The Impact of Structural Changes on Predictions of Diffusion in Networks

Mayank Lahiri, Arun S. Maiya
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607, USA
{mlahiri, amaiya}@cs.uic.edu

Rajmonda Sulo
Department of Mathematics,
Statistics and Computer Science
University of Illinois at Chicago
Chicago, IL 60607, USA
rsulo1@uic.edu

Habiba, Tanya Y. Berger Wolf
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607, USA
{hhabiba, tanyabw}@cs.uic.edu

Abstract

In a typical realistic scenario, there exist some past data about the structure of the network which are analyzed with respect to some possibly future spreading process, such as behavior, opinion, disease, or computer malware. How sensitive are the predictions made about spread and spreaders to the changes in the structure of the network? We investigate the answer to this question by considering seven real-world networks that have an explicit timeline and span a range of social interactions, from celebrity sightings to animal movement. For each dataset, we examine the results of the spread analysis with respect to the changes that occur in the network as the time unfolds as well as introduced random perturbations. We show that neither the estimates of the extent of spread for each individual nor the set of the top spreaders are robust to structural changes. Thus, analysis performed on historic data may not be relevant by the time it is acted upon.

1. Introduction

Prediction of the course and extent of processes spreading in social networks and identification of the top spreading individuals have become important issues in many contexts, from epidemiology to viral marketing. In a typical realistic scenario, there exists some past data about the structure of the network which is analyzed with respect to the future spread of some process, such as behavior, opinion, disease, or computer malware. The important tasks are (1) estimating the possible number of affected individuals once the process starts, (2) predicting who those individuals may be, (3) identifying the most effective spread initiators, and (4) identifying individuals that can effectively block the spread of the process. However, by the time the outcomes of such analysis are acted upon, such as by selecting marketing targets or vaccination candidates, time has elapsed and the network structure may have changed significantly from what was used for the initial analysis. The effectiveness of the marketing scheme [23] or epidemiological response may be sabotaged if analysis results are sensitive to such structural changes.

In this paper, we focus mainly on the tasks of estimating the extent of spread and identifying the top spreaders. These are the individuals that, when used as the start of a spread, affect the largest proportion of the population. We ask how sensitive the predictions made about spread and spreaders are to changes in the structure of the network. To answer this overall question, we formulate three specific questions:

1. **How much does the relative spreading ability of individuals change?** Most algorithms for estimating the extent of spread and for identifying the top spreaders fundamentally rely on estimates of the spreading ability of each individual. Thus, it is important to know how reliable those estimates are both in terms of actual numbers and in the ranking they impose on individuals.

2. **How much does the identity of the top spreaders change?** While the first question asks whether our predictions hold for all the individuals in the population,

this question focuses only on the top spreaders. The set of top spreaders may be more or less robust than the rest of the individuals, yet it is typically more critical to the impeding action.

3. **How does the spreading ability of the top spreaders from the past compare with that of the top spreaders after the change?** While the identity of the top spreaders may change as the network changes, the previous set of top spreaders may still perform well. Although it may not be the best set of top spreaders in the new network, we ask whether it is good enough.

We investigate the answers to these questions by considering seven real-world networks (Section 5) that have an explicit timeline and span a range of social interactions, from celebrity sightings to animal movement. For each dataset, we examine the results of the spread analysis with respect to the changes that occur in the network as time unfolds, as well as introduced random perturbations (Section 4). We show that neither the estimates of the extent of spread for each individual nor the set of the top spreaders are robust to structural changes (Section 6). Thus, analyses performed on historic data may not be relevant by the time they they are acted upon if the network changes substantially in the meantime.

2 Related Work

Many phenomena such as diseases, opinions, information, fads, and behavior have been modeled as diffusion processes in a social network, and have been studied in a number of domains including epidemiology [2, 9, 12, 16, 19, 29, 31, 32], diffusion of technological innovations and adoption of new products [5, 6, 10, 15, 22, 24], phenomena such as voting, strikes, rumors [17, 28, 36] and numerous others. Several previous results have also addressed the problem of identifying influential individuals affecting the spread of a phenomenon in a network [3, 4, 10, 12, 19, 22, 25].

The problem of identifying the *set* of top k spreaders in social networks has been shown to be NP-complete under various formulations [4, 18, 22] but allows a simple greedy $(1 - 1/e)$-approximation. Later results by Mossel and Roch [30] show that the general case of influence maximization is NP-hard with the approximation guarantee $(1-1/e-\varepsilon)$. The algorithms for picking the k-best spreaders in a network rely on first approximating the spreading ability of each individual, usually through stochastic simulations. For large networks, this can be very computationally intensive. Furthermore, strong inapproximability results for several other variants of influence maximization in social networks have been shown by Chen [7].

Several recent studies explore various network properties as proxies for the spreading ability of a node *e.g.* [21, 27]), yet those results are not generally conclusive. For all these approaches, the sample data used for analysis does not take into consideration possible future network changes. In the next section we formalize the distinction between networks that change in time (dynamic networks) and their aggregate or static view (Section 3.1). We also state the mathematical models of spread in networks in Section 3.2.

3 Definitions

3.1 Static and Dynamic Networks

A social network is defined as a graph $G = (V, E)$ where the nodes V correspond to a set of unique individuals and the edges $E \subseteq V \times V$ represent interactions or relationships between these individuals. We differentiate between two types of social networks – those that change and evolve over time, and those that are inherently unchanging. Examples of the former include human contact networks, as the patterns of interaction between people are likely to change over the time. The top-level Internet router topology is an example of a network that does not change, or changes very little with time.

