A Community-Based Model of Online Social Networks

Leendert Botha
MIH Media Lab
Stellenbosch University
South Africa
lwbotha@ml.sun.ac.za

Steve Kroon
Computer Science Division
Stellenbosch University
South Africa
kroon@sun.ac.za

ABSTRACT

A major concern expressed by the research community at the 2009 Social Network Mining and Analysis workshop in Paris, was the lack of benchmark data sets for social network analysis. The quality of existing data sets was also criticized for incompleteness, sampling bias and the lack of temporal data. In this paper, we focus on these issues by proposing a random graph model for the generation of artificial social network data. Our community-based model simulates the growth of social networks over time, and we investigate the quality of this model's fit on two complete temporal data sets of proprietary online social networks. We compare the model's performance with existing random graph models for social networks, focussing on three distinguishing properties of social networks: the average separation, the clustering present in such networks, and the degree distribution of the networks. We also provide a set of fully temporal artificial data sets generated by our model for use in social network analysis.[1]

Categories and Subject Descriptors

G.2.2 [**Discrete Mathematics**]: Graph Theory; E.1 [**Data Structures**]: Graphs and Networks; J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms

Measurement, Experimentation

1. INTRODUCTION

Social network analysis is applied in a variety of fields such as primatology [13], epidemiology[30], economics[19, 12], and anti-terrorism [33]. Online social networks are typically used for representing individuals and some form of relationship between them. As of July 2009, two thirds of all Internet users had joined at least one online social network, making them the platform of choice for creating and sharing content on the Internet [29]. Following this surge in popularity, researchers have started analyzing the structure of online social networks. A deep understanding of the structure of social networks is crucial for the design and evaluation of algorithms and data structures for online social networks. However, progress has been hampered by *inter alia* the lack of complete publicly available data sets. Due to security and privacy concerns, most organizations managing online social networks are hesitant to release data to the research community. Thus, most currently available data sets are partial networks scraped from major online networking sites such as Orkut [24], MySpace [22], and Twitter [28]. However, these data sets do not accurately represent the true networks, since they are typically subject to sampling bias (see Section 2.4). Leskovec et. al. note that almost all of the publicly available temporal data sets also suffer from the problem of the *missing past* [18], meaning that temporal data, when available, usually does not stretch back to the birth of the network. These problems highlight the importance of random graph models for data set generation.

To generate random social networks, a random graph model that accurately incorporates all the important characteristics of such networks is needed. A random graph model is a probabilistic model that uses a set of distributions and parameters to generate graphs with common characteristics. A large amount of work has focused on creating *static* random graph models, i.e. models that generate static social network configurations successfully. However, very few *dynamic* models exist that aim at modeling the development of the network.

Most existing random graph models employ a bottom-up approach, aimed at adding nodes and edges in such a way that the resulting graph resembles the community structure of a social network. We propose a top-down approach, in which we model the community structure and then translate that model into a social graph. A major advantage of this approach is that it is very intuitive, easily translating to real-world behavior where people meet new friends through the communities they belong to.

Although there exists sufficient evidence to suggest that social network evolution is a complex process [18, 2], advances in this area of study have been severely limited by the lack of publicly available temporal data sets.

[1] Datasets and more results are available from `www.ml.sun.ac.za/~lwbotha`

The 4th SNA-KDD Workshop '10 (SNA-KDD'10) July 25, 2010, Washington DC, USA

2. SOCIAL NETWORK STRUCTURE

Most research about social network structure has focused on three distinguishing properties of social networks: the 'small-world' phenomenon, the clustering present in such networks, and the degree distribution of the networks. In what follows, we refer to the users in the network as *nodes* and to the connections between them as *edges*.

2.1 Small World Phenomenon

Small-world networks are networks with a small *average separation* i.e. a small average distance between a random pair of nodes in the network. Kochen and de Sola Pool began investigating the small world problem in the early 1950s [25]. Motivated by his interaction with Pool and Kochen, social psychologist Stanley Milgram designed an experiment to measure the average degree of separation between people in the United States. He gave letters to random subjects who each were instructed to pass the letter on to an acquaintance who they thought might know the addressee. He found the average number of people required for the letter to reach its destination to be only about six [20], which sparked the social phrase "Six Degrees of Separation".

