pregibon, and C. Volinsky (2003, December). Computational methods for dynamic graphs. *Journal of Computational and Graphical Statistics* 12(4), 950–970.
- Coyle, M. and B. Smyth (2007). On the community-based explanation of search results. In *IUI '07: Proceedings of the 12th international conference on Intelligent user interfaces*, New York, NY, USA, pp. 282–285. ACM.
- Crandall, D., D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri (2008). Feedback effects between similarity and social influence in online communities. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 160–168. ACM.
- Dacin, T. M., M. J. Ventresca, and B. D. Beal (1999). The embeddedness of organizations: Dialogue and directions. *J. Management* 25(3), 317–356.
- D’Amore, R. (2004). Expertise community detection. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp. 498–499. ACM.
- De Choudhury, M., H. Sundaram, A. John, and D. Seligmann (2008a). Dynamic prediction of communication flow using social context. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, New York, NY, USA, pp. 49–54. ACM.
- De Choudhury, M., H. Sundaram, A. John, and D. D. Seligmann (2007). Contextual prediction of communication flow in social networks. In *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, Washington, DC, USA, pp. 57–65. IEEE Computer Society.
- De Choudhury, M., H. Sundaram, A. John, and D. D. Seligmann (2008b). Can blog communication dynamics be correlated with stock market activity? In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, New York, NY, USA, pp. 55–60. ACM.
- De Choudhury, M., H. Sundaram, A. John, and D. D. Seligmann (2008c). Multi-scale characterization of social network dynamics in the blogosphere. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, New York, NY, USA, pp. 1515–1516. ACM.
- De Choudhury, M., H. Sundaram, A. John, and D. D. Seligmann (2009a). Predicting efficient communication paths in social networks leveraging user context. *ACM Transactions on The Web*. Under review.

- De Choudhury, M., H. Sundaram, A. John, and D. D. Seligmann (2009b). Social synchrony: Predicting mimicry of user actions in online social media. In *Social-Com '09: Proceedings of the 2009 International Conference on Social Computing*, Washington, DC, USA, pp. 151–158. IEEE Computer Society.
- De Choudhury, M., H. Sundaram, A. John, and D. D. Seligmann (2009c). What makes conversations interesting?: themes, participants and consequences of conversations in online social media. In *WWW '09: Proceedings of the 18th international conference on World wide web*, New York, NY, USA, pp. 331–340. ACM.
- De Choudhury, M., H. Sundaram, A. John, and D. D. Seligmann (2010). “birds of a feather”: How does attribute homophily impact information diffusion in social media? In *KDD '10: 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Under review.
- De Choudhury, M., H. Sundaram, A. John, and D. D. Seligmann (2011). Extraction, characterization and utility of prototypical communication groups in the blogosphere. *ACM Transactions on Information Systems*. in press, forthcoming.
- De Choudhury, M., H. Sundaram, Y.-R. Lin, A. John, and D. D. Seligmann (2009). Connecting content to community in social media via image content, user tags and user communication. In *ICME 2009. IEEE International Conference on Multimedia and Expo, 2009.*, Washington, DC, USA, pp. 1238 – 1241. IEEE Computer Society.
- Delre, S.A., J. W. and M. Janssen (2007). Diffusion dynamics in small-world networks with heterogeneous consumers. *Comput. Math. Organ. Theory* 12(2), 185–202.
- Dixon, M. and V. J. Roscigno (2003, May). Status, networks, and social movement participation: The case of striking workers. *The American Journal of Sociology* 108(6), 1292–1327.
- Domingos, P. and M. Richardson (2001). Mining the network value of customers. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 57–66. ACM.
- Doreian, P. (1980). Linear models with spatially distributed data. *Sociological Methods and Research* 9, 29–60.
- Du, N., B. Wu, X. Pei, B. Wang, and L. Xu (2007). Community detection in large-scale social networks. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, New York, NY, USA, pp. 16–25. ACM.
- Dubinko, M., R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins (2006). Visualizing tags over time. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, New York, NY, USA, pp. 193–202. ACM.
- Duda, R. O., P. E. Hart, and D. G. Stork (2000, November). *Pattern Classification (2nd Edition)* (2 ed.). Wiley-Interscience.

