Models of Communication Dynamics for Simulation of Information Diffusion

Konstantin Mertsalov, Malik Magdon-Ismail, Mark Goldberg
Rensselaer Polytechnic Institute
Department of Computer Science
110 8th Street, Troy, NY
{mertsk2,magdon,goldberg}@cs.rpi.edu

***Abstract*—We study information diffusion in real-life and synthetic dynamic networks, using well known threshold and cascade models of diffusion. Our test-bed is the communication network of the LiveJournal Blogosphere. We observe that the dynamic and static versions of the Blogograph, yield very different behaviors of the diffusion. It was earlier discovered that the communication dynamics of the Blogograph is quite high - over 60% of the links each week were not present in the previous week, though the size of the node set is relatively stable. Our models of the Blogograph evolution reproduce general stable statistics of the real-life Blogograph. We discover that the diffusion footprint on our models closely approximate the diffusion footprint of the real-life dynamic network.**

I. Introduction

In recent years blogs and other Internet social media have become the major transmitter of information such as news, rumors and even intentional misinformation. Diffusion through blogs and forums on the web has reached the scope where it is no longer a local phenomenon of limited interest to online communities, but it has very real impact on the well-being of the offline world. Online channels of information diffusion were used by perpetrators to spread misinformation that resulted in significant loses by physical entities such as commercial banks [1]. Considering the low cost of attacks through diffusion of false rumors it is only reasonable to believe that it will continue and grow.

We need to track the extent of a diffusion over a dynamic network such as the blogosphere. Specifically, given the initial set of infected nodes and the diffusion laws (the basic properties of the diffusion such as how people get infected along links and at what rates) how will the diffusion spread. In order to be able to *predict* macroscopic properties of the diffusion dynamics (such as the rate at which the network is getting infected), it is necessary to have a model of the (communication) link dynamics of the network, because as we will show, having a static view of the network leads to a drastically different result.

Mathematical epidemiology and lately computer science has expended significant effort in developing and studying models of disease spread [2], [3], [4]. Typically such study has been on static networks. Information diffusion has similar properties to disease spreading, and the question we address is how different models of diffusion, in particular the cascade and threshold models, behave on *dynamic networks*. Our results indicate that the dynamics of the diffusion depends strongly on the type of diffusion model, and the dynamic view of the network versus the static view can have an even more drastic effect on how the diffusion spreads. Thus, given initial conditions, in order to predict how the diffusion will spread, it is necessary to have an accurate model of the dynamic evolution of the network. Our main goal is to study how different models of the communication dynamics of the social network fare in reproducing the observed diffusion footprint on the real dynamic network.

II. Related Work

The information diffusion in social networks was analyzed from theoretical [5] and empirical perspectives [6]. Newman [3] provides theoretical analysis of the spreading of disease in various networks. Kempe et al. [2] provide a framework for reasoning about spread of influence in a social network.

Leskovec et al. [7] study the cascading behavior of diffusion in the network of interlinked blog post. They propose the generative model of cascade formation which results in information cascades with similar properties as observed. The major difference between their generative model and the variation of the model that we apply to dynamic graphs is the ability of the nodes to be infected multiple times.

Lahiri et al. [8] investigate the impact of structural changes of the network on the individual ability of the nodes to facilitate the information diffusion. They found that results obtained from a static representation of the network had little correspondence with results from a dynamic representation. We investigate dynamic and static views of the networks generated by the models of link dynamics and we find that same phenomena is present as well.

Habiba and Berger-Wolf [9] extend the work of Kempe et al. [2] to study the diffusion in dynamic graphs. They solve the problem of selecting the set of individuals for initial infection so that resulting extend of the spread in the dynamic network is maximized. They observed the significant difference between the extend of the spread in aggregate and dynamic views of networks. We also came across a drastic difference in the scope of diffusion and we

further looked into the rate of diffusion under dynamic and static network representations.