For those networks that do vary with time, a *dynamic network* is a convenient representation for explicitly modeling temporal changes. While conventional methodology involves observing interactions for a period of time and representing them as a single graph, a dynamic network is a time-series of graphs where each graph represents interactions over a small time period.

Definition 1. A *dynamic network* is a time-series of labeled graphs $\mathcal{G} = \langle G_1, ..., G_T\rangle$, where $G_t = (V_t, E_t)$ is the graph of interactions taking place at timestep t. $V_t \subseteq V$ is the set of individuals observed at timestep t, and an edge (v_1, v_2) exists in E_t if v_1 and v_2 were observed interacting in that time period.

The question of how much actual time should be quantized into a 'timestep' is beyond the focus of this paper. However, we note that many types of social systems have natural time quantizations such as hours or days. Figure 1 shows a dynamic network of interactions between four individuals.

One can aggregate a range of timesteps in a dynamic network into a single *static* graph, or an *aggregate network*. This is done by accumulating vertices and edges present in a given range of timesteps. Note that aggregating the entire range of timesteps results in a traditional social network, i.e. a single graph of interactions without any temporal information.

Definition 2. Given a dynamic network $\mathcal{G}$, we build a *static*, or an *aggregate network* $G_{[i,j]}$ from a range $[i, j]$ timesteps of $\mathcal{G}$ by accumulating vertices and edges in that range:

$$V(G_{[i,j]}) = \bigcup_{i \le t \le j} V_t \qquad E(G_{[i,j]}) = \bigcup_{i \le t \le j} E_t$$

If $i = 1$ and $j = T$, then the resultant aggregate network $G_{[1,T]}$ is equivalent to the traditional social network G.

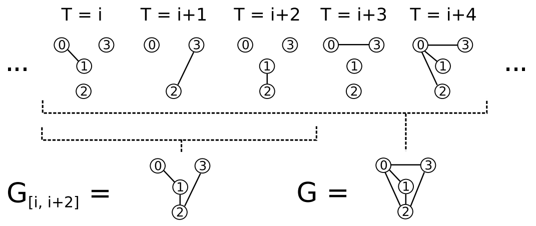

Figure 1. A dynamic network (top), an aggregated network (bottom left), a traditional social network (bottom right).

In traditional approaches, network analysis is performed on the aggregate network rather than the original explicitly dynamic network. One might immediately recognize a problem here: paths in an aggregate network might not correspond to valid propagation paths in the original dynamic network. Since the propagation of a rumor or virus must proceed along a sequence of edges that are increasing in time, and since an aggregate network has no temporal information, modeling the spread of a process without considering time can lead to grossly inaccurate results [18].

3.2 Network Diffusion

A process spreading in a network can be described formally using many models of transmission. For this paper we use a model that has been extensively studied in the context of social networks and viral marketing, the *Linear Threshold* model [17]. While our analysis and conclusions are applicable to a related simpler *Independent Cascade* model [9, 10, 26, 31, 32], more commonly used in epidemiology, we omit it in this paper for focus and brevity.

Models of collective behavior are developed for situations where individuals choose between alternatives based on how many other individuals connected to them choose those alternatives. The key concept is that of a “threshold” which is the cumulative number of neighbors of an individual that must make a decision before the individual does so [17]. Such models are most appropriate for product adoption or behavior propagation.

The Linear Threshold diffusion model describes the spread over two sets of individuals, active and inactive. Each inactive individual has a certain susceptibility to become active, which is denoted by the individual’s threshold. Each active individual has a certain weight of influence over each of its inactive neighbors. An individual becomes active if the accumulated weight of all its active neighbors become larger than the individual’s susceptibility threshold.

More formally, the linear threshold model is defined by two parameters. For each individual v, a threshold $\theta_v \le 1$ indicates the latent tendency of this individual to be activated. For each edge $(u, v) \in E$ the weight $b_{u,v}$ is the influence of the individual u on v, that is, u’s ability to activate v. For each v, $\sum b_{u,v} \le 1$. In a dynamic network, the weight b_{u_t,v_t} may be time-dependent.

The spreading process described by linear threshold model starts with a given set of thresholds θ_v assigned to each individual. The initial set of active individuals is A_0. The process unfolds in discrete timesteps. In a dynamic network, we assume for simplicity that the timesteps are synchronized with the timesteps of the network itself, $1 \ldots T$. At each step t, each inactive individual v is influenced by the set of its active neighbors. The inactive individual v becomes active at timestep $t + 1$ if $\sum b_{u,v} \ge \theta_v$. If $\sum b_{u,v} < \theta_v$ then v remains inactive and at every subsequent timestep, a new attempt is made to activate it by the set of its neighbors active at time $t + i$. Each attempt is independent of any previously made attempts. The outcomes of the process is the set of individuals A_f active after T timesteps or until no more activations are possible, and the size of that set, $|A_f|$. We denote by $\sigma(A_0) = A_f$ the correspondence between the initial set A_0 and the resulting set of active individuals A_f. We call this process *static linear threshold spread* when it unfolds over an aggregate, or static, network.

The spreading process in the dynamic network graph is different from the aggregate network, where each active individual attempts to activate each of its inactive neighbors whenever it interacts with the inactive neighbor. We consider two variants of linear threshold model for dynamic networks, *memoryless* and *with memory*. In the memoryless model an individual v becomes active at time $t + 1$ if the total weight of its active neighbors *at time* t exceeds its threshold: $\sum b_{u_t,v_t} \ge \theta_v$. This variant models the process of adoption of impulsive behavior, influenced by peers present at the moment. In the linear threshold model with memory an individual v becomes active at time $t + 1$ if the total *cumulative* weight of its active neighbors *up until time* t exceeds its threshold: $\sum_{i=0}^{t} b_{u_i,v_i} \ge \theta_v$.

