Web Mining on Political Homophily in the Congressional Twitter Community

Alex Patton, William Strong, and Shieu-Hong Lin[1]
Department of Mathematics and Computer Science, Biola University, California, USA

Abstract. Homophily is the social phenomenon that people who are similar in some aspect interact at a significantly higher rate. We conducted web mining based on Twitter's application programming interfaces to investigate political homophily in the congressional Twitter community. We examined the structures of public Twitter communications among active members of the US congress. Our findings showed strong evidence of homophily with respect to party affiliation and a significant higher rate of communication for members in the same party. In an entirely distinct differentiation, we discovered moderate evidence of inverse homophily with respect to seniority and a significant higher rate of communication between members at different seniority levels.

Keywords. Homophily, Social Networks, Web Mining

Introduction

Homophily is the phenomenon that social interactions among people similar in some aspect occur at a significantly higher than the rate among dissimilar ones [5, 9, 10, 11]. As a contrast, inverse homophily is the phenomenon that social interactions among people similar in some aspect occur at a significantly lower rate than the rate among dissimilar ones [5, 9, 10, 11]. Homophily has been an interesting subject in modern sociological research for decades [9, 10, 11]. Homophily is closely related to the evolution of social networks fostered by the modern social media platforms. On the one hand, homophily happens because it is more likely for people with similar interests to interact and become connected through the social media platforms [3, 5]. On the other hand, people tend to draw their friends to join their online communities and communicate through the social media platforms, which fosters the phenomenon of homophily [5, 7].

In the previous research on political homophily in social networks [1], political blog sites in the web network structure in U.S. prior to the 2004 U.S. Presidential election were examined, and the result revealed strong homophily along two well-separated clusters of blog sites with heavy intra-cluster hyperlink connections and light inter-cluster connections. In this research, we investigated political homophily in the U.S. congressional Twitter community as a contrast. We studied public Twitter messages posted by active members of US congress to explore the presence of political homophily in these tweets with respect to party affiliation and seniority respectively.

Social media platforms such as Twitter and Facebook have been playing critical roles in the social life of the modern society, leading to thriving online communities

[1] Corresponding Author. Email address: shieu-hong.lin@biola.edu

connected as social networks with their own characteristics of communications [5, 16]. Twitter and other social media platforms provide application programming interfaces (APIs) for assess of public communication data generated by the online communities [16]. This enables researchers to explore a plethora of public Twitter messages (tweets) to study various phenomena of the underlying social networks [6, 8]. In this research, we implemented a web mining prototype based on Twitter's application programming interfaces (APIs) for the Python language to investigate political homophily in the congressional Twitter community. Our web mining approach for data collection was comparable to the frameworks depicted in [8, 16], but was implemented from scratch to target specific needs of data analysis on political homophily. Section 1 below describes the data collection framework and the raw data collected while Section 2 depicts an in-depth statistical analysis to examine the underlying political homophily.

1. Data Collection and Visualization of the Congressional Social Network

We implemented a web mining prototype for data collection to examine the congressional social network on Twitter based on their publicly posted tweets. For each congressman X active on Twitter, the prototype gathered and maintained the information of (i) the Twitter user name, (ii) the party affiliation, (iii) the number of years of service, and (iv) the Twitter user names of all the congressmen explicitly mentioned in congressman X's tweet messages. We can then visualize the congressional social network, in which vertices correspond to congressman X active on Twitter while edges indicate communication links between congressman. The following is an overview of key modules in the web mining prototype and a summary of the data collected.

