Stable Statistics of the Blogograph

Mark Goldberg, Malik Magdon-Ismail, Stephen Kelley, Konstantin Mertsalov

Rensselaer Polytechnic Institute
Department of Computer Science

Abstract. The primary focus of this paper is to describe stable statistics of the blogosphere's evolution which convey information on the social network's dynamics. In this paper, we present a number of non-trivial statistics that are surprisingly stable and thus can be used as benchmarks to diagnose phase-transitions in the network. We believe that stable statistics can be used to identify anomalous behavior at all levels: that of a node, of a local community, or of the entire network itself. Any substantial change in those stable statistics must alert the researchers and analysts to the need for further investigation. Furthermore, the usage of these or similar statistics that are based solely on the communication dynamics and not on the communication content, allows one to diagnose anomalous behavior with minimal intrusion of privacy.

1 Introduction

Large social networks, such as the Blogosphere, are now channels for a significant portion of information flow. One would expect important social events to manifest themselves on such social networks as changes to the information flow dynamics, slightly before, during and after the events. More specifically, suppose one tracks social groups which are identified based solely on the pattern of their communication. One might ask whether a particular group gains in popularity and has the potential for becoming a large movement, so that a thorough study of this group is warranted. In order to answer questions like this, a picture of what *normal* group dynamics and behavior look like is needed as a benchmark against which hypotheses might be tested.

Our goal is to develop a framework for detecting anomalous behavior in blogosphere-like social networks. In particular, we take the first step in this direction by describing normal behavior against which anomalous behavior can be calibrated. As our test-bed, we take data from the LiveJournal blogosphere. There are certainly many parameters that can be extracted from the data. However, for any *statistic* of the social network's evolution to be useful as a diagnostic tool of anomalous behavior, the statistic should be *stable* during the normal functioning of the network. Only then can we identify a change in the statistic, and possibly connect it to some underlying change in its fundamental behavior.

The primary focus of this paper is to describe stable statistics of the blogosphere's evolution which convey information about the social network's dynamics. We categorize such statistics as follows:

(i) Individual statistics: statistics relating to properties of individual nodes, such as in-degree and out-degree distributions;
(ii) Relational statistics: statistics describing communication links (edges) in the network, such as the persistence of edges, correlation, and clustering coefficients;
(iii) Global statistics: statistics relating to global properties of the network, such as the size and diameter of its largest ("giant") component and total communication density;
(iv) Community statistics: statistics relating to the community (social group) structure in the network; and
(v) Evolution statistics: statistics relating to evolution in the social network; for example, the average lifespan of a social group.

We are interested in the dynamics of such statistics, in particular, their stability. In this paper, we present a number of non-trivial statistics that are surprisingly stable and thus can be used as benchmarks to diagnose phase-transitions in the network. The stability of these statistics is surprising because even though the network size is stable, the network dynamics itself is far from stable–our experiments show that close to 60% of the edges in the network change from week to week. We believe that stable statistics can be used to identify anomalous behavior at all levels: that of a node, of a local community, or of the entire network itself. Any substantial change in those stable statistics must alert the researchers and analysts to the need for further investigation. Furthermore, the usage of these or similar statistics that are based solely on the communication dynamics and not on the communication content, allows one to diagnose anomalous behavior with minimal intrusion of privacy.

2 LiveJournal Blog Data

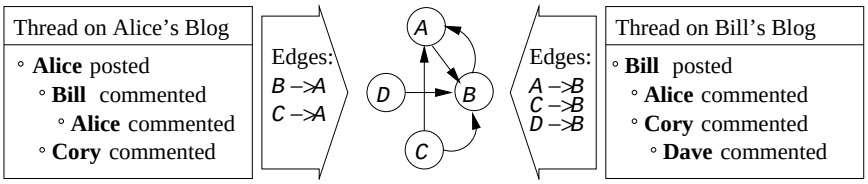

Fig. 1. Blogograph generation example. Vertices are placed for every blogger who posted or commented, the edges are placed from the author of the comment to the author of the post (the blog owner). Parallel edges and loops are not allowed.

