Communication Dynamics of Blog Networks

Mark Goldberg
CS Department, RPI
110 8th Street
Troy, NY
goldberg@cs.rpi.edu

Stephen Kelley
CS Department, RPI
110 8th Street
Troy, NY
kelles@cs.rpi.edu

Malik Magdon-Ismail
CS Department, RPI
110 8th Street
Troy, NY
magdon@cs.rpi.edu

Konstantin Mertsalov
CS Department, RPI
110 8th Street
Troy, NY
mertsk2@cs.rpi.edu

William (Al) Wallace
DSES Department, RPI
110 8th Street
Troy, NY
wallaw@rpi.edu

ABSTRACT

We study the communication dynamics of Blog networks, focusing on the Russian section of LiveJournal as a case study. Communications (blogger-to-blogger links) in such online communication networks are very dynamic: over 60% of the links in the network are new from one week to the next, though the set of bloggers remains approximately constant. Two fundamental questions are: (i) what models adequately describe such dynamic communication behavior; (ii) how does one detect changes in the *nature* of the communication dynamics. We approach these questions through the notion of stable statistics. We give strong experimental evidence for the fact that, despite the extreme amount of communication dynamics, several non-trivial aggregate statistics are remarkably stable. We use stable statistics to test our models of communication dynamics: any good model should produce values for these statistics which are both stable and close to the observed ones. Stable statistics can also be used to identify phase transitions, since any change in a normally stable statistic indicates a substantial change in the nature of the communication dynamics.

Our model for the communication dynamics in large social networks is based on the locality of communication: a node's communication energy is spent mostly within it's local social "area." By varying the definition of a nodes' social area, our model can be used for a variety of social networks. Our results with different definitions of locality show that the best approximation to the stable statistics observed on the blog network supported by LiveJournal occurs when the social locality is defined as the union of clusters (social groups) containing the node, and when nodes communicate within their locality using a preferential attachment strategy.

The 2nd SNA-KDD Workshop '08 (SNA-KDD'08) August 24,2008 Las Vegas, Nevada, USA

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences; H.1 [**Models and Principles**]: General

Keywords

Social Networks, Blogs, Modeling, Community Detection, Community Evolution, Stable Statistics.

1. INTRODUCTION

The structure of large social networks, such as the WWW, the Internet, and the Blogosphere, has been the focus of intense research during the last decade (see [1, 7, 8, 12, 17, 19, 20, 21, 22]. One of the main foci of this research has been the development of dynamic models of network creation ([2, 11, 22, 18]) which incorporates two fundamental elements: network growth, with nodes arriving one at a time; and some form of preferential attachment in which an arriving node is more likely to attach itself to a more prominent existing node than a less prominent one one (*the rich get richer*).

Once a network has grown and stabilized in size, how does it evolve? Such an evolution is governed by the communication dynamics of the network: links being broken and formed as social groups form, evolve and disappear. The *communication dynamics* of these networks have been studied much less, partially because the typical networks studied (the WWW, the Internet, collaboration networks) mainly exhibit growth dynamics, and not communication dynamics. Clearly, as a network matures, the growth (addition of new users) becomes a minor ingredient of the total change (see Figure 1). Further, links in a socially dynamic network such as the Blogosphere should not be interpreted as static. The posts made by a blogger a week ago may not be reflective of his/her current interests and social groups. In fact, blog networks display an extreme communication dynamics. Over the 20 week period shown in Figure 1, in a typical week, 510,000 pairs of bloggers communicated via blog comments. Out of those about 380,000 are between pairs of bloggers who *did not* communicate the week before, i.e. over 70% of the communications are new. What models adequately describe the dynamics of the communications in such networks which have more or less stabilized in terms of growth?

To begin to address this question, one must first develop

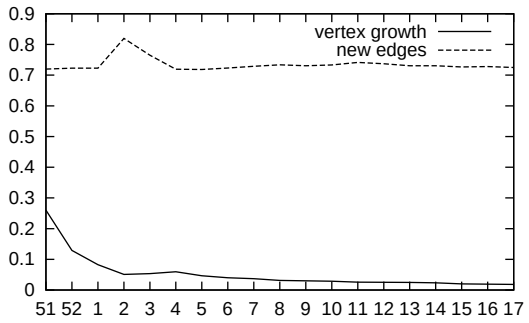

Figure 1: Edge and vertex dynamics. Clearly the rate of growth is decreasing however the fraction of new edges which appear in a week remains approximately constant at over 70%.

methods for testing the validity of a model. In such an environment of extreme stochastic dynamics, one cannot hope to replicate the dynamics of the individual communications, but rather the evolution of interesting macroscopic properties of the communication dynamics. Particularly interesting macroscopic properties are those which are time invariant. We refer to such properties as *stable statistics*. As we demonstrate, even in such an active environment, certain statistics are remarkably stable. For example: the power-law coefficient for the in-degree distribution, the clustering coefficient, and the size of the giant component (see Table 1).

1.1 Our Contributions

Our goal is two-fold. First, to demonstrate experimentally that a (non-exhaustive) set of non-trivial statistics are in fact stable in the Russian section of LiveJournal. Such stable statistics may then be used to validate models, characterize networks and identify phase transitions (testing for when the model changes). Second, to present a locality based model for communication dynamics. We show (through simulation) that our model stabilizes to an equilibrium in which the aggregate statistics of the communication dynamics aggregate are stable. Further, among the set of models we tested, the values for of the stable statistics from our model best reproduces the statistics.

Stable Statistics.

