Structure and Evolution of Online Social Networks

Ravi Kumar Jasmine Novak Andrew Tomkins
Yahoo! Research, 701 First Ave, Sunnyvale, CA 94089.
{ravikumar, jnovak, atomkins}@yahoo-inc.com

ABSTRACT

In this paper, we consider the evolution of structure within large online social networks. We present a series of measurements of two such networks, together comprising in excess of five million people and ten million friendship links, annotated with metadata capturing the time of every event in the life of the network. Our measurements expose a surprising segmentation of these networks into three regions: singletons who do not participate in the network; isolated communities which overwhelmingly display star structure; and a giant component anchored by a well-connected core region which persists even in the absence of stars.

We present a simple model of network growth which captures these aspects of component structure. The model follows our experimental results, characterizing users as either passive members of the network; inviters who encourage offline friends and acquaintances to migrate online; and linkers who fully participate in the social evolution of the network.

Categories and Subject Descriptors: H.2.8 [**Data Management**]: Database Applications—*Data Mining*

General Terms: Measurements, Theory

Keywords: graph mining, small-world phenomenon, graph evolution, social networks, stars

1. INTRODUCTION

In this paper, we study the evolution of large online social networks. To our knowledge, this is the first detailed evaluation of the growth processes that control online social networks in the large.

The power of people interacting with people in an online setting has driven the success or failure of many companies in the internet space. Social media applications such as flickr (`flickr.com`) or myspace (`www.myspace.com`) have exploded in popularity, shocking pundits and realigning the online landscape. Similarly, classically successful online destinations that deal largely in the buying and selling of physical items owe much of their success to the power of online networks; consider for instance the product reviews of Amazon (`amazon.com`) or the reputation mechanism of Ebay (`ebay.com`). In fact, social networks have become the subject of numerous startup companies in their own right, offering each user the promise of managing her own social network as a valuable resource to be shepherded and grown.

As the stock of social networks has grown, so too has interest in the academic community. Offline networks have been the subject of intense academic scrutiny for many decades, but the availability of large online social networks has raised new sets of questions. Much work to date has focused on the structure of a static snapshot of an evolving social network. In this paper, we have access to the entire lifetime of two large social networks, and hence we are able to study their dynamic properties. We study the social network of Flickr, and Yahoo! 360.

SUMMARY OF FINDINGS. We begin with a study of the overall properties of the network. We show that the density of the network, which measures the amount of interconnection per person, follows the same unexpected pattern in both networks: rapid growth, decline, and then slow but steady growth. We postulate based on the timing of the events that the pattern is due to the activities of early adopters who create significant linkages in their exploration of the system, followed by a period of rapid growth in which new members join more quickly than friendships can be established, settling finally into a period of ongoing organic growth in which both membership and linkage increases.

Next, we classify members of a social network into one of three groups: the singletons, the giant component, and the middle region, as follows.

Singletons. The singletons are degree-zero nodes who have joined the service but have never made a connection with another user in the social network. They may viewed as loners who do not participate actively in the network.

Giant component. The giant component represents the large group of people who are connected to one another through paths in the social network. These people find themselves connected directly or indirectly to a large fraction of the entire network, typically containing most of the highly active and gregarious individuals.

Middle region. The middle region is the remainder. It consists of various isolated communities, small groups who interact with one another but not with the network at large. We will show that this group may represent a significant fraction of the total population.

We begin with a detailed study of the middle region, which represent about 1/3 of the users of Flickr and about 10% of the users of Yahoo! 360. We show first that over significant periods of time, and significant fractions of growth in the network (exceeding 10x), the fraction of users who exist in isolated communities of a particular size remains remarkably stable, even though the particular users change dramatically.

We study the migration patterns of isolated communities, seeking insight to how these communities grow and merge. Our find-

KDD'06, August 20–23, 2006, Philadelphia, Pennsylvania, USA.

ings are quite surprising. The likelihood that two isolated communities will merge together is unexpectedly low. Evolution in the middle region is characterized by two processes: isolated communities grow by a single user at a time, and then may eventually be merged into the giant component; these processes capture the majority of activity within the middle region. Furthermore, we present a structural finding showing that almost all the isolated communities are in fact *stars*: a single charismatic individual (in the online sense) linked to a varying number of other users who have very few other connections.

We study the formation of these stars and show that they grow rapidly, and then either merge into the giant component or cease growth when the individual holding the community together loses focus on growing the network.

Next, we turn to the structure of the giant component. We show that, in this region, the merging of stars does not represent the defining structural characteristic of the giant component. Instead, merging stars represent a sort of outer layer of the region, around a much more tightly-connected core of active members who are the heart of the entire social network. Removal of all stars from the giant component has no significant impact on the connectivity of the remaining nodes.

Over time, the average distance between users in the giant component is seen to fall. This surprising result has been observed in other settings [22]; we show it here for online social networks.