- **Python-twitter module**: We implemented this module based on Twitter's application programming interfaces (APIs) for the Python language. It retrieved tweets as raw data for the project. Twitter's APIs set a limit on the number of tweets allowed to retrieve per 15-minute block of time and only allowed the program to pull up to 3200 tweets belonging to a given Twitter user. The module worked under these constraints to retrieve public tweets of U.S. congressmen as raw data for further analysis.
- **Twitter-controller module**: This module parsed and transformed the tweets retrieved by the python-twitter module into structural data in the JSON data format. For each congressman X, this module automatically identified the Twitter user names of other congressmen mentioned in congressman X's tweets and retain the information for further analysis.
- **Database-controller module**: This module managed the structural data gathered by the twitter-controller module and stored the information under a persistent MongoDB database on an AWS server. This allowed us to conveniently access the processed data from the MongoDB database for subsequent analysis, avoiding the slow process of repeatedly pulling data from Twitter when needed for analysis.
- **Summary of the data collected and its visual representations**: At the time of our study in the end of 2017, we were able to identify 54 US congressmen who were active on Twitter and retrieved their public tweet messages for analysis. Using the data gathered by our web mining prototype, Figure 1 and Figure 2 below show different views of the underlying social network rendered by the

graph-tool python library [15]. There are 54 vertices in the social network representing the 54 congressmen active on Tweeter. There are 427 edges in the social network. An edge between a pair of congressmen X and Y indicated X mentioned Y or Y mentioned X in at least one tweet. In other words, out of the 54 congressmen there are 427 pairs of congressmen who did show evidence of communication based on the public tweets retrieved from Twitter.

Figure 1 labels the vertices in the social network into two groups according to the party affiliation of the congressmen. Figure 1(a) presents the social network in a force directed layout [2, 15] clustered by party. Figure 1(b) presents the same network based on the nested block model in a radial layout [2, 15] clustered by party, with the size of each vertex scaled up proportional to the number of edges associated with the vertex. Figure 1(b) depicts the center of the radial layout and the centers of the two clusters respectively along the orientation axis in the middle. Two clusters are clearly visible largely along the party line. Both layouts in Figure 1 visually show strong homophily with respect to party affiliation since we can see a lot more communication links occurring between members of the same party than between members of different party.

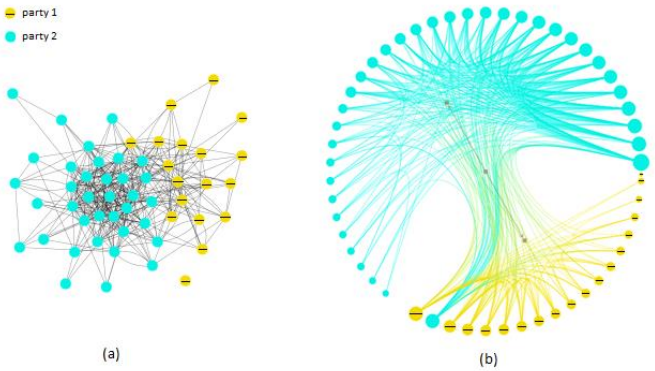

Figure 1. Two visual presentations of the congressional social network based on party affiliation

Figure 2 labels the vertices into two groups according to the seniority of the congressmen respectively. The junior congressmen have served less than ten years in the congress while the senior congressmen have served at least ten years. Figure 2 presents the social network in a force directed layout. As a contrast to homophily shown in Figure 1(a) with respect to party affiliation using the direct-force layout, Figure 2 does not show clear homophily with respect to seniority using the same direct-force layout.

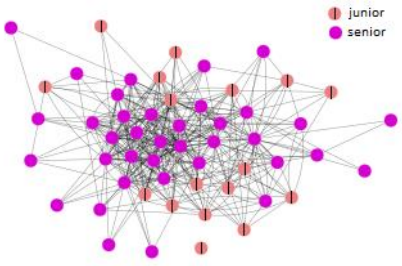

Figure 2. Visual presentation of the congressional social network based on seniority

We did not find good rendering results based on seniority to render the social network in a radial layout comparable to Figure 1(b) using the graph-tool python library [15] either. In the next section, we conduct further statistical analysis over the congressional social network above to rigorously examine the significance of homophily with respect to seniority and party affiliation respectively.