Our case study was the Russian section of LiveJournal. Over an observed period of 20 weeks, roughly 153,000 users are active in any week period. The size of this set is quite stable (changes typically from 1 to 2%), although the makeup of the set changes drastically from week to week. Surprisingly many aggregated statistics computed for the Blogograph show strong stability. Among those stable statistics are: the distribution of the in-degrees and the out-degrees of the nodes; the (overlapping) coalition distribution as described by the cluster density and size; and, the coalition lifespan distribution.

The nodes of the Blogograph represent bloggers and the directed edges between them represent all pairs $\{A, B\}$ where blogger A visited the blog of B during the week in question, and left a comment to a specific post already in the blog. We consider the following five types of stable statistics:

(i) **Individual Statistics**: properties of individual nodes such as the in-degree and out-degree distributions for the graph

(ii) **Relational Statistics**: properties of edges in the graph such as the persistence of edges and clustering coefficients

(iii) **Global Statistics**: properties reflecting global information such as the size and diameter of the largest component and total density

(iv) **Community Statistics**: properties relating to group structure such as the community size and density distributions

(v) **Evolution Statistics**: properties related to the evolution of graphs such as the average lifespan of communities

The purpose of collecting these statistics is two-fold. First, they create a baseline which describes the normal behavior of individuals, communities, and the network as a whole. Once this base has been established, anomalous behavior at each of these levels can be identified and investigated further. Second, stable statistics can be used for testing any model of the network dynamics, as any model which attempts to replicate the communication dynamics must, in particular, be able to reproduce these statistics. Furthermore, the quality of a model can be measured by how well the statistics computed from the network generated by the model (in equilibrium) replicate those observed in the real network.

Locality based Models of Communication Dynamics.

Existing growth-based models fail to adequately replicate the observed stable statistics, as they do not capture communication dynamics. We consider models for communication dynamics which take as input: (a) The current (observed) communication graph; and, (b) each user's out-degree (communication energy) at the next time step (or a distribution over for the user's out-degree). These two inputs are standard for existing growth models (such as the preferential attachment growth model). Such models are only applicable when the communications are open (observable to all nodes). The output is the communication graph at the next time step, based on the model for probabilistic attachment of each node's out-edges.

We discuss intuitive extensions of growth models for modeling communication dynamics and illustrate that these extensions are inadequate for modeling the observed stable statistics. We present a locality based model which relies on two fundamental principles to more accurately reflect the observed communication dynamics. First, our concept of *locality* reduces the set of nodes a node can attach to in the next time step (a week in our case). This locality is based on structural properties of the current (observable to all) communication graph. The locality represents a semi-stable set of "neighbor" nodes that an individual is highly likely to connect to, and can be interpreted as that individuals view of the communities she belongs to. We test various structural (graph theoretic) definitions of a node's social locality, ranging from trivial localities such as the entire graph to

notions of a node's neighborhood (e.g. the 2-neighborhood; the clusters to which a node belongs). Second, after obtaining a node's locality, one must specify the *attachment* mechanism, the mechanism used by the individual to select the nodes in her locality to which she will connect at the next time step. We test a number of different attachment mechanisms which one could consider, ranging from uniform attachment to some form of preferential attachment. Thus, we present results using each of the various choices for the locality and attachment mechanism.

Such probabilistic models are Markov chains, and we test a model's performance by comparing the values it produces for the stable statistics after it has equilibrated. We find experimentally that the mixing times are small and the equilibrium statistics are independent of the starting state (the chains are ergodic), hence the equilibrium distribution is unique. Our results indicate that our locality based model with locality defined as the union of clusters to which a node belongs and a preferential attachment mechanism produces the best values for the stable statistics.

2. CLUSTERS

The notion of a social community is crucial to our model of a Blog network. The underlying idea of our model is that every user selects the nodes to visit (to leave a comment) from the set of nodes that belong to a relatively small "area" of a node. Our experiments with different definitions of the local area of the node show that the best approximation to the observed statistics is achieved if the area is taken as the union of *clusters* containing a given node. Our definition of network clusters is borrowed from [4, 5, 6] with an important specification of the notion of the density of a set of nodes in a network.

Definition. Given a graph $G(V.E)$ let function D, called the *density*, be defined on the set of all subsets of V. Then, a set $C \subseteq V$ is called a cluster if it is locally maximal w.r.t. D 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 $D(C)$.

The idea of the definition matches the common understanding of a social community as a set of members that forge more communication links within the set than that with the outside the set. The function D is not specified by the definition, but its precise formulation is crucial in "catching" the nature of social communities. The density function considered in [3] is as follows:

$$D(C) = \frac{w_{in}}{w_{in} + w_{out}}, \tag{1}$$

where w_{in} is the number of edges xy with $x, y \in C$ and w_{out} is the number of edges xy with either $x \in C$ & $y \notin C$ or $x \notin C$ & $y \in C$ (to allow for directed graphs). The main deficiency of the definition of a cluster as a computational representation of a social community is that it is easy to find examples of networks that permits very large and loosely connected clusters, that intuitively are not representing any community. The idea of our modification of 1 is to introduce an additional parameter which represents the edge probability in the set

$$D(C) = \frac{w_{in}}{w_{in} + w_{out}} + \lambda \frac{2w_{in}}{|C|(|C| - 1)}, \tag{2}$$

where the parameter λ depends on the specific network under the consideration, and is supposed to be selected by the researcher. For our experiments, we selected $\lambda = 0.125$.

