User Interactions in Social Networks and their Implications

Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P. N. Puttaswamy, and Ben Y. Zhao

Computer Science Department, University of California at Santa Barbara

{bowlin, bboe, alessandra, krishnap, ravenben}@cs.ucsb.edu

Abstract

Social networks are popular platforms for interaction, communication and collaboration between friends. Researchers have recently proposed an emerging class of applications that leverage relationships from social networks to improve security and performance in applications such as email, web browsing and overlay routing. While these applications often cite social network connectivity statistics to support their designs, researchers in psychology and sociology have repeatedly cast doubt on the practice of inferring meaningful relationships from social network connections alone. This leads to the question: *Are social links valid indicators of real user interaction? If not, then how can we quantify these factors to form a more accurate model for evaluating socially-enhanced applications?* In this paper, we address this question through a detailed study of user interactions in the Facebook social network. We propose the use of *interaction graphs* to impart meaning to online social links by quantifying user interactions. We analyze interaction graphs derived from Facebook user traces and show that they exhibit significantly lower levels of the "small-world" properties shown in their social graph counterparts. This means that these graphs have fewer "supernodes" with extremely high degree, and overall network diameter increases significantly as a result. To quantify the impact of our observations, we use both types of graphs to validate two well-known social-based applications (RE [Garriss 2006] and SybilGuard [Yu 2006]). The results reveal new insights into both systems, and confirm our hypothesis that studies of social applications should use real indicators of user interactions in lieu of social graphs.

Categories and Subject Descriptors C.2.4 [*Distributed Systems*]: Distributed Applications

General Terms Measurement, Performance

EuroSys'09, April 1–3, 2009, Nuremberg, Germany.

1. Introduction

Social networks are popular infrastructures for communication, interaction, and information sharing on the Internet. Popular social networks such as MySpace and Facebook provide communication, storage and social applications for hundreds of millions of users. Users join, establish *social links* to friends, and leverage their social links to share content, organize events, and search for specific users or shared resources. These social networks provide platforms for organizing events, user to user communication, and are among the Internet's most popular destinations.

Recent work has seen the emergence of a class of socially-enhanced applications that leverage relationships from social networks to improve security and performance of network applications, including spam email mitigation [Garriss 2006], Internet search [Mislove 2006], and defense against Sybil attacks [Yu 2006]. In each case, meaningful, interactive relationships with friends are critical to improving trust and reliability in the system.

Unfortunately, these applications assume that all online social links denote a uniform level of real-world interpersonal association, an assumption disproven by social science. Specifically, social psychologists have long observed the prevalence of low-interaction social relationships such as Milgram's "Familiar Stranger" [Milgram 1977]. Recent research on social computing shows that users of social networks often use public display of connections to represent status and identity [Donath 2004], further supporting the hypothesis that social links often connect acquaintances with no level of mutual trust or shared interests.

This leads to the question: *Are social links valid indicators of real user interaction? If not, then what can we use to form a more accurate model for evaluating socially-enhanced applications?* In this paper, we address this question through a detailed study of user interaction events in Facebook, the most popular social network in the US with over 110 million active users. We download more than 10 million user profiles from Facebook, and examine records of user interactions to analyze interaction patterns across large user groups. Our results show that user interactions do in fact deviate significantly from social link patterns, in terms of factors such as time in the network, method of interaction, and types of users involved.

We make three key contributions through our study. First, we present, to the best of our knowledge, the first large-scale study of the Facebook social network. Unlike Orkut, YouTube or Flickr, Facebook's strong focus on user privacy has generally prevented researchers from "crawling" their network of user profiles. We present detailed analysis of our data set with particular emphasis on user interactions (Section 4), and show that users tend to interact mostly with a small subset of friends, often having no interactions with up to 50% of their Facebook friends. This casts doubt on the practice of extracting meaningful relationships from social graphs, and suggests an alternative model for validating user relationships in social networks. Second, we propose the *interaction graph* (Section 5), a model for representing user relationships based on user interactions. An interaction graph contains all nodes from its social graph counterpart, but only a subset of the links. A social link exists in an interaction graph if and only if its connected users have interacted directly through communication or an application. We construct interaction graphs from our Facebook data and compare their salient properties, such as clustering coefficient and average path lengths, to their social graph counterparts. We observe that interaction graphs demonstrate significantly different properties from those in standard social graphs, including larger network diameters, lower clustering coefficients, and higher assortativity.