- Eagle, N., A. Pentland, and D. Lazer (2009). Inferring social network structure using mobile phone data. *PNAS* 106(36), 15274–15278.
- Ehrlich, K., C.-Y. Lin, and V. Griffiths-Fisher (2007). Searching for experts in the enterprise: combining text and social network analysis. In *GROUP '07: Proceedings of the 2007 international ACM conference on Supporting group work*, New York, NY, USA, pp. 117–126. ACM.
- Falkowski, T., J. Bartelheimer, and M. Spiliopoulou (2006). Mining and visualizing the evolution of subgroups in social networks. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, Washington, DC, USA, pp. 52–58. IEEE Computer Society.
- Feld, S. L. (1981). The focused organization of social ties. *American Journal of Sociology* 86(5), 1015–1035.
- Fernandess, Y. and D. Malkhi (2008). On spreading recommendations via social gossip. In *SPAA '08: Proceedings of the twentieth annual symposium on Parallelism in algorithms and architectures*, New York, NY, USA, pp. 91–97. ACM.
- Fiore, A. T. and J. S. Donath (2005). Homophily in online dating: when do you like someone like yourself? In *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, New York, NY, USA, pp. 1371–1374. ACM.
- Fisher, D. and P. Dourish (2004). Social and temporal structures in everyday collaboration. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, New York, NY, USA, pp. 551–558. ACM.
- Frank, O. (1978). Sampling and estimation in large social networks. *Social Networks* 1(91), 101.
- Galaskiewicz, J. and W. Bielefeld (1998). *Nonprofit organizations in an age of uncertainty: A study of organizational change*. New York: Aldine de Gruyter.
- Garg, N. and I. Weber (2008). Personalized, interactive tag recommendation for flickr. In *RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems*, New York, NY, USA, pp. 67–74. ACM.
- Gill, P., M. Arlitt, Z. Li, and A. Mahanti (2007). Youtube traffic characterization: a view from the edge. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, New York, NY, USA, pp. 15–28. ACM.
- Girvan, M., D. S. Callaway, M. E. J. Newman, and S. H. Strogatz (2002, Mar). Simple model of epidemics with pathogen mutation. *Phys. Rev. E* 65(3), 031915.
- Gjoka, M., M. Sirivianos, A. Markopoulou, and X. Yang (2008). Poking facebook: characterization of osn applications. In *WOSP '08: Proceedings of the first workshop on Online social networks*, New York, NY, USA, pp. 31–36. ACM.
- Goldenberg, J., L. B. and E. Muller (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters* 12(3), 211–223.

- Gómez, V., A. Kaltenbrunner, and V. López (2008). Statistical analysis of the social network and discussion threads in slashdot. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, New York, NY, USA, pp. 645–654. ACM.
- Gonzalez, M.C., L. P. and H. Herrmann (2006). Model of mobile agents for sexual interaction networks. *European Physics Journal B* 49, 371–376.
- Goyal, A., F. Bonchi, and L. V. Lakshmanan (2008). Discovering leaders from community actions. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, New York, NY, USA, pp. 499–508. ACM.
- Granovetter, M. S. (1973). The strength of weak ties. *The American Journal of Sociology* 78(6), 1360–1380.
- Gruhl, D., R. Guha, R. Kumar, J. Novak, and A. Tomkins (2005). The predictive power of online chatter. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, New York, NY, USA, pp. 78–87. ACM.
- Gruhl, D., R. Guha, D. Liben-Nowell, and A. Tomkins (2004). Information diffusion through blogspace. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, New York, NY, USA, pp. 491–501. ACM.
- Gyöngyi, Z., H. Garcia-Molina, and J. Pedersen (2004). Combating web spam with trustrank. In *VLDB '04: Proceedings of the Thirtieth international conference on Very large data bases*, pp. 576–587. VLDB Endowment.
- Hammer, M. (1980). Social access and the clustering of personal connections. *Social Networks* 2(4), 305–325.
- Haythornthwaite, C. (1996). Social network analysis: An approach and technique for the study of information exchange. *Library & Information Science Research* 18(4), 323–342.
- Hofmann, T. (1999). Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp. 50–57. ACM.
- Honeycutt, C. and S. C. Herring (2009). Beyond Microblogging: Conversation and Collaboration via Twitter. In *HICSS '09. 42nd Hawaii International Conference on System Sciences, 2009*, pp. 1–10.
- Hoppen, C. and N. Wormald (2008, May). Induced forests in regular graphs with large girth. *Comb. Probab. Comput.* 17, 389–410.
- Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine* 14(6), 1–4.
- Hu, M., A. Sun, and E.-P. Lim (2007). Comments-oriented blog summarization by sentence extraction. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, New York, NY, USA, pp. 901–904. ACM.

