CS224W Final Report
Emergence of Global Status Hierarchy in Social Networks

Group 10: Yue Chen, Jia Ji, Yizheng Liao

December 10, 2012

1 Introduction

Social network analysis provides insights into a wide range of social computing applications. The fundamental element of social network is the relations between people. By analyzing the relations between people in a network, we can reveal interesting structures and useful insights about the network.

Relations between people in real world networks often exhibit two prominent features: (1) it is a mixture of positive (e.g. supporting) and negative (e.g. opposing) interactions; (2) it evolves through time. While many people have studied relations between people and network structure, most researches on social networks have focused on positive interactions between people. Only recently people have taken negative interactions into account. Moreover, few studies have been presented on the evolution of signed networks. In addition, many works are often limited to theoretical analysis under strong assumptions.

In this course project, we address the above problems by analyzing the evolution of real world signed network and building models for signed network evolution. Interesting insights are discussed by analyzing real world network and evaluating our models based on it.

In the first part of the project, we analyzed how status hierarchy evolves and converges to a true global status hierarchy in real world signed network. We designed algorithms to compute status hierarchy of a network. Given an evolving signed network, we assumed that the status hierarchy, which was computed in the final state, represented the true underlying global status hierarchy. We used normalized discounted cumulative gain score, which was widely used in information retrieval community, to analyze how the status hierarchy followed the global status hierarchy over time. Our analysis shows that the signed edges of Epinions network follow the global status hierarchy more and more.

In the second part of the project, we proposed different models of network evolution. The models are based on local status information (i.e. how the two nodes involved in the link formation process give and receive signed edges) as well as global status information (i.e. the status hierarchy of the network at the time of link formation, and the community's overall link creation behavior). We evaluated our models on real world network and the experiment results show that users create signed links based mostly on local information, rather than global information.

The rest of this paper is organized as follows: Section 2 introduces related work; Section 3 presents the evolution analysis algorithms and status prediction models; Section 4 investigates the experiment and simulation results for network evolution and status prediction. Section 5 highlights the key results and concludes the paper.

2 Prior Work

Relations between people in real world networks are often a mixture of positive (e.g. supporting) and negative (e.g. opposing) interactions. [1] is the first literature that discussed both trusting and distrusting in the social network. Although [1] mentioned the importance of the negative edge in signed network, it did not clearly address the reason why the unbalanced trial made sense in the real signed network. Jure et.al. [2] covered this part. Jure et.al. [2] investigated two theories of signed social network from social psychology - balance and status, and conducted first large-scale evaluations using three online dataset.

Jure et.al. [3] built on their work on signed networks and the theories of balance and status to predict edge signs and provided insight into some of the fundamental principles that drives the formation of signed links in networks.

Massa et.al. [4] concentrated on the controversial users who were judged by other users in very diverse way and compared the technique using Global Trust Matrix and Local Trust Matrix in answering questions such as Should I trust this user?. The results demonstrated that the local Trust Metric was able to significantly reduce the prediction error for controversial users, while retaining a good coverage.

Antal, et.al. [5] have proposed and analyzed two models of link dynamics - local triad dynamics (LTD) and constrained triad dynamics (CTD), which were essentially optimizing local, global consistency with respect to the balance theory. Marvel et.al. [6] treated a signed network as a graph with certain energy describing the consistency of the graph with respect to certain properties (e.g. balance theory or status theory), and the dynamics of the network were thus naturally viewed as a way for energy.

3 Algorithm

3.1 Status Hierarchy Computation

In this section we will discuss three algorithms we used to compute the status hierarchy: PageRank, Support Number, Hopfield. Each algorithm will give score and status for each node on each day based on node's information or global connection information. Then we compute the normalized discounted cumulative gain for each day by comparing the node scores on day k with those on last day.

3.1.1 PageRank

PageRank [7] is a link analysis algorithm that assigns a numerical weighting to each element of a graph, with the purpose of "measuring" its relative importance within the set. We can easily assign status of each node based on their PageRank score. PageRank is a probability distribution used to represent the likelihood that a random walk through the edges will arrive at a particular node. The equation of calculating PageRank is as follows:

$$r_j = \sum_{i \to j} \beta \frac{r_i}{d_i} + (1-\beta)\frac{1}{n} \tag{1}$$

where d_i denotes the out-degree of node i, r_i denotes the rank, and β denotes the probability of jumping to a random node.