Finally, we examine in Section 6 the impact of using different graph models in evaluating socially-enhanced applications. We conduct simulated experiments of the Reliable Email [Garriss 2006] and SybilGuard [Yu 2006] systems on both social and interaction graphs derived from our Facebook data. Our results demonstrate that differences in the two graph models translate into significantly different application performance results.

2. The Facebook Social Network

Before describing our methodology and results, we first provide background information on Facebook's social network. With over 150 million active users (as of February 2009), Facebook is the largest social network in the world, and the number one photo sharing site on the Internet [Facebook 2008]. Facebook allows users to set up personal profiles that include basic information such as name, birthday, marital status, and personal interests. Users establish bidirectional social links by "friending" other users. Each user is limited to a maximum of 5,000 total friends.

Each profile includes a message board called the "Wall" that serves as the primary asynchronous messaging mechanism between friends. Users can upload photos, which must be grouped into albums, and can mark or "tag" their friends in them. Comments can also be left on photos. All Wall posts and photo comments are labeled with the name of the user who performed the action and the date/time of submission. Another useful feature is the Mini-Feed, a detailed log of each user's actions on Facebook over time. It allows each user's friends to see at a glance what he or she has been doing on Facebook, including activity in applications and interactions with common friends. Other events include new Wall posts, photo uploads and comments profile updates, and status changes. The Mini-Feed is ordered by date, and only displays the 100 most recent actions.

Unlike other social networking websites in which all users exist in a global search-space, Facebook is designed around the concept of "networks" that organizes users into membership-based groups. Each network can represent an educational institution (university or high school), a company or organization (called work networks), or a geographic (regional network) location. Facebook authenticates membership in college and work networks by verifying that users have a valid e-mail address from the associated educational or corporate domain. Users can authenticate membership in high school networks through confirmation by an existing member. In contrast, no authentication is required for regional networks. Users can belong to multiple school and work networks, but only one regional network, which they can change twice every sixty days.

A user's network membership determines what information they can access and how their information is accessed by others. By default, a user's profile, including birthday, address, contact information, Mini-Feed, Wall posts, photos, and photo comments are viewable by anyone in a shared network. Users can modify privacy settings to restrict access to only friends, friends-of-friends, lists of friends, no one, or all. Although membership in networks is not required, Facebook's default privacy settings encourage membership by making it very difficult for non-members to access information inside a network.

3. Data Set and Collection Methodology

In this section, we briefly describe our methodology for collecting our Facebook data set. We also present experimental validation of the completeness of our network crawl and describe the types of user interaction data that form the basis for our later examination of interaction graphs.

Data Collection Process. As we mentioned, Facebook is divided into networks that represent schools, institutions, and geographic regions. Membership in regional networks is unauthenticated and open to all users. Since the majority of Facebook users belong to at least one regional network, and most users do not modify their default privacy settings, a large portion of Facebook's user profiles can be accessed by crawling regional networks. As of Spring 2008, Facebook hosted 67 million user profiles, 66.3% of whom (44.3 million) belonged to a regional network. Statistics for regional networks have since been removed.

While other studies of social networks rely on statistical sampling techniques [Mislove 2007] to approximate graph coverage of large social networks, Facebook's partitioning of the user population into networks means that subsets of the social graph can be completely crawled in an iterative fashion. Our primary data set is composed of profile, Wall and photo data crawled from the 22 largest regional networks on Facebook between March and May of 2008. We list these networks and their key characteristics in Table 1. For user interaction activity at finer time granularities, we also performed daily crawls of the San Francisco regional network in October of 2008 to gather data specifically on the Mini-Feed.

To crawl Facebook, we implemented a distributed, multi-threaded crawler using Python with support for remote method invocation (RMI) [Boe 2008]. Facebook provides a feature to show 10 randomly selected users from a given regional network; we performed repeated queries to this service to gather 50 user IDs to "seed" our breadth-first searches of social links on each network. Two dual-core Xeon servers were generally able to complete each crawl in under 24 hours, while averaging roughly 10 MB/s of download traffic. Our completed data set is approximately 500 GB in size, and includes full profiles of more than 10 million Facebook users.

Completeness of Graph Coverage. Prior research on online social networks indicates that the majority of user accounts in the social graph are part of a single, large, weakly connected component (WCC) [Mislove 2007]. Since social links on Facebook are undirected, breadth first crawling of social links should be able to generate complete coverage of the WCC, assuming that at least one of the initial seeds of the crawl is linked to the WCC. The only inaccessible user accounts should be ones that lie outside the regional network of the crawl, have changed their default privacy settings, or are not connected to the WCC.

