Stability of Individual and Group Behavior in a Blog Network

Stephen Kelley, Mark Goldberg, Malik Magdon-Ismail, Konstantin Mertsalov
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180
{kelles, goldberg, magdon, mertsk2}@cs.rpi.edu

***Abstract*—This work experimentally examines different notions of stability of the behavior of individuals and groups in a network of blogs. Our experiments are conducted on data collected from LiveJournal. All stability notions aim to locate stable behavior within an individual's area, which is defined in a variety of manners. Our experiments confirm an earlier observation of the highly dynamic nature of the network. Roughly 70% of the communication of a typical week was not observed in the previous week. Depending on the definition of stability and area used, we find small, but highly stable, sets of individuals with stable behavior in the network.**

I. Introduction

The emergence of vast, easily observable networks such as those formed by the WWW, email activity, social networking sites, and blog communications has enabled a large amount of research focusing on network dynamics. Recent work has shown that some networks have connections which are very dynamic while vertex sets remain comparatively static [1][2]. Given these intense reconnection dynamics, the identification of stability becomes important, as it enabes classification, prediction, and understanding of individual and group behavior within the network.

Previous research has focused on locating stability in the sea of statistics that can be generated from evolving, dynamic graphs [3]. This has led to a better understanding of universal trends such as network size during a given period, what portion of edges remain constant, and how out-degree and in-degree relate in a given snapshot of a network. However, this work did not aim to locate individual users or groups whose behavior is stable. Using a series of snapshots of communication patterns in a blog network, this work will present and locate various notions of behavioral stability at both the individual and collective level. Rather than focus on the stability of social groups or topics, an individual's behavior will be examined through the concept of an area. Conceptually, areas are defined as a subset of the graph where a specific user is likely to attach edges in the future. The stability of a variety of areas are examined further in the text, offering glimpses into the stability of the network's behavior at different levels of granularity.

The dynamics of the blogograph, a series of networks formulated from actual communications within a blog provider, make it incredibly difficult to pin down a singular notion of stability. Individuals may be inactive for long stretches of time, though their behavior when they do appear may be incredible stable. In order to circumvent this, we examine multiple flavors of stability:

- **Universal Stability**: A real value which quantifies how stable an individual's area is when viewed from the perspective of the entire network. Vertices are considered unstable if they do not appear in the graph regularly.
- **Conditional Stablility**: A real value which quantifies how stable an individual's area is when viewed from his or her perspective. Stability is measured without taking into account user inactivity.
- **Parameterized Stability**: A binary classification where an individual's area is described as stable or unstable based on some behavioral thresholds.

Each of these metrics have some drawbacks and advantages which will be explored as well as statistics showing the distribution of stability among users of the network.

The formalization and observation of stability within networks such as the blogosphere have various implications. First, any identification of stability will result in better models of the overall behavior of individuals in these networks. Better models will result in better results for predicting who will communicate with whom and how ideas will spread between them. Identification of stability also gives a global picture of expected behavior in a network. If, all of a sudden, distributions and statistcs change drastically, either for a stable individual or for the network as a whole, an observer would know that some significant event has happened with respect to the individual in question or the network as a whole.

II. Data

The popular blogging service LiveJournal has grown quite rapidly over the years. As of Janurary 2009, 18 million blogs have been created since the service's inception in 1999[4]. LiveJournal offers users the ability to create their own blogs, as well as declare friends and interests, join community centered groups, and discuss other user's blog posts through comments. Our research focuses on the Russian language subset of LiveJournal and uses commenting and posting activity to generate graphs representing weekly snapshots of the communications dynamics in the network. One of the advanced features of LiveJournal is an RSS feed which publishes newly created posts as they appear in the system. Our collection software records every post published in this feed and stores a record of each comment that appears in response to that post.

