Microscopic Evolution of Social Networks

Jure Leskovec* Lars Backstrom† Ravi Kumar‡ Andrew Tomkins‡
*Carnegie Mellon University †Cornell University ‡Yahoo Research
jure@cs.cmu.edu lars@cs.cornell.edu {ravikuma, atomkins}@yahoo-inc.com

ABSTRACT

We present a detailed study of network evolution by analyzing four large online social networks with full temporal information about node and edge arrivals. For the first time at such a large scale, we study individual node arrival and edge creation processes that collectively lead to macroscopic properties of networks. Using a methodology based on the maximum-likelihood principle, we investigate a wide variety of network formation strategies, and show that edge locality plays a critical role in evolution of networks. Our findings supplement earlier network models based on the inherently non-local preferential attachment.

Based on our observations, we develop a complete model of network evolution, where nodes arrive at a prespecified rate and select their lifetimes. Each node then independently initiates edges according to a "gap" process, selecting a destination for each edge according to a simple triangle-closing model free of any parameters. We show analytically that the combination of the gap distribution with the node lifetime leads to a power law out-degree distribution that accurately reflects the true network in all four cases. Finally, we give model parameter settings that allow automatic evolution and generation of realistic synthetic networks of arbitrary scale.

Categories and Subject Descriptors: H.2.8 **[Database Management]**: Database applications—*Data mining*

General Terms: Measurement, Experimentation

Keywords: Social networks, Graph generators, Network evolution, Maximum likelihood, Triadic closure, Transitivity

1. INTRODUCTION

In recent years a wide variety of models have been proposed for the growth of complex networks. These models are typically advanced in order to reproduce statistical network properties observed in real-world data. They are evaluated on the fidelity with which they reproduce these global network statistics and patterns. In many cases, the goal is to define individual node behaviors that result in a global structure such as power law node degree distributions; in other cases, the goal is to match some other network property such as small diameter.

For example, the observation of heavy-tailed degree distributions [10] led to hypothesis about edge creation processes (e.g., preferential attachment [1]) that could lead to this observation. In fact, there are several edge creation processes that all lead to heavy-tailed degree distributions and it is not clear which among them captures reality best.

Here we take a different approach. Instead of only focusing on the global network structure and then hypothesizing about what kind of microscopic node behavior would reproduce the observed macroscopic network structure, we focus directly on the microscopic node behavior *per se*. For the first time at such a large scale, we study a sequence of millions of individual edge arrivals, which allows us to directly evaluate and compare microscopic processes that give rise to global network structure.

Evaluation based on likelihood. Given that the microscopic behavior of nodes solely determines the macroscopic network properties, a good network model should match real-world data on global statistics, while maximizing the likelihood of the low-level processes generating the data. Towards this goal, we propose the use of model likelihood of individual edges as a way to evaluate and compare various network evolution models.

Likelihood has not been considered to date in the analysis of evolution of large social networks mainly due to lack of data and computational issues. Many early network datasets contained only a single or a small number of snapshots of the data, making likelihood computations for evolutionary models infeasible. We study four large social networks with *exact* temporal information about individual arrivals of millions of nodes and edges. Here we are therefore able to consider edge-by-edge evolution of networks, and hence efficiently compute the likelihood that a particular model would have produced a particular edge, given the current state of the network. In contrast to previous work on evolution of large networks that used a series of snapshots to consider patterns at global scale, we study the exact edge arrival sequence, which means we are able to *directly* observe and model the fine-grained network evolutionary processes that are directly responsible for global network patterns and statistics.

A likelihood-based approach has several advantages over approaches based purely on global statistics:

(1) Models may be compared directly in a unified way, rather than arguing whether faithful reproduction of, e.g., diameter is more important than clustering coefficient and so forth.

(2) As our understanding of real-world networks improves, the evaluation criterion, i.e., likelihood, remains unchanged while the generative models improve to incorporate the new understanding. Success in modeling can therefore be effectively tracked.

(3) Models may be meaningfully distinguished based on as-yet-undiscovered properties of real-world data.

Part of this work was done while the first and second authors were visiting Yahoo! Research.

KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.
Copyright 2008 ACM 978-1-60558-193-4/08/08 ...$5.00.

Data and model structure. We consider four large online social network datasets — FLICKR (`flickr.com`, a photo-sharing website), DELICIOUS (`del.icio.us`, a collaborative bookmark tagging website), YAHOO! ANSWERS (`answers.yahoo.com`, a knowledge sharing website), and LINKEDIN (`linkedin.com`, a professional contacts website) — where nodes represent people and edges represent social relationships. These networks are large with up to millions of nodes and edges, and the time span of the data ranges from four months to almost four years. All the networks are in early stages of their evolution with the connected component being small and the clustering coefficient increasing over time.

We consider models that can be decomposed into three core processes, namely, the node arrival process (governs the arrival of new nodes into the network), the edge initiation process (determines for each node when it will initiate a new edge), and the edge destination selection process (determines the destination of a newly initiated edge). Our networks do not include removal of nodes or edges, so we do not model deletion (although we do model the “death” of a node in the sense that it ceases producing new edges).

Our results. We begin with a series of analyses of our four networks, capturing the evolution of key network parameters, and evaluation of the extent to which the edge destination selection process subscribes to preferential attachment. We show that the inherently non-local nature of preferential attachment is fundamentally unable to capture important characteristics in these networks. To the best of our knowledge, this is the first direct large-scale validation of the preferential attachment model in real networks.

Next, we provide a detailed analysis of the data in order to consider parsimonious models for edge destination selection that incorporate locality. We evaluate a wide variety of such models using the maximum-likelihood principle and choose a simple triangle-closing model that is free of parameters. Based on the findings, we then propose a complete network evolution model that accurately captures a variety of network properties. We summarize our model based on the three processes listed earlier.

Node arrival process. We find large variation in node arrival rates over the four networks, ranging from exponential to sub-linear growth. Thus we treat node arrival rate as input to our model.

Edge initiation process. Upon arrival, a node draws its lifetime and then keeps adding edges until reaching its lifetime, with edges inter-arrival rate following a power law with exponential cut-off distribution. We find that edge initiations are *accelerating* with node degree (age), and prove that this leads to power law out degree distributions. The model produces accurate fits and high likelihood.

Edge destination selection process. We find that most edges (30%–60%) are local as they close triangles, i.e., the destination is only two hops from the source. We consider a variety of triangle-closing mechanisms and show that a simple scheme, where a source node chooses an intermediate node uniformly from among its neighbors, and then the intermediate node does the same, has high likelihood.

Our model is simple and easy to implement. It precisely defines the network evolution process, and we also give parameter settings that allow others to generate networks at arbitrary scale or to take a current existing network and further evolve it. We show that our model produces realistic social network evolution following the true evolution of network properties such as clustering coefficient and diameter; our purely local model gives rise to accurate global properties.

