Controversy Analysis:

Clustering and Ranking Polarized Networks with Visualizations

by

Sedat Gokalp

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

Approved April 2015 by the
Graduate Supervisory Committee:

Hasan Davulcu, Chair
Arunabha Sen
Huan Liu
Mark Woodward

ARIZONA STATE UNIVERSITY

May 2015

ABSTRACT

US Senate is the venue of political debates where the federal bills are formed and voted. Senators show their support/opposition along the bills with their votes. This information makes it possible to extract the polarity of the senators. Similarly, blogosphere plays an increasingly important role as a forum for public debate. Authors display sentiment toward issues, organizations or people using a natural language.

In this research, given a mixed set of senators/blogs debating on a set of political issues from opposing camps, I use signed bipartite graphs for modeling debates, and I propose an algorithm for partitioning both the opinion holders (senators or blogs) and the issues (bills or topics) comprising the debate into *binary opposing camps*. Simultaneously, my algorithm scales the entities on a *univariate scale*. Using this scale, a researcher can identify moderate and extreme senators/blogs within each camp, and polarizing versus unifying issues. Through performance evaluations I show that my proposed algorithm provides an effective solution to the problem, and performs much better than existing baseline algorithms adapted to solve this new problem. In my experiments, I used both real data from political blogosphere and US Congress records, as well as synthetic data which were obtained by varying polarization and degree distribution of the vertices of the graph to show the robustness of my algorithm.

I also applied my algorithm on all the terms of the US Senate to the date for longitudinal analysis and developed a web based interactive user interface **www.PartisanScale.com** to visualize the analysis.

US politics is most often polarized with respect to the left/right alignment of the entities. However, certain issues do not reflect the polarization due to political parties, but observe a split correlating to the demographics of the senators, or simply receive consensus. I propose a hierarchical clustering algorithm that identifies groups of bills that share the same polarization characteristics. I developed a web based

interactive user interface **www.ControversyAnalysis.com** to visualize the clusters while providing a synopsis through distribution charts, word clouds, and heat maps.

Dedicated to my mother and my father.

ACKNOWLEDGMENTS

Martin Luther King, Jr. once said "The ultimate measure of a man is not where he stands in moments of comfort and convenience, but where he stands at times of challenge and controversy." He was talking about a measure in quality. But as I was studying machine learning, I realized it could be possible to quantify the measurement. That inspired me to apply machine learning techniques to have explanatory models for politics.

I would like to express my gratitude to the people who made this possible. First and foremost, my advisor Dr. Hasan Davulcu had been an inspiring leader. His open minded approach to research has a great impact on me. His patience in the process, ambition in delivering, and joy of success took me through this long journey. I enjoyed celebrating every milestone with him.

I also would like to thank my committee members Dr. Arunabha Sen, Dr. Huan Liu and Dr. Mark Woodward for the amazing classes they taught, meetings we had to push our projects further, and their feedback on my dissertation.

One of the key figures in my life has been Sukru Tikves. He was a big brother and a great mentor to me during high school. After college, we ended up going to the same university for graduate studies, where he had been a unique friend whom I walked this path together with.

None of this would be possible if it wasn't for my beloved mother, father and sister. Regardless of how much I tried to hide the bumps on the road, my mother would immediately feel something would upset me from thousands of miles away, and sympathize. And my father is my consultant in life. He would always provide great insight about life choices, and help me make decisions that is best for me.

Last, but not the least, I would like to thank my dear wife Ayfer, and my son Irfan. I am the luckiest man in the world to have them in my life. Ayfer provided me her uninterrupted support in good days and bad. She patiently bore with me when I endlessly spoke my mind, while still providing feedback. And Irfan is my buddy who always makes my day with his smile after long hours.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Research Overview

Blogosphere plays an increasingly important role (Drezner and Farrell, 2008) as a forum of public debate, with knock-on consequences for the media, politics, and policy. Hotly debated issues span all spheres of human activity; from liberal vs. conservative politics, to extremist vs. counter-extremist religious debate, to climate change debate in scientific community, to globalization debate in economics, and to nuclear disarmament debate in security. There are many applications (Mullen and Malouf, 2006; Malouf and Mullen, 2007; Thomas *et al.*, 2006; Bansal *et al.*, 2008; Lin and Hauptmann, 2006) for recognizing politically-oriented sentiment in texts. Adamic and Glance (2005) studied linking patterns and discussion topics of political bloggers by measuring the degree of interaction between liberal and conservative blogs, and to uncover their differences. In this research, given a mixed set of blogs debating a set of related issues from two opposing camps, I propose an algorithm to determine (i) which blog lies in which camp, (ii) what are the contested issues, and, (iii) who are mentioned as the key individuals within each camp.

Bipartite graphs have been widely used (Deng *et al.*, 2009; Rege *et al.*, 2006; Zha *et al.*, 2001) to represent relationships between two sets of entities. I use bipartite graphs to model the relationships between blogs and issues (i.e. topics, individuals, etc.) mentioned within blogs. I use signed weighted edges to represent opinion strengths, where positive edges denote support, and negative edges denote opposition between a blog and an issue.

I develop algorithms to solve the following problems on signed bipartite graphs modeling blog debates:

1. Partitioning of both the blogs, and the underlying issues mentioned in blogs, into two opposing camps;

2. Scaling of both the blogs and the underlying issues on a univariate scale such that the position of a vertex is closer to the positions of the vertices it is connected with positive edges, and further away from the positions of the vertices it is connected with negative edges.

Using this scale, a researcher can identify both the moderate and extreme blogs within each camp, and the polarizing vs. unifying issues. Partitioning and scaling help a researcher to better understand the structure of a social, political or economic debate, or even the details of an emerging geopolitical conflict in the world. While extremist ends of a scale, may represent blogs with irreconcilable viewpoints, in some cases, moderate blogs may represent viewpoints that are more amenable to engage in a constructive dialog through a set of unifying issues. Moderates may sympathize with some of the claims and grievances of the other side. Longitudinal analysis using my proposed algorithms could reveal interesting dynamics, such as, moderates from opposing camps could be in the process of forming a coalition by making the necessary compromises to reach a consensus. All the while, moderates may be alienating extremists in their own camps who may choose to focus on polarizing issues only, and lash out violent or demonizing rhetoric on everyone else who do not share their exclusivist viewpoints.

Similarly, the current political party system in the United States is a two-party system, which suggests a bipolar nature for both the senators and the bills; such that, there exists two polarized camps of senators that oppose each others views, and two

sets of bills that polarize the senators. It can be presumed that these camps would purely split according to the political parties of the senators, or the political parties of the sponsors of the bills. Although this is true to a certain extent, my analysis show that the actual behaviors can be different for a minority.

Senators show their support/opposition along the bills with their votes. This information makes it possible to extract the polarity of the senators. I use signed bipartite graphs for modeling the opposition, and I used my previous work ANCO-HITS algorithm for partitioning both the senators, and the bills into two polarized camps. Simultaneously, my algorithm scales both the senators and the bills on a univariate scale. Using this scale, a researcher can identify moderate and partisan [1] senators within each camp, and polarizing vs. unifying bills.

Partitioning and scaling help a researcher to better understand the structure of political debates in the Senate. While partisan ends of a scale may represent senators with irreconcilable viewpoints, moderate senators may represent viewpoints that are more amenable to engage in a constructive dialog through a set of unifying issues. Moderates may sympathize with some of the claims and grievances of the other side. Longitudinal analysis using my proposed algorithms could reveal interesting dynamics, such as, moderates from opposing camps could be in the process of forming a coalition by making the necessary compromises to reach a consensus.

1.2 Contributions

To the best of my knowledge, simultaneous scaling on signed weighted bipartite graphs has not been studied in the literature, and this research is the first attempt to introduce the problem and provide an effective solution and evaluation strate-

[1] *Partisanship* can be defined as being devoted to or biased in support of a party.

gies. Similarly, hierarchical clustering of signed bipartite graphs preserving structural equilibrium is first attempted in this research.

Major contributions of this research are:

1. an iterative algorithm, named **ANCO-HITS** (Alternatingly Normalized CO-HITS), to propagate the scores on a signed bipartite graph to solve the partitioning and scaling problems described above;

2. a convergence proof for the proposed ANCO-HITS algorithm;

3. definition of a new coefficient to measure *structural equilibrium* for signed bipartite graphs using the multiplicative transitivity property presented by Kunegis *et al.* (2009) exemplified by the phrase *the enemy of my enemy is my friend*;

4. executing the ANCO-HITS algorithm on two real-world and one synthetic data sets:

   (a) scaling analysis of a total of 112 US Congress voting records having Republicans/Democrats and their *roll call votes* [2]

   (b) scaling analysis of top 22 liberal and conservative blogs, and the most influential individuals mentioned in these blogs.

   (c) performance evaluations of scaling algorithms using synthetic data sets which were obtained by varying polarization and degree distribution of signed bipartite graphs

5. a hierarchical clustering algorithm for signed bipartite graphs to identify subsets of entities preserving structural equilibrium

6. two web based interactive user interfaces

[2] `http://thomas.loc.gov/home/rollcallvotes.html`

(a) www.PartisanScale.com visualizes the output of the ANCO-HITS algorithm for the US Senate with year-over-year displacement analysis

(b) www.ControversyAnalysis.com visualizes the hierarchical clusters of US Senate, US House of Representatives, and the United Nations General Assembly, providing cluster synopsis through various visualizations.

In my experiments, variance in polarization relates to the distributions of the ratio of vertices corresponding to extremes vs. moderates.

Alongside my proposed ANCO-HITS, I also evaluated two baseline algorithms, namely CO-HITS (Deng *et al.*, 2009) and spectral clustering (Luxburg, 2007). Although Co-HITS was designed for scaling unsigned bipartite graphs, it can be directly applied for scaling signed bipartite graphs, and partitioning by considering the signs of vertex values. Spectral clustering algorithm was designed for partitioning of graphs, and it can also produce a scale by using the component values of the eigenvector associated with the second smallest positive eigenvalue of the graph Laplacian (Ng *et al.*, 2001; Shi and Malik, 2000).

My experiments showed that the ANCO-HITS algorithm is the only robust algorithm in the presence of variance in polarization and vertex degrees.

1.3 Problem Formulation

There are many applications (Mullen and Malouf, 2006; Malouf and Mullen, 2007; Thomas *et al.*, 2006; Bansal *et al.*, 2008; Lin and Hauptmann, 2006) for recognizing political orientation, and bipartite graphs (Deng *et al.*, 2009; Rege *et al.*, 2006; Zha *et al.*, 2001) have been widely used to represent relationships between two sets of entities. I use bipartite graphs to model the relationships between the senators and

the bills. I use signed edges to represent the votes, where positive edges denote support, and negative edges denote opposition on a bill by a senator.

Definition 1 (Co-Scaling problem for signed bipartite graphs)**.** *Given*

- $G = (U \cup V, A)$ *is a bipartite graph consisting of two disjoint sets of vertices* U *and* V*, and a signed adjacency matrix* A
- $U = \{u_1, u_2, \ldots, u_m\}$*, a set of* m *vertices*
- $V = \{v_1, v_2, \ldots, v_n\}$*, a set of* n *vertices*
- $A \in \mathbb{R}^{m \times n}$*, where* a_{ij} *represents the signed edge between* u_i *and* v_j

Find

- $X = (x_1, x_2, \ldots, x_m)$*, where* $x_i \in \mathbb{R}$ *is the assigned value of the vertex* u_i
- $Y = (y_1, y_2, \ldots, y_n)$*, where* $y_i \in \mathbb{R}$ *is the assigned value of the vertex* v_i

such that

- $sgn(x_i)$ *and* $sgn(y_i)$ *shall determine the* polarity *of the vertices i.e.* -1 *and* $+1$ *as the opposing polarities*
- x_i *value for a vertex* u_i *should be closer to the* y_j *values of the vertices that it supports (connects positively), and further away from the* y_k *values of the vertices that it opposes (connects negatively). The magnitudes of* x_i *and* y_j *denote the* extremity *of the nodes* u_i *and* v_j*. i.e. magnitudes closer to* 0 *meaning* more moderate *and larger magnitudes meaning* more extreme.