Given these findings, we draw some high-level behavioral conclusions about the structure and evolution of online social networks. First, there are two distinct ways that people join the network: they may register by actively seeking out the network, or they may be invited by a friend or colleague. The stars in the middle region are largely characterized by invitations, and the individuals performing the invitations are typically motivated more by migrating and existing offline social network into an online setting, rather than building new connections online. On the other hand, the members of the well-connected core of the giant component are the reverse: they are highly focused on the evolution of the internal network of which they are perhaps the key piece.

MODEL. Based on these observations, we propose a rudimentary model of network evolution in which we attempt to capture the salient properties of our measurements using as small a parameter space as possible. Our model uses a notion of biased preferential attachment which introduces a disparity between the relative ease of finding potential online connections within the giant component, and the relative difficulty of locating potential connections out in the isolated communities. The model accurately reproduces the quantitatively very different component structure of Flickr and Yahoo! 360.

ORGANIZATION. The paper is organized as follows. In Section 2, we discuss the related work on theoretical and experimental analysis of large-scale social and other related networks. In Section 3, we describe our experiments and observations about the Flickr and Yahoo! 360 social networks. In Section 4, we outline the biased preferential attachment model for online social network evolution. In Section 5, we discuss our findings and outline thoughts for future work. Finally, Section 6 concludes the paper.

2. RELATED WORK

Large real-world graphs such as the world-wide web, internet topology, phone call graphs, social networks, email graphs, biological networks, and linguistic networks have been extensively studied from a structural point of view. Typically, these studies address properties of the graph including its size, density, degree distributions, average distance, small-world phenomenon, clustering coefficient, connected components, community structures, etc. We briefly outline some of the work in this area. Faloutsos, Faloutsos, and Faloutsos [13] made a crucial observation showing that the degree distribution on the internet follow a power law. Subsequently, an intense body of work followed in both computer science and physics communities, aimed at studying properties of large-scale real-world graphs. Power law degree distributions were also noted on the graph defined by the world-wide web [21, 4]. Broder et al [8] studied the world-wide web from a connectivity point of view and showed that it has a large strongly connected component. Several other studies have also shown that the average diameter of the web is quite small [8, 3]. Online friendship and email graphs have been studied in the context of explaining and analyzing friendships [18] and demonstrating the small-world and navigability properties of these graphs [23, 9, 1]. For surveys of analysis of large graphs, the readers are referred to [29, 28, 2, 25, 11, 10, 16].

Many of these above studies were performed on static graphs whereas most real-world graphs are evolving in nature. In fact, there are very papers that study the evolution of real-world graphs; this is partly because of the difficulty in obtaining temporal information about every node/edge arrival in an evolving real-world graph. A typical way this problem is addressed is to take snapshots of the graph at various points in time and use these snapshots to make inferences about the evolutionary process. This approach was used to study the linkage pattern of blogs and the emergence of bursty communities in the blogspace [19]. Structural properties of different snapshots of the world-wide web graph was studied by Fetterly et al and Cho et al [14, 27]. Recently, Leskovec, Kleinberg, and Faloutsos [22] considered citation graphs and showed that these exhibit densification and shrinking diameters over time.

A parallel body of work is concerned with developing tractable mathematical models for massive graphs. Because of their evolutionary nature and their power law degree distributions, these graphs cannot be modeled by traditional Erdö–Rényi random graphs [12, 6]. However, there have been a few alternate models that are more faithful to observed properties. One is the so-called configuration model, which chooses a graph uniformly at random from all graphs with a prescribed degree distribution [5, 24, 26]; the degree distribution can be set to match practical observations and is usually a power law. Another approach is to use a generative model to describe the evolution of graphs. A typical example is the copying or the preferential attachment model [20, 4]: nodes arrive one by one, and link themselves to a pre-existing node with probability proportional to the degree of the latter. This "rich get richer" principle can be analytically shown to induce power-law degree distributions. Kleinberg [17, 15] proposed a model to explain the small-world phenomenon and navigability in social networks; see also [30]. Leskovec et al [22] proposed a forest-fire graph model to explain the decreasing diameter phenomenon observed in citation graphs. For a survey of mathematical analysis of some of these models, the readers are referred to [7, 16].

3. MEASUREMENTS

In this section we detail our study on two online social networks at Yahoo!. Each social network is presented as a *directed time graph* $G = (V, E)$, i.e., every node $v \in V$ and directed edge $\langle u, v\rangle \in E$ in the graph G has an associated time stamp v_t and $\langle u, v\rangle_t$ indicating the exact moment when the particular node v or the edge e became part of the graph [19]. In particular, for any time t, there is a natural graph G_t that comprises all the nodes and edges that have arrived up until time t; here we assume that the end

points of an edge always arrive during or before the edge itself. We use timegraph to refer to properties that are specific to the evolution and use graph to refer to the graph G_{Jan2006} as the *final graph*. We note that our study of timegraphs is of much finer granularity than almost all of previous such studies in that we know the *exact* moment of each node/edge arrival.