- Huang, J., Z. Zhuang, J. Li, and C. L. Giles (2008). Collaboration over time: characterizing and modeling network evolution. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, New York, NY, USA, pp. 107–116. ACM.
- Ino, H., M. Kudo, and A. Nakamura (2005). Partitioning of web graphs by community topology. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, New York, NY, USA, pp. 661–669. ACM.
- Iwata, T., K. Saito, and T. Yamada (2007). Modeling user behavior in recommender systems based on maximum entropy. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, New York, NY, USA, pp. 1281–1282. ACM.
- Iwata, T., T. Yamada, Y. Sakurai, and N. Ueda (2010). Online multiscale dynamic topic models. In *KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 663–672. ACM.
- Janssen, M. A. and W. Jager (2001, December). Fashions, habits and changing preferences: Simulation of psychological factors affecting market dynamics. *Journal of Economic Psychology* 22(6), 745–772.
- Jerrum, M. and A. Sinclair (1990, June). Fast uniform generation of regular graphs. *Theor. Comput. Sci.* 73, 91–100.
- Joyce, E. and R. E. Kraut (2006). Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication* 11, 2006.
- Jung, K. C., S. M. Yoon, and H. J. Kim (2000). Continuous hmm applied to quantization of on-line korean character spaces. *Pattern Recogn. Lett.* 21(4), 303–310.
- Kaltenbrunner, A., V. Gomez, and V. Lopez (2007). Description and prediction of slashdot activity. In *LA-WEB '07: Proceedings of the 2007 Latin American Web Conference*, Washington, DC, USA, pp. 57–66. IEEE Computer Society.
- Karger, D. R. (1994). Random sampling in cut, flow, and network design problems. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, STOC '94, New York, NY, USA, pp. 648–657. ACM.
- Kempe, D., J. Kleinberg, and E. Tardos (2003). Maximizing the spread of influence through a social network. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146.
- Kennedy, L., M. Naaman, S. Ahern, R. Nair, and T. Rattenbury (2007). How flickr helps us make sense of the world: context and content in community-contributed media collections. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, New York, NY, USA, pp. 631–640. ACM.
- Kennedy, L. S. and M. Naaman (2008). Generating diverse and representative image search results for landmarks. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, New York, NY, USA, pp. 297–306. ACM.