This very important result means that even in large social networks, a relatively short path can be found between any pair of nodes. This result has been confirmed by a number of independent studies [32, 3, 1], including a recent study of the Microsoft Messenger Instant-Messaging System performed by Leskovec and Horvitz [17], in which they found the average separation to be 6.6 in a social graph containing 180 million nodes. Thus it is important that the networks generated by a random graph model closely match the average separation in real-life networks.

A further manifestation of the small world phenomenon is that the *network diameter*[2] $D(n)$, as a function of n, the number of nodes in the network, scales very slowly. Empirical results range from $D(n) \approx \log(n)$ [3] to $D(n) \approx \log \log n$ [18], with some networks even showing shrinking diameters [18].

2.2 Clustering

A common property of social networks is that highly connected clusters form, representing groups of people in which many members are acquainted with each other. We refer to these highly connected clusters as *communities*, and they often correspond to real-life behavior in which most of one's acquaintances come from a small set of social circles such as one's family, school, work colleagues or sport club.

In 1998, Watts and Strogatz introduced the *clustering coefficient* [32], a measure that quantifies the degree of clustering in a network. For a given node i, with degree $k_i > 1$, the clustering coefficient is defined to be the ratio of the number of links that exist between node i's neighbors and the total number of potential links that could exist between them. When $k_i \leq 1$, the clustering coefficient of the node is zero. Formally, if E_i is the number of links that actually exist between the k_i neighbors of node i, then the clustering coefficient of the network, defined as the average of all individual clustering coefficients, is given by

$$CC(n) = \frac{1}{n} \sum_{i:k_i>1} \frac{2E_i}{k_i(k_i-1)}$$

Note that we view the clustering coefficient as a function of n, since we want to analyze the evolution of this measure as the network grows. In social networks, the clustering coefficient is, in general, several orders of magnitude greater than in other networks. Since the clustering coefficient is defined to be 0 for isolated nodes and leaf nodes, using only the giant component for analysis will result in an over-estimation of the clustering coefficient. Therefore, it is very important that a random graph model should produce the right percentage of isolated nodes.

2.3 Degree Distribution

The degree distribution of a graph is a distribution function $P(k)$ that gives the probability that a randomly selected node in the graph has degree k. In a purely random Erdős-Rényi (ER) graph [9], the degree distribution is binomial, so that the vast majority of the nodes have degree close to the mean degree. However, empirical results [3] show that for social networks, the degree distribution follows a power law of the form

$$P(k) \propto k^{-\alpha} .$$

Graphs with degree distributions of this form are called *scale-free*: They typically consist of a large number of nodes with low degree and a small percentage of nodes with unusually high degree.

2.4 Problems with Sampling from Social Networks

Due to their complex nature, no standard sampling technique seems to preserve the above three properties of social networks [15, 21]. The most used sampling technique, *snowball sampling* (also referred to as 'crawling' a network), is often the only option for online social networks. Snowball sampling starts from a pre-selected node and follows edges from this node, recursively adding all nodes and edges it encounters to the sampled network. Due to their highly connected nature, dense communities are over-sampled, producing connected networks with significantly higher clustering coefficients and shorter average path lengths than the original networks [15]. Also, this method is extremely likely to only sample from the *giant component*[3] of the network and gives no indication of how many other connected components or isolated nodes there are in the network.

3. RELATED WORK

The theory of random graphs was pioneered in the late 1950s by Paul Erdős and Alfréd Rényi [9]. Despite the widespread use of their model in a variety of fields, it has been shown that it does not capture any of the important characteristics of social networks [3, 16, 32]. In 1998, Watts and Strogatz (WS) introduced a model aimed at generating small world networks to address this shortcoming [32]. Their model produced highly clustered networks, but the degree distribution is still Poisson, unlike that of real-world social networks. Much research on social networks followed, including work by Barabasi and Albert [3] who introduced the concept of preferential attachment (PA) to reproduce the degree distributions observed in social networks, in 1999.

[2] The diameter of a network is the maximal shortest path length between any two nodes in the network.

[3] Most sufficiently dense random graphs have a connected component comprising a high proportion of the nodes, known as the giant component.

Model	Small-world	Degree Distribution	Clustering	Dynamic	Community overlap
Erdős-Rényi (ER)	No	No	No	No	-
Watts and Strogats (WS)	Yes	No	Yes	No	-
Preferential Attachment (PA)	Yes	Yes	No	Yes	-
Triangle Closure	Yes	Yes	Yes	No	-
Birmelé	Yes	Yes	Yes	No	No
Guillaume and Latapy (GL)	Yes	Yes	Yes	Yes	No

Table 1: Properties captured by the main current models.