3.1 Datasets

The dataset consists of two online social networks at Yahoo! — Flickr and Yahoo! 360. Each of these social networks is presented as a timegraph. For privacy reasons, all the data used in the paper were provided to us after appropriate anonymization. For confidentiality reasons, we do not specify the exact number of nodes or edges in these timegraphs but only provide a ball-park estimate — this will not in any way affect the presentation of our results or the inferences that can be drawn.

Flickr (`www.flickr.com`) is an active and popular online photo sharing and social networking community. Flickr users can upload and tag photos and share them with their friends or publicly. Each user in Flickr can invite a new friend to Flickr or can add a pre-existing Flickr user as a friend. In Jan 2006, the Flickr timegraph consisted of around one million nodes and around eight million directed edges. The dataset we used had the following anonymized information about each Flickr user: the time when the user became a Flickr member and the list of friends he/she has on Flickr, and for each friend, the time when the user befriended the person. Even though we had the entire Flickr timegraph available, for our experiments, we focused only on the evolution of the timegraph since the Flickr website was publicly launched (Feb 2004); this amounted to about 100 weeks worth of data. We made this decision in order to avoid the initial phase before the public launch when Flickr usage was mostly limited to internal users and the user/friendship addition processes were too skewed to lead to meaningful conclusions.

Yahoo! 360 (`360.yahoo.com`) is a social networking website that is part of the Yahoo! user network. Users of Yahoo! 360 can add contacts and invite other users to the 360 network. Yahoo! 360 is primarily used to share a blog or photo albums among the friends of a user. In Jan 2006, the Yahoo! 360 timegraph consisted of around five million nodes and a around seven million directed edges. As in Flickr, we used an anonymized timegraph and as before, chose to discard the initial segment of the timegraph in order to filter out pre-launch noise/bias. This resulted in about 40 weeks worth of data.

3.2 Basic timegraph properties

In this section we consider three basic properties of these timegraphs. The first property we consider is the *reciprocity* of a directed graph, that is, the fraction of directed edges $\langle u, v\rangle$ such that $\langle v, u\rangle$ also exists in the graph. The goal is to understand:

Are friendships reciprocal in online social networks?

The reciprocity of the Flickr final graph is around 70.2%, and that of the Yahoo! 360 final graph is around 84%. Thus, friendship edges are highly mutual. In fact, a finer analysis shows that not only are many friendship edges reciprocal but in fact many reciprocal edges are formed almost simultaneously. Figure 1 shows for reciprocal edges $\langle u, v\rangle_t$ and $\langle v, u\rangle_{t'}$ in the Flickr final graph, the distribution of $|t - t'|$, i.e., the delay (in days) of the reciprocity. We see that an overwhelming fraction of reciprocal edges arrive within a day of each other. A similar phenomenon is also seen in the Yahoo! 360 final graph. From these observations, we conclude that for the purposes of analysis and for simplicity of exposition, we can pretend that the graph is undirected. So, for the remainder of the paper, we deal only with undirected graphs and treat the Flickr and Yahoo! 360 graphs to be undirected by removing all uni-directional edges.

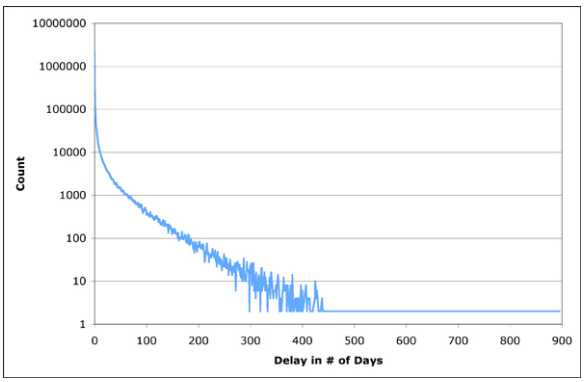


Figure 1: Delay (in days) of reciprocity in Flickr final graph.

Next, we look at the *density* of these graphs, that is, the ratio of undirected edges to nodes, of the timegraphs. In a recent work, Leskovec et al [22] observed that certain citation graphs became denser over time. We wish to ask a similar question for online social networks:

How does the density of online social networks behave over time?