- Kleinberg, J. (2002). Bursty and hierarchical structure in streams. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 91–101. ACM.
- Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM* 46(5), 604–632.
- Klov Dahl, A., Z. Dhofier, G. Oddy, J. O'Hara, S. Stoutjesdijk, and A. Whish (1977). Social networks in an urban area: First canberra study. *Journal of Sociology* 13(2), 169–172.
- Koh, J., Y.-G. Kim, B. Butler, and G.-W. Bock (2007, February). Encouraging participation in virtual communities. *Commun. ACM* 50, 68–73.
- Kondor, R. and J. Lafferty (2002). Diffusion kernels on graphs and other discrete structures. In *Proc. Int. Conf. on Machine Learning*, pp. 315–322.
- Kossinets, G., J. Kleinberg, and D. J. Watts (2008). The structure of information pathways in a social communication network. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 435–443. ACM.
- Kossinets, G. and D. J. Watts (2009, September). Origins of homophily in an evolving social network. *American Journal of Sociology* 115(2), 405–450.
- Kumar, R., J. Novak, P. Raghavan, and A. Tomkins (2003). On the bursty evolution of blogspace. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, New York, NY, USA, pp. 568–576. ACM.
- Kumar, R., J. Novak, and A. Tomkins (2006). Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 611–617. ACM.
- Lampe, C. and E. Johnston (2005). Follow the (slash) dot: effects of feedback on new members in an online community. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, GROUP '05, New York, NY, USA, pp. 11–20. ACM.
- Lerman, K. and A. Galstyan (2008). Analysis of social voting patterns on digg. In *WOSP '08: Proceedings of the first workshop on Online social networks*, New York, NY, USA, pp. 7–12. ACM.
- Leskovec, J., L. A. Adamic, and B. A. Huberman (2007). The dynamics of viral marketing. *ACM Trans. Web* 1(1), 5.
- Leskovec, J., L. Backstrom, and J. Kleinberg (2009). Meme-tracking and the dynamics of the news cycle. In *KDD '09: Proceedings of the 15th ACM SIGKDD*, New York, NY, USA, pp. 497–506. ACM.
- Leskovec, J., L. Backstrom, R. Kumar, and A. Tomkins (2008). Microscopic evolution of social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 462–470. ACM.

- Leskovec, J. and C. Faloutsos (2006). Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 636. ACM.
- Leskovec, J. and E. Horvitz (2008). Planetary-scale views on a large instant-messaging network. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, New York, NY, USA, pp. 915–924. ACM.
- Leskovec, J., J. Kleinberg, and C. Faloutsos (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 187. ACM.
- Leskovec, J., A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance (2007). Cost-effective outbreak detection in networks. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 420–429. ACM.
- Leskovec, J., M. Mcglohon, C. Faloutsos, N. Glance, and M. Hurst (2007). Cascading behavior in large blog graphs. In *In SDM*.
- Li, B., S. Xu, and J. Zhang (2007). Enhancing clustering blog documents by utilizing author/reader comments. In *ACM-SE 45: Proceedings of the 45th annual southeast regional conference*, New York, NY, USA, pp. 94–99. ACM.
- Liben-Nowell, D. and J. Kleinberg (2008). Tracing information flow on a global scale using internet chain-letter data. *PNAS* 105(12), 4633–4638.
- Liben-Nowell, D., J. Novak, R. Kumar, P. Raghavan, and A. Tomkins (2005, August). Geographic routing in social networks. *PNAS* 102(33), 11623–11628.
- Liljeros, F., C. R. Edling, and L. A. N. Amaral (2003). Sexual networks: implications for the transmission of sexually transmitted infections. *Microbes and Infection* 5(2), 189 – 196.
- Lin, Y.-R., H. Sundaram, Y. Chi, J. Tatemura, and B. L. Tseng (2007). Blog community discovery and evolution based on mutual awareness expansion. In *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, Washington, DC, USA, pp. 48–56. IEEE Computer Society.
- Ling, X., Q. Mei, C. Zhai, and B. Schatz (2008). Mining multi-faceted overviews of arbitrary topics in a text collection. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 497–505. ACM.
- Liu, Y., X. Huang, A. An, and X. Yu (2007). Arsa: a sentiment-aware model for predicting sales performance using blogs. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp. 607–614. ACM.
- Liu, Z. and R. Laganière (2007). Phase congruence measurement for image similarity assessment. *Pattern Recogn. Lett.* 28(1), 166–172.