We define the blogograph to represent the communication within a fixed time-period. For our experiments, this period is one week. The blogograph is a directed unweighted graph with a node for every blogger and a directed edge

w	$\|V\|$	$\|E\|$	GC	C	d	α
35	111,248	376,878	96.0%	0.0788	5.336	2.87
36	118,664	411,294	96.0%	0.0769	5.327	2.74
37	120,549	410,735	96.0%	0.0752	5.375	2.79
38	119,451	386,962	95.8%	0.0728	5.455	2.82
39	113,296	323,062	95.2%	0.0666	5.641	2.80
40	124,329	430,523	96.3%	0.0764	5.332	2.77
41	121,609	380,773	95.9%	0.0705	5.471	2.81
42	124,633	415,622	96.2%	0.0739	5.349	2.74
43	123,898	403,309	96.5%	0.0713	5.425	2.81

Fig. 2. Statistics for observed blogograph: order of the graph ($|V|$), graph size ($|E|$), fraction of vertices that are part of giant component (GC size), clustering coefficient (C), average separation (d), power law exponent (α)

from the author of any comment to the owner of the blog where the comment was made during the observed time period. Parallel edges are not allowed and a comment is ignored if the corresponding edge is already present in the graph. To study the evolution dynamics, we considered consecutive weekly snapshots of the network. The communication graph contains the bloggers that either posted or commented during this week and the edges represent the comments that appeared during the week. An example blogograph is given on Figure 1.

Our data was collected from the popular blogging service LiveJournal. LiveJournal imposes few restrictions on communication. What makes this network particularly interesting for our purposes is that bloggers typically make decisions to communicate and join social communities without strong influence from the outside. For this reason we believe the network observed at LiveJournal has a natural communication structure as the steady state of the network evolution. This makes the LiveJournal Blogosphere an attractive domain for our research.

Much of the communication in LiveJournal is public, which allows for easy access, especially given the real time RSS update feature provided by LiveJournal that publishes all open posts that appear on any of the hosted blogs. In our experience, the overwhelming majority of comments appear on a post within two weeks of the posting date. Thus, our screen-scraping program visits the page of a post after it has been published for two weeks and collects the comment threads. We then generate the communication graph.

We have focused on the Russian section of LiveJournal. as it is reasonable but not excessively large (currently close to active 250,000 bloggers) and almost self contained. We identify Russian blogs by the presence of Cyrillic characters in the posts. Technically, this also captures the posts in other languages with a Cyrillic alphabet, but we found that the vast majority of the posts are actually Russian. The community of Russian bloggers is very active. On average, 32% of all LiveJournal posts contain Cyrillic characters. Our work is based on data collected during September and October of 2006.

3 Global, Individual, and Relational Statistics

The observed communication graph has interesting properties. The graph is very dynamic (on the level of nodes and edges) but quite stable if we look at some aggregated statistics. For any week, about 70% of active bloggers will also be active in the next week. Further, about 40% of edges that existed in a week will also be found in the next week. A large part of the network changes weekly, but a significant part is preserved. Some of the important parameters of the blogograph illustrating their stability are presented in Figure 2. The giant component (GC) is the largest connected subgraph of the undirected blogograph. A giant component of similar size has been observed in other large social networks [5, 4]. The clustering coefficient (C) refers to the probability that the neighbors of a node are connected. The clustering coefficient of a node with degree k is the ratio of the number of edges between it's neighbors and $k(k-1)$. The clustering coefficient of the graph is defined to be the average of the node clustering coefficients. The observed clustering coefficient is stable over multiple weeks and significantly different from the clustering coefficient in a random graph with the same out-degree distribution, which is 0.00029. The average separation (d) is the average shortest path between two randomly selected vertices of the graph. We computed it by sampling 10,000 random pairs of nodes and finding the undirected shortest path between them. The blog communication graph is not significantly different with respect to this parameter than other observed social networks [5, 6].

The in-degree of a node describes its popularity in a network. The popularity is determined thought the interaction of network participants and depends on the properties of the participants and the network structure. Many large social networks [1, 4] have a power law in-degree distribution, $P(k) \propto k^{-\alpha}$, where $P(k)$ is the probability a node has degree k. Figure 3 shows the in-degree distribution averaged over the observed period. We observed a power law tail with parameter $\alpha \approx 2.81$, which is stable from week to week. This value was computed using the maximum likelihood method described in [3] and Matlab code provided by Aaron J. Clauset.