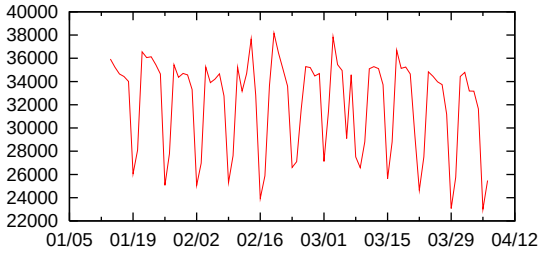

Figure 2: Number of posts per day that appeared between January 14, 2008 and April 6th, 2008. The periodic drops in the number of posts correspond to Saturdays and Sundays.

3. DATA

We define the blogograph as a directed, unweighted graph representing the communication of the blog network within a fixed time-period. There is a vertex in the blogograph representing each blogger and a directed edge 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. Loops, comments on a bloggers own blog, are ignored as well. To study the communication dynamics, we consider consecutive weekly snapshots of the network; the communication graph contains the bloggers that either posted or commented during a week and the edges represent the comments that appeared during the week. We chose to split graphs into one week periods due to highly cyclic nature of activity in the blogosphere (see Figure 1 and Figure 2). An illustration of the blogograph's construction is given on Figure 3.

The data used for our research was collected from the popular blogging service LiveJournal. As of May 2008, there are more than 15 million user for the whole network; the number of posts during a 24 hour period is approximately 191,000 (see html://www.livejournal.com/). Much of the communication in LiveJournal is public, which allows for open access. LiveJournal provides a real time RSS update feature that publishes all open posts that appear on any of the hosted blogs. In our experience, the overwhelming majority of comments appear on these posts 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 580,000 bloggers out of the total 15 million) 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 in Russian. The network of Russian bloggers is very active. On average, 32% of all posts contain Cyrillic characters. LiveJournal blogging has become a cultural phenomenon in Rus-

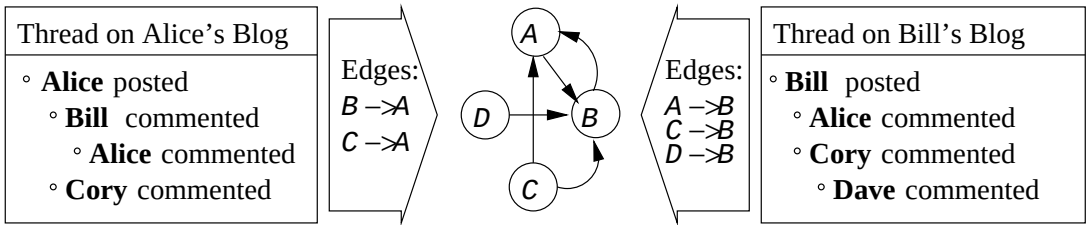

Figure 3: Blogograph generation example. Vertices are placed for every blogger who posted or commented, the edges are placed from the author of the comment to the author of the post (the blog owner). Parallel edges and loops are not allowed.

week	$\|V\|$	$\|E\|$	GC	C	d	α
49	155,615	530,160	95.88%	0.0639	5.333	2.63
50	156,026	532,189	95.91%	0.0644	5.327	2.66
51	155,093	527,364	95.62%	0.0635	5.316	2.65
52	151,559	516,483	95.62%	0.0635	5.316	2.71
1	118,979	327,356	93.55%	0.0573	5.777	2.92
2	142,478	444,457	95.14%	0.0587	5.392	2.68
3	159,436	559,506	96.16%	0.0629	5.268	2.68
4	158,429	550,436	95.60%	0.0631	5.224	2.67
5	156,144	534,917	95.49%	0.0627	5.293	2.72
6	156,301	526,194	95.70%	0.0615	5.338	2.72
7	154,846	523,235	95.44%	0.0622	5.337	2.69
8	156,064	528,363	95.59%	0.0609	5.320	2.69
9	156,362	524,441	95.58%	0.0602	5.377	2.68
10	154,820	523,304	95.48%	0.0593	5.368	2.68
11	155,267	516,280	95.13%	0.0600	5.356	2.68
12	156,872	514,269	95.20%	0.0590	5.367	2.63
13	155,338	510,070	95.42%	0.0601	5.342	2.71
14	155,099	506,892	95.19%	0.0607	5.309	2.73
15	153,440	504,850	95.32%	0.0601	5.303	2.73
16	154,012	512,094	95.34%	0.0599	5.298	2.60
17	151,427	503,802	95.30%	0.0611	5.288	2.75

Table 1: 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)

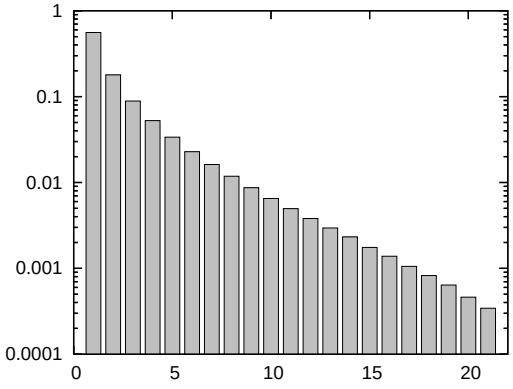

Figure 4: Edge stability: distribution of number of weeks an edge appeared in. 60% of all edges only appeared once.