Figure 1.1 depicts a perfectly polarized bipartite graph. The two axes X and Y represent the univariate scale for the nodes in U and V. The vertices to the right

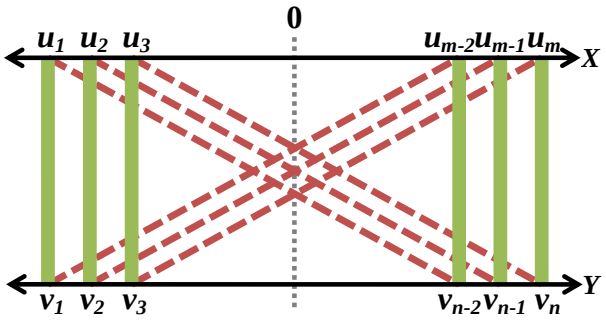


Figure 1.1: Perfectly Polarized Bipartite Graph

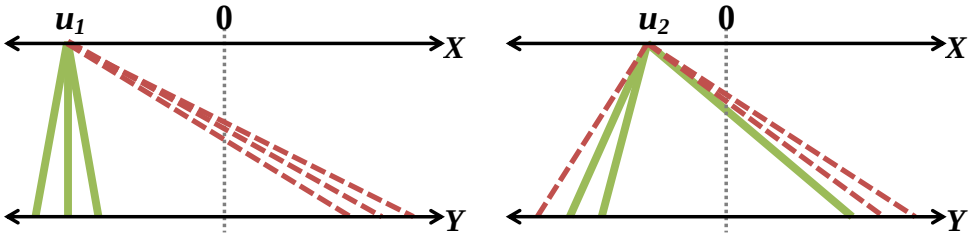


Figure 1.2: Extreme vs. Moderate Vertices

of zero have positive values, and the vertices to the left have negative values on the scale. A green solid line between the nodes u_i and v_j represents support, and a red dashed line represents opposition.

Figure 1.2 shows an example of two vertices; u_1 being extreme and u_2 being more moderate. u_1 supports the vertices of same polarity, and opposes the vertices of the opposite polarity. However, u_2 has mixed support and opposition.

Although partitioning algorithms can be utilized to detect the polarity of vertices, it is not possible to distinguish extremes from moderates. Scaling overcomes this problem and makes it possible to compare two vertices of same polarity. In this research, I am not only able to compare pairs of vertices, but also provide the exact locations on the scale, therefore providing valuable information about the shape of the distribution as well.

On the other hand, direct relations between the vertices may not be available, but indirect relations can be induced through intermediate vertices of different type. For example, signed relations between blogs may not be available, however through their polarized views towards people, one can identify the polarization between individual blogs.

To solve this co-scaling problem, I present two baseline methods. The first one is a common modification (Zha *et al.*, 2001; Fern and Brodley, 2004) of the well-known *Spectral Clustering* approach to work on graphs with signed edges. The second one is the CO-HITS (Deng *et al.*, 2009) algorithm, that is a modification of the well-known HITS algorithm, designed for bipartite graphs.

Finally, I compare these baseline methods with a novel algorithm I developed for co-scaling problem, named Alternatingly Normalized CO-HITS (ANCO-HITS).

Chapter 2

LITERATURE REVIEW

Scaling vertices of a graph based on the network structure rather than individual properties has been of great interest for more than a decade. Two most well-known algorithms are the PageRank (Page *et al.*, 1999) and the HITS (Kleinberg, 1999) algorithms. They were designed to rank the vertices of graphs with positive weighted edges. Spectral analysis show that both PageRank and HITS algorithms converge. An important distinction between the two algorithms is that; the HITS algorithm provides two different types of rankings corresponding to *hubs* and *authorities*, whereas PageRank provides only a single ranking.

Many data types from data mining applications can be modeled as bipartite graphs, examples include terms and documents in a text corpus, customers and items purchased in market basket analysis and bloggers writing about current issues.

Based on variations of HITS and PageRank, many researchers have proposed algorithms. Deng *et al.* (2009) propose a modification of the HITS algorithm to work on bipartite graphs called CO-HITS. The main difference between HITS and CO-HITS is that; HITS provides two scores for each vertex, whereas CO-HITS provides one score for each type of vertex. In this research, I use CO-HITS as one of the baseline algorithms, and in order to overcome its deficiencies, I extend it with normalization steps.

Data mining methods such as clustering have been used quite extensively for exploratory data mining applications (Dhillon *et al.*, 2001; Slonim and Tishby, 2000). Clustering analysis (Berkhin, 2006) provides a partitioning of the data into subsets, called clusters such that the objects in a cluster are more similar than those in distinct

clusters. Spectral clustering (Dhillon *et al.*, 2004; Luxburg, 2007; Ng *et al.*, 2001) is a powerful clustering method that is able to outperform K-means clustering (Hartigan and Wong, 1979) in many cases, especially when the clusters are non convex. The method is based on computing the eigenvalues of the normalized version of the graph Laplacian, and has theoretical connections with the normalized cut of the graph. In particular when clustering a bipartite graph into two balanced clusters, the second smallest positive eigenvalue (Ng *et al.*, 2001) is the solution to the normalized cut of the graph. In recent years, several authors have used spectral clustering to analyze bipartite graphs (Zha *et al.*, 2001). Furthermore, some work has been done to take into account a signed adjacency matrix by using an augmented adjacency matrix (Kunegis *et al.*, 2010). In this research, spectral clustering was also used as one of my baseline methods for partitioning and scaling signed bipartite graphs.

The *clustering coefficient* was first introduced by Watts and Strogatz (1998) to measure how much multiplicative transitivity property the graph exhibits, which reflects the tendency of the vertices to form small groups. Kunegis *et al.* (2009) define a new coefficient using the multiplicative transitivity for signed graphs to measure structural equilibrium. In this research, I define another coefficient through multiplicative transitivity for signed bipartite graphs.

SocialAction project (Perer and Shneiderman, 2006) integrates visualization for social network analysis. *Modularity* (Newman and Girvan, 2004), on the other hand, defines a measure for community detection purposes. Gómez *et al.* (2009) and Traag and Bruggeman (2009) further extend it to be used for signed graphs based on the assumption of structural balance. These approaches identify opposing communities. However, they lack scaling the individuals within the communities.

2.1 Spectral Clustering

Spectral clustering (Ng *et al.*, 2001) uses linear algebra methods for clustering purposes. The eigenvectors of the normalized Laplacian of the adjacency matrix are used to partition the graph into clusters. Spectral clustering is able to outperform K-means clustering in many situations, especially in the presence of non-convex groups of data. This method has close connections with the normalized cut (Luxburg, 2007) of the graph. In particular when clustering a bipartite graph into two balanced clusters, the second smallest positive eigenvalue of the Laplacian matrix (Dhillon *et al.*, 2004) is the solution to the problem of minimizing the normalized cut of the graph.

Spectral clustering uses an adjacency matrix with all positive entries. However, my problem assumes a signed adjacency matrix. One of the common techniques to circumvent this problem is to augment the matrix into a bigger matrix (Zha *et al.*, 2001; Fern and Brodley, 2004; Dhillon, 2001), such that all entries are positive. The first half of the augmented matrix is reserved for the entries with positive values, and the second half is reserved for the entries with negative values.

Define $\tilde{A} \in \mathbb{R}^{m \times 2n}$ such that $\tilde{A} = [A^+, A^-]$ where

$$a_{ij}^+ = \begin{cases} a_{ij}, & \text{if } a_{ij} > 0 \\ 0, & \text{otherwise} \end{cases} \qquad a_{ij}^- = \begin{cases} -a_{ij}, & \text{if } a_{ij} < 0 \\ 0, & \text{otherwise} \end{cases}$$

In order to partition and scale the nodes $u_i \in U$ and $v_i \in V$, I define the following matrix:

$$B = \begin{pmatrix} 0_{m \times m} & \tilde{A} \\ \\ \tilde{A}^T & 0_{2n \times 2n} \end{pmatrix}$$

I define the Laplacian of B as $L = D - B$ where D is the diagonal degree matrix and $d_{ii} = \sum_{j=1}^{m+2n} b_{ij}$. I further compute the normalized Laplacian $L_{sym} = D^{-1/2} L D^{-1/2}$. It should be noted here that both L and L_{sym} are positive semi-definite.

Let the eigenvalues of L_{sym} have the values $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{m+2n}$ with associated eigenvectors $v_1, v_2, \ldots, v_{m+2n}$, my univariate scale being the eigenvector v_2.

The first m components of v_2 are set to be the X vector, and the following n components are set to be the Y vector, solutions of the co-scaling problem.

2.2 CO-HITS

Deng *et al.* (2009) modify the well-known HITS (Kleinberg, 1999) algorithm and propose the CO-HITS algorithm which is used to rank vertices of a bipartite graph. Even though the adjacency matrix has only positive values in the original HITS paper, the theory still holds for adjacency matrices with signed entries.

Algorithm 1 describes the steps of the CO-HITS algorithm for the co-scaling problem.

The update functions for x and y are defined as follows:

$$x_i^{<k>} = \sum_{j=1}^{n} a_{ij} y_j^{<k-1>}, \qquad y_j^{<k>} = \sum_{i=1}^{m} a_{ij} x_i^{<k>} \tag{2.1}$$

and convergence is achieved when

$$\left\| x^{<k>} - x^{<k-1>} \right\|_2 < \epsilon \tag{2.2}$$

with ϵ a small positive value. In my experiments, I used $\epsilon = 10^{-20}$.

The intuition behind having such update rules is that each vertex should have positive edges towards the vertices with the same sign, and negative edges towards the vertices with the opposite sign.

Algorithm 1 Iterative update procedure for CO-HITS

procedure CO-HITS(A)
  $y^{<0>} \leftarrow (1, 1, \ldots, 1)$
  $k \leftarrow 0$
  repeat
    $k \leftarrow k + 1$
    Update $x^{<k>}$
    Update $y^{<k>}$
  until x vector converges
  return $x^{<k>}, y^{<k>}$
end procedure

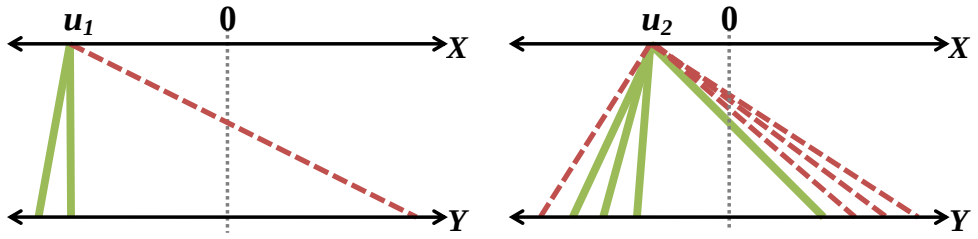


Figure 2.1: Extremity vs Degree

The drawback of this method is its sensitivity to the degree of each vertex, in the sense that the higher degree the vertex has, the higher score it will be assigned on the scale.