Their model was the first dynamic model for small-world, scale-free networks. In the PA model, nodes are added one-by-one, with each being connected to a pre-defined number of existing nodes, chosen with linear bias based on the degree of the nodes. The PA model has been extended in various ways, such as by associating an additional *fitness* value with each node [4], by using non-linear preferential attachment [14], and by removing some edges to model network decay [7]. Although these models typically generate small-world, scale-free networks, they fail to reproduce the high level of clustering present in social networks [11].

p^* models [31, 26] are another class of models which model the prevalence of certain node configurations in the network using maximum likelihood principles. However, these complex models fail to provide an intuitive interpretation of how the networks develop, and they often require expensive computations of high-dimensional integrals.

To keep the computational complexity of such models manageable, many authors have focused solely on the most prominent of the node configurations: the triangle. In social networks, people with common friends are more likely to become friends, resulting in many triangles occurring in the network. This contributes to the high degree of clustering observed in online social networks. Models building networks through triangle closures have been proposed by various authors [27, 23], and, in general, they produce highly clustered networks, although they are mostly aimed at generating static data sets.

A very promising new class of models use a two-level, or bipartite structure. This effectively adds a second layer of nodes to the network which is used, in different ways, to build the social network. Birmelé presented a static model [5] which uses the top nodes to represent cliques (fully connected communities) so that all the bottom nodes connected to the same top node are connected in the social network. Guillaume and Latapy (GL) [10, 11] use the same structure, but they propose a growing model based on preferential attachment. The GL model takes as input the degree distribution of the top nodes: If this distribution follows a Poisson law, then the generated networks exhibit social network structure. They show that their model fits six static scale-free data sets fairly well, not including an online social network. Although this simple model yields an intuitive and accurate way of building the network, it has two shortcomings:

1. The way the communities form is not entirely consistent with real-world behavior. In online social networks, communities are more accurately modeled using *quasi-cliques*, i.e. groups of nodes which are highly connected, but not completely. Guillaume and Latapy's model generates networks using fully connected communities, generally resulting in a higher level of clustering than desired.
2. As they note, their model fails to incorporate *bipartite clustering*, with communities showing very little neighborhood overlap. This is a major drawback, since, in real-world networks, if two communities have one node in common, they are likely to have more. In fact, real-world networks show a hierarchical community structure, with a recent study on a mobile social network revealing up to six levels in the community hierarchy [6].

Table 1 summarizes the discussed models, in terms of their performance in reproducing the most important properties of social networks. Our work is, in principle, an extension of the model of Guillaume and Latapy, in which we attempt to address the two shortcomings above and focus on the development of the network.

4. OUR MODEL

All of the models presented above, except Birmele's model, use a 'bottom-up approach', adding nodes and edges at the *microscopic* level in a certain way in order to mimic social network structure on a *macroscopic* level. One very important characteristic of this macroscopic structure is the potentially very complex way in which communities evolve and overlap. We use a 'top-down approach' by first modeling the growth and overlap of the community structure. This is a very intuitive approach which translates directly to real-life behavior, where one makes friends through interactions in the communities one belongs to. It is very important to realize that the 'communities' we refer to here are communities as defined by the *implicit* structure of the social network, corresponding to 'real-life' communities, i.e. families, schools, clubs, social circles, etc. These communities are often identified as clusters in the social network. For an analysis of the evolution and structure of *explicitly* defined communities in social networks, i.e. communities/groups that the user joins on the network, refer to the work of Backstrom et. al. [2].

Once we have modeled the macroscopic community structure, we sample the microscopic interaction between the nodes from the community model. Socially, there are many factors that influence the probability that two people will befriend each other in a community. First, the activity of the two people in the community are important factors for this is directly linked to the probability of them meeting and befriending in the community. We incorporate this into the model by associating a *commitment value* for each community a user belongs to. Furthermore, the intimacy of the community is important since some communities, such as