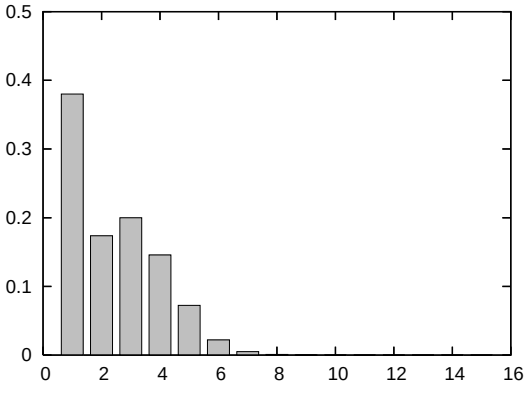

Figure 5: Edge history: the distribution of the shortest undirected distance of end points of an edge in a previous time cycle.

sia. Discussion threads often contain intense and interesting discussions which encourage communication through commenting. Our work is based on data collected during between December 2007 and April 2008. The basic statistics about the size of obtained data are presented in Table 1. A simpler set of statistics on a smaller set of observed data is presented in [16].

4. STABLE STATISTICS

The observed communication graph has interesting properties. The graph is very dynamic on the level of nodes and edges but has stable aggregated statistics. About 75% of active bloggers will also be active in the next week. Further, about 28% 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. The stability of various statistics of the blogograph is presented in Table 1. The giant component (GC) is the largest connected (not necessarily strongly connected) subgraph of the undirected blogograph. A giant component of similar size has been observed in other large social networks [18], [14]. 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 observed value in the blogograph is similar to what has been found in many other social networks ([18], [23]).

Many large social networks ([2], [14]) display a power law in the degree distribution, $P(k) \propto ck^{-\alpha}$, where $P(k)$ is the probability a node has degree k. Figure 6 shows the mean in-degree distribution of the collected blogographs. In these graphs, we observed power law tail with parameter $\alpha \approx 2.70$ which is stable from week to week. This value was computed using maximum likelihood method described in [10] and Matlab code provided by Aaron J. Clauset.

To evaluate the dynamic in the observed communication we considered the change in the set of links or edges from one week to another. Figure 4 shows the distribution of number of weeks a particular pair of bloggers communicated. It is evident from this plot that vast majority of communication does not re-occur, yet some links reappear every week. We also looked at the past relationship between the bloggers who communicated. We defined the history of the edge (i, j) that appeared in time cycle t to the the shortest undirected distance between i and j in the graph of the time cycle $t-1$. Figure 5 presents the distribution of the edge histories of all observed edges of all time cycles. The edge history distribution of the particular observed weeks is very close to the presented distribution (the variation at each point is less then 2%). As figure suggests, the majority of communicating vertices were less or at 3 hops away in the network on the previous time cycle. This provides evidence for the strong locality of communication that occurs in the observed network.

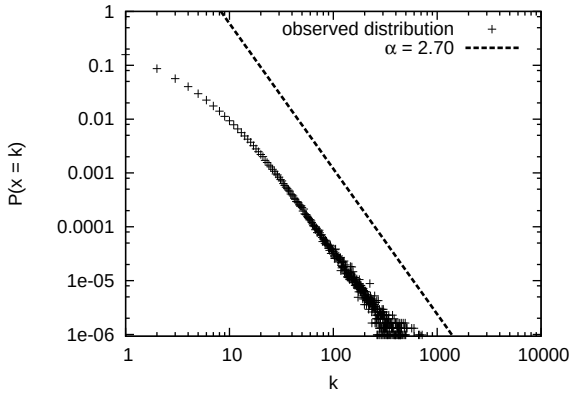

Figure 6: Average in-degree distribution in the blogograph observed over 21 weeks from Dec. 03, 2007 and Apr. 28, 2008

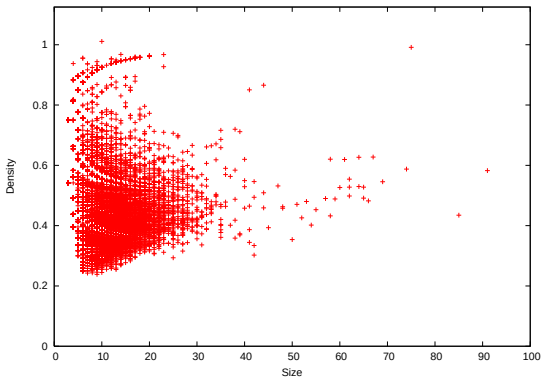

Figure 7: A size vs density plot for week 5 of the observed data. The x-axis is a measure of the community size while the y-axis shows the value of δ. Each point represents a community.

In addition to looking for stability in structural statistics, it is also useful to examine stable community behavior. Using the notion of clusters discussed previously in this text, we find locally optimal communities using each edge in the graph as a seed. Once all seeds are optimized, duplicates and clusters of size 2 are removed. Statistics of the remaining clusters are showing in Table 2. A size vs density plot is also given in Figure 7. The general shape and scale of this plot is replicated across all observed weeks.

Since the evolution of communication dynamics is being examined, it makes sense to also consider the evolution of communities. We define evolution as follows. Initially, the Iterative Scan algorithm optimizes a set of seed communities. In this case, we seed using every edge in the graph. After optimization, communities of size 2 and duplicate communities are filtered out. The resulting communities are then placed in the next graph as seeds.