A number of models of dynamic graph generation were proposed over the last decade (see [10], [11], [12], [13], [14], [15], [16]). These models capture the growth of the network where nodes arrive one at a time, attach and never leave the network or break existing edges. In our earlier work [17] we proposed the model that also incorporated the link dynamics, where edges are formed and broken as network evolution proceeds. We use these models of network evolution in this work.

III. Data

The data for this study was collected from the Russian section of LiveJournal. In this section we describe the collection process and the obtained network.

We chose to focus on the Russian section of LiveJournal as it is reasonable but not excessively large (currently close to 580,000 bloggers out of the total 15 million) and almost self-contained. This network already serves as a major carrier for variety of news, rumors and other information.

We refer to the LiveJournal network as the blogograph. We define the blogograph as a directed, unweighted graph representing communication of the bloggers within a fixed time-period. A vertex in the blogograph represents a blogger and a directed edge represent the communication by the means of commenting on a post. In particular, for every comment we place an edge from the vertex corresponding to the author of the comment to the author of the post. Parallel edges and loops are not allowed and a comment is ignored if the corresponding edge is already present in the graph. An illustration of the blogograph's construction is given on Figure 1.

The data was collected between December of 2007 and May of 2008 using a real time RSS update feature of LiveJournal that publishes all posts as they appear on any of the hosted blogs. We obtain the comments to the posts by "screen-scraping" them from the HTML code of the blog two weeks after the post was up.

We aggregate the edges corresponding to comments into the weekly snapshots of the network; for each week the communication graph contains the bloggers that either posted or commented during a week. We chose to split graphs into one week periods due to highly cyclic nature of activity in the blogosphere - the activity of bloggers on the weekends is much lower than on weekdays. For study of diffusion we create three views of this data - a static snapshots of each week, the union of ten snapshots and the dynamic graph that we define to be a sequence of ten consecutive snapshots. Therefore, the 21 observed weeks yield 12 union and dynamic graphs.

On average a weekly snapshot of blogograph contains about 153,000 vertices and 510,000 edges. About 70% of edges in a graph of a given week did not appear in a graph of a previous week. In-depth analysis and statistics for this network were reported in [17].

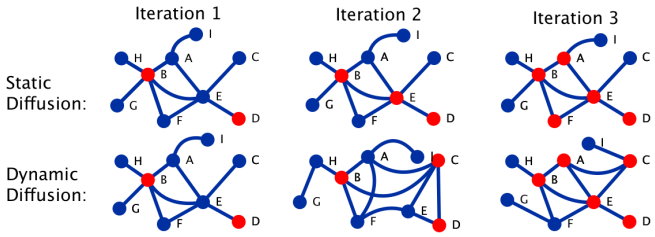


Figure 2. Diffusion in static and dynamic networks. Red nodes show the propagation of infection in static and dynamic graphs. In static case the network structure remains the same as diffusion progresses, but in dynamic case it changes.

IV. Models

We use models of communication dynamics to simulate the network evolution and the models of diffusion to mimic the diffusion of information. We evaluated two models of diffusion with seven models of network evolution.

A. Models of diffusion

The model of diffusion specifies the set of individuals that will be infected in the next time cycle given the structure of the graph and the set of individuals who are currently infected. Both cascade and threshold models start off by infecting a fraction of all vertices which are selected uniformly at random from the set of all vertices. The algorithm then determines the nodes that will be infected in the next iteration and then iterates. The set of newly infected nodes is determined as follows:

1) **Linear Threshold Model:** at the initialization the succeptability threshold θ_i is assigned to every node i and influence threshold $b_{i,j}$ is assigned to every pair (i, j). In the undirected case that we used any node i for which $\sum_j (b_{i,j} + b_{j,i}) > \theta_i$ will become infected. In our experiments we set all $b_{i,j} = 1$ and we randomly sampled θ_i for every node from uniform continuous distribution between 0 and 1. The threshold was sampled once per node and did not change in the duration of the experiment.