To validate our data collection procedure and ensure that our crawls are reaching every available user in the WCC, we performed five simultaneous crawls of the San Fransisco regional network. Each crawl was seeded with a different number of user IDs, starting with 50 and going up to 5000. The difference in the number of users discovered by the most and least revealing crawls was only 242 users out of approximately 169,000 total (a difference of only 0.1%). Keep in mind that Facebook is a dynamic system and the graph topology may be changing during a crawl, and thereby can influence crawl results. We have performed near-time repeated crawls of our data, which uncovered an extremely low amount of variation. Furthermore, the 242 variable users display uniformly low node degrees of 2 or less, indicating that they are outliers to the WCC that were only discovered due to the addition of more seeds to the crawl. This experiment verifies that our methodology effectively reaches all nodes in the large WCC in each network within a negligibly small margin of error. This testing procedure is the same one used in [Mislove 2007] to verify their crawling methodology.

Description of Collected Data. We collected the full user profile of each user visited during our crawls. In addition to this, we also collected full transcripts of Wall posts and photo comments for each user. For the remainder of this paper, we will refer to Wall posts and photo comments collectively as "interactions."

While Facebook profiles do not include a "Date Joined" field, we can estimate this join date by examining each user's earliest Wall post. The Wall is both ubiquitous and the most popular application on Facebook, and a user's first Wall post is generally a welcome message from a Facebook friend. Thus we believe a user's earliest Wall post corresponds closely with their join date. We also collected photo tags and comments associated with each user's photo albums, since this is another prevalent form of Facebook interaction, and gives us insight into users who share physical proximity as well as online friendships.

While the Wall and photo comments are in no way a complete record of user interactions, they are the oldest and most prevalent publicly viewable Facebook applications. Our recent data sets from crawls of user Mini-Feeds show that they are also the two most popular of the built-in suite of Facebook applications by a large margin. Most of the other applications are recent additions to Facebook, and cannot shed light on user interactions from Facebook's earlier history. For example, the Wall was added to Facebook profiles in September 2004, while the Notes application was not introduced until August 2006.

To obtain interaction data on Facebook at a more fine-grained level, we performed crawls of Mini-Feed data from the San Francisco regional network. Unlike Wall posts and photo comments, which are stored indefinitely, the Mini-Feed only reports the last 100 actions taken by each user. Thus, we repeated our crawl of San Francisco daily in the month of October to ensure that we build up a complete record of each user's actions on a day-to-day basis. Given time and manpower constraints, performing daily crawls of all our sampled networks for Mini-Feed data was not feasible, so we focused solely on the relatively small San Francisco network ($\sim$400K users).

4. Analysis of Social Graphs

In this section, we present high level measurement and analysis results on our Facebook data set. First, we analyze general properties of our Facebook population, including user connectivity in the social graph and growth characteristics over time. We use these results to compare the Facebook user population to that of other known social networks, as well as accepted models such as small-world and scale-free networks. Second, we take a closer look at the different types of user interactions on Facebook, including how interactions vary across time, applications, and different segments of the user population. Finally, we present an analysis of detailed user activities through crawls of user Mini-Feed from

Network	Users Crawled (%)	Links (%)	Rad.	Diam.	PathLen.	C. Coef.	Assort.
London, UK	1,241K (50.8)	30,725K (26.5)	11	15	5.09	0.170	0.23
Australia	1,215K (61.3)	121,271K (71.4)	10	14	5.13	0.175	0.17
Turkey	1,030K (55.5)	42,799K (56.7)	13	17	5.10	0.133	0.06
France	728K (59.3)	11,219K (34.6)	10	13	5.21	0.172	0.11
Toronto, ON	483K (41.9)	11,812K (21.9)	10	13	4.53	0.158	0.21
Sweden	575K (68.3)	17,287K (44.8)	8	11	4.55	0.157	0.18
New York, NY	378K (45.0)	7,225K (15.7)	11	14	4.80	0.146	0.18
Colombia	565K (71.7)	10,242K (31.7)	9	12	4.94	0.136	0.08
Manchester, UK	395K (55.5)	11,120K (35.2)	11	15	4.79	0.195	0.21
Vancouver, BC	314K (45.1)	35,518K (59.3)	9	14	4.71	0.170	0.23
Total/Average [Std. Dev.]:	10,697K (56.3)	408,265K (43.3)	9.8 [1.34]	13.4 [1.84]	4.8 [0.41]	0.164	0.17 [0.07]
Orkut [Mislove 2007]	1,846K (26.9)	22,613K	6	9	4.25	0.171	0.072

Table 1. High level statistics and social graph measurements for the ten largest regional networks in our Facebook data set.

