Connecting Users with Similar Interests for Group Understanding

by

Xufei Wang

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

Approved March 2013 by the
Graduate Supervisory Committee:

Huan Liu, Chair
Subbarao Kambhampati
Hari Sundaram
Jieping Ye

ARIZONA STATE UNIVERSITY

May 2013

ABSTRACT

In most social networking websites, users are allowed to perform interactive activities. One of the fundamental features that these sites provide is to connecting with users of their kind. On one hand, this activity makes online connections visible and tangible; on the other hand, it enables the exploration of our connections and the expansion of our social networks easier. The aggregation of people who share common interests forms social groups, which are fundamental parts of our social lives. Social behavioral analysis at a group level is an active research area and attracts many interests from the industry.

Challenges of my work mainly arise from the scale and complexity of user generated behavioral data. The multiple types of interactions, highly dynamic nature of social networking and the volatile user behavior suggest that these data are complex and big in general. Effective and efficient approaches are required to analyze and interpret such data. My work provide effective channels to help connect the like-minded and, furthermore, understand user behavior at a group level. The contributions of this dissertation are in threefold: (1) proposing novel representation of collective tagging knowledge via tag networks; (2) proposing the new information spreader identification problem in egocentric soical networks; (3) defining group profiling as a systematic approach to understanding social groups. In sum, the research proposes novel concepts and approaches for connecting the like-minded, enables the understanding of user groups, and exposes interesting research opportunities.

DEDICATION

To my parents, my wife and our baby on the way

ACKNOWLEDGEMENTS

Thanks everyone who has helped me in the last five years.

First and foremost, I would like to thank my advisor, Dr. Huan Liu, for his consistent support during my Phd study. Professor Huan Liu helped to cultivate my research strength, provide constant invaluable advices in the last five years. Together with him, I learned to write my first paper and accomplish many challenging problems. More importantly, I learned from him the work attitude, ethic and disciplines that I will benefit for a life long. I feel very lucky to have him as my advisor, a teacher, and a life long friend.

I would like to express my gratitude to my thesis committee, professor Subbarao Kambhampati, professor Hari Sundaram and professor Jieping Ye. Dr. Kambhampati is more than a good advisor but a thinker. Dr. Sundaram is always able to provide guidance and new thinking on my research work. With Dr. Ye, I am more comfortable and capable of solving any research problem. I would particularly like to thank Dr. Wei Fan, Dr. Lei Wang and Dr. Baoxin Li for many thoughtful discussions.

The Data Mining and Machine Learning Lab backs me up on everything at anytime. It is a great pleasure working with them at our best times in our lives, particularly, Lei Tang, Zhen Zhao, Nitin Argarwal, Magdiel Galan, Geoffery Barbier, Bill Cole, Sai Moturu, Reza Zafarani, Shamanth Kumar, Huiji Gao, Xia Hu, Jiliang Tang, Zhuo Feng, Salem Alelyani, Ali Abbasi, Fred Morstatter, Pritam Gundecha and Isaac Jones. This would never be possible without their constant support and discussions.

Finally, to my parents and my lovely wife Liyuan Chen, for their understanding, patience and support in the last five years.

TABLE OF CONTENTS

Page

CHAPTER

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

INTRODUCTION

The growing popularity of social networking services enables online interactions between the social media users. Online activities have become an even more important ingredient in our social lives than ever before. From the individual's point of view, the need to connect with other people arises. Then social groups form naturally as people selectively connect with others, i.e., forming a community structure. Understanding social groups becomes an emergent task in social and behavior science, impacting many applications such as targeted advertisement, trend prediction, group dynamics modeling, etc.

1.1 Background

Social networking sites enable the building of social networks or connections among people who make friends, share interests, activities and their likes. On social networking sites, people can interact freely, sharing and discussing information about each other and their lives, using multiple types of media such as text, photos, videos, and taking various kinds of activities that are provided by these sites.

Social media appears in many different forms including blogs and microblogs, forums and message boards, social bookmarking, tagging, social networking, reviewing, questioning and answering, data and content sharing, etc. Many social networking sites serve some features mentioned above.

As more and more people are involved, social media has become an integral part of our social lives. Social media is now a platform for maintaining our relationships and serves as a new dimension of our identities. We also use social media as a new channel for self expression, for sharing interests, worries and needs, for communication and interaction with other people.

The rise of social media provides many research and business potentials in the years to come. It is a multidisciplinary research area which requires knowledge in social science, physics, mathematics and computational science, involving many different cultural aspects. Compared to data in traditional social science, the availability of the big behavioral data in social media presents new challenges in processing, analyzing and modeling, which is attributed by the complexity of the data. It also presents even greater opportunities to study online human behaviors at arbitrary resolutions, answer questions that are beyond reach in the past, gain insight and knowledge, make use of the data to improve productivity, explore business opportunities, etc.

One fundamental problem in analyzing the big behavioral data is to form groups of users with similar interests and to understand the unique characteristics of a group. They are two interconnected aspects of the problem. With the aggregation of the like-minded, social groups with specific characteristics form naturally. At a bigger scenario, ultimately, we attempt to understanding (both explicit and implicit) social groups. The knowledge could be harnessed to explain group formation and evolution, provide insights in designing and improving social services with practical significance.

The first aspect of the problem has become an important component in social media websites as they grow. For example, Facebook and LinkedIn provide a function (i.e., “People You May Know” or PYMK) to recommend other potential friends, Twitter and Google+ have a similar function called “Who to Follow”. The second aspect of the proposed problem is not yet well developed but is important in many different perspectives. Studying a group of users who have similar interests or tastes differs from studying individuals. It is usually impractical to study individual users as the social networking sites host hundereds of millions of users. Group analysis is more tangible without losing fine granularity. The group level analysis plays a key role in social science, “the founders of sociology claimed that the causes of social phenomena were to be found by studying groups rather than individuals" [40]. In practice, understanding social groups helps to provide insights into group formation and evolution, explaining various social phenomena, monitoring and tracking group dynamics, predicting future trends, behavioral targeting [71] and improving social services.

In this dissertation, we study the problem of connecting users who have similar interests in a social network, and propose novel approaches to facilitate the understanding of such groups. It is organized into two interrelated components: connecting the like-minded and understanding social groups. Next we introduce each component.

Connecting the Like-minded

In the social media era, users are consumers and producers simultaneously. As a consumer, users read articles and posts, receive messages and updates from their online contacts. As a producer, users write blogs, post updates, use tags to organize online resources, initiate interactions with their online contacts, etc. The changing role of social media users brings new challenges and opportunities in academia and industry.

The long tail distribution of social networks implies that the majority of users (e.g., 80%) have only few links. Similarly, users in the long tail produce less content than users in the short head. These challenges are not easily captured by the traditional data mining approaches (e.g., Collaborative Filtering). For instance, it is hard to follow links and find the like-minded users who are several hops away in the social network. It is meaningful to clarify the differences between the Collaborative Filtering and the proposed approach. Collaborative Filtering is designed for recommending items instead of people in social networks, assuming that similar people would likely to have similar tastes. Furthermore, user generated content (e.g., tags) is produced in a free style, meaning that synonyms and polysemy co-exist. Capturing the semantic correlation is not a trivial task in general. We enable the measure of semantic relevance among different terms by introducing the novel concept of tag network.

I demonstrate that the collective tagging knowledge can be captured by the introduction of tag networks. Furthermore, I demonstrate that identifying like-minded users via tag networks is a more effective methods than several baseline methods. Details will be discussed in Chapter 2.

Identifying users with similar interests via tag networks. We propose to utilize tag networks to effectively connect users with similar interests. A tag network is the "wisdom of a crowd" or collective wisdom. It organizes user generated tags into a graph, which is able to capture the semantic correlation between tags with different forms. Based on the tag network, we are able to infer who are the like-minded in a social network.

We set forward to studying the interaction among the like-minded, especially the spread of information. A direct important question is to identifying information spreaders, i.e., the key persons who have similar interests and relay information in a social network.

Identifying information spreaders. Use Twitter as an example, we propose to utilize user generated Tweets to find information spreaders in the Twitter follower networks. An information spreader is defined as a person who relays information (i.e., tweets) from her friends and share them with her own followers. A set of feasible approaches are proposed and compared with each other of their effectiveness of identifying information spreaders. Interesting findings are reported with detailed discussions.

Understanding Social Groups

Social groups form naturally for a multitude of reasons. A major reason is for some people to achieve common goals or satisfy some form of need. Besides, the edge distribution in a social network suggests a group structure with high concentrations within a set of neighboring nodes and low concentrations between these two sets of neighboring nodes. In this dissertation, groups and communities are used interchangeably. Examples of communities or groups in the real world include families, relatives, labmates, etc. A prominent feature of such community structure is that they are generally overlapped, i.e., one person belongs to one or more communities.

Hypotheses are in place to explain why communities are formed in social networks: “similarity breeds connection”, or the homophily effect [77]. The homophily principle states that people within a community are homogeneous such that they share a lot of commons in terms of sociodemographic, behavioral, and interpersonal characteristics.

Communities in social networks have different forms, i.e., communities are disjointed [111], overlapped [117], or hierarchical [118]. To identify meaningful communities, some methods make use of one type of interaction (e.g., links [109, 111], content [117]), while some other methods integrate mutiple types of (heterogeneous) relations [112]. Multiple mechanisms such as graph partition, objective function maximization, and statistical inferences are applied to detect communities, as well summarized in the survey [33].

Although community detection is an important task with various applications, it is even more important to understanding social groups, which helps to reveal group formation and evolution, identify group sentiment, predict future group dynamics, etc. Therefore, our work propose mechanisms to extract the unique characteristics of social groups. Details will be presented in Chapter 3.

Co-clustering users and tags. We propose a user-tag co-clustering framework, which takes advantage of networking information between users and tags in social media, to discover overlapping communities. In the network, users are connect to tags and tags to users, thus forming a bipartite graph. This explicit representation of users and tags in a same group, entailing who are interested in what, is useful for group understanding.

Co-clustering users and tags is a constraint scenario which demonstrates the feasibility of group understanding by leveraging community detection technologies. To generalize, with the presence of social groups (either explicit or implicit), we propose the group profiling as a systematic approach for group understanding.

Interpreting communities via group profiling. Group profiling is a task to extract most meaningful keywords that describe a group. Provided with representative keywords, we are able to understand what the group of people are interested in. We explore different group-profiling strategies to construct descriptions of a group. This research can assist network navigation, visualization and analysis, etc.

1.2 Problem Formulation

Let $G(U, E)$ represent a social network, where $U = \{u_1, u_2, \ldots, u_{|U|}\}$ is the set of users and $E = \{e_1, e_2, \ldots, e_{|E|}\}$ is the set of edges. The cardinality of a set represents the size of the set, e.g., $|U|$ is the number of users and $|E|$ is the number of edges. An edge or a connection could be directed (e.g., representing a following relationship) or undirected (e.g., representing a friend relationship).

A user could be connected with other users, or contacts. Contacts could be followers (i.e., connections from others), followees (i.e., connections to others), friends (i.e., undirected and positive links), foes (i.e., negative links), or a combination of some of the specific relationships, depending on the specific social network site. A user could generate certain content such as user profiles, bookmarks, posts, likes, blogs, tags, etc. A specific social networking website provides some of these features.

Given the necessary definitions, the problem of our study is defined as follows, *in a social network, we aim to connecting users with similar interests, furthermore, to understanding the unique characteristics of social groups with the most descriptive attributes.*

The problem consists of several interrelated subtopics which will be given specific definitions and discussed in detail in each chapter of the dissertation.

1.3 Contributions

Most work in this dissertation are closely connected to real applications in social media. They are developed to addressing real world needs, therefore some of them could be leveraged to improve user experiences in large scale social networking platforms such as LinkedIn, Facebook, Twitter, etc.

In addition, the proposed work address fundamental problems (e.g., identifying information spreaders, group profiling) in the scope of social media, contributing to the active research area in the near future. We believe that these work will have wide impact on relevant research areas including but not limited to collective knowledge representation and utilization, community detection and understanding, etc. Below is a summary of contributions of this dissertation:

- proposing tag networks as a novel representation of collective tagging knowledge and an effective approach to connecting the like-minded;
- proposing and solving the new problem of identifying information spreaders in egocentric social networks;
- proposing a co-clustering framework to both detect and interpret social groups; and
- proposing group profiling as a systematic approach to extract the most representative keywords for understanding social groups.

1.4 Organization

The dissertation consists of two major parts: connecting the like-minded and interpreting social groups. In Chapter 2, we demonstrate approaches that are based on user generated content to connect the like-minded. Two subtopics are studied. One utilizes tag networks and the other one utilizes user generated tweets in Twitter. In Chapter 3, we attempt to interpret online groups. We propose a user-tag co-clustering framework to detect and interpret communities. Then we generalize the group understanding problem via group profiling techniques. The related work is summarized in Chapter 4. We conclude the dissertation and point out promising research directions in Chapter 5.

Chapter 2

CONNECTING THE LIKE-MINDED

One of the most popular activities that social media users perform is to connect with other users, especially with those who share things in common. This is an active research area as the findings could be potentially applied to social networking websites for recommending future connections.

We study two sub-topics in this section. In the first task, we propose to connect users with similar interests in social media websites, utilizing tag networks as a new representation of collective tagging knowledge. In the second task, we study the novel problem of identifying information spreaders who have similar interests, relay information and share with their own contacts.

2.1 Learning from Tag Network Inference

Networking via social media is increasingly becoming an integral part of social life in which friend recommendation is an important feature. There are many successful applications of leveraging link information or connectivity in social networking environments. However, in identifying users with similar interests, there are also limitations that come with links: following links is inefficient and could be incomplete. For instance, the space complexity of an exhaustive search is exponential; an incomplete search risks not being able to find anybody of interest. The long tail users who only have few connections, could be difficult to find, and in certain scenarios, some of them are disconnected from the largest component of a social network. Therefore, link based approaches could fail.

Nonetheless, connecting people with similar interests is an important task. For instance, these like-minded could be treated as a source of future friends. Besides, in problem solving areas, we would have a better chance to solve an issue if we can find someone who has worked on similar tasks. In addition, understanding behavior

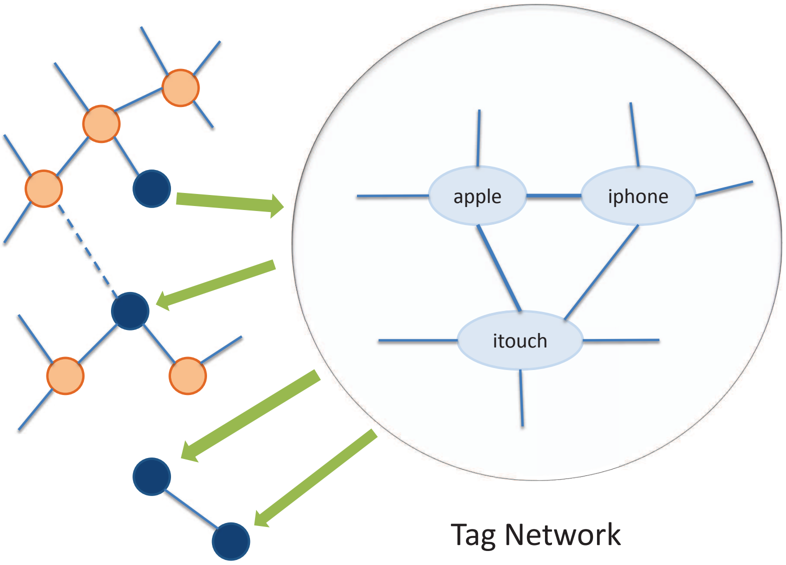


Figure 2.1: Connecting Like-minded Users in a Tag Network Approach

of users with similar interests could help gain better insights on interpreting group level behaviors. Connecting to "people like you" has psychological edges: "a sense of self-worth and fulfillment, being reassured of their worth and value, a sense of belonging to a community, the need to both seek help from and provide help to others, etc" [48].

Challenges of connecting users with similar interests are summarized below. First of all, people only have an egocentric view of the social network, i.e., users only see their immediate contacts. Secondly, the scale of a social network website like Facebook, Twitter, or LinkedIn makes manual search unrealistic. Therefore, inventing more effective and efficient tools is a necessity. Thirdly, as shown earlier, link information has innate limitations due to the long tail distribution of social networks.

Connecting via Tag Network Inference

We propose to connect users of like-minded via tag network inference. The basic idea is illustrated in Figure 2.1 in a simplified way. Nodes with different colors represent users of different kinds in a social network. Some users are in the largest component of the network, whereas other users are disconnected, thus either isolated or in small groups. A solid link represents two users are connected. Dashed link represents two users are not directly connected, but reachable from one to another. The four nodes highlighted in blue (dark) are, for example, fans of Apple products such as iPhone, iTouch, etc. Thus, the four users are deemed “like-minded”. The right part of the figure represents a tag network in which each node represents a tag, and the weight between two tags corresponds to users who use the two tags simultaneously.

Providing the “wisdom of the crowd”, tag network can be utilized to describe the semantic relationships among tags (more details later in this Chapter). Based on the tag network, the similarity between two users can be measured by their tag usage similarity. Take Figure 2.1 as an illustrative example, assume we want to connect other Apple fans to the upper left user in blue (dark). Instead of traversing links, we turn to the tag network, and return the other three Apple fans in the lower left.

Notations and Formulation

A social network $\mathcal{G} = (\mathcal{U}, \mathcal{E})$ is represented as an undirected graph, in which $\mathcal{U} = \{u_1, u_2, \ldots, u_n\}$ represents a set of n users and $\mathcal{E} = \{e_1, e_2, \ldots, e_\ell\}$ represents ℓ connections amongst the set of users. Each user subscribes to a certain number of tags. We denote the tag subscription relationship as a matrix $U \in R^{t \times n}$, in which each entry represents the number of times a tag is used by a given user. Let the number of unique tags associated to u_i be $\|u_i\|$. Denote $u_i^{interest}$ as a set of interests (e.g., categories specified by users on BlogCatalog) explicitly declared by the i-th user. Two users are said to be like-minded if they share some interests, e.g., both of them are fans of Apple iPhone.

$$u_i^{interest} \cap u_j^{interest} \neq \emptyset, 1 \leq i, j \leq n \tag{2.1}$$

However, two Apple fans may not necessarily use same tags, e.g., one person likes to use *iPhone* as a tag, the other person prefers to use *apple iPhone*.

Table 2.1: Notations

Notations	Description
$\mathcal{G}$	Social Network
U	User Tag matrix
W	Tag Network
u_i	The i-th user in $\mathcal{U}$
$u_i^{interest}$	Interests of the u_i
$\|u_i\|$	Number of unique tags of u_i
S_i	The set of top k most similar users of u_i
k	Number of users to be selected
K_β	Diffusion kernel with parameter β
MSI_j	Mean Shared Interests between u_i and the j-th user in S_i averaged on all u_is in $\mathcal{G}$

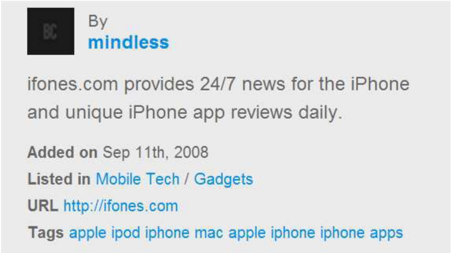
A tag network $W \in R^{t \times t}$ is a symmetric graph in which each node represents a tag that could be a word or a phrase, a non-zero entry w_{ij} in W represents the number of users who use the two corresponding tags simultaneously. A diffusion kernel K_β defined on a tag network is utilized to measure the tag similarities, where β is the parameter which controls the speed of diffusion. Table 2.1 summarizes the notations. The problem is then defined as follows:

- **Input:** Given a social network $\mathcal{G}$, a user u_i $(1 \leq i \leq n)$, a tag network $W \in R^{t \times t}$, and a scalar k.
- **Output:** top k most similar users from $\mathcal{G}$.

We next introduce the construction of tag networks from user generated tags, then design the novel approach, utilizing tag networks, for identifying users with similar interests.

Tag Network Construction

Tagging is an activity for organizing various objects like bookmarks and blogs for future browsing, management, and sharing using informal vocabularies. Tags can be words or phrases, and *informal* implies that they may not be found in any dictionary. Figure 2.2



By
mindless

ifones.com provides 24/7 news for the iPhone and unique iPhone app reviews daily.

Added on Sep 11th, 2008
Listed in Mobile Tech / Gadgets
URL http://ifones.com
Tags apple ipod iphone mac apple iphone iphone apps

Figure 2.2: A Snapshot of a Blog Description

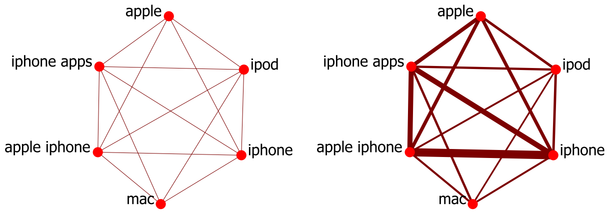
is a snapshot of a description for a blog on BlogCatalog with tags[1]. As shown in the figure, the blog, which is a news and review website on iPhone and iPhone applications, was added September 2008. It has a primary category *Mobile Tech* and a secondary category (or sub-category) *Gadgets*. Categories indicate the owner's interests. Six semantically relevant tags (i.e., *apple*, *ipod*, *iphone*, *mac*, *apple iphone*, *iphone apps*) are specified by the owner such that other readers can easily discover the topics of the blog without browsing hundreds of articles within it.

Tagging is a sort of knowledge that reflects labels on various web resources [42]. Collective wisdom emerges when many people's tag knowledge are aggregated together. The underlying hypothesis is that collective tagging naturally brings semantically relevant tags closer. For example, if two tags (e.g., iPhone and Apple iPhone) are used simultaneously by many people, there could be a semantic relevance between them. We represent the connectivity of tags in a network format: *Tag Network*.

We illustrate the steps to construct a tag network on the BlogCatalog data set.

- For each object (e.g. blog) and its descriptive tags, we connect the tags as a clique as shown in Figure 2.3 (a);

[1] http://www.blogcatalog.com/blogs/apple-iphone-news-and-app-reviews-ifonescom#



(a) Tag Network of a Blog (b) Tag Network of a Site

Figure 2.3: Examples of Tag Networks

- For each person, we combine all cliques corresponding to the objects she owns and form one or more **unweighted** tag networks, since her tags may or may not be connected in a tag network;

- We construct a **weighted** tag network by aggregating all tag networks belonging to each person. In the weighted tag network, tags correspond to the union of all users' vocabularies, and the weight of each link represents the number of users who use both tags simultaneously.