2. RELATED WORK

Many studies on online social networks, world wide web, and biological networks focused on macroscopic properties of static networks such as degree distributions, diameter, clustering coefficient, communities, etc; work in this area includes [10, 21, 2, 18, 8, 7]. Similarly, macroscopic properties of network evolution, like densification and shrinking diameters, were examined [16, 11, 19, 13].

Given that the classical Erdös–Rényi model cannot capture the above network characteristics, a number of alternate network models have been proposed. The copying [14] and the preferential attachment [1] models belong to this category. The Forest Fire model [16] attempts to explain the densification and decreasing-diameter phenomena observed in real networks. See [6] for a topic survey.

Recently, researchers examined the finer aspects of edge creation by focusing on a small set of network snapshots. The role of common friends in community formation was analyzed by Backstrom et al. [3]. Kleinberg and Liben-Nowell [17] studied the predictability of edges in social networks. The role of triangle closure in social networks was long known to sociologists. Simmel theorized that people with common friends are more likely to create friendships and Krackhardt and Handcock [12] applied this theory to explain the evolution of triangle closures. A network model based on closed triangles was proposed by Shi et al. [20].

The maximum-likelihood principle has been typically used to estimate network model parameters [15, 22, 23] or for model selection [4], which often requires expensive computations of high dimensional integrals over all possible node arrival sequences. In contrast, we use the likelihood in a much more direct way to evaluate and compare different modeling choices at a microscopic level.

3. PRELIMINARIES

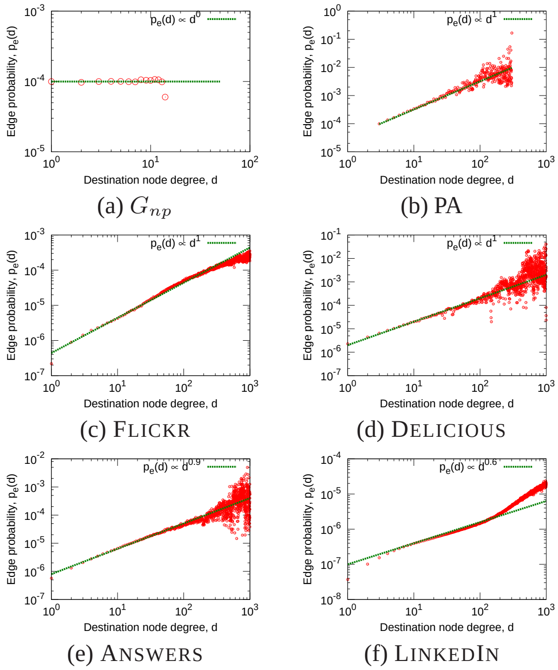
Datasets. For each of our four large network datasets, we know the exact time of all the node/edge arrivals. Table 1 gives the basic statistics of the four networks. All the networks slowly densify with a densification exponent [16] $\rho \approx 1.2$. All the networks, except DELICIOUS, have shrinking diameter. In FLICKR, ANSWERS, and LINKEDIN, the effective diameter reaches the maximum value of 10 when the network has around 50,000 nodes, and then slowly decreases to the around 7.5; in DELICIOUS, the diameter is practically constant. Also, in all the networks, a majority of edges are bidirectional (column E_b). The reciprocity is 73% in FLICKR, 81% in DELICIOUS, and 58% in ANSWERS; LINKEDIN is undirected, but we know the edge initiator. The fraction of nodes that belongs to the largest weakly connected component is 69% in FLICKR, 72% in DELICIOUS, 81% in ANSWERS, and 91% in LINKEDIN.

Notation. Let N, E, and T denote the total number of nodes, edges, and the span of the data in days. Let G_t be a network composed from the earliest t edges, $e_1, \ldots, e_t$ for $t \in \{1, \ldots, E\}$. Let $t(e)$ be the time when the edge e is created, let $t(u)$ be the time when the node u joined the network, and let $t_k(u)$ be the time when the k^{th} edge of the node u is created. Then $a_t(u) = t - t(u)$ denotes the age of the node u at time t. Let $d_t(u)$ denote the degree of the node u at time t and $d(u) = d_T(u)$. We use $[\cdot]$ to denote a predicate (takes value of 1 if expression is true, else 0).

Maximum-likelihood principle. The maximum-likelihood estimation (MLE) principle can be applied to compare a family of parameterized models in terms of their likelihood of generating the observed data, and as a result, pick the “best” model (and parameters) to explain the data. To apply the likelihood principle, we consider the following setting: we evolve the network edge by edge,

Network	T	N	E	E_b	E_u	$E_\triangle$	%	ρ	κ
FLICKR (03/2003–09/2005)	621	584,207	3,554,130	2,594,078	2,257,211	1,475,345	65.63	1.32	1.44
DELICIOUS (05/2006–02/2007)	292	203,234	430,707	348,437	348,437	96,387	27.66	1.15	0.81
ANSWERS (03/2007–06/2007)	121	598,314	1,834,217	1,067,021	1,300,698	303,858	23.36	1.25	0.92
LINKEDIN (05/2003–10/2006)	1294	7,550,955	30,682,028	30,682,028	30,682,028	15,201,596	49.55	1.14	1.04

Table 1: Network dataset statistics. E_b is the number of bidirectional edges, E_u is the number of edges in undirected network, $E_\triangle$ is the number of edges that close triangles, % is the fraction of triangle-closing edges, ρ is the densification exponent ($E(t) \propto N(t)^\rho$), and κ is the decay exponent ($E_h \propto \exp(-\kappa h)$) of the number of edges E_h closing h hop paths (see Section 5 and Figure 4).



(a) G_{np} (b) PA

(c) FLICKR (d) DELICIOUS

(e) ANSWERS (f) LINKEDIN

Figure 1: Probability $p_e(d)$ of a new edge e choosing a destination at a node of degree d.

and for every edge that arrives into the network, we measure the likelihood that the particular edge endpoints would be chosen under some model. The product of these likelihoods over all edges will give the likelihood of the model. A higher likelihood means a "better" model in the sense that it offers a more likely explanation of the observed data. For numerical purposes, we use log-likelihoods.

4. PREFERENTIAL ATTACHMENT

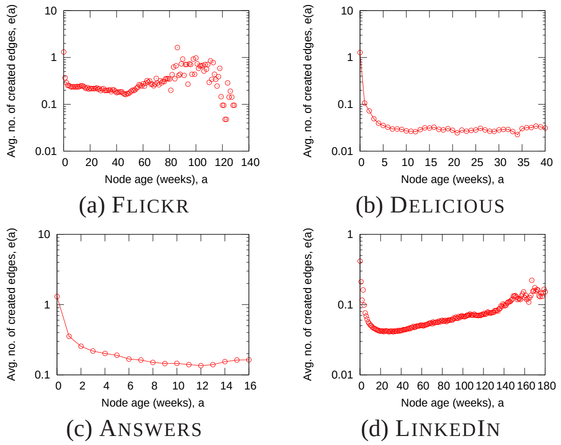
In this section we study the bias in selection of an edge's source and destination based on the degree and age of the node.