2) **Independent Cascade Model:** at the initialization every pair of nodes (i, j) in the graph is assigned the infection probability $p_{i,j}$ and this value does not change throughout the experiment. When a node becomes infected either during the initial random infection or from another node in the graph, it will be contagious and able to spread infection to its adjacent neighbors during the next iteration. The infection is passed over the edge (i, j) with probability $p_{i,j}$ that

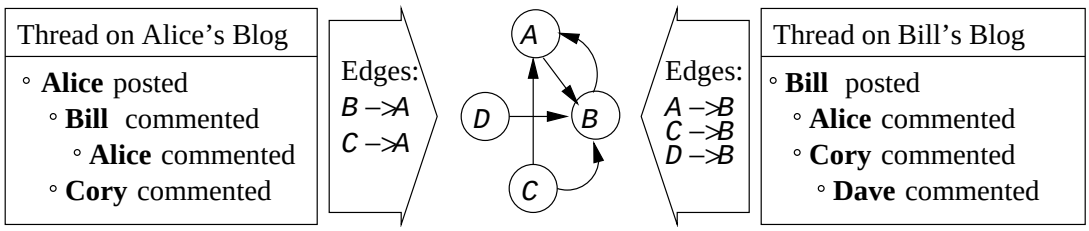


Figure 1. Blogograph generation example. Vertices are placed for every blogger who posted or commented, the edges are placed from the author of the comment to the author of the post (the blog owner). Parallel edges and loops are not allowed.

is assigned to this edge. Whether or not the node infects any of its neighbors during the contagious state it will not enter contagious state in the following iterations again. In our experiments we sampled $p_{i,j}$ from continuous probability distribution with probability density function $f(x) = \frac{1}{3}x^{-2/3}$ (cube of a random number sampled from uniform continuous distribution between 0 and 1).

B. Models of communication dynamics

The models of communication dynamics specify the structure of the graph of the next iteration given the structure of current graph and the distribution of out-degrees. The models generate the sequence of graphs recursively using the graph of the last iteration to produce the graph of the next. In the remainder of this section we will briefly describe these models. We refer the reader to [17] for detailed description and analyses.

These models of communication dynamics are based on the principle of locality: every node can attach to nodes within their local neighborhood using some method of selection of nodes for attachment. Full model of communication dynamics is specified by the definition of local area and the rule for attachment with in local area.

We considered three rules of attachment:

1) **Uniform attachment**: node can attach to any other node in its neighborhood with uniform probability.

2) **Preferential attachment (in-degree)**: the node is selected from the local area with probability proportional to its in-degree.

3) **Preferential attachment (out-degree)**:the node is selected from the local area with probability proportional to its out-degree.

We considered three definitions of local area:

1) **Global attachment:** every node is aware of the full network and its local area contains all nodes. As described in [17] global area definition only produces reasonable graphs with uniform attachment and attachment proportional to out-degrees. Therefore, this definition of area is only combined with two attachment models.

2) k**-neighborhood:** the nodes can attach to other nodes that are not further then k hops away from them. Similarly to a global area definition, k-neighborhood only produces meaningful graphs when combined with uniform attachment and attachment proportional to out-degree. We found that $k = 3$ produces the best models and we use 3-neighborhood area definition in our experiments.

3) **Union of clusters:** the area of a node is defined to be the union of all clusters [18] in the graph that contain this node. This area definition works well with all methods of attachment and therefore yields three models of communication dynamics.

For the first iteration, the models were given the graph with random assignment of edges. The models were also configured with out-degree distribution observed in LiveJournal.

We found that graphs produced by the combination of clusters area definition and preferential attachment were most similar to observed network when compared by the ensemble of parameters as shows in Table I. Graphs produced by combination of 3-Neighborhood area definition with preferential attachment were also quite similar to the observed ones. The summary of parameters of the models of communication dynamics is given in Table I.

The models of communication dynamics were used to generate the dynamic network. We experiment with three views of this network - the static snapshot graphs of each iteration, a union of ten snapshots and dynamic graph that is made up of a sequence of ten snapshot graphs.