For example, let us consider two vertices u_1 and u_2 with u_1 having a smaller degree than u_2. However, let u_1 be more polarized than u_2 as shown in Figure 2.1. In this scenario, the corresponding scale values for u_1 and u_2 should satisfy $|x_1| > |x_2|$. But, this will not be the case with the CO-HITS. This suggests a better algorithm that accounts for the negative impact of degree variation through some normalization mechanism.

One can consider normalizing the adjacency matrix A only by its rows, such that the sum of each row adds up to 1. However, this will not take into account normalizing the other dimension.

Chapter 3

ANCO-HITS

3.1 Co-Scaling using ANCO-HITS

According to my problem formulation, the values of the vertices on the scale shall not be sensitive to their degrees, but rather be sensitive to what kind of relations they have with the other set of vertices.

For this purpose, I propose **ANCO-HITS** (Alternatingly Normalized CO-HITS) algorithm, which introduces a normalization mechanism to address the issue of degree sensitivity of CO-HITS. The proposed method uses the same iteration procedure described in Algorithm 1. The update functions for x and y vectors are modified such that they are normalized as follows:

$$x_i^{<k>} = \frac{\sum_{j=1}^{n} a_{ij} y_j^{<k-1>}}{\sum_{j=1}^{n} |a_{ij}|}, \qquad y_j^{<k>} = \frac{\sum_{i=1}^{m} a_{ij} x_i^{<k>}}{\sum_{i=1}^{m} |a_{ij}|} \tag{3.1}$$

Section 3.2 covers the proof that ANCO-HITS algorithm will have x and y vectors converge to the principal eigenvectors of M and N matrices which are derived from the original A matrix.

This research uses a modified normalization scheme than the original ANCO-HITS algorithm.

$$x_i^{<k>} = \frac{\sum_{j=1}^{n} a_{ij} y_j^{<k-1>}}{\sum_{j=1}^{n} |a_{ij} y_j^{<k-1>}|} \qquad y_j^{<k>} = \frac{\sum_{i=1}^{m} a_{ij} x_i^{<k>}}{\sum_{i=1}^{m} |a_{ij} x_i^{<k>}|} \tag{3.2}$$

The update functions for x and y vectors are modified such that the vectors x and y would converge not only in direction, but also in value. Furthermore, the

convergence values will satisfy $-1 \leq x_i, y_j \leq +1$. The results of all the datasets satisfied these conditions.

3.2 Proof of Convergence

Theorem 1. *ANCO-HITS algorithm will converge for any matrix $A \in \mathbb{R}^{m \times n}$, with $|A|$ having non-zero row-sums and column-sums.*

Proof. Let $B \in \mathbb{R}^{m \times m}$ and $C \in \mathbb{R}^{n \times n}$ diagonal matrices with positive entries, where

$$B_{ij} = \begin{cases} \dfrac{1}{\sum_{j=1}^{n} |a_{ij}|}, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

similarly,

$$C_{ij} = \begin{cases} \dfrac{1}{\sum_{i=1}^{m} |a_{ij}|}, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

I should note here that both B and C matrices are *symmetric* and *positive definite*. The update rules for x and y vectors can be written in matrix notation as follows:

$$x^{<k>} = BAy^{<k-1>} \tag{3.5}$$

$$y^{<k>} = CA^T x^{<k>} \tag{3.6}$$

Therefore,

$$x^{<k>} = (BACA^T)x^{<k-1>} \tag{3.7}$$

If there exists a vector x^* that the $x^{<t>}$ will converge in direction, it has to satisfy the equation:

$$cx^* = (BACA^T)x^* \tag{3.8}$$

Even though this is an eigenvalue equation, the eigenvalues may not be real, because the matrix $(BACA^T)$ is not symmetric. But if I multiply each side of the equation with $B^{-1/2}$, which exists since B is positive definite, I will get:

$$cB^{-1/2}x^* = B^{1/2}ACA^TB^{1/2}B^{-1/2}x^* \tag{3.9}$$

Define $M \in \mathbb{R}^{m\times m}$ to be $M = B^{1/2}ACA^TB^{1/2}$ and $z \in \mathbb{R}^{m\times 1}$ to be $z = B^{-1/2}x^*$, I will get

$$cz = Mz \tag{3.10}$$

which is again an eigenvalue equation. However, in this case M is a symmetric matrix, and can be shown to be positive semi-definite with z as an eigenvector c as an eigenvalue. The M matrix has a set of m eigenvectors that are all unit vectors and all mutually orthogonal; that is, they form a *basis* for the space $\mathbb{R}^m$.

Let us denote the eigenvalues of the M matrix by $c_1, c_2, \ldots, c_m$ sorted in such a way that $c_1 \geq c_2 \geq \ldots \geq c_m \geq 0$, with the eigenvectors $z_1, z_2, \ldots, z_m$ respectively.

Using Equation (3.7), we can write a compact form for the k^{th} update iteration of x as follows:

$$x^{<k>} = (BACA^T)^k x^{<0>} \tag{3.11}$$

We can rewrite the above equation in terms of M matrix

$$x^{<k>} = B^{1/2}M^kB^{-1/2}x^{<0>} \tag{3.12}$$

Any vector $v \in \mathbb{R}^m$ can be written as a linear combination of the eigenvectors $z_1, z_2, \ldots, z_m$. Therefore,

$$B^{-1/2}x^{<0>} = (a_1z_1 + a_2z_2 + \ldots + a_mz_m) \tag{3.13}$$

which will lead to

$$\begin{aligned} x^{<k>} &= B^{1/2} M^k (a_1 z_1 + a_2 z_2 + \ldots + a_m z_m) \\ &= B^{1/2} (a_1 M^k z_1 + a_2 M^k z_2 + \ldots + a_m M^k z_m) \\ &= B^{1/2} (a_1 c_1^k z_1 + a_2 c_2^k z_2 + \ldots + a_m c_m^k z_m) \end{aligned}$$

As k goes to infinity, the $x^{<k>}$ vector will converge to a multiple of the $B^{1/2} z_1$ vector.

$$\lim_{k \to \infty} \frac{x^{<k>}}{c_1^k} = a_1 B^{1/2} z_1 \tag{3.14}$$

Similarly, the convergence for the $y^{<k>}$ can be proved in the same fashion:

$$y^{<k>} = C^{1/2} N^k C^{-1/2} y^{<0>} \tag{3.15}$$

where $N \in \mathbb{R}^{n \times n}$ is $N = C^{1/2} A^T B A C^{1/2}$ and the $y^{<k>}$ vector will converge to a multiple of the $C^{1/2} q_1$ vector, with q_1 being the principal eigenvector of the N matrix. □

3.3 Experiments & Evaluations

To validate my algorithm, I have used two different datasets that are *US Congress* and *political blogosphere*. In addition to real data, I introduced a model to generate synthetic data to analyze the performance of the algorithms for various parameters.

Table 3.1 provides descriptive summarizes of the real-world data sets, as well as the partitioning accuracies of the algorithms. I cannot provide scaling accuracies due to the lack of quantitative information about how partisan/moderate each senator are. However, qualitative information regarding the partisanship of the senators can be obtained through a variety of resources. I will report this analysis in the corresponding section.

Table 3.1: Real-World Datasets Descriptive Summaries and Partitioning Accuracies

	111th US Senate	**111th US House**	**Political Blogs**
Vertices in U	64 Democrat 42 Republican Senators	268 Democrat 183 Republican Representatives	13 Liberal 9 Conservative Blogs
Vertices in V	696 Bills	1655 Bills	34 People
Graph Density	88.36%	91.23%	39.04%
Str. Equilibrium	39.47%	39.37%	87.21%
Spectral Clustering	100.00%	99.11%	75.39%
CO-HITS	100.00%	99.56%	98.21%
ANCO-HITS	100.00%	99.56%	98.21%

3.3.1 US Congress

The United States has a bicameral legislature that comprises the US Senate as the upper house, and the US House of Representatives. The terms of the US Senate last for two years, and the senators serve three terms (six years) each. The terms are staggered in such a way that approximately one-third of the seats are up for election every two years.

The Senate meets in the United States Capitol in Washington, D.C. to form and debate on motions, or bills. When debates conclude, the bill in question is put to a vote, where senators respond either 'Yea' (in favor of the bill) or 'Nay' (against the bill). For most of the bills, only the total number of 'Yea' and 'Nay' votes are recorded, except for the roll call votes. According to The Library of Congress [1] ,

[1] `http://thomas.loc.gov/home/rollcallvotes.html`

A roll call vote guarantees that every Member's vote is recorded, but only a minority of bills receive a roll call vote.

The US Congress has been collecting data since the very first congress of the US history. This data has been encoded as XML files and publicly shared through the govtrack.us project [2] .

To illustrate my results, I use the *roll call votes* for the 111^{th} US Congress which includes The Senate and The House of Representatives and covers the years 2009-2010. The 111^{th} Senate has 108 [3] senators and the data contains their votes on 696 bills, and The 111^{th} House has 451 representatives and the data contains their votes on 1655 bills.

I extracted the adjacency matrix $A \in \{-1, 0, 1\}^{|U|\times|V|}$, with U vertices representing the congressmen, and the V vertices representing the bills. The values a_{ij} are 1 if the congressman u_i votes 'Yea' for the bill v_j, -1 if the congressman votes 'Nay', and 0 if he did not attend the session.

The aforementioned scaling algorithms will scale both the congressmen and the bills. In presence of partisanship [4] in the Congress, the sign of the scale values for the congressmen should correspond to the Democrat and Republican parties, and the magnitude of the scale values should represent the amount of partisanship.

The first two columns of Table 3.1 provide information about this data and the partitioning accuracies of the algorithms.

Figure 3.1 depicts the vote matrices of the 111^{th} US Senate data, where rows representing the senators and the columns representing the bills. Also, the light green color represents 'Yea' votes, and dark red represents 'Nay' votes. Scaling these

[2] `http://www.govtrack.us/data`

[3] Normally, each congress has 100 senators (2 from each state), however in many of the congresses, there are unexpected changes on the seats caused by displacements or deaths.

[4] *Partisanship* can be defined as being devoted to or biased in support of a party.

