The Spread and Quarantine of Anti-Vaccination Sentiment in Social Networks

Chan, Chi Ling (chiling@stanford.edu)
Liu, Raymond (rpliu@stanford.edu)

---

1. Introduction

It is posited, in Condorcets Jury Theorem and more recent emergence of ideas related to the wisdom of the crowds, that the exchange of dispersed information will enable socially beneficial aggregation of information. While this process of information exchange frequently leads to the formation of more accurate beliefs, society is not immunized from systematic biases and the spread of misinformation.

In the case of healthcare, misinformation can cost lives. In recent years, unsubstantiated concerns over the safety of vaccines, have led to anti-vaccination rhetoric to spread rapidly over the Internet. This has resulted in outbreaks - such as the whooping cough outbreak in California in 2010 that killed 10 children - that might otherwise be preventable.

As online social media provide access to data that can help identify target areas for communication and intervention efforts, there is increasing interest in the dynamics of health behaviors in social networks and how they affect collective public health outcomes. Our project is interested in examining the diffusion of misinformation in online social networks - specifically, how individual health behaviors are modulated by social networks.

2. Problem Definition

We define two central problems of interest in this paper:

1. **The formation of opinion clusters**, i.e, *homophily* - the assortative mixing of users with a qualitatively similar sentiments/opinions. In a network of opinionated users, the question of community structure naturally arises: are there communities where positive or negative attitudes dominate, and to what degree of homogeneity/heterogeneity? If there are such communities, what are their structures, and are there any structural differences? In particular, we wish to closely examine the structure of communities in our network which harbor the greatest level of anti-vaccination sentiment to better understand how to prevent the dissemination of anti-vaccination sentiment.
2. **Information diffusion and Negative Sentiment Reduction** - How does information reach a node, and what affects information adoption? Of interest is also the role of "influencers" - we want to characterize how the presence of well-connected, forceful agents interfere with information aggregation in a network. Specific to this problem, we wish to closely examine the role of influencers in spreading anti-vaccination sentiment, and the extent to which targeting nodes with negative sentiment may quarantine the spread of such sentiments.

3. Literature Review

The literature concern the two questions discussed above. The first set of papers analyzed different ways of detecting community structure in very large datasets.

3.1 Louvain Community Detection

The Louvian method extracts the community structure of large networks using a heuristic method based on modularity optimization. The principle advantage of this method over the Girvan-Newman is its ability to find high modularity partitions of large networks within a much shorter time.

The algorithm operates through two phases. In the first phase each node is assigned a distinct community so that there are as many partitions as there are nodes. For each node i the algorithm considers its neighbors j and computes the gain of modularity that results from removing i from the community and placing it in community j. Node i is then placed in the community that would optimize modularity gain. This process is iterated until no further modularity gain is possible.

In the second phase, a new network is constructed based on nodes with communities found in the first phase. Weights of links between the new nodes are given by the sum of the weight of links between nodes in corresponding two communities.

We chose to use the Louvian method for community detection due to the advantages that this method confers. Principally, the run time of this algorithm on a dataset as large as ours is significantly shorter. Since possible gains in modularity are easy to compute and the number of communities decreases drastically after several passes, running time is concentrated on the first iterations. As a result, its complexity is linear, which lends much greater time-efficiency. Furthermore, the steps for implementing this algorithm are intuitive and the outcome is unsupervised.

3.2 Statistical Properties of Community Structure in Large Networks

PageRank, the Google algorithm covered in class, will be used as a quantification of the amount of influence a node has within the network. The more "influential" the node, the higher in-degree (number of retweets) the node will have, meaning that the PageRank of the node will be greater.

NCP (Network Community Profile) plots, as defined by Leskovec (2008), are plots that measure the quality of the best possible community in a large network as a function of the size of the purported community. We use the NCP Plot implementation given by SNAP, which approximates the Minimum Conductance Cut Problem, an NP-hard problem.

The magnitude of local minima of an NCP plot at k tells us roughly how separable the graph is into communities of size k (Leskovec, 2008). In reality, we may find that the suggested values of k are not necessarily useful for determining precisely what our community sizes may be. For example, if k is too small (i.e. 5), then our communities will lack the structure of larger communities; on the other hand, if k is too large then we have fewer communities to work with.