It turns out that the density of social networks as a function of time is non-monotone. Figure 2 shows the density of the Flickr and Yahoo! 360 timegraphs. In both the plots there are three clearly

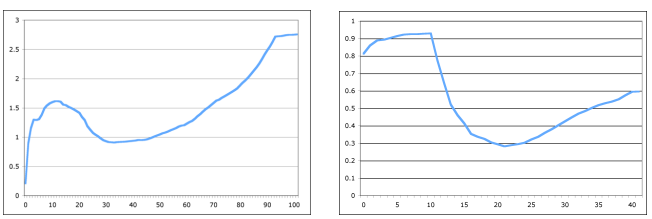


Figure 2: Density of Flickr and Yahoo! 360 timegraphs, by week.

marked stages: an initial upward trend leading to a peak, followed by a dip, and the final gradual steady increase. We believe that this is due to the following social phenomenon. Right after the launch, there is an initial euphoria among a few enthusiasts who join the network and frantically invite many of their friends to join; this gives rise to the *first stage* that culminates in a peak. The *second stage* corresponds to a natural dying-out of this euphoria and this leads to the dip. The *third stage* corresponds to true organic growth of the network (when more and more people know about the network). This growth takes takes over the node/edge creation activities, slowly overwhelms the dip, and eventually leads to a steady increase in density. To the best of our knowledge, this phenomenon has not been observed before in real social networks (again, perhaps due to the lack of suitable data).

For completeness, we also look at the degree distribution of these graphs. Figure 3(A) shows the degree distribution of the Flickr final graph in log-log scale. As expected, it is a power law. The Yahoo! 360 final graph exhibits an almost identical degree distribution. It is interesting to note the non-monotone shape of this plot for the first three values of the degree (i.e., degree = 0, 1, 2). This peak occurs because of the "invite" option that is often used in adding new

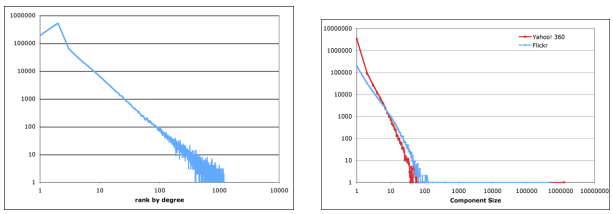


Figure 3: (A) Degree distribution in Flickr final graph. The x-axis is the ranked degree and the y-axis is the number of nodes at this rank. (B) Component size distribution for Flickr and Yahoo! 360 final graph.

people to these networks. Typically, many users join via invitation, and arrive with a single edge already in place. Degree zero nodes have explicitly joined the network without an invitation, and are a smaller fraction of the total user base. We will return to this issue in Section 4.

3.3 Component properties

In this section we study the component structure of the graph in detail. Our goal is to understand the connectivity structure of the graph as it evolves over time. In particular, we ask:

> What is the dynamics of component formation and evolution in social networks?

We apply a simple connected components algorithm on the timegraph by considering the instance at every week. The results for the Flickr and Yahoo! 360 timegraphs are in Figure 4. This plot shows the fraction of nodes in components of various sizes. The intervals representing various horizontal bands were chosen so that the top band represents the largest connected component, which we will call the *giant component*, while the bottom band represents the total number of *singleton* nodes in the graph, with no links in the social network at all. The rest of the bands constitute the *middle region*, consisting of nodes which exist in small isolated neighborhoods. While there are quantitative differences between the plots for Flickr and Yahoo! 360, both the plots share two particularly interesting properties.

1. The fraction of singletons, the fraction of nodes in the giant component, and fraction of nodes in the middle region remain almost constant once a steady state has been reached, despite significant growth of the social network during the period of steady component structure. For example, the Flickr social network grew by a factor of over 13x from the period $x = 40$ to $x = 100$ in the graph, with very little visible change in the fraction of users who occupied components of a certain size. This steady state corresponds to the third stage observed in Figure 2.

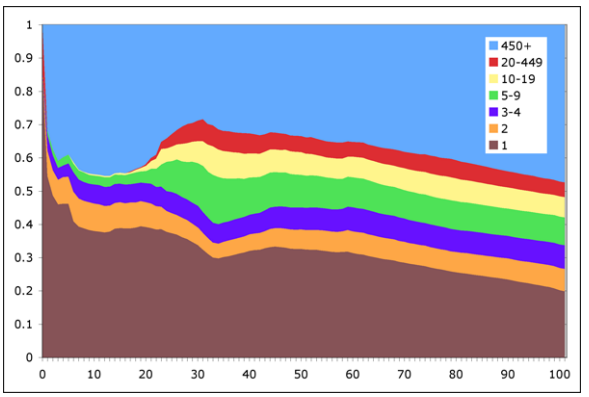
2. In the middle region, each band of the diagram appears fairly constant. In fact, as Figure 3(B) shows, the component size distribution for both datasets follows a power law with exponent -2.74 for the Flickr graph, and -3.60 for Yahoo! 360.

3.4 Structure of the middle region

We now proceed to investigate the formation and structure of the middle region. Our first question was motivated by the evolutionary aspect of the timegraph:

> How do components merge with each another as nodes and edges arrive in social networks?