A snapshot of the weighted tag network is demonstrated in Figure 2.3 (b). Note that other tags and the corresponding links are not shown. We count *the number of users* instead of *the number of times* two tags cooccur as the weight of each link to discount bias from spam use rs, i.e., those who may use automated tools to assign the same group of tags many times. However, it could be interesting to consider user influence in assigning link weights as future work. Tags are available on most social networking sites in different forms such as user interests, bookmarks, labels, etc. Thus, the construction process can be easily adapted.

The tag network enables us to measure the similarities between any pair of tags within it. The simplest measure of similarity between two tags is the shortest path distance. However, the shortest path distance is susceptible to change in graph structure, i.e., newly added or removed tags and links might dramatically affect the distance between two nodes. Therefore, we prefer to average all path distances between two given tags for a more robust similarity measure, which leads to the idea of random walk with varying steps, equivalent to a diffusion kernel on a network [53, 97]. The concept of diffusion kernels is well established, thus readers who are familiar with it can simply skip.

Given a tag network $W \in R^{t \times t}$, where t represents the number of unique tags in a social network, we define a matrix L, whose negation is called Laplacian matrix, as follows,

$$L = W - D, \tag{2.2}$$

where D is a diagonal matrix in which the i-th diagonal entry corresponds to the summation of the entries in i-th column of matrix W. Let I represent the identity matrix, the diffusion kernel K_β of a tag network is defined as follows,

$$e^{\beta L} = \lim_{s \to \infty} (I + \frac{\beta L}{s})^s, \tag{2.3}$$

where $\beta \geq 0$ is a user specified parameter which controls the speed of diffusion. A larger β value means a faster information diffusion speed on the network; and there is no diffusion when β is set to 0. The diffusion kernel is positive semi-definite, thus is a valid kernel for measuring similarity between any pair of two tags [97].

The computation of a diffusion kernel requires an eigen-decomposition of L such that $L = V \Sigma V^\top$,

$$\begin{aligned} K_\beta &= e^{\beta L} \\ &= I + \beta L + \frac{(\beta L)^2}{2!} + \frac{(\beta L)^3}{3!} + \ldots \\ &= V\big(I + \beta\Sigma + \frac{\beta^2}{2!}\Sigma^2 + \frac{\beta^3}{3!}\Sigma^3 + \ldots)V^\top \\ &= V e^{\beta\Sigma} V^\top \end{aligned} \tag{2.4}$$

where the columns of V are the eigenvectors, Σ is a diagonal matrix whose diagonal entries are eigenvalues, and $(e^{\beta\Sigma})_{ii} = e^{\beta\Sigma_{ii}}$, other non-diagonal elements are all zeros.

Let u_i be a seed user, K_β be the kernel, the goal is to select the top k most relevant users in terms of similarity from the social network. The similarity between two users is aggregated on the pair-wise tag similarity given below,

$$sim(u_i, u_j) = \sum_{t \in u_i, t' \in u_j} \frac{u_i(t)}{\sqrt{\|u_i\|}} \cdot K_\beta(t, t') \cdot \frac{u_j(t')}{\sqrt{\|u_j\|}}, \tag{2.5}$$

where $u_i(t)$ represents the number of times the tag t is used by the i-th user and two normalization terms $\sqrt{\|u_i\|}$ and $\sqrt{\|u_j\|}$ are applied to the two users, respectively. The normalization is necessary because it prevents selecting spammers who use a large number of tags. But users who share more semantically relevant tags are credited thus we use the square root for both normalization terms. The intuition of Equation (2.5) is that two users are more *like-minded* if they share more semantically relevant tags.

Denote Z as a diagonal matrix whose diagonal entries are $Z_{ii} = \frac{1}{\sqrt{\|u_i\|}}$. We rewrite the similarity between u_i to other users in the social network as follows,

$$sim(u_i, \cdot) = u_i^\top \cdot K_\beta \cdot U \cdot Z \tag{2.6}$$

We discard the normalization term $\|u_i\|$ since it does not affect the final ranking. Without prior knowledge, determining parameter β is difficult in practice. However, tag network does provide heuristics for β selection. Tags that are frequently used simultaneously are semantically relevant, which is also the basic idea behind Latent Semantic Indexing (LSI) which leverages term co-occurrence in articles [25]. In a tag network, many semantically relevant tags are close or even immediate neighbors, thus it is desirable to select *small* values of βs.

Table 2.2: Statistics on BlogCatalog

Measure	BlogCatalog
Nodes	88,784
Edges	1,409,112
Average Contacts	49
Unique Tags	5,713
Average Tags	4.0

Data Collection and Experiments

BlogCatalog[2] is an online blog service which enables bloggers to register, manage, share, and connect blogs. A blog in BlogCatalog is associated with various pieces of information such as the categories that the blog is listed under, blog level tags, blog statistics such as the average rating and recent viewers, posts within the blog, and reviews from peer bloggers. A blogger also connects to other bloggers to form her social circle on BlogCatalog. A blogger's interests could be gauged by the categories (e.g. arts, business, education, etc) she publishes her blogs in. We obtained in total 60 categories in the processed BlogCatalog data set. We notice that a blogger can specify more than one category for each blog. On average each blogger lists their blog under 1.69 categories. In the rest of the paper, categories are treated as bloggers' interests. Bloggers in this social network form the largest component, thus any blogger can be connected to any other blogger through some intermediate bloggers. The social network is undirected. After post processing, we obtain a data set with 88,784 bloggers, 5,713 unique tags[3], and 60 categories. The BlogCatalog data set is shared with the public and can be downloaded from this link: http://dmml.asu.edu/users/xufei/datasets.html

[2] www.blogcatalog.com

[3] Tags that are used by less than 10 users are removed. This process helps to reduce noisy tags or typos in tags.

Baseline Methods

Two baseline approaches are selected. One is based on connectivity and the other one is based on latent semantic indexing.

Triadic Closure seeks to find similar users in terms of the number of mutual friends, and is solely based on links. This approach returns the top k people who are two hops away (friends of friends) in a social network. Note that it may return potential friends, but not necessarily return the most similar users.

Latent Semantic Indexing (LSI) is used to capture semantic correlation by applying Singular Value Decomposition (SVD). This approach computes the cosine similarity between an arbitrary pair of users in the latent space and can connect like-minded users who are far apart in a social network.

Evaluation Metrics

The quality is evaluated by the number of shared interests between the seed user and the selected users. More specifically, if the users selected by approach A share more interests with the seed user than those by approach B, intuitively, we say approach A is better.

On BlogCatalog data set, each individual has explicit categories (or interests) which serve as the ground truth for evaluation purposes. The metric, Mean Shared Interests (MSI), is formally defined in Equation (2.7),

$$MSI(j) = \frac{1}{n} \sum_{i=1}^{n} \| u_i^{interest} \cap S_i(j)^{interest} \|, \; 1 \leq j \leq k, \tag{2.7}$$

where u_i represents the seed user, $S_i(j)$ $(1 \leq j \leq k)$ represents the j-th recommended user for u_i, noting each user set S_i (ranked in descending order) depends on u_i. We average the shared interests over all users in a social network.

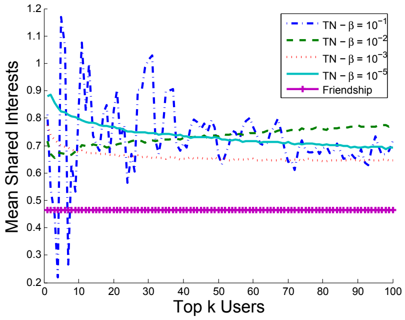


Figure 2.4: Shared interests v.s. selection of β

Comparative Study

The diffusion parameter β is sensitive to the outcomes. Figure 2.4 shows the MSI values with respect to different β values range from 10^{-1} to 10^{-5}. The performance stabilizes when βs are set to smaller than or equal to 10^{-5}. The x-axis represents the top 100 users sorted in descending order in terms of similarity with the seed user. The y-axis denotes the MSI values between the j-th selected user (excluding the seed user's immediate contacts) with the seed user. The plots suggest that the best performance is achieved when β is set to 10^{-5}, since we often recommend few users as candidates, e.g., 10 or 15. We also notice that large β values cause large variations. For instance, when β is set to 0.1, the performance is not stable. As a baseline measure, we compute the average shared interests between the user and her immediate neighbors, denoted by the lower solid line in Figure 2.4. The higher MSI values of the proposed approach suggest that more like-minded users could be returned.

Theoretically, in a connected network, there is a path from any user to any other user. Thus, it is possible to connect all like-minded users by following links. However, exhaustive search is expensive and inefficient for a contemporary social network which can have hundreds of millions of nodes. As an alternative, applying triadic closure only searches for candidates up to two hops away. Therefore, the search by triadic closure principle is incomplete.

For comparison, we include all three approaches: triadic closure, LSI, and tag network with a specified parameter. The results are plotted in Figure 2.5. The LSI approach does provide improvement to some extent compared to the baseline measure as indicated by Friendship. It should be noted that the best performance for LSI is obtained when the latent dimension is set to 200 for the studied data set. The proposed method outperforms the LSI approach significantly under t-test (p < 0.001). In computing the MSI values for above two approaches, the seed user's immediate contacts are excluded. The approach based on triadic closure is not as effective as the other two approaches, as indicated by the bottom curve in Figure 2.5. Comparing to the baseline methods (or measures), on average, the relative improvements of the tag network approach are 27%, 60%, and 108% for LSI, Friendship, and Triadic Closure, respectively.

Further Discussions Tag network and Latent Semantic Indexing are both capable of capturing the semantic correlation between tags, but diffusion on tag network appears to be more capable than LSI. The probable reasons for this are (1) the collective wisdom from the crowd brings the semantically relevant tags close to each other in terms of the number of hops; (2) although LSI also leverages the tag co-occurrence for dimension reduction, the diffusion kernel is more capable of measuring the similarity between any pair of two tags. We interpret the difference between LSI and diffusion on tag networks: LSI uses one path (i.e., the co-occurrence of two tags), whereas diffusion kernel combines all paths between any two tags (i.e., combining many different paths but discounted by distance).

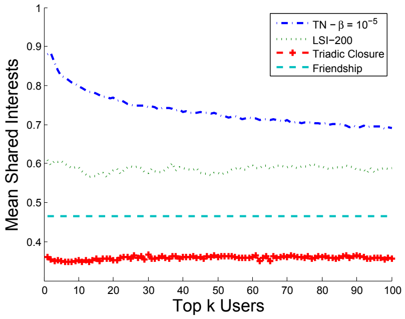


Figure 2.5: Shared interests v.s. different approaches

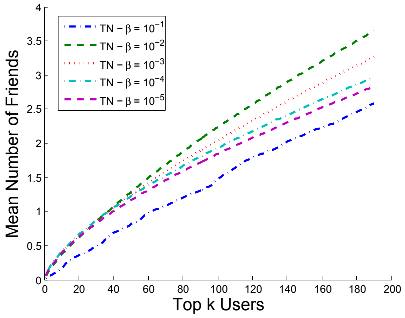


Figure 2.6: Correlation between friends and the like-minded

Correlation Analysis

In this section, we demonstrate the overlap between the true friends of the seed user and the top k most similar users. We find that a small set of selected users are actually the user's friends. The correlation between the friends and the returned top k users are presented in Figure 2.6. The x-axis represents top k most similar users sorted in descending order; y-axis represents the number of users who are actually friends, noting that y-axis values are averaged over all users in the social network. We found most similar users (around 98%), thoese who share interests with the seed users, are not her immediate friends. We evalute different kernels but they all show very similar performance.

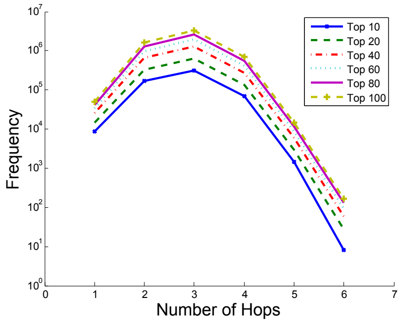


Figure 2.7: Distance distribution of the like-minded

Distance Distribution from a Seed User

We observe on the BlogCatalog data set that users that are multiple hops away could be like-minded. Thus, we compute the number of hops between the seed users and their top k most similar users. The computation is done by a breadth first search starting from a seed user, then each of the top k users is assigned the number of hops from the corresponding seed user. Finally we aggregate the number of users by hop distance from their corresponding seed users.

The distance distribution is presented in Figure 2.7, in which the curves from bottom to top represent top k (k = 10, 20, ..., 100) users who are considered. As shown in this figure, statistically, the majority of the most similar users are 2, 3, and 4 hops away. A small number of users who are 5 or 6 hops away from the seed users, (the diameter of the BlogCatalog social network is only 7) are also suggested as like-minded. The percentages of users with different hops from the seed users are summarized in Table 2.3. The immediate friends who are 1-hop away from a seed user account for less than 2%. The above results demonstrate that the tag network approach is capable of returning distant like-minded people for future interactions.

Table 2.3: Distance Distribution of Top k Candidates

# of Hops	1	2	3	4	5	6
Top 10	1.555%	20.197%	**55.825**%	12.160%	0.263%	0.002%
Top 50	1.062%	29.035%	**57.406**%	12.229%	0.264%	0.003%
Top 100	0.875%	28.528%	**58.072**%	12.260%	0.261%	0.003%

2.2 Identifying Information Spreaders in Egocentric Social Networks

The microblogging service Twitter has exploded in popularity in recent years by providing the ability for users to share information with one another in the form of short posts, called “tweets”. One feature that distinguishes Twitter from other social networking platforms is the ability to “retweet” another user’s tweet. Retweeting is a powerful way of disseminating information in the Twitter follower social network, becoming the key mechanism for information diffusion in Twitter [104]. Recently, a number of research efforts have studied the factors that affect retweeting [11, 104], retweet patterns and behaviors [59, 80], predicting retweets [82, 89, 131] and information diffusion [114, 126].

An important yet unaddressed question in retweet analysis is to identifying the people who retweet information from their friends (a.k.a. followees) and share with their own followers, or *the identifying information spreader problem*. A direct impact of this work is to increase user engagement at Twitter. When a user posts an update, Twitter can send messages (e.g., email or SMS) to the information spreaders, then they might follow back or their followers might join the discussion. The second impact is viral marketing. Later in this paper we demonstrate that information spreaders are not influentials. Investing in information spreaders rather than the influentials could be more effective in increasing the exposure of a product to more potential buyers. The next impact is to help understand the information diffusion in Twitter. Identifying the information spreaders discovers the backbone of information pathways, which helps visualize and understand the information flow in a network.

Next we define the concept of information spreader and point out the important difference between this work and previous work.

The Information Spreader Identification Problem

Previous studies show that the vast majority of users are information consumers [79]. However there exists some small set of users who are *information spreaders* who retweet from his or her friends and share information with his or her own followers. How information spreads from sources to the silent majority has been the subject of many research efforts [21, 65, 93, 115, 126]. However, identifying information spreaders in Twitter and other social networks is not formally defined. Furthermore, we demonstrate a set of feasible approaches for identifying information spreaders.

Though some prior problems may seem similar, there are substantial differences between prior work and ours. Those problems either operate at the global level (e.g., information diffusion and identifying influential people) or narrow the scope down to only one tweet (e.g., retweet prediction). The problem of information spreader identification asks *who* among a user's direct contacts spreads information.

Knowing who spreads ideas in social networking is important in many fields. By identifying these people, information diffusion can be expedited, access to information can be increased, new ideas can be adopted more quickly, and the backbone of information pathway can be discovered. This work proposes the problem of identifying information spreaders, shows that this problem is different from finding influential people, and empirically evaluates a set of feasible approaches.

Our primary focus is to understand retweet patterns between pairs of following users in the Twitter social network, the effectiveness of features which are originated from both the social network and user generated content, and various approaches in predicting who are the willing-to-retweet followers. Next we will introduce the necessary notations and the formal definition of the novel problem.

Notations and Formulation

We first introduce the notations to be used in this section. The Twitter social network can be modeled as a directed graph $G = \{U, E\}$, where $U = \{u_1, u_2, \ldots, u_n\}$ is the set of users and E is the following relationship between users. A typical Twitter user u has a set of followers $(Follower(u))$ and friends $(Friend(u))$ which is known as followees before. We denote contacts ($Contact(u)$) as the union of the user's followers and friends, that is,

$$Contact(u) = Follower(u) \cup Friend(u) \tag{2.8}$$

The friends, followers and contacts are called neighbors of a user as they are connected in a certain manner. The cardinality of a set represents the size of the set, e.g., $|Friend(u)|$ represents the number of friends of user u.

Common friends CFR refer to the set of users who are followed by two users u_i and u_j. Similarly, we define the common followers CFO and common contacts CCO as the users who are shared by the two corresponding user sets. That is,

$$\begin{aligned} CFR(u_i, u_j) &= Friend(u_i) \cap Friend(u_j) \\ CFO(u_i, u_j) &= Follower(u_i) \cap Follower(u_j) \\ CCO(u_i, u_j) &= Contact(u_i) \cap Contact(u_j) \end{aligned} \tag{2.9}$$

We aggregate all tweets that are owned by user u, then form a term-frequency vector $t(u)$, excluding stop words. Similarly, the set of hashtags and URLs that are associated to user u are represented as term-frequency vectors $ht(u)$ and $url(u)$, respectively.

Given a user u and her followers, our primary focus is to predict which of the followers would like to retweet her tweets, considering a wide range of features from

Table 2.4: Parameters Used to Data Collection

Country	Keywords/Hashtags	Geo-Boundary
Egypt	#egypt,#muslimbrotherhood,#tahrir,#mubarak, #cairo,#jan25,#july8,#scaf,#noscaf	(22.1,24.8),(31.2,34.0)
Syria	#syria,#assad,#aleppovolcano,#alawite,#homs	(32.8,35.9),(37.3,42.3)
Libya	#libya,#gaddafi,#benghazi,#brega,#misrata, #nalut,#nafusa,#rhaibat	(23.4,10.0),(33.0,25.0)
Bahrain	#bahrain,#bah	(50.4,25.8),(50.8,26.3)
Yemen	#yemen,#sanaa,#lbb,#taiz,#aden,#saleh, #hodeidah,#abyan,#zanjibar,#arhab	(12.9,42.9),(19.0,52.2)

the Twitter social network and user generated content. It can be modeled as a ranking, prediction, or regression problem depending on the specific context,

$$\begin{aligned} &\max_{\{f_i\}_{i=1}^{k}} \sum_{i=1}^{k} P(f_i|u) \\ &s.t.\;\; f_i \in Follower(u) \end{aligned} \tag{2.10}$$

Collection Methodology

In order to assemble our data set, we collected tweets, user profiles and network data through the Twitter API using the system described in [58]. The collection of data was restricted through the use of keywords, hashtags, and geographic regions. We collected more than $660,000$ users and 16 million tweets published from or concerned Egypt, Syria, Libya, Bahrain and Yemen. The tweets were crawled using the streaming API over a period of 7 months starting February 1st, 2011 and ending August 31st, 2011. A full list of the parameters used is presented in Table 2.4. Column 3 lists the geographic bounding box used to crawl all the geo-located tweets from each country in that region. The inspected Tweets during this period account for approximately 10% of all tweets hosted by Twitter[4].

As expected, the node degree distribution follows a power law distribution. Consistent with other studies on the Twitter network [59], only around 20% of links are

[4] We verifed this claim with Twitter's "firehose" data which cannot be directly used in this paper for legal reasons.

Table 2.5: Statistics of the Twitter Data Set

Measure	**Value**	**Measure**	**Value**
Users	666,168	Mean Friends	130.20
Mean Followers	130.20	Mean Contacts	217.09
Links	86,710,704	Bidir. Links	19.9%
Tweets	16,043,422	Retweets	3,874,449
URL	6,531,602	URL Ratio	40.33%
Hashtag	37,276,618	Hashtag Ratio	97.88%
Reply	472,160	Reply Ratio	3.98%
Mention	972,042	Mention Ratio	5.49%

reciprocated. We computed several other relevant statistics, in particular, the retweet ratio is around 24%, suggesting that information diffusion in the collected data set is prevalent. Around 40% of tweets contain URLs and interactive tweets only account for a small part of the data, around 4% and 5% of the tweets are replies and mentions, respectively. Table 2.5 summarizes many other data set statistics.

Pair-wise Retweet Analysis

An intuitive idea for identifying information spreaders is to look at the retweet history. Next, we present some interesting findings on aggregated retweeting behavior concerning retweet history in egocentric social networks. That is, a user and her followers, or pair-wised retweet analysis.

We take a closer look at the retweeting pairs, that is, the two involved users in an instance of retweeting. More than 75% of users only retweeted once in the entire 7 months of data collection and 95% of the users have less than or equal to 5 retweets in the whole duration. Figure 2.8 shows the retweeter count distribution. For each individual user, we compute the number of his or her followers who retweet at least once in the seven month time window. The distribution shows that more than 50% of users have been retweeted by only one follower and 80% of users have been retweeted by less than or equal to 5 followers. These two observations that are plotted in Figure 2.8 reveal that retweeting is not a daily activity for the vast majority of users.

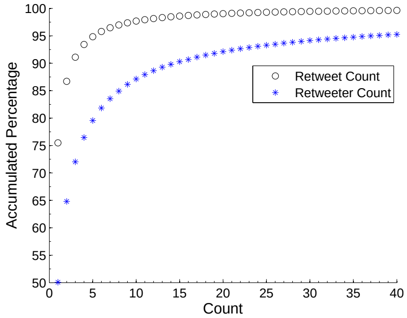


Figure 2.8: Retweet and Retweeter Count Distribution

We conducted an empirical analysis on retweet likelihood with respect to the users' retweeting history. The 7-month data is split into seven time frames by month; February, March, . . ., August. Then we study the correlation between retweet history and any future retweets, by comparing August retweet behavior to retweet behavior occurring in previous months.

The first empirical study reveals the extent to which users stop retweeting in the last month, August, compared to previous months of historical retweeting behavior. The measure *inactive ratio* is thus in place to represent the percentage of people who have at least one retweet in the previous months but stop retweeting in August. Results are demonstrated in Table 2.6, in which each row represents different length of historical retweeting data that is considered and the last column represents the percentage of people who stop retweeting in August. This table shows that retweet history only tells part of the users' retweeting "story": within the set of users who retweet from their friends in February, only 25.8% of them retweet again in August. Even when we consider the retweet history over all six months, over one third stop retweeting.