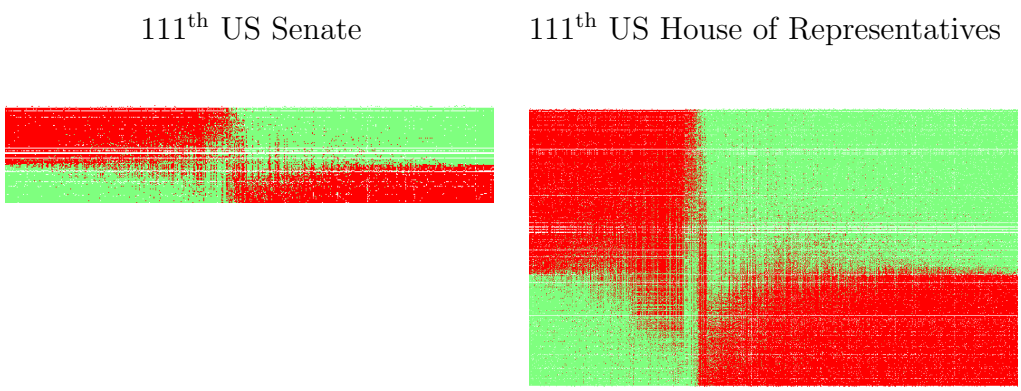


Figure 3.1: Vote Matrix After ANCO-HITS

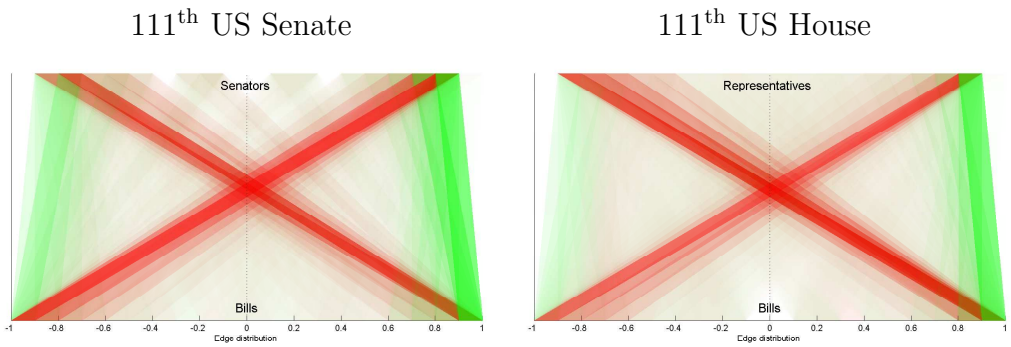


Figure 3.2: Bipartite Graph After ANCO-HITS

graphs leads to a re-ordering of the rows and columns such that senators and bills are co-clustered together.

I analyzed the congressmen that have been assigned to be moderate by each algorithm. I observed that the baseline algorithms tend to have the congressmen with less number of votes (i.e. lesser degree) to be moderate regardless of their partisanship. On the other hand, when I queried the names assigned to be most moderates by the ANCO-HITS, for both Democrats and Republicans, I was able to identify a number of supporting articles matching the ANCO-HITS scaling (Dennis, 2011; Newton-Small, 2009; Wikipedia, 2012; Coalition, 2012).

Figure 3.2 represents the bipartite graph of the 111^{th} US Congress data after scaling both the congressmen and the bills with ANCO-HITS. The light green colored edges represent 'Yea' votes, and dark red represents 'Nay' votes. Similar to my motivating Figure 1.1, this figure also shows partisan behavior in the 111^{th} US Congress.

In order to have a more extensive evaluation of the algorithms on real data sets, I executed the algorithms for each of the 111 terms in the US Senate. The analysis for the two major parties of each term show that ANCO-HITS algorithm partitions the senators with a higher accuracy than the baseline algorithms [5] : Spectral Clustering($\mu = 83.9\%, \sigma = 13.9\%$), CO-HITS($\mu = 85.9\%, \sigma = 13.7\%$), ANCO-HITS($\mu = 86.1\%, \sigma = 13.7\%$).

3.3.2 Political Blogosphere

As Web 2.0 platforms gained popularity, it became easy for web users to be a part of the web and express their opinions, mostly through blogs. In this study, I focus on a set of popular political liberal or conservative blogs that have a clearly declared positions. These blogs contain discussions about social, political, economic issues and related key individuals. They express positive sentiment towards individuals whom they share ideologies with, and negative sentiment towards the others. In these blogs, it is common to see criticism of people within the same camp, and also support for people from the other camp.

In this experiment, I collected a list of 22 most popular liberal and conservative blogs from the Technorati [6] rankings. For each blog, I fetched the posts for the 6 months before the 2008 US presidential elections (May - October, 2008) due to the

[5] μ: Mean accuracy, σ: Standard deviation

[6] `http://technorati.com/`

intensity of the debates and discussions. Table 3.2 shows the list of blogs with their URLs, political camps and the number of posts for the given period.

I use AlchemyAPI [7] to run a named entity tagger to extract the people names mentioned in the posts, and an entity-level sentiment analysis which provided us with weighted and signed sentiment (positive values indicating support, and negative indicating opposition) for each person. This information was used to synthesize a signed bipartite graph, where the blogs and people correspond to the two sets of vertices U and V. The a_{ij} values of the adjacency matrix A are the cumulative sum of sentiment values for each mention of the person v_j by the blog u_i.

To get a gold standard list of the most influential liberal and conservative people, I used The Telegraph List [8] for 2007. The third column of Table 3.1 provides information about this data and the partitioning accuracies.

3.3.3 Synthetic Data

The actual partitioning information for the real datasets were available, which made it possible to check the partitioning accuracy of the algorithms. However, to thoroughly check the scaling accuracy of the algorithms, I developed a method to generate random bipartite graphs with the following property:

- The degrees and the scores for the vertices in U and V follow independent probability distribution with varying parameters and shapes.

Algorithm 2 describes the method to generate random graphs.

I picked a normal probability distribution $\mathcal{N}(\mu, \sigma)$ for D_{degree} with values $\mu = 50$ and $\sigma = 5, 10, 15$. Similarly, for the probability distribution D_{scale}, I selected

[7] `http://www.alchemyapi.com/`

[8] `http://www.telegraph.co.uk/news/uknews/1435447/The-top-US-conservatives-and-liberals.html`

Table 3.2: List of Political Blogs

Blog name	URL	Political Camp	Posts
Huffington Post	www.huffingtonpost.com	Liberal	3959
Daily Kos	www.dailykos.com	Liberal	1957
Boing Boing	www.boingboing.net	Liberal	1576
Crooks and Liars	www.crooksandliars.com	Liberal	1497
Firedoglake	www.firedoglake.com	Liberal	1354
AMERICABlog	americablog.com	Liberal	1297
Think Progress	thinkprogress.org	Liberal	1197
Talking Points Memo	www.talkingpointsmemo.com	Liberal	1081
Wonkette	wonkette.com	Liberal	1064
Balloon Juice	www.balloon-juice.com	Liberal	923
Digby's Hullabaloo	digbysblog.blogspot.com	Liberal	553
Informed Comment	www.juancole.com	Liberal	179
Truthdig	www.truthdig.com	Liberal	159
Hot Air	hotair.com	Conservative	1579
Reason - Hit and Run	reason.com/blog	Conservative	1563
Little green footballs	littlegreenfootballs.com	Conservative	787
Atlas shrugs	atlasshrugs2000.typepad.com	Conservative	773
Stop the ACLU	www.stoptheaclu.com	Conservative	741
Wizbangblog	wizbangblog.com	Conservative	621
Michelle Malkin	michellemalkin.com	Conservative	532
Red State	www.redstate.com	Conservative	311
Pajamas media	pajamasmedia.com	Conservative	97

Algorithm 2 Procedure to generate random graphs

procedure RANDOMGRAPH(m, n, D_{degree}, D_{scale})
  $U = \{u_1, u_2, \ldots, u_m\}$ ▷ set of m vertices
  $V = \{v_1, v_2, \ldots, v_n\}$ ▷ set of n vertices
  $D_{u_i} \leftarrow$ RANDOM(D_{degree}), with $1 \leq D_{u_i} \leq m$ ▷ random node degrees
  $D_{v_j} \leftarrow$ RANDOM(D_{degree}), with $1 \leq D_{v_j} \leq n$ ▷ random node degrees
  repeat
    $i \leftarrow$ RANDBETWEEN$(1, m))$ ▷ random node pair
    $j \leftarrow$ RANDBETWEEN$(1, n))$ ▷ random node pair
    if $D_{u_i} > 0$ and $D_{v_j} > 0$ **then**
      $D_{u_i} \leftarrow D_{u_i} - 1$
      $D_{v_j} \leftarrow D_{v_j} - 1$
      if RANDBETWEEN$(0, 1) > (1 - |x_i|)(1 - |y_j|)$ **then**
        $a_{ij} \leftarrow sgn(x_i) \times sgn(y_j)$ ▷ consistent edge
      else
        $a_{ij} \leftarrow -sgn(x_i) \times sgn(y_j)$ ▷ inconsistent edge
      end if
    end if
  until D = 0
end procedure

Beta distribution (Rohatgi and Saleh, 2008) $Be(\alpha, \beta)$ with the parameters $\alpha = \beta = 0.1, 0.2, 0.5$.

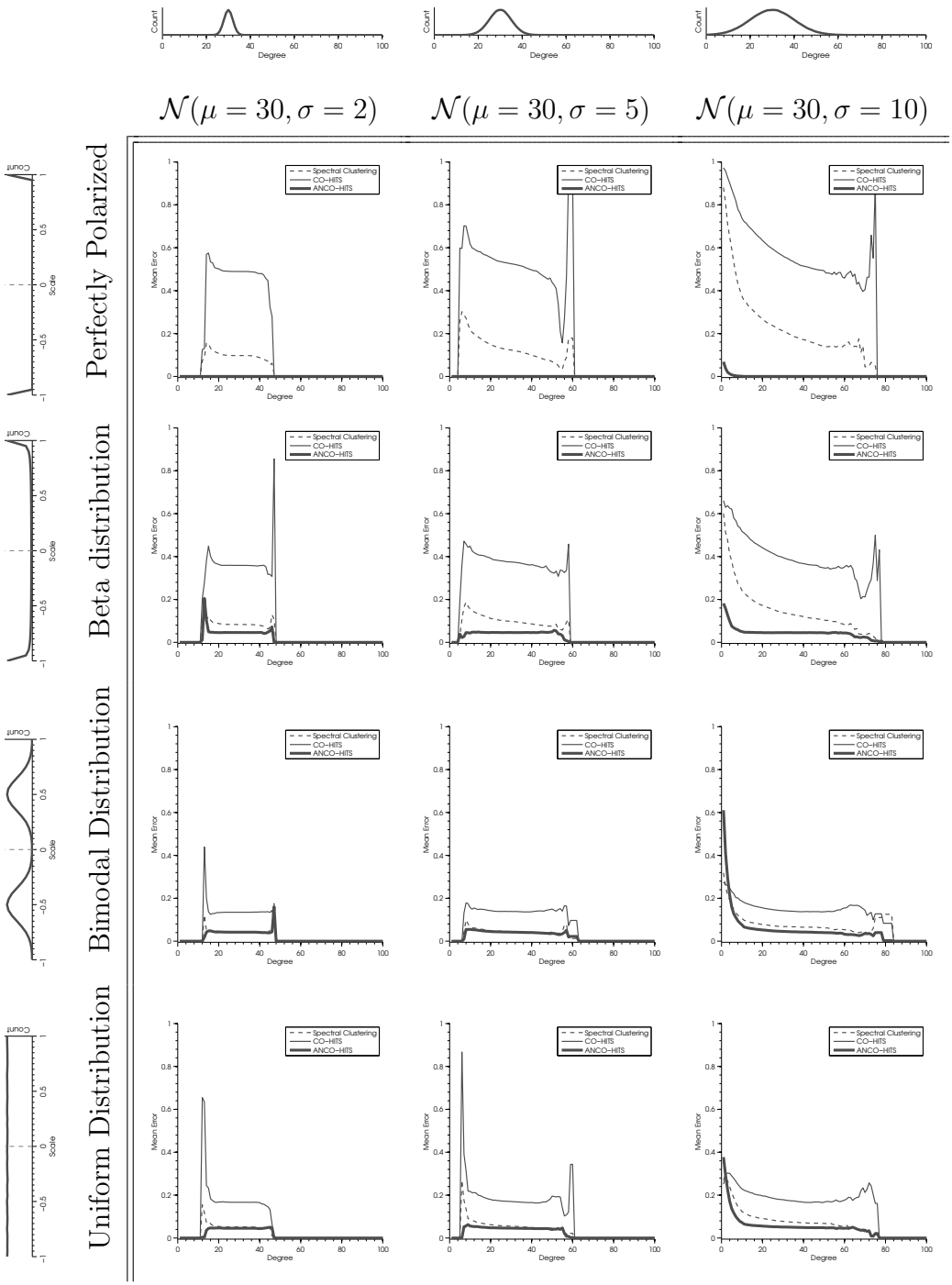
The difference between the scale obtained for a vertex by executing the scaling algorithm and its scale assigned by the random graph generator algorithm defines the error for that vertex. Table 3.3 shows the mean error vs vertex degrees plots for each algorithm applied to 12 different synthetic data sets.