From this collected data, weekly snapshots of activity within LiveJournal are created. Weeks were considered to be a natural until of time based on the cyclic nature of activity levels in the data, where weekends result in much fewer comments and posting than weekdays. From the collected data for each week, a weighted, directed graph is created in the following manner. If an individual, A, comments on one of the posts of user B, the graph will contain an edge from A to B. The edges are weighted based on the

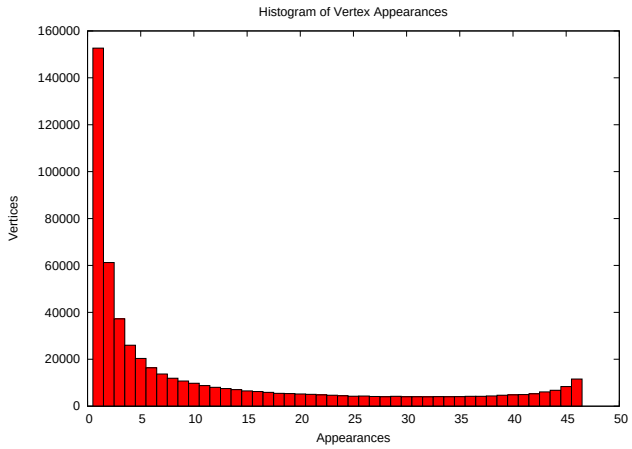


Fig. 2. A histogram showing how many vertices appear in a given number of weeks of the observed data.

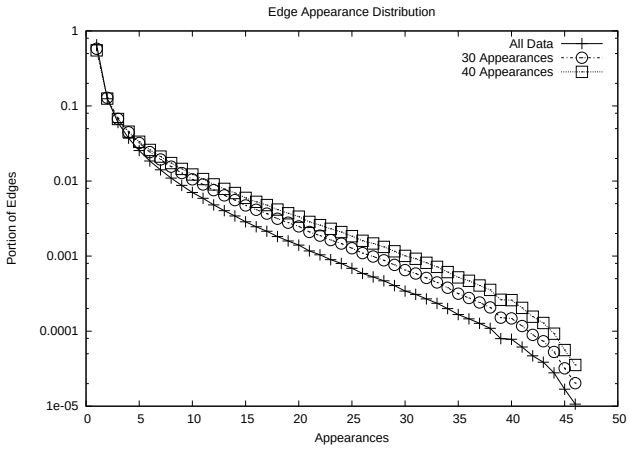


Fig. 3. Distribution of how often edges repeat for the various graph formulations.

number of posts of B user A has commented on. For example, if A comments 10 times on one of B's posts, the resulting edge will have weight 1. However, if user A comments once on 10 of B's posts, the edge from A to B would have weight 10. An example of this construction is given in Figure 1. Detailed descriptions of the statistics of these graphs are given in [1] and [3].

The most significant feature of the LiveJournal data collected is its instability. Over the 46 weeks of observed data, the presence of an individual in a given week's graph indicates that at the very least, the user in question posted a comment or received a comment from another user on one of his or her posts. If a user does not participate in either of these activities, they are an isolate in the graph formulated from the week's postings. These users are considered inactive for the given week. The ability of bloggers to become inactive from week to week causes difficulties in the search for stability. Figure 2 is a histogram showing how many vertices appear a given number of weeks in the observed data. A proportionally large number of individuals appear only a handful of weeks.

Due to this instability in vertex set, it makes sense to examine the stability of specific subsets of the vertices. Thus, in this analysis we examine graphs based on the data. First, we consider the set consisting of all vertices. This graph is constructed as described above. We also consider the subsets of vertices which appear at least 30 weeks and those which appear at least 40 weeks. This restricts the graph to only active and ultra-active individuals. In the graphs analyzed for vertices with at least 30 appearances and 40 appearances, edges from the original graph are dropped if they do not contain endpoints in the same set. This removes vertices below the appearance threshold from the analysis.

The connections between individuals in the data are also highly dynamic. For any week's graph, 60% of the edges will not be present in the next week's graph. Many of these edges do not reappear at any point in the observed 46 weeks. Figure 3 shows how often individual edges appear in the observed data for each of the three graphs described above. Note that the number of edges shown on the y-axis is a raw count in order to show that the values converge as values along the x-axis increase. Proportionally however, the graphs with appearance thresholds predictably contain a larger number of stable edges.

III. Area

In a network such as the one described above, the dynamics of the network complicate the search for stable behavior. In order to combat this, the notion of an *area* is defined. In [1], an area is defined as a region of the graph from which a user is more likely to reconnect its edges in the next evolution of the graph. Simply put, an individual's area in a given snapshot of the graph is defined by some criteria which indicates the belief that a user has more of a connection to individuals within their area than individuals in the rest of the graph. The criteria used to define this set can be varied and range from simple definitions such as the 1-neighborhood to complex definitions such as the union of social groups which contain an individual. This text will examine the stability of these two area definitions, an individual's one neighborhood and the union of their social groups, over all users in the observed networks.