4.1 Edge attachment by degree

The preferential attachment (PA) model [1] postulates that when a new node joins the network, it creates a constant number of edges, where the destination node of each edge is chosen proportional to the destination's degree. Using our data, we compute the probability $p_e(d)$ that a new edge chooses a destination node of degree d; $p_e(d)$ is normalized by the number of nodes of degree d that exist just before this step. We compute:

$$p_e(d) = \frac{\sum_t [e_t = (u,v) \wedge d_{t-1}(v) = d]}{\sum_t |\{u : d_{t-1}(u) = d\}|}.$$

First, Figure 1(a) shows $p_e(d)$ for the Erdős–Rényi [9] random network, G_{np}, with $p = 12/n$. In G_{np}, since the destination node is chosen independently of its degree, the line is flat. Similarly, in the PA model, where nodes are chosen proportionally to their degree, we get a linear relationship $p_e(d) \propto d$; see Figure 1(b).



(a) FLICKR (b) DELICIOUS

(c) ANSWERS (d) LINKEDIN

Figure 2: Average number of edges created by a node of age a.

Next we turn to our four networks and fit the function $p_e(d) \propto d^\tau$. In FLICKR, Figure 1(c), degree 1 nodes have lower probability of being linked as in the PA model; the rest of the edges could be explained well by PA. In DELICIOUS, Figure 1(d), the fit nicely follows PA. In ANSWERS, Figure 1(e), the presence of PA is slightly weaker, with $p_e(d) \propto d^{0.9}$. LINKEDIN has a very different pattern: edges to the low degree nodes do not attach preferentially (the fit is $d^{0.6}$), whereas edges to higher degree nodes are more "sticky" (the fit is $d^{1.2}$). This suggests that high-degree nodes in LINKEDIN get super-preferential treatment. To summarize, even though there are minor differences in the exponents τ for each of the four networks, we can treat $\tau \approx 1$, meaning, the attachment is essentially linear.

4.2 Edges by the age of the node

Next, we examine the effect of a node's age on the number of edges it creates. The hypothesis is that older, more experienced users of a social networking website are also more engaged and thus create more edges.

Figure 2 plots the fraction of edges initiated by nodes of a certain age. Then $e(a)$, the average number of edges created by nodes of age a, is the number of edges created by nodes of age a normalized by the number of nodes that achieved age a:

$$e(a) = \frac{|\{e = (u,v) : t(e) - t(u) = a\}|}{|\{t(u) : t_\ell - t(u) \geq a\}|},$$

where t_ℓ is the time when the last node in the network joined.

Notice a spike at nodes of age 0. These correspond to the people who receive an invite to join the network, create a first edge, and then never come back. For all other ages, the level of activity seems to be uniform over time, except for LINKEDIN, in which activity of older nodes slowly increases over time.

4.3 Bias towards node age and degree

Using the MLE principle, we study the combined effect of node age and degree by considering the following four parameterized models for choosing the edge endpoints at time t.

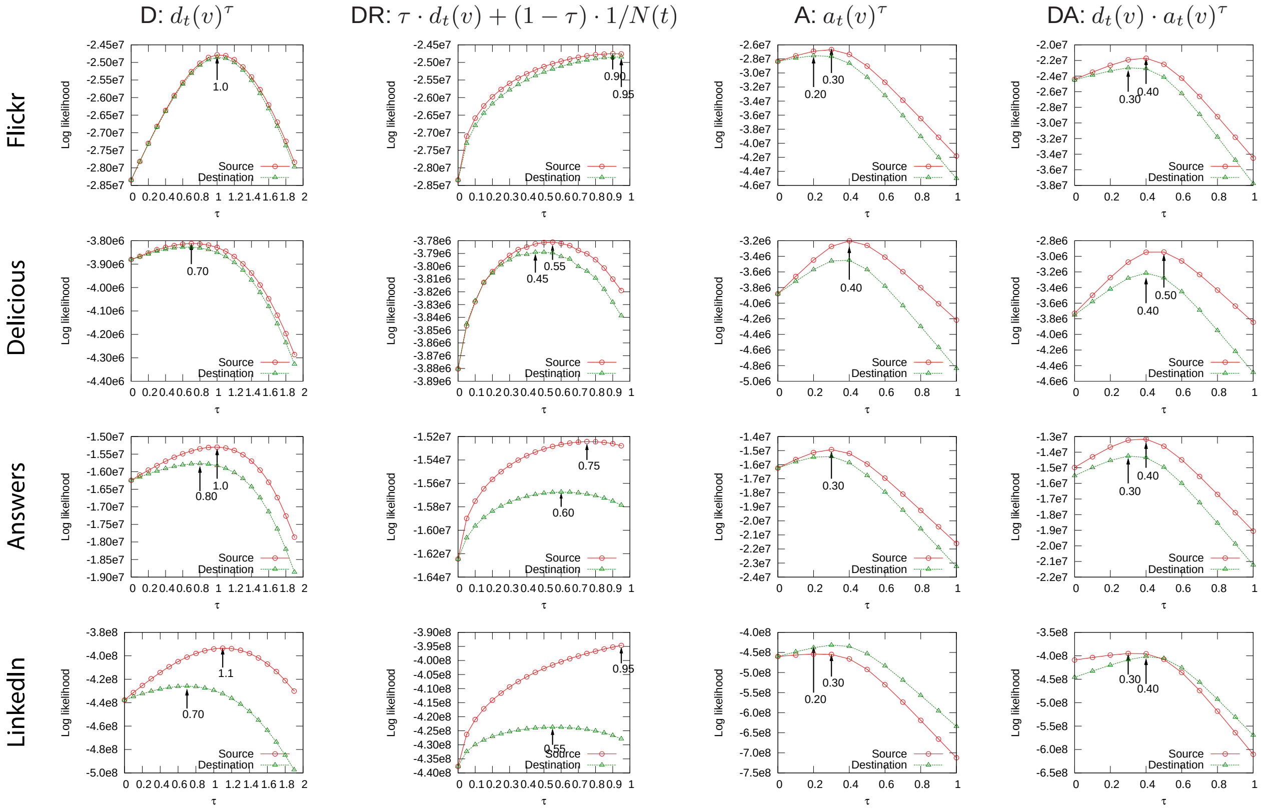


Figure 3: Log-likelihood of an edge selecting its source and destination node. Arrows denote τ at highest likelihood.