Figure 6 shows the average cumulative in-degree distribution over 9 weeks of observed data with an envelope that shows the maximum and minimum curves over the same 9 weeks shown with grey area. The envelope curves appear very close to the average value, clearly showing the stability of the in-degree distribution.

The out-degree distribution of the network describes the activity levels of the participants. Figure 7 shows the average cumulative out-degree distribution over 9 weeks of data with an minimum and maximum curve envelope. As with the in-degree distribution, the envelope curves of the out-degree distribution appear very close to the average value and illustrate the stability of the out-degree distribution.

We use edge stability and edge history to evaluate the evolutionary dynamics of individual edges in the snapshots of the evolving network. Edge stability measures the number of the observed time periods that contained a particular edge and the edge history measures how close the end points of the edge were

in the previous iteration conditioned on the activity level of the source of the edge. Both edge history and edge stability can be measured in the directed or undirected graph. We found that directed version to be more informative for edge stability evaluation and the undirected version to be more informative for edge history. Figure 4 presents the edge stability distribution shown on a log scale. As shown, the majority of the edges appear only once or twice in the observed period, but the network also contains some edges that are very stable and appear almost in every observed week.

We define the history H_{ij}^{T} of an edge (i, j) found in iteration T to be the geodesic distance between vertices i and j in the graph of iteration $T - 1$. The average edge history with minimum and maximum curve envelopes over nine observed weeks of data is presented in Figure 5. This plot shows the average portion of edges whose end points had a geodesic distance one, two, three, etc in the previous observed week for each activity level (out-degree). The lower line on the plot shows the portion of the edges in time period T that were present in the the graph of time period $T - 1$ and therefore had geodesic distance one. The second line from the bottom shows the portion of edges whose end points had geodesic distance at most two, third line is for portion of edges with geodesic distance three, etc. The minimum and maximum curves for each line bound the envelope around it. Clearly, the envelope is very close to the line itself. This suggests that the edge histories are stable in the observed period. It is surprising to see that the portion of the edges that repeat week to week conditioned on the out-degree of the edge source is so stable. As Figure 5 shows the portion of edges repeated in the next week is around 40% for vertices with out-degree five, 45% for vertices with out-degree ten, and 47% for vertices with out-degree fifteen. Furthermore, the portion of edges for which the end points had geodesic distance greater then one follows the same trend.

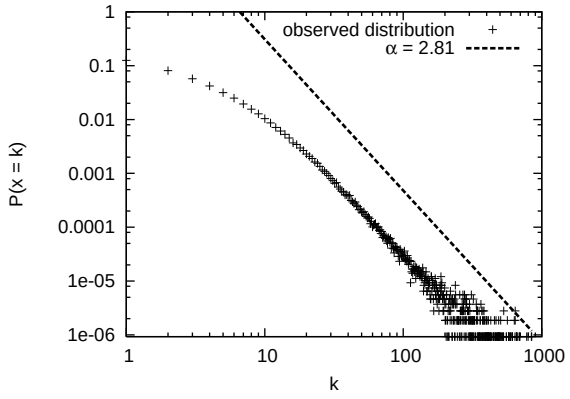

Fig. 3. Average in-degree distribution

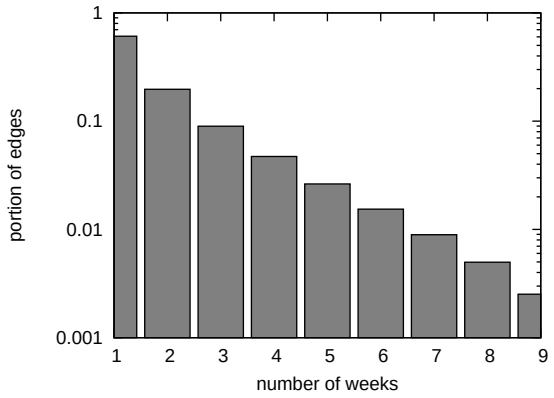

Fig. 4. Edge stability

4 Community and Evolution Statistics

Beyond statistics centering on individual vertices and edges, statistics of groups may also be examined. In order to determine groups in the data, the Iterative Scan algorithm presented in [2] was used. This algorithm produces sets of vertices which are locally optimal with respect to some density function. A formal definition of communities discovered is given as follows.