In particular, it was our assumption when we began this experiment that the non-giant components would grow organically, with a size three component linking to a size four component to form a new



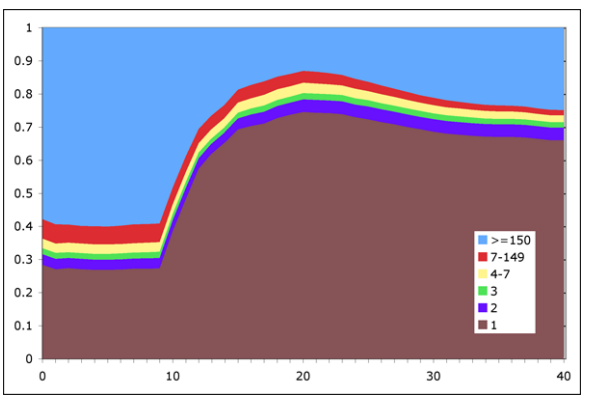


Figure 4: Fraction of nodes in components of various sizes within Flickr and Yahoo! 360 timegraph, by week.

	1	2	3-4	5-9	10-19	20-449	450+
1	205.1						
2	55.9	0.8					
3-4	64.2	0.5	0.3				
5-9	70.8	0.4	0.3	0.2			
10-19	43.9	0.2	0.1	0.1	0.09		
20-449	2.6	0.1	0.01	0.07	0.04	0.03	
450+	315.3	11.5	7.1	5.0	2.4	1.0	0

	1	2	3	4	5-7	8-149	150+
1	584.3						
2	126.1	5.9					
3	69.2	2.6	1.2				
4	43.6	1.5	0.6	0.4			
5-7	66.9	2.3	1.0	0.6	0.9		
8-149	72.6	2.3	1.1	0.6	0.9	1.1	
150+	767.3	54.9	22.4	12.2	15.7	13.0	0.1

Table 1: Sizes of components in Flickr and Yahoo! 360 timegraphs when merging, in 1000's of nodes.

component of size 7, and so forth. Table 1 shows how component merges happen in both Flickr and Yahoo! 360 timegraphs. The (i, j)-th entry of this symmetric table gives the number of times during the evolution of the timegraph that a component of size i merges with a component of size j.

Strikingly, almost all the mass in this table is in the bottom row and the left column, implying that the component merges are of primarily two types: singletons merging with the current non-giant components and the giant component, and non-giant components, including singletons, merging with the giant component.

That is, it is surprisingly rare during the evolution of the timegraph that two non-giant components merge to produce another non-giant component.

Our next goal is to understand the consequences of this observed phenomenon and its impact on the structure of the middle region. Indeed, if most of the component merges are characterized by the above two types, it is natural to speculate that this is caused by some special node in the non-giant component that serves to "attract" the incoming singleton. Notice that if this were to happen, it would lead to many middle region *stars*, that is, components with a center of high degree and many low-degree nodes connected to the center. We ask:

> Do the components in the middle region have any special structure, and in particular, are they stars?

First, to be able to observe this phenomenon, we need a reasonably robust definition of what a star is. We define a star to be connected component with the following two properties: it has one or two nodes (centers) that have an edge to most of the other nodes in the component and it contains a relatively large number of nodes that have an edge solely to one of these centers. More formally, let U be the nodes in a connected component that is not the giant component. Trivially, U is a star if $|U| = 2$. Otherwise, let $C \subseteq U$ be the set of nodes with degree more than $|U|/2$ and let $T \subseteq U$ be the set of nodes with degree equal to one. For a parameter $k \in (0, 1)$, we define U to be a *star* if $|C| \in \{1, 2\}$ and $|T|/|U \setminus C| > k$; we call C the *centers* of the star and $|T|$ the *twinkles*. In our experiments, we set $k = 0.6$ in the above definition.

Based on this definition of a star, we analyze the final graphs of both Flickr and Yahoo! 360. In the Flickr final graph, 92.8% of the middle region was composed of stars; in total there were 69,532 centers and 222,564 twinkles. In the Yahoo! 360 final graph, 88.7% of the middle region was composed of stars; there were 147,071 centers and 264,971 twinkles. Thus, there is an overwhelming number of stars in the middle region, validating our hypothesis that each component in the middle region has a center and the singleton node joins the center to become a twinkle. We will make heavy use of this characterization in order to develop a generative model which produces an appropriate middle region.