Since different types of graphs (i.e. Erdos-Renyi, Small-World, copying, etc.) lend themselves to different types of communities, NCP plots are also useful for identifying the general structure of a network and the category of graph it happens to fall under (Leskovec, 2008).

3.3 Network Homophily

A key problem, as defined earlier, is the measurement of network homophily. Homophily is the robust tendency of people to associate more with those who are similar to them on some dimensions than with those who are not.

3.4 Data

Our dataset consists of publicly available tweets (short messages of ≤ 140 characters) collected from Twitter between August 2009 and January 2010 in the United States, a time in which pandemic influenza A(H1N1) was spreading nationwide and a vaccination had been developed. 477,768 tweets containing keywords relating to vaccination were collected by a team of researchers (Salath, 2011), of which a subset of 318,379 relevant tweets made it into the dataset. A machine learning algorithm was trained on the manually rated tweets to classify tweets as expressing a negative, positive or neutral sentiment towards influenza A(H1N1) vaccination.

Twitter data is ideal for two reasons: (1) It contains not only what is being broadcasted, but to whom it is broadcasted, and thereby allows us to study processes such as the spread of information, behaviors, opinions etc as well as the social structure on which these processes occur; (2) It provides a very structured mechanism for tracing exposure and contagions.

5. Methodology

5.1 Sentiment

Since each node has tweeted more than once, we sum up the relevant sentiments of their tweets, then divide the sentiment value by the number of tweets to determine the average sentiment of each node (herein, sentiment will be used for clarity, while average sentiment will refer to the arithmetic mean of the sentiments for every particular node). When calculating sentiment, irrelevant tweets (score -10) are not taken into account.

The average sentiment is calculated by taking the weighted average of every node's average sentiment values, with irrelevant tweets discarded. Two separate weighting functions were used: outdegree and PageRank.

In practice, the two strategies produced remarkably similar results. This is likely due to our dataset's representation of the community as an undirected graph, which complicates the identification of authorities under the PagRank algorithm. Note also that most communities tend to have average sentiments very close to neutral (0) as the majority of tweets have neutral sentiment.

Of the fifteen communities, the vast majority were positive, but the community with the greatest absolute difference from neutral was negative. It appears that negative communities have a greater absolute difference from neutral, but this result may be hard to generalize considering our small sample size of communities.

5.2 Homophily

Homophily is a measure of the variance of opinion between two nodes with an edge between them. For two nodes a and b with sentiments A, and B respectively, we define homophily to be the absolute

sum of A and B (i.e. $|A + B|$). We assign this homophily value to the edge which connects a and b; this edge is undirected.

Note that, for any two nodes, the homophily measure takes on values in the interval (0, 2); with 2 indicating that the two nodes share the same sentiment, and 0 indicating that the two nodes have completely opposing sentiments. Since the sentiments of a and b are not guaranteed to be integers, homophily is not necessarily a discrete-valued function.

The average homophily of a cluster may be said to be analogous to the variance of opinion within the cluster; the average sentiment would be analogous to the mean. We calculate the average homophily of a graph to be the sum of homophilies across all edges, divided by the number of edges. The average homophily gives us a general of idea of similarity between the sentiments of nodes in a cluster.

If we randomly-assign sentiments to nodes, we will find that the expected value of homophily in a graph should approach some constant c. Note that $c = 1$ for Erdos-Renyi graphs, but may not necessarily be 1 depending on the structure of the graph in question. In our Twitter communities, for example, the average c value was empirically determined to be around 1.11.

For the same graph, if the value of the true homophily h significantly differs from the mean, then we know that either the nodes in the graph tend to share the same sentiment ($h > c$) or contrasting sentiments ($h < c$).

5.3 Deletions and Interventions

The goal of a policy intervention is to correct anti-vaccination misinformation by encouraging pro-vaccination sentiments in the network. In this simulation, we simulate several intervention strategies and measure the efficacy of a policy by the overall change in average sentiment of the network.

For our network, we will simulate this suppression of negative posts with a series of "interventions". If we choose to intervene on some node v, we set the sentiment of v to neutral (0); essentially, we prevent v from saying anything negative.