- Lomi, A. and P. Pattison (2006). Manufacturing relations: An empirical study of the organization of production across multiple networks. *Organ. Sci.* 17(3), 313–332.
- Lorenz, D. and D. Raz (1999). *A simple efficient approximation scheme for the restricted shortest paths problem*. Bell Labs Technical Memorandum.
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision* 60(2), 91–110.
- Loy, G. and A. Zelinsky (2003). Fast radial symmetry for detecting points of interest. *IEEE Trans. Pattern Anal. Mach. Intell.* 25(8), 959–973.
- Lu, Y., C. Zhai, and N. Sundaresan (2009). Rated aspect summarization of short comments. In *WWW '09: Proceedings of the 18th international conference on World wide web*, New York, NY, USA, pp. 131–140. ACM.
- Mani, A. and H. Sundaram (2007). Modeling user context with applications to media retrieval. *Multimedia Systems* 12(4), 339–353.
- McDonald, D. W. (2003). Recommending collaboration with social networks: a comparative evaluation. In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, New York, NY, USA, pp. 593–600. ACM.
- McPherson, M., L. S. Lovin, and J. M. Cook (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27(1), 415–444.
- McPherson, M. and L. Smith-Lovin (1987). Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American sociological review* 52(3), 370–379.
- Mei, Q., D. Cai, D. Zhang, and C. Zhai (2008). Topic modeling with network regularization. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, New York, NY, USA, pp. 101–110. ACM.
- Mei, Q., X. Ling, M. Wondra, H. Su, and C. Zhai (2007). Topic sentiment mixture: modeling facets and opinions in weblogs. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, New York, NY, USA, pp. 171–180. ACM.
- Mei, Q., C. Liu, H. Su, and C. Zhai (2006). A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, New York, NY, USA, pp. 533–542. ACM.
- Mei, Q. and C. Zhai (2005). Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, New York, NY, USA, pp. 198–207. ACM.
- Milgram, S. (1967). The small world problem. *Psychology Today* 2, 60–67.

- Miller, A. D. and W. K. Edwards (2007). Give and take: a study of consumer photo-sharing culture and practice. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '07, New York, NY, USA, pp. 347–356. ACM.
- Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller (1990, January). Introduction to wordnet: an on-line lexical database*. *Int J Lexicography* 3(4), 235–244.
- Mishne, G. and N. Glance (2006). Leave a reply: An analysis of weblog comments. In *In Third annual workshop on the Weblogging ecosystem*.
- Mor Naaman, Jeffrey Boase, C.-H. L. (2010). Is it really about me? message content in social awareness streams. In *CSCW '10: Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, New York, NY, USA. ACM.
- Murphy, K. (2002, July). *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph. D. thesis, UC Berkeley, Computer Science Division.
- Naaman, M., J. Boase, and C.-H. Lai (2010). Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, New York, NY, USA, pp. 189–192. ACM.
- Nakajima, S., J. Tatemura, Y. Hino, Y. Hara, and K. Tanaka (2005). Discovering important bloggers based on analyzing blog threads. In *Proceedings of WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem*.
- Nallapati, R. M., A. Ahmed, E. P. Xing, and W. W. Cohen (2008). Joint latent topic models for text and citations. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 542–550. ACM.
- Negoescu, R. A. and D. Gatica-Perez (2008). Analyzing flickr groups. In *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval*, New York, NY, USA, pp. 417–426. ACM.
- Negoescu, R. A., A. C. Loui, and D. Gatica-Perez (2010). Kodak moments and flickr diamonds: how users shape large-scale media. In *Proceedings of the international conference on Multimedia*, MM '10, New York, NY, USA, pp. 1027–1030. ACM.
- Newcomb, T. M. (1961). *The acquaintance process*. New York, NY: Holt, Rinehart and Winston.
- Newman, M. E. (2001, July). Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys Rev E Stat Nonlin Soft Matter Phys* 64(1 Pt 2).
- Newman, M. E. J. (2002, Jul). Spread of epidemic disease on networks. *Physical Review E* 66(1), 016128+.