The community detection algorithm used to locate social groups to use as areas is Iterative Scan, with input seeds determined by Link Aggregate. The specifics of the algorithm are described in [5]. The algorithm works by taking a set of seed communities and adding or removing vertices until each group is conditionally optimal with respect to some defined density function. For this application, the density function used was

$$density = \frac{e_{in}}{e_{in} + e_{out}} + \lambda e_p \tag{3.1}$$

where e_{in} is the number of edges within the community, e_{out} is the number of edges connecting vertices within the community to vertices outside, λ is a weighting parameter, and e_p is the edge probability within the cluster. The term involving edge probability was introduced to restrict community size in sparse graphs. The term discourages vertices from being added which are not significantly connected to the rest of the community being

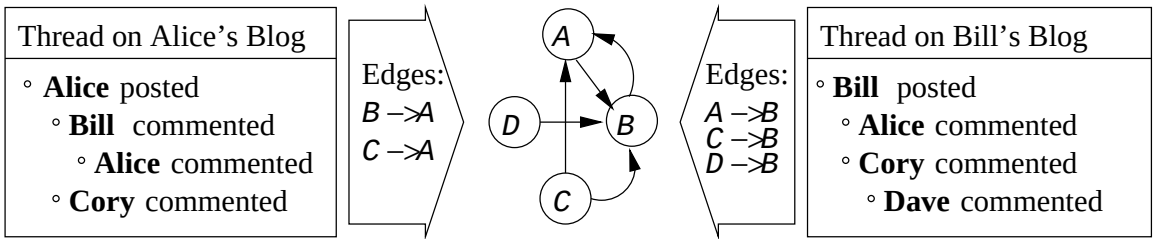


Fig. 1. Blogograph generation example. Vertices are placed for every blogger who posted or commented, the edges are placed from the author of the comment to the author of the post (the blog owner). Parallel edges and loops are not allowed.

λ	Clusters	Avg Size	AvgDensity
1	102124	2.0	1.12408
0.5	102121	2.00314	0.625279
0.25	101899	2.05851	0.383007
0.125	101309	2.65638	0.293328
0.0625	99838	5.80228	0.30531
0.03125	98086	10.1978	0.339646
0	95603	33.3089	0.379738

Fig. 4. Table showing cluster and area statistics as λ changes.

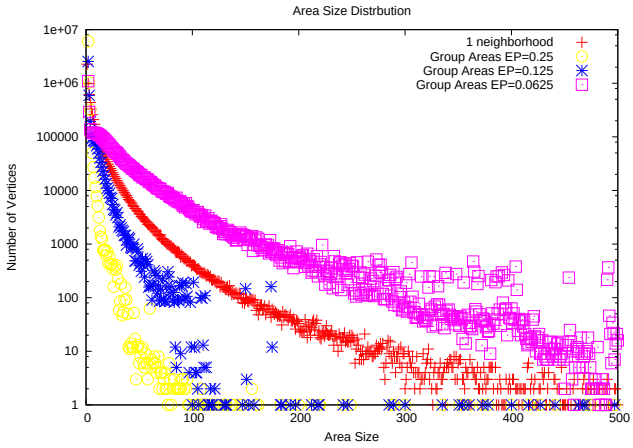


Fig. 5. Area size distribution for various areas. Note that if the 1neighborhood is used as a baseline, at some point, increasing λ causes the social group area definition's distribution of sizes to fall below that of the 1neighborhood.

optimized. This value can have a significant effect on the results of the algorithm. The average community sizes and densities are shown in Figure 4 for a week of the observed data for various values of λ.

Now, we consider two area definitions on the same graph. Figure 5 shows the distribution of area sizes for the areas defined by the 1-neighborhood and by the union of communities with various values of λ. The difference in the size distributions for different area definitions show that as the value of λ increases, areas get smaller. This is a consequence of favoring groups with high edge probability. As λ grows, the edge probability is weighted so heavily in the density function, that adding aditional vertices to a community of size 2 becomes incredibly difficult since a community of size two with an edge between the two vertices has a high edge probability. Figure 5 shows the distribution in detail. However, in the true distribution for 1 neighborhood, there is a huge tail extending out to an area size of 2500, while no group based areas approach that size. This is because when optimizing groups, the algorithm shies away from placing a vertex of degree 2500 in a group. Since the density function described penalizes for edges cut by the community boundary, adding a user with a high degree to a group requires that a large portion of vertices adjacent to the user already be in the group, which is unlikely barring the presence of intense community structure between the high degree vertex and all of his neighbors.