Consider some node w who, in the original graph, has retweeted v's negative post. If v cannot post anything negative, then w could not have retweeted v's post and would perhaps not have spread these negative sentiments. Thus, it is possible that interventions will incur cascading effects, but it is also possible that w will spread negative sentiments without v having tweeted them first. Ergo, we will model this intervention process via probabilistic contagions. To further simulate the case where w already harbors negative sentiments, we will allow w to only catch the contagion exactly once; any negative node will only be subject to at most a single intervention.

In addition, we will perform targeted deletions on the most influential nodes within the negative communities to examine the overall structure of communities with negative sentiment.

This gives us the following three strategies with which we will experiment and observe impacts on average network sentiment and network homophily.

1. **VIN Assassination**: Delete Very Influential Nodes (disregarding their individual sentiments) within predominantly anti-vaccination communities, starting from the most influential nodes (by PageRank scores). We graph the average sentiments and homophily versus the proportion of nodes deleted; the slope of the sentiment graph tells us what kinds of nodes are deleted. If

the slope is positive, we are deleting negative nodes and vice-versa. If the slope is 0, we are deleting neutral nodes.

2. **Random Intervention**: Convert negative nodes to neutral nodes at random, disregarding PageRank scores. Each initial target proceeds to convert surrounding negative nodes with 0.10 probability.

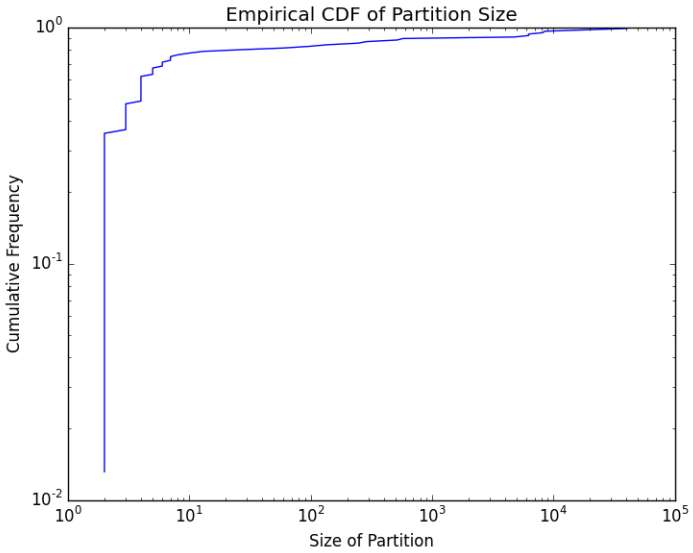
3. **Targeted Intervention**: Convert negative nodes to neutral nodes, starting from the most influential nodes with highest PageRank scores. Each initial target proceeds to convert surrounding negative nodes with 0.10 probability.

Note that 0.10 probability was chosen based on the observation that there is no significant change in overall average sentiment beyond the 0.10 point.
Of the fifteen communities we identified, we have three negative communities (1, 6, 14); we will run the above tests on communities 1 and 6, which also happen to be our largest communities (see appendix). We refrain from reporting the results of our tests on Community 14, which, despite having the most negative average sentiment of all our communities, consists of only 27 sparsely-connected nodes. Because Community 14 has so few nodes, slight alterations create extreme deviations, making any results inconclusive.

6. Findings

6.1 Community Detection

Applying the Louvian method for community detection, 31 communities were detected. The following is the empirical cumulative distribution function of partition size in our network.