- D: The probability of selecting a node v is proportional to its current degree raised to power τ: $d_t(v)^\tau$.
- DR: With probability τ, the node v is selected preferentially (proportionally to its degree), and with probability $(1-\tau)$, uniformly at random: $\tau \cdot d_t(v) + (1-\tau) \cdot 1/N(t)$.
- A: The probability of selecting a node is proportional to its age raised to power τ: $a_t(v)^\tau$
- DA: The probability of selecting a node v is proportional the product of its current degree and its age raised to the power τ: $d_t(v) \cdot a_t(v)^\tau$.

Figure 3 plots the log-likelihoods under different models, as a function of τ. The red curve plots the log-likelihood of selecting a source node and the green curve for selecting the destination node of an edge.

In FLICKR the selection of destination is purely preferential: model D achieves the maximum likelihood at $\tau = 1$, and model DA is very biased to model D, i.e., $\tau \approx 1$. Model A has worse likelihood but model DA improves the overall log-likelihood by around 10%. Edge attachment in DELICIOUS seems to be the most "random": model D has worse likelihood than model DR. Moreover the likelihood of model DR achieves maximum at $\tau = 0.5$ suggesting that about 50% of the DELICIOUS edges attach randomly. Model A has better likelihood than the degree-based models, showing edges are highly biased towards young nodes. For ANSWERS, models D, A, and DR have roughly equal likelihoods (at the optimal choice of τ), while model DA further improves the log-likelihood by 20%, showing some age bias. In LINKEDIN, age-biased models are worse than degree-biased models. We also note strong degree preferential bias of the edges. As in FLICKR, model DA improves the log-likelihood by 10%.

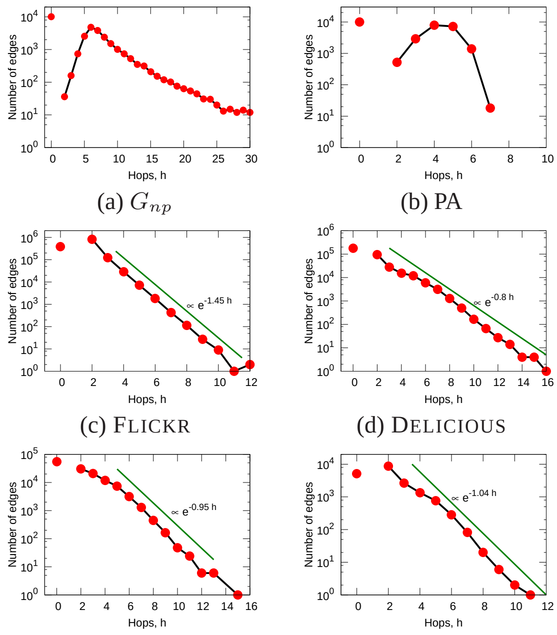
We notice that selecting an edge's destination node is harder than selecting its source (the green curve is usually below the red). Also, selecting a destination appears more random than selecting a source — the maximum likelihood τ of the destination node (green curve) for models D and DR is shifted to the left when compared to the source node (red), which means the degree bias is weaker. Similarly, there is a stronger bias towards young nodes in selecting an edge's source than in selecting its destination. Based on the observations, we conclude that PA (model D) performs reasonably well compared to more sophisticated variants based on degree and age.

5. LOCALITY OF EDGE ATTACHMENT

Even though our analysis suggests that PA is a reasonable model for edge destination selection, it is inherently "non-local" in that edges are no more likely to form between nodes which already have friends in common. In this section we perform a detailed study of the locality properties of edge destination selection.

We first consider the following notion of edge locality: for each new edge (u, w), we measure the number of hops it spans, i.e., the length of the shortest path between nodes u and w immediately before the edge was created. In Figure 4 we study the distribution of these shortest path values induced by each new edge for G_{np} (with $p = 12/n$), PA, and the four social networks. (The isolated dot on the left counts the number of edges that connected previously disconnected components of the network.)

For G_{np} most new edges span nodes that were originally six hops away, and then the number decays polynomially in the hops. In the PA model, we see a lot of long-range edges; most of them span four hops but none spans more than seven. The hop distributions corresponding to the four real-world networks look similar to one another, and strikingly different from both G_{np} and PA. The



(a) G_{np} (b) PA

(c) FLICKR (d) DELICIOUS

(e) ANSWERS (f) LINKEDIN

Figure 4: Number of edges E_h created to nodes h hops away. $h = 0$ counts the number of edges that connected previously disconnected components.

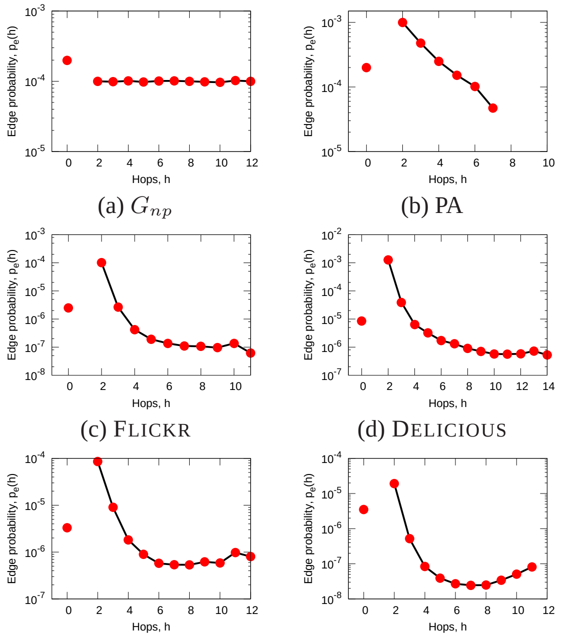
number of edges decays exponentially with the hop distance between the nodes (see Table 1 for fitted decay exponents κ). This means that most edges are created locally between nodes that are close. The exponential decay suggests that the creation of a large fraction of edges can be attributed to locality in the network structure, namely most of the times people who are close in the network (e.g., have a common friend) become friends themselves.

These results involve counting the number of edges that link nodes certain distance away. In a sense, this overcounts edges (u, w) for which u and w are far away, as there are many more distant candidates to choose from — it appears that the number of long-range edges decays exponentially while the number of long-range candidates grows exponentially. To explore this phenomenon, we count the number of hops each new edge spans but then normalize the count by the total number of nodes at h hops. More precisely, we compute

$$p_e(h) = \frac{\sum_t [e_t \text{ connects nodes at distance } h \text{ in } G_{t-1}]}{\sum_t (\#\text{ nodes at distance } h \text{ from the source node of } e_t)}.$$

First, Figures 5(a) and (b) show the results for G_{np} and PA models. (Again, the isolated dot at $h = 0$ plots the probability of a new edge connecting disconnected components.) In G_{np}, edges are created uniformly at random, and so the probability of linking is independent of the number of hops between the nodes. In PA, due to degree correlations short (local) edges prevail. However, a non-trivial amount of probability goes to edges that span more than two hops. (Notice the logarithmic y-axis.)