3.1.2 Number of Supporter

The second method we used is the support number, i.e. how many nodes can reach the node n in the network by following forward positive edges or backward negative edges. Status of node is then ranked by the support number of each node. We define adjacency matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ to represent a signed network as follows: for directed graph $\mathbf{G}$, if there exists an edge from node i to node j, then $\mathbf{A}(i,j) = 1$. Otherwise, $\mathbf{A}(i,j) = 0$. Please notice that $\mathbf{A}(i,i) \equiv 0$. We are also interested in how the nodes are connected via some intermediate nodes, this can be computed by multiplying the adjacency matrix $\mathbf{A}$ by itself for several times.

However, the matrix multiplication is expensive, therefore, we decompose the graph into a directed acyclic graph (DAG) on its strongly connected components (SCCs). Then the support number computation can be decomposed into the computation of support number in a strongly connected graph and the computation of support number in a DAG. Hence, we can do it recursively.

To decompose the graph, in each iteration, we find the largest SCC in the graph, replace the nodes in the SCC with one node, and use a hash table to keep track of the original nodes and their links. Iteration stops when the largest SCC only has one node.

After decomposing the graph, we can compute the support number of each node n in following steps: (1) find the corresponding SCC$_k$ of the node, (2) find the set of SCCs, $\mathbf{K}$, that support SCC$_k$(3) The support number of node n is

$$\text{Support number of node } n = \sum_{i=1}^{|\mathbf{K}|} |\text{SCC}_i| - 1 \quad (2)$$

3.1.3 Hopfield

The idea of Hopfield is to use stochastic gradient ascent to find a rank (status hierarchy) that is (locally) optimal for status consistency[8]. We say that an edge is consistent with respect to status and a given ranking. A positive edge points to a node with higher rank and a negative edge points to a node with lower rank. Given a graph and node ranking, we can then define the status consistency of the ranking with respect to the graph as the percentage of edges in the graph that is consistent with respect to status and the ranking.

In order to compute a (locally) optimal ranking and score using Hopfield, we start with an random node ranking, then we randomly flip the ranking of two nodes if it will result in a higher status consistency, this process continues until it reaches a local maximum.

3.2 Normalized Discounted Cumulative Gain

Normalized Discounted Cumulative Gain (NDCG) is a measurement widely used in information retrieval community to evaluate the relavance of a list of result returned by a search engine [9]. We borrowed the idea of NDCG in our project to compare the similarity of two status hierarchies, by treating a status hierarchy as a ranked list of results. Discounted Cumulative Gain (DCG) measures the quality of the results in a ranked list, where items in that list are graded in some way. DCG accumulated at a particular rank position p is defined as:

$$\text{DCG}_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2 i} \quad (3)$$

where $rel_i = \left(1 - \frac{|r_i - \hat{r}_i|}{N}\right)^{\alpha}$, r_i is the training rank, $\hat{r}_i$ is the estimated rank, and N is the number of samples. α is a control variable.

Normalized DCG (NDCG) is a way to calculate this measure across many independent queries, the result lists for which might all be of different length. NDCG is computed as:

$$\text{NDCG}_p = \frac{(DCG)_p}{IDCG_p} \quad (4)$$

where IDCG, ideal DCG is the DCG value of the ideal ordering ranked list.

3.3 Network Evolution Prediction

To investigate the underlying mechanism of signed link formation, we proposed five models to predict the evolution of the signed network. They are:

- Model 1: Global sign distribution model
- Model 2: Destination status model
- Model 3: Source status model
- Model 4: Destination and source status model
- Model 5: Global status model

For these five models, we can group them into three classes: global sign distribution based model, local status based model, and global status based model. We used the proposed models to predict the sign of all the new created edge of day i based on the information of previous $i-1$ days.