The second empirical study reveals how the active retweeter behaves in the last month considered in this paper, August of 2011. We compute the retweet likelihood in August with respect to the number of retweets that were performed in the first seven months. The distribution is plotted in Figure 2.9, in which the x-axis is log-scaled. A

Table 2.6: Retweet Inactive Ratio

Time Span	Test Month	Inactive Ratio
Feb	August	74.2%
Feb - Mar	August	66.7%
Feb - Apr	August	59.7%
Feb - May	August	56.9%
Feb - Jun	August	50.4%
Feb - Jul	August	36.2%

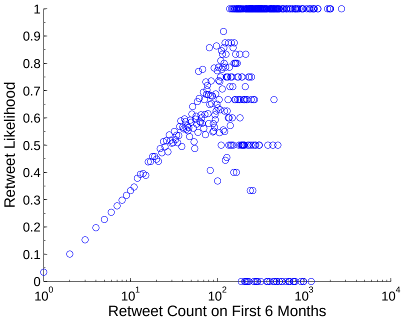


Figure 2.9: Retweet Likelihood Analysis.

roughly positive correlation (Pearson coefficient $r = 0.21$) between the retweet likelihood and the number of historical retweets is observed. If a user retweets a lot from the same friend (e.g., more than 100 retweets in the last six months), it is likely that she will retweet again from that friend in the future. However, 7.8% of the users who retweeted significantly in the last six months do not continue to retweet in the seventh month.

The three observations show that retweeting behavior is highly dynamic and ephemeral. Many people stop retweeting and other people start to retweet at any time. The study suggests that active retweeters are also likely to be information spreaders. However, the limitation of utilizing retweeting history for identifying information spreaders is obvious: the silent majority are infrequent retweeters and are infrequently retweeted by their followers. For these people, historical data is either absent or limited in usefulness, meaning more sophisticated approaches must be incorporated to identify information spreaders in egocentric networks.

Methods for Identifying Information Spreaders

In this section, we attempt to automatically rank a user's followers by their likelihood for future retweeting. Our hypothesis is that retweet behavior of a given user's followers can be learned from the follower's other online behavior.

We propose to do this by extracting features that may contribute to the follower's likelihood of retweeting. These features include user similarity, online interaction, structural features, and profile features. Some features are well discussed by prior work such as [82, 89, 104, 131]. These features are summarized in Table 2.7 with descriptions in the last column.

- *Proximity-based features* measure the similarity between an arbitrary pair of following users u_i and u_j, relative to the network topology. These features are extracted from the Twitter following network and thus give no indication of content of tweets or retweets. Features include common friends, common followers, common contacts, social status, etc.
- *Content-based features* measure the similarity of the user-generated content between two users. The set of features used in this paper are common hashtags, common URLs, and tweet similarity.
- *Interaction-based features* indicate the frequency that two persons interact with one another. We extract the number of replies and mentions between a pair of users as the interaction features.
- *Profile-based features* include statistics related to each user: the number of tweets; followees, followers and contact counts; the number of lists that user appears on; the language a person uses; and the account creation date.

Table 2.7: Feature Description

Group	Feature	Description
Proximity	Common Followers	The number of users who follow both users
	Common Friends	The number of users who are followed by both users
	Common Contacts	The union of followers and friends
	Mutual Link	Indicator of whether two users follow each other
	Social Status	PageRank values
Content	Common Hashtags	The number of common hashtags
	Common URL	The number of common URLs
	Tweet Similarity	The cosine similarity
Interaction	Reply	The number of replies
	Mention	The number of mentions
Profile	Status	The number of Tweets of a user
	Lists	The number of lists that belongs to a user
	Language	The preferred language of a user
	Account	The date that the user's account is created
	Friends	The number of friends
	Followers	The number of followers
	Contacts	The number of contacts

Feature Extraction

For each tweet, we extract the following information where possible: the owner of the tweet, the hashtag(s), URL(s), mentioned user(s) and the reply-to user. Then, we form the previously discussed term-frequency vectors $t(u)$, $ht(u)$, and $url(u)$. We found that an average Twitter user uses the same small set of tweet terms and hashtags repeatedly. However, URL usage statistics are very different. Although the average number of hashtags and URLs are relatively large, the majority of the users use very few of them, as indicated by the median numbers in Table 2.8. Most Twitter users have used certain amount of tweet terms and hashtags within tweets. The last column "NZ" (Not Zero) highlights the fact that hashtag usage is substantially more prevalent than URL usage.

Table 2.8: Feature Statistics

Measure	Unique		Duplicate		NZ
	Mean	**Median**	**Mean**	**Median**	
Terms	52.8	13	147.0	13	91.1%
Hashtag	9.1	4	52.4	4	92.7%
URL	14.1	1	16.4	1	52.9%

Methods for Ranking Followers

In this section, we summarize the set of approaches that are potentially suitable for ranking a user's followers by their likelihood of retweeting. All these methods assign a score to an arbitrary following relationship, i.e., $P(f_i|u) \in [0,1], f_i \in Follower(u)$. Some methods are very well developed but are also applicable in other tasks. To simplify notations, we always use the hashtag $ht(u)$ as an example to derive the proposed approaches. The definitions can be generalized to the other features easily. Assume u_i and u_j are two Twitter users that have a following relationship, e.g., u_i is a follower of u_j.

- **Shared Feature Counting.** Countable features in this data set include shared followers, followees, and contacts, shared hashtags and URLs. This approach is reasonable because shared features and retweet likelihood are positively correlated. However, the statistical results are not presented in this paper due to space limitations.

$$|ht(u_i) \cap ht(u_j)| \tag{2.11}$$

- **Jaccard Index** measures the extent to which two sets overlap. It is a normalized similarity measure.

$$\frac{|ht(u_i) \cap ht(u_j)|}{|ht(u_i) \cup ht(u_j)|} \tag{2.12}$$

- **Adamic/Adar Index** assigns more weights to shared features that are rarely used by other people [1]. We consider the hashtags and URLs that are used by Twitter users in the paper to compute this index. Let u_i and u_j be two users, z be a shared hashtag, $F(z)$ represents the number of users who used the feature z in the data set, the Adamic/Adar index between two users is given by

$$\sum_{z \in ht(u_i) \cap ht(u_j)} \frac{1}{\log F(z)} \tag{2.13}$$

We also consider a variation (i.e., Weighted Adamic/Adar Index) which takes into account the number of times that a hashtag has been shared by two users. Let z_{u_i} be the number times that a hashtag z is used by user u_i, we define the weighted Adamic/Adar index in this way

$$\sum_{z \in ht(u_i) \cap ht(u_j)} \frac{\min(z_{u_i}, z_{u_j})}{\log F(z)} \tag{2.14}$$

- **Tweet Similarity** is computed by modeling each user as a term-frequency vector. The similarity of two users is thus given by their cosine similarity.

$$\frac{t(u_i) \cdot t(u_j)}{\|t(u_i)\| \cdot \|t(u_j)\|} \tag{2.15}$$

- **Regression Models** are used to investigate the relationship between a dependent variable and one or more independent variables. In this paper, the dependent variable is the occurrence of retweeting (more details in next section), and the independent variables are the features with z-score normalization. Two regression models are considered: logistic regression and random forest regression.

  Logistic Regression [44] is widely used in many fields. Given a pair of two users f_i and u, $f_i \in Follower(u)$, the likelihood that a user f_i will retweet from user u can be estimated by

$$p(f_i|u) = \frac{1}{1 + e^{-(w^\top x_i + b)}},\ f_i \in Follower(u) \tag{2.16}$$

where w and b represent the weight of the features and offset, respectively, vector x_i is a feature vector that is associated with f_i and u.

Random Forest [12] is an ensemble learning method which consists of many decision trees and can be used in both prediction and regression tasks. It takes advantages of high accuracy, efficiency, and robustness to noise [92].

Table 2.9: Precision Performance of Various Methods

Method		**Top k Retrieved Followers**								
		1	5	10	20	30	40	50	100	500
Hashtag	Common Tags	.29	.18	.15	.13	.11	.11	.10	.09	.07
	Jaccard Index	.26	.16	.13	.11	.10	.10	.09	.08	.07
	Adamic/Adar	.33	.20	.16	.13	.12	.11	.10	.09	.07
	Weighted Adamic/Adar	.29	.18	.15	.12	.11	.11	.10	.09	.07
URL	Common URLs	.42	.25	.19	.15	.13	.12	.11	.09	.07
	Jaccard Index	.41	.24	.18	.14	.12	.11	.10	.09	.07
	Adamic/Adar	**.47**	**.28**	**.21**	**.16**	**.14**	.12	.11	.09	.07
	Weighted Adamic/Adar	.47	.28	.21	.16	.13	.12	.11	.09	.07
Neighbor	Common Friends	.09	.07	.07	.07	.07	.07	.07	.06	.06
	Jaccard Index (CFR)	.15	.10	.09	.08	.07	.07	.07	.07	.06
	Common Followers	.11	.09	.08	.08	.08	.07	.07	.07	.06
	Jaccard Index (CFO)	.15	.11	.10	.09	.08	.08	.08	.07	.06
	Common Contacts	.10	.08	.08	.07	.07	.07	.07	.07	.06
	Jaccard Index (CCO)	.16	.11	.09	.08	.08	.07	.07	.07	.06
Interaction	Reply	.15	.13	.13	.12	.12	.12	.12	.12	.12
	Mention	.18	.15	.14	.14	.14	**.14**	**.14**	**.13**	**.13**
Similarity	Tweet	.37	.21	.16	.13	.12	.11	.11	.10	.08
Regression	Logistic	.23	.15	.13	.11	.10	.10	.09	.08	.07
	Random Forest	.42	.24	.18	.14	.12	.11	.10	.09	.07

It is possible to design even more sophisticated models which integrate retweet history (e.g., the number of retweets) and other relevant features. In this work, we use the retweet history as ground truth, thus it is not used as a feature. In addition, our primary focus of this work is to introduce the information spreader prediction problem. Therefore, there are plenty of future research opportunities along this direction, e.g., designing more sophisticated approaches and verifying their effectiveness in prediction.

Experimental Results

We first introduce the ground truth construction and the measure that will be used to evaluate the performance of above methods. Then we present the experimental results.

Ground Truth Construction. The emergence of retweet between a user and her friends is deemed as ground truth. More specifically, if a user retweets at least once from her friends, then the directed link her to the friend is labeled as positive (i.e., '+1'), whereas, if no retweet occurs during the seven-month time frame, this link is labeled as negative (i.e., '–1'). Thus, for each user, followers are in two categories: the positive set in which all followers retweet at least once and the negative set in which all followers never retweet.

Evaluation Strategy. We evaluate the performance of different methods by the measure precision which is widely used in information retrieval tasks. More specifically, for each user, we rank the followers by their likelihood in retweeting from the user in descending order, then compare the top-k ranked users with the ground truth. In the following experiments, the number k is chosen as 1, 5, 10, 20, 30, 40, 50, 100 and 500. The precision that is averaged over all users in the Twitter social network is reported.

Experimental Results. Table 2.9 lists the precision performance of the different methods. Each column represents the top k users that are retrieved, e.g., column 1 indicates that we only consider the first user who is recommended by the corresponding methods.

The URL-based methods outperform the other methods, especially when the selected number k is small. For example, the best performance of URL-based approach is 11.9% better than the second best approach when $k = 1$. We also notice that different features have different strengths in predicting retweets: URL is the best, followed by tweet similarity and hashtags. Statistically, when comparing the best performances of

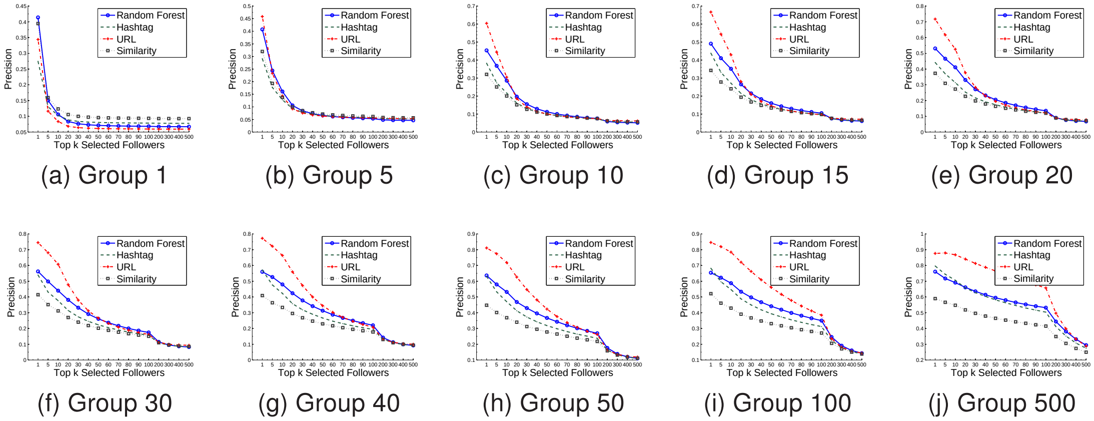
URL-based methods to those of feature based methods, the relative improvements are 30.5% and 72.4%, respectively. This result is consistent with prior studies that tweets with URLs are more likely to be retweeted by others [59, 82, 104].

There are several observations of the different treatments of the features: (1) the Adamic/Adar Index consistently outperforms the other approaches, (2) applying weights to the Adamic/Adar index does not improve the performance at all, suggesting information spreaders are likely to be infrequent retweeters, and (3) the performance of common feature counting is comparable to that of the Jaccard Index.

We found interaction features are not suitable for predicting which followers are likely to retweet because there are too few interactions in the data, e.g., only around 4% and 5% of the tweets are related to reply and mention, respectively. On the other hand, since more than 90% of users have at least one tweet, the tweet similarity is a relatively strong feature for retweet prediction.

Regression models that take all relevant features into account do not improve the retweet prediction any further. Logistic regression is less effective than the random forest approach. For both regression models, we randomly sample a certain amount of data as training data. Different sizes of instances (i.e., from $1,000$ to $20,000$) that are used to train the regression models are tried, and we find sizes are insensitive to the prediction performance. The results are not presented due to space limitation.

Determine the Best Strategy. For the studied Twitter users who have been retweeted at least once by their followers, the majority of them are retweeted by a very small number of followers. Figure 2.8 shows that around 50% of Twitter users are retweeted by only one follower. We assign users into different groups by the number of retweeters, then study which methods might be appropriate for diffrent user groups. For example, “group 1” represents the group in which users are retweeted by only one



(a) Group 1 (b) Group 5 (c) Group 10 (d) Group 15 (e) Group 20

(f) Group 30 (g) Group 40 (h) Group 50 (i) Group 100 (j) Group 500

Figure 2.10: Precision performance of different approaches

follower, and “group 10” represents that these users are retweeted by more than 5 but at most 10 followers. These groups have different characteristics and would deserve different treatments.

We consider four methods in this experiment: Adamic/Adar Index on hashtag, Adamic/Adar Index on URL, Tweet Similarity and Random Forest. Results are presented in Figure 2.10 in which each figure represents the precision performance on the corresponding user group. In order to return the top 10 most likely to retweet followers, we find in “group 1”, it is preferable to use Random Forest or Tweet Similarity for retweet prediction, for “group 5” and “group 10”, both Random Forest and URL-based approaches are good candidates. Otherwise, URL-based approach is preferred. We conjecture that for user groups with an extremely small number of retweeters, users might not share any of the single features (e.g., hashtag, URL), so it is imperative to take other information (e.g., tweets or other features) into account.

Are Information Spreaders Important Persons? Important Persons (IP) or influential persons in online social networks are usually characterized by their Pagerank values [59]. For each Twitter user, two ranked lists are present: the list of important per-

Table 2.10: Comparing information spreaders to important people

Measures and Methods		Top k Information Spreaders											
		1	5	10	15	20	25	30	35	40	45	50	100
nDCG	URL	.01	.03	.05	.06	.07	.08	.09	.10	.11	.12	.13	.18
	HashTag	.02	.05	.06	.07	.09	.10	.11	.12	.13	.14	.14	.20
	Similarity	.02	.03	.05	.06	.07	.08	.09	.10	.11	.12	.13	.18
	Random Forest	.01	.03	.05	.06	.07	.08	.09	.10	.11	.12	.13	.18
Jaccard Index	URL	.01	.03	.04	.06	.07	.09	.10	.11	.11	.12	.13	.17
	HashTag	.02	.03	.04	.06	.07	.08	.09	.10	.10	.11	.12	.16
	Similarity	.02	.03	.04	.05	.07	.08	.09	.10	.11	.11	.12	.16
	Random Forest	.01	.02	.04	.05	.06	.07	.08	.09	.10	.10	.11	.15

sons (IP), and the list of information spreaders (IS). Both ranked lists are in descending order either by their Pagerank values or the likelihood of retweeting. Comparing the IS list to the IP list is able to answer the question. Two measures the discounted cumulative gain (nDCG) and the Jaccard Index are used to quantify the difference between the two lists. In nDCG, the relevance score is binary and is determined in the following way: if the i-th user $IS(i)$ appears in the first i users in the IP list, the relevance value is 1, otherwise, it is 0. That is,

$$rel_i = \begin{cases} 1 & IS(i) \in \{IP(1), IP(2), \ldots, IP(i)\} \\ 0 & \text{otherwise} \end{cases} \tag{2.17}$$

Both measures fall between 0 and 1. Value 0 represents that two lists are completely different, and value 1 represents that the two lists are exactly the same. So if the information spreaders are equal to the important persons in each user' follower networks, we would expect that the mean nDCG value and Jaccard Index that are averaged over all Twitter users are close to 1. Results in Table 2.10 disprove this statement: in fact, the small values suggest that information spreaders are very unlikely to be the important persons in the egocentric networks, and even unlikely to be important persons globally. The results are obtained on the four best strategies: Adamic/Adar Index on URL and hashtag, tweet similarity and Random Forest.

In social networking websites, as more and more people are connected with their kinds, groups form naturally. In recent years, many work are dedicated to find groups from the network structure, user generated content or the combination of both, ignoring the essential task of understanding these social groups. Next we introduce our novel work for group understanding.

Chapter 3

UNDERSTANDING SOCIAL GROUPS

In social network analysis, analysis at group level attracts increasingly interests from social science and applied research such as behavioral targeting [71]. One of the urgent tasks is to understanding social groups that are formed in social media websites, which is the focus of this Chapter.

We first propose a novel framework to co-clustering users and tags into groups. This representation of groups entails who are interested in what, helping to answer questions such as “who these people are”, “why they form a group”, etc. To generalize, with the presence of social groups (both explicit and implicit), we propose the group profiling as a systimatic approach for group understanding.

3.1 Co-clustering Users and Tags

Community detection, which is generally based on link analysis, attempts to return a community structure, but ignores the interpretation of these communities. That is, there is no straightforward proof showing the focus of a group or what a group is about. On the other hand, social network sites usually provide both link information and various user generated content (e.g., tags). Can we obtain social groups with meaningful descriptions such that the groups can be easily interpreted.

We propose to co-clustering users and tags to obtain ‘meaningful’ community structure. Let us demonstrate the high level idea with a toy example which is shown in Figure 3.1 with two communities. Vertices $u_1 - u_5$ on the left represent users, $t_1 - t_4$ on the right represent tags and edges represent tag subscription relation between users and tags. Based on the graph structure, it is more reasonable to have two overlapping clusters $(u_1, u_2, u_3, t_1, t_2)$ and $(u_3, u_4, u_5, t_3, t_4)$, in which the users’ interests of each cluster can be summarized using t_1, t_2, and t_3, t_4, respectively.

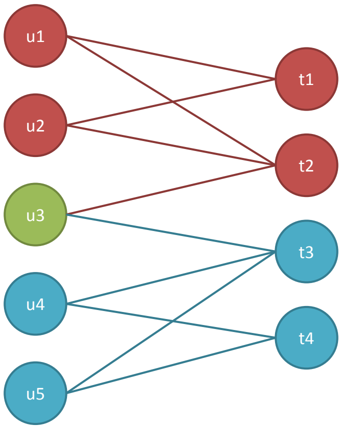


Figure 3.1: A 2-community toy example

A user usually has multiple types of relationships, therefore, groups usually overlap. An interesting observation in social life is that a connection is often associated with one affiliation [107]. For instance, a person likes or dislikes a movie, he/she is or is not a member of special interest group, and so on. Instead of clustering vertices, clustering edges seems more appropriate and obtains overlapping communities.

Notations and Formulation

Let $\mathcal{U}$ = $(u_1, u_2, \ldots, u_m)$ denote the user set, $\mathcal{T}$ = $(t_1, t_2, \ldots, t_n)$ the tag set. A community $C_i (1 \leq i \leq k)$ is a subset of users and tags, where k is the number of communities. As mentioned above, communities usually overlap, i.e., $C_i \bigcap C_j \neq \varnothing$ (1 $\leq$ i, j $\leq$ k). On the other hand, users and their subscribed tags form a user-tag matrix M, in which each entry $M_{ij} \in \{0, 1\}$ indicates whether user u_i subscribes to tag t_j. So it is reasonable to view a user as a sparse vector of tags, and each tag as a sparse vector of users.

We formulate the overlapping co-clustering problem as follows:

- **Input:** A user-tag subscription matrix $M_{N_u \times N_t}$, where N_u and N_t are the numbers of users and tags, respectively, and a scalar k.
- **Output:** k overlapping communities which consist of both users and tags.

The observation that a user is usually involved in several affiliations but *a link is usually related to one community* enlightens us to cluster edges instead of nodes. After obtaining edge clusters, communities can be recovered by replacing each edge with its two vertices, i.e., a node is involved in a community as long as any of its connection is in the community. Then the obtained communities are often highly overlapped.

In a user-tag network, each edge is associated with a user vertex u_i and a tag vertex t_p. If we take an edge-centric view by treating each edge as an instance, and two vertices as features, each edge can be represented as a sparse vector. The length of vector is $N_u + N_t$, in which the first N_u entries correspond to users, and the other N_t entries correspond to tags. For example, the edge between u_1 and t_1 in Figure 3.1 can be represented as $(1, 0, 0, 0, 0, 1, 0, 0, 0)$, in which only entries for vertices u_1 and t_1 are non-zero.