In my experiments, the number of vertices of the graph is $m = n = 100$. I used four different distributions for varying polarization. These were perfectly polarized, *Beta*, bimodal and uniform distributions (Rohatgi and Saleh, 2008). Perfectly polarized distribution was obtained by mapping all vertices to the extremes of both sides with equal probability. I used three different normal distributions for varying the degree distributions of vertices. Degree distributions were obtained by $\mathcal{N}(\mu = 30, \sigma = 2)$, $\mathcal{N}(\mu = 30, \sigma = 5)$ and $\mathcal{N}(\mu = 30, \sigma = 10)$ in order to evaluate the effect of degree variance on the performance of the algorithms. I also experimented with different μ values of 10, 30, and 50 in order to measure the effect of density variations of the graph on the performance of the algorithms, which did not show any significant impact.

For each polarization and degree distribution I tested the performance of two baseline algorithms and my proposed algorithm. Table 3.3 presents these experimental results corresponding to 12 scenarios. In this table, columns correspond to the variance in degrees, and rows correspond to the polarization distributions.

In order to better visualize the effect of the degree of the vertex in determining its scaling position I used the mean error as an aggregate score. In the scatter plots, x-axis corresponds to the degree of vertices of a graph, and y-axis corresponds to mean scaling error. There are some error peaks at the boundary degree values due to their low frequencies.

Table 3.3: Synthetic Data Performances



From the table, we can make the following observations:

- Across all polarizations, as the vertex degree variance increases, overall errors for baseline algorithms increase due to their sensitivity to vertex degrees.
- Between the baseline algorithms, spectral clustering consistently outperforms CO-HITS.
- Even though spectral clustering performs almost as good as my proposed ANCO-HITS for bimodal and uniform polarization distributions, when the polarization is high, as in the other two distributions, its performance degrades.
- As polarization increases, from U-shaped to perfect polarization, ANCO-HITS performance increases. In case of perfect polarization, ANCO-HITS has almost no error.

Overall, in every single case my proposed ANCO-HITS algorithm outperforms the baselines.

3.4 Structural Equilibrium

The relation phrased as *the enemy of my enemy is my friend* is observed on various networks. This relation in general can be formalized for graphs by constraining any cycle of arbitrary length to have even number of negative edges (Hage and Harary, 1984). (Kunegis *et al.*, 2009) relate this constraint with the *multiplicative transitivity* property of an adjacency matrix, which can be measured using a modification of the clustering coefficient introduced by Watts and Strogatz (1998).

The structural equilibrium(SE) can be measured by checking the consistency of the edges forming cycles of length three. The relative signed clustering coefficient

calculates the ratio of balanced cycles among all possible cycles of length three.

$$SE(A) = \frac{\|A \circ A^2\|_+}{\left\|\bar{A} \circ \bar{A}^2\right\|_+}$$

where

- $\bar{A}$ is the absolute adjacency matrix such that $\bar{a}_{ij} = |a_{ij}|$
- $C = A \circ B$ is defined as the Hadamard product (element-wise product) for two matrices, such that $c_{ij} = a_{ij} * b_{ij}$
- $\|A\|_+$ is defined as the sum of all matrix elements, such that $\|A\|_+ = \sum_i \sum_j a_{ij}$

Bipartite graphs do not have cycles of odd length. Therefore, the structural equilibrium cannot be measured as formalized before. But it can be extended to calculate the ratio of balanced cycles of length four. For this purpose, I define the multiplicative transitivity for bipartite graphs as follows.

A signed bipartite graph exhibits multiplicative transitivity when a path of three edges tend to be completed by a fourth edge having a sign equal to the product of the three edges' signs.

This can be rephrased as *the enemy of my enemy of my enemy is my enemy*, or *the enemy of my friend of my enemy is my friend*, etc. Figure 3.3 depicts two cycles with odd number of edges (a and c), and two cycles with even number of edges (b and d). By definition, the cycles with odd number of negative edges do not satisfy multiplicative transitivity.

Hence, the corresponding relative signed clustering coefficient can be reformulated for bipartite graphs as follows:

$$SE(A) = \frac{\left\|A \circ AA^TA\right\|_+}{\left\|\bar{A} \circ \bar{A}\bar{A}^T\bar{A}\right\|_+}$$

For my experimental datasets, I report the corresponding SE values.

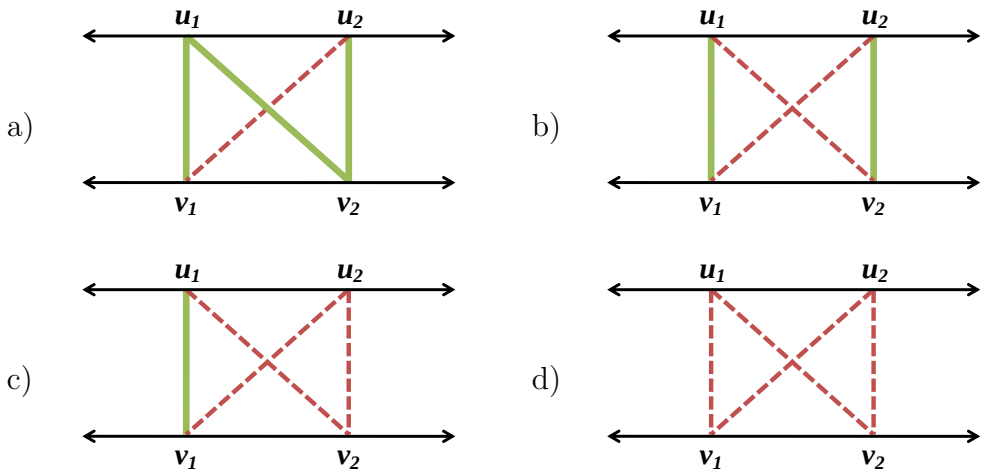


Figure 3.3: Cycles of Four with Negative Edges

3.5 Clustering Bills to Reveal Polarizing Issues

The behavior in US politics mostly correlates with the political party that the legislative is affiliated with. A senator from the democratic party will often support the bills proposed by other members of the democratic party, while often opposing to the bills proposed by the members of the republican party.

Although this is the most prominent behavior in US politics, it is still possible that a senator votes against the majority in his party. This breaks the structural equilibrium as defined in section 3.4, resulting in an imperfect bi-polarization. Political systems where there are multiple parties observe this more than US which has a two-party system. Even more so where there are no parties and each entity is independent from each other, such as in United Nations.

Each bill promoted for voting relate to one or more of the topics that the chamber is responsible with. Health care, foreign relations, budget and immigration are some of the topics covered in US Senate. Similarly, nuclear disarmament, human rights and environment are some of the topics covered in United Nations.

The overall voting behavior in a chamber usually doesn't present a perfect structural equilibrium. Nevertheless, it is possible to identify clusters of bills that has near-perfect bipolar nature. The bills in each cluster share the same characteristics not only in the way entities vote them, but also with the topics that they relate to.

Identification of bill clusters that polarize the legislators in different ways would help better analyze the political behavior in each chamber. Selection of an appropriate clustering algorithm plays an important role in achieving meaningful results. The data carries equivalent gravity as it can be assembled from various aspects.

One way to cluster bills could be through their descriptions. Content analysis could reveal clusters of bills based on their topics. However, it is possible that bills in one larger topic may not observe structural equilibrium, and it needs to be split into smaller clusters. Similarly, bills relating to two or more different topics may be polarizing the chamber in the same way, and should be contained in one cluster. Therefore, the content of the bills is not best aspect of data to use in clustering.

A better approach to clustering the bills is through analyzing the voting behavior. Collaborative filtering can unfold more interesting clusters than content filtering as it will expose the relation between the voters and the bills. Clusters with high structural equilibrium will have more descriptive power.

The bipartite graph described in Section 1.3 can be clustered to achieve high inter-cluster structural equilibrium using a range of algorithms. The classical k-means (MacQueen, 1967) algorithm can be modified with the following assignment and update steps.

Assignment: Assign each bill to the cluster which preserves highest structural equilibrium. Each cluster of bills $U^i \subset U$ with $i \in \{1, 2, ..., k\}$ is assigned as:

$$U^i = \{u_t : SE([A^i \;\; u_t]) \geq SE([A^j \;\; u_t]), \forall j \in \{1, 2, ..., k\}\}$$

Update: Derive voting matrix $A^i \in \mathbb{R}^{m \times n}$ from the original voting matrix A for each cluster $i \in \{1, 2, ..., k\}$ as:

$$a^i_{tk} = \begin{cases} a_{tk}, & \text{if } u_t \in U^i \\ 0, & \text{otherwise} \end{cases}$$

The algorithm assigns each bill to the cluster where other bills polarize the legislators in the same fashion, resulting with very informative clusters. Although very powerful, k-means is known to have various drawbacks such as sensitivity to noise, a priori knowledge of k, constraint on fixed density, etc.

Density-based spatial clustering of applications with noise (DBSCAN) (Ester *et al.*, 1996) addresses many of these issues, except for flexibility to varying densities. DBSCAN is a widely accepted clustering algorithm with various modifications available. Ordering points to identify the clustering structure (OPTICS) (Ankerst *et al.*, 1999) algorithm can be viewed as a generalization of DBSCAN in the sense that it can capture meaningful clusters when the data is of varying density in nature.

I modified the OPTICS algorithm in such a way that the notion of density is replaced by structural equilibrium. Legislatures are initially linearly ordered and than hierarchically clustered based on the agreement of polarizations which is measured by structural equilibrium. Finally, the hierarchy is represented in a dendrogram to visualize the clusters.

Chapter 4

APPLICATIONS

4.1 www.PartisanScale.com

I collected the *roll call votes* of the US Senate for the terms 1 through 112, covering the years 1789-2011. I ran the ANCO-HITS algorithm for each individual term. The sign of the ANCO-HITS values are arbitrary; therefore, I aligned consecutive terms by mirroring the scale if necessary. By analyzing more than 3,000,000 votes, I produced the web based interactive user interface **www.PartisanScale.com** that allows the users to navigate through the history of the US Senate.

Figure 4.1 shows a screenshot of the user interface. Each term of the senate is shown as a column in the figure. The top row shows the terms and the years for each senate with the incumbent US president shown below. The senators are represented by boxes which are colored according to their political parties.