Since a cluster may become disconnected when placed into the next graph, we begin optimization on the largest con-

week	$\|C\|$	avg size	δ_{avg}	e_p
51	19631	10.0183	0.456677	0.253212
52	19520	10.0615	0.453763	0.252101
1	23187	10.0915	0.473676	0.248130
2	20970	9.98412	0.458161	0.251843
3	17986	9.86184	0.448757	0.254203
4	18510	9.71891	0.453578	0.257481
5	18808	9.88255	0.455823	0.254305
6	19318	9.79242	0.454656	0.253901
7	19343	9.80381	0.456364	0.255236
8	19796	9.83113	0.453577	0.252818
9	20136	9.95401	0.473693	0.252607
10	19670	9.71678	0.45449	0.255778
11	20212	9.66842	0.456908	0.256098
12	20415	9.70331	0.461118	0.255819
13	20030	9.78058	0.455676	0.254681
14	19893	9.74936	0.455234	0.254384
15	19392	9.73407	0.455365	0.254687
16	19113	9.74787	0.454531	0.254721
17	18737	9.72333	0.455775	0.255658

Table 2: 19 weeks of communities from the Russian section of Live Journal. $|C|$ is the number of communities, δ_{avg} is the average density, and e_p is the average edge probability within the communities.

lifespan	week2	week3	week4
1	0.9895	0.9918	0.9869
2	0.009633	0.007561	0.01156
3	0.0008584	0.0006672	0.001243
4	0.000004768	0	0.0001607
5	0	0	0.0001081
6	0	0	0.00005403

Table 3: A table showing the lifespan in weeks of communities "born" in a given week. The values are a normalized portion of all communities initially discovered in the indicated week. Lifespan is simply the number of consecutive weeks the community is considered to be "alive" as defined previously in this text.

nected component in the new graph. We consider a community to be alive if

$$\frac{C_T \cap C_{T+1}}{C_T \cup C_{T+1}} \geq t$$

where C_T is the optimized community at time-step T and t is a threshold value. The threshold used for this paper is $\frac{1}{3}$, which corresponds to half of the vertices from time-step T being in the community at time-step $T+1$ if there is no change in community size. If this threshold is not reached, the community is considered dead.

In Table 3, we show the lifespan distribution for communities which are born in 3 weeks of the observed data. From these results, it can be seen that the number of communities which persist for longer than a few weeks is quite small. This relatively expected due to the changes in the node set covered by the graph on a week to week basis, which, as stated previously, may result in 25% of the graph being different in one week from the next.

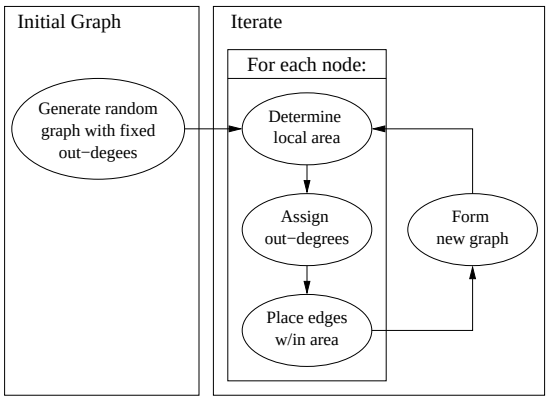

Figure 8: Model execution flow

5. MODELING

As previously stated, networks with such strong communication dynamics have not been well modeled. Much of the previous work aims to replicate the growth phase of a network's life-cycle, ignoring the evolution of communication once the network's size stabilizes. Models which replicate these dynamics would be useful as a sand-box within which social hypotheses on information diffusion, the emergence of leaders, and group formation and dissolution can be tested. To be considered useful, any model should create a set of graphs whose statistics come as close as possible to mirroring the statistics of the observed data presented previously.

Before delving into the creation of a new model, let us first consider the modification of a previously existing one. The simplest method of producing a set of evolving graphs is to grow each week's graph using a known network growth algorithm. Vertices can be assigned an out-degree based on the observed data and connected to each other via preferential attachment for each of the weeks. If done correctly, this would yield a set of graphs whose in-degree and out-degree distributions come close to matching the observed data's power law distributions.

Despite this initial positive result, examining the rest of the statistics demonstrates that the model is insufficient. Relational statistics such as edge stability, edge history, and clustering coefficient all significantly depart from the observed values, which we will show in detail further in the paper. This model's inability do recreate these statistics is expected, since it generates each graph independently.

Below, we propose a model which performs its edge connection within some *locality* in an effort to more closely mirror the edge stability, edge history, clustering coefficient, and community based statistics of the network.

6. A LOCALITY BASED MODEL

The goal of our model is to produce a sequence of graphs which simulate the connection and reconnection of vertices. Our model specifies how nodes update their edges in response to the observed communication activity. In specifying this model of evolution, we take as input the out-degree distribution of the blogograph. The justification for this is that, while the out-degree distribution would be an interesting object to model, it mainly reflects the individual properties of the users in the network such as the level of energy

Algorithm 1 Evolution Model.

Function **Model** (T, **OutDeg**, **Area**, **Prob**)
// **Output:** Blogographs $G_1, \ldots, G_T$.
$\{k_0^1, \ldots, k_0^n\} \leftarrow$ **OutDeg**
Initialize G_0 (e.g. to a random graph)
for $t = 1$ to T **do**
  $E_t \leftarrow \emptyset$; $\{k_t^1, \ldots, k_t^n\} \leftarrow$ **OutDeg**
  for $i = 1$ to n **do**
    $A_t^i \leftarrow$ **Area**(i, G_{t-1}); $p_t^i \leftarrow$ **Prob**(i, A_t^i, G_{t-1})
    $E_t^i \leftarrow$ **Attach**$(i, , A^i, p_t^i, k_t^i)$; $E_t \leftarrow E_t \cup E_t^i$.
  end for
  $G_{t+1} \leftarrow (V, E_t)$
end for

Algorithm 2 Edge attachment algorithm.