In fact, our hypothesis is further strengthened when we examine this process more closely. Call a star *non-trivial* if it has more than two nodes and let u be the center of a non-trivial star. Figure 5(A) shows the distribution of the time lag between first twinkle u and the last twinkle u' to join the star, i.e., the distribution of $t' - t$ (in weeks) where $\langle u, v \rangle_t$ is the edge that adds the first twinkle and $\langle u', v \rangle_{t'}$ is the edge that adds the last twinkle. As we see, the

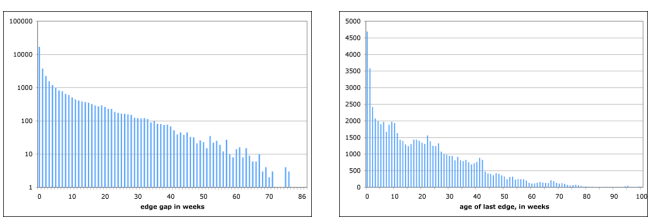


Figure 5: (A) Distribution of time lag (in weeks) between the first and last twinkle addition to non-trivial stars in the Flickr final graph. (B) Age of non-trivial stars in the Flickr final graph.

distribution is sharply decreasing, suggesting that stars are formed rather quickly. We next analyze the *age* of stars, which is the time since the last edge arrival in the star. Figure 5(B) shows the age of stars in the Flickr final graph. Again, a large fraction of stars are more than 10 weeks old. This suggests that the middle section consists of stars that are formed quickly but have not been absorbed into the giant component yet.

Similar results were also observed for Yahoo! 360 final graph. For sake of brevity, we do not present these results.

3.5 Structure of the giant component

In this section we analyze the structure of the giant component. The most natural question to ask is:

> How does the diameter of the social network behave as a function of time?

We study the diameter of the giant component. Formally, the diameter is the maximum over all pairs in the giant component of the shortest path connecting the pair. This measure is not robust in general, as a single long path in the component could result in an enormous diameter. Thus, we turn instead to the *average diameter*, which is defined as the length of the shortest path between a random pair of nodes. For comparison, we also consider the *effective diameter*, which is defined as the 90-th percentile of the shortest path lengths between all pairs of nodes; this quantity was used in [22]. We estimate both these quantities by sampling sufficiently many pairs of nodes in the giant component uniformly at random.

For the giant component in the Flickr final graph, we compute the average diameter to be 6.01 and the effective diameter to be 7.61. For the giant component in the Yahoo! 360 final graph, the corresponding values are 8.26 and 10.47 respectively. Notice that these are slightly higher values than the one suggested by the "six-degrees of separation" folklore. Figure 6 shows diameter as a function of time in the Flickr and Yahoo! 360 timegraphs. The shape of this curve has high correlation with that of density over time, which exhibited three distinct stages in the evolution of the timegraph. We note that the three stages in Figure 6 exactly correspond to the three stages in Figure 2. In the first stage, the diameter is almost flat. In the next stage, where the edge density drops, the diameter grows till it reaches a peak. In the third stage, when the edge density starts increasing, the diameter starts decreasing.

A similar phenomenon of shrinking diameter was recently observed by Leskovec et al [22] in citation graphs. Our study shows that diameter shrinking happens in social networks as well. Again, to the best of our knowledge, this is the first instance of such an observation for online social networks. Well-known models of network growth based on preferential attachment [20, 4] do not have this property (see [7] for details).

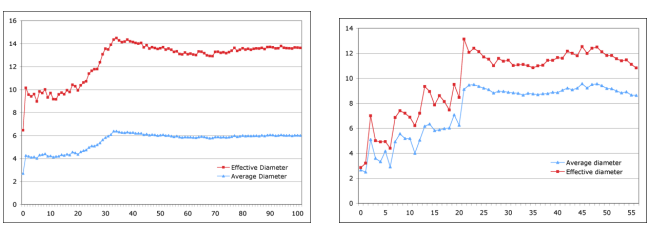


Figure 6: Average and effective diameter of the giant component of Flickr and Yahoo! 360 timegraphs, by week.

We then investigate the structure to see if we can explain the diameter values that were observed. In particular, we ask:

> Does the giant component have a reasonably small core of nodes with high connectivity?

By computing the degree distribution of the nodes in the giant component, we observe that in the Flickr final graph, about 59.7% of

the nodes in the giant component have degree 1. The corresponding number for the Yahoo! 360 final graph was 50.4%. These degree 1 nodes therefore contribute to the increase in diameter values. Suppose we discard these degree 1 nodes in the giant component and analyze the remaining *core*. For the core of the Flickr final graph, the average diameter is 4.45 and the effective diameter is 5.58. For the core of the Yahoo! 360 final graph, the corresponding numbers are 6.52 and 7.95 respectively. This suggests that there is a small core inside the giant component of extremely high connectivity.

Stars are the dominant explanation of the structure outside the giant component. Given the presence of this small core of well-connected nodes, one might naturally ask the following question:

> Are stars merging into the giant component also responsible for the highly-connected core of the giant component?

We identify all stars throughout the life of the time graph, and track them as they merge into the giant component. Based on this tracking, we remove all star centers, and both the original twinkles belonging to that star, and all new degree-1 nodes connected to that star, and ask whether any fragmentation results. In fact, the giant component remains extremely well connected.

Thus, we conclude that the stars represent the primary form of structure outside the giant component, but represent only a thin layer of structure at the outside of the giant component. The true characteristic of the giant component is the well-connected core at the center. Later we will discuss some possible implications of this observation.