Definition Given a graph $G(V, E)$ let function δ, called the *density*, be defined on th set of all subsets of V. Then, a set $C \subseteq V$ is called a cluster if it is locally maximal w.r.t δ in the following sense: for every vertex $x \in C$ (resp. $x \notin C$), removing x from C(resp. adding x to C) creates a set whose density is smaller than $\delta(C)$.

This definition is compatible with social science observations that a community is a set of individuals with more communications within the set than outside of it. Within the description above, the formulation of the density function δ is left to the algorithm user. In these experiments, the definition is

$$\delta = \frac{E_{in}}{E_{in} + E_{out}} + \lambda e_p$$

where E_{in} and E_{out} are the numbers of edges within the community and cut by the community boundary respectively, e_p is the edge probability within the community, and λ is a parameter which can either increase or decrease the amount of weight placed on the edge probability of a community. This weighting was added to the density function improve the intuitive "quality" of clusters in sparse graphs such as the one detailed in this paper. Without this term, sparse areas of the graph can be added to a cluster quite easily resulting in very large communities with high diameters.

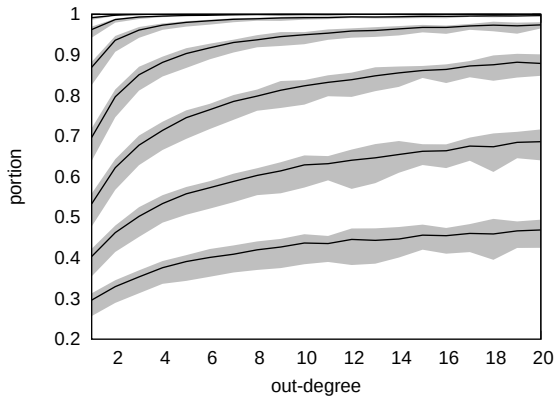

Fig. 5. Average edge histories with envelopes. The bottom line presents the portion of edges that existed in the previous iteration; every next line shows the portion of the current edges whose endpoints in the previous iteration were on the distance not exceeding the corresponding value.

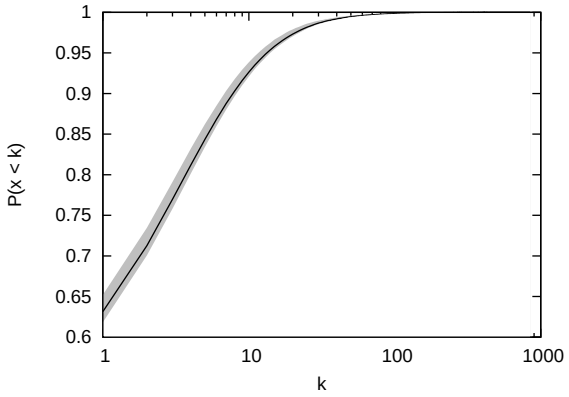

Fig. 6. Cumulative in-degree distribution with envelope for nine observed weeks

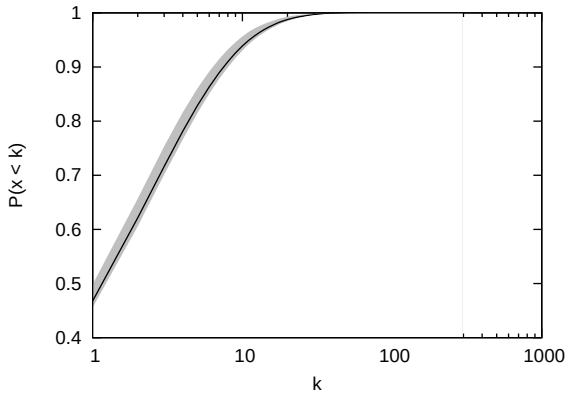

Fig. 7. Cumulative out-degree distribution with envelope for nine observed weeks

The algorithm was seeded using the Link Aggregate algorithm described in [2]. The number of clusters produced after optimization via Iterative Scan, their average size, average density, and average edge probability are all shown in Figure 8. Further, two plots showing size and density are given in Figure 9. Note the similarities in both scale and shape of these plots. Also in Figure 9 is a plot showing the boundary of each week's plot. Here, each point is defined as the largest 5% of the clusters in a given range along the y-axis of the plot. The portion of this plot where the lines are furthest apart are areas of few communities. However, it can be seen that each of the plots has an upper portion similar to those observable in the preceding weekly plots. The plots also show that each week has a number of low density communities of size 2. These communities are merely seeds which optimization did not modify. They can be filtered out based on some domain specific criteria, but in this case, were left in the data to get a more general sense of the algorithm's performance without obscuring details.