Defining an individual's area to be the union of his or her social groups does have a slight wrinkle to it in that occasionally, indidividuals do not belong to any social groups as found by the algorithm. In this case, this analysis cannot define an area for the person. Therefore, he or she is simply considered to be inactive in the graph. Figure 6 shows the number of vertices that have defined areas and the average area size for various area definitions. Notice again, that reducing the size of λ increases area size due to increased cluster size. Also, when many large clusters are discovered, the number of vertices with defined area in the analysis approaches the number of vertices when defining areas using 1-neighborhood. This value is the maximum possible, as every vertex which is observe must, at some point, be connected to some other vertex.

IV. Stability

Stability can be measured in a variety of ways. Since stability is, at its heart, the similarity between two sets, it makes sense to use the Jaccard index between vectors representing an individual's area in two time periods. Formally, the Jaccard index is given as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

where in this case, A and B are sets containing the vertices in an individual's area. This value will be 1 when the sets A and B are equal and is 0 if they do not overlap. This index can be utilized in different ways to capture some of the subtleties of stability.

The simplest notion of stability is that of *universal stability*. In this case, the stability is measured with respect to the rest of

Area Type	All Appearances	30 Appearances	40 Appearances
1-neighborhood	555853 (**6.691**)	89539 (**8.684**)	47851 (**9.654**)
Groups $\lambda = 1$	470326 (**2.319**)	89498 (**2.300**)	47840 (**2.290**)
Groups $\lambda = 0.5$	469972 (**2.323**)	89498 (**2.298**)	47840 (**2.288**)
Groups $\lambda = 0.25$	471074 (**2.452**)	89499 (**2.307**)	47839 (**2.287**)
Groups $\lambda = 0.125$	478218 (**4.382**)	89497 (**2.232**)	47839 (**2.816**)
Groups $\lambda = 0.0625$	505289 (**25.83**)	89506 (**28.52**)	47844 (**26.35**)
Groups $\lambda = 0.03125$	529049 (**156.0**)	89510 (**206.3**)	47843 (**217.4**)

Fig. 6. Table showing the number of vertices with defined areas in the observed 46 weeks of data. The average area size for all discovered areas is also shown in parentheses and bold.

the graph. That is to say, that if an individual does not appear in the graph in a given week, the network continues its normal behavior. From the network's perspective, the individual should be penalized for being unstable, even though the individual's area might be the same in all weeks in which he or she appears in the graph. Using the Jaccard index, we formalize the described situation as

$$S_U(t_0, t_n) = \frac{\Sigma_{i=0}^{n-1} J(A_{t_i}, A_{t_{i+1}})}{n-1}$$

From a practical standpoint, this value is perhaps the most indicative of stability. Given an arbitrary graph, it shows how stable one would expect an individual to be in the next graph. However, based on the appearance histogram shown in Figure 2, vertices in the blogoraph enter and leave the network regularly. Using this measure, a vertex with the same area over 10 weeks would be classified as unstable even though he or she embodies some notion of stability.

As a complement to universal stability, *conditional stability* can be used. In this measure, stability is measured from the user's perspective. Here, if the user does not appear in the network, the metric delays its similarity computation until the individual appears again. It then takes the average of the Jaccard index of these adjacent appearances. Consider the chronologically ordered set $T = \{t_0, t_1, t_2, ..., t_{n-1}, t_n\}$ of time steps in which a given vertex appeared in the network.

$$S_C(T) = \frac{\Sigma_{i=0}^{t-1} J(t_i, t_{i+1})}{|T| - 1}$$

These two stability metrics can be used together to help gain an understanding of the stability of an individual. A person with a high conditional stability value has a stable area when he appears in the network. A person with a high universal stability has a stable area as well a stable activity profile; he or she regularly is a part of the network. It would also be useful to identify stable sets of individuals with respect to certain parameters and to observe the size of the stable sets increase or decrease as the parameters change. This is the third notion of stability we will present. We identify a vertex as being stable if he or she has some number $T_{stable} \geq T_{thresh}$ of adjacent appearances with Jaccard index $J(t_j, t_{j+1}) \geq t$, where s_{thresh} and t are user defined parameters.