3.3.1 Global Sign Distribution Based Model

In the global sign distribution based model, we use global sign distribution to predict sign of edges. For given signed network G, we first traversed all the edges in G and counted the number of positive sign edge. Then we can calculate the probability of positive sign edge by dividing the total number of positive sign edges by the total number of edges. Then we predict the signs of new added edges using an independent and identical (i.i.d) random variable from a Bernoulli Distribution with success probability p equals to the probability of positive sign edges, i.e. sign $\sim$ Bern (p).

3.3.2 Local Status Based Model

In the node information based model, we use destination node and/or source node information to predict the sign of edges. Here, we have three submodels: destination status based model, source status based model, and destination and source status based model.

For each node in the signed network, n_{in+} denotes the number of edge with positive sign point to it; n_{in-} denotes the number of edge with negative sign point to it; n_{out+} denotes the number of edge with positive sign from it; n_{out-} denotes the number of edge with negative sign from it. Now for the destination node, we define the destination node information as follows:

$$I_{dst} = n_{in+} - n_{in-} - n_{out+} + n_{out-}. \tag{5}$$

For the source node, we define its information as follows:

$$I_{src} = n_{out+} - n_{out-} - n_{in+} + n_{in-}. \tag{6}$$

For model 2, let $c = I_{dst}$; for model 3, let $c = I_{src}$; for model 4, let $c = I_{dst} + I_{src}$. Now, for the local status based model, the sign decision rule is:

$$\text{sign} = \begin{cases} 1 & c > 0 \\ -1 & c < 0 \\ X & c = 0 \end{cases} \tag{7}$$

where $X \sim \text{Bern}(p)$ based on the global sign distribution.

3.3.3 Global Status Based Model

In the global status based model, we use the PageRank of the source node and destination node to predict the signs of edges.Given the signed network G, we can calculate the pagerank of each node using the algorithm described in Section 3.1. Then, we can predict the sign of edge A to B as follows:

$$\text{sign} = \begin{cases} 1 & \text{PR(A)} > \text{PR(B)} \\ -1 & \text{PR(A)} < \text{PR(B)} \\ X & \text{PR(A)} = \text{PR(B)} \end{cases} \tag{8}$$

where $X \sim \text{Bern}(p)$ based on the global sign distribution.

4 Experiments

4.1 Dataset

We use a dataset of Epinions.com in this project. Epinions.com is a who-trust-whom online social network, members of the site write consumer reviews and decide whether to trust each other. The trust relationship on Epinions.com is represented signed edges, and the entire social network is represented as a signed network. The dataset consists of 131,585 nodes (users) and 838,985 signed edges (trust/distrust relationship) over 940 days. We discretize the time into 94 timestamps and study how status hierarchy evolves over the 94 timestamps.

4.2 Evolution of Status Hierarchy

In the first experiment, we analyze how status hierarchy evolves over time. Given an evolving signed network, we assumed that the status hierarchy in the final state (timestamp) represents the true underlying global status hierarchy. We analyze how the status hierarchy followed the global status hierarchy over time, by comparing the status hierarchy at a timestamp to the status hierarchy at the final timestamp. We use NDCG described in Section 3 to compare two status hierarchies.

Three algorithms (PageRank, Number of Supporter and Hopfield) were used to compute status hierarchy at each timestamp. In order to see how well the status hierarchy evolves toward the true status hierarchy, we use two randomly generated directed graph as baselines. The first random network, Gnm Network, is generated to have the same number of nodes and edges with the Epinion Network at each timestamp, every possible edge has the same probability to be generated. The second random network, Random Sign Network, is generated to have the same number of nodes and the same edges with the Epinion Network at each timestamp, and also it has the same edge sign distribution with the Epinion Network.

Below, we shows the experiments results using PageRank, Number of Supporter and Hopfield respectively.