- Newman, M. E. J. (2003, Feb). Mixing patterns in networks. *Physical Review E* 67(2), 026126+.
- Ng, A. Y., M. I. Jordan, and Y. Weiss (2001). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pp. 849–856. MIT Press.
- Nov, O., D. Anderson, and O. Arazy (2010). Volunteer computing: a model of the factors determining contribution to community-based scientific research. In *Proceedings of the 19th international conference on World wide web, WWW '10*, New York, NY, USA, pp. 741–750. ACM.
- Nov, O., M. Naaman, and C. Ye (2010, March). Analysis of participation in an online photo-sharing community: A multidimensional perspective. *J. Am. Soc. Inf. Sci. Technol.* 61, 555–566.
- Ohira, M., N. Ohsugi, T. Ohoka, and K.-i. Matsumoto (2005). Accelerating cross-project knowledge collaboration using collaborative filtering and social networks. *SIGSOFT Softw. Eng. Notes* 30(4), 1–5.
- Palla, G., A.-L. Barabasi, and T. Vicsek (2007, April). Quantifying social group evolution. *Nature* 446(7136), 664–667.
- Potthast, M. (2009). Measuring the descriptiveness of web comments. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp. 724–725. ACM.
- Qiu, F. and Y. Cui (2010). An analysis of user behavior in online video streaming. In *Proceedings of the international workshop on Very-large-scale multimedia corpus, mining and retrieval, VLS-MCMR '10*, New York, NY, USA, pp. 49–54. ACM.
- Rabiner, L. R. (1990). A tutorial on hidden markov models and selected applications in speech recognition. pp. 267–296.
- Robins, G., T. Snijders, P. Wang, M. Handcock, and P. Pattison (2007). Recent developments in exponential random graph (p^*) models for social networks. *Social Networks* 29(2), 192–215.
- Rusmevichientong, P., D. Pennock, S. Lawrence, and C. Giles (2001). Methods for sampling pages uniformly from the world wide web. In *AAAI Fall Symposium on Using Uncertainty Within Computation*, pp. 121–128.
- Saito, K., N. R. and M. Kimura (2008). Prediction of information diffusion probabilities for independent cascade model. *Knowledge-based Intelligent Information and Engineering Systems* 3, 67–75.
- Schein, A. I., A. Popescul, L. H. Ungar, and D. M. Pennock (2002). Methods and metrics for cold-start recommendations. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp. 253–260. ACM.

- Schuth, A., M. Marx, and M. de Rijke (2007). Extracting the discussion structure in comments on news-articles. In *WIDM '07: Proceedings of the 9th annual ACM international workshop on Web information and data management*, New York, NY, USA, pp. 97–104. ACM.
- Shevade, B., H. Sundaram, and L. Xie (2007). Modeling personal and social network context for event annotation in images. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, New York, NY, USA, pp. 127–134. ACM.
- Shi, X., J. Zhu, R. Cai, and L. Zhang (2009). User grouping behavior in online forums. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 777–786. ACM.
- Sigurbjörnsson, B. and R. van Zwol (2008). Flickr tag recommendation based on collective knowledge. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, New York, NY, USA, pp. 327–336. ACM.
- Singh, V. K., R. Jain, and M. S. Kankanhalli (2009). Motivating contributors in social media networks. In *WSM '09: Proceedings of the first SIGMM workshop on Social media*, New York, NY, USA, pp. 11–18. ACM.
- Smith, M., V. Barash, L. Getoor, and H. W. Lauw (2008). Leveraging social context for searching social media. In *SSM '08: Proceeding of the 2008 ACM workshop on Search in social media*, New York, NY, USA, pp. 91–94. ACM.
- Song, X., Y. Chi, K. Hino, and B. Tseng (2007). Identifying opinion leaders in the blogosphere. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, New York, NY, USA, pp. 971–974. ACM.
- Song, X., B. L. Tseng, C.-Y. Lin, and M.-T. Sun (2006). Personalized recommendation driven by information flow. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp. 509–516. ACM.
- Soroka, V. and S. Rafaeli (2006). Invisible participants: how cultural capital relates to lurking behavior. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, New York, NY, USA, pp. 163–172. ACM.
- Stewart, A., L. Chen, R. Paiu, and W. Nejdl (2007). Discovering information diffusion paths from blogosphere for online advertising. In *ADKDD '07: Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising*, New York, NY, USA, pp. 46–54. ACM.
- Stumpf, M. P. H., C. Wiuf, and R. M. May (2005, March). Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proceedings of the National Academy of Sciences of the United States of America* 102(12), 4221–4224.