week	$\|C\|$	$size_{avg}$	δ_{avg}	e_p^{avg}
35	7700	9.1827	0.5606	0.30258
36	7602	9.2324	0.5495	0.30522
37	7688	9.1895	0.5521	0.30258
38	8647	9.0304	0.5516	0.30259
39	9965	9.1915	0.5389	0.29669
40	7908	9.0282	0.5519	0.30556
41	9094	9.1223	0.5348	0.29901
42	8240	9.1368	0.5379	0.30066
43	8768	9.0991	0.5357	0.30282

Fig. 8. Cluster Statistics

Now that communities are clearly defined, the question of how they evolve over time arises. For this paper, we have defined community evolution as follows. The Iterative Scan algorithm takes as input a set of seeds and produces optimized output communities. The output from running the algorithm on one week can be used as input to the next week's optimization. This causes some difficulty as sets of connected vertices taken from one graph may not be connected in the next. In order to get around this, the set of vertices that make up the optimized community are placed into the next graph and the largest connected component of this set in the new graph is used as a seed. A second difficulty is the definition of when a community actually succeeds another. Given two successive communities C_t and C_{t+1} discovered in the manner described above, we consider cluster C_{t+1} to be a continuation of cluster C_t if

$$\frac{|C_t \cap C_{t+1}|}{|C_t \cup C_{t+1}|} > t$$

where t is a threshold value indicating how strong we require the continuation to be. We define the lifespan of some initial community as the number of consecutive graphs in which the initial community exists or one of its continuation communities exists.

We measure these lifespans with respect to some initial set of communities which are discovered in the manner presented at the start of this section. Figure 10 shows a histogram of the lifespans with respect to three different starting weeks in the 9 week data. These numbers appear to be quite stable.

5 Conclusion

In the observed graph, communication patterns are dynamic. Even with these changes in the linkage of individual nodes, general statistics appear to be quite stable. Beyond this, link evolution and community evolution present another set of statistics which are stable. We propose that each of these sets of base statistics can be used as a foundation upon with future mechanisms for detecting anomalous individuals and communities can be built. In the future, this work will be expanded to a variety of structurally different social networks. In these explorations, additional in-depth statistics will also be examined.

Acknowlgements. This material is based upon work partially supported by the U.S. National Science Foundation (NSF) under Grant Nos. IIS-0621303, IIS-0522672, IIS-0324947, CNS-0323324, NSF IIS-0634875 and by the U.S. Office of Naval Research (ONR) Contract N00014-06-1-0466 and by the U.S. Department of Homeland Security (DHS) through the Center for Dynamic Data Analysis for Homeland Security administered through ONR grant number N00014-07-1-0150 to Rutgers University. The content of this paper does not necessarily reflect the position or policy of the U.S. Government, no official endorsement should be inferred or implied.