4.2.1 PageRank

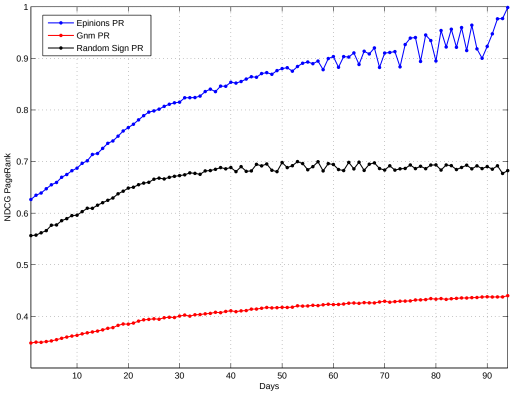

Figure 1: NDCG of PageRank

Fig. 1 shows the evolution of the NDCG scores using PageRank. From this figure, we can find that the NDCG score of the Epinion Network grows with time from around 0.62 to 1. The NDCG score increases quickly at the first 30 time periods, slow down the growth gradually and has more fluctuation after 70 time periods, and converges to 1. For the Random Sign Network, the NDCG score starts at around 0.55 and also grows fast at first 30 time periods and converges to around 0.7. For the Gnm network, the NDCG score increases from around 0.35 to 0.44. It is very low compared with the Epinion Network. We can find that the NDCG score of the Epinions Network is much better than the Gnm Network and Random Sign Network. From this figure, we can conclude that we can use the final graph hierarchy status as the ground truth status hierarchy and the network converges to the ground truth status.

4.2.2 Number of Supporter

Fig. 2 shows the evolution of NDCG scores using Number of Supporter. Comparing Fig. 1 and Fig. 2, we can find that the NDCG scores also grows with time, the scores of Epinions Network are much better than the Gnm Network and Random Sign Network. But the NDCG scores fluctuate more and the scores are lower compared with the scores using PageRank. So this also proves the conclusion we got in Section 4.2.1, but PageRank is a better way to construct the status hierarchy than Number of Supporter.

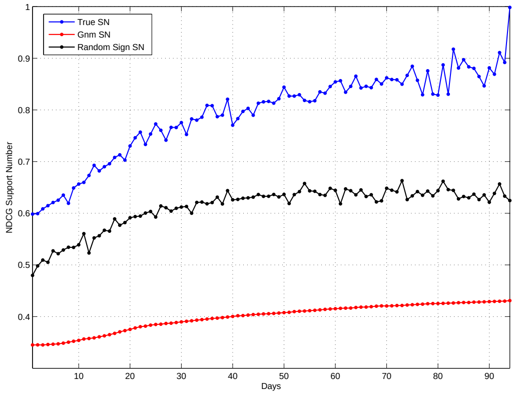

Figure 2: NDCG of Support Number

4.2.3 Hopfield

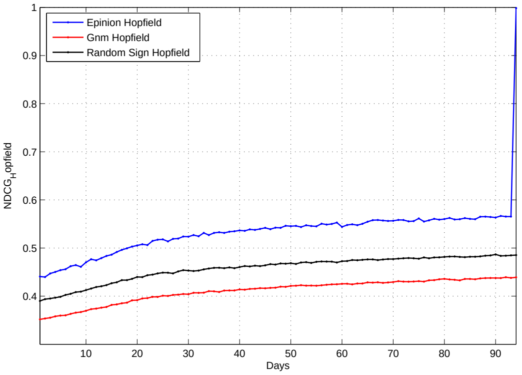

Figure 3: NDCG of Hopfield

Fig. 3 shows the evolution of NDCG scores using Hopfield, we can find that the scores are very low (the best score achieved is only 0.57 except the final status that is 1 for sure. So Hopfield is not a very good way of calculating the status hierarchy. This is probably because that the Hopfield algorithm is a stochastic gradient ascend method, and the status consistency is non-convex, so the algorithm can be easily converged into local minimum. Also the way Hopfield flip edges optimize for the consistency with respect to status, which doesn't optimize for the similarity (NDCG score) with the true status hierarchy (e.g. flipping the edge between two nodes in the top of status hierarchy has similar gain as flipping the edge between two nodes in the bottom of status hierarchy, but they will result in dramatically different NDCG scores).