- Sun, E., I. Rosenn, C. Marlow, and T. Lento (2009, May). Gesundheit! modeling contagion through facebook news feed. In *Proceedings of the Third International Conference on Weblogs and Social Media*, San Jose, CA.
- Tang, J., J. Sun, C. Wang, and Z. Yang (2009). Social influence analysis in large-scale networks. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 807–816. ACM.
- Toivonen, R., J.-P. Onnela, J. Saramäki, J. Hyvönen, and K. Kaski (2006, November). A model for social networks. *Physica A: Statistical and Theoretical Physics* 371(2), 851–860.
- Valente, T. W. (1996, January). Social network thresholds in the diffusion of innovations. *Social Networks* 18(1), 69–89.
- Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc.
- Vijay Mahajan, E. M. and F. M. Bass (1990). New product diffusion models in marketing: A review and directions for research. *The Journal of Marketing* 54(1), 1–26.
- Viswanath, B., A. Mislove, M. Cha, and K. P. Gummadi (2009). On the evolution of user interaction in facebook. In *WOSN '09: Proceedings of the 2nd ACM workshop on Online social networks*, New York, NY, USA, pp. 37–42. ACM.
- Wan, X. and J. Yang (2007). Learning information diffusion process on the web. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, New York, NY, USA, pp. 1173–1174. ACM.
- Wasserman, S. and K. Faust (1994, November). *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)* (1 ed.). Cambridge University Press.
- Watts, D. J. (2002). A simple model of global cascades on random networks. *PNAS* 99, 5766–5771.
- Watts, D. J. (2003). *Six Degrees: The Science of a Connected Age*. New York, New York: W.W. Norton and Company, Inc.
- Watts, D. J. and P. S. Dodds (2007, December). Influentials, networks, and public opinion formation. *Journal of Consumer Research* 34(4), 441–458.
- Watts, D. J., P. S. Dodds, and M. E. J. Newman (2002, May). Identity and search in social networks. *Science* 296(5571), 1302 – 1305.
- Watts, D. J. and S. H. Strogatz (1998, June). Collective dynamics of 'small-world' networks. *Nature* 393(6684), 440–442.
- Wu, F. and B. A. Huberman (2008). Popularity, novelty and attention. In *EC '08: Proceedings of the 9th ACM conference on Electronic commerce*, New York, NY, USA, pp. 240–245. ACM.

- Xiao, Z., Z. Hou, C. Miao, and J. Wang (2005). Using phase information for symmetry detection. *Pattern Recogn. Lett.* 26(13), 1985–1994.
- Zachary, W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* 33, 452–473.
- Zhai, C. and J. Lafferty (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.* 22(2), 179–214.
- Zhang, F. (2000). An approach to incremental svm learning algorithm. In *ICTAI '00: Proceedings of the 12th IEEE International Conference on Tools with Artificial Intelligence*, Washington, DC, USA, pp. 268. IEEE Computer Society.
- Zhou, D., E. Manavoglu, J. Li, C. L. Giles, and H. Zha (2006). Probabilistic models for discovering e-communities. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, New York, NY, USA, pp. 173–182. ACM.
- Zhou, Y., X. Guan, Z. Zhang, and B. Zhang (2008). Predicting the tendency of topic discussion on the online social networks using a dynamic probability model. In *WebScience '08: Proceedings of the hypertext 2008 workshop on Collaboration and collective intelligence*, New York, NY, USA, pp. 7–11. ACM.

BIOGRAPHICAL SKETCH

Munmun De Choudhury is a doctoral candidate in Computer Science at the Arizona State University, Tempe. She is affiliated to the Reflective Living group at the School of Arts, Media & Engineering, ASU and has interned at three of the premier research labs in the country: Microsoft Research, Yahoo! Research and Avaya Labs. Munmun's research interests span the interdisciplinary area of computational social science. Specifically her focus centers around large-scale computational modeling of macroscopic social processes. In addition to publications in high quality conferences and journals (including, World Wide Web Conference, ACM Trans. On Information Systems), Munmun's research has received extensive press coverage ranging from the Wall Street Journal to the MIT Technology Review. She has also been a finalist of the Facebook fellowship program in 2010, and has been a Grace Hopper scholar thrice between 2007 and 2010.