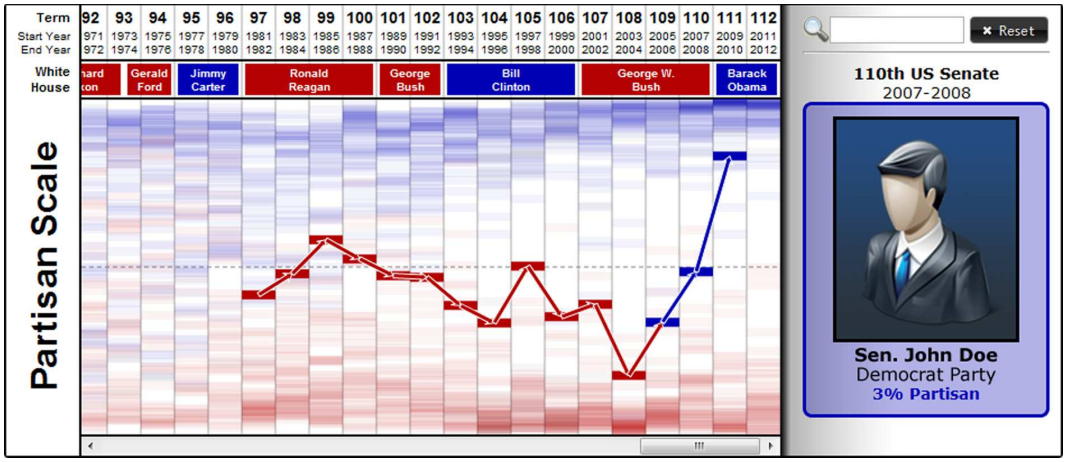


Figure 4.1: A Screenshot from PartisanScale.com

The vertical axis of the scale represents the bipolar nature of the US Senate. The polarity of each senator is represented by the location of each box. The dashed line shows the zero point. Senators around this point are calculated to be moderate, and the senators away from the dashed line are calculated to be more polarized. Hovering along these boxes will show the picture, the political party, and the amount of partisanship for the senator in focus. Clicking on the scale will further filter the figure to show the partisanship history. This filtering can also be done with the quick search tool on the top right corner. The auto-completion feature will help the users easily select the senator.

For example, Figure 4.1 shows a senator that is calculated to be moderate for the 110th term. It can be seen that this senator was first elected in 1981 and served for 15 terms until the year 2010. It also shows us that after 12 terms of service as a republican, he switches membership to the Democratic Party for the last 3 terms of his service.

The ANCO-HITS algorithm calculates the polarities for both the senators as well as the bills. However, based on feedback from area experts, I chose to display only the senators' partisanship along time and omit the bills. The decision is finalized after several iterations of design choices to give the visualization the strongest interpretability power.

The visualization can focus on individual senators through two different methods. Users can hover with the mouse over the region which corresponds to the years of interest, and partisanship of interest. By clicking on any block on the UI, the corresponding senator will be focused, and his/her partisanship history will be displayed. Similarly, users can use the search box to find senators by their name. The auto-complete feature will further help find the names faster.

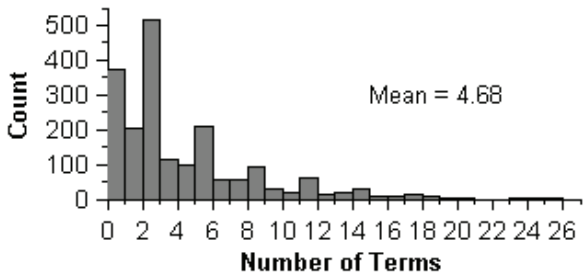


Figure 4.2: Longevity of Service

An introductory screen cast video that shows the usage of the system can be found at **http://www.youtube.com/watch?v=zCTiScyaPuw**.

4.1.1 Longevity of Service

Elected senators in United States serve staggered six-years, and they can race for re-election as many times as they want. In case of deaths, expulsions, or other reasons for early termination, mid-term vacancies are filled by special elections.

Figure 4.2 shows the histogram for the number of terms each senator served. Most senators are elected only once, and they complete their service by the end of the sixth year, hence the peak in the histogram at point 3. The average number of terms the senators served is 4.68, and the longest run is 26 terms.

4.1.2 Partisanship Displacement Distribution

Figure 4.3 shows the partisanship displacement distribution for three ΔT values on a semi-log scale. Partisanship displacement is defined as the absolute distance of partisan scale values for a senator between two terms T_1 and T_2. $C_{\Delta T}(d)$ is the number of displacements $\geq d$ between any two terms T_1 and T_2 satisfying $\Delta T = T_1 - T_2$.

This figure shows three plots of $C_{\Delta T}$ values for $\Delta T = 1$, $\Delta T = 2$ and $\Delta T = 3$. It can be clearly seen that the plots on the semi-log scale form a linear function, which suggests an exponential distribution. This implies that most senators have

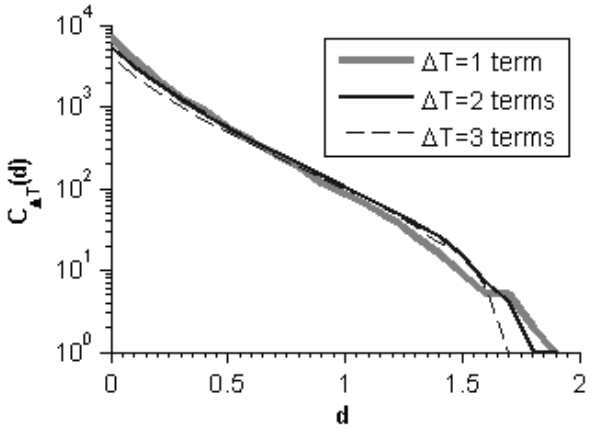


Figure 4.3: Partisanship Displacement Distribution

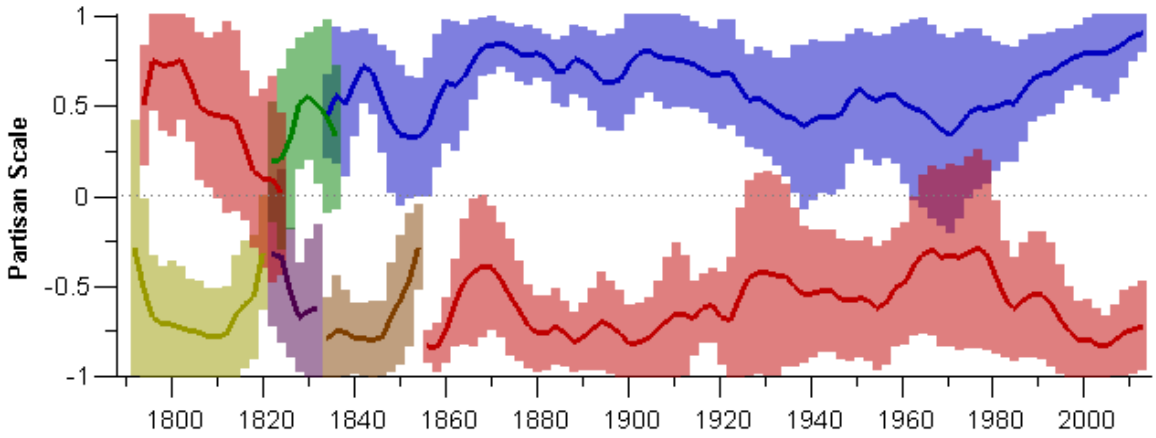


Figure 4.4: Aggregated Party Partisanship

a strong tendency to display consistent partisanship through their careers whereas fewer senators have experienced a complete turnaround.

4.1.3 Aggregated Party Partisanship

Figure 4.4 aggregates the party polarities. The mean partisanship values of the senators from each party is shown as a solid line. The shaded areas show 1 standard deviation along the mean for each term. This figure is helpful to identify the times of partisan politics within the US Senate.

4.2 www.ControversyAnalysis.com

The first application **www.PartisanScale.com** calculates the partisanship of each senator along with their polarities. However, it lacks the explanatory power needed to have better understanding of the analysis.

I conducted usability tests with political scientists and other area experts to pinpoint the second application **www.ControversyAnalysis.com** as shown in Figure 4.5. This work addresses the users' requests by

- identifying the groups of bills which polarize the legislatives in different fashions
- providing a synopsis of the polarization for each cluster of bills, and
- providing a synopsis of each cluster of bills by identifying the subjects covered

Below are few of the most common ways US legislatives are polarized in the congress:

- Consensus bills, or bipartisan bills, receive approval from the majority of the congress
- Partisan bills split the votes in the congress such that Democrats oppose Republicans
- Semi-partisan bills that receive opposition from a minority group of members of the sponsor party. These bills reveal factions within the large parties.
- Bills that split the congress in a different fashion than the political parties. These splits can be regional (i.e. south vs. north), religious (i.e. Catholics vs others), gender oriented (i.e. male vs. female), etc.

The ultimate measure of a man
is not where he stands in moments of ***comfort and convenience***,
but where he stands at times of ***challenge and controversy***.
Martin Luther King, Jr.

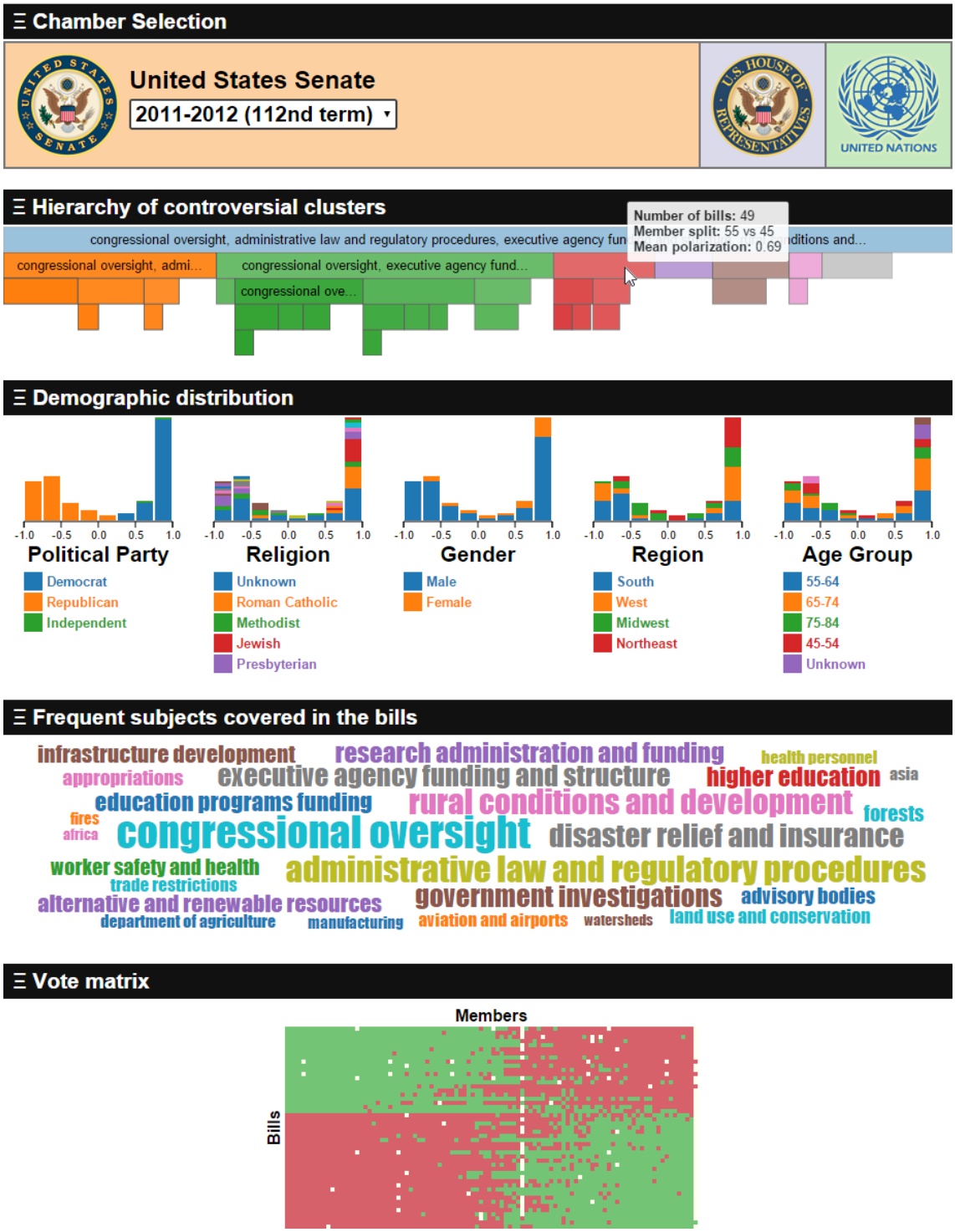


Figure 4.5: Web User Interface for www.ControversyAnalysis.com

4.2.1 Data

This application contains data from three chambers: US Senate, US House of Representatives, and the United Nations General Assembly. Details of the data from the first two chambers are explained in detail in Section 3.3.1. Only the last chamber will be spelled out in this section.