Figures 5(c)–(f) show the plots for the four networks. Notice the probability of linking to a node h hops away decays double-exponentially, i.e., $p_e(h) \propto \exp(\exp(-h))$, since the number of edges at h hops increases exponentially with h. This behavior is drastically different from both the PA and G_{np} models. Also note that almost all of the probability mass is on edges that close length-two paths. This means that edges are most likely to close triangles, i.e., connect people with common friends.



(a) G_{np} (b) PA

(c) FLICKR (d) DELICIOUS

(e) ANSWERS (f) LINKEDIN

Figure 5: Probability of linking to a random node at h hops from source node. Value at $h = 0$ hops is for edges that connect previously disconnected components.

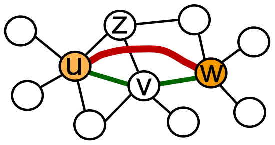


Figure 6: Triangle-closing model: node u creates an edge by selecting intermediate node v, which then selects target node w to which the edge (u, w) is created.

Column $E_\triangle$ in Table 1 further illustrates this point by presenting the number of triangle-closing edges. FLICKR and LINKEDIN have the highest fraction of triangle-closing edges, whereas ANSWERS and DELICIOUS have substantially less such edges. Note that here we are not measuring the fraction of nodes participating in triangles. Rather, we unroll the evolution of the network, and for every new edge check to see if it closes a new triangle or not.

5.1 Triangle-closing models

Given that such a high fraction of edges close triangles, we aim to model how a length-two path should be selected. We consider a scenario in which a source node u has decided to add an edge to some node w two hops away, and we are faced with various alternatives for the choice of node w. Figure 6 illustrates the setting. Edges arrive one by one and the simplest model to close a triangle (edge (u, w) in the figure) is to have u select a destination w randomly from all nodes at two hops from u.

To improve upon this baseline model we consider various models of choosing node w. We consider processes in which u first selects a neighbor v according to some mechanism, and v then selects a neighbor w according to some (possibly different) mechanism. The edge (u, w) is then created and the triangle (u, v, w) is closed. The selection of both v and w involves picking a neighbor of a node. We consider five different models to pick a neighbor v of u, namely, node v is chosen

FLICKR	random	deg$^{0.2}$	com	last$^{-0.4}$	comlast$^{-0.4}$
random	13.6	13.9	14.3	16.1	15.7
deg$^{0.1}$	13.5	14.2	13.7	16.0	15.6
last$^{0.2}$	14.7	15.6	15.0	17.2	**16.9**
com	11.2	11.6	11.9	13.9	13.4
comlast$^{0.1}$	11.0	11.4	11.7	13.6	13.2

DELICIOUS	random	deg$^{0.3}$	com	last$^{-0.2}$	comlast$^{-0.2}$
random	11.7	12.4	13.8	13.2	15.1
deg$^{0.2}$	12.2	12.8	14.3	13.7	15.6
last$^{-0.3}$	13.8	14.6	16.0	15.3	17.2
com	13.6	14.4	15.8	15.2	17.1
comlast$^{-0.2}$	14.7	15.6	16.9	16.3	**18.2**

ANSWERS	random	deg$^{0.3}$	com	last$^{-0.2}$	comlast$^{-0.2}$
random	6.80	10.1	11.8	9.70	13.3
deg$^{0.2}$	7.18	10.5	12.2	10.1	13.7
last$^{-0.3}$	9.95	13.4	15.0	12.8	**16.4**
com	6.82	10.3	11.8	9.80	13.4
comlast$^{0.2}$	7.93	11.5	12.9	10.9	14.5

LINKEDIN	random	deg$^{0.1}$	com	last$^{-0.1}$	comlast$^{-0.1}$
random	16.0	16.5	18.2	17.2	18.5
deg$^{0.1}$	15.9	16.4	18.0	17.0	18.4
last$^{-0.1}$	19.0	19.5	21.1	20.0	**21.4**

Table 2: Triangle-closing models. First pick intermediate node v (fix column), then target node w (fix row). The cell gives percent improvement over the log-likelihood of picking a random node two hops away (baseline).

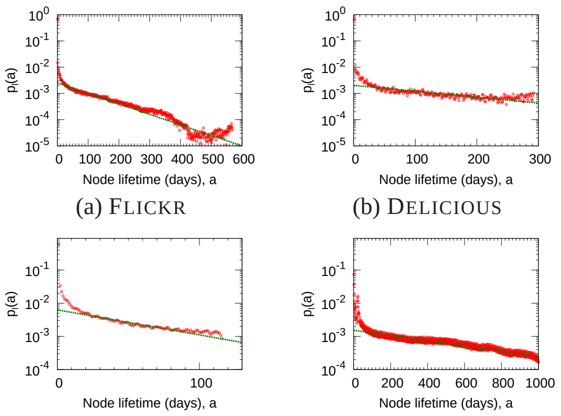
- random: uniformly at random,
- deg$^{\tau}$: proportional to degree raised to power τ, $d(v)^{\tau}$,
- com: prop. to the number of common friends $c(u, v)$ with u,
- last$^{\tau}$: proportional to the time passed since v last created an edge raised to power τ,
- comlast$^{\tau}$: proportional to the product of the number of common friends with u and the last activity time, raised to power τ.

As stated before, we can compose any two of these basic models to choose a two-hop neighbor, i.e., a way to close the triangle. For instance, the last$^{0.1}$-com model will work as follows: u will employ the last$^{0.1}$ model to select node v, v will employ the com model to select node w, and then u will add an edge to w, closing the triangle (u, v, w). We consider all 25 five possible composite models for selecting a two-hop neighbor and evaluate them by the likelihood that the model generated all the edges that closed length-two paths in the real network.

Table 2 shows the percent improvement of various triangle-closing models over the log-likelihood of choosing a two-hop neighbor uniformly at random as a destination of the edge (the baseline). The simplest model, random-random, works remarkably well and has many desirable properties. It gives higher probability to nodes with more length-two paths, discounting each path by roughly $1/d(v)$. Moreover, it is also biased towards high-degree nodes, as they have multiple paths leading towards them.

The deg$^{1.0}$-random model weighs each node w by roughly the number of length-two paths between u and w. However, we find that it performs worse than random-random. For the more general deg$^{\tau}$-random, the optimal value of τ varies from 0.1 to 0.3 over all the four networks, and this model provides meaningful improvements only for the ANSWERS network.

The com model considers the strength of a tie between u and v, which we approximate by the number of common friends $c(u, v)$ of nodes u and v; the larger the value, the stronger the tie. By selecting v with probability proportional to $c(u, v)$, we get a substantial gain in model likelihood. A factor that further improves the model is the recency of activity by v, captured by last$^{\tau}$. By selecting nodes that have recently participated in a new edge with higher probability, we get another sizable improvement in the model likelihood. These two capture the finer details of network evolution.