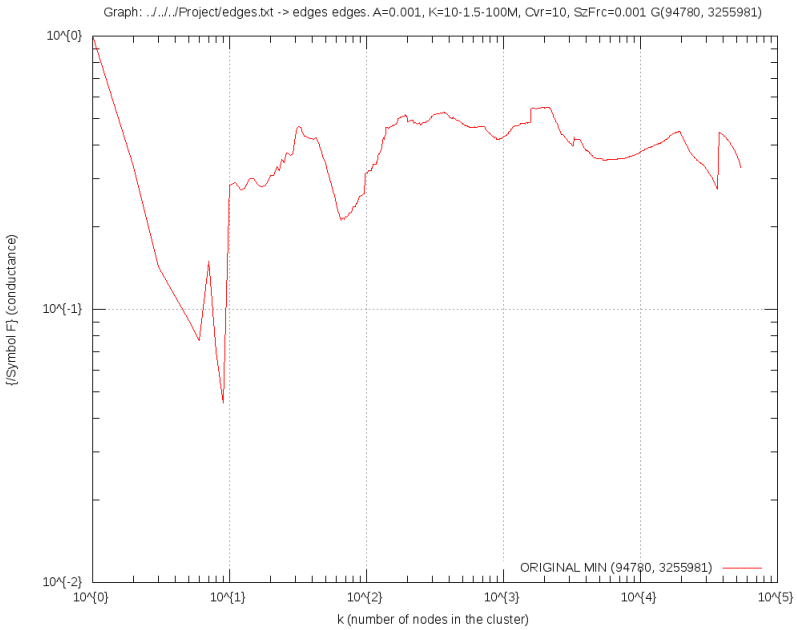
As is evident in this network, there is a significant number of small-sized communities. We consider a usable, non-trivial community to be a community with a node count of >20 nodes. Based on this standard, we derive 15 usable communities. The largest community has 40217 nodes, while the smallest has 27 nodes.

For each community, we compute the average sentiment using two weighting functions (degree and pagerank). We find that a vast majority of communities have a positive average sentiment, and only

3 out of the 15 have a negative average sentiment, a finding consistent for both weighting functions.

Apart from average sentiment, we also compute the following metrics for each community: node count, edge count, network density, sentiment ratio and percentage of neutral nodes.

6.2 NCP Plot Analysis



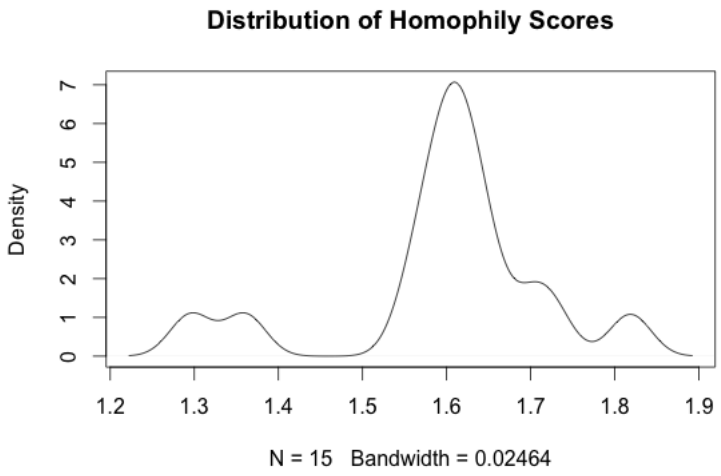
The NCP Plot generated on our Twitter data appears to approximate a copying graph, a class graph modelled upon the copying mechanism, in which nodes and their edges are imperfectly copied. This type of graph models a graph in which new nodes are interested in the same existing content (topics) as other nodes, therefore causing new nodes to have connections cosmetically similar to those of existing nodes.

Intuitively, it appears to make sense that the Twitter network would be a copying graph. The most popular tweets on any particular subject are disseminated via retweets and can be viewed by groups of people with particular interests, who may choose to further retweet these tweets.

The NCP suggests various community sizes (i.e. the values of k which are local minima); most notably around 10, 70, 850, and 38,000. None of these sizes other than 850 would be particularly feasible; 10 and 70 yield far too many communities (10000 and 1500 communities respectively) while 38,000 yields too few (about 3). Nevertheless, given the viral nature of Twitter and the arbitrary size of its communities (just as there are people with hundreds of followers, there are people with thousands and even millions), we will not impose any limits on community size.