4 Methodology

To answer the three questions posed in Section 1, we recall a typical scenario for network analysis: the network is observed for some time, then is analyzed as one aggregated social network. The results of the analysis are then deployed in the network, which has changed in the meantime. In order to determine the effect of the changes on the results of the historical analysis, we use the following overall experimental template:

1. Consider (part of) the dynamic network as an aggregate "historical data" network G_h.

2. Perform analysis on G_h: estimate the spreading ability of each individual and identify the top spreaders.

3. Extract the changed network G_f. We do this in two ways, both by considering the actual future segment of the network and by randomly perturbing the network to introduce changes.

4. Perform analysis on G_f: estimate the spreading ability of each individual and identify the top spreaders. Compare the results of the analysis on G_h with the results on G_f. This will answer questions 1 and 2: how much does the relative spreading ability of each individual and the identify of the top spreaders changes.

5. Recall that the third question was how well do the top spreaders from the past perform in the future relative to the best spreaders of the future. To answer this question, we simulate the spread in G_f (changed network) starting both from the top spreaders of G_f and the top spreader of G_h and compare their performance.

As we have pointed out, there are at least two ways to consider the changes that may happen in the structure of the network as the network evolves with time. First, we may look at the actual dynamic network and aggregate a portion of it into an initial "historical" segment used for analysis. Subsequent segments are designated "future" data and used to validate the results of the analysis. The changes in the structure of the network then are the actual changes that are recorded in the data. This is the approach of *temporal cross-validation*, which is an adaptation of the well-known statistical technique.

Temporal cross-validation involves dividing the timeline of the dynamic network into several segments and aggregating each segment into a single graph. Any analysis technique performed on the graph of one segment should then produce similar results in another segment, given that it is the same underlying network (and presumably the same dynamics) being modeled. If this is not the case, then we can conclude that either the analysis technique is not robust, or that the underlying dynamics of the network are changing. In either case, the particular analysis technique is then unlikely to produce actionable results. For this study, for temporal cross-validation we divide each dynamic network into five segments of equal duration.

While the temporal cross-validation approach examines the robustness of the analysis with respect to actual recorded network changes, these changes may not be representative of the changes that may, *in principle,* happen in the network. Thus, for the second way to introduce changes into the structure of the network, we take the aggregate network and randomly remove and add edges to introduce possible perturbations and provide the answer to the *expected* robustness of the analysis.

4.1. Experimental Setup

We initiate the experiments with the "historical" data network G_h. This is the aggregate network of either a segment of or the entire dynamic network. That is, $G_h = (V_h, E_h)$, where V_h are the nodes present in timesteps $i \leq t \leq j$, $E_h = \bigcup_{t=i}^{j} E_t$. Here i and j are either the first and the last timestep of a particular network segment or $i = 0$ and $j = T$.

We define an *objective function* for each vertex v, denoted $spread(v)$, which is the proportion of the population that v eventually activates if it starts as the only active node in the network. This is determined, as in earlier approaches [22], by Monte Carlo simulations. We simulate the linear threshold spreading process on the network starting with v as the only active node, for each node $v \in V$. The threshold values Θ_v are chosen randomly for each iteration, and all three variants of the linear threshold spreading process are simulated – aggregate spread on G_h, and dynamic spread with and without memory on the underlying dynamic network $\langle G_i, ..., G_j \rangle$.

In all cases, we simulate the spreading process for $j - i + 1$ timesteps. We used 500 iterations of each spread simulation, which was sufficient to produce consistent results. After simulating the spread from each individual, in each iteration, we note the number of activated individuals $\sigma(v) = |A_f|$. The overall spreading capacity of v is then the average over all iterations of the size of the active set proportionally to V_h: $spread(v) = \frac{1}{500} \sum \sigma(v)/|V_h|$.

We then rank the individuals in the order of their spreading ability: $spread(v_1) \geq spread(v_2) \geq \ldots \geq spread(v_{|v_h|})$. We call the first k individuals in this order the "top k spreaders". While as a set, this may not be the best *set* of k individuals from which to start a spreading process, individually they are the top k performers. We investigate whether they remain in the top k as the network changes.

As mentioned earlier, we obtain the changed network of "future" data G_f in two ways. In the temporal cross-validation setting this is one of the segments that follows the segment of G_h. That is, $G_f = (V_f, E_f)$ is the aggregate network of a latter segment. For random perturbations, we remove a fraction p of existing edges uniformly at random and add the same number of edges that were not in the network, preserving the overall number of edges. We use the range of $p = \{0.05, 0.1, 0.3\}$, that is, changing $5, 10$ and 30 percent of the edges. Note, that if the density of a network is d, one cannot change more than $1 - d$ fraction of the edges in this scheme.

To answer the first question about the change in the relative spreading ability of individuals, we measure the correlation between the orderings imposed on the individuals by their spreading ability. That is, given the ordering imposed by $spread(v)$ function in G_h and G_f, we measure the difference in the orderings using Spearman's rank correlation coefficient [34]. For the second question, we measure the difference in the identity of the top k spreaders for $k = \{5, 10\}$ by measuring the Jaccard similarity [20] of those sets in networks G_h and G_f.