Communities that aggregate similar users and tags together can be detected by maximizing intra-cluster similarity, which is shown in Eq. (3.1).

$$\arg\max_{C} \frac{1}{k} \sum_{i=1}^{k} \sum_{x_j \in C_i} S_c(x_j, c_i) \tag{3.1}$$

where k is the number of communities, C = {C_1, C_2, . . ., C_k}, x_j represents an edge, and c_i is the centroid of community C_i. This formulation can be solved by using k-means. However, k-means is not efficient for large scale data sets. We propose to use EdgeCluster which is a k-means variant and is a scalable algorithm to extract communities for sparse social networks [107]. EdgeCluster maintains an indexing structure which significantly reduces the number of comparisons between instances and the centroids. It is reported to be able to cluster a sparse network with more than 1 million nodes into thousands of clusters in tens of minutes. The clustering quality is comparable to modularity maximization but the time and space reduction is significant. It should be noted that the network in [107] is 1-mode, but the user-tag network is 2-mode.

The expected density of the user-tag network is shown in Eq. (3.2), which guarantees an efficient solution by applying EdgeCluster.

$$\text{density} \approx \frac{\gamma - 1}{2 - \gamma} \cdot (d^{2-\gamma} - 1) \cdot \frac{1}{N_u} \tag{3.2}$$

where d is the maximum tag degree, N_u is the number of users in this graph and γ is the exponent of the power law distribution, which usually falls between 2 and 3 in social networks [84]. The maximum degree d is usually large in a power law distribution. Thus, the density is approximately inverse to the number of users.

A key step in clustering edges is to define edge similarity (centroids can be viewed as edges as well). Given two edges $e(u_i, t_p)$ and $e'(u_j, t_q)$ in a user-tag graph, the similarity between them can be defined in Eq. (3.3):

$$S_e(e, e') = \alpha S_u(u_i, u_j) + (1 - \alpha) S_t(t_p, t_q) \tag{3.3}$$

where $S_u(u_i, u_j)$ is the similarity between two users, and $S_t(t_p, t_q)$ is the similarity between two tags. This is reasonable because the edge similarity should be dependent on both user and tag similarity. And parameter α ($0 \leq \alpha \leq 1$) controls the weights of users and tags. Considering the balance between user similarity and tag similarity, α is set to 0.5 in our experiments.

In the following sections, we show that our framework can cover different similarity schemes.

Independent Learning Independence assumption is a popular way to simplify the problem we want to solve. If two tags are different, their similarity can be defined as 0, and 1 if they are the same. Thus the similarity can be represented by an indicator function which can be shown by Eq. (3.4).

$$\delta(m,n) = \begin{cases} 1 & m = n \\ 0 & m \neq n \end{cases} \tag{3.4}$$

The user-user similarity is also defined in a similar way. Cosine similarity is widely used in measuring the similarity between two vectors. Given two edges $e(u_i, t_p)$ and $e'(u_j, t_q)$, their cosine similarity can be rewritten in Eq. (3.5).

$$S_e(e, e') = \frac{1}{2}\left(\delta(u_i, u_j) + \delta(t_p, t_q)\right) \tag{3.5}$$

Following Eq. (3.3), we can define the similarity between two edges as in Eq. (3.5), which is essentially the cosine similarity between two edges.

Normalized Learning In online social networks, the tag usage behavior differs one user to another. For example the tag usage distribution follows a power law: some tags are shared by a small group of people, which might suggest a higher likelihood that they form a community. On the other hand, popular tags may not be discriminative in inferring group structures. Thus there is a need to differentiate the importance of different users and tags.

Let d_{u_i} denote the degree of the user u_i, and d_{t_p} represent the degree of tag t_p in a user-tag network. After applying normalization, edge $e(u_i, t_p)$ can be represented by $(0, \ldots, 0, \frac{1}{d_{u_i}}, 0, \ldots, 0, \frac{1}{d_{t_p}}, 0, \ldots, 0)$. Given two edges $e(u_i, t_p)$ and $e(u_j, t_q)$, the cosine similarity after normalization between them can be written in Eq. (3.6).

$$S_e(e, e') = \frac{d_{t_p} d_{t_q} \delta(u_i, u_j) + d_{u_i} d_{u_j} \delta(t_p, t_q)}{\sqrt{d_{u_i}^2 + d_{t_p}^2}\sqrt{d_{u_j}^2 + d_{t_q}^2}} \tag{3.6}$$

Setting α to 0.5, $S_u(u_i, u_j)$ and $S_t(t_p, t_q)$ given by Eq. (3.7), we can derive Eq. (3.6) from Eq. (3.3). Thus normalized edge similarity is consistent with the proposed framework.

$$\begin{aligned} S_u(u_i, u_j) &= \frac{2d_{t_p}d_{t_q}\delta(u_i, u_j)}{\sqrt{d_{u_i}^2 + d_{t_p}^2}\sqrt{d_{u_j}^2 + d_{t_q}^2}} \\ S_t(t_p, t_q) &= \frac{2d_{u_i}d_{u_j}\delta(t_p, t_q)}{\sqrt{d_{u_i}^2 + d_{t_p}^2}\sqrt{d_{u_j}^2 + d_{t_q}^2}} \end{aligned} \tag{3.7}$$

It is noticed that the similarity between two users is not only related to users, but also the tags they are associated with. Eq. (3.5) and Eq. (3.6) both assume tags (users) are independent, which is not true in real applications. We next propose a similarity measurement based on correlation.

Correlational Learning Users often use more than one tag to describe the main topic of a bookmark. Grouped tags indicate their correlation. For instance, the tags *car information*, *auto info* and *online cars info*, are used to describe a blog[1] registered on BlogCatalog, are different, but semantically close.

In a user-tag network, a user can be viewed as a vector by treating tags as features. On the other hand, a tag can also be viewed as a vector by treating users as features. Representing users in a latent semantic space captures the correlation between tags, for example, mapping several semantically close tags to a common latent dimension. Let $\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_m$ be the orthogonal basis of a latent semantic sub-space for tags, user vectors in the original space can be mapped to new vectors in the latent space, which is shown in Eq. 3.8.

$$\tilde{u}_i(\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_m) = \mathcal{M}(u_i(t_1, t_2, \ldots, t_n)) \tag{3.8}$$

where $\mathcal{M}$ is a linear mapping from the original space to the latent sub-space. Singular Value Decomposition (SVD) is one of the ways to obtain the set of orthogonal basis.

[1] http://www.blogcatalog.com/blogs/online-cars-info-auto-info-car-news.html

The singular value decomposition of user-tag network M is given by $M = U\Sigma V^\top$, where columns of U and V are the left and right singular vectors and Σ is the diagonal matrix whose elements are singular values. User vectors in the latent space can be formulated in Eq. (3.9).

$$\begin{aligned} & u_i(t_1, t_2, \ldots, t_n) = \{U\Sigma\}_i V^\top \\ \Leftrightarrow & u_i(t_1, t_2, \ldots, t_n) = \tilde{u}_i(\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_m) V^\top \\ \Leftrightarrow & \tilde{u}_i(\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_m) = u_i(t_1, t_2, \ldots, t_n) V \end{aligned} \tag{3.9}$$

where $u_i(t_1, t_2, \ldots, t_n)$ and $\tilde{u}_i(\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_m)$ are the user vectors in the original and latent space, respectively.

However, only a small set of right singular vectors $V' = (v_2, v_3, \ldots, v_m)$ are necessary to be computed. Dhillon [23] suggests that it be $\lceil \log_2 k \rceil + 1$. Recent experimental evaluation in text corpus suggests the dimension between 50 and 1,000 depending on the corpus size and the problem being studied [61]. Another reason of taking a relatively small m is to reduce noise in the data. The user vectors in the latent space can be represented by pluging V' into Eq. (3.9). We set m to 300 for social media data sets. The user similarity and tag similarity are then defined by the corresponding vectors in the latent space.

$$\begin{aligned} S_u(u_i, u_j) &= \frac{\tilde{u_i} \cdot \tilde{u_j}}{\|\tilde{u_i}| \, \|\tilde{u_j}\|} \\ S_t(t_i, t_j) &= \frac{\tilde{t_i} \cdot \tilde{t_j}}{\|\tilde{t_i}| \, \|\tilde{t_j}\|} \end{aligned} \tag{3.10}$$

The above treatment is related to spectral clustering on graphs [74].

$$Lz = \lambda W z \tag{3.11}$$

where z solves the generalized eigenvectors of above equation, L is the laplacian matrix and W is the adjacency matrix, their definitions are shown in Eq. (3.12) in which D_1

and D_2 are diagonal matrix whose non-zero entries are user degrees and tag degrees, respectively.

$$\begin{aligned} L &= \begin{bmatrix} D_1 & -M \\ -M^\top & D_2 \end{bmatrix} \\ W &= \begin{bmatrix} 0 & M \\ M^\top & 0 \end{bmatrix} \end{aligned} \tag{3.12}$$

Let $Z = \begin{bmatrix} U \\ V \end{bmatrix}$ denote the eigenvectors of Eq. (3.11). The generalized eigenvector problem can be rewritten by:

$$\begin{bmatrix} D_1 & -M \\ -M^\top & D_2 \end{bmatrix} \begin{bmatrix} U \\ V \end{bmatrix} = \lambda \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} \begin{bmatrix} U \\ V \end{bmatrix} \tag{3.13}$$

After simple algebraic manipulation, we obtain

$$\begin{aligned} M &= (1-\lambda)V^\top D_1 U \\ M^\top &= (1-\lambda)U^\top D_2 V \end{aligned} \tag{3.14}$$

Thus eigenvectors Z are actually the right and left singular vectors of adjacency matrix M. Thus top singular vectors (except the principle singular vector) of the adjacency matrices contain partition information [23, 74, 124]. Since the user-tag graph studied in this paper is connected, the principle singular vector is discarded.

Data Collection and Statistics

BlogCatalog is a social blog directory where the bloggers can register their blogs under predefined categories. We crawled user names, user ids, their friends, blogs, the associated tags and blog categories. For each blog, users are allowed to specify several tags as a short description. These tags are usually correlated with each other. We crawled more than 10,000 users. Users who have no tags are removed from the data

Table 3.1: Statistics of BlogCatalog and Delicious

	BlogCatalog	Delicious
# of users	8,797	11,285
# of unique tags	7,418	13,592
# of links	69,045	112,850
density	1.1 $\times 10^{-3}$	7.3 $\times 10^{-4}$
maximum tag usage	165	10
minimum tag usage	1	10
average tag usage	7.8	10

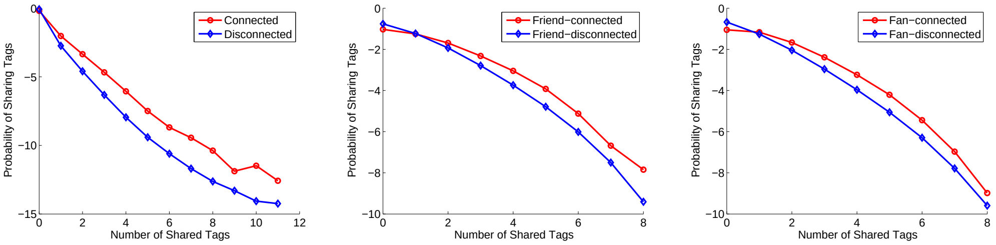
set, and tags that were used by less than two persons were removed as well. Finally, we obtained a data set with 8,797 users and 7,418 tags.

Delicious is a social bookmarking website, which allows users to tag, manage, and share online resources (e.g., articles). For each resource, users are asked to provide several tags to summarize its main topic. We crawled 11,285 users whose information include user name, user id, their friends and fans, their subscribed resources and tags for each resource. The top 10 most frequent tags of each person are kept, which is 13,592 in total. In contrast to BlogCatalog, two kinds of links are formed in Delicious. Fans are the connections from other people (in-links) and friends are the links point to others (out-links). Thus, the connections are directional in Delicious.

The statistics of both data sets are summarized in Table 3.1. The most important difference between the two data sets is that BlogCatalog has category information which can be served as a ground truth for clustering distribution.

Interplay between Link Connection and Tag Sharing

There exist explicit and implicit relations between users. Examples of explicit relations are friends or fans people choose to be. Examples of implicit relations are tag sharing, i.e., people who use the same tags. Are there any correlation between the two different relations? What drives people connect to others? Is it a random operation? We conducted statistical analysis between user-user links and tag sharing.



(a) BlogCatalog Friend (b) Delicious Friend (c) Delicious Fan

Figure 3.2: Tag sharing v.s. connectivity

In the first study, we fix users who have or have no connection with others, then show the tag sharing probabilities. Figure 3.2 shows the tag sharing probabilities in BlogCatalog and Delicious data sets. For Delicious data, the friends network and fans network are evaluated separately. All three graphs show a similar pattern that the tag sharing probability is higher among users who are connected than users who are not. This can be explained by the homophily principle that people tend to connect with those who are like-minded.

Figures 3.3 and 3.4 are the probability that two users being connected if they share tags in BlogCatalog and Delicious, respectively. In Figure 3.3, the probability of a link between two users increases with respect to the number of tags they share. In Delicious, similar pattern is observed. It is also intriguing to show the probability that two users are connected is higher in fans network than that in friends network, which implies users are more *similar* to their fans than their friends.

Clustering Evaluation

The clustering evaluation consists of three studies. First, cross-validation is performed to demonstrate the effectiveness of different clustering algorithms in BlogCatalog data set. Then we study the correlation between user connectivity and co-occurrence in extracted communities. Finally, concrete examples illustrate what clusters are about.

Table 3.2: Cross Validation Performance on BlogCatalog (Micro-F1)

Training Ratio	10%	20%	30%	40%	50%	60%	70%	80%	90%
Correlational Learning	**38.45**	**37.75**	**40.53**	**38.84**	**41.92**	**41.30**	**43.77**	**43.15**	**44.88**
Independent Learning	33.96	36.15	35.07	34.72	35.36	37.32	42.12	41.83	43.09
Normalized Learning	23.89	28.10	29.22	32.14	34.52	35.19	35.79	35.74	37.62
EdgeCluster(user-user)	24.85	25.55	26.27	25.18	25.28	24.80	24.11	23.94	22.22
Dhillon's Co-clustering	23.18	24.18	24.11	24.30	24.34	24.23	24.18	24.15	23.97

Table 3.3: Cross Validation Performance on BlogCatalog (Macro-F1)

Training Ratio	10%	20%	30%	40%	50%	60%	70%	80%	90%
Correlational Learning	**28.85**	**26.83**	**27.68**	**28.52**	**28.18**	**29.69**	28.60	**30.16**	**29.96**
Independent Learning	23.84	25.32	24.34	23.81	25.06	26.28	**29.05**	27.27	26.84
Normalized Learning	14.76	17.61	16.85	18.78	21.66	21.80	22.07	22.39	24.20
EdgeCluster(user-user)	14.24	15.16	16.43	15.75	15.96	16.08	15.42	15.78	14.99
Dhillon's Co-clustering	4.95	5.06	5.11	5.19	5.07	5.18	5.17	5.23	4.66

Comparative Study

In BlogCatalog, categories for each blog are selected by the blog owner from a predefined list. A category is treated as a community or group which suggests the common interest of people within the group. For example, category "Blog Resources" is related to the gadgets used to manage blogs or to communicate with other social media sites. Around 90% of bloggers had joined two categories, and few bloggers had more than 4 categories.

With category information, certain procedures such as cross validation (e.g., treating categories as class labels, cluster memberships as features) can be used to show the clustering quality. Linear SVM [30] is adopted in our experiments since it scales well to large data sets. As recommended by Tang et al. [107], 1,000 communities are used in our experiments. We vary the fraction of training data from 10% to 90% and use the rest as test data. The training data are randomly selected. This experiment is repeated for 10 times and the average Micro-F1 and Macro-F1 measures are reported.

Tables 3.2 and 3.3 show five different clustering methods and their prediction performance. In this table, the fourth algorithm EdgeCluster [107] uses user-user network rather than the user-tag network. Dhillon's co-clustering algorithm is based on Singular Value Decomposition (SVD) of the normalized user-tag matrix. As shown in Tables 3.2 and 3.3, Correlational Learning consistently performs better, especially when the training set is small. And normalization does not improve performance. This suggests normalization should be taken cautiously. Dhillon's co-clustering method which can only deal with non-overlapping clustering does not perform well compared to other methods.

It is also interesting to notice that clustering based on user-tag is significantly better than user-user connection which suggests that meta data (e.g., tags) rather than connection is more accurate in measuring the homophily between users. The clustering difference between meta data and links also reveals promising applications of the framework in link prediction systems. Next, we try to interpret clustering results.

Connectivity Study

We study the correlation between user co-occurrence in extracted communities and the actual social connections between them. We also study the connectivity between users who are in the top similar list. 1,000 overlapping communities are extracted by Correlational Learning.

In Table 3.4, first row represents the number of communities two users co-occur, and each entry in this table is the probability that two users have a connection established in actual social networks. The last column lists the probability if two users are connected randomly. Higher probability than randomness suggests that users within communities are similar to each other. As observed in Table 3.4, frequent co-occurrence of users in different communities implies that they are more likely to be connected. Therefore, it is reasonable to state that higher co-occurrence frequency

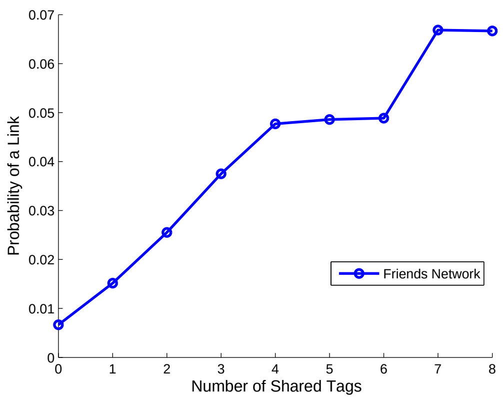


Figure 3.3: Link probability w.r.t tag sharing in BlogCatalog

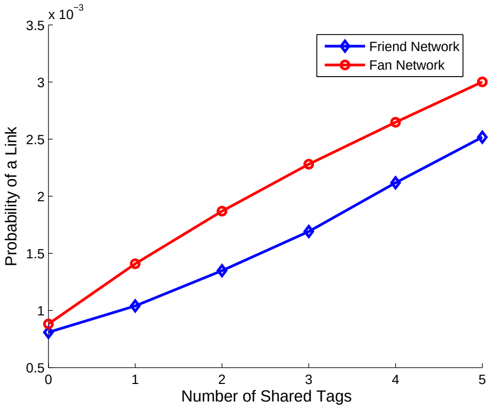


Figure 3.4: Link probability w.r.t tag sharing in Delicious

suggests that two users are more similar. Similar patterns are observed in the other two methods.

We compute pairwise cosine similarity between users (in the latent space) and sort them in descending order, then study the dis-connectivity between users who are most similar. Figure 3.5 shows that the probability of being disconnected is higher than 96% and 99% in BlogCatalog and Delicious, respectively, which means that the majority of homogeneous users are not connected in actual social networks. For example, users *marama*[2] and *ameer157*[3] both are interested in the online game "World of War-

[2] http://www.blogcatalog.com/user/marama

[3] http://www.blogcatalog.com/user/ameer157

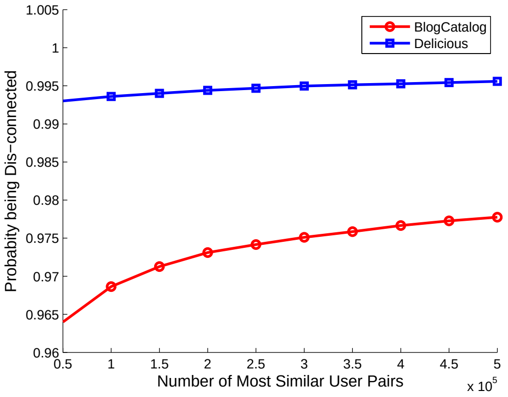


Figure 3.5: Probability being Dis-connected between Top Similar Users

Table 3.4: Co-occurrence vs. Connectivity

# of Co-occurrence	1	2	3	4	5	Random
BlogCatalog($\times 10^{-2}$)	1.64	2.78	4.27	4.43	4.48	0.74
Delicious($\times 10^{-3}$)	2.52	3.83	3.94	3.97	3.45	0.35

craft". Their tags highly overlap, but there is no connection between them. In online social networks, most users are scattered in the long tail, and are usually unreachable by following their and their friends' links. But it is possible to connect them with our Correlational Learning.

Illustrative Examples

Below we use "category" to represent the ground truth and use "cluter" to represent the groups that we obtained via the proposed Correlational Learning. Two clusters *cluster-health* and cluster-nutrition are sub-groups of the Health category. The two clusters are different as suggested by the tag clouds and, meanwhile, they overlap with each other to some extent.

Health is the second largest category (the largest is *personal*) in BlogCatalog, a hot topic that attracts lots of cares. To visualize communities, we create tag clouds using Wordle[4]. In a tag cloud, size of a tag is representative of its frequency or impor-

[4] http://www.wordle.net/

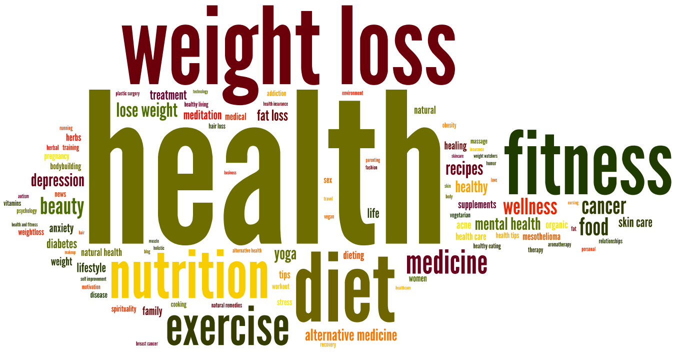


Figure 3.6: Tag cloud for category-health in BlogCatalog

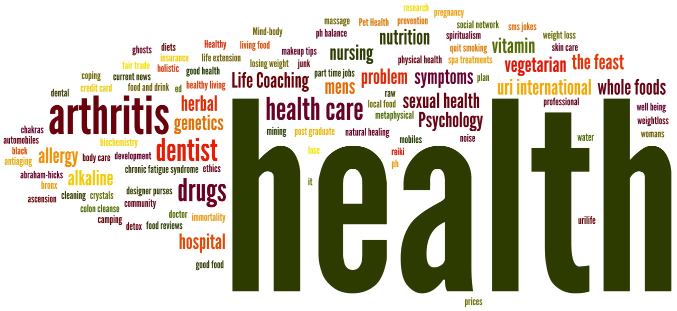


Figure 3.7: Tag cloud for cluster-health in BlogCatalog

tance in a set of tags or phrases. Figure 3.6 shows the tag cloud for Category Health (*category-health*) including all tags of this category. The most frequent 5 tags, *health*, *weight loss*, *diet*, *fitness* and *nutrition*, are all about health.