Function **Attach** (i, A_t^i, p_t^i, k_t^i)
// **Output:** E_t^i: edges in G_t originating at i
while $k_t^i > 0$ **do**
  if $(\sum_{v \in A_t^i} p_t^i(v) > 0)$ **then**
    Select node $v \in A_{t-1}^i$ with probability $p_t^i(v)$
    $p_t^i(v) \leftarrow 0$; renormalize p_t^i
  else
    Select node $v \in V \setminus A_{t-1}^i$ with uniform probability
  end if
  $k_t^i = k_t^i - 1$
  $E_t^i = E_t^i \cup (i, v)$
end while

and involvement of the user. Such quantities tend to be innate to a user. Different people have different social habits; some manage to communicate with hundreds of people while others interact with only a small group. Hence, out-degrees should be specified either *ab initio* (e.g. from social science theory) or extracted directly from the observed data. We will take the latter approach to specifying the out-degree distribution when it comes to testing our model. An early version of this model with preliminary results is presented in [15].

Given the out-degrees for all nodes, the task is now to specify how to attach the out-edges of the nodes and to obtain the in-degree distribution. It is the in-degree distribution that characterizes the global communication structure of the network (for example, who is considered by others to be important). Clearly, the out-degree distribution of a graph alone does not determine its in-degree distribution. Algorithms for generating undirected random graphs with a prescribed degree distribution are well known (see [9, 13, 24]). However, even if those algorithms are expanded to the domain of directed graphs, they will still be insufficient for our purpose of modeling evolution, which requires repeated generation of the next graph given the previous one.

To summarize, we are interested in models which reproduce the observed evolution *given* the out-degrees of the nodes. Thus, all our locality models assume that a node when deciding where to attach its communication links, has some fixed budget of emanating edges which it can attach. The main task of our model is to develop an evolution mechanism that re-creates an in-degree distribution close to the observed one.

We will use standard graph theory terminology in describing our model (see for example [25]). The sequence of blogographs are represented by directed graphs $G_0, G_1, G_2, \ldots$, where at every time step t, $G_t = (V, E_t)$. V is the common vertex set of all known bloggers, $V = \{v_1, ..., v_n\}$. An edge (v_i, v_j) is in the edge set E_t if blogger v_i commented on a post by v_j during the time period t. One time period covers one week, which appears to be the natural time scale in the blogosphere.

The input to the model is the set of out-degrees at time t for each vertex, $\{k_t^1, \ldots, k_t^n\}$ and G_{t-1}, the blogograph at time $t-1$. The output of the model is G_t, the blogograph at time t. Our model is locality based. At time t, every node v_i identifies its area, and assigns it out-edges with destinations in its area.

More formally, denote the area of v_i at time t by $A_t^i \subseteq V$. A_t^i represents the locality of node v_i at time t. Typically, a node's locality at time t will depend on G_{t-1}, the blogograph at time $t-1$. The attachment mechanism is probabilistic for each node. Node v_i attaches its k_t^i out-edges according to its own probability distribution p_t^i, where $p_t^i(v)$ specifies the probability for node v_i to attach to node v for $v \in V$. The probability distribution p_t^i may depend on A_t^i and G_{t-1} (e.g. higher degree nodes may get higher probabilities). In particular, we assume that $\sum_{v \in A_t^i} p_t^i(v) = 1$, which corresponds to the assumption that every nodes expends all its communication energy within its local area. Since we do not allow parallel edges, if $k_t^i > |A_t^i|$, it is not possible for node v_i to expend all its communication energy within its local area A_t^i. In this case, we assume that $k_t^i - A_t^i$ edges are attached uniformly at random to nodes outside its area and the remaining edges are attached within its area. The precise algorithm for distributing the edges given the probability distribution p_t^i is given in Algorithm 2.

The evolution model is illustrated in Figure 8. In more detail, the evolution model first obtains the out-degrees (which are exogenously specified). From G_{t-1}, it computes A_t^i and p_t^i for all nodes $v_i \in V$. For all nodes, it then attaches edges according to Algorithm 2. This entire process is iterated for a user specified number of time steps. This process is given in Algorithm 1. The inputs to the model are the procedure **OutDeg** which specifies the out-degrees (assumed to be exogenous), the procedure **Area** which identifies the local areas of the nodes given the previous graph, and the procedure **Prob** which specifies the attachment probabilities according to the attachment model. We will now discuss some approaches to defining the areas and the attachment probabilities. When testing our model, we will also need the procedure for obtaining the out-degrees, which will be discussed in Section 7

6.1 Locality Models

A node expends the majority of it's communication energy within it's local area. This captures the intuition that people mostly communicate with in a small group that contains friends, family, colleagues, etc. We propose the following definitions of an area:

1. **Global.** Every node v_i is aware of the whole network, the local area of v_i is $A_t^i = V$ at every time period t.

2. **k-neighborhood.** The local area A_t^i of node v_i at time t consists of all v_j such that undirected shortest distance $\delta_t(v_i, v_j) \leq k$.

3. **Clusters.** Using the definition of a cluster presented earlier in this paper, we define the local area of a blogger as the union of all clusters which he/she is a member of. Intuitively, this restricts a blogger's activity to the set of individuals in groups which they have shown interest in previously.

6.2 Attachment Models

Given the local area A_t^i of the node v_i at time t, the attachment model describes the probability $p_{t+1}^i(v_j)$ of occurrence of an edge (v_i, v_j) at time $t+1$ for $v_j \in V$. We propose the following attachment modes:

1. **Uniform.** Node v_i attaches to any $v_j \in A_t^i$ with equal probability

$$p_t^i(v_j) = \frac{1}{|A_t^i|}$$

and for $v_j \notin A_t^i$, $p_t^i(v_k) = 0$.