4.2.4 Top 10 Users PageRank

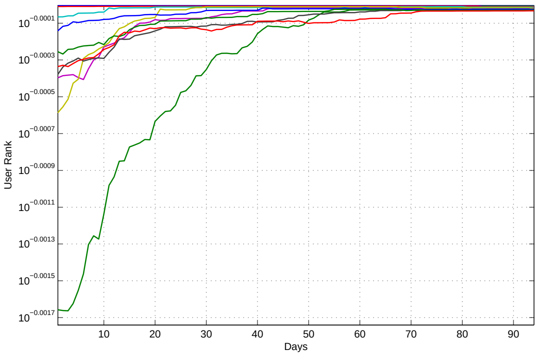

Figure 4: The Rank Evolution of Top 10 Users over Time

Fig. 4 shows the evolution of PageRank for the top 10 users who have the highest PageRank at the final status. It shows that the high PageRank users have got very high PageRank at the early stage of the formation of the signed network. And the PageRank of them change very little over time.

4.3 Model of Network Evolution

To evaluate the models of network evolution we proposed in Section 3.3, we simulated the evolution of the graph using the models. For each model, given the Epinions Network graph on the first time period G, we predict the signs of edges added to the network on time period $t+1$ using the information from simulated graph we got till time period t. Then, we add the new edges to the simulated graph.

Fig. 5 shows the cumulative accuracy of the prediction models. We can get some interesting information from this figure. Firstly, the model 2, 3 and 4 works very well in the prediction and among them Model 4 did best. The model 1 and model 5 are worse than the model 2, 3 and 4, and model 5 is the worst. So using local node information can predict the signs best compared with using global information or status. From this, we can find that people decide their trust or mistrust toward other people mainly depending on the local information of himself and the target people.

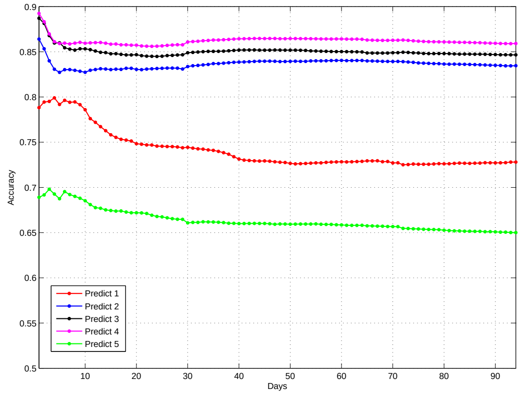

Figure 5: Cumulative Accuracy of Prediction

Furthermore, we find that the cumulative accuracy starts relatively high, and drops in the first several time periods, then remains at some certain levels. Since the prediction of each day is based on the predicted graph we got on the previous day, so once the simulated graph has make some difference with the real Epinions graph, it will make further impact on the prediction of the sign of other new edges. So the precision starts relatively high because at the start, the given graph is exactly the same graph as the Epinions real graph, so the information is accurate. But as more and more predicted edges added to the predicted graph, the information of each node has got some changes, so the accuracy drops. This drop also implies that while nodes with high true status established their status initially, the status of most nodes are random compared to their final status. The final stable accuracy shown in the figure implies that the status of the predicted graph and real graph has got a relatively stable status, which means the status of both graphs have converged to their final status.

Fig. 6 shows the NDCG scores of the simulated graph compared with the Epinions Network final graph using PageRank as described in 3.1. From this figure, model 2, 3 and 4 also did better than

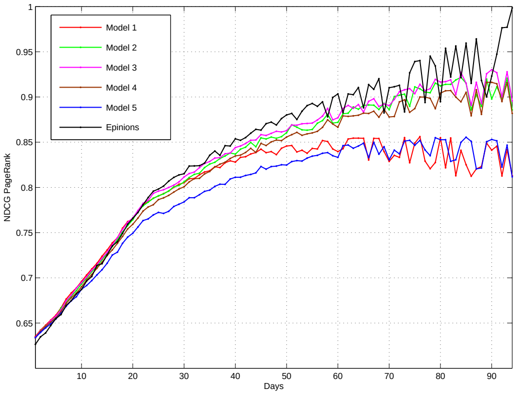

Figure 6: NDCG of Predicted Network PageRank

model 1 and model 5 in the NDCG scores. Model 2, 3, and 4 can achieve the NDCG score of around 0.9. And model 1 and 5 can achieve NDCG scores of around 0.83. The local status based model have very similar growth rate with the real Epinions network, and outperforms the global sign distribution based model, this further proves that people make their decision mainly depending on the local status information.