Finally, we answer the last question by taking the top k spreaders from G_h and using them as the set of initially active individuals in G_f. We denote the average proportion of activated individuals as APX. We compare that number to the same process repeated with the initial set being the top k spreaders from G_f itself. We denote the average proportion of activated individuals in the latter case by OPT. We then measure the performance of the historical top spreaders in the changed network as the fraction APX/OPT.

We perform our analysis across different datasets representing a wide range of types of interactions. A summary of results obtained from our analysis is presented in Section 6.

5 Datasets

For our experiments, we used real dynamic networks spanning the range of interactions from animal behavior to celebrity sightings.

5.1. Animal Social Networks

Grevy's Zebra. The Grevy's dataset consists of social interactions among Grevy's zebra (*Equus grevyi*) recorded by biologists over the period of June through August of 2002 in the Laikipia region of Kenya [35]. Predetermined census loops were driven approximately twice per week and individual zebra were identified by unique stripe patterns. Upon a sighting, the zebra's GPS location was taken. In the resulting dynamic network, each node represents an individual zebra and two animals are interacting (i.e. an edge exists between the nodes) if their GPS locations are in close proximity. The dataset consists of 28 zebra.

Plains Zebra. Plains zebra (*Equus burchelli*) are another species of zebra. The data were collected in a similar fashion to that of the Grevy's dataset. The data were collected through visual scans (approximately once per day) over a period of several months [13]. Each entity is a Plains zebra and the interactions represent spatial proximity as determined by ecologists based on GPS locations. It should be noted that this similarity between the Plains Zebra dataset and the Grevy's Zebra dataset should *not* be taken to mean that the social interaction patterns will also be the same. There is evidence to indicate that different species of zebra can exhibit very different interaction patterns [35]. The Plains-1 dataset represents data from observations of 282 individuals from 12th July 2003 to 19th September 2006. The Plains-2 dataset represents observations of a different population of 313 individuals from 5th January 2004 to 3rd July 2007.

5.2. Mobile P2P

A number of different datasets of human and group interactions recorded through wireless hand-held devices have been made available. We use two such datasets in this experiment.

MIT Reality Mining. The MIT Reality Mining dataset consists of social interactions among 100 students and faculty over a nine month period at Massachusetts Institute of Technology [11]. Interactions were inferred from recorded Bluetooth connections between Nokia 6600 smartphones distributed to the participants. Our processed dynamic network consists of 96 vertices. The quantization was chosen as 4 hours [8].

Haggle Infocomm. The Haggle Infocomm dataset consists of social interactions among attendees at an IEEE Infocomm conference in the Grand Hyatt Miami [33]. There were 41 participants and the duration of the conference was 4 days. The time quantization period was 10 minutes.

5.3. Enron Email Network

The Enron e-mail corpus is a publicly available database of e-mails sent by and to employees of the now defunct Enron corporation [1]. The corpus was made available in 2003 by the Federal Energy Regulatory Commission during their investigation of the company. We use a cleaner version of the original dataset with fewer integrity issues[1]. Timestamps, senders and lists of recipients were extracted from message headers for each e-mail on file. We chose a day as the quantization timestep, with a directed interaction present if at least one e-mail was sent between two individuals on a particular day.

[1] Available at http://www.cs.cmu.edu/∼enron/

5.4. IMDB Photo Network

The Internet Movie Database (IMDB)[2] maintains a large archive of tagged and dated photographs of individuals associated with the production of commercial entertainment, including actors, directors and musicians. One might reasonably assert that people tagged on a popular online movie information repository are 'recognizable' to the general public, and that a degree of social association exists between people photographed together. Thus, similar to the methodology of the Plains Zebra sightings, we collected metadata on 45,477 photos with two or more people, which collectively represents a partial structure of the social network of people associated with the entertainment industry. The quantization period was one day.

Table 1 summarizes the basic statistics of the datasets.

Dataset	Vertices	Timesteps	Density
Grevy's	28	44	0.304
Plains-1	282	1166	0.788471
Plains-2	313	1,276	0.654358
Reality Mining	100	2,940	0.681579
Haggle	41	576	0.967073
Enron	82,614	2,588	0.000465963
IMDB	15,011	13,967	0.00042449

Table 1. Dataset characteristics. Density is the edge density of the aggregate network

6 Results

We now describe the results of the experiments and analysis that address each of the three questions, in turn, posed in Section 1. In all the figures, the datasets are shown in the order of increasing network density.

6.1 The change in the relative spreading ability of individuals

Recall that to answer the question of how valid the predictions about the spreading capacity of each individual in the network are as the network changes over time, we compare the rankings of the individual's spreading capacity. We calculated the Spearman correlation coefficient between those rankings in the original and changed networks. Figure 2 shows the correlations of the ranking by static linear threshold spread in the aggregate network of a given segment versus the rankings by the static and the two dynamic linear threshold spread within the same and all future segments of a dynamic network. For example, the bottom row of the plots shows the static spread ranking in the first segment versus all the rankings in each of the five segments, while the top row shows the ranking of the fourth segment versus the rankings of the fourth and fifth segments. As expected, the only perfect correlation is between the static spread ranking with itself within the same segment. What is unexpected, however, is how little correlation there exists between the dynamic and the static spread models and how quickly the correlation deteriorates as time unfolds.

To measure how much the network actually changes with time, we calculated the distance between the aggregate networks of every two segments. We measure this distance as the complement of the Jaccard similarity of the edge sets of the two networks. Figure 3 shows the scatter plot of the distance between segment networks versus the Spearman correlation coefficient between the spread rankings of the individuals in the two networks. The surprising feature of the plots is how little correspondence there is between the network similarity and the consistency of the rankings. Moreover, note that in most datasets the distance between any two segments is at least .4 and often reaches .8, which means networks typically change very fast with time.