The United Nations General Assembly (UNGA) is a chamber where each member of United Nations have equal representation. UNGA meets in regular yearly sessions since 1946. Starting with 51 member nations, UNGA had a varying list of member states over the years.

UN states bring forth resolutions that they sponsor, which may cover issues from peace and security to industrialization, from diplomacy to UN administration and budget. These resolutions are generally non-binding on member states, but carry considerable political weight.

UNGA does not have a party system, therefore each member has an independent voice in voting resolutions. Although UN states do not explicitly have left or right wing alignment as in national politics, it is still possible to see the world nations form bi-polar camps in various issues.

I collected UNGA roll call votes for the dates 1946 through 2000. I grouped the data for every 5 years to enable temporal analysis. Similar to the US Senate data, I extracted the adjacency matrix $A \in \{-1, 0, 1\}^{|U|\times|V|}$, with U vertices representing the UN states, and the V vertices representing the resolutions. The values a_{ij} are 1 if the member nation u_i votes 'Yea' for the resolution v_j, -1 if the member nation votes 'Nay', and 0 if they did not attend the session, or was not a member of UN at the time.

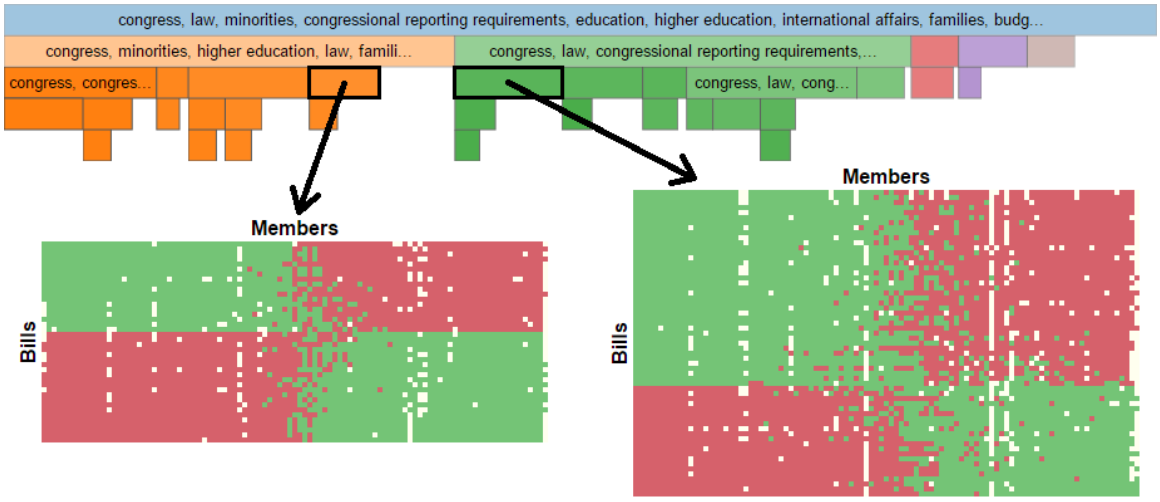


Figure 4.6: Two Clusters of Bills with High Structural Equilibrium Within and Low Structural Equilibrium Across

I did my analysis for the US Senate and the US House of Representatives covering 112 terms starting with the very first term (1789-1790) up to the 112th term (2011-2012). My analysis for UNGA contains 11 5-year time slots for years 1946-2000. The *Chamber Selection* section of the user interface allows users to select from the three chambers, and further narrow down to the particular time range of interest.

4.2.2 Hierarchical Clustering

Legislative chambers have bipolarity in their nature as the possible votes are either positive or negative. If we consider bills individually, they polarize the legislatives into two camps being supporters and opposers. It is very common that multiple bills polarize the legislature in the same fashion. These bills can be clustered together in such a way that clusters would have high structural equilibrium within, and low structural equilibrium across.

I used the hierarchical clustering approach explained in Section 3.5 to cluster the bills in each chamber within each time range. The cluster structure is visualized as

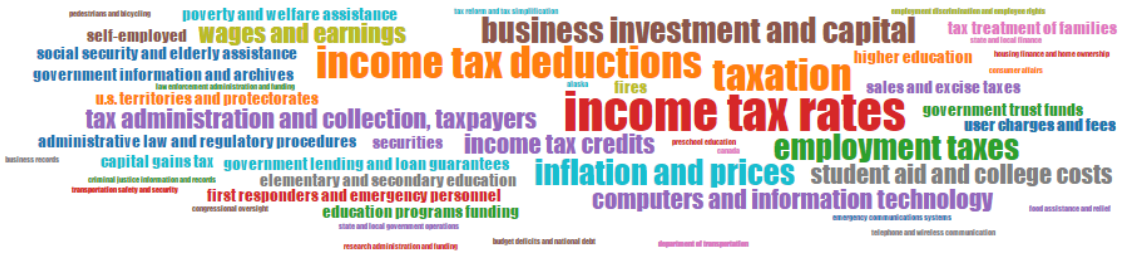


Figure 4.7: Word Cloud of Subjects Covered in a Cluster

a dendrogram in the *Hierarchy of controversial clusters* section of the user interface. Each block represents a cluster, and each main branch of the cluster tree is shown in a different color. The saturation of each block corresponds to the structural equilibrium value within the clusters.

Figure 4.6 shows an example dendrogram as well as the voting matrices corresponding to two different clusters. The vote matrices look similar in structure, however the order of the members in columns are different, i.e. legislators are polarized in a very different fashion. For example, one legislator identified to be moderate in the first cluster shows a polarized behavior in the second cluster. Detailed analysis reveals that the first cluster contains bills regarding internal affairs, wheres the second one is on foreign relations. The overall difference in polarization between the two clusters is so high that they are laid out on different main branches in the cluster tree.

4.2.3 Cluster Synopsis

The clustering algorithm in this analysis clusters only one side of the bipartite graph, i.e. each cluster contains different sets of bills, but all of the legislators. I run the ANCO-HITS algorithm as described in Section 3.1 on each cluster to derive a polarization value for the entities ranging between -1.0 and +1.0.

Figure 4.8: Distribution of Legislative Demographics Along the Bipolar Scale

There are two types of entities in this analysis, the bills and the legislatives. And I provide cluster synopsis from the two perspectives, the summary of bills and the polarity map of the legislatives.

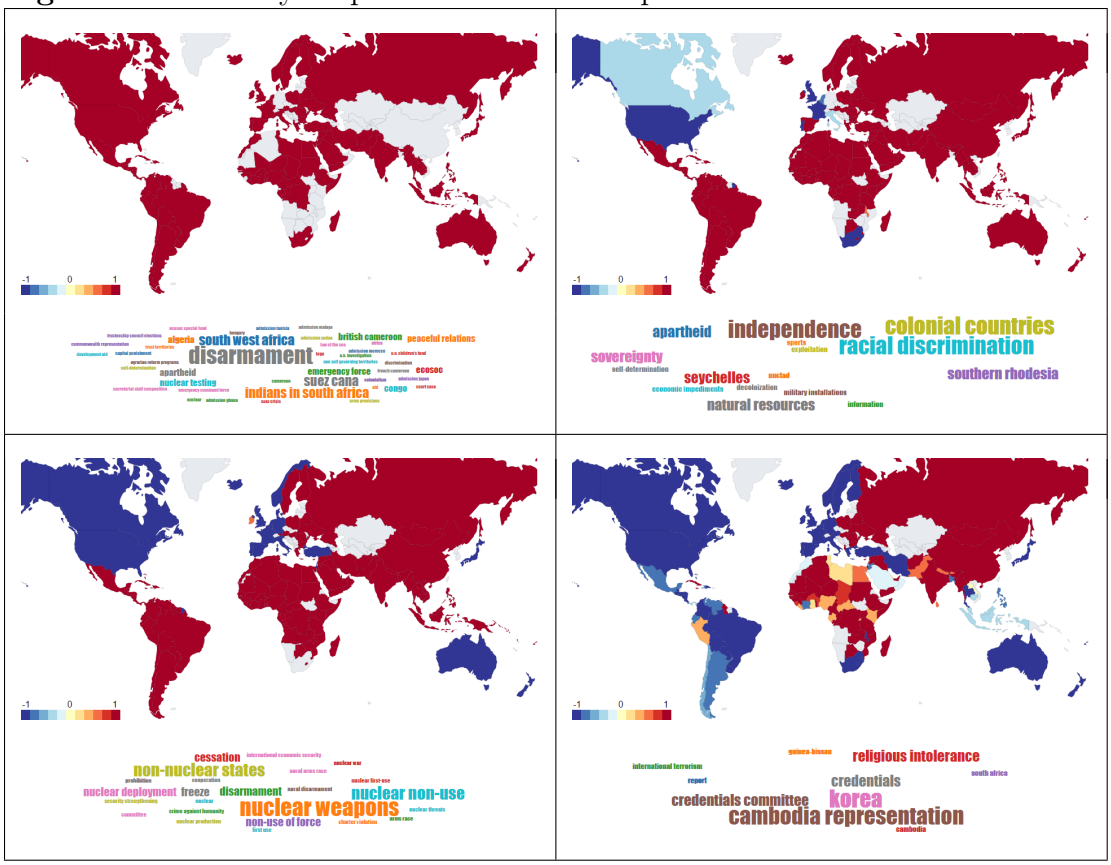
The subset of bills covered in each cluster polarize the chamber in a consistent way. And the legislatives take side on each topic. The combination of the sides taken by each legislator is often consistent across the bills that are on similar topics. Although not so often, it is also possible to observe consistent polarization across the bills that are on different topics.

The user interface section *Frequent subjects covered in the bills* provides a synopsis of the bills with a word cloud of the topics prominent in the bills. The topic sizes are weighed by the frequency observed in the cluster. Figure 4.7 shows an example cluster synopsis where the most prominent topic is *finance* and *taxation* in various industries.

Every single UNGA bill is marked with the prominent topics, however the US Senate and the House of Representatives data provide the topic information only for the dates after 1973. Therefore, this module is not available predating that.

The polarity map of the legislatives are visualized differently for US and UNGA. The entities in the US Senate and House of Representatives are persons, and their

Figure 4.9: Polarity Maps and Prominent Topics in Various Clusters in UNGA



distribution along the bipolar scale will be represented in terms of their demographics. Whereas the entities in UNGA are countries, and they will be laid out on a world map showing the polarization on a color scale.