C. Combining the models

The models of diffusion were applied to every view of observed and modeled networks. In case of static snapshot and union graphs the diffusion progressed through the same graph at every iteration. In case of dynamic view the structure of the graph changed as iterations of diffusion

(a) Static Parameters

Area	Attch	GC	C	d	E	g_{err}
Observed		0.9545	0.0613	5.34	0.0289	0.00144
Global	Uniform	0.9867	5.2×10^{-6}	7.86	1.075	0.04215
Global	P.A. (in)	-	-	-	-	-
Global	P.A. (out)	0.9688	0.00018	5.21	0.427	0.01189
3-Neighb.	uniform	0.8939	0.00045	5.30	0.4331	0.01792
3-Neighb.	P.A.(in)	-	-	-	-	-
3-Neighb.	P.A.(out)	0.9776	0.00133	4.53	0.1412	0.03504
Clusters	uniform	0.9646	0.00252	6.73	0.7267	0.03484
Clusters	P.A. (in)	0.9643	0.00149	6.88	0.1713	0.03811
Clusters	P.A. (out)	0.9523	0.03156	6.56	0.5320	0.02034

(b) Dynamic Parameter

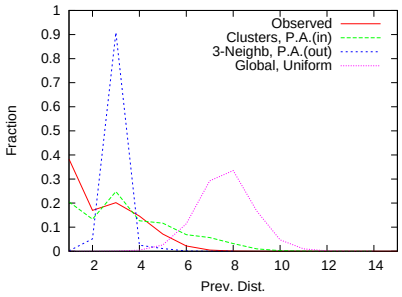


Table I

COMPARISON OF NETWORK PARAMETERS OF OBSERVED AND MODELED NETWORKS. TABLE (A) SHOWS THE STATIC PARAMETERS OF GRAPHS: GC - SIZE OF A GIANT COMPONENT, C - CLUSTERING COEFFICIENT, d - THE AVERAGE DIAMETER, E - THE DIFFERENCE OF OUT-DEGREE DISTRIBUTIONS, g_{err} - THE DIFFERENCE IN CLUSTER STRUCTURES OF THE GRAPHS. FIGURE (B) SHOWS THE EDGE HISTORY, WHICH MEASURES HOW CLOSE THE END POINTS OF THE OBSERVED EDGE WERE IN THE GRAPH OF PREVIOUS TIME CYCLE. THESE STATISTICS ARE COVERED IN MORE DEPTH IN [17]

progressed. The first iteration of diffusion occurred over the first snapshot in the dynamic graph, then the edges of the graph were changed in accordance with second snapshot and the second iteration of diffusion was executed and so on. Diffusion in static and dynamic views of network is illustrated in Figure 2.

V. RESULTS

We compare the rate of spread of the diffusion over models of the communication dynamics with diffusion over the real observed network. Given an initial fraction of randomly selected infected nodes, and the initial graph, we track at time step $t = 1, 2, \ldots$ the expected fraction of the network which is infected. The expectation is computed by a sample average over at least 20 runs (where we randomize over the set of infected nodes and the diffusion process itself).

First, we compare the diffusion on the dynamic and static views of real observed network. Figure 3 shows the rate of spread under cascade and threshold models of diffusion in dynamic and various static views of the network. The snapshot view takes the network at some time-step as a static network. The union static view aggregates all edges over ten time steps and considers this aggregated network as a static network. In dynamic view the network changes while diffusion progresses as discussed in the previous section. The initial infection is a randomly selected fraction of the entire network, and we experimented with different sizes for this fraction (0.1%, 1%, 5% and 10%). The results in all cases are qualitatively similar, and we only present the results for an initial infected fraction of 1%. It can be seen that the dynamic view leads to a drastic change in the diffusion footprint, as compared to either static view. In-fact, in the threshold model for the diffusion, the real dynamic network seems to add significant diffusive power.