The result of random perturbations of the edges in networks are shown in Figure 4. Here we only compare the rankings of the static linear threshold spread on the aggregate networks before and after the perturbation. We do not consider the dynamic spread models since the perturbations are not explicitly dynamic and we do not control the timesteps in which the random edges are perturbed. These results also show that the quality of the predictions of the spreading capacity of the individuals deteriorates rapidly with the increase in the amount of perturbation.

6.2 The change in the identity of the top spreaders

The second question we asked was whether, despite the fact that overall the relative predictions about the spreading capacity of individuals may not be robust, the identities of the top spreaders remain relatively constant. We compare the sets of the top five and top ten ranked individuals in the network before and after the spread. We measure the Jaccard similarity of the sets of the top spreaders. Figure 5 shows the similarity of the sets of the top five spreaders between the first segment and all subsequent segments. We omit the comparison between all other pairs of segments for brevity, but note that they show a similar trend. As the results show, the identity of the top five spreaders changes drastically as time unfolds. Figure 6 shows the scatter plot of the similarity between the top five sets versus the amount of change in the network. The results for the top ten ranked individuals are similar and we omit them due to space constraints.

[2] http://www.imdb.com

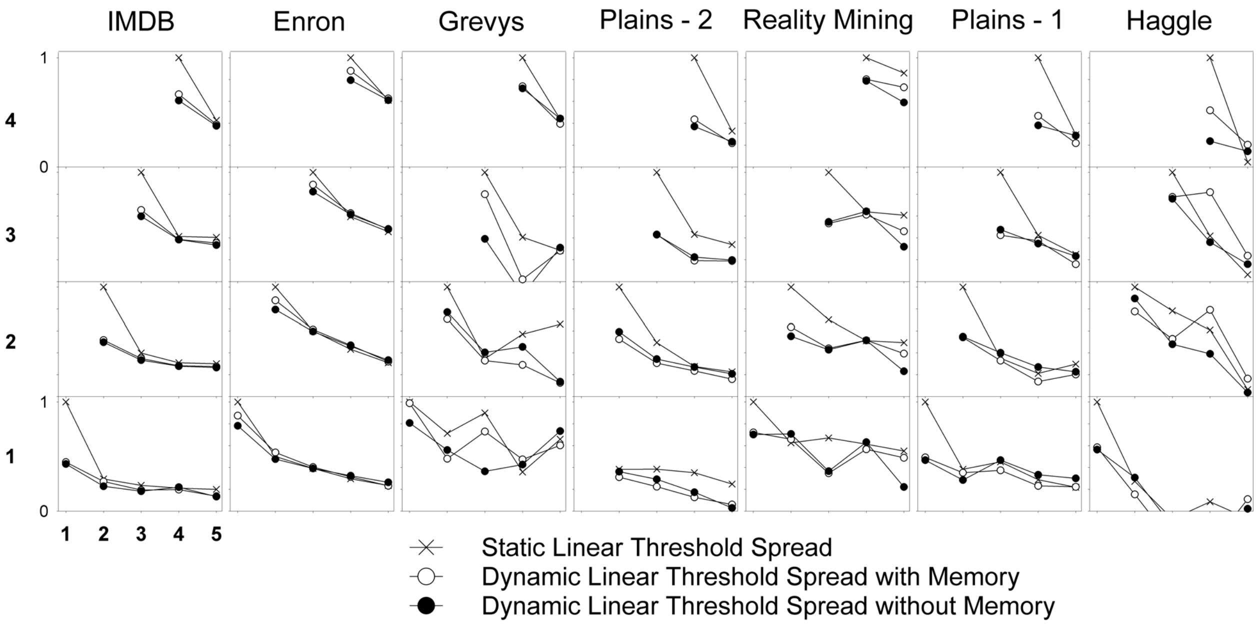

Figure 2. Spearman's correlation coefficient comparing the ranking of individuals (by estimated spreading capacity) across different segments. Comparisons are made both within the same segment and between current and future segments. Results for three different spread models are shown.

Even in the identity of the top ranked individuals there is little correspondence between the amount of change in the network and the consistency of the results. Recall that the datasets are ordered by their density. There is a possible trend that in sparser networks the top ranked individuals remain more consistently in the top, but before we draw any conclusions, this trend must be investigated further. Figure 4 shows the similarity of the top spreaders for the randomly perturbed networks. Even at 5% perturbation these sets are almost entirely different. Thus, for example, even small changes in the network may invalidate the predictions about the potentially good marketing targets.

6.3 The relative performance of the historical spreaders

Finally, despite the fact that, as we saw, the identity of the top spreaders may have changed, we asked whether the old top spreaders would still "perform" sufficiently well in the new, changed network. That is, if the top spreaders from the original network are used to initiate the spread in the new, changed network, how would the number of individuals affected by this spread compare to the extent of spread initiated by the new set of the top spreaders in the new network? The answer to this question directly implies the answer to whether the actions based on past predictions are valid as the network changes and lead to sufficiently good results.