V. Results

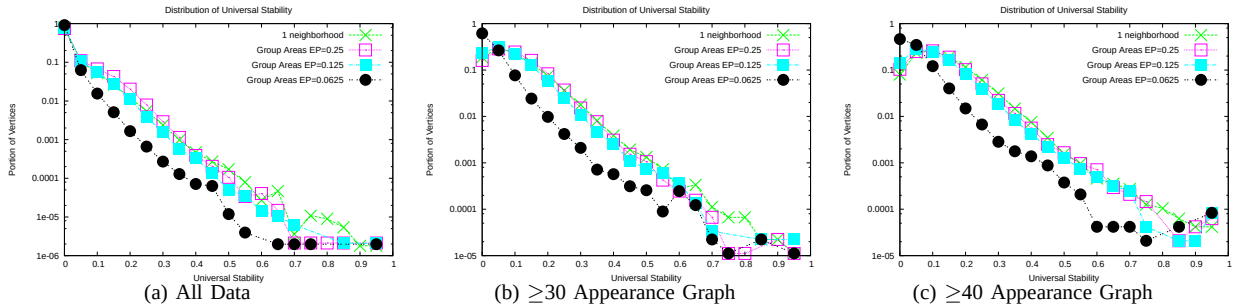
The universal stability results for each of the three graph formulations are shown below in Figure 7. As would be expected using such a strict measure, all of the stability values are fairly low for each of the different area definitions. For the 30 appearance and 40 appearance graphs, an increase in stability across all area definitions is seen. Again, this is expected, as universal stability will penalize vertices for not periods of inactivity. These graphs have had inactive vertices, to a certain extent, removed. The general trend using this metric is that 1-neighborhood appears to provide the most stability followed by the areas which are defined by taking the union of social groups with the smallest size.

Figure 8 shows the distribution of conditional stabilities. Here, a much different distribution is observed, as individuals many more individuals have higher stability due to the relaxation of metric. Using the union of social groups to define areas with this measure provides more stability for λ values producing smaller areas than the 1-neighborhood. Looking at the 30 and 40 appearance graphs, the difference in conditional stability defined by low λ group areas, high λ group areas, and 1-neighborhoods becomes even more significant.

Figures 9, 10, and 11 examine parameterized stability. They show, for each of the graphs, the number of vertices that are "stable" for a given number of weeks along the x-axis given one of three stability thresholds: 0.1, 0.3, 0.7. These values are meant to examine a low, medium, and high threshold requirement. Looking at Figure 9, using an area defined by the 1-neighborhood of an individual results in more stability across all thresholds, though using groups with a high value of λ as the area shows significant stability with a threshold of 0.3 for 5-15 weeks. In Figures 10 and 11 showing the 30 appearance graph and 40 appearance graph, this continues to be the case, except when the stability threshold reaches 0.7. Here, the union of groups with a high value of λ results in an a similar amount of stability in the 30 appearance graph and more stability in the 40 appearance graph than using the 1-neighborhood.

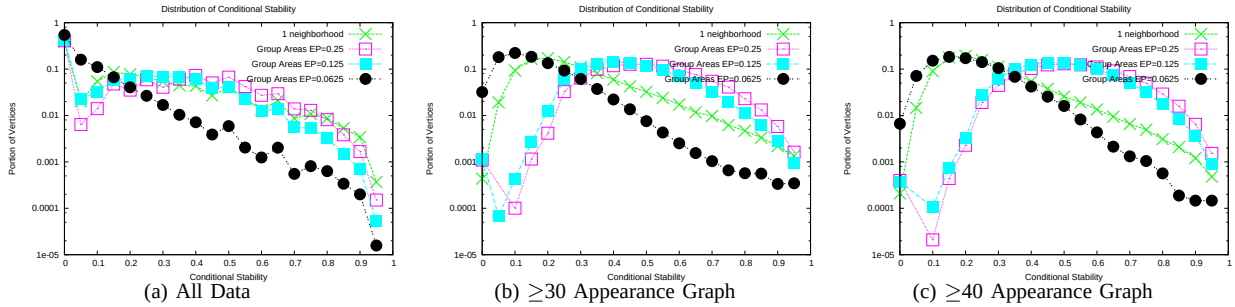
Based on all of this data, it is apparent that, with few exceptions, using areas as defined by the 1-neighborhood of an individual result in more stability. Given this, one can ask how the sets that are considered stable compare. Specifically, how well does the set of stable individuals found using 1-neighborhood cover the, generally smaller, set of stable individuals found using the union of social groups. Across all combinations of stable parameters, approximately 50% of stable individuals found in group based areas are also stable when using their 1-neighborhood as an area. This indicates that the 2 definitions find different sets of stable individuals.