The largest cluster about Health obtained by Correlational Learning is *cluster-health* with 127 users and 102 tags. The cluster that has the maximum user overlapping with *cluster-health* is *cluster-nutrition* with 83 users and 25 tags. Their tag clouds are shown in Figures 3.7 and 3.8. Between the two clusters, there are 18 users and 3 tags *health*, *nutrition* and *weight loss* in common. Both clusters are related to health but the first has an emphasis on physical health, highlighted by tags *arthritis*, *drugs*, *food*, *dentist*, and the second is more about *nutrition*.

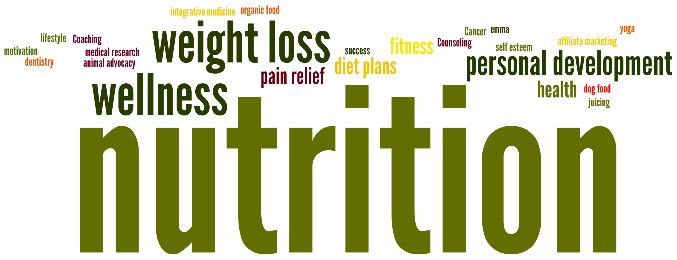


Figure 3.8: Tag cloud for cluster-nutrition in BlogCatalog

We also study the tag overlapping between *category-health* and *cluster-health*, and between *category-health* and *cluster-nutrition*. The top 102 tags of category-health are compared to the tags of *cluster-health* and the top 25 tags of *category-health* to those of *cluster-nutrition*. The numbers of shared tags are 16 for *cluster-health* and 9 for *cluster-nutrition*. The overlapping analysis indicates that tags of the two clusters differ (with only 3 tags in common), the tags of the two clusters are not the same as those of *category-health*, and each cluster represents a new concept (or a sub-topic of health) that is buried in the tags of *category-health*.

3.2 Group Profiling for Understanding

Recently, a surge of work has reported statistical patterns presented in complex networks across many domains [83, 16]. The majority of work studies global patterns presented in a static or an evolving network [57, 67]. Microscopic patterns such as individual interaction patterns are also attracting increasing attention [66]. We, alternatively, focus on meso-level or group-level analysis of a network. A variety of community detection (a.k.a. finding cohesive subgroups [119]) methods have been proposed to capture such social structures in a network [85, 87, 33].

While a large body of work has been devoted to discovering groups based on network topology, few systematically delve into extracted groups to understand the formation of a group. Some fundamental questions remain intriguing:

> How to *understand* a social structure emanated from a network? What is the particular reason that binds group members together?

Some pioneering work attempts to understand group formation based on statistical structural analysis. [7] studied prominent online groups in the digital domain, aiming at answering some basic questions about evolution of groups. One of them is: what are the *structural features* that determine which group an individuals will join. They found that the number of friends in a group is the most important factor to determine whether a new actor would join the group. This result is interesting, though not surprising. It provides a global level of structural analysis to help understand how communities attract new users. [68] observed that spectral clustering (a popular method used for community detection) always finds tight and small-scale but almost trivial communities, i.e., the community is connected to the remaining network via one single edge. Both papers above focus on a global (statistical) picture of communities. Further research is required to understand the formation of a particular group.

In social media, people are likely to interact with each other if they share certain similarity (a.k.a. *homophily* [77]), resulting in assorted communities. Various reasons lead to the formation of a community. For example, some users may interact with each other because they attend the same university; some users form a group as they are enrolled in an event. Users can also coalesce if they share the same political view. In this work, we attempt to understand a group from *a descriptive* aspect, which helps explain the group formation.

- Given individual attributes, can we find out group-level shared commonalities?
- If so, what are the effective approaches?

We aim to extract group attributes that help understand a group. For the aforementioned examples, the group attributes, ideally, should indicate the university, the event, and the political view, respectively.

Extracting descriptive attributes for a group of people is referred as *group profiling* [110]. To construct a group profile, we study strategies to extract attributes for a group when individual attributes are available. This is especially applicable in social media since individuals might share their profiles as well as user activities, such as blog posts, status updates, comments, visited web pages, clicked ads, and so on. This large number of noisy individual traces pose a challenge to extract useful information to describe a group. In this work, three sensible methods are presented for comparative study: *aggregation, differentiation, and egocentric differentiation* based group profiling. Another challenge is that evaluation usually requires extensive human efforts to delve into group member activities to figure out the shared similarity among them. We carefully designed experiments to alleviate human burden for evaluation. Extensive experiments with concrete case studies on two social media domains demonstrate the effectiveness of group profiling based on (egocentric) differentiation. We also enclose a discussion of potential applications based on group profiling, paving the way for in-depth network analysis at large as well as effective group search and retrieval.

Group profiling is to construct a descriptive profile for a provided group. In this section, we motivate this task and formally define the problem.

Motivation

According to the concept of Homophily [77], a connection occurs at a higher rate between similar people than dissimilar people. Homophily is one of the first characteristics studied by early social network researchers [6, 121, 10], and holds for a wide variety of relationships [77]. Homophily is also observed in social media [31, 113, 62]. In this work, we study the "inverse" problem: given a group of users, can we figure out why they are connected? Or what is their shared similarity?

It is impossible to answer these questions if no information other than a social network is available. Luckily, social media often provides more information than just a network. In blogosphere, users post blogs and upload tags. On Facebook, users chat with each other, update their status, leave comments and share interesting stories. These different activities reflect online social life of users, and thus can be used to answer the aforementioned questions.

Social media sites often come with a social network between users. For instance, in Twitter[5], there is a following-follower network. Some community detection methods can be applied to find out the *implicit groups* hidden beneath the interactions. Group profiling, in this case, can be used to understand the extracted communities, facilitating the network analysis and community tracking.

[5] http://twitter.com/

At some other sites like Livejournal[6], Flickr[7], YouTube[8], and Facebook[9], users are allowed to form *explicit groups*. Various explicit groups, besides implicit groups, have cropped up. Some might suspect that the group name and description already provide enough information to peek into one explicit group. Unfortunately, this is not necessarily true. In Livejournal, one of the data sets we studied in the experiments, we encountered a large number of communities whose profile page provides little information on the group. For instance, the community profile of *fruits*[10] does not say much about the exact topic of the community. Group name might provide some hints, but can be misleading in certain cases. Take *fruits* as an example again. A first glimpse at the community name led us to think that this community is composed of people who are fond of fruits. However, after we conduct group profiling[11] on this community, we obtain the following top-ranking tags for this group:

fruits, japan, hello kitty, sanrio lolita, fashion, Japanese street fashion.

Except the first tag that coincides with the group name, all the other tags indicate this group is more about Japanese fashion. Though this group starts with *fruits*, some characters in animes and mangas like *hello kitty*[12] are often discussed as well. It is known that *hello kitty* is a very popular character used in Japanese fashion. Group profiling can help understand implicit communities extracted based on network topology as well as explicit communities formed by user subscriptions. Besides understanding social structures, group profiling also assists network visualization and navigation, tracking the topic shift of a group, group modeling, event alarming, direct marketing and

[6] http://www.Livejournal.com/community/

[7] http://www.flickr.com/groups/

[8] http://www.youtube.com/groups_main

[9] http://www.facebook.com/

[10] http://community.Livejournal.com/fruits/profile

[11] More details in later parts.

[12] http://www.sanrio.com/

connecting the dots. As for direct marketing, it is possible that the online consumers of products naturally form several groups, and each group posts different comments and opinions on the product. If a profile can be constructed for each group, the company can design new products accordingly based on the feedback of various groups. It is noticed that an online network (e.g., blogosphere) can be divided into three regions [57]: singletons who do not interact with others, isolated communities, and a giant connected component. Isolated communities actually occupy a very stable portion of the entire network, and the likelihood of two isolated communities to merge is very low as a network evolves. If group profiles are available, it is possible for one group or a singleton to find other similar groups and make connections of segregated groups of similar interests.

Problem Statement

In order to understand an emerging structure in social media, we aim to build a group profile that illustrates the concerns of a group. This *group profiling* problem can be stated formally as follows:

Given:

- A social network $G = (V, E)$ where V is the vertex (actor) set, and E the edge (connection) set;
- A particular group $g = (V_g, E_g)$ where $V_g \subseteq V$, and $E_g \subseteq V_g \times V_g$, $E_g \subseteq E$.
- Individual attributes $A \in \{0, 1\}^{n \times d}$ where n is the number of nodes in the network G, and d is the total number of attributes;
- The number of group attributes to pick k.

Output:

- A list of top-k descriptive attributes of group g.

Here we assume the attributes of individual users are boolean. For instance, one attribute can denote the gender of actors, or their attitude toward abortion. It can also represent whether a word occurs in an actor's status update, blog post or recently uploaded tags. In some real-world applications, individual attributes might be categorical rather than boolean, e.g., a user's favorite color, location, age, etc. For this kind of attributes, we can convert them into multiple boolean features. For example, if the color attribute contains three values $\{red, yellow, green\}$, we can convert it into three boolean features A_{red}, A_{yellow}, and A_{green}. So $A_{red} = 1$ means the user likes *red*. Thereafter, we just focus on boolean attributes. For convenience, we say a node has attribute A_i if $A_i = 1$ for the node.

It is desirable if a group profiling method satisfies the following properties:

- Descriptive. The selected attributes for a group should reflect the foundation of a group and the shared interest or the associated affiliation.
- Robust. Mountains of data are produced each day in social media. These data tend to be very noisy. The group profiling method should be robust to noise.
- Scalable. In social media, a network of colossal size is the norm. Typically, one network involves hundreds of thousands or millions of actors. E.g., Livejournal has more than 27 million registered users and around 140,000 users updated their journals in last 24 hours[13]. Twitter has 190 million users and tweets 65 million times a day[14]. And Facebook even has more than 500 million active users, and on average, each user creates 90 pieces of content in a month[15]. Meanwhile, networks are highly dynamic. Each day, new users join a network, and

[13] http://www.Livejournal.com/stats.bml

[14] http://techcrunch.com/2010/06/08/twitter-190-million-users/

[15] http://www.facebook.com/press/info.php?statistics

Table 3.5: Statistics on group and attribute

group	$+$	$-$
$A = 1$	tp	fp
$A = 0$	fn	tn

new interactions occur between exiting ones. Users engage in various activities, producing rich user interactions and overwhelming user-generated content. This also presents a challenge for a group profiling method to be scalable and efficient.

Following the guidelines above, we next present several possible strategies for group profiling.

Profiling Strategies

Suppose there are n nodes in a social network G, and d attributes $\{A_1, A_2, \cdots, A_d\}$. For a specified group g, we are interested in the most descriptive features to explain the group formation. We can treat the group as the positive class (denoted as "$+$") and some other nodes not belonging to the group as the negative class (denoted as "$-$"). The instances (nodes) of positive (negative) class are called positive (negative) instances, respectively.

Given a feature A, we have the following statistics as summarized in Table 3.5:

- true positive (tp) is the number of positive instances containing feature A.
- true negative (tn) is the number of negative instances not containing feature A.
- false positive (fp) is the number of negative instances containing feature A.
- false negative (fn) is the number of positive instances not containing feature A.

Given these statistics above, we can compute the conditional probability of an attribute occurring in a group as follows:

- true positive rate (tpr) is the conditional probability of a feature occurring in a group. In particular,

$$tpr = P(A|+) = \frac{tp}{tp + fn} \tag{3.15}$$

- false positive rate (fpr) is the conditional probability that a feature associated with the nodes that are not of the group. Specifically,

$$fpr = P(A|-) = \frac{fp}{fp + tn}. \tag{3.16}$$

We now present the methods for group profiling (GP).

Aggregation-based Group Profiling (AGP)

Since group profiling aims to find features that are shared by the whole group, a natural and straightforward approach is to find attributes that are most likely to occur within the group. This aggregation-based group profiling (AGP) essentially solves the problem below:

$$\max_{\{A_i\}_{i=1}^k} \sum_{i=1}^{k} P(A_i|+) \tag{3.17}$$

We can simply aggregate individual attributes in the group and pick the top-k most-frequent features. Note that this aggregation-based profiling is widely used in current tagging systems in forms of tag clouds. Tag clouds are widely used in social media to show the popularity of a tag by its font size. If the whole network is considered a group, a tag cloud is produced based on aggregation.

However, this method can be sensitive to certain (dumb) features. For instance, words like *world*, *good* and *2009* in blog posts or status updates can be very frequent. They do not contribute to characterizing a group. Even the wisdom of crowds such as user shared tags may not help much following this aggregation strategy. Take one community named *photography*[16] in Livejournal as an example. It is not difficult to

[16] http://community.Livejournal.com/photography/profile

figure out the shared interests among the group members. If we look at those interests that occur most frequently in profiles of users the group, we have the following list:

> *photography, art, music, movies, reading, writing, love, books, painting, poetry*

Except the first two, other tags are actually not good group descriptors. This is because these tags are shared by a large number of people, thus in this group as well. Directly aggregating these tags is biased towards selecting popular tags, rather than those that can characterize this group.

Differentiation-based Group Profiling (DGP)

Instead of aggregating, we can select features which differentiate one group from others in the network. Hence, the group profiling problem amounts to feature selection [70] in a 2-class classification problem with the group being the positive class and the remaining nodes in the network as the negative class. The goal is to find out those top-k *discriminative* features that are representative of a group.

Note that a particular group is fairly small compared with the whole network. For instance, the Livejournal data set that we collected has 16,444 users, and the first two largest groups have around 5,000 and 1,500 members respectively. The majority (90.1%) of the groups are in the long tail, each with less than 100 members. This results in a highly unbalanced class distribution [106]. With this skewed class distribution, Binormal separation (BNS) [32] is an effective method that outperforms other feature selection methods [32, 106] such as information gain and χ^2 square statistic. The BNS score of an attribute is defined as

$$BNS = \left|F^{-1}(tpr) - F^{-1}(fpr)\right|, \tag{3.18}$$

where F^{-1} is the inverse cumulative probability function of a standard normal distribution. A difference of discriminative group profiling and feature selection is that we only care about features that are descriptive of a group (the positive class). Thus we enforce the following constraint for selected attributes:

$$tpr_{A_i} > fpr_{A_i} \tag{3.19}$$

In other words, feature A_i should better explain the positive class rather than the negative class.

Combining the BNS criterion in Eq. (3.18) and the constraint in Eq. (3.19), we have the following formulation for differentiation-based group profiling (DGP):

$$\begin{aligned} &\max_{\{A_i\}_{i=1}^k} \sum_{i=1}^{k} \left| F^{-1}(tpr_{A_i}) - F^{-1}(fpr_{A_i}) \right| \\ &s.t.\ \ tpr_{A_i} \geq fpr_{A_i} \end{aligned} \tag{3.20}$$

Since F^{-1} is a monotonic increasing function, the objective can be reformulated as follows:

$$\max_{\{A_i\}_{i=1}^k} \sum_{i=1}^{k} \left(F^{-1}(tpr_{A_i}) - F^{-1}(fpr_{A_i}) \right) \tag{3.21}$$

Essentially, we select those features that appear frequently in one group but rarely outside the group.

Egocentric Differentiation-based Group Profiling (EDGP)

In the previous differentiation strategy, all the nodes outside a group are deemed as belonging to negative class. However, it might be a luxury to have this global view of all the nodes in a network. Scalability can also be a concern. Most popular online social networks are very huge. For instance, Facebook claims to have more than 500 million active users as of January 10, 2011. Livejournal has more than 25 million registered accounts[17]. It's either time consuming or impractical to retrieve all the information of

[17] http://www.Livejournal.com/stats.bml

a real-world network. In some applications, only an egocentric view is available. In other words, we only know our friends but little knowledge about the people who are strangers to us. Is it possible to describe a group by its members and the members' network structure without knowing the global network topology?

Instead of differentiating a group from the whole network, we propose to differentiate the group from the neighbors of its members, i.e., group profiling based on the egocentric view (EDGP). Group neighbors refer to nodes outside a group that are connected to at least one group member as in Figure 3.9. Egocentric differentiation follows the same objective function as in Eq. (3.21). The key difference is that the egocentric approach treats only the group neighbors, instead of the whole network, as the negative class. Given the huge size difference of the negative classes between DGP and EDGP, one wonders if this egocentric approach suffices in finding discriminative features.

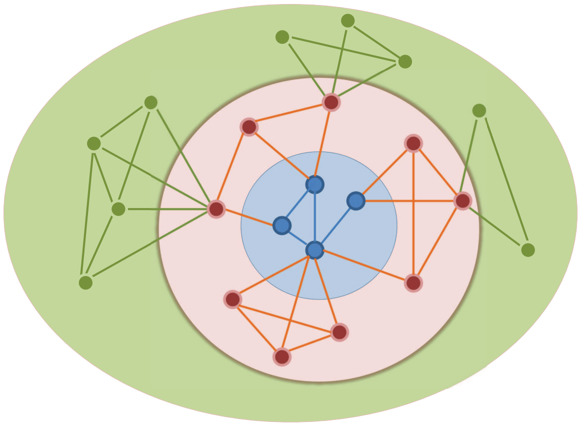


Figure 3.9: Neighbors of a group

Experimental Evaluation

Evaluation Methodology

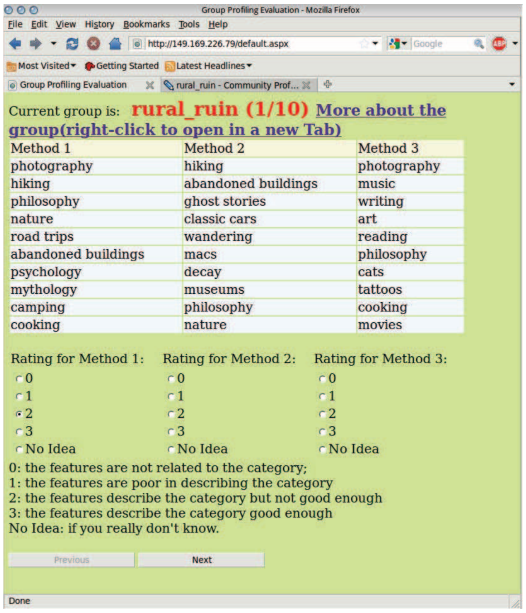
Group profiling outputs a list of features to describe groups. The quality of the extracted profile depends on the group profiling method being used. There are several challenges to perform the comparison. We will address them one by one.

1) How can we obtain group information? For evaluation purpose, we use explicit communities in social media as the group information. In certain social media sites, users can subscribe to one or more interested groups. Explicit communities come with their group names and sometimes descriptions as well. These information can help human subjects to find out the ground truth for evaluation. Of course, this evaluation strategy does not limit the group profiling approach to be applied to implicit groups extracted from a network. As shown later, most explicit online groups also demonstrate a much higher link density than expected.

2) What kind of individual attributes should we look into to extract group profiles? In social media sites, users can share their profiles, upload tags, post blogs and update status. All these activities provide certain information. We treat user interests in profiles or words and tags occurring in their posts as attributes, and find out those key attributes to describe groups.

3) How to evaluate the quality of extracted group profiles? Since there is no ground truth information available, we invite people with different backgrounds to evaluate the result.

   We launched a website with a user-friendly interface for evaluators to log in and rate. A screenshot of the website after a user log in is shown in Figure 3.10. For each group, we use the three proposed approaches (AGP, DGP, EDGP) to select top k ($k = 10$ in our experiments) most representative features. On each evaluation page, the profile features extracted based on each method were listed in a column from top to bottom by importance in descending order (denoted as method 1, method 2 and method 3, respectively in the screenshot). It should be emphasized that evaluators do not know what the group profiling methods are and which column is generated by which method. To avoid the bias associated with the column position, the presentation order of group profiles is also randomized


Group Profiling Evaluation - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://149.169.226.79/default.aspx
Google
Most Visited
Getting Started
Latest Headlines
Group Profiling Evaluation
rural_ruin - Community Prof...
Previous
Next
Done

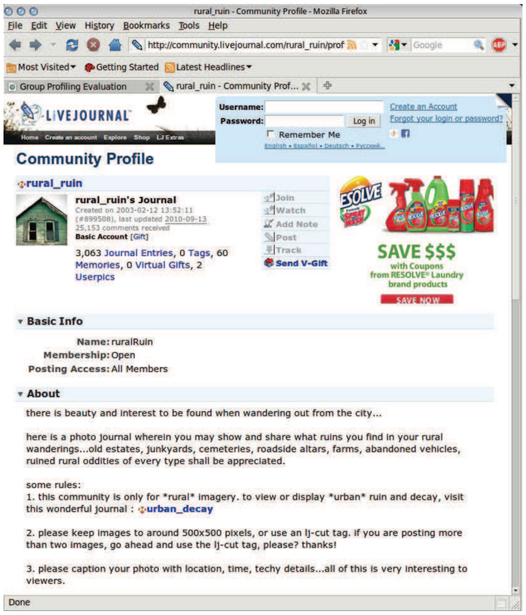
Current group is: **rural_ruin (1/10)** More about the group(right-click to open in a new Tab)

Method 1	Method 2	Method 3
photography	hiking	photography
hiking	abandoned buildings	music
philosophy	ghost stories	writing
nature	classic cars	art
road trips	wandering	reading
abandoned buildings	macs	philosophy
psychology	decay	cats
mythology	museums	tattoos
camping	philosophy	cooking
cooking	nature	movies

Rating for Method 1:	Rating for Method 2:	Rating for Method 3:
○0	○0	○0
○1	○1	○1
●2	○2	○2
○3	○3	○3
○No Idea	○No Idea	○No Idea

0: the features are not related to the category;
1: the features are poor in describing the category
2: the features describe the category but not good enough
3: the features describe the category good enough
No Idea: if you really don't know.