Figure 7 shows the relative performance of the top five ranked individuals in segment 1 compared to the top five individuals in each subsequent segment. The results for other segments and the top ten individuals are similar and we omit them due to space constraints. Despite the fact that the identity of the top individuals changes, we see that the old top spreaders perform as well as the new top spreaders in many cases. However, this is true only for the networks where spread saturation is easily achieved and spread initiated from almost any set of individuals reaches everybody in the population. In sparser networks like IMDB and Enron, the performance of the old top spreaders deteriorates with time. Figure 8 shows the relationship between the performance of the top ranked sets and the amount of change between the old and the new networks which, again, demonstrates that there is little correspondence between the amount of perturbation and the performance of the top individuals.

Surprisingly, on randomly perturbed networks, the original top 5 set of individuals performed always nearly as well as the top set from the perturbed network. This is despite the fact that the sets themselves have few individuals in common. Thus, random perturbations may not be representative of changes in real networks and it is, then, particularly important to distinguish true patterns of network evolution

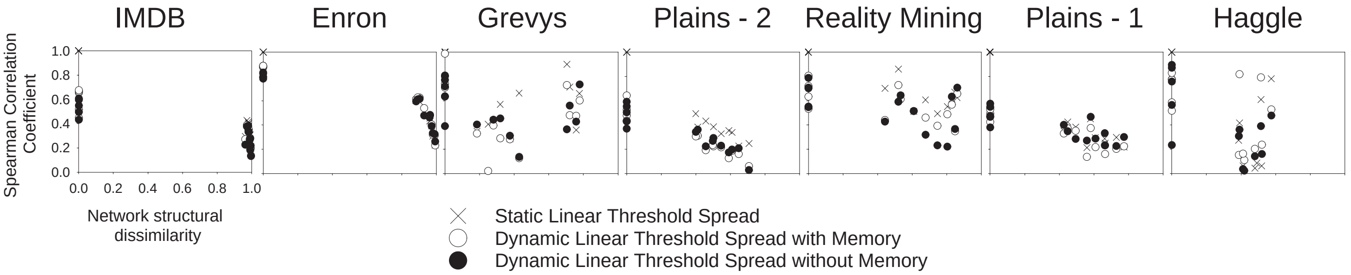

Figure 3. Spearman correlation coefficient between the ranking of individuals (by their estimated spreading capacity) as a function of the dissimilarity between the underlying networks. Dissimilarity (x-axis) is measured as the complement of the Jaccard similarity on the edge sets of the networks.

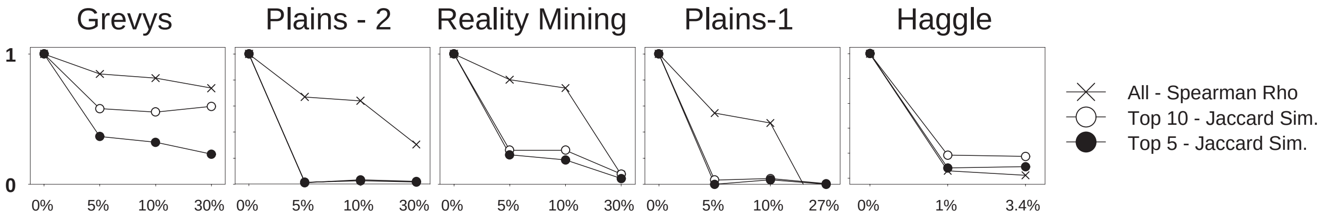

Figure 4. Spearman's correlation coefficient between the ranking of all individuals and the Jaccard similarity between the 5 and 10 top ranked individuals in the original and perturbed networks, both as functions of the percent of randomly perturbed edges in the network. The datasets are ordered by increasing density. The amount of perturbation cannot exceed the complement of the density.

and noise.

7 Conclusions and Future Work

Most social network analysis is performed on historical and typically aggregate data, and the possible structural changes that happen as the network evolves are not taken into consideration. Thus, by the time the analysis is completed and acted upon, its results may not be valid if the network indeed has changed in the meantime. In this paper, we asked how much such changes can affect the results of network analysis in the context of diffusion in networks. Specifically, we asked three questions: (1) whether the predictions about the relative spreading capacity of each individual are robust; (2) whether the sets of the top spreaders are relatively unaffected; and (3) whether the performance of the top spreaders in terms of the extent of spread they may cause remains good enough even after the changes. In the process of answering these questions, we also compared the predictions made on an the traditional aggregate, static representation of a network to the explicitly dynamic view of social interactions.

We found that in real dynamic networks the predictions about the relative spreading capacity of individuals and the identity of the top spreaders are sensitive even to minimal changes in the network. Moreover, we found that networks change significantly with time, often by as much as 40% of edges in a short time period. Surprisingly, we also found that there is little correspondence between the amount of change in the network and the robustness of the predictions. Finally, while in the real timeline, the performance of the top spreaders from the past did not compare well with the performance of the current top spreaders, in randomly perturbed networks past top spreaders typically did well even after the perturbations.