2. Statistical Analysis of Political Homophily in the Congressional Social Network

Consider a binary factor *f* (such as seniority or party affiliation in this study) with two possible statuses and a social network *G* of *m* edges. To conduct statistical analysis of homophily in *G* with respect to *f*, we first define the following terms and notation to refer to the characteristic parameter values involved in the statistical analysis.

- We use *p* and *q* to denote the fractions of vertices in G associated with the two different statuses with respect to *f* respectively. For a binary factor *f*, we have $p+q=1$.
- We refer to those edges that connect vertices with the same status with respect to *f* with respect to *f* as homogeneous edges in *G* with respect to *f*. We use m_1 to denote the number of homogeneous edges in *G*.
- We refer to those edges that connect vertices with different statuses with respect to *f* as heterogeneous edges in *G* with respect to *f*. We use m_2 to denote the number of heterogeneous edges in *G*.

For the congressional social network *G* in our study, Table 1 below shows the characteristic values of *G* with respect to party affiliation and seniority respectively. Note that with respect to party affiliation the number of homogeneous edges is much higher than the number of heterogeneous edges. As a contrast, with respect to seniority the number of homogeneous edges is close to the number of heterogeneous edges.

Table 1. Characteristic values of the congressional social network with respect to two factors

Characteristic parameter of *G*	Parameter value with respect to party affiliation	Parameter value with respect to seniority
p	18/54	37/54
q	36/54	17/54
number of edges: *m*	427	427
homogeneous edges: m_1	350	219
heterogeneous edges: m_2	77	208

2.1. Measuring statistical significance of homophily or inverse homophily

Assuming factor *f* plays no role in the formation of the *m* edges (communication links) in the social network *G*, we can independently pick *m* pairs of vertices from *G*, with each pair determining an edge, to model the formation process of these *m* edges. Under this hypothesis, each edge (as a pair of vertices independently drawn from *G*) would form a heterogeneous edge with probability *2pq* and the expected number of heterogeneous edges out of *m* edges would be *2mpq* [3, 13, 14]. If the number of heterogeneous edges in *G* with respect to factor *f* is significantly lower (higher) than *2mpq*, then there is evidence for homophily (inverse homophily) with respect to *f* in the social network [3].

Equivalently, a pair of vertices independently drawn from G would form a homogeneous edge with probability *1-2pq* and the expected number of heterogeneous edges out of m edges would be *m-2mpq*. If the number of homogeneous edges in G with respect to factor f is significantly higher (lower) than *m-2mpq*, then there is evidence for homophily (inverse homophily) with respect to f in the social network or [3]. This allows us to use the binomial test [4, 12] to measure the statistical significance of homophily or inverse homophily with respect to a factor f according to the following procedure.

- When the number of homogeneous edges m_1 is significantly higher than the expected *m-2mpq*, measure the statistical significance of homophily in G with respect to factor f by conducting a one-tailed binomial test to determine the probability of seeing m_1 or more successes from m independent binomial trials, each trial with a probability of *1-2pq* for success.
- When the number of homogeneous edges m_1 is significantly lower than the expected *m-2mpq*, measure the statistical significance of inverse homophily in G with respect to factor f by conducting a one-tailed binomial test to determine the probability of seeing m_1 or fewer successes from m independent binomial trials, each trial with a probability of *1-2pq* for success.
- The probabilities determined above are the p-values [4, 12] for judging the statistical significance of evidences regarding the presence of homophily or inverse homophily. The lower the p-values are the stronger the statistical significance is. It is common practice to consider a p-value of 5% or lower as statistically significant [12].

2.2. Findings about homophily and inverse homophily in the congressional network

For the congressional social network G in our study, Table 2 below shows the resulting statistics from the procedure above regarding homophily and inverse homophily in G with respect to party affiliation and seniority respectively. Note that with respect to party affiliation the actual number of homogeneous edges is much higher than the expected number of homogeneous edges. Therefore, we only calculate the p-value to determine the statistical significance of homophily respect to party affiliation, but not for inverse homophily. Similarly, with respect to seniority the actual number of homogeneous edges is lower than the expected number of homogeneous edges. Therefore, we only calculate the p-value to determine the statistical significance of inverse homophily respect to seniority, but not for homophily.