(a) FLICKR (b) DELICIOUS

(c) ANSWERS (d) LINKEDIN

Figure 7: Exponentially distributed node lifetimes.

In summary, while degree helps marginally, for all the networks, the random-random model gives a sizable chunk of the performance gain over the baseline (10%). Due its simplicity, we choose this as the triangle-closing model for the rest of the paper.

Note that the above methodology could be extended to edge creations other than triangle-closing. We chose to focus on the triangle-closing edges for two reasons. First, a high fraction of all edges created fall into this category, and hence an understanding of triangle-closing edges is an important first step towards understanding the overall network evolution. Second, with the exception of quite simplistic models, it is computationally infeasible to compute the likelihood at a distance greater than two hops as the number of nodes and possible paths increases dramatically.

6. NODE AND EDGE ARRIVAL PROCESS

In this section we turn our focus to the edge initiation process that determines which node is responsible for creating a new edge (Section 6.1), and then to the process by which new nodes arrive into the network (Section 6.2).

6.1 Edge initiation

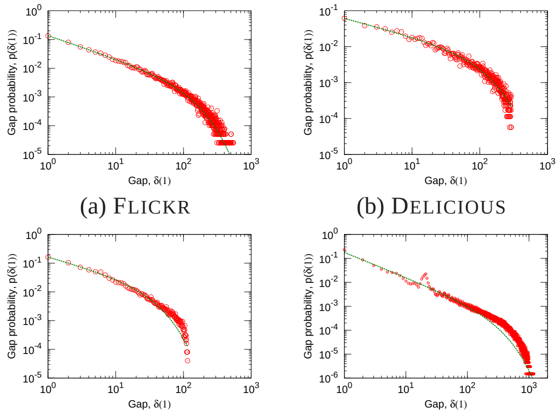
In the following we assume that the sequence and timing of node arrivals is given, and we model the process by which nodes initiate edges. We begin by studying how long a node remains active in the social network, and then during this active lifetime, we study the specific times at which the node initiates new edges.

6.1.1 Node lifetime

To avoid truncation effects, we only consider those nodes whose last-created edge is in the first half of all edges in the data. Recall that the lifetime of a node u is $a(u) = t_{d(u)}(u) - t_1(u)$. We evaluate the likelihood of various distributions and observe that node lifetimes are best modeled by an exponential distribution, $p_{\ell}(a) = \lambda \exp(-\lambda a)$. Figure 7 gives the plot of the data and the exponential fits, where time is measured in days. In Table 5, the row corresponding to λ gives the values of fitted exponents. We note that the exponential distribution does not fit well the nodes with very short lifetimes, i.e., nodes that are invited into the network, create an edge and never return. But the distribution provides a very clean fit for nodes whose lifetime is more than a week.

degree d	power law	power law exp. cutoff	log normal	stretched exp.
1	9.84	**12.50**	11.65	12.10
2	11.55	**13.85**	13.02	13.40
3	10.53	**13.00**	12.15	12.59
4	9.82	**12.40**	11.55	12.05
5	8.87	**11.62**	10.77	11.28
avg., $d \leq 20$	8.27	**11.12**	10.23	10.76

Table 3: Edge gap distribution: percent improvement of the log-likelihood at MLE over the exponential distribution.



(a) FLICKR (b) DELICIOUS

(c) ANSWERS (d) LINKEDIN

Figure 8: Edge gap distribution for a node to obtain the second edge, $\delta(1)$, and MLE power law with exponential cutoff fits.

6.1.2 Time gap between the edges

Now that we have a model for the lifetime of a node u, we must model that amount of elapsed time between edge initiations from u. Let $\delta_u(d) = t_{d+1}(u) - t_d(u)$ be the time it takes for the node u with current degree d to create its $(d+1)$-st out-edge; we call $\delta_u(d)$ the *edge gap*. Again, we examine several candidate distributions to model edge gaps. Table 3 shows the percent improvement of the log-likelihood at the MLE over the exponential distribution. The best likelihood is provided by a power law with exponential cutoff: $p_g(\delta(d); \alpha, \beta) \propto \delta(d)^{-\alpha} \exp(-\beta\delta(d))$, where d is the current degree of the node. (Note that the distribution is neither exponential nor Poisson, as one might be tempted to assume.) We confirm these results in Figure 8, in which we plot the MLE estimates to gap distribution $\delta(1)$, i.e., distribution of times that it took a node of degree 1 to add the second edge. In fact, we find that all gaps distributions $\delta(d)$ are best modeled by a power law with exponential cut-off (Table 3 gives improvements in log-likelihoods for $d = 1, \ldots, 5$ and the average for $d = 1, \ldots, 20$.)

For each $\delta(d)$ we fit a separate distribution and Figure 9 shows the evolution of the parameters α and β of the gap distribution, as a function of the degree d of the node. Interestingly, the power law exponent $\alpha(d)$ remains *constant* as a function of d, at almost the same value for all four networks. On the other hand, the exponential cutoff parameter $\beta(d)$ increases *linearly* with d, and varies by an order of magnitude across networks; this variation models the extent to which the "rich get richer" phenomenon manifests in each network. This means that the slope α of power-law part remains constant, only the exponential cutoff part (parameter β) starts to kick in sooner and sooner. So, nodes add their $(d+1)^{st}$ edge faster than their d^{th} edge, i.e., nodes start to create more and more edges (sleeping times get shorter) as they get older (and have higher degree). So, based on Figure 9, the overall gap distribution can be modeled by $p_g(\delta|d; \alpha, \beta) \propto \delta^{-\alpha} \exp(-\beta d\delta)$.

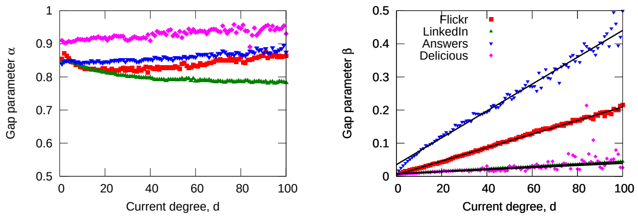
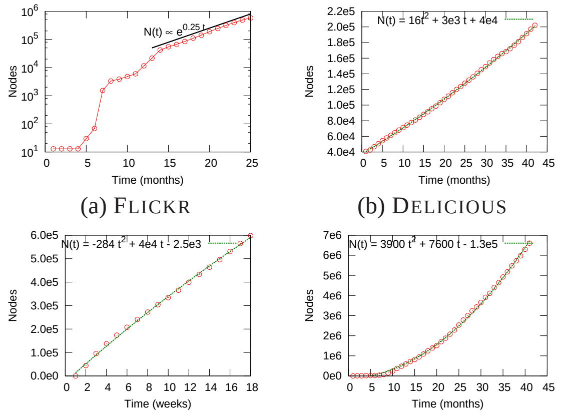


Figure 9: Evolution of the α and β parameters with the current node degree d. α remains constant, and β linearly increases.