6.3 Community Homophily

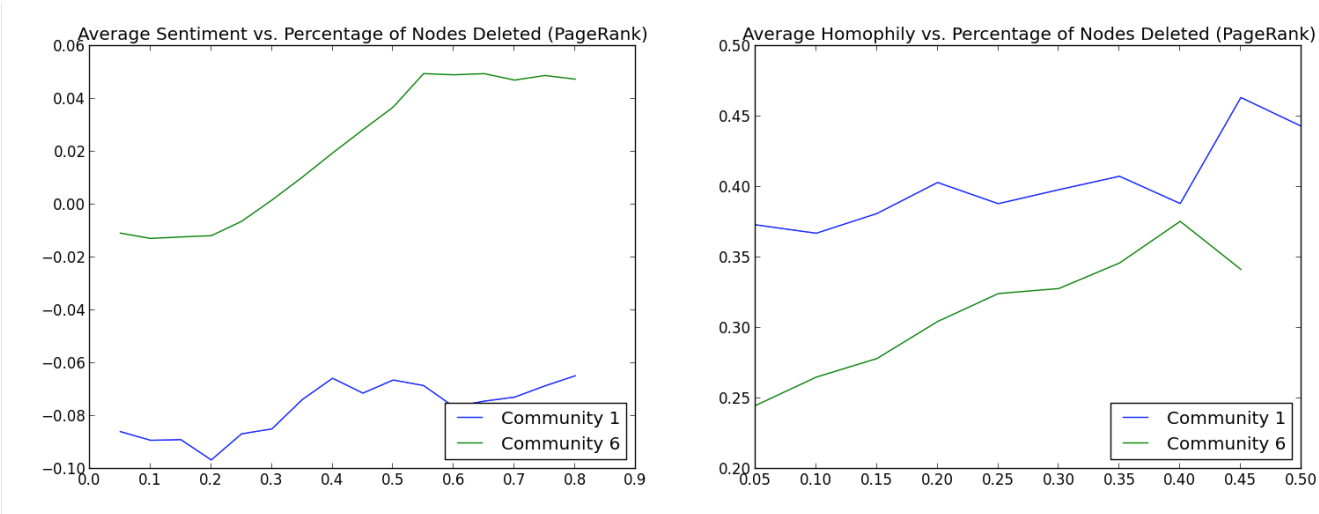
Computing the homophily score for all 15 communities we get a homophily score ranging from 0.181 to 0.704. Below is the density plot for our homophily measure. In all communities the true homophily scores deviated significantly from random (see Appendix).



Among communities that are overall negative in sentiment, the homophily score is 0.425, indicating a slightly higher degree of homophily than those that are overall positive in sentiment (average homophily score = 0.395). This first-cut examination suggests that communities with negative sentiments have a higher degree of homophily overall. Additionally, the two largest negative communities have exceptionally high densities, with homophily scores lower than the average homophily score of the 15 communities. Note that this relationship between community density and sentiment is not necessarily correlated, as 15 communities is too small a sample size.

6.4 VIN Assassination on Negative Communities

As mentioned in our methodology, VIN Assassination was applied with the view of understanding changes in network structure as influencers are removed. We begin with communities 1 and 6, both communities with overall negative sentiments.



6.4a Sentiment

The above figure demonstrates that the nodes with highest influence mostly hold negative sentiments, since average sentiment increases as the most influential nodes get deleted. Community 1 is always more homophilic than Community 6, which makes sense given that Community 6 has the strange property of having negative average sentiment overall but more positive nodes (this means

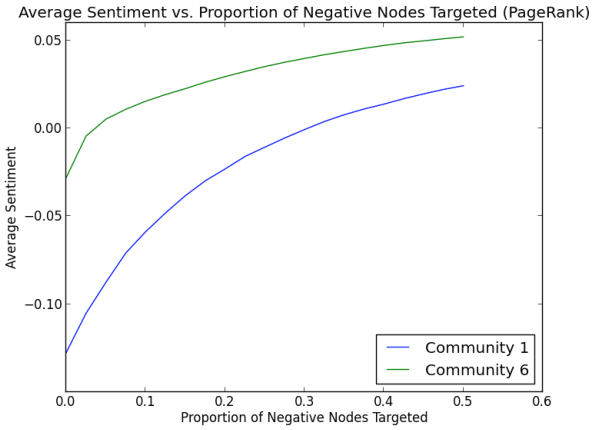
the negative nodes must be more influential).