Thus, overall, we found that not only do predictions from the past not hold well into the future, these predictions do not deteriorate gracefully either. This implies that we cannot estimate the robustness of our predictions by measuring the amount of structural change in the network. Moreover, since random changes do not diminish the relative spreading ability of the top spreaders as much as the changes with real passage of time, we conclude that a few critical edges can make a big difference. Thus, we need methods for identi-

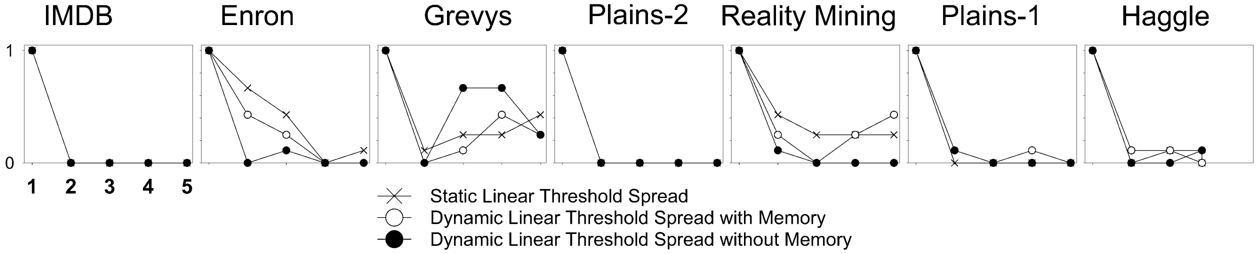

Figure 5. Jaccard similarity comparing the top 5 individuals (ranked by spreading capacity) in segment 1 with the top 5 individuals in subsequent segments. The x-axis is the current segment to which the top five set from segment 1 is being compared.

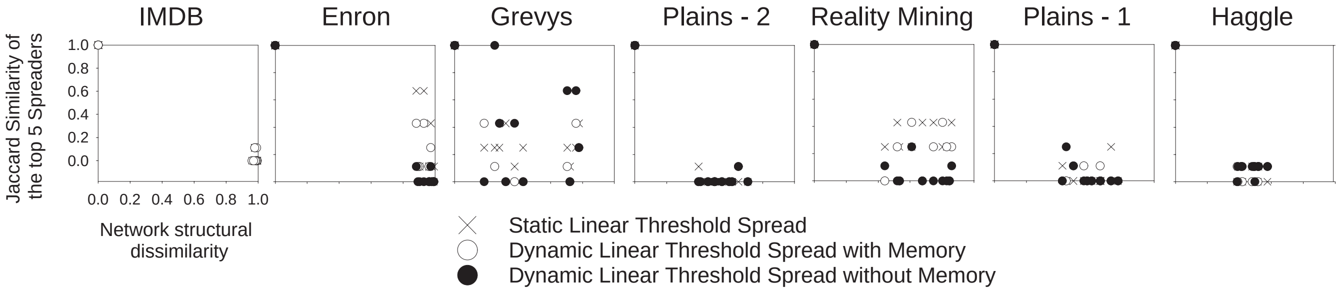

Figure 6. Jaccard similarity between the sets of top 5 individuals (ranked by spreading capacity) as a function of the dissimilarity between the underlying networks.

fying edges that are critical to the robustness of the predictions. We must also develop analysis techniques that take possible future network changes into consideration.

Finally, in almost all experiments, the analysis performed in an aggregate network using a static diffusion model had little correspondence to the explicitly dynamic models of spreading processes simulated on dynamic networks. Thus, in explicitly dynamic networks we must use analysis methods that explicitly take the dynamic nature of interactions into consideration.

8. Acknowledgements

We are grateful to Dan Rubenstein, Ilya Fischhoff, and Siva Sundaresan of the Department of Ecology and Evolutionary Biology at Princeton University for sharing the Plains Zebra data. Their work was supported by the NSF grants CNS- 025214 and IOB-9874523. This work is supported by NSF grants IIS-0705822 and CAREER IIS-0747369.

References

[1] J. I. Adibi. Enron email dataset. Downloaded from http://www.isi.edu/ adibi/Enron/Enron.htm.

[2] M. Ancel, M. E. J. Newman, M. Martin, and S. Schrag. Applying network theory to epidemics: Control measures for mycoplasma pneumoniae outbreaks. *Emerging Infectious Diseases*, February 2003.

[3] E. Berger. Dynamic monopolies of constant size. *Journal of Combinatorial Theory Series B*, 83:191–200, 2001.

[4] S. Bharathi, D. Kempe, and M. Salek. Competitive influence maximization in social networks. pages 306–311. 2007.

[5] K. Carley. Communicating new ideas: The potential impact of information and telecommunication technology. *Technology in Society*, 18(2):219–230, 1996.

[6] L. Chen and K. Carley. The impact of social networks in the propagation of computer viruses and countermeasures. *IEEE Tras. Syst., Man and Cybernetics*, 2005.

[7] N. Chen. On the approximability of influence in social networks. *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1029–1037, 2008.

[8] A. Clauset and N. Eagle. Persistence and periodicity in a dynamic proximity network.

[9] Z. Dezsö and A.-L. Barabási. Halting viruses in scale-free networks. *Physical Review E*, 65(055103(R)), 2002.

[10] P. Domingos. Mining social networks for viral marketing. *IEEE Intelligent Systems*, 20:80–82, 2005.

[11] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, V10(4):255–268, May 2006.

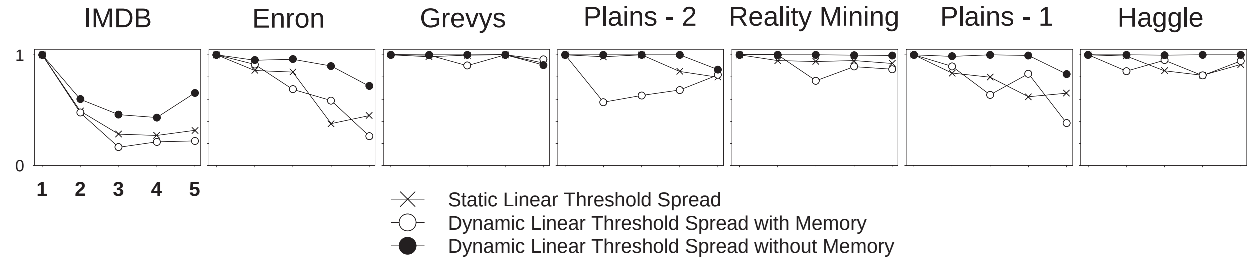

Figure 7. Relative performance of the top five sets of individuals from each segment in subsequent segments. Each datapoint with the x-coordinate value x is the ratio of the estimated extent of spread in segment x initiated by the top five individuals from segment 1 to that initiated by the top five individuals from segment x itself.