The rate of diffusion in graphs generated by models of communication dynamics were compared to rate of spread in observed graphs. Table II provides some quantitative comparison of resulting curves of rate of diffusion. The results in Figure 4 consider the 2×3 matrix of possible diffusion scenarios, for the 3 possible views of the network (static snapshot, static union and dynamic) and the 2 possible diffusion models (cascade and threshold). In each plot the diffusion footprint for the real graph is grey, the area representing the range of observed behaviors (an error bar). We compare the observed true footprint with the diffusion footprints of four of the models of the network dynamics which were introduced in the previous section:

1. Cluster based areas with preferential attachment according to out-degree.

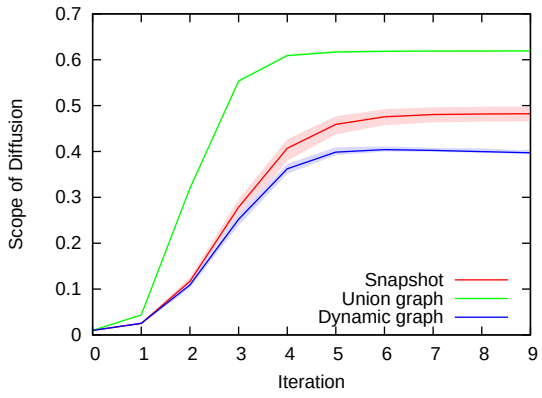
2. Global area with preferential attachment according to out-degree.

3. 3-neighborhood area with uniform attachment.

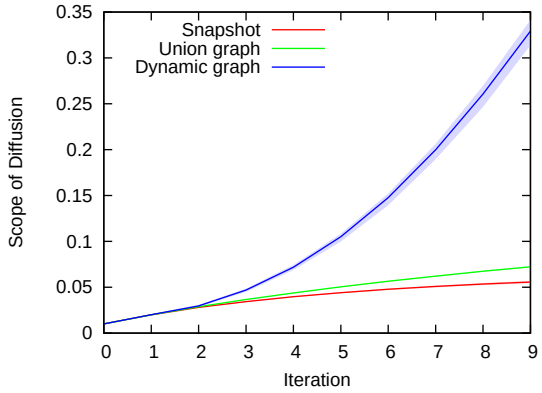
4. Global area with uniform attachment.

Model (4) is a benchmark evolving random graph which typically has a diffusion footprint that is nothing like the observed one. In general the other three models yield diffusion footprints in all the 6 scenarios which match the observed diffusion well. By our estimate the 3-neighborhood model with uniform attachment seems best. The largest deviation between the models and the observed is for the aggregate union view. We hypothesize that this is because the 1-neighborhoods in the real graph are much more stable than the 1-neighborhoods in any of the models. The results indicate that these models are good approximations to the reality in large social networks and hence provide a viable test bed for studying diffusion in social networks, especially networks which are as dynamic as the blogosphere.

We observe that the cascade diffusion model stabilized in all graphs after just a few iterations although it reached different final infection rate based on the view of the graph (static or dynamic). In the case of the threshold models



(a) Cascade Model



(b) Threshold Model

Figure 3. The rate of diffusion in observed LiveJournal blogograph for two static views of the network and the dynamic view.

the diffusion stabilized quickly in the static snapshot and union views of the network but progressed much more aggressively in dynamic view, infecting at a much higher rate. This observation compliments the results from [9], [8] where authors found that the influential nodes (sets which achieve maximal spread) in the static and dynamic views of the network are not consistent. It has also been corroborated in other settings such as ad-hoc network routing where mobility in the network (which results in link dynamics) can significantly increase the throughput [19].