(a) FLICKR (b) DELICIOUS

(c) ANSWERS (d) LINKEDIN

Figure 10: Number of nodes over time.

Given the above observation, a natural hypothesis would be that nodes that will attain high degree in the network are in some way a priori special, i.e., they correspond to "more social" people who would inherently tend to have shorter gap times and enthusiastically invite friends at a higher rate than others, attaining high degree quickly due to their increased activity level. However, this phenomenon does not occur in any of the networks. We computed the correlation coefficient between $\delta(1)$ and the final degree $d(u)$. The correlation values are -0.069 for DELICIOUS, -0.043 for FLICKR, -0.036 for ANSWERS, and -0.027 for LINKEDIN. Thus, there is almost no correlation, which shows that the gap distribution is independent of a node's final degree. It only depends on node lifetime, i.e., high degree nodes are not a priori special, they just live longer, and accumulate many edges.

6.2 Node arrivals

Finally, we turn to the question of modeling node arrivals into the system. Figure 10 shows the number of users in each of our networks over time, and Table 4 captures the best fits. FLICKR grows exponentially over much of our network, while the growth of other networks is much slower. DELICIOUS grows slightly superlinearly, LINKEDIN quadratically, and ANSWERS sublinearly. Given these wild variations we conclude the node arrival process needs to be specified in advance as it varies greatly across networks due to external factors.

7. A NETWORK EVOLUTION MODEL

We first take stock of what we have measured and observed so far. In Section 6.2, we analyzed the node arrival rates and showed that they are network-dependent and can be succinctly represented by a node arrival function $N(t)$ that is either a polynomial or an exponential. In Section 6.1, we analyzed the node lifetimes and showed they are exponentially distributed with parameter λ. In Section 4.1, we argued that the destination of the first edge of a

Network	$N(t)$
Flickr	$\exp(0.25t)$
Delicious	$16t^2 + 3000t + 40000$
Answers	$-284t^2 + 40000t - 2500$
LinkedIn	$3900t^2 + 76000t - 130000$

Table 4: Node arrival functions.

node is chosen proportional to its degree (i.e., preferentially attached). In Section 6.1, we analyzed the time gaps between edge creation at a node and showed they can be captured by a power law with exponential cutoff, with parameters α, β. In Section 5, we showed that most of the edges span two hops, and the simple random-random triangle-closing model works well.

Motivated by these observations, we now present a complete network evolution model. Our model is parameterized by $N(\cdot), \lambda, \alpha, \beta$, and operates as follows.

1. Nodes arrive using the node arrival function $N(\cdot)$.
2. Node u arrives and samples its lifetime a from the exponential distribution $p_\ell(a) = \lambda \exp(-\lambda a)$.
3. Node u adds the first edge to node v with probability proportional to its degree.
4. A node u with degree d samples a time gap δ from the distribution $p_g(\delta|d; \alpha, \beta) = (1/Z)\delta^{-\alpha} \exp(-\beta d \delta)$ and goes to sleep for δ time steps.
5. When a node wakes up, if its lifetime has not expired yet, it creates a two-hop edge using the random-random triangle-closing model.
6. If a node's lifetime has expired, then it stops adding edges; otherwise it repeats from step 4.

The values of $N(\cdot)$ for the four networks are given in Table 4 and the values of α, β, λ are given in Table 5.

Note that one could also use more sophisticated edge placement techniques, like random surfer model [5] or other triangle-closing techniques as discussed in Section 5.1. For example, in step 5, a node u can pick a sequence of nodes $(u = w_0, w_1, \ldots, w_k = w)$, where each w_i is picked uniformly from the neighbors of w_{i-1}, and the sequence length k is chosen from the distribution in Figure 4. Node u then links to w.

7.1 Gaps and power law degree distribution

We now show that our model, node lifetime combined with gaps, produces power law out-degree distribution. This is interesting as a model of temporal behavior (lifetime plus gaps) gives rise to a network property.

THEOREM 1. *The out-degrees are distributed according to a power law with exponent* $1 + \frac{\lambda\Gamma(2-\alpha)}{\beta\Gamma(1-\alpha)}$.

PROOF SKETCH. We first compute the normalizing constant Z of the gap distribution $p_g(\delta|d; \alpha, \beta)$:

$$Z = \int_0^\infty \delta^{-\alpha} e^{-\beta d \delta} \mathrm{d}\delta = \frac{\Gamma(1-\alpha)}{(\beta d)^{1-\alpha}}. \tag{1}$$

Let a be the lifetime sampled from the exponential distribution $p_\ell(a) = \lambda \exp(-\lambda a)$. Recall the edge creation process: a node adds its first edge and samples the next gap $\delta(1)$ according to $p_g(\cdot)$, sleeps for $\delta(1)$ time units, creates the second edge, samples a new gap $\delta(2)$ according to $p_g(\cdot)$, sleeps for $\delta(2)$ units, and so on until it uses up all of its lifetime a. This means that for a node u with lifetime $a = a(u)$ and final degree $D = d(u)$, we have

$$\sum_{d=1}^{D} \delta(k) \le a. \tag{2}$$

	Flickr	Delicious	Answers	LinkedIn
λ	0.0092	0.0052	0.019	0.0018
α	0.84	0.92	0.85	0.78
β	0.0020	0.00032	0.0038	0.00036
true	1.73	2.38	1.90	2.11
predicted	1.74	2.30	1.75	2.08

Table 5: Predicted by Theorem 1 vs true degree exponents.

Analogous to (1), we obtain the expected time gap $E(\delta|d; \alpha, \beta)$ for a node of degree d:

$$E(\delta|d; \alpha, \beta) = \frac{\Gamma(2-\alpha)}{\Gamma(1-\alpha)}(\beta d)^{-1}. \tag{3}$$

Combining (2) and (3), we relate the lifetime a and the expected final degree D of a node:

$$\sum_{d=1}^{D} \frac{\Gamma(2-\alpha)}{\Gamma(1-\alpha)}(\beta d)^{-1} = \frac{\Gamma(2-\alpha)}{\Gamma(1-\alpha)}\beta^{-1} \sum_{d=1}^{D} d^{-1} \le a. \tag{4}$$

Notice that $\sum_{d=1}^{D} d^{-1} = \Theta(\ln D)$. From (4), the final degree D of a node with lifetime a is

$$D \approx \exp\Big(\frac{\Gamma(1-\alpha)}{\Gamma(2-\alpha)}\beta a\Big).$$

Thus, D is an exponential function of the age a, i.e., $D = r(a) = \exp(\mu a)$, where $\mu = \frac{\Gamma(1-\alpha)}{\Gamma(2-\alpha)}\beta$. Since node lifetimes are exponentially distributed with parameter λ, we now compute the distribution of D as a function of λ and μ as follows:

$$D \sim p_\ell(r^{-1}(D))\Big|\frac{\mathrm{d}r^{-1}(D)}{\mathrm{d}D}\Big| = \frac{\lambda}{\mu D} e^{-(\lambda/\mu)\log D} = \frac{\lambda D^{-(1+\lambda/\mu)}}{\mu}.$$