Table 2. Statistical significance of homophily and inverse homophily in the congressional social network

	Statistics with respect to party affiliation	Statistics with respect to seniority
p	18/54	37/54
q	36/54	17/54
1-2pq	0.556	0.569
Total number of edges (m_1)	427	427
Actual homogeneous edges (m_1)	350	219
Expected homogeneous edges	237.22	242.79
Actual heterogeneous edges (m_2)	77	208
Expected heterogeneous edges	189.78	184.21
Homophily: P value	< 0.00000001	
Inverse homophily: P value		0.011

3. Conclusions and Future Work

We implemented a web mining prototype in Python to collect information of the congressional social network on Twitter based on Twitter's application programming interfaces. Visual representations of the congressional social network in Section 1 clearly show evidence of homophily with respect to party affiliation. But homophily or inverse homophily with respect to seniority is not evident from the visual representations alone. Based on the statistical analysis depicted in Section 2 and the statistics shown in Table 2, we see strongly significant evidence of homophily with respect to party affiliation in the congressional social network (p-value < 0.00000001). We also see moderately significant evidence of inverse homophily with respect to seniority in the congressional social network (p-value around 0.01). In other words, with significant statistical evidence the findings indicate that (i) congressmen in the same party tend to interact more closely and (ii) congressmen at different seniority levels tend to interact more closely.

In addition to the study of homophily in the congressional social network, we had extended the web mining prototype to collect data on common Twitter followers of top news sites in US, such as Economist, Wall Street Journal, and New York Times. We had preliminarily examined around 2 million Twitter followers of these news sites to determine the numbers of common Twitter followers of pairs of new sites. We plan to expand the empirical study of political homophily to explore the correlation between (i) the numbers of common followers of news sites on Tweeter and (ii) the degrees of similarity in political orientations among news sites. The goal is to collect more data and quantitatively analyze the data to gain insight into political homophily regarding the behavior Tweeter followers of news sites.

References

[1] L. Adamic 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*, (2005), 36–43.

[2] G. Battista, P. Eades, R. Tamassia, and I. Tollis, *Graph Drawing: Algorithms for the Visualization of Graphs*, Prentice Hall PTR, 1998.

[3] R. Breiger, The duality of persons and groups, *Social Forces* **53** (1974), 181–190.

[4] M. DeGroot and M. Schervish, *Probability and Statistics (4th Ed.)*, Pearson, 2011.

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

[6] A. Giachanou and F. Crestani, Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *ACM Computing Surv*ey, **49:2** (2016), 1–41.

[7] D. Kandel, Homophily, selection, and socialization in adolescent friendships. *American Journal of Sociology*, **84** (1978), 427–436.

[8] S. Kumar, F. Morstatter, and H. Liu, *Twitter Data Analytics*, Springer, 2014.

[9] P. Lazarsfeld and R. Merton, Friendship as a social process: A substantive and methodological analysis. *In M. Berger, T. Abel, and C. Page, editors, Freedom and Control in Modern Society*, Van Nostrand, 1954.

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

[11] J. Moody, Race, school integration, and friendship segregation in America. *American Journal of Sociology*, **107** (2001), 679–716.

[12] H. Motulsky, *Intuitive Biostatistics (4th Ed.)*, Oxford University Press, 2018.

[13] M. Newman, Mixing patterns in networks, *Physical Review E*, **67: 026126**, (2003).

[14] M. Newman, D. Watts, and S. Strogatz, Random graph models of social networks. *Proc. Natl. Acad. Sci. USA*, **99** (2002), 2566–2572.

[15] T. Peixoto, The graph-tool python library, *figshare*, DOI: 10.6084/m9.figshare.1164194 [sci-hub, @tor], (2014).

[16] M. Russell and M. Klassen, *Mining the Social Web (3rd Ed.)*, O'Reilly Media, 2018.