VI. Discussion

There are two main conclusions which our study supports. Standard models of diffusion have different spread properties on the *real dynamic* LiveJournal network as compared to various static views of the network (we considered the snapshot view and the aggregated union of edges view in addition to the dynamic view). Hence it is important to take into account interaction between the link dynamics and the diffusive process if one is to have an accurate picture of the spread. In fact we see that the dynamic graph for

	Obs.	k-N.PA	Glob.Un.	Clst.PA
Threshold				
Dynamic	0.40,0	0.52,0.49	0.85,1.72	0.62,0.84
Union	0.13,0	0.25,7.90	0.23,7.12	0.22,5.74
Cascade				
Dynamic	0.39,0	0.45,0.50	0.16,2.25	0.50,0.76
Union	0.61,0	0.66,1.32	0.99,8.92	0.75,3.25

Table II
Comparison of rate of diffusion in dynamic and union views of the graph under cascade and threshold models with three models of communication dynamics. For each combination of a graph view, model of diffusion and model of communication dynamics we provide values x,y - x: the fraction of graph infected and y: the difference of area under the curve of rate of diffusion (Figure 4) of the modeled and observed networks.

certain diffusion models (eg. the threshold model) results in faster spread than even the union graph which aggregates all observed edges into a single static graph. The dynamics *increases* the diffusive power. This is a very surprising observation and we point out that similar phenomena such as the increase in throughput of a mobile ad-hoc network versus a static ad-hoc network have also been observed [19].

Since dynamics has a big impact on the diffusion, it follows that in order to predict the future spread, one needs to have a model for the link dynamics. We showed that for the LiveJournal network certain models are bad (for example random link dynamics) whereas certain models are very good at reproducing the observed diffusion dynamics of the real network. In particular, uniform attachment within the 3-neighborhood, global preferential attachment according to out-degree and cluster-area based preferential attachment according to out-degree produce diffusion dynamics which are faithful to the real dynamic network's diffusion dynamics.

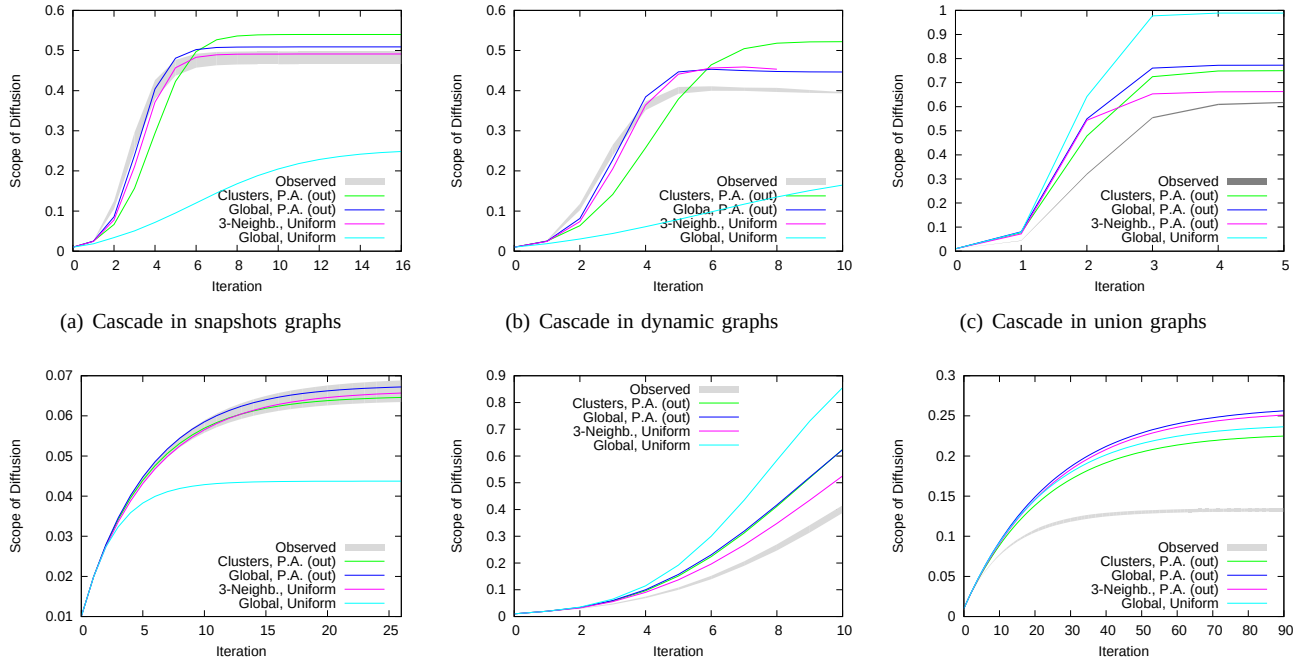
Our work has studied one particular aspect of the diffusion dynamics, namely the rate at which the network gets infected. It would be interesting to also study how good these models are at reproducing some other properties of the nodes which get infected (such as degree distributions) and how the dynamics may change the set of influential nodes.