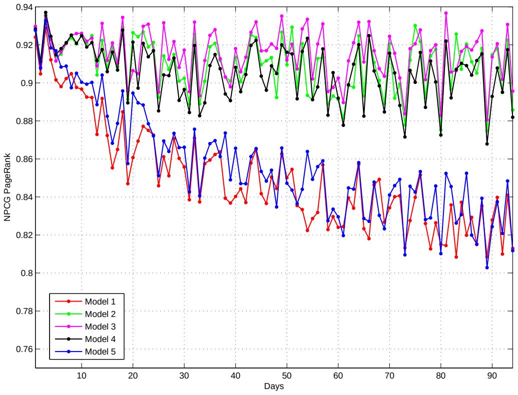

Figure 7: NDCG of Predicted Network PageRank on Each Day

Fig. 7 shows the NDCG score of the simulated graph compared with the Epinions Network graph on the same day. Here, we compared the similarity of predicted graph with the Epinions network graph on that same day. We can find that the model 2, 3, and 4 are very stable at around 0.9 and the model 1 and 5 NDCG score keeps dropping, but the dropping rate decreases over time. This further improves that the local status model can describe mechanism of the network evolution since it has very similar evolution process with the real Epinions network.

5 Conclusion

In this project, we analyzed how the status hierarchy evolves and converges to a true global status hierarchy in real world signed network. We designed algorithms to compute status hierarchy of a network. From our experiments results, we found that PageRank and Support Number are good indicators of network hierarchy. They also show that the status hierarchy converges to the true status. The evolution of PageRank for the top 10 users implies that the nodes with high true status established their status initially.

Then, we proposed different models of network evolution and used the models to simulate the evolution of the signed network. From the simulation results, we found that while nodes with high true status established their status initially, the status of most nodes are random compared to their final status. We also found that local status based prediction models have high prediction accuracy, which is about 88% at the end of evolution. Network evolution predicted by these models are close to the real network evolution. This implies that users create signed links based mostly on local information, rather than global information in social networks.

There are many future research areas related to our work. For example, it will be very interesting to discover if there exist user cluster or user community in the signed network and how it evolves. In addition, it will be interesting to explore how structural properties relate to the link formation process. We believe that our work will provide a foundation for many future works on the status hierarchy evolution of social networks.

References

[1] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, “Propagation of trust and distrust,” in *Proceedings of the 13th international conference on World Wide Web*, pp. 403–412, ACM, 2004.

[2] J. Leskovec, D. Huttenlocher, and J. Kleinberg, “Signed networks in social media,” in *Proceedings of the 28th international conference on Human factors in computing systems*, pp. 1361–1370, ACM, 2010.

[3] J. Leskovec, D. Huttenlocher, and J. Kleinberg, “Predicting positive and negative links in online social networks,” in *Proceedings of the 19th international conference on World wide web*, pp. 641–650, ACM, 2010.

[4] P. Massa and P. Avesani, “Controversial users demand local trust metrics: An experimental study on epinions. com community,” in *Proceedings of the National Conference on Artificial Intelligence*, vol. 20, p. 121, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[5] T. Antal, P. Krapivsky, and S. Redner, “Social balance on networks: The dynamics of friendship and enmity,” *Physica D: Nonlinear Phenomena*, vol. 224, no. 1, pp. 130–136, 2006.

[6] S. Marvel, S. Strogatz, and J. Kleinberg, “Energy landscape of social balance,” *Physical review letters*, vol. 103, no. 19, p. 198701, 2009.

[7] L. Page, S. Brin, R. Motwani, and T. Winograd, “The pagerank citation ranking: bringing order to the web.,” 1999.

[8] J. Hopfield, “Neural networks and physical systems with emergent collective computational abilities,” *Proceedings of the national academy of sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[9] K. Järvelin and J. Kekäläinen, “Cumulated gain-based evaluation of ir techniques,” *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.