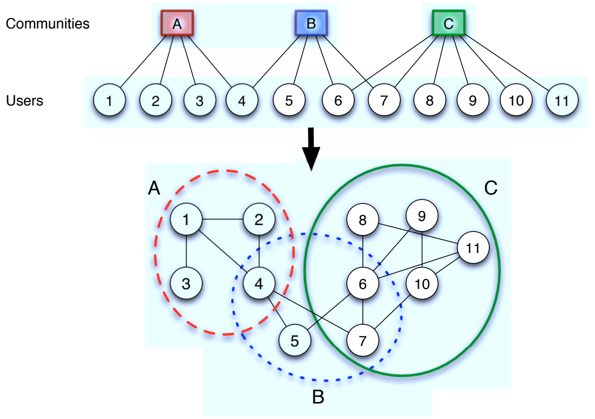

Figure 1: An example of a bipartite community structure (above) and a possible sampled social network (below).

families, are denser in personal relationships than others, like college classes. We model this by associating a *density* value with each community. This is analogous to the 'level of activity' Backstrom et. al. uses for explicitly defined online communities [2]. The densities and commitment values, as well as the *user activities*, which are an indication of how likely users are to join a community, can be sampled from arbitrary distributions. For a description of the distributions we use in this study, see Table 3.

4.1 Community Model Construction

We grow a community structure, represented as a bipartite graph, over a series of time steps as follows. During each time step:

1. With probability β a new community node c_i, with no connections, is added to the network. A *density* value, $d(c_i)$, is associated with c_i and this refers to the 'intimacy' of the community, an indication of how likely it is that two members of c_i are connected.

2. With probability γ a new user node u_j is added to the network. An *activity* value, $a(u_j)$, is associated with each user node. Upon joining the network, the node is connected to one community, chosen uniformly at random.

3. An existing user node u_j is connected to an existing community node c_i:

   - The node, u_j, is chosen preferentially based on activity, i.e. the probability of node j being chosen is equal to
   $$p_j = \frac{a(u_j)}{\sum_{k=1}^{n} a(u_k)} \tag{1}$$
   - The community c_i is chosen using a two-step process: First, a community c is selected using preferential attachment based on the community commitments of u_j. Then, c_i is selected from the set of communities u_j is not a member of, using preferential attachment based on the *overlap* between c and these communities. The overlap $\theta(c, c_k)$ is defined as the number of mutual members of c and c_k.

   Whenever a user node u_i is connected to a community node c_k, a weighted edge is inserted in the bipartite graph between u_i and c_k. The edge weight δ_{ik} indicates the user node's commitment to the community node. As was mentioned above, the density values $d(c_i)$, activity values $a(u_j)$, and commitment values δ_{ik} are sampled from arbitrary distributions. In this study, we sampled these values from power-law distributions for which the parameters are shown in Table 3.

4.2 Social Network Construction

Multiple growing social networks can be sampled from the bipartite graph B during its construction. This is done as follows. For each social network G being sampled:

- All user nodes of B are nodes in G.
- Whenever a user node u_i is connected to a community node c_k in B, u_i is connected in G to each member u_j of c_k with probability
$$f(\delta_{ik}, \delta_{jk}, d_k) = \frac{1}{w}\exp\left(\frac{-\phi}{\delta_{ik}}\right)\cdot\exp\left(\frac{-\phi}{\delta_{jk}}\right)\cdot\exp\left(\frac{-\phi'}{d_k}\right) \tag{2}$$
if u_i and u_j are not already connected. For the purposes of this study, we chose $\phi = \phi' = 1$, yielding
$$f(\delta_{ik}, \delta_{jk}, d_k) = \frac{1}{w}\exp\left[-\left(\frac{1}{\delta_{ik}} + \frac{1}{\delta_{jk}} + \frac{1}{d_k}\right)\right]$$
where w is a scaling constant.

Thus, whenever a user node is connected to a community node in B, it is connected in G to each other node already in that community with probability based on the density of the community and the commitments of the two nodes to the community. This probabilistic model simplifies to the deterministic version used in the GL model when $f(\delta_{ik}, \delta_{jk}, d_k) = 1$. This corresponds to the limiting case of our model where $w = 1$ and $\delta_{ik}, \delta_{jk}, d_k \to \infty$, i.e. all user nodes have an infinitely strong commitment to all of their communities, and all communities are infinitely intimate.

A very important property of the construction process is that the bipartite graph grows independently of the social networks sampled from it. Furthermore, we can also sample static data sets after building the bipartite graph B, which will be structurally equivalent to the networks sampled during the construction of B. To do this, we:

1. Add all the user nodes of the bipartite graph B to the social network G.

2. Connect each pair of nodes u_i and u_j in G with probability:
$$P(e_{ij}) = \sum_{k=1}^{r}\left[f(u_i, u_j, c_k)\cdot\prod_{l=1}^{k-1}(1 - f(u_i, u_j, c_l))\right]$$

where the sum is over all r mutual communities of u_i and u_j.