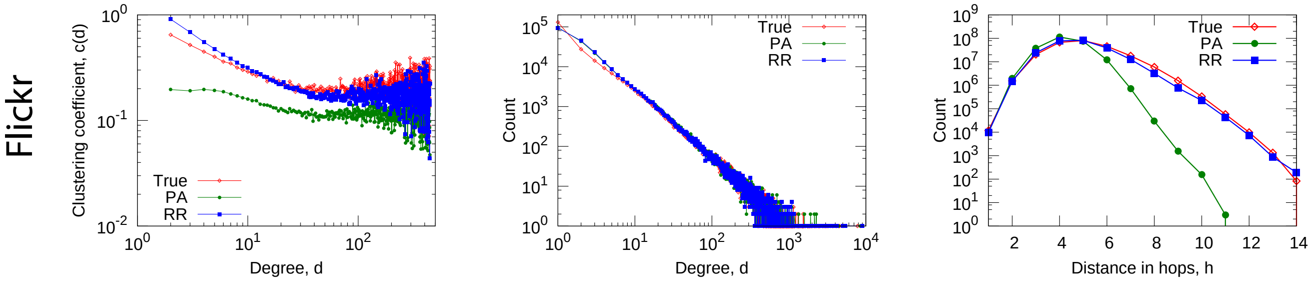
Thus, the degree distribution in our gap model follows a power law with exponent $1 + \lambda/\mu$, completing the proof. □

Validation of the model. We validate the accuracy of our modeling assumptions by empirically estimating the lifetime λ, and gap distribution α, β parameter values for each network. We then apply Theorem 1, which yields the power-law degree exponents produced by our model. Then we empirically measure the true power law degree exponents of the four networks and compare them to predictions of Theorem 1. Table 5 shows the results. Note the predicted degree exponents remarkably agree with the true exponents, validating our model. This is interesting as we specified the model of temporal node behavior (lifetime+gaps) that results in a accurate structural network property (power-law degree distribution).

7.2 Unfolding network evolution

To further our understanding of the network evolution, especially the edge creation process, we perform the following semi-simulation. We consider the real network $G_{T/2}$ and evolve it from $t = T/2, \ldots, T$ using the random-random model to obtain a network G'_T. At the end of the evolution, we compare the macroscopic properties of G'_T and G_T. For completeness, we also compare the results to the vanilla PA model.

More precisely, we evolve $G_{T/2}$ by considering all the edges that were created after time $T/2$ between the nodes in $G_{T/2}$. (We do not allow new nodes to join $G_{T/2}$.) We consider two different processes to place these new edges. In the first process (PA), we select two nodes preferentially, with probabilities proportional to their degrees, and add an edge. In the second process (RR), we use the random-random triangle-closing model, i.e., we first select a node preferentially and then pick a node two hops away using the random-random model.



(a) Clustering coefficient (b) Degree distribution (c) Geodesic distance

Figure 11: We take FLICKR network at first half of its evolution. Then we simulate the evolution using our model and PA for the second half, and compare the obtained networks with the real FLICKR network. Notice our model matches the macroscopic statistical properties of the true FLICKR network very well, and in fact much better than PA.

Figure 11 shows results for FLICKR: clustering coefficient, degree distribution, and pairwise distance histogram for the true data, and the two simulations. The random-random model matches the true network well and outperforms than the PA process. Similar results also hold for other networks; we omit these plots for brevity.

8. CONCLUSIONS

In this paper we present a microscopic analysis of the edge-by-edge evolution of four large online social networks. The use of the maximum-likelihood principle allows us to quantify the bias of new edges towards the degree and age of nodes, and to objectively compare various models such as preferential attachment. In fact, our work is the first to quantify the amount of preferential attachment that occurs in networks.

Our study shows that most new edges span very short distances, typically closing triangles. Motivated by these observations, we develop a complete model of network evolution, incorporating node arrivals, edge initiation, and edge destination selection processes. While node arrivals are mostly network-specific, the edge initiation process can be captured by exponential node lifetimes and a “gap” model based on a power law with exponential cutoff. We arrive at an extremely simple yet surprisingly accurate description of the edge destination selection in real networks. Our model of network evolution can be used to generate arbitrary-sized synthetic networks that closely mimic the macroscopic characteristics of real social networks.

9. REFERENCES

[1] R. Albert and A.-L. Barabasi. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[2] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74, 47, 2002.

[3] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. In *12th KDD*, pages 44–54, 2006.

[4] I. Bezáková, A. Kalai, and R. Santhanam. Graph model selection using maximum likelihood. In *23rd ICML*, pages 105–112, 2006.

[5] A. Blum, H. Chan, and M. Rwebangira. A random-surfer web-graph model. In *ANALCO*, 2006.

[6] B. Bollobas and O. Riordan. Mathematical results on scale-free random graphs. In S. Bornholdt and H. Schuster, editors, *Handbook of Graphs and Networks*, pages 1–37. Wiley–WCH, 2002.

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

[8] S. Dorogovtsev and J. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford Univ. Press, 2003.

[9] P. Erdős and A. Rényi. On the evolution of random graphs. *Mathematical Institute of the Hungarian Academy of Science*, 1960.

[10] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262, 1999.

[11] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. *Software Practice and Experience*, 34(2):213–237, 2004.

[12] D. Krackhardt and M. Handcock. Heider vs. Simmel: Emergent features in dynamic structure. In *Statistical Network Analysis: Models, Issues, and New Directions*, pages 14–27, 2007.

[13] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *12th KDD*, pages 611–617, 2006.

[14] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *41st FOCS*, pages 57–65, 2000.

[15] J. Leskovec and C. Faloutsos. Scalable modeling of real graphs using Kronecker multiplication. In *24th ICML*, pages 497–504, 2007.

[16] J. Leskovec, J. M. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM TKDD*, 1(1):2, 2007.

[17] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *12th CIKM*, pages 556–559, 2003.

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

[19] A. Ntoulas, J. Cho, and C. Olston. What’s new on the web? The evolution of the web from a search engine perspective. In *13th WWW*, pages 1–12, 2004.

[20] X. Shi, L. A. Adamic, and M. J. Strauss. Networks of strong ties. *Physica A*, 378(1):3347, 2007.

[21] S. Strogatz. Exploring complex networks. *Nature*, 410, 2001.

[22] S. Wasserman and P. Pattison. Logit models and logistic regressions for social networks. *Psychometrika*, 60:401–425, 1996.

[23] C. Wiuf, M. Brameier, O. Hagberg, and M. P. Stumpf. A likelihood approach to analysis of network data. *PNAS*, 103(20), 2006.