2. **Preferential Attachment.** Node v_i attaches to any $v_j \in A_t^i$ with probability

$$p_t^i(v_j) \propto indeg_{t-1}(v_j) + \gamma \quad (3)$$

where $indeg_{t-1}(v_j)$ is the in-degree of vertex v_j in graph G_{t-1} and γ is a constant.

3. **Markov Chain.** To obtain the attachment probabilities for vertex v_i we simulate the particle traveling over undirected edges of graph G_t starting from the node v_i and randomly selecting edges to travel over until it arrives at first node $v_e \notin A_t^i$. Every time the particle arrives at some node $v_j \in A_t^i$, the counter c_j^i is incremented. After this simulation is repeated with out resetting the counters c_j^i, $\forall\ v_j \in A_t^i$ a number of times, we determine the attachment probability

$$p_t^i(v_j) \propto c_j^i$$

.

4. **Inverse distance.** Node v_i attaches to some node $v_j \in A_t^i$ with probability

$$p_t^i(v_j) \propto \frac{1}{\delta_{t-1}^{\rho}(i,j)} \quad (4)$$

where $\delta_{t-1}(i,j)$ is the shortest undirected distance between vertices v_i, v_j in graph G_{t-1} and ρ is a constant.

The combination of the locality model and attachment model specifies the evolution model that, given the out-degree distribution, will produce a series of graphs that represent the blogograph at different time periods.

7. EXPERIMENTS AND RESULTS

In this section we present the results of execution of few of the models and the evaluation of their performance.

To evaluate the performance of the models, we compare the sequence of graphs produced by the model to the sequence of graphs produced by the observed communication in LiveJournal. In particular, we compare the clustering coefficient, the size of the giant component, average separation between two nodes and in-degree distributions. To compare the in-degrees, we compute the point-wise difference of the normalized distributions. Formally, for each graph G_i we compute the normalized distribution $D_i(k) = \frac{k}{|V_i|}$, where k is the degree and $|V_i|$ is the number of vertices in the graph. The differences between distributions of observed graph G_o and generated graph G_g is

$$E = \sum_d |D_o(d) - D_g(d)|.$$

Notice, $E \in [0, 2]$ and lower value of E corresponds to a closer match.

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

Table 4: The stable parameters of graphs generated by various models compared to the parameters of the observed data.

To evaluate a particular model we execute enough iterations to let the model stabilize. We determine the stabilization by inspection of plots of the major parameters (including in-degree distribution, clustering coefficient, etc). Then, we compare the sequence of the graphs produced by the model after the stabilization to the sequence mined from LiveJournal.

Table 4 contains the results of execution of models with various combinations of local area and attachment mechanisms compared to average parameters of graphs of different observed weeks. Note, for observed data average parameter E is computed by comparing distributions of graphs corresponding to various observed weeks. Figure 9 compares the observed degree distribution to the ones generated by some of the best area/attachment combinations. As defined in Section 4, edge history conveys information about how close the end points of the observed edge were in the previous time cycle and therefore measures the significance of locality in the communications. Figure 10 compares the observed edge history with the edge histories produces by the best models. The following is the discussion of these results.

7.1 Global Area Model

First, we considered the model with global area where vertices are aware of and can connect to any other vertex in the network.

In the case of a uniform attachment, the resulting model is very similar to the Erdös-Rényi model. The in-degree distribution and other parameters generated by such model are predictably very different from the power law degree distribution in the observed graph.

Global area with preferential attachment strictly proportional to the in-degree of the vertices in the graph of the

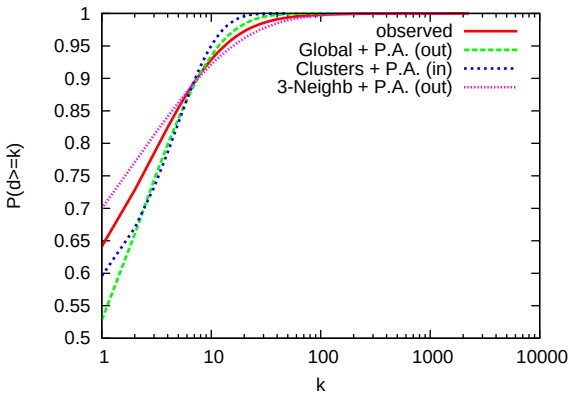

Figure 9: Cumulative in-degree distributions for various models

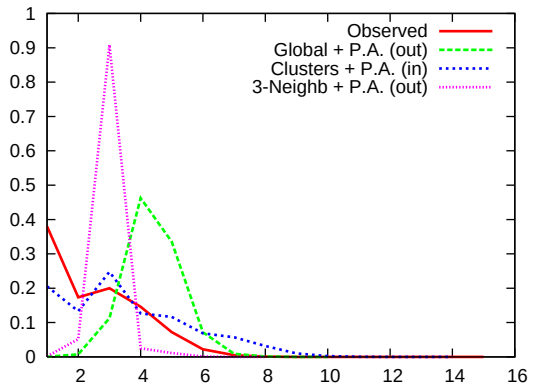

Figure 10: Edge history distribution of the various models and observed network

previous iteration results in a formation of a power house - small set of vertices with very high in-degree that attract all of out-degree. This effect is caused directly by preferential attachment; since vertices with zero in-degree will never be attached to, any vertex that receives no incoming vertices at some iteration will not receive any incoming vertices in any of the following iterations. Clearly, the graph with small set of vertices that attract all of the in-degree is very different from the observed graph.

The combination of global area and preferential attachment proportional to the out-degree of the vertices in the graph of the previous iteration produced results that were more similar to the observed network than the other global models, but the results were also significantly worse compared to models with other area definitions (k-neighborhood and union of clusters). Since this model allows for random selection of the end points of edges from the whole graph, the edge history (Figure 10) is very different from the one observed in the real network.