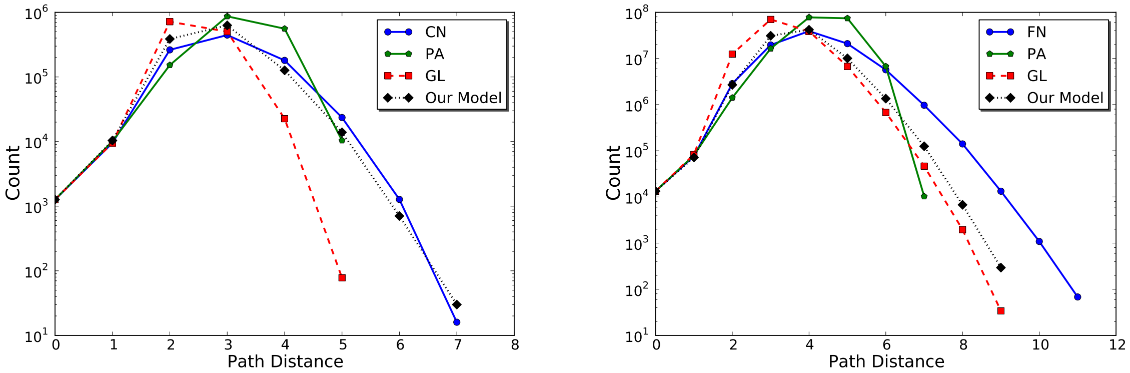

Figure 2: Pairwise distance histogram of the CN (left) and the FN (right) with those of the fitted models.

5. RESULTS

Our first temporal data set is from a proprietary corporate social network owned by a multi-national holding company. It is a closed network in which employees can connect with colleagues in other companies owned by the parent company. Although the network is small ($N = 1265$ nodes), it is a mature network, having being adopted by most of the individual companies since it's launch in 2008. For this network, we have exact temporal information for the arrival of all the nodes and the edges. We refer to this network as the *Corporate Network* (CN).

The second network is a South African social network attracting young people through a local presence in entertainment venues. The data set contains exact temporal information for $N = 13,295$ nodes and $M = 40,679$ connections between them. We refer to this network as the *Friendship Network* (FN).

	N	M	$\alpha(N)$	$CC(N)$	$AS(N)$	$D(N)$
CN	1265	4753	1.63	0.29	2.24	7
FN	13295	40679	1.87	0.021	2.95	11

Table 2: Information about the networks, showing the total number of nodes (N) and edges (M), with the power-law parameter (α), the clustering coefficient (CC), the average separation (AS) and the network diameter (D).

Although both these networks strongly exhibit all the properties of social networks, they are structurally very different. The separate companies in the CN form very highly connected clusters, whereas the FN shows a much lower degree of clustering: the clustering coefficient of the CN is 14 times that of the FN. Despite this, the average degree of nodes in the FN is only 20% less than those in the CN. Modeling both of these networks accurately poses an interesting challenge, since it requires the model to show a great deal of flexibility.

To fit our model to the data sets, a grid search was performed over the parameter space in order to find the parameter choice that generated a network development most visually similar to the data sets. Since our model is probabilistic, we performed a set of simulations using the resulting parameter estimates (shown for the two networks in Table 3), and the mean of the results are compared to that of the true data. Note that the deviation of each simulation from the mean is almost unnoticeable on the scale of the presented plots, and error bars are therefore omitted.

	β	γ	α_1	α_2	α_3	w
CN	0.2	0.29	2.0	2.5	1.2	1.385
FN	0.005	0.3	1.7	2.5	2.0	66.67

Table 3: Our model's parameter estimates for the CN and the FN. We sample the user activities, the commitments, and the community densities from power-law distributions with parameters α_1, α_2 and α_3 respectively.

For the GL model, we used the method described by Guillaume and Latapy [10] to obtain the optimal parameters. The PA model takes only one parameter, m_0, the number of connections a node creates upon its entry. We chose m_0 to be half the desired average degree. In this way, all three models were configured to produce the same number of nodes and approximately the same number of connections as in the true data.