Figure 3.10: Screenshot of the Evaluation System


rural_ruin - Community Profile - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://community.livejournal.com/rural_ruin/prof
Google
Most Visited
Getting Started
Latest Headlines
Group Profiling Evaluation
rural_ruin - Community Prof...
LIVEJOURNAL
Home Create an account Explore Shop LJ Extras
Username:
Password:
Log in
Create an Account
Forgot your login or password?
Remember Me
English • Español • Deutsch • Русский...
Join
Watch
Add Note
Post
Track
Send V-Gift
ESOLVE
SAVE $$$
with Coupons
from RESOLVE® Laundry
brand products
SAVE NOW
Done

Community Profile

rural_ruin

rural_ruin's Journal
Created on 2003-02-12 13:52:11 (#899508), last updated 2010-09-13
25,153 comments received
Basic Account [Gift]

3,063 Journal Entries, 0 Tags, 60 Memories, 0 Virtual Gifts, 2 Userpics

Basic Info

Name: ruralRuin
Membership: Open
Posting Access: All Members

About

there is beauty and interest to be found when wandering out from the city...

here is a photo journal wherein you may show and share what ruins you find in your rural wanderings...old estates, junkyards, cemeteries, roadside altars, farms, abandoned vehicles, ruined rural oddities of every type shall be appreciated.

some rules:
1. this community is only for *rural* imagery. to view or display *urban* ruin and decay, visit this wonderful journal : **urban_decay**

2. please keep images to around 500x500 pixels, or use an lj-cut tag. if you are posting more than two images, go ahead and use the lj-cut tag, please? thanks!

3. please caption your photo with location, time, techy details...all of this is very interesting to viewers.

Figure 3.11: Group Profile Page for Reference

for each page. Suppose for one group the three columns are generated by AGP, DGP, EDGP, respectively. The next time this group or another group is chosen, the three columns might correspond to methods in a totally different order.

We also highlighted the title of the studied group and provided a link to the particular online group profile page, so that evaluators are encouraged to get general group information before making a decision. For instance, by clicking on the link at the top of the screenshot in Figure 3.10, one will be directed to the group page as in Figure 3.11. This profile page contains some description of the group and links to the activities and journal posts insides the group. Hopefully, this can help a subject to make the right decision.

Each evaluator will rate for the resultant profiles on how well they are describing this group. The rating is ranged from 0 to 3, respectively representing “irrelevant”, “partly related”, “reasonable” and “very good”. An evaluator can also decline to give a rating (by choosing a “no idea” option) if he is not sure. As we noticed in

Table 3.6: Statistics on BlogCatalog and Livejournal

	BlogCatalog	Livejournal
# Bloggers	70,086	16,444
# Links	1,706,146	131,846
Link Density	6.9 $\times 10^{-4}$	9.8 $\times 10^{-4}$
Average Links	49	16
Diameter	5	8
Group Title	Category Name	Community Name
Group Numbers	344	100, 441
Average Groups Joined	1.9	32.6

one pilot study, subjects tend to assign random ratings if the task takes too much time. To assure the quality of evaluation, each person was asked to evaluate only 10 group profiles in one session, which can be finished in few minutes.

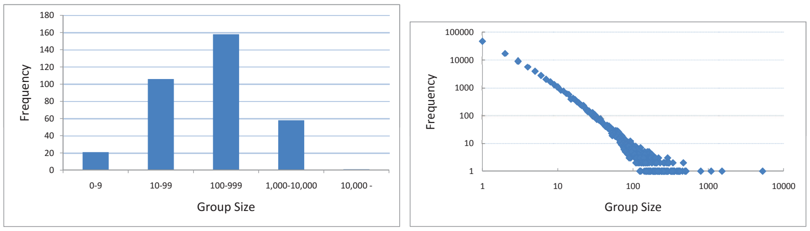
Social Media Data

As mentioned above, we need data sets with groups as well as rich individual attributes. Hence, we select two social media sites for data collection: BlogCatalog[18] and Livejournal[19]. BlogCatalog is a social blog directory where bloggers can register their blogs under specified categories. Livejournal is a virtual community where users can keep a blog, journal or diary. Both websites serve as a platform for users to connect and communicate with others. At both sites, users can engage in social activities like adding friends, joining groups, commenting, tagging and so on.

On BlogCatalog, we crawled blogger's name, friends, the blogs belonging to him/her, tags, categories and most recent 6 snippets. We treat blog categories as groups. After removing the non-English blogs, we obtained 70,086 bloggers and 344 groups. The total friendship links are 1,706,145, and each blogger has 49 friends on average. On Livejournal, we started with a popular blogger *just_ducky*, and crawled bloggers that are reachable in 4 hops away from this seed by following their friendship

[18] http://www.blogcatalog.com/

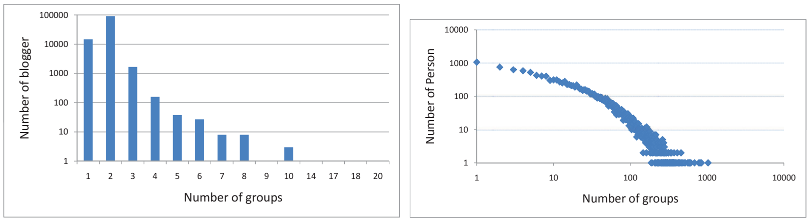
[19] http://www.Livejournal.com/



(a) Blogcatalog

(b) Livejournal

Figure 3.12: Group Size Distribution



(a) Blogcatalog

(b) Livejournal

Figure 3.13: Group subscription distribution

connections. We collected blogger's name, friends, posts, interests specified in his/her profile and the communities the blogger subscribes to. Each user-created community is considered a group. Finally the data set has 16,444 bloggers, more than 130 thousand pairs of friendship links and 100,441 different communities. The statistics of these two data sets are summarized in Table 3.6. One key difference between these two social media websites is that Livejournal bloggers can create communities freely. BlogCatalog users, however, can only specify categories from a predefined list. This explains why there is a much larger number of groups in Livejournal.

These two sites demonstrate different statistical patterns. The group size distributions at both sites are plotted in Figure 3.12, in which, the x-axis represents the group size and the y-axis the frequency. Since the number of groups is very limited in BlogCatalog, we plot the distribution in histogram instead of scatter plot. The group size distribution in BlogCatalog is more like a bell curve, possibly because of the different

Table 3.7: Selected Groups on BlogCatalog

Group	Size	Density	Group	Size	Density
personal	11478	1.3‰	dogs	173	8.0‰
blogging	7727	2.7‰	adult education	139	1.3‰
entertainment	4671	1.9‰	buddhism	96	11.0‰
health	3877	2.4‰	hunting	86	41.0‰
shopping	2687	2.1‰	sailing	71	8.9‰
sports	2529	2.0‰	lawn&garden	55	8.9‰
computers	1934	2.4‰	music industry	47	6.1‰
animals	1357	5.6‰	natural	41	10.0‰
investing	906	3.8‰	city guides	40	32.0‰
science	826	2.4‰	anarchism	29	34.0‰
home cooking	564	3.7‰	auto repair	23	4.3‰
hardware	424	1.2‰	earth science	22	16.0‰
pop	254	2.5‰	aqua. fish	19	17.0‰
stock&bond	245	7.1‰	choreography	13	26.0‰
cultural	229	4.5‰	extinct birds	3	0.0‰

Table 3.8: Selected Groups on Livejournal

Group	Size	Density	Group	Size	Density
photography	320	13.0‰	ontd_startrek	139	12.0‰
sextips	297	1.8‰	behind_the_lens	134	16.0‰
mp3_share	288	2.1‰	tvshare	132	5.2‰
art_nude	232	33.0‰	ru_portrait	131	76.0‰
ourbedrooms	216	12.0‰	knitting	124	2.3‰
houseepisode	211	6.2‰	girl_gamers	121	3.6‰
fruits	205	16.0‰	wow_ladies	115	2.0‰
free_manga	205	9.1‰	art_links	113	50.0‰
ucdavis	189	39.0‰	weddingplans	110	4.7‰
photographie	188	12.0‰	doctorwho_eps	109	25.0‰
cooking	181	2.3‰	ru_travel	108	20.0‰
hot_fashion	161	25.0‰	blythedoll	108	110.0‰
naturalliving	157	3.8‰	rural_ruin	105	14.0‰
topmodel	155	2.8‰	supernatural_tv	103	15.0‰
photocontest	147	1.5‰	animeicons	102	5.0‰
cheaptrip	142	29.0‰	gossipgirltv	101	8.1‰

mechanism for creating groups as we mentioned above. On the contrary, group size in Livejournal follows a power law distribution as observed in many large-scale networks.

On the other hand, the number of groups one blogger joins is shown in Figure 3.13. In BlogCatalog, most bloggers join 2 groups, but a few bloggers (0.23%) join more than 3 groups. In Livejournal, the distribution is different, with 82.3% bloggers joining at least 4 groups. One blogger even has joined 1,032 groups. The average number of groups that one single blogger subscribes to are 1.9 and 32.6 on these two sites, respectively.

In the experiment, we would like to test group profiling methods with different noise level and investigate how each method performs. Typically, words in blog posts are much more noisy than tags or user interests listed in users' profile pages. Hence, we created 4 data sets: BlogCatalog based on tags (BC-Tag) or blog posts (BC-post), and Livejournal based on user interests (LJ-Interest) or journal posts (LJ-post). We expect Livejournal to be more noisy than BlogCatalog as the communities there are user-generated rather than pre-specified.

Since the evaluation involves human efforts, it is impractical to evaluate exhaustively over all groups. We select a subset of representative groups with varying sizes and densities as listed in Tables 3.7 and 3.8. In particular, 30 groups from BlogCatalog and 32 groups from Livejournal. For evaluation purpose, here we use explicit groups, i.e., in which the membership is determined by subscription. But we would like to point out that the density of most groups is much higher than the network density suggesting frequent within-group interactions. Their neighborhood size versus the group size is also plotted in Figure 3.14. Because each node has a plurality of connections, thus the neighborhood size is typically much larger and increasing with respect to the group size.

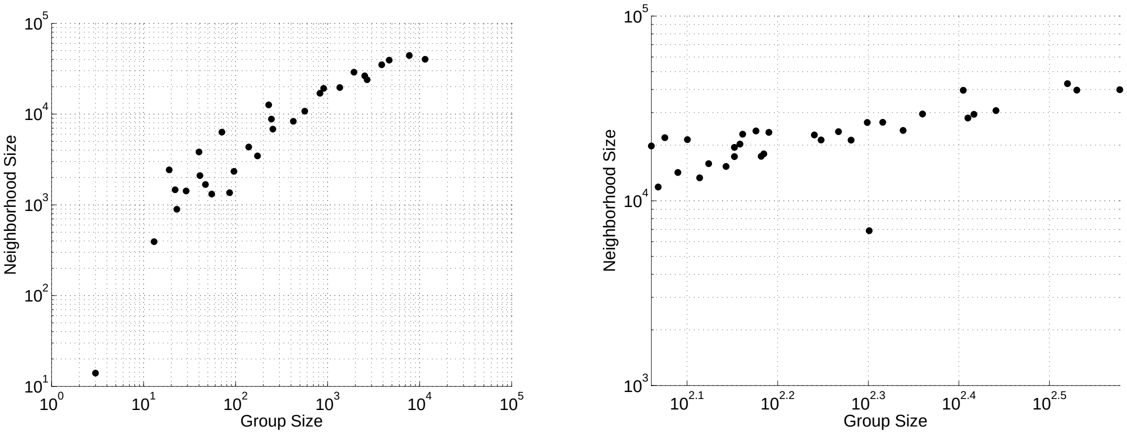


Figure 3.14: Group size v.s. neighborhood size

Empirical Results

52 people with assorted backgrounds (undergraduate, graduate students,university faculty and employees) participated in our evaluation. In total, $2,028$ ratings were collected, of which 101 ratings were “no idea”. So only the remaining $1,927$ ratings were used in our analysis. On average, each group was evaluated 32 times and the average ratings were reported.

Comparative Study

The average ratings for each method on different data sets are shown in Table 3.9. On BC-Tag, three methods are comparable, however the aggregation-based approach deteriorates when we use words in blog posts as features. A similar pattern is observed on Livejournal, though the ratings drop sharply. On both data sets, DGP and EDGP consistently outperform AGP. This is most observable when individual attributes are noisy. That is, a large number of attributes are associated with individuals, among which only few of them are relevant to the group topic (say, when words appearing in blog posts are used as attributes).

Table 3.9: Ratings averaged over all groups

Data set	AGP	DGP	EDGP
BC-Tag	2.55	**2.62**	**2.62**
BC-Post	1.92	**2.35**	2.26
LJ-Interest	1.53	1.91	**2.00**
LJ-Post	0.54	**1.42**	1.35

This result is more visible in Figure 3.15, where we plot the probability of each group profiling method being the winner. It is computed as the frequency of one method winning over the total number of evaluations. One method wins when it receives the highest rating among the three. It is noticed that ties often occur during evaluation. For example, if the ratings for AGP, DGP and EDGP are 2, 3, 3, then we consider both DGP and EDGP win. On BC-Tag, all three methods yield a similar performance. But on the other data sets, DGP and EDGP are consistently better than AGP, and the difference between the former and the latter increases as the noise level increases (Livejournal is more noisy than BlogCatalog as communities are not pre-specified, and posts are more noisy than tags or user-specified interests).

The performance of DGP and EDGP are comparable, with the former slightly better. This demonstrates that little information is lost if we only compare a group with its adjacent neighbors, rather than with all users. With only an egocentric view, the computation cost of profiling a particular group can dramatically drop because of a much smaller number of involved bloggers. In BlogCatalog, the number of 1-hop away bloggers averaged on the selected groups is 8,274, or around 11.8% of the whole network. On Livejournal, for groups whose sizes are larger than 50, the average number of 1-hop away bloggers is 1,016, or around 6.2% of all the bloggers. The egocentric differentiation method is favorable in dynamic and evolving huge networks, because updating features is easy. Only the local information instead of the whole network is required.

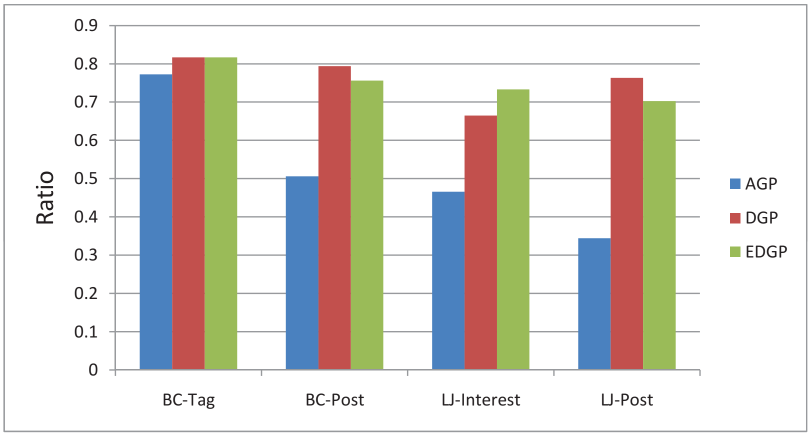


Figure 3.15: Probability of receiving highest rating

Table 3.10: Profiles for *health* group

BC-Tag			BC-Post		
AGP	DGP	EDGP	AGP	DGP	EDGP
health	health	health	people	health	health
fitness	fitness	fitness	health	people	people
diet	diet	diet	body	body	body
weight loss	weight loss	weight loss	life	life	weight
nutrition	nutrition	nutrition	world	weight	life
exercise	exercise	exercise	weight	disease	disease
beauty	cancer	cancer	long	diet	diet
medicine	medicine	medicine	find	food	treatment
cancer	beauty	mental health	back	healthy	food
mental health	mental health	wellness	important	treatment	healthy

Table 3.11: Profiles for *blythedoll* group

LJ-Interest			LJ-Post		
AGP	DGP	EDGP	AGP	DGP	EDGP
blythe	blythe	blythe	love	blythe	blythe
photography	dolls	dolls	back	doll	doll
sewing	sewing	sewing	ll	flickr	dolly
japan	japan	blythe dolls	people	ebay	dolls
dolls	blythe dolls	super dollfie	work	dolls	ebay
cats	super dollfie	japan	things	photos	sewing
art	hello kitty	hello kitty	thing	dolly	flickr
music	knitting	toys	feel	outfit	blythes
reading	toys	knitting	life	sell	outfit
fashion	junko mizuno	re-ment	pretty	vintage	dollies

Case Studies

To have a tangible understanding of the outcome of different methods, here we show two concrete examples: *health* group in BlogCatalog and *blythedoll* group in Livejournal.

Health group has 2,607 members. The topics covered in this group are *medicine, diet, weight loss, men's and woman's health,* and so on. Table 3.10 presents profiles extracted to describe the group based on tags and posts, respectively. The features are sorted by importance in descending order. In BC-Tag, features extracted by all the three methods are related to health. Only the order of some keywords are different. In BC-post, the result of AGP becomes worse. Some features like *world*, *long*, *find*, and *important*, seem irrelevant to health. By looking at the features generated by DGP and EDGP, it is not difficult to figure out that they are about health. These two methods demonstrate subtle difference. Only the order of some features differs.

Table 3.11 shows profiles for *blythedoll* group on Livejournal. Blythedoll was first created in 1972 by U.S. toy company Kenner. Later it spread out to the world. In LJ-Interest, some of the features extracted by AGP method are very frequently used words, e.g., photography, art and music, and we can hardly connect them to blythedoll. In LJ-Post, the AGP result is even worse. There is almost no connection to the blythedoll group. The other two methods, DGP and EDGP, perform consistently better than simple aggregation. This example demonstrates the superiority of DGP and EDGP with noisy data.

Similarity Between Profiles of Different Methods

In previous experiments, we have shown that (egocentric) differentiation-based group profiling tend to outperform the aggregation-based method. In this subsection, we systematically examine the similarity of the profiles produced by the three methods. It is noticed that DGP and EDGP receive similar ratings as reported in Section 3.2. Is this due to the effect that they often select similar features to construct group profile?

As each method outputs a ranked list of attributes, we use Kendall's Tau(τ) rank correlation coefficient [52] to measure the difference of the ordering. Kendall Tau Coefficient measures the agreement between two ranked list. In our experiments, only ten terms are selected for each group, we first construct two ranked lists by assigning a rank for each term. Given two rankings R_1 and R_2 concerning the same set of elements, let x_1 and x_2 denote the rank of element x in R_1 and R_2 respectively. Two elements x and y are a *concordant pair* when the ranks for both elements agree, i.e., if $x_1 < y_1$ and $x_2 < y_2$, or $x_1 > y_1$ and $x_2 > y_2$. x and y form a discordant pair if the relative rank of the two does not agree, i.e., if $x_1 < y_1$ yet $x_2 > y_2$, or $x_1 > y_1$ yet $x_2 < y_2$. The Kendall τ coefficient is defined as

$$\tau = \frac{\text{number of concordant pairs} - \text{number of discordant pairs}}{\frac{1}{2}n(n-1)}.$$

Its value is between -1 (one ranking is the reverse of another) and +1 (two rankings are the same). Two ranks have no correlation if their Kendall Tau Coefficient is 0.

The τ coefficients on all the four data sets are listed in Table 3.12, with entries in bold face to denote the highest similarity in each column. It is observed all methods demonstrate a positive correlation. Among them, DGP and EDGP often output similar rankings. It is noticed that the coefficient on Livejournal data is much smaller than that on BlogCatalog. This might due to the noisy nature as embedded in the Livejournal data.

The ordering effect is ignored, one might be only interested in the set of top-ranking attributes. Thus, we computed the Jaccard similarity [45] between the top-

	BC-Tag	BC-Post	LJ-Interest	LJ-post
AGP / DGP	0.48	0.18	0.10	0.14
AGP / EDGP	0.42	0.08	**0.11**	0.11
DGP / EDGP	**0.60**	**0.31**	0.10	**0.15**

Table 3.12: Mean Kendall's Tau Rank Coefficient

	BC-Tag	BC-Post	LJ-Interest	LJ-post
AGP / DGP	0.80	0.42	0.22	0.04
AGP / EDGP	0.73	0.32	0.07	0.01
DGP / EDGP	**0.85**	**0.71**	**0.31**	**0.14**

Table 3.13: Jaccard Index

ranking attributes output by different methods. Given two sets A and B, Jaccard similarity is defined as

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{3.22}$$

Its range is between 0 and 1. The average Jaccard similarity between the top-10 attributes as selected by different profiling methods are reported in Table 3.13.

Again, DGP and EDGP are quite similar, especially on the BlogCatalog data. This explains why their ratings are similar as reported in Section 3.2. It also suggests that by comparing one group with its neighborhood, rather than the whole network, it is often sufficient to extract a discriminative group profile.

Further Analysis

Understanding Evaluation Results

We noticed that different groups receive quite distinctive ratings even for the same group profiling method. What might be the reason leading to this differences? Is there any connection between group size and ratings? Figure 3.16 plots individual group ratings of EDGP on BC-Post. The groups are sorted, from left to right, by group sizes in a descending order. No evident correlation is found between the group size and the quality of group profiling. Large groups such as "personal" can receive low ratings, and

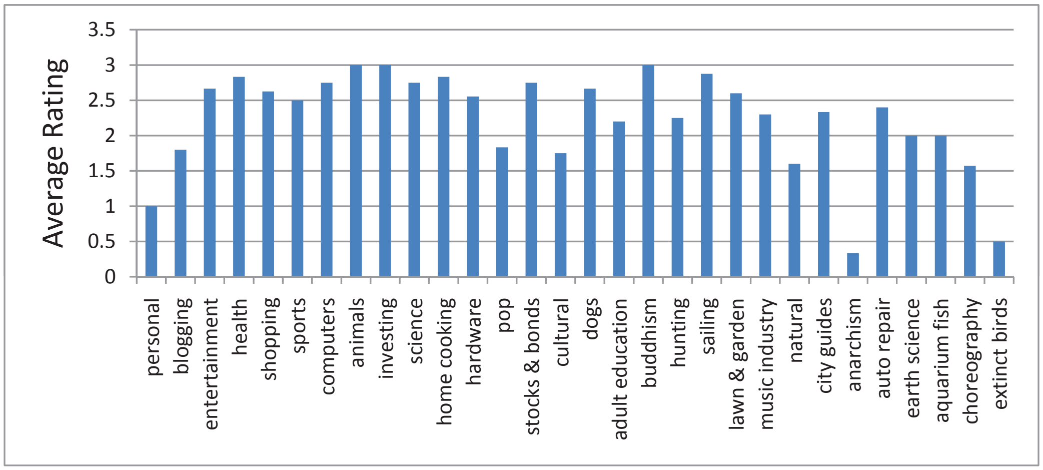


Figure 3.16: Rating of individual groups

small groups like “auto repair” can have high ratings. We observed similar patterns on other data sets with different profiling methods.