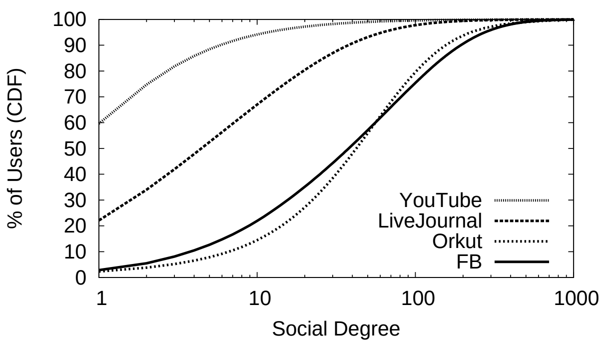

Figure 1. Comparing social degree in Facebook to those of Orkut, YouTube and LiveJournal.

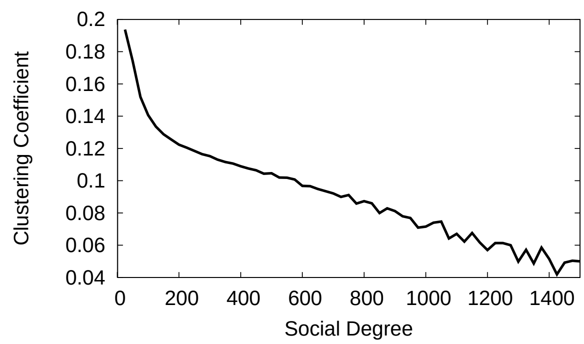

Figure 2. Clustering coefficient of Facebook users as a function of social degree.

the San Francisco network, paying special attention to social network growth and interactions over fine-grained time scales.

4.1 Social Network Analysis

Through our measurements, we were able to crawl roughly 10 million users from the 22 largest regional networks on Facebook, which represents 56% of the total user population of those networks. The remaining 44% of users could not be crawled due to aforementioned issues, such as restrictive privacy policies or disconnection from the WCC of the network. Our complete data set includes just over 940 million social links and 24 million interaction events. Table 1 lists statistics on the ten most populous networks that we crawled, as well as the totals for our entire data set.

Social Degree Analysis. In Figure 1, we compare the social degree (*i.e.* number of friends) of Facebook users against prior results obtained for three other social networks: Orkut, YouTube and LiveJournal [Mislove 2007]. Connectivity among Facebook users most closely resembles those of users in Orkut, likely because both are sites primarily focused on social networking. In contrast, YouTube and LiveJournal are content distribution sites with social components, and exhibit much lower social connectivity. Facebook users are more connected than Orkut users: 37% of Facebook users have more than 100 friends, compared to 20% for Orkut.

As expected of a social network, social degrees on Facebook scale based on a power-law distribution [Barabasi 1999]. Using the method described in [Clauset 2007], we compute that the power-law curve fitting the social degree CDF presented in Figure 1 has an alpha value of 1.5, with fitting error of 0.554. This is identical to the alpha value derived for the Orkut data in [Mislove 2007], although their fitting error was slightly higher at 0.6.

Social Graph Analysis. To evaluate specific graph properties that have an important bearing on social network analysis, we construct a social graph for each crawled regional network. Some of the social links in our data set were not followed, because they point to users that are either not members of the specified regional network, or have modified their default privacy settings. Since we do not have complete social linkage information on these users, we limit our social graphs to only include links for which users at both endpoints were fully visible during our crawls. This prevents incomplete information on some users from biasing our results. As shown in Table 1, 43% of all social links observed

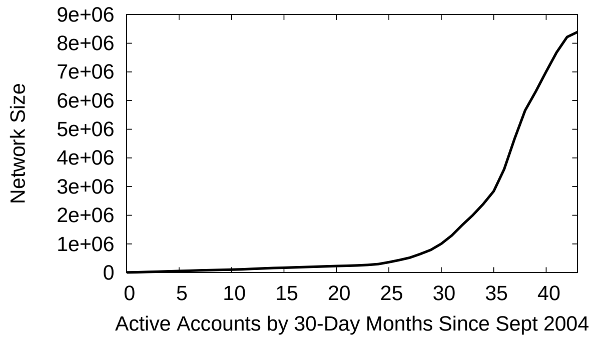

Figure 3. The growth of users in our sample set, starting in September 2004.