Using parameterized stablility, the result of each analysis is a partitioning of the vertex set into "stable" and "unstable"



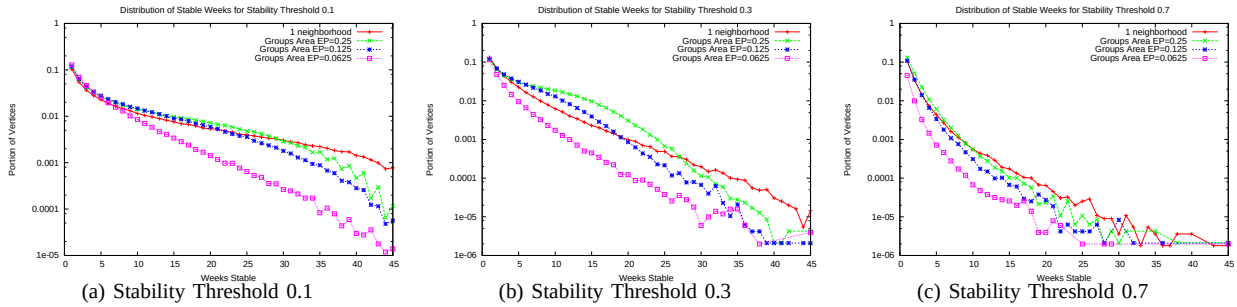
(a) All Data (b) $\geq$30 Appearance Graph (c) $\geq$40 Appearance Graph

Fig. 7. Three plots detailing the universal stability distribution across different area definitions on the 3 formulated graphs.



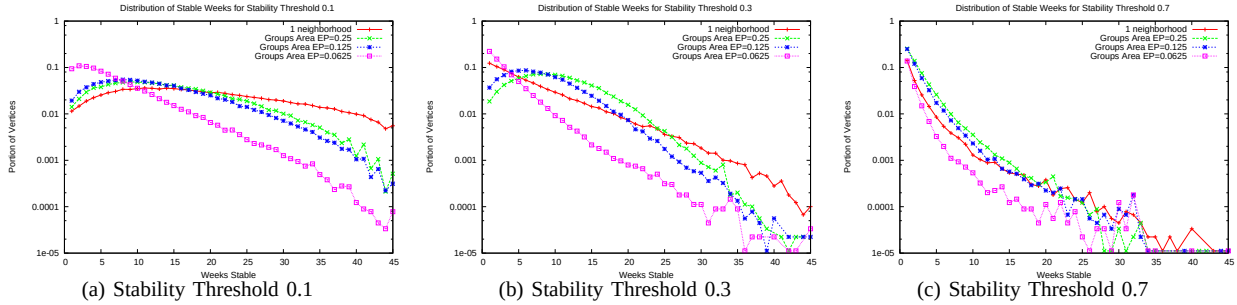
(a) All Data (b) $\geq$30 Appearance Graph (c) $\geq$40 Appearance Graph

Fig. 8. Three plots detailing the average conditional stability of 4 area definitions on the 3 formulated graphs. Subfigure (a) uses a logscale along its y-axis. For all graphs, using areas composed of the union of social groups which on average are smaller than the areas defined by 1neighborhood locate more stable vertices. The elimination of infrequent users in the 30 and 40 appearance graphs also results in a change in distribution shape, indicating an increase in the proportion of vertices with higher stability values.