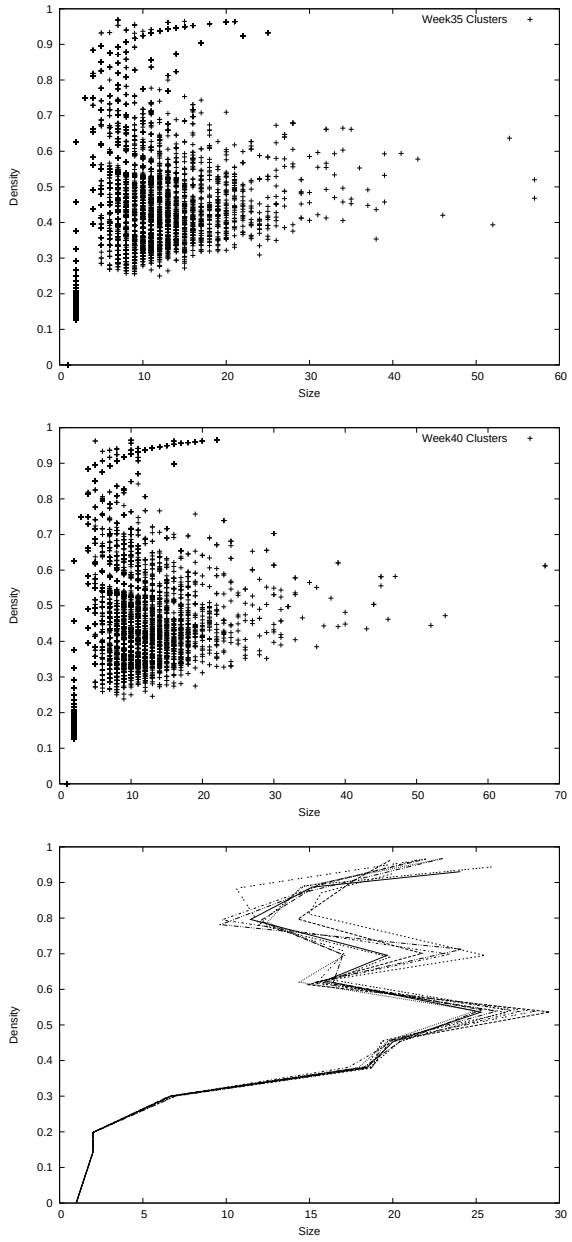

Fig. 9. The top figures show a size-density plot for weeks 35 and 40. Each point represents one discovered community. The bottom figure shows a line representing the boundaries of each week's plot.

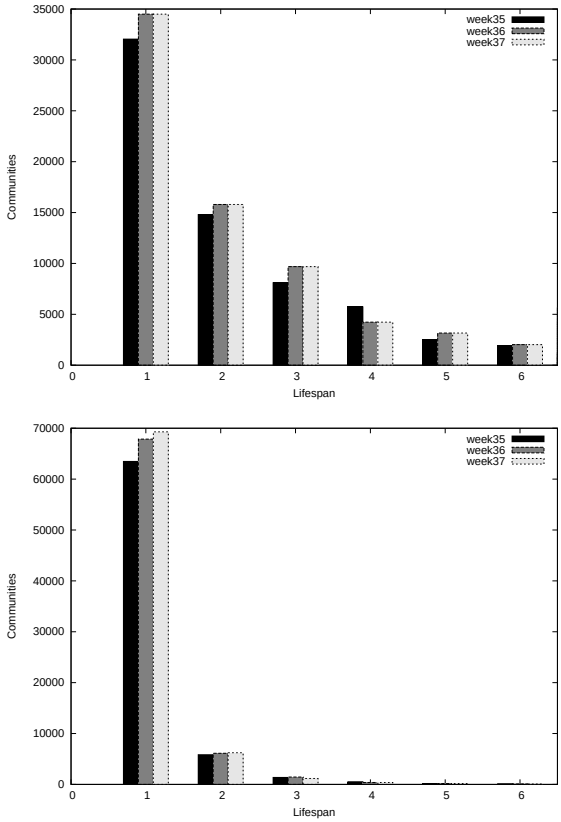

Fig. 10. Lifespan based on continuation results. The first image has $t = 0.3$ while the second has $t = 0.4$.

References

1. A. Barabási, J. Jeong, Z. Nëda, E. Ravasz, A. Shubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica*, A 311(590-614), 2002.
2. J. Baumes, M. K. Goldberg, and M. Magdon-Ismail. Efficient identification of overlapping communities. In P. B. Kantor, G. Muresan, F. Roberts, D. D. Zeng, F.-Y. Wang, H. Chen, and R. C. Merkle, editors, *ISI*, volume 3495 of *Lecture Notes in Computer Science*, pages 27–36. Springer, 2005.
3. A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data, 2007.
4. K.-I. Goh, Y.-H. Eom, H. Jeong, B. Kahng, and D. Kim. Structure and evolution of online social relationships: Heterogeneity in unrestricted discussions. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 73(6):066123, 2006.
5. G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311:88–90, 2006.
6. M. E. J. Newman. The structure of scientific collaboration networks. *PROC.NATL.ACAD.SCI.USA*, 98:404, 2001.