One interesting finding is that the more specific a group is, the higher the rating it receives. For instance, the largest group “personal” contains $11,478$ members but has an average rating of 1. Group “auto repair” with only 234 members receives a rating of 2.4. This result agrees with intuition that it is more difficult to describe general concepts, but easier to describe a specific one.

We further analyze the user evaluation behavior. We show the groups of BlogCatalog in Figure 3.17 sorted by their average ratings. The red circles in the curve highlight those groups receiving “no idea” during evaluation, with their sizes indicating the relative probability. It is noticed that the markers tend to reside at the tail of the curve, i.e., when the rating is relatively low. When it is difficult for a human to judge what a particular group is about, it is not surprising the performance of group profiling decreases as well.

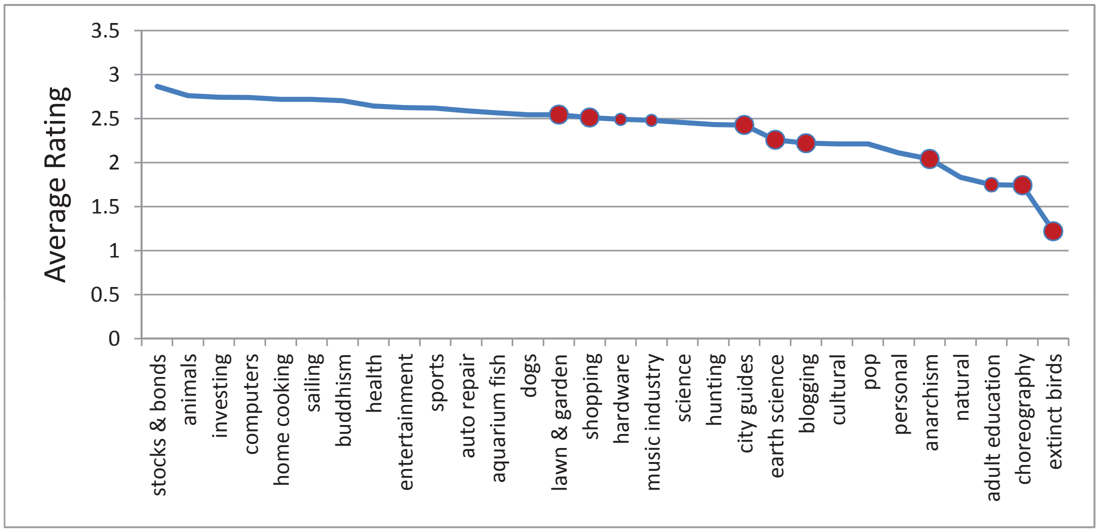


Figure 3.17: Average ratings of groups

Exploiting Group Internal Structures

For all our studied methods, we do not exploit the internal structure inside a group. Presumably, all groups have their influentials [3]. These are opinion leaders, and may play a more important role to reflect the peculiarity of a group. There are many ways to define the importance of a node. Commonly used ones include degree centrality, closeness centrality, betweenness centrality or eigenvector centrality [119]. Here, we take degree centrality as an indicator of a node's importance inside a group. The more connections he has inside a group, the more central role he plays in the group. The number of one node's connections inside a group is used as a weight when we compute the statistics as in Table 3.5.

After applying this simple weighting for profiling, we observe the top ranking features are changed for many groups. Table 3.14 shows the average Jaccard similarity between methods without and with weighting. For DGP and EDGP, the weighting can change the profile a lot. Nevertheless, AGP is not affected as much by the weighting.

Table 3.14: Mean profile similarity

	BC-Tag	BC-Post	LJ-Interest	LJ-post
AGP_{wo}/AGP_w	0.35	0.30	0.94	0.59
DGP_{wo}/DGP_w	0.20	0.05	0.06	0.01
$EDGP_{wo}/EDGP_w$	0.21	0.06	0.05	0.01

Table 3.15: Profiles for *City Guides* group

Without Weighting	With Weighting
olympic games	singapore sights
travel	singapore food
country	singapore recommendations
california	singapore places
tourism	singapore parks
islam	travel products
people	boutique hotels
lifestyle	travel deals
culture	travel style
reviews	luxury resorts

It is noted that the group profiles with a weighting scheme demonstrate some interesting patterns. Those more specific attributes might appear in a profile. For example, Table 3.15 shows the DGP profiles for group *City Guides* with or without weighting. Both types of profiles are sensible. The profiles without weighting seem to be more general whereas some specific terms related to Singapore appear frequently on the right column as the central node is quite interesting in visiting there. It is difficult to conclude which type is better. But it is clear that the group internal structures can play a role in the construction of different informative profiles. We expect that the group internal structure as well as connections to members outside the group can affect the profiling output, and requires further research.

Potential Applications of Group Profiling

Group profiling can help describe groups. The group description can be further used in various types of applications. For instance, group profiles can be used to enrich user profiles. User profiling [101] is one fundamental task in targeting and advertising. However, some users might have very few features. In this case, borrowing features from their group profiles can help improve targeting [100]. Group profiles can also be used to understand the formation of implicit groups, assist community tracking, and group search. Below, we showcase two applications of group profiling: one for understanding implicit groups, and the other for group search and retrieval.

Understanding Implicit Groups

Social media provides tremendous data of network interactions, providing opportunities to study human interactions on an unprecedented scale. These large-scale networks present strong community structures [16]. Group profiling can help understand those implicit groups behind these diverse interactions. Here, we show some interesting findings of group profiling applied to a Flickr network.

Flickr[20] is a photo sharing website where photos are organized in a collaborative way such that both the owner and browsers can upload tags to them. We crawled user names, their contacts, and tags associated with their uploaded photos, ending up with 39,933 users and more than 3.59 million connections after 2 weeks. We applied the EdgeCluster algorithm [108] to find overlapping communities inside the network. EdgeCluster defines a community as a set of edges, rather than a set of nodes like the majority of existing work. By partitioning edges into disjoint sets, it allows the resultant communities to overlap. We obtained 171 clusters with varying sizes. After applying group profiling methods to those clusters, we have several interesting observations.

- People are usually gathered together by their nationality. Flickr is an international social media site, people from different countries might speak different languages. This is intuitive since people tend to tag places, events in their

[20] http://www.flickr.com/

own languages. We found groups extensively focused on Italian, Arabic, Indian, Malaysian, Farsi, Spanish, and so on. A representative profile for an Italian group is shown below (only top 15 keywords are included):

> *bimba, italians, Italians, ritratto, amicizia, ombrello, abbandono, autunno, viaggio, luce, amica, dolcezza, colori, nuvole, gambe*

All keywords except *italians* and *Italians* are all Italian. For instance, *bimba* means infant, ritratto means picture or portrait. The other words starting from *amicizia* can be translated as friendship, umbrella, neglect, autumn, travel, light, friend (female), sweetness, colors, clouds, legs, respectively. The topic is not focused yet at such a large community. But based on group profiling, we know that the communication at a high level is mainly between people speaking the same language. We can also apply group profiling to sub-communities to understand each community in a finer granularity.

- People connect to like-minded peers. Their shared interests are reflected in group profiles. For example, the top keywords for one of these groups is shown below:

  > *TheUnforgettablePictures, TopShots, platinumphoto, SuperShot, GoldStarAward, RubyPhotographer, NaturesElegantShots, ourmasterpiece, SOE, Cubism, GoldDragon, AnAwesomeShot, ABigFave, WorldWideLandscapes*

  These keywords are highly similar in semantics, reflecting users' consensus in their preference. We found most of them are actually titles of some *explicit* interest groups in Flickr. Though people subscribe to different interest groups with different titles, they interact with each other frequently, thus forming an *implicit* group with similar interests. This indicates the usefulness of group profiling in understanding community structures in social media.

Group Search and Retrieval

On social networking sites, users may want to subscribe to different groups. Some groups might match their interests, but with a misleading group name. In this case, it is difficult for a user to locate these groups. On the other hand, advertisers would like to launch campaigns target those groups with desired properties, such as age, gender, education level, interest, etc. Group profiling, by providing an expanded and discriminative description of groups, can be used to build a better group recommendation system. As a proof of concept, we present one example to show how to retrieve and rank related groups to a query based on the result of group profiling. More advanced techniques may be borrowed from the tasks in BlogTrec [76].

A query can have multiple words $q = \{w_1, w_2, \cdots, w_\ell\}$. Given a group profile, i.e., the ranked list of top-k features, we deem a group relevant if at least one word in q appears in the list. We determine each word's ranking score $r(w_i)$ by its position in the group profile. That is, $r(w_i) = m$ if a word w_i appears in the m-th position of the profile. If the word does not appear in the profile, we enforce a penalty by setting $r(w_i) = k+1$. Then, we can compute the proximity of the query and the group:

$$P(q, g) = \sum_{i=1}^{\ell} r(w_i) \tag{3.23}$$

Those groups with lower proximity can be returned as recommended. For instance, in Livejournal data set, the search of "street fashion" results in the following top-ranking groups:

> photo_loli, fott, flammable_live, the cutters, fashion_fucks, books_and_knits, neon_haul, thriftybusiness, alt_boutique, print_project, ru_york, girl_style, egl_glamour, pansy_club, purple_hair, the_chic

Most are reasonable by looking at the group names. Some like *thriftybusiness*[21] seem irrelevant at first glimpse. But once we look at the pictures uploaded by its members, we notice that the majority of the uploaded pictures are indeed about clothes and accessories, confirming the relevance of the group to the query. This example showcases the power of group profiling. The Livejournal website also provides a group search engine. It sorts returned groups by recency of one group being active. The group profiling strategy can find groups based on relevance. In practice, ranking can be accomplished following a hybrid criterion of group activeness and group-query relevance can be explored.

[21] http://community.Livejournal.com/thriftybusiness

Chapter 4

LITERATURE REVIEW

Related work of this dissertation include two parts: connecting the like-minded and understanding social groups in online social networks. Next we give a literature review for each component.

4.1 Connecting the Like-Minded

This section includes two sub-topics that aim to identifying users with similar interests. The first task demonstrates the power of using tag networks in finding the most alike users. The second task shows various feasible approaches to predict information spreaders who are willing to retweet and share novel information with her own followers on Twitter follower networks.

Learning from tag network inference

Closely related work to this problem include collaborative filtering, link prediction and utilization of tags in social network analysis.

Collaborative Filtering (CF) is widely used in many modern recommendation systems. The underlying assumption of collaborative filtering is that people who agreed in the past tend to agree in the future. Therefore, we could leverage past known information to predict future known information [103]. One of the important applications is to recommend items such as products, movies and books that a user could be interested in using different models and knowledge [2, 55, 127]. Recently, social network information is also incorporated into collaborative filtering [50, 54]. One application is to recommend “People You May Know” (or PYMK) in social networking websites such as LinkedIn and Facebook by the number of mutual friends or triadic closure [119].

Link prediction is to infer future interactions between users in a social network with the knowledge at current time stamp. The key idea of link prediction is to recommend potential friends, in terms of proximity, for a seed user [69]. There are several lines of work in predicting future links. The mainstream method is to measure the proximity or similarity between two users, then recommend user pairs with highest proximity scores. The proximity between two users is usually based on structural features [27] such as Common Neighbors [56], Salton Index [96], Jaccard Index [45], Leicht-Holme-Newman Index [64], Hub Promoted Index [91], Adamic-Adar Index [1], etc. The second line of work attemps to modeling the network structure by likelihood maximization [19, 36]. Typically, a probabilistic model is first learnt from the observed network, then is applied to predict missing links. Example models include Probabilistic Relational Model [19], Probabilistic Entity Relationship Model [41], and Stochastic Relational Model [129]. Other approaches for link prediction utilize multitude types of information such as user profile, activity, interaction, user generated content [18, 43, 75, 98], interest [88], or features extracted from above [39], etc.

Tagging on the web is a collective effort that helps to promote information sharing, managing and organizing. The crowd wisdom can be utilized in applications such as tag recommendation [125], social bookmarking, web navigation and browsing [116], query expansion [73], etc. Comparing to domain experts, normal users can tag with reasonably high quality [42], suggesting that collective tagging could be a high quality source of collective human knowledge. However, semantic relevance between tags is rarely addressed in prior work. We focus on measuring the semantic correlation between user generated tags via diffusion kernels that are defined on tag networks. The (diffusion) kernel matrix is required to be positive semi-definite (PSD) and can be viewed as a similarity matrix. There are many successful applications of diffusion kernels in biomedical informatics [63, 97, 105], image retrieval [4], etc. Diffusion kernel is closely related to random walks on graphs [53]. Recent studies show that learning tasks that combining multiple kernels (linearly) often outperforms using a single kernel [60]. Some work are designed to learn the weights among different kernels with the availability of extra information [105].

Identifying information spreaders on Twitter

Twittering becomes a hot research topic recently. We briefly introduce the most relevant work with regard to identifying information spreaders, including the retweet pattern and retweetability analysis, retweet prediction, information diffusion, friend and influential user recommendation, etc.

Retweet is deemed as an effective means to relay information to users who are not necessary direct followers. Kawk et al. studied several interesting topics related to retweet patterns, e.g., the audience size of retweet, retweet tree, temporal aspects of retweet [59]. They found the distributions of the height of retweet trees and the number of participating uses in retweet trees follow a power law: with a small set of retweet trees aggregate a large number of people and spread to longer distances, but most tweet trees only involve a few persons and short distances. They also found that retweeting is time sensitive, i.e., half of retweeting occur within an hour, and 75% within a day. However, they also point out that around 10% of retweets take place a month later.

Many researches analyze factors that might affect the retweetability of a tweet. Boyd et al. interpret the retweeting practice as a way of conversation in which Twitter participants “retweet others and look to be retweeted” [11]. Based on user feedback of reasons why they retweet and on what they retweet most, they find that there are diverse motivations such as “to amplify or spread tweets to new audiences” and “to entertain or inform a specific audience”. More specifically, Suh et al. find that URLs and hashtags have strong correlation with retweetability [104], i.e., a tweet with URLs or hashtags are more likely to be rewteeted than one without.

Retweet prediction, which attempts to predict the occurrences of retweeting, attracts a number of research interests [20, 82, 89, 102, 131, 128]. Naveed et al. view the likelihood of retweetability as a function of interestingness and propose to predict retweeting based on content-based characteristics of tweets [82]. Petrović et al. also attempt to predict whether a tweet is likely to be retweeted by considering a set of social features and tweet features. They claim that the automatic retweet prediction performance is as good as the human prediction. They also found that social features dominate the performance, while the tweet features also add a substantial boost [89]. Zaman et al. propose to predict whether a person will retweet a given tweet from another user by using a collaborative filtering approach [131].

Information diffusion is observed when information flow on the Twitter follower networks. Both retweeting and the spread usage of hashtags are treated as information diffusion on Twitter [21, 65, 93, 115, 126]. Compared to the spread of hashtags, retweeting depends more on the Twitter social network. It is long believed that weak ties are more likely to be sources of novel information, rather than strong ties [35]. Romero et al. examine the hashtags that are spread on Twitter and observe significant variations on the spread of hashtags on different topics. They conclude that the repeated exposure to hashtags have significant marginal effects on their adoption by other users [93]. Tsur and Rappoport show that the combination of content features with temporal and topological features all contribute to predicting the spread of an idea in a given time frame [115].

Other relevant applications on Twitter include recommending friends or followees [13] via link prediction techniques [69] and social collaborative filtering [14]. The information spreader problem is also related to quantifying influence and identifying influential users [8, 15, 38, 122]. Kwak et al. claimed that influential users on Twitter are mostly overlapped with users who have the largest number of followers [59]. Though influential people are important people in a social network, we find that information spreaders are not influentials at all. There are some other relevant work in understanding the factors that affect response such as reply or retweet [20], the usage of Twitter [47, 132], etc.

4.2 Understanding Social Groups

This section is related to work in group discovery and understanding. Social networks show several prominent properties such as high clustering coefficient and small characteristic length (or “small-world networks”) [120] and community structure [33, 34], i.e., groups of nodes are more densely connected internally than with the rest of the network. The majority of work have been contributed to discover implicit groups, rare are focused on understanding these groups.

Group Discovery

Many early work in community detection attempt to discover disjoined communities by maximizing various measurements and objectives [33]. Representative approaches include graph partition [74], modularity maximization [85, 123], random walk [90, 133], etc.

Later on, overlapping communities detection, which allows one user to be associated in one or more communities, attracts more attention. Fuzzy clustering or soft clustering is one of the ways for overlapping community detection, in which each node will be assigned a membership score to a community [86, 130]. Soft clustering returns a dense representation of matrix, which requires an extensive memory footprint to hold the data. Another way of overlapping community detection, which is more popular, is towards discrete assignment. CFinder [87] first enumerates all k-cliques and combines them if there is a high overlapping (e.g., they share k-1 nodes) between two cliques. Cliques are fully connected sub-graphs and a node may belong to several cliques. This method can discover overlapping communities, but it is computationally expensive. EdgeCluster [107] views the graph in an edge-centric angle, i.e., edges are treated as instances and nodes are treated as features. It also shows that a user is usually involved in multiple affiliations, but an edge is usually only related to a specific group. Thus, they propose to cluster edges instead of nodes. This discrete assignment of nodes in a graph gives a clear definition on the community of nodes. Evans et al. [29] proposes to partition links of a line graph to uncover the overlapping community structure. A line graph can be constructed from the original graph, i.e., each vertex in the line graph corresponds to an edge in the original graph and the links in the line graph represents the adjacency between two edges in the original graph, for instance, two vertices in line graph are connected if the corresponding edges in the original graph share a vertex. But it is difficult to scale up to large data sets because of the memory requirement.

Recently, hierarchical clustering approaches are utilized for community detection at multiple resolutions [87, 91, 95, 99, 118]. This line of work first attempts to find communities at the finest resolution (i.e., base communities), then combine communities that are most similar in an aggregated approach until certain constraints are met (e.g., the number of communities). Thus, a hierarchical structure of communities form. Wang et al. found that communities that are discovered at different resolutions all contribute to predicting users' online behavior [118].

Co-clustering involves two sets of relational objects, which are often represented as a bipartite graph, and assigns both sets of objects into different groups with certain constraints. Dhillon et al. [23] propose to co-cluster documents and terms. At first, a bipartite graph between documents and terms is constructed, but partitioning documents and words in this graph is NP-hard, thus it is relaxed to a spectral co-clustering problem. Then top singular vectors (except the principle singular vector) of the document-word bipartite graph are clustered by k-means algorithm. The work above does not take the document-document correlation into account. Java et al. [46] advance this method by adding link structures between entities. For example, links between academic papers in terms of citation are added to the paper-word bipartite graph. The basic idea of Zha et al. [124] is close to Dhillon's work. The bipartite graph partition problem is solved by computing a partial singular vector decomposition (SVD) of the weight matrix. Furthermore, Zha et al. also show that the normalized cut problem is connected to correspondence analysis in multivariate analysis. Similar to [23], this problem is also relaxed to spectral clustering, then k-means is run on the eigenvectors to discover clusters. Compared to [23], this method requires more memory and are computationally more expensive. Information-theoretic co-clustering [24] maximizes mutual information between document clusters and term clusters.

Group profiling

Group profiling describes the shared characteristics of a group of people. It can be applied for policy-making, direct marketing, trend analysis, group search and tracking. Tang et al. [110] present the group profiling problem in terms of topics shared by the group. They propose to classify online documents associated with groups, and then aggregate the class labels to represent the shared group interests. To capture latent semantic relationship between different groups, topics are organized in a hierarchical manner, represented as a taxonomy. As the semantics of different topics can vary in an evolving online environment, they propose to adapt the taxonomy accordingly when new content arrive. Note that the work [110] concentrates on topic taxonomy adaptation. Group profiles are constructed by aggregation.

Group profiling is also applied by sociologists to understand politics and culture in the Persian blogosphere [51]. In the study, bloggers are first clustered based on their link structure. Then, human beings are hired to assign topics and write a short summary for each blog site. Based on the description, the authors analyze profiles associated with each group. They also count frequencies of Iranian related terms occurring in each group and report patterns associated with each group, including which terms occur frequently in one particular group, what are the common terms shared by two different groups. All above analysis require a lot of human effort. That is where our automatic group profiling techniques can help to extend the analysis to a much larger scale.

Selecting the set of representative keywords could be modeled as a feature selection problem, which chooses a subset of features to represent the original high dimensional data, in order to improve prediction performance or reduce time and space complexity [37]. It has been widely used in various domains. Different metrics are used to measure the importance of features. Take text as an example, term frequency, document frequency, tf-idf weight [49], χ^2 statistics, information gain, and mutual information are commonly used to select terms from text. Term frequency selects most frequent terms. Similarly, document Frequency (DF) measures the number of documents a term appears. Tf-idf weighting is a combination of term frequency and document frequency to balance between term specialty and popularity, widely used in information retrieval and text mining applications. χ^2 statistics (CHI) measures the divergence between a term and a category from the χ^2 distribution if one assumes the independence of the term and category. This measure is not reliable for extremely infrequent terms [26]. Information Gain (IG) chooses feature with maximal information increment for classification.

Another relevant line of research is to extract annotations from relational data. For instance, Roy et al. [94] construct a hierarchical annotation structure with a generative model. The model complexity and scalability hinder its application to large-scale networks. Chan et al. [17] propose NUBBI (Networks Uncovered By Bayesian Inference) to infer descriptions of entities in a text corpora. In addition, they also annotate relationships between these entities. Another close branch of relevant work is text summarization which is the creation of shortened version of a text, within the natural language processing community. It has two different forms: single document summarization and multi-document summarization [22].

Some other work extend topic models to extract groups based on network and text information together. Conventionally, a collection of documents are modeled as a set of latent topics, and each topic represents a distribution of words. Link-LDA [28] treats citations of papers the same way as normal words, i.e., the citation is generated based on a multinomial distribution over documents. Pairwise Link-LDA [81] essentially combines the topic model [9] and the mixed membership stochastic block model [5] by sharing the same latent mixture of communities for both word topics and relation topics. Link-PLSA-LDA [81] extends the model link-LDA one step further by modeling the citation as a mixture of latent topics instead of a multinomial distribution. Mei et al. [78] treats connections between documents in a different fashion. It enforces the connected documents to share similar topics and use the network information as regularization to extract topics. Topic-Link LDA [72] models the probability of connections between two nodes as depending on their similarities in terms of both latent topics and latent community memberships.