5.1 Average Separation

Figure 2 shows a histogram of the shortest path lengths for the true data together with the data generated by the three models for $n = N$. Our model matches the histograms noticeably better than the other two models, both of which overproduce shorter paths and fail to produce paths of longer length. In fact, in both cases, the other models generate networks with significantly lower diameters than that of the true data, whereas our model matches the diameter more accurately.

In general, the Guillaume model produces communities much denser than that of the true data, since all communities are built from fully connected cliques. The community density parameter in our model allows for less dense communities to form, creating the longer paths between some nodes that match those observed in the true data. In the PA model, since each node is connected to at least m_0 other nodes, there exist much shorter paths between most pairs of nodes than in the true data.

5.2 Clustering Coefficient

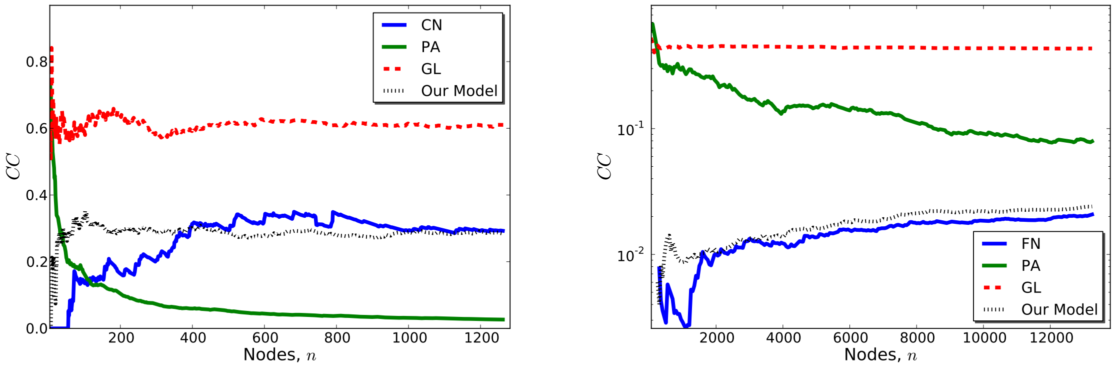

Figure 3: Evolution of the clustering coefficient of the CN (left) and the FN (right), compared to those of the three models.

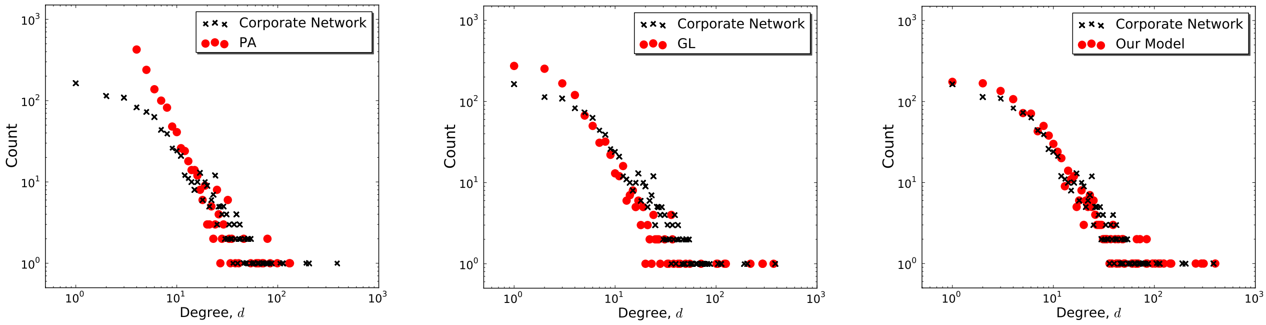

Figure 4: Degree distribution of the CN, compared to those of the three models for $n = 1265$.

Figure 3 shows the evolution of the clustering coefficients of the two networks and their fitted models. The PA model yields a clustering coefficient that decreases with network size, a trend that is clearly not observed in the true data. The CN shows an initial increase in clustering and starts to stabilize for $n > 600$. For the FN, the increasing trend is still continuing (note the logarithmic y-axis). Our model is the only model to capture this initial growth period of the networks in which the clustering increases with network size.

For the highly clustered CN network, the clustering coefficient yielded by the PA model is about ten times lower than desired. The Guillaume model, on the other hand, generates an even more clustered network with a clustering coefficient double that of the CN. Our model fits the clustering coefficient well throughout the evolution of the network, being within 20% of the true value 75% of the time.