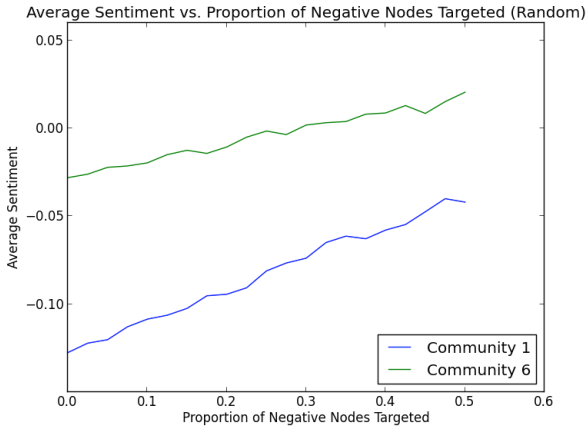
6.4b Homophily

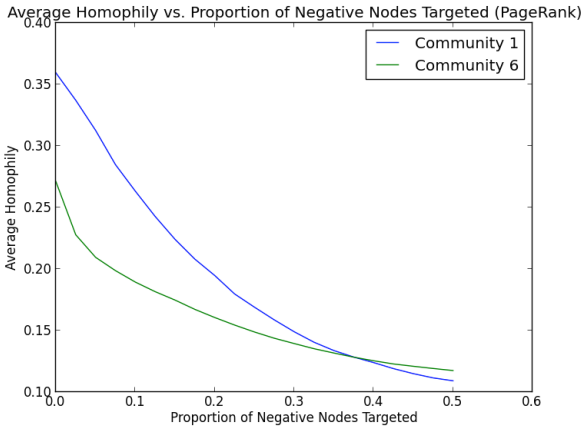
The above figure shows that, for both graphs, as the percentage of deleted nodes increases, the average homophily increases until it reaches the 0.45 mark. This is a sign that the network is centered around controversy as influential nodes are removed, communities actually become tighter. Thus, it would appear that the community is actually formed of several smaller neighborhoods of communities which harbor similar sentiments; these neighborhoods are connected via the most influential nodes, which appear to mainly be negative. In addition to having very influential negative nodes, the positive nodes in Community 6 are actually *less* influential than the neutral nodes (the graph plateaus, meaning after around 0.45 of nodes deleted, we begin deleting mostly neutral nodes). This explains why Community 6 has more positive nodes, but a negative sentiment overall.

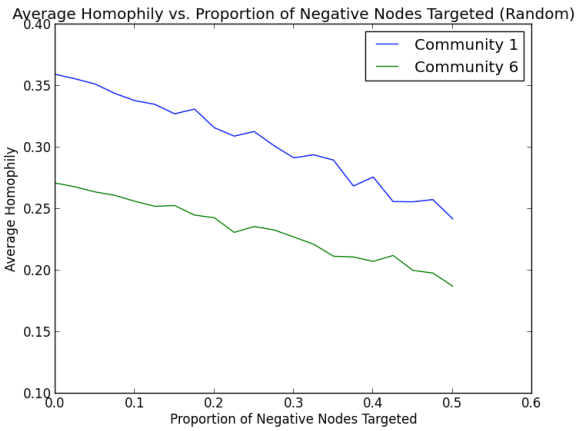
6.5 Random and Targeted Interventions on Negative Communities

As described in the discussion of our methodology, we target negative nodes for conversion using two policies: (1) targeting the negative nodes with the highest PageRank first, and (2) randomly, and graphed the average sentiment and homophily versus the size of the initial set of nodes targeted for conversion:









6.5a Sentiment

The curves for both communities look very similar under the same policy; the increase in sentiment for random interventions is roughly linear while it is logarithmic for PageRank-targeted interven-

tions. These are both expected results as random interventions would tend to increase the average sentiment by roughly the same amount every time, whereas the sentiment would increase quickly at first for the PageRank-targeted interventions as we choose the most "influential" nodes first.

Targeted interventions appear significantly more effective than random interventions; with targeted interventions, by initially converting only around 0.075 of the nodes with the highest PageRank we can successfully make the overall sentiment of Community 6 positive, compared to about 0.375 of randomly-chosen negative nodes. Similarly, we can successfully make Community 1 positive after initially converting 0.300 of all negative nodes (admittedly, this is already unrealistically high) whereas we need to randomly convert more than half the nodes to achieve the same effect.

It appears that PageRank-targeted interventions are effective at quarantining anti-vaccination sentiment for Community 6, which happens to be our largest community as well.

6.5b Homophily

In both graphs, the homophily measures decrease, which means that negative nodes have connections to each other. Alternatively, it is also possible that the converted nodes are connected to neutral nodes; our homophily metric currently calculates the homophily between two neutral nodes to be 0 (other values make less sense). It is unlikely that negative nodes have many connections to positive nodes; the homophily value would increase if this were the case, as the homophily value between a neutral node and a positive node is always greater than the homophily between a negative node and the same positive node.