These work differ from group profiling as they aim to extract latent topics of a collection of documents, while group profiling aims to extract representative attributes that are descriptive of a given group. After extracting topics, it remains unanswered which topic or which words from the topics should be chosen to represent the given group. However, we agree that the two approaches are relevant to some extent. For instance, the group profiling techniques discussed here can be applied to select topics for each group as well.

Chapter 5

CONCLUSIONS AND FUTURE WORK

Social networking services have eased personal communication since its origin in the Web 2.0 era. In these online social networks, connecting users with similar interests adds extra value to both the social networking platforms and the interacting individuals. Next we conclude the dissertation and point out several promising future lines of work.

5.1 Conclusions

In the following sections, we set forward to conclude each of the two components: connecting the like-minded and understanding social groups.

Connecting the like-minded

Social media users not only consume but also produce content simultaneously. The user generated content are indicators of users' intent or interests in the virtual world.

In our first attempt, we propose the new concept (i.e., Tag Networks) to represent the collective tagging knowledge that is produced spontaneously by social media users. A tag network is a graph in which each node represents a tag, a weighted edge between two tags represents the number of users who used the two tags for describing an object (e.g., article, photo, blog). The hypothesis is that co-occurrence counts between two tags describe the semantic correlation between them. We measure the semantic similarity by defining a diffusion kernel on tag networks. With tag networks, we are able to measure the similarity between an arbitrary pair of users. Compared to other popular approaches (e.g., Triadic Closure) to connecting users that are alike, our approach achieved a 108% improvement. We also demonstrated that tag networks are more capable of capturing the semantic relation between tags than Latent Semantic Indexing (LSI), with an improvement of 27% on the studied social media data set BlogCatalog.

The second attempt is to identify information spreaders on Twitter follower social networks. An information spreader is a follower who is (more) likely to retweet a tweet and share it with his or her own followers. Information cascade on the Twitter follower networks by the aggregated efforts of information spreaders. We propose the new problem of identifying information spreaders, which is remain unaddressed. Our work helps to bridge the gap between analyzing the retweetability and understanding information diffusion. By analyzing the user generated content (i.e., tweets), we proposed a number of feasible approaches based on proximity, content, interaction and profile features. We found simple methods outperform complex methods for the information spreader identification problem, i.e., hashtags and URLs are strong features for identifying information spreaders. Combining multiple features is necessary in scenarios where users have only small number of followers. Furthermore, we also found that information spreaders have very small overlapping with the influential people in a social network, suggesting that information spreaders are unlikely to be influential people.

Understanding Social Groups

Link creation in online social media platforms is a fundamental activity. Groups with focused interests are likely to form in these social networks. Identifying and interpreting (overlapping) groups in social networks becomes an urgent and important research topic recently.

Social groups that are identified by various community detection algorithms are usually difficult to interpret as they are extracted from link information. We proposed a co-clustering framework to identify and understand groups simultaneously. We first construct an undirected bipartite graph in which users are connected to tags, and tags to users. Compared to other state-of-the-art community detection algorithms, the identified groups by our approach are easier to be understood by looking at who are interested in what. Empirical results show that the co-clustering framework is able to produce groups with more like-minded users than other approaches based on link information.

To generalize, we propose group profiling as a systematic approach for group understanding when groups are present. This is an emergent field that requires more research in years to come. We explore different strategies to construct descriptive features of a group, e.g., aggregation, differentiation and egocentric differentiation. Empirical evaluations show that the differentiation strategy which is based on Bi-normal Separation [32] produces the most satisfied results, and its egocentric version helps to save significant computational power, maintaining comparable profiling quality.

5.2 Future Work

Online relationships have become an integral part of our social lives. The importance of online relationships is increasingly strengthened as more and more people accept and get involved in online interactions. Recommending users with similar interests is one of the most important components in popular social networking websites. Though we have addressed some problems in the context of social media, there are many meaningful work to be done as the web continue to evolve.

In social networking websites, multiple types of online interactions (e.g., connecting, posting, liking, etc) co-exist. Integrating the multiple heterogeneous data sources and knowledge is a challenging and meaningful work that is worth further explorations, especially in the area of theoretical modeling and analysis. The problem becomes even more challenging when negative relationship is introduced.

The second tangible work is to detect topical like-minded users. Users usually have multiple types of interests (i.e., multi-faceted) and they could be interested more in some and less in others (i.e., with preferences). Multi-faceted interests reflect user preferences more accurately, providing further potentials for finding users with closer preferences.

Improve scalability is imperative as social networks grow significantly larger in recent years. There are much space that we can improve our approaches to scale up to social networks with millions or even hundreds of millions users. New techniques such as sub-optimal approximation and cloud computing are the working directions to adapt the learning approaches to cope with big data in years to come.

It is also intriguing to study temporal variations of user interests in online social networks, and to study the evolutionary group behavior of users with similar interests. As user groups may change over time dramatically, it is interesting to analyze the interplay between the group constitution and common interests. Many interesting questions yet to be answered such as the group evolution drives the change of the common interests or in the other way around.

A user leaves traces on each social networking website that she likes to visit. By default, there is no connection among the set of activities, which might be significantly different (or similar). Each part of user activities form the online identity of a user in social networking environment. Therefore, whatever analysis done on one social networking site is incomplete. It would be interesting to see how different the user behaviors across multiple social networking sites and whether the knowledge on one site helps to infer a user's behavior on another. One direct question is that how the collective wisdom (i.e., tag networks) can be generalized to other websites, i.e., whether the tag networks from different websites are equivalent and to what extent that they are similar.

Besides, I am working on several other pieces of work that are not closely connected to the thesis, but are very relevant to my current work. One is to predicting the

Twitter trends, by analyzing various factors and validating different models. Preliminary results show that behavioral factors, which are rarely addressed, such as activeness are critical in trend prediction. The second work is to learn negative relationships from the link structure, leveraging the PU (partially supervised learning) framework. The third work is to predicting query intent by integrating heterogeneous types of user generated data such as history clicks, user profile and friendship, etc. Next ongoing work is to designing new models for improved search experience in social networking environments.

REFERENCES

[1] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[2] D. Agarwal and B.-C. Chen. Regression-based latent factor models. In *the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'09)*, 2009.

[3] N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In *Proceedings of the international conference on Web search and web data mining*, pages 207–218, New York, NY, USA, 2008. ACM.

[4] R. Agrawal, W. Grosky, F. Fotouhi, and C. Wu. Application of diffusion kernel in multimodal image retrieval. In *The Third IEEE International Workshop on Multimedia Information Processing and Retrieval (IEEE-MIPR 2007)*, 2007.

[5] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

[6] J. C. Almack. The influence of intelligence on the selection of associates. *School and Society*, 16:529 – 530, 1922.

[7] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54, 2006.

[8] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web Search and Data Mining*, pages 65–74, 2011.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[10] H. Bott. Observation of play activities in a nursery school. *Genetic Psychology Monographs*, 4:44–88, 1928.

[11] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the 43rd Hawaii International Conference on Social Systems*, 2010.

[12] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[13] M. J. Brzozowski and D. M. Romero. Who should i follow? recommending people in directed social networks. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[14] X. Cai, M. Bain, A. Krzywicki, W. Wobcke, Y. S. Kim, P. Compton, and A. Mahidadia. Collaborative filtering for people to people recommendation in social networks. In *Proceedings of the 23rd Australasian Conference on Artificial Intelligence*, pages 476 – 485, 2010.

[15] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[16] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1):2, 2006.

[17] J. Chang, J. Boyd-Graber, and D. M. Blei. Connections between the lines: augmenting social networks with text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 169–178, New York, NY, USA, 2009. ACM.

[18] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy. “make new friends, but keep the old” — recommending people on social networking sites. In *Proceedings of the 27th international conference on Human factors in computing systems (CHI’09)*, 2009.

[19] A. Clauset, C. Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98 – 101, 2008.

[20] G. Comarela, M. Crovella, V. Almeida, and F. Benevenuto. Understanding factors that affect response rates in twitter. In *Proceedings of the 23rd ACM conference on Hypertext and hypermedia*, 2012.

[21] E. Cunha, G. Magno, G. Comarela, V. Almeida, M. A. Goncalves, and F. Benevenuto. Analyzing the dynamic evolution of hashtags on twitter: a language-

based approach. In *Proceedings of the Workshop on Languages in Social Media*, page 58Ű65, 2011.

[22] D. Das and A. F. Martins. A survey on automatic text summarization, November 2007.

[23] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274, New York, NY, USA, 2001. ACM Press.

[24] I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 89–98. ACM New York, NY, USA, 2003.

[25] S. T. Dumais. Enhancing performance in latent semantic indexing (lsi) retrieval. Unpublished manuscript, September 1992.

[26] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *COMPUTATIONAL LINGUISTICS*, 19(1):61–74, 1993.

[27] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*, chapter Strong and Weak Ties. Cambridge University Press, 2010.

[28] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101(90001):5220–5227, 2004.

[29] T. S. Evans and R. Lambiotte. Line graphs, link partitions and overlapping communities. *Physical Review E*, 80:016105, 2009.

[30] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[31] A. T. Fiore and J. S. Donath. Homophily in online dating: when do you like someone like yourself? In *extended abstracts on Human factors in computing systems*, pages 1371–1374, 2005.

[32] G. Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305, 2003.

[33] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3 - 5):75 – 174, 2010.

[34] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821 – 7826, 2002.

[35] M. Granovetter. The strength of weak ties. *American Journal of Socialogy*, 78(6):1360–1380, May 1973.

[36] R. Guimer$\acute{a}$ and M. Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073 – 22078, 2009.

[37] I. Guyon. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[38] J. Hannon, M. Bennett, and B. Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the fourth ACM conference on Recommender systems*, 2010.

[39] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *Proceedings of the Workshop on Link Discovery: Issues, Approaches and Applications*, 2005.

[40] M. Hechter. *Principles of Group Solidarity*. University of California Press, 1988.

[41] D. Heckerman and C. Meek. Probabilistic entity-relationship models, prms, and plate models. In *In Proceedings of the 21st International Conference on Machine Learning*, 2004.

[42] P. Heymann, A. Paepcke, and H. Garcia-Molina. Tagging human knowledge. In *Third ACM International Conference on Web Search and Data Mining (WSDM'10)*, 2010.

[43] D. Horowitz and S. D. Kamvar. The anatomy of a large-scale social search engine. In *The 19th International World Wide Web Conference (WWW'10)*, 2010.

[44] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley-Interscience Publication, 2000.

[45] P. Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547 – 579, 1901.

[46] A. Java, A. Joshi, and T. Finin. Detecting commmunities via simultaneous clustering of graphs and folksonomies. In *WebKDD 2008 Workshop on Web Mining and Web Usage Analysis*, August 2008.

[47] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In *Procedings of the Joint Ninth WebKDD and First SNA-KDD Workshop*, 2007.

[48] M. Joel. *Six Pixels of Separation*. Business Plus, 2009.

[49] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):1121, 1972.

[50] H. Kautz, B. Selman, and M. Shah. Referralweb: Combining social networks and collaborative filtering. *Communications of the ACM*, 40:63 – 65, 1997.

[51] J. Kelly and B. Etling. *Mapping Iran's Online Public: Politics and Culture in the Persian Blogosphere*. Cambridge: Berkman Center for Internet and Society at Harvard University, 2008.

[52] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–89, 1938.

[53] R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *the 19th International Conference on Machine Learning (ICML 2002)*, 2002.

[54] I. Konstas, V. Stathopoulos, and J. M. Jose. On social networks and collaborative recommendation. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 195 – 202, 2009.

[55] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *the 14th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'08)*, 2008.

[56] G. Kossinets. Effects of missing data in social networks. *Social Neworks*, 28(3):247 – 268, 2006.

[57] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–617, New York, NY, USA, 2006. ACM.

[58] S. Kumar, G. Barbier, M. A. Abbasi, and H. Liu. Tweettracker: An analysis tool for humanitarian and disaster relief. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[59] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International World Wide Web Conference*, pages 591–600, 2010.

[60] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27 – 72, 2004.

[61] T. K. Landauer and S. T. Dumais. Latent semantic analysis. *Scholarpedia*, 3(11):4356, 2008.

[62] H. Lauw, J. C. Shafer, R. Agrawal, and A. Ntoulas. Homophily in the digital world: A livejournal case study. *IEEE Internet Computing*, 14(2):15–23, 2010.

[63] H. Lee, Z. Tu, M. Deng, F. Sun, and T. Chen. Diffusion kernel-based logistic regression models for protein function prediction. *OMICS: A Journal of Integrative Biology*, 10(1):40 — 50, 2006.

[64] E. A. Leicht, P. Holme, and M. E. J. Newman. Vertex similarity in networks. *Physical Review E*, 73(2):026120, 2006.

[65] K. Lerman and R. Ghosh. Information contagion: an empirical study of the spread of news on digg and twitter social networks. In *Proceedings of Fourth International Conference on Weblogs and Social Media*, 2010.

[66] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, New York, NY, USA, 2008. ACM.

[67] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1):2, 2007.

[68] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In *Proceedings of the 17th international conference on World Wide Web*, pages 695–704, 2008.

[69] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of 12th International Conference on Information and Knowledge Management*, 2003.

[70] H. Liu and H. Motoda. *Feature selection for knowledge discovery and data mining*, volume 454. The Springer International Series in Engineering and Computer Science, 1998.

[71] K. Liu and L. Tang. Large-scale behavioral targeting with a social twist. In *Proceedings of the 20th International Conference on Information and Knowledge Management*, pages 1815 – 1824, 2011.

[72] Y. Liu, A. Niculescu-Mizil, and W. Gryc. Topic-link lda: Joint models of topic and author community. In *in Proceedings of the 26th International Conference on Machine Learning,*, 2009.

[73] Z. Lu, W. Kim, and W. J. Wilbur. Evaluation of query expansion using mesh in pubmed. *Information Retrieval Boston*, 12(1):69 – 80, 2009.

[74] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395 – 416, 2007.

[75] H. Ma, H. Yang, M. R. Lyu, and I. King. Sorec: Social recommendation using probabilistic matrix factorization. In *The 17th ACM International Conference on Information and Knowledge Management (CIKM'08)*, 2008.

[76] C. Macdonald, R. L. Santos, and I. O. I. Soboroff. Blog track research at trec. *ACM SIGIR Forum*, 44(1):58–75, 2010.

[77] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[78] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In *Proceedings of the 17th international conference on World Wide Web*, pages 101–110, 2008.

[79] E. Mustafaraj, S. Finn, C. Whitlock, and P. T. Metaxas. Vocal minority versus silent majority: Discovering the opinions of the long tail. In *the IEEE Third International Confernece on Social Computing*, 2011.

[80] M. Nagarajan, H. Purohit, and A. Sheth. A qualitative examination of topical tweet and retweet practices. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[81] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 542–550, 2008.

[82] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd International Conference on Web Science*, 2011.

[83] M. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[84] M. Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–352, 2005.

[85] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, Feb 2004.

[86] M. E. J. Newman and Leicht. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 104:9564–9, June 2007.

[87] G. Palla, I. Der$\acute{e}$yi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, June 2005.

[88] M. Pennacchiotti and S. Gurumurthy. Investigating topic models for social media user recommendation. In *20th International World Wide Web Conference*, 2011.

[89] S. Petrovi$\acute{c}$, M. Osborne, and V. Lavrenko. Rt to win! predicting message propagation in twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 586–589, 2011.

[90] P. Pons and M. Latapy. Computing communities in large networks using random walks. *J. of Graph Alg. and App. bf*, 10:284–293, 2004.

[91] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabàsi. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551 – 1555, 2002.

[92] M. Robnik-Sikonja. Improving random forests. In *the 15th European Conference on Machine Learning*, pages 359–370, 2004.

[93] D. M. Romero and B. M. J. Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704, 2011.

[94] D. M. Roy, C. Kemp, V. K. Mansinghka, and J. B. Tenenbaum. Learning annotated hierarchies from relational data. In *In Advances in Neural Information Processing Systems*, 2006.

[95] M. Sales-Pardo, R. Guimer$\acute{a}$, A. A. Moreira, and L. A. N. Amaral. Extracting the hierarchical organization of complex systems. *Proceedings of the National Academy of Sciences*, 104(39):15224–15229, September 2007.

[96] G. Salton. *Introduction to Modern Information Retrieval*. MuGraw-Hill Auckland, 1983.

[97] B. Sch$\ddot{o}$lkopf, K. Tsuda, and J.-P. Vert, editors. *Kernel Methods in Computational Biology*, chapter Diffusion Kernels, pages 171 – 192. The MIT Press, 2004.

[98] R. Schifanella, A. Barrat, C. Cattuto, B. Markines, and F. Menczer. Folks in folksonomies: Social link prediction from shared metadata. In *Third ACM International Conference on Web Search and Data Mining*, 2010.

[99] H. Shen, X. Cheng, K. Cai, and M.-B. Hu. Detect overlapping and hierarchical community structure in networks. *Physica A: Statistical Mechanics and its Applications*, 388(8):1706–1712, 2009.

[100] X. Shi, K. Chang, V. K. Narayanan, V. Josifovski, and A. J. Smola. A compression framework for generating user profiles. In *In ACM SIGIR workshop on feature generation and selection for information retrieval*, 2010.

[101] M. Shmueli-Scheuer, H. Roitman, D. Carmel, Y. Mass, and D. Konopnicki. Extracting user profiles from large scale data. In *Proceedings of the 2010 Workshop on Massive Data Analytics on the Cloud*, 2010.

[102] K. Starbird and L. Palen. Will the revolution be retweeted? information diffusion and the 2011 egyptian uprising. In *Computer Supported Cooperative Work*, 2012.

[103] X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009, 2009.

[104] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, pages 177–184, 2010.

[105] L. Sun, S. Ji, and J. Ye. Adaptive diffusion kernel learning from biological networks for protein function prediction. *BMC Bioinformatics*, 9:162, 2008.

[106] L. Tang and H. Liu. Bias analysis in text classification for highly skewed data. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, 2005.

[107] L. Tang and H. Liu. Scalable learning of collective behavior based on sparse social dimensions. In *Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1107–1116, New York, NY, USA, 2009. ACM.

[108] L. Tang and H. Liu. Scalable learning of collective behavior based on sparse social dimensions. In *In 18th ACM Conference on Information and Knowledge Management*, 2009.

[109] L. Tang and H. Liu. *Community Detection and Mining in Social Media*. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers, 2010.

[110] L. Tang, H. Liu, J. Zhang, N. Agarwal, and J. J. Salerno. Topic taxonomy adaptation for group profiling. *ACM Transactions on Knowledge Discovery from Data*, 1(4):1–28, 2008.

[111] L. Tang, X. Wang, and H. Liu. Uncovering groups via heterogeneous interaction analysis. In *ICDM*, Miami, FL, USA, Dec. 6-9 2009.

[112] L. Tang, X. Wang, and H. Liu. Scalable learning of collective behavior. *IEEE Transaction of Knowledge and Data Engineering (TKDE)*, 2012.

[113] M. Thelwall. Homophily in myspace. *Journal of the American Society for Information Science and Technology*, 60(2):219–231, 2009.

[114] E. Tonkin, H. D. Pfeiffer, and G. Tourte. Twitter, information sharing and the london riots? *Bulletin of the American Society for Information Science and Technology*, 38(2):49–57, 2012.

[115] O. Tsur and A. Rappoport. What's in a hashtag? content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 643–652, 2012.

[116] X. Wang, S. Kumar, and H. Liu. A study of tagging behavior across social media. In *In SIGIR Workshop on Social Web Search and Mining (SWSM)*, 2011.

[117] X. Wang, L. Tang, H. Gao, and H. Liu. Discovering overlapping groups in social media. In *the 10th IEEE International Conference on Data Mining series (ICDM2010)*, Sydney, Australia, December 14 - 17 2010.

[118] X. Wang, L. Tang, H. Liu, and L. Wang. Learning with multi-resolution overlapping communities. *Knowledge and Information Systems (KAIS)*, 2012.

[119] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[120] D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393(6684):440 – 442, June 1998.

[121] B. Wellman. The school child's choice of companions. *The Journal of Educational Research*, 14(2):126–132, 1926.

[122] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, pages 261–270, 2010.

[123] S. White and P. Smyth. A spectral clustering approach to finding communities in graphs. In *SIAM 2005 Conference on Data Mining*, April 2005.

[124] H. Z. Xiaofeng, X. He, C. Ding, H. Simon, and M. Gu. Bipartite graph partitioning and data clustering. In *10th International Conference of. Information and Knowledge Management*, pages 25–32, 2001.

[125] Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: Collaborative tag suggestions. In *Proceedings of Collaborative Web Tagging Workshop at 15th International World Wide Web Conference*, 2006.

[126] J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in twitter. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[127] S. H. Yang, B. Long, A. Smola, N. Sadagopan, and Z. Z. H. Zha. Like like alike — joint friendship and interest propagation in social networks. In *the 20th International World Wide Web Conference (WWW'11)*, Hyderabad, India., March 2011.

[128] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su. Understanding retweeting behaviors in social networks. In *The 19th ACM International Conference on Information and Knowledge Management*, 2010.

[129] K. Yu and W. Chu. Stochastic relational models for discriminative link prediction. In *In Proceedings of Neural Information Processign Systems*, page 1553, 2006.

[130] K. Yu, S. Yu, and V. Tresp. Soft clustering on graphs. In *Advances in Neural Information Processing Systems*, page 05, 2005.

[131] T. R. Zaman, R. Herbrich, J. V. Gael, and D. Stern. Predicting information spreading in twitter. In *Computational Social Science and the Wisdom of Crowds Workshop*, 2010.

[132] D. Zhao and M. B. Rosson. How and why people twitter: The role that microblogging plays in informal communication at work. In *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 243–252, 2009.

[133] H. Zhou. Distance, dissimilarity index, and network community structure. *PHYSICAL REVIEW E*, 67:061901, 2003.

BIOGRAPHICAL SKETCH

Xufei Wang earned his Bachelor of Science degree in Physics from Zhejiang University in 2002. He received his Masters degree in Software Engineering from Tsinghua University in 2008. In the Fall semester of 2008, he entered the graduate college at Arizona State Univeristy to pursue his doctorate in Computer Science. He joined LinkedIn in 2012. He was invited as a program committee (PC) member for ICDM 2011 and IJCAI 2013, and an active reviewer for top tier conferences and journals.