7.2 k-Neighborhood Area Model

We experimented with different values of k ($k \in \{2, 3, 4\}$) and determined that $k = 3$ produced the best models.

The combination of 3-neighborhood area and uniform attachment produced a model that showed mediocre results when compared to the observed parameters. The combination of 3-neighborhood area and preferential attachment proportional to the in-degree produced a graph with a small power house in just a few iterations. 3-neighborhood area with preferential attachment proportional to the out-degree produced a model that generated graphs with in-degree distributions very similar to the observed graph. In particular, the power law tail resembled the tail of the observed graph. The edge history of this model was quite different from the observed, since most of the end points for new edges are selected such that their distance in the previous iteration's graph was 3.

7.3 Union of Clusters Area Model

An area constructed via the union of clusters in combination with preferential attachment proportional to the in-degree produced in-degree distributions more similar to the observed than other attachment mechanisms combined with this area definition. This model also produced a sequence of graphs with an edge history that was closest to the observed as evident from Figure 10.

Models with this area definition were the only ones that produced non-trivial edge stability defined as the likelihood of a repetition of a recently observed edge. To evaluate this stability, we considered the number of edges that appeared more then once in 21 iterations of the model after stabilization. Models with global and 3-neighborhood area definitions that did not result in the formation of power houses produced a set of graphs such that less then 1% of edges that appeared more the once in all of the graphs. Models with an area defined by the union of clusters produced a set of graphs in which, on average, 14% of edges appear more then once in 21 iterations. In particular, a combination of this area definition with preferential attachment proportional to the in-degree produced a sequence of graphs with 18% of edges appear more than once, while in the observed network, 40% of edges (see Figure 4) appear more then once during 21 observed weeks.

After considering all of the parameters of the models, we determined the combination of an area defined by the union of clusters with preferential attachment proportional to the in-degrees of the vertices to be the best model to describe the dynamics of communication in the observed network.

8. CONCLUSION

We have presented a set of statistics which display strong stability even for such a dynamic network as the blogosphere. Our list of stable statistics is not exhaustive however they represent a comprehensive set of interesting properties of a network that any model for communication dynamics should capture.

Our experiments have shown that the communication dynamics of large social networks is best explained as a result of *local* communication, where the majority of members communicate within their social locality, a relatively small set of nodes reflective of their interests or communities. The best approximation to this locality, among the models we evaluated on LiveJournal data was the one determined by the union of clusters a node belonged to. Our notion of a cluster is a set of nodes which locally maximized a cluster density. This notion of a cluster has the important property that it

allows clusters to overlap, which is important if a cluster is to represent a community or coalition.

Many possibilities exist for enhancing the definitions of locality and the attachment mechanisms. One direction which we intend to pursue as future research is the combination of local with global attachment mechanisms.

Acknowledgment 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.

9. REFERENCES

[1] R. Albert and A.-L. Barabási. Statictical mechanics of complex networks. *Reviews of Modern Physics*, 74(47-97), 2002.

[2] 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.

[3] J. Baumes, H.-C. Chen, M. Francisco, M. Goldberg, M. Magdon-Ismail, and W. Wallace. Dynamics of bridging and bonding in social groups, a multi-agent model. In *Third conference of the North American Association for Computational Social and Organizational Science (NAACSOS 05)*, Notre-Dame, Indiana, June, 26-28, 2005.

[4] J. Baumes, M. Goldberg, M. Krishnamoorthy, M. Magdon-Ismail, and N. Preston. Finding comminities by clustering a graph into overlapping subgraphs. *Proceedings of IADIS International Conference, Applied Computing 2005*, pages 97–104, 2005.

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

[6] J. Baumes, M. Goldberg, M. Magdon-Ismail, and W. Wallace. Identification of hidden groups in communications. *Handbooks in Information Systems, Volume 2: National Security, 2007*, 2007.

[7] T. Berger-Wolf and J. Saia. A framework for analysis of dynamic social networks. *DIMACS Technical Report*, 28, 2005.

[8] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stat, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer Networks*, 33(1-6):309–320, 2000.

[9] F. Chung and L. Lu. Connected components in random graphs with given degree sequence. *Annals of Combinatoreics*, 6:125–1456, 2002.

[10] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data, 2007.

[11] P. Doreian and E. F.N. Stokman. Evolution of social networks. *Gordon and Breach*, 1997.

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

[13] C. Gkantsidi, M. Mihail, and E. Zegura. The markov chain simulatiopn methods for generating connected power law random graphs. *Proc. of ALENEX'03*, pages 16–50, 2003.

[14] 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.

[15] M. Goldberg, S. Kelley, M. Magdon-Ismail, and K. Mertsalov. A locality model for the evolution of blog networks. In *IEEE Information and Security Informatics (ISI)*, 2008.

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

[17] J. M. Kleinberg and S. Lawrence. The structure of the web. In *Science*, pages 1849–1850, 2001.

[18] G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311:88–90, 2006.

[19] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. Structure and evolution of blogospace. *Communications of the ACM*, 33(1-6):309–320, 2004.

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

[21] M. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

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

[23] M. E. J. Newman. The structure of scientific collaboration networks. *PROC.NATL.ACAD.SCI.USA*, 98:404, 2001.

[24] A. O. Stauffer and V. C. Barbosa. A study of the edge-switching markov chain method for the generation of random graphs. `arxiv: cs. DM/0512105`, 2006.

[25] D. B. West. Introduction to graph theory. *Prentice Hall, Upper Saddle River, NJ*, 2003.