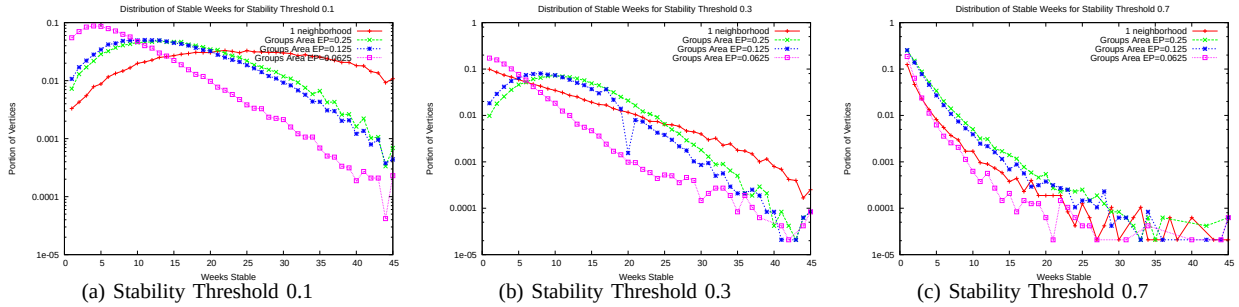
(a) Stability Threshold 0.1 (b) Stability Threshold 0.3 (c) Stability Threshold 0.7

Fig. 9. Plot showing the portion of vertices that appear stable for exactly the number of weeks listed along the x-axis for various stability thresholds. The underlying graphs used in this evaluation consist of all communication data.



(a) Stability Threshold 0.1 (b) Stability Threshold 0.3 (c) Stability Threshold 0.7

Fig. 10. Plot showing the number of vertices that appear stable for exactly the number of weeks listed along the x-axis for various stability threshold. The underlying graphs used in this evalution consist only of individuals appearing at least 30 times in the observed data.



(a) Stability Threshold 0.1 (b) Stability Threshold 0.3 (c) Stability Threshold 0.7

Fig. 11. Plot showing the number of vertices that appear stable for exactly the number of weeks listed along the x-axis for various stability threshold. The underlying graphs used in this evalution consist only of individuals appearing at least 40 times in the observed data.

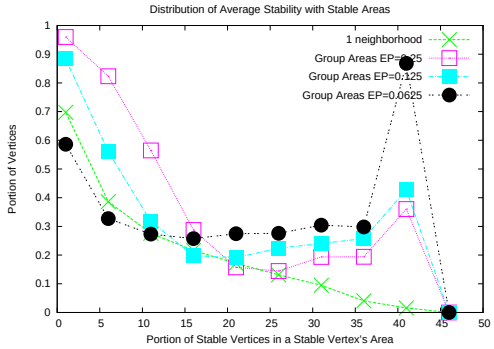


Fig. 12. A plot showing the average portion of vertices in a stable vertex's area which are also stable.

vertices with respect to the specified parameters. How these two sets of vertices relate gives additional insight into the stability and character of the networks. Looking at each stable individual, the portion of his or her areas composed of stable vertices can be computed. A high portion of stable vertices in a stable individual's area indicates the existence of a pocket of stablility within the graph. As this number becomes larger relative to the same portion computed over unstable vertices, the degree to which the stable area of the graph is isolated increases. Figure 12 shows a set of distributions of this portion for different area definitions. As shown in the previous plots, as the value of λ decreases the number of stable vertices decreases, but the size of their area increases. However, in this figure, we see an increase in the portion of stable vertices in a stable individual's area as λ decreases for large numbers of stable weeks.

VI. Conclusion

In highly dynamic networks such as the blogograph, understanding the nuances of behavioral stability is a difficult task. Using the concept of an individual's area, we are able to observe how stability distributions change under different stability metrics. Once these metrics have been applied to each data set, we can then consider how individual stability compares under definitions and how stable individuals interact with unstable individuals. In addition to analyzing the full observed data, we have also examined the graphs composed of only active and ultra-active individuals in an attempt to study the active "core" of the graph. Such analyses are important as they provide a set of observations which can be used to enhance current models of dynamic network behavior used in link prediction, diffusion, etc.

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.

References

[1] M. K. Goldberg, M. Magdon-Ismail, S. Kelley, and K. Mertsalov, "A locality model of the evolution of blog networks," in *ISI*. IEEE, 2008, pp. 191–193.

[2] M. Goldberg, M. Magdon-Ismail, S. Kelley, K. Mertsalov, and W. Wallace, "Communication dynamics of blog networks," in *The 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008)*.

[3] M. Goldberg, S. Kelley, M. Magdon-Ismail, and K. Mertsalov, "Stable statistics of the blogograph," in *Interdisciplinary Studies in Information Privacy and Security*, 2008.

[4] Http://www.livejournal.com.

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