The FN shows an extremely low degree of clustering compared to the CN, although it's clustering coefficient is still about 60 times that of a purely random Erdős-Rényi graph. Neither of the existing models are able to produce networks with a clustering coefficient this low. On average, Guillaume's model exceeds the true clustering coefficient by a factor 20. The PA model provides a closer match, yielding a clustering coefficient 6.5 times greater on average, although the trend does not match that of the true data at all. Our model matches the evolution of the clustering coefficient very accurately, being within 25% of the true value 84% of the time.

5.3 Degree Distribution

Figure 4 shows the degree distributions of the CN and the three fitted models at the end of the simulation, that is for $n = 1265$. All three of the models produce power-law degree distributions. However, since the minimum degree in the PA model is $m_0 = 4$, the distribution has a much shorter tail than that of the CN. In fact, it has been shown that the PA model can only produce networks for which the degree distribution has parameter $\alpha = 2.9 \pm 0.1$ [8]. Our model yields the closest match, with the GL model also providing a good fit, but with a slightly shorter tail than the degree distribution of the true data. The FN showed similar results, which we omit for brevity.

Since we are interested in the development of the network, we present the evolution of the power-law parameter, α, for the two networks in Figure 5. In both networks, α decreases before stabilizing, although in the CN, α shows a slight increase towards the end. Thus, both networks start out with a degree distribution with a shorter tail, and this tail gradually grows before α stabilizes. Our model also shows this downward trend in α, although the decrease is much more rapid. The GL model does not show this trend, yielding a power-law parameter which seems to be quite stable throughout the simulation, although their degree distribution very closely matches that of the FN network for $n = N$. Our model deviates from the evolution of the power-law pa-

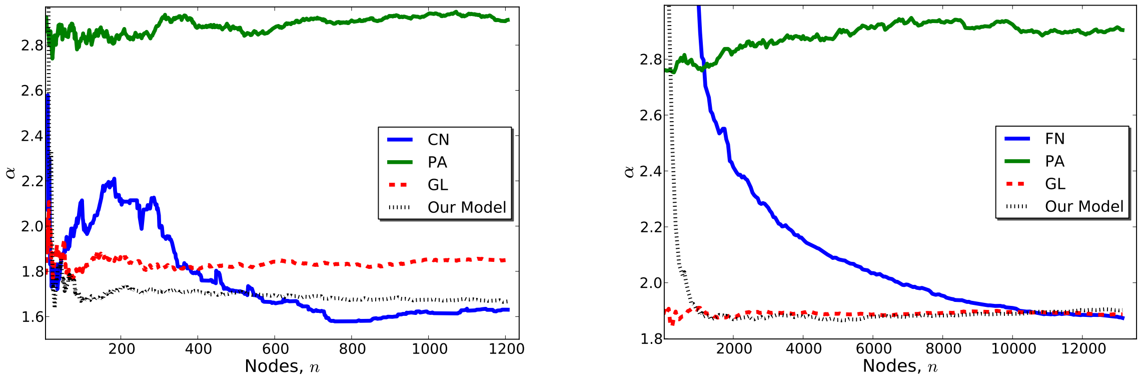

Figure 5: Evolution of the power-law parameter of the CN (left) and the FN (right), compared to those of the three models.

rameter of the CN for $n < 500$, and thereafter provides the best fit to the data. For the FN, our model also produces a growing tail in the degree distribution, but with the speed of growth exceeding that of the true data before stabilizing to approximately the same value.

6. FUTURE WORK

As noted, research on models for social networks is restricted by the lack of complete temporal data sets. If possible, we would like to analyze more data sets to provide further evidence of the quality of our model.

There are a number of interesting questions that arise from this work. All of the existing models for social networks assume 'natural' growth of the network without external influence. How would the evolutionary process change if the users were given various incentives to be active on the network, or to draft new users? How should one incorporate such factors in a random graph model? It would also be interesting to know what causes some networks to be so tightly clustered compared to others. How much of the difference between the clustering coefficients in our two networks is due to the differing 'culture' of the two sites, and how much of the difference can simply be explained by the different degree distributions?

Envisaged future work on our proposed model includes deriving an improved parameter estimation procedure, as well as a more thorough theoretical analysis of the properties of the model and the effects of the various model parameters (including ϕ and $\phi^{'}$ in (2)). This knowledge would make our model a potentially useful tool for social network growth prediction and algorithm analysis.