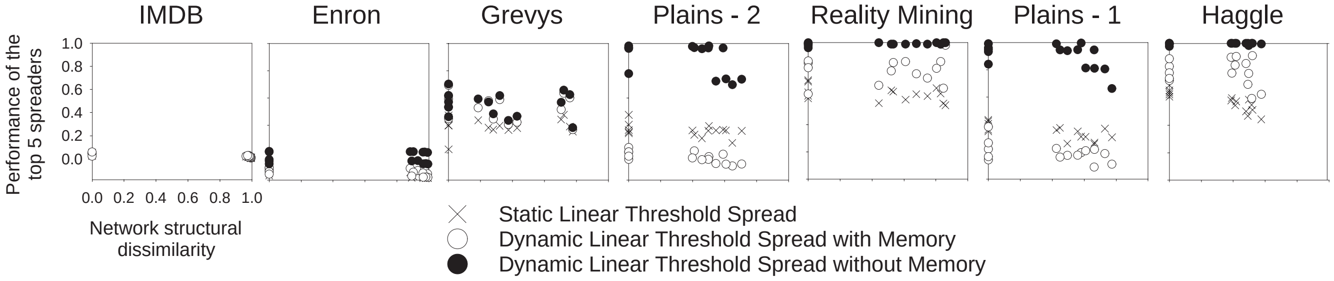

Figure 8. Relative performance of the top five sets of individuals as a function of the amount of change in the network. Each datapoint with x-coordinate x is the ratio of the estimated extent of spread initiated in the changed network by the top five individuals from the original network to that initiated by the top five individuals from the changed network itself, where the fraction of edges differing between the original and the changed networks is x.

[12] S. Eubank, H. Guclu, V. Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429:429:180–184., Nov 2004. Supplement material.

[13] I. R. Fischhoff, S. R. Sundaresan, J. Cordingley, H. M. Larkin, M.-J. Sellier, and D. I. Rubenstein. Social relationships and reproductive state influence leadership roles in movements of plains zebra, equus burchellii. *Animal Behaviour*, 73(5):825–831, May 2007.

[14] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12(3):211–223, 2001.

[15] J. Goldenberg, B. Libai, and E. Muller. Using complex systems analysis to advance marketing theory development. *Academy of Marketing Science Review*, 2001.

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

[17] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.

[18] Habiba and T. Berger-Wolf. Maximizing the extent of spread in a dynamic network. Tech. Rep. 2007-20, DIMACS, 2007.

[19] P. Holme. Efficient local strategies for vaccination and network attack. *Europhys. Lett.*, 68(6):908–914, 2004.

[20] P. Jaccard. The distribution of flora in the alpine zone. *The New Phytologist*, 11(2):37–50, 1912.

[21] F. Jordán and J. Benedek, Z.and Podani. Quantifying positional importance in food webs: A comparison of centrality indices. *Ecological Modelling*, 205:270–275, 2007.

[22] D. Kempe, J. Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD'03*, pages 137–146, 2003. ACM.

[23] P. Korkki. For marketers, viruses just won't cooperate. The New York Times, July 6, 2008.

[24] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. In *EC '06*, pages 228–237, 2006.

[25] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, and J. VanBriesen. Cost-effective outbreak detection in networks. In *KDD'07*, 2007.

[26] R. M. May and A. L. Lloyd. Infection dynamics on scale-free networks. *Physical Review E*, 64(066112), 2001.

[27] A. Miklas, K. k. Gollu, K. K. W. S. S. Chan, K. P. Gummadi, and L. E. Exploiting social interactions in mobile systems. In *The 9thth International Conference on Ubiquitous Computing (UbiComp 2007)*, pages 409–428. Springer-Verlag Berlin Heidelberg 2007, 2007.

[28] Y. Moreno, M. Nekovee, and A. F. Pacheco. Dynamics of rumor spreading in complex networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 69(6):066130, 2004.

[29] M. Morris. Epidemiology and social networks:modeling structured diffusion. *Sociological Methods and Research*, 22(1):99–126, 1993.

[30] E. Mossel and S. Roch. On the submodularity of influence in social networks. In *The Annual ACM Symposium on Theory of Computing(STOC)*, 2007.

[31] M. E. Newman. Spread of epidemic disease on networks. *Physical Review E*, 66(016128), 2002. DOI: 10.1103/PhysRevE.66.016128.

[32] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86(14):3200–3203, Apr 2001.

[33] J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau. CRAWDAD trace cambridge/haggle/imote/infocom (v. 2006-01-31). Downloaded from http://crawdad.cs.dartmouth.edu/cambridge/haggle/imote/infocom, Jan. 2006.

[34] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101, 1904.

[35] S. R. Sundaresan, I. R. Fischhoff, J. Dushoff, and D. I. Rubenstein. Network metrics reveal differences in social organization between two fission-fusion species, grevy's zebra and onager. *Oecologia*, September 2006.

[36] D. H. Zanette. Dynamics of rumor propagation on small-world networks. *Phys. Rev. E*, 65(4):041908, Mar 2002.