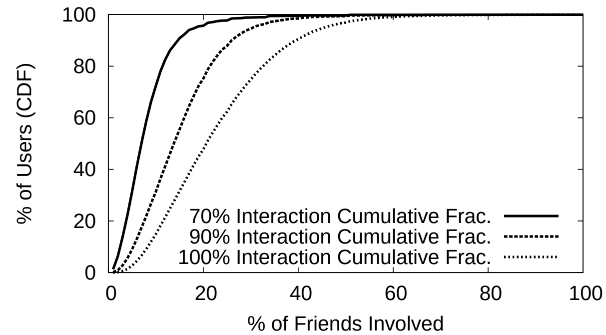

Figure 4. The distribution of users' interaction among their friends, for different % of users' interactions.

during our crawl remained in our social graphs after applying this limiting operation.

For each regional social graph, we display the radius, diameter, and average path length in Table 1. Radius and diameter are calculated using the eccentricity of each node in the social graph. Eccentricity is defined as the maximum distance between a node and any other node in the graph. Radius is defined as the minimum of all eccentricities, while diameter is the maximum. Average path length is simply the average of all-pairs-shortest-paths on the social graph. Note that given the size of our social graphs, calculating all-pairs-shortest-paths is computationally infeasible. Our radius, diameter, and average path lengths are estimates based on determining the eccentricity of 1000 random nodes in each graph. The radius should be viewed as an upper bound and the diameter as a lower bound.

The average path length is 6 or lower for all 22 regional networks, lending credence to the six-degrees of separation hypothesis for social networks [Milgram 1967]. The radius and diameter of each graph is low when compared to other large network graphs, such as the World Wide Web [Broder 2000], but similar to the values presented for other social networks [Mislove 2007].

Clustering Coefficient Measurements. Clustering coefficient is a measure to determine whether social graphs conform to the small-world principle [Watts 1998]. It is defined on an undirected graph as the ratio of the number of links that exist between a node's immediate neighborhood and the maximum number of links that could exist. For a node with N neighbors and E edges between those neighbors, the clustering coefficient is $(2E)/(N(N-1))$. Intuitively, a high clustering coefficient means that nodes tend to form tightly connected, localized cliques with their immediate neighbors.

Table 1 shows that Facebook social graphs have average clustering coefficients (column label C. Coef) between 0.133 and 0.211, with the average over all 22 regional networks being 0.167. This compares favorably with the average clustering coefficient of 0.171 for Orkut. Graphs with average clustering coefficients in this range exhibit higher levels of local clustering than either random graphs or random power-law graphs, which indicates a tightly clustered fringe that is characteristic of social networks [Mislove 2007].

Figure 2 shows how average clustering coefficient varies with social degree on Facebook. Users with lower social degrees have high clustering coefficients, again providing evidence for high levels of clustering at the edge of the social graph. This fact, combined with the relatively low average path lengths and network diameters in our data, is a strong indication that Facebook is a small-world network [Watts 1998].

Assortativity Measurements. The assortativity coefficient, r, of a graph measures the probability for nodes in a graph to link to other nodes of similar degree. It is calculated as the Pearson correlation coefficient of the degrees of node pairs for all edges in a graph, and returns results in the range $-1 \leq r \leq 1$. Assortativity greater than zero indicates that nodes tend to connect with other nodes of similar degree, while assortativity less than zero indicates that nodes connect to others with dissimilar degrees. The assortativity coefficients for our Facebook graphs, shown in Table 1, are uniformly positive, implying that connections between high degree nodes in our graphs are numerous. This well-connected core of high degree nodes form the backbone of small-world networks, enabling the highly clustered nodes at the edge of the network (see Figure 2) to achieve low average path lengths to all other nodes. Our assortativity coefficient values closely resemble the those for other large social networks [Mislove 2007, Newman 2003].

Growth of Facebook over Time. Since users typically receive a Wall message shortly after joining Facebook, we use the earliest Wall post from each profile as a conservative estimate of each profile's creation date. From this data, we plot the historical growth of the user population in our sample set. The results plotted in Figure 3 confirm prior measurements of Facebook growth [Sweney 2008]. Note that Facebook opened its services to the general public in September 2006 (month 24), which explains the observed subsequent exponential growth in network size. We can also