The 5 demographics contained in the data are elected political party, religion, gender, region of elected state, and age group. Figure 4.8 shows an example distribution for a cluster that highly polarizes the US Senate. In the first glance, it can be thought that the entities are mostly polarized according to their political parties, i.e. most democrats are on the positive polarity, whereas most republicans are on the negative polarity. When paid further attention, the region demographic reveals that entities from south and west states position against the northeast states in this cluster of bills.

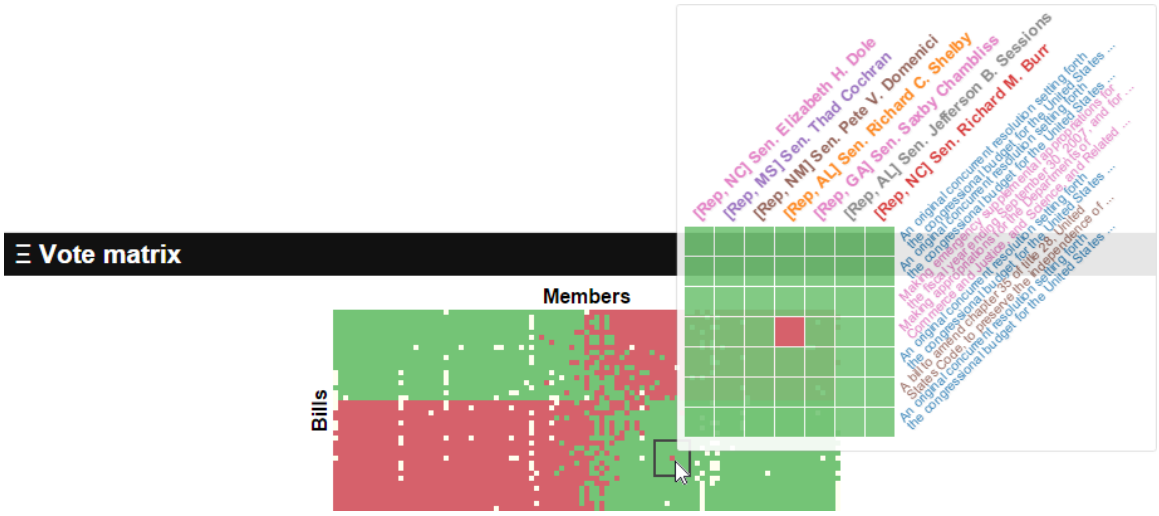


Figure 4.10: Vote Matrix with the *Microscope* Feature

Figure 4.9 depicts 4 clusters from UNGA with the corresponding polarity maps and word clouds. The top-left cluster is an example to a consensus case where all the UNGA members are on the same page. Prominent topics for the bills in this cluster are *disarmament* and *African countries*. The top-right cluster is about *colonialism* and few countries separate from the rest of the world. The bottom-left cluster about *nuclear weapons* displays another example of strong polarization whereas the bottom-right cluster observes lower structural equilibrium with a mixed set of topics.

Various synopsis modules provide high-level information regarding the clusters. However, *the devil is in the detail*. At this point, the *Vote matrix* module provides a visualization of the raw data to enable researchers to do low-level analysis. Rows and columns correspond to bills and members respectively. The cell colors represent the vote, i.e. green for 'Yea', red for 'Nay, and white when no vote was cast.

Both rows and columns are ordered by the ANCO-HITS score of the entity. And each cluster is generated to have high structural equilibrium. Therefore, the vote matrices always have a *quadrant* shape where top-left and bottom-right quadrants

are always green, and the top-right and bottom-left quadrants are always red. This is the picture of a well polarized chamber. Consensus clusters fit the quadrant shape as well, with the right half having length zero.

The *microscope* feature enables users to zoom-in and see which vote was cast on which bill by which member. Figure 4.10 shows this feature at work.

Chapter 5

CONCLUSIONS

In this research, I introduced a new problem for scaling and partitioning signed weighted bipartite graphs. I adapted two existing algorithms, and proposed a new algorithm to solve this problem. I used both real data from political blogosphere and US Congress records, as well as synthetic data to evaluate these algorithms. My experiments showed that my proposed algorithm is very effective and outperforms the two other baselines.

I see diverse applications of this not only on social network analysis, but also on survey analysis, with the respondents and the questions being the vertices, and their responses (agree/disagree) being the signed edges of the bipartite graph.

I developed an interactive visualization accessible at **www.PartisanScale.com** for longitudinal analysis of the US Congress. The system shows ANCO-HITS scales covering all voting records since the 1st US Senate.

The algorithms in source code and the test data is available online at www.PartisanScale.com/paperdata

I further designed and implemented a hierarchical clustering algorithm to identify various polarizations within each chamber. The interactive visualization accessible at **www.ControversyAnalysis.com** displays the cluster structure and synopsis.

The evaluations of the algorithm using officially maintained structured data shows superior performance. In the future, this algorithm can be used as part of a social network analysis pipeline to reveal disagreements within entities, and how it leads them to polarize. These analysis can be correlated with census or survey data for evaluation. Furthermore, it can even be utilized as a predictor for elections.

REFERENCES

Adamic, L. A. and N. Glance, "The political blogosphere and the 2004 u.s. election: divided they blog", in "Proceedings of the 3rd international workshop on Link discovery", LinkKDD '05, pp. 36–43 (ACM, New York, NY, USA, 2005).

Ankerst, M., M. M. Breunig, H. peter Kriegel and J. Sander, "Optics: Ordering points to identify the clustering structure", pp. 49–60 (ACM Press, 1999).

Bansal, M., C. Cardie and L. Lee, "The power of negative thinking: Exploiting label disagreement in the min-cut classification framework", Proceedings of COLING: Companion vol: Posters pp. 13–16 (2008).

Berkhin, P., "Survey of clustering data mining techniques", Grouping Multidimensional Data: Recent Advances in Clustering pp. 25–71 (2006).

Coalition, B. D., "Blue dog coalition", `http://bluedogdems.ngpvanhost.com/content/blue-dog-membership-1` (2012).

Deng, H., M. Lyu and I. King, "A generalized co-hits algorithm and its application to bipartite graphs", in "Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining", pp. 239–248 (ACM, 2009).

Dennis, S. T., "Senate moderates look for more influence", `http://www.rollcall.com/issues/56_90/-203808-1.html` (2011).

Dhillon, I., J. Fan and Y. Guan, "Efficient clustering of very large document collections", Data mining for scientific and engineering app. pp. 357–381 (2001).

Dhillon, I., Y. Guan and B. Kulis, "Kernel k-means: spectral clustering and normalized cuts", in "Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining", pp. 551–556 (ACM, 2004).

Dhillon, I. S., "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", KDD '01, pp. 269–274 (ACM, New York, NY, USA, 2001), URL `http://doi.acm.org/10.1145/502512.502550`.

Drezner, D. and H. Farrell, "The power and politics of blogs", Public Choice 134, 15–30 (2008).

Ester, M., H. peter Kriegel, J. S and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", pp. 226–231 (AAAI Press, 1996).

Fern, X. and C. Brodley, "Solving cluster ensemble problems by bipartite graph partitioning", in "Proceedings of the twenty-first international conference on Machine learning", p. 36 (ACM, 2004).

Gómez, S., P. Jensen and A. Arenas, "Analysis of community structure in networks of correlated data", Phys. Rev. E 80, 016114 (2009).

Hage, P. and F. Harary, *Structural Models in Anthropology* (Cambridge University Press, 1984), URL http://ebooks.cambridge.org//ebook.jsf?bid=CBO9780511659843.

Hartigan, J. and M. Wong, "Algorithm as 136: A k-means clustering algorithm", Journal of the Royal Statistical Society. Series C (Applied Statistics) 28, 1, 100–108 (1979).

Kleinberg, J., "Authoritative sources in a hyperlinked environment", Journal of the ACM (JACM) 46, 5, 604–632 (1999).

Kunegis, J., A. Lommatzsch and C. Bauckhage, "The slashdot zoo: mining a social network with negative edges", in "Proceedings of the 18th international conf. on World wide web", pp. 741–750 (ACM, 2009).

Kunegis, J., S. Schmidt, A. Lommatzsch, J. Lerner, E. De Luca and S. Albayrak, "Spectral analysis of signed graphs for clustering, prediction and visualization", in "Proc SDM", (Citeseer, 2010).

Lin, W. and A. Hauptmann, "Are these documents written from different perspectives?: a test of different perspectives based on statistical distribution divergence", in "Proc. of the 21st International Conf. on Computational Linguistics", pp. 1057–1064 (Assoc. for Computational Linguistics, 2006).

Luxburg, U. V., "A tutorial on spectral clustering", (2007).

MacQueen, J., "Some methods for classification and analysis of multivariate observations.", Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 281-297 (1967). (1967).

Malouf, R. and T. Mullen, "Graph-based user classification for informal online political discourse", in "Proceedings of the 1st Workshop on Information Credibility on the Web", (2007).

Mullen, T. and R. Malouf, "A preliminary investigation into sentiment analysis of informal political discourse", in "AAAI symposium on computational approaches to analysing weblogs", pp. 159–162 (2006).

Newman, M. E. J. and M. Girvan, "Finding and evaluating community structure in networks", Phys. Rev. E 69, 026113, URL http://link.aps.org/doi/10.1103/PhysRevE.69.026113 (2004).

Newton-Small, J., "Can ben nelson get a bipartisan stimulus win", http://www.time.com/time/politics/article/0,8599,1877535,00.html (2009).

Ng, A., M. Jordan and Y. Weiss, "On spectral clustering: Analysis and an algorithm", in "Advances in Neural Information Processing Systems 14: Proceeding of the 2001 Conference", pp. 849–856 (2001).

Page, L., S. Brin, R. Motwani and T. Winograd, "The pagerank citation ranking: Bringing order to the web.", Technical Report 1999-66, Stanford InfoLab (1999).

Perer, A. and B. Shneiderman, "Balancing systematic and flexible exploration of social networks", Visualization and Computer Graphics, IEEE Transactions on 12, 5, 693–700 (2006).

Rege, M., M. Dong and F. Fotouhi, "Co-clustering documents and words using bipartite isoperimetric graph partitioning", in "6th International Conference on Data Mining, 2006. ICDM'06.", pp. 532–541 (2006).

Rohatgi, V. and A. Saleh, *An introduction to probability and statistics* (Wiley-India, 2008).

Shi, J. and J. Malik, "Normalized cuts and image segmentation", Pattern Analysis and Machine Intelligence 22, 8, 888–905 (2000).

Slonim, N. and N. Tishby, "Document clustering using word clusters via the information bottleneck method", in "Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval", pp. 208–215 (ACM, 2000).

Thomas, M., B. Pang and L. Lee, "Get out the vote: Determining support or opposition from congressional floor-debate transcripts", in "In Proceedings of EMNLP", pp. 327–335 (2006).

Traag, V. A. and J. Bruggeman, "Community detection in networks with positive and negative links", Phys. Rev. E 80, 036115, URL `http://link.aps.org/doi/10.1103/PhysRevE.80.036115` (2009).

Watts, D. and S. Strogatz, "Collective dynamics of small-world networks", Nature 393, 6684, 440 (1998).

Wikipedia, "Factions in the republican party (united states)", `http://en.wikipedia.org/wiki/Factions_in_the_Republican_Party_(United_States)` (2012).

Zha, H., X. He, C. Ding, H. Simon and M. Gu, "Bipartite graph partitioning and data clustering", in "Proceedings of the tenth international conference on Information and knowledge management", pp. 25–32 (ACM, 2001).