7. CONCLUSIONS

We have presented an intuitive model for generating random social networks and showed that the generated networks fit the important properties of social networks well, namely: the average separation between the nodes, the clustering coefficient and the degree distribution. To demonstrate the model's flexibility we fitted it to two quite different online social networks, and we showed that our model outperforms existing random graph models in modeling the development of these two real-world networks. Although our model is more complex than the other two and uses a richer parameter space, we feel that the amount of information present in the network data justifies a richer parameter space, a claim supported by the results presented. We also used our model to generate a set of fully temporal social network data sets, which we are releasing for public use in social network analysis.

8. ACKNOWLEDGEMENTS

The authors wish to express their deep gratitude towards Myriad International Holdings (MIH), without whom this research project would not be possible. We would also like to thank Amy Becht for her very useful comments on previous drafts of this paper.

9. REFERENCES

[1] L. Amaral, A. Scala, M. Barthelemy, and H. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21):11149, 2000.

[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. *Proceedings of the 12th ACM SIGKDD International Conference*, page 54, 2006.

[3] A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[4] G. Bianconi and A. Barabasi. Competition and multiscaling in evolving networks. *Europhysics Letters*, 54(4):436–442, 2001.

[5] E. Birmele. A scale-free graph model based on bipartite graphs. *Discrete Appl. Math.*, 157:1–23, Jun 2009.

[6] V. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008:P10008, 2008.

[7] S. Dorogovtsev and J. Mendes. Scaling behaviour of developing and decaying networks. *EPL (Europhysics Letters)*, 52:33–39, 2000.

[8] R. Durrett. *Random Graph Dynamics*. Cambridge Series in Statistical and Probabilistic Mathematics, 2006.

[9] P. Erdös and A. Rényi. On random graphs, i. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

[10] J. Guillaume and M. Latapy. Bipartite structure of all complex networks. *Information processing letters*, 90(5):215–221, 2004.

[11] J. Guillaume and M. Latapy. Bipartite graphs as models of complex networks. *Physica A: Statistical Mechanics and its Applications*, 371(2):795–813, 2006.

[12] Y. Ioannides. Random graphs and social networks: An economics perspective. *Unpublished manuscript, Tufts University*, 2006.

[13] C. Kasper and B. Voelkl. A social network analysis of primate groups. *Primates*, 2009.

[14] P. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Physical Review Letters*, 85(21):4629–4632, 2000.

[15] S. H. Lee, P.-J. Kim, and H. Jeong. Statistical properties of sampled networks. *Physical Review E*, 73(1):1–7, Jan 2006.

[16] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. *Proceeding of the 14th ACM SIGKDD International Conference*, pages 462–470, 2008.

[17] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. *Proc. of the 17th Intl. Conference on WWW*, pages 915–924, 2008.

[18] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2, 2007.

[19] A. Mayer. Online social networks in economics. *Decis. Support Syst.*, 47(3):169–184, 2009.

[20] S. Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

[21] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. *Proceedings of the 7th ACM SIGCOMM Conference on Internet measurement*, pages 29–42, 2007.

[22] Myspace. `www.myspace.com`.

[23] M. Newman. Random graphs with clustering. *Physical Review Letters*, 103(5):58701, 2009.

[24] Orkut. `www.orkut.com`.

[25] I. Pool and M. Kochen. Contacts and influence. *Social Networks*, 1:1–48, 1978.

[26] G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2):173–191, 2007.

[27] X. Shi, L. Adamic, and M. Strauss. Networks of strong ties. *Physica A: Statistical Mechanics and its Applications*, 378(1):33–47, 2007.

[28] Twitter. `www.twitter.com`.

[29] UniversalMcCann. Power to the people - social media tracker wave 4. `http://www.universalmccann.com/wave/`, 2009. [Online; accessed 26-January-2010].

[30] M. Vélez, J. Ospina, and D. Hincapié. Tutte polynomials and topological quantum algorithms in social network analysis for epidemiology, bio-surveillance and bio-security. In *BioSecure '08: Proceedings of the 2008 International Workshop on Biosurveillance and Biosecurity*, pages 74–84, Berlin, Heidelberg, 2008. Springer-Verlag.

[31] S. Wasserman and P. Pattison. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*. *Psychometrika*, 61(3):401–425, 1996.

[32] D. Watts and S. Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 393(6684):440–442, 1998.

[33] C. C. Yang and M. Sageman. Analysis of terrorist social networks with fractal views. *J. Inf. Sci.*, 35(3):299–320, 2009.