Homophily decreases linearly for the random tests and power-law for the PageRank-targeted tests. These results are expected as randomly converting more nodes should decrease the homophily at approximately a constant rate while the PageRank-targeted tests should affect the homophily to a great extent even at lower probabilities, since we select nodes based on their influence.

Interestingly, the homophily curves of Community 1 and Community 6 cross each other on the PageRank graph when the proportion of initially-targeted nodes is around 0.375 but do not cross each other on the random graph (i.e. the average homophily of Community 1 is strictly greater). This means that the PageRank of negative nodes is likely to be power-law distributed.

References

[1] E. Eaton, R. Mansbach. "A Spin-Glass Model for Semi-Supervised Community Detection". Retrieved from: http://www.seas.upenn.edu/ eeaton/papers/Eaton2012SpinGlass.pdf.

[2] J. Leskovec, K.J. Lang, A. Dasgupta, M.W. Mahoney. "Statistical Properties of Community Structure in Large Social and Information Networks". Retrieved from: http://dl.acm.org/citation.cfm?id=1367591.

[3] S. Myers, C. Zhu, J. Leskovec. "Information Diffusion and External Influence in Networks". Retrieved from: http://cs.stanford.edu/people/jure/pubs/ext-kdd12.pdf

[4] S. Myers, J. Leskovec. "Clash of the Contagions: Cooperation and Competition in Information Diffusion". Retrieved from: http://cs.stanford.edu/people/jure/pubs/topicmix-icdm12.pdf Newman M E J and Girvan M, 2004 Phys. Rev. E 69 026113 Pons P and Latapy M, 2006 Journal of Graph Algorithms and Applications 10 191. Wu F and Huberman B A, 2004 Eur. Phys. J. B 38 331

Appendix: Community Statistics

Community	Avg. Sentiment (Degree)	Avg Sentiment (PageRank)	Homophily	Avg Rand Homophily (20 runs)	Std Dev	Node Count	Positive Nodes	Negative Nodes	Edge Count	Density
0	0.04938207398	0.06002066307	0.4112556082	0.8881428955	0.007033697967	8188	2154	1296	182796	22.32486566
1	-0.1343504567	-0.1275100697	0.3615884652	0.8903923237	0.01277426763	6254	701	1953	301335	48.18276303
2	0.1066601088	0.1222594744	0.370822599	0.8856958959	0.008314096305	6248	1991	693	71772	11.4871959
3	0.01880424095	0.05866375979	0.4261485206	0.8850820086	0.01134750356	4753	1336	804	72776	15.31159268
4	0.05644366264	0.06900298616	0.3897585771	0.8905115091	0.006779096728	40217	11439	5698	433006	10.76674043
5	0.06619357961	0.05898218403	0.4416971791	0.8758854167	0.06274195787	97	28	15	639	6.587628866
6	-0.03709637591	-0.02802360753	0.274309907	0.8893424182	0.005928237042	18356	3237	3026	759668	41.38526912
7	0.03155382662	0.04448116255	0.1814621273	0.8873886755	0.009405158736	8480	1909	778	182308	21.49858491
8	0.1217464092	0.1250403921	0.3923022367	0.8908581368	0.02472422684	515	151	40	4911	9.53592233
9	0.343099839	0.4057333442	0.7041635688	0.8983643123	0.03438780265	577	319	101	1035	1.793760832
10	0.08684834763	0.08509115721	0.4013641102	0.8865780231	0.04523227301	138	36	20	1345	9.746376812
11	0.0153349768	0.04975093673	0.3865829237	0.890751866	0.03031706693	290	80	38	3448	11.88965517
12	0.08588024771	0.1137629139	0.3026729035	0.8925570539	0.0375832592	246	79	23	1338	5.43902439
13	0.1031323877	0.09812141818	0.349726776	0.8938979964	0.05902559858	60	14	5	376	6.266666667
14	-0.45703125	-0.4286411546	0.6397058824	0.9132352941	0.0949146706	27	5	13	64	2.37037037
TOTAL AVG.	**0.03044010787**	**0.04711570404**	**0.4022374257**	**0.8905789217**	**0.03003392758**	**94446**	**23479**	**14503**	**2016817**	**224.5864172**