4. MODEL

In this section we present a model of the evolution of online social networks. Our goal in developing this model is to explain the key aspects of network growth in as simple a manner as possible, obviating the need for more complex behavioral explanations.

The properties we will seek to reproduce are the following.

Component structure. The model should produce an evolving component structure similar to that of Figure 4. The fraction of users who are singletons, those in the middle region, and those in the giant component should reflect the underlying data. The non-giant component of each size should capture a fraction of the users which matches the empirical observations and should analytically match the observed power law.

Star structure. The non-giant components should be predominantly star-like. Their growth rates should match the growth of the actual data.

Giant component structure. The nodes making up the giant component should display a densely-connected core and a large set of singleton hangers-on, and the relationship between these regions should explain the average distance of the giant component.

4.1 Description of the model

Our model is generative, and informally proceeds as follows. There are three types of users: passive, linkers, and inviters. *Passive users* join the network out of curiosity or at the insistence of a friend, but never engage in any significant activity. *Inviters* are interested in migrating an offline community into an online social network, and actively recruit their friends to participate. *Linkers* are full participants in the growth of the online social network, and actively connect themselves to other members.

At each timestep, a node arrives, and is determined at birth to be passive, linker, or inviter according to a coin toss. During the same timestep, ε edges arrive and the following happens for each edge. The source of the edge is chosen at random from the existing inviters and linkers in the network using preferential attachment; that is, the probability that a particular node is chosen is proportional to its degree plus a constant. If the source is an inviter, then it invites a non-member to join the network, and so the destination is a new node. If the source is a linker, then the destination is chosen from among the existing linkers and inviters, again using preferential attachment. The parameters controlling the model are shown below.

	Description of the parameter
p	User type distribution (passive, inviter, linker)
γ	Preference for giant component over the middle region
ε	Edges per timestep

More formally, the model proceeds as follows. We incrementally build a timegraph $G = (V, E)$. At any point in time, let the set of passives, inviters, and linkers be denoted by P, I, and L respectively, such that $V = P \cup I \cup L$. Let $d(u)$ denote the degree of node u.

At each timestep, a new node arrives, and is assigned to P, I, or L according to the probabilities in p. Let $\beta > 0$ be a parameter. We will define probability distribution D^β over V representing the probability of selecting a node u via a *biased preferential attachment*, as follows:

$$D^\beta(u) \propto \begin{cases} \beta \cdot (d(u)+1) & u \in L \\ d(u)+1 & u \in I \\ 0 & \text{otherwise} \end{cases}$$

Then ε undirected edges arrive, as follows. For each edge (u, v), u is chosen from D^0, where the bias parameter is set to 0. If u is an inviter, then v is a new node, assigned to P. If u is a linker then v is chosen from D^γ. Notice that the initiator of a link is chosen from all non-passive nodes based only on degree. However, once a linker decides to generate a node internal to the existing network, the destination of that node is biased towards other linkers by γ. This reflects the fact that the middle region is more difficult to discover when navigating a social network.

4.2 Simulations

We now evaluate the model with respect to the three families of conditions we hope it will fulfill. We choose suitable parameters for our model and simulate the model. We then examine the properties of the graph created by our model and see how closely it matches that of Flickr and Yahoo! 360 timegraphs. The following table shows the appropriate parameter choices.

	p	γ	ε
Flickr	(0.25, .35, .4)	15	6
Yahoo! 360	(.68, .22, .1)	2	1

We refer to the graphs generated by simulation as Flickr.model and 360.model. We start with the component structure of these simulations and compare them against the actual data. The following table shows the exact match of the fraction of nodes in each of the three main regions.

Data	Singletons	Middle region	Giant component
Flickr	.2	.33	.47
Flickr.model	.20	.33	.47
360	.66	.9	.25
360.model	.66	.9	.25

We now refine the middle region further and compare the simulated versus the actual data.

	1	2	3-4	5-9	10-19	20-449	≥ 450
Flickr	.2	.07	.07	.08	.06	.05	.47
Flickr.model	.2	.06	.08	.08	.06	.03	.47

	1	2	3	4-6	7-149	≥ 150
360	.66	.038	.016	.02	.016	.25
360.model	.66	.04	.02	.02	.01	.25

From our simulation, we see that in terms of components and the structure of the middle region, our model can accurately capture the properties of Flickr and Yahoo! 360 graphs, when the parameters are well-chosen.

5. DISCUSSION AND FUTURE WORK

There are several key takeaway points from our experiments. The first is that online social networks often contain more than half their mass outside the giant component, and the structure outside the giant component is largely characterized by stars. The creation of stars is largely a result of the dynamics of invitation, in which many people are invited to the social network, but only a small fraction choose to engage more deeply than simply responding to an invitation from a friend.