References

[1] V. Pavlov, "(in russian) vojny on-line s dostavkoy (online wars delivered)." *Den'gi 2.0*. [Online]. Available: http://www.kommersant.ru/doc.aspx?DocsID=1098179

[2] D. Kempe, J. Kleinberg, and Éva Tardos, "Maximizing the spread of influence through a social network," in *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2003, pp. 137–146.

[3] M. E. J. Newman, "The spread of epidemic disease on networks," *Physical Review*



(a) Cascade in snapshots graphs (b) Cascade in dynamic graphs (c) Cascade in union graphs

(d) Threshold in snapshots graphs (e) Threshold in dynamic graphs (f) Threshold in union graphs

Figure 4. The rates of diffusion in graphs generated with different models of communication dynamics.

E, vol. 66, p. 016128, 2002. [Online]. Available: http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0205009

[4] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Physical Review Letter*, vol. 86(4), 2001.

[5] X. Song, Y. Chi, K. Hino, and B. L. Tseng, "Information flow modeling based on diffusion rate for prediction and ranking," in *WWW '07: Proceedings of the 16th international conference on World Wide Web*. New York, NY, USA: ACM, 2007, pp. 191–200.

[6] daniel gruhl, r. guha, david liben nowell, and andrew tomkins, "information diffusion through blogspace," in *www '04: proceedings of the 13th international conference on world wide web*. new york, ny, usa: acm, 2004, pp. 491–501.

[7] J. Leskovec, M. Mcglohon, C. Faloutsos, N. Glance, and M. Hurst, "Cascading behavior in large blog graphs," in *In SDM*, 2007.

[8] M. Lahiri, A. S. Maiya, R. Sulo, Habiba, and T. Y. B. Wolf, "The impact of structural changes on predictions of diffusion in networks," in *ICDM Workshop on Analysis of Dynamic Networks*, December 2008.

[9] Habiba and T. Berger-Wolf, "Maximizing the extent of spread in a dynamic network," *DIMACS Technical Report*, vol. 20, 2007.

[10] R. Albert and A.-L. Barabási, "Statictical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 47-97, 2002.

[11] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stat, A. Tomkins, and J. Wiener, "Graph structure in the web," *Computer Networks*, vol. 33, no. 1-6, pp. 309–320, 2000.

[12] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *SIGCOMM*, 1999, pp. 251–252.

[13] J. M. Kleinberg and S. Lawrence, "The structure of the web." in *Science*, 2001, pp. 1849–1850.

[14] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," in *KDD'06*, 2006.

[15] M. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[16] M. Newman, A.-L. Barabási, and D. Watts, "The structure and dynamics of networks," *Princeton University Press*, 2006.

[17] M. Goldberg, S. Kelley, M. Magdon-Ismail, K. Mertsalov, and W. A. Wallace, "Communication dynamics of blog networks," in *Proceedings SIGKDD Workshop on Social Network Mining and Analysis*, 2008.

[18] J. Baumes, M. Goldberg, and M. Magdon-Ismail, "Efficient identification of overlapping communities," *IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 27–36, May 2005.

[19] M. Grossglauser and D. Tse, "Mobility increases the capacity of ad-hoc wireless networks," *IEEE/ACM Transactions on Networking*, vol. 10, pp. 477–486, 2001.