The second key takeaway is that online social networks appears to travel through distinct stages of growth, characterized by specific behavior in terms of density, diameter, and regularity of component structure. We have observed these changes by studying the time graphs of two very different social networks, but we do not yet have a more detailed characterization of the root cause for this progression. It would be attractive to develop a more detailed theory of the adolescence of a social network.

Third, Figure 4 shows a surprising macroscopic component structure in which the total mass of individuals is well spread across a broad range of sizes of isolated communities (or from a graph theoretic perspective, smaller components). We feel that a deeper understanding of the behavior of "middle band" activity versus "core" activity may reveal that the dichotomy is a meaningful reflection of two active by very different types of participants.

Finally, we have presented a simple model which is surprisingly accurate in its ability to capture component growth. It will be interesting to do a more detailed analysis of the model to show that it also predicts diameter of the giant component, in addition to structure of the middle region. Similarly, the model itself is optimized to be the simplest possible approach to reproducing particular aspects of social network structure rather than a detailed model built from the data in order to provide predictive power. Nonetheless, it is interesting to ask whether the best fitting model parameters may be taken as descriptive of the social network in any sense. For example, in the model, Yahoo! 360 displays a smaller relative fraction of active members, compared to the Flickr community, but at the same time offers fewer barriers to discovering isolated sub-communities and incorporating them into the giant component. Is this representative of the underlying reality?

6. CONCLUSIONS

In this paper we studied the structure and evolution of two popular online social networks, namely Flickr and Yahoo! 360. Our study analyzes these graphs from an evolutionary point of view, by keeping track the precise moments when each node and edge arrives in the graph. We show that these quantitatively different graphs share many qualitative properties in common. In particular, we analyzed the structure and evolution of different-sized components and showed the prevalence of "stars", an intriguing feature of online social networks. Based on these empirical observations, we postulated a very simple evolving graph model for social networks and showed by simulation that this model faithfully reflects the observed characteristics. Since our model is fairly simple, we believe it is amenable to mathematical analyses.

Our work raises a number of questions about the behavioral characteristics of the users who contribute to these various different network regions.

Acknowledgments. We are grateful to the Flickr and Yahoo! 360 teams at Yahoo! for their support in data gathering, data analysis, and direction. In particular, we would like to thank Stewart Butterfield, Catarina Fake, Serguei Mourachov, and Neal Sample.

7. REFERENCES

[1] L. A. Adamic and E. Adar. How to search a social network. *Social Networks*, 27(3):187–203, 2005.
[2] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74, 47, 2002.
[3] R. Albert, H. Jeong, and A.-L. Barabasi. Diameter of the world wide web. *Nature*, 401:130–131, 1999.
[4] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.
[5] B. Bollobás. A probabilistic proof of an asymptotic formula for the number of labeled regular graphs. *European Journal of Combinatorics*, 1:311–316, 1980.
[6] B. Bollobas. *Random Graphs*. Cambridge University Press, 2001.
[7] B. Bollobas and O. Riordan. *Mathematical results on scale-free random graphs*, pages 1–37. Wiley–WCH, 2002.
[8] A. Broder, S. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener. Graph structure in the web. *WWW9 / Computer Networks*, 33(1-6):309–320, 2000.
[9] P. S. Dodds, R. Muhamad, and D. J. Watts. An experimental study of search in global social networks. *Science*, 301:827–829, 2003.
[10] S. Dorogovtsev and J. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, 2000.
[11] S. Dorogovtsev and J. Mendes. Evolution of networks. *Advances in Physics*, 51, 2002.
[12] P. Erdös and A. Rényi. On random graphs I. *Publications Mathematics Debrecen*, 6:290–297, 1959.
[13] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262, 1999.
[14] 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.
[15] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *32nd STOC*, pages 163–170, 2000.
[16] J. Kleinberg. Complex networks and decentralized search algorithms. In *Intl. Congress of Mathematicians*, 2006.
[17] J. M. Kleinberg. Navigation in a small world. *Nature*, 406:845, 2000.
[18] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. Structure and evolution of blogspace. *CACM*, 47(12):35–39, 2004.
[19] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. *World Wide Web Journal*, 8(2):159–178, 2005.
[20] 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.
[21] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. *WWW8 / Computer Networks*, 31:1481–1493, 1999.
[22] J. Leskovec and J. K. C. Faloutsos. Graphs over time: Densification laws, shrinking diameters, and possible explanations. In *11th KDD*, pages 177–187, 2005.
[23] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *PNAS*, 102(33):11623–11628, 2005.
[24] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 1995.
[25] M. Newman. The structure and function of complex networks. *SIAM Review*, 45, 2:167–256, 2003.
[26] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physics Reviews E*, 64, 2001.
[27] 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.
[28] S. Strogatz. Exploring complex networks. *Nature*, 410, 2001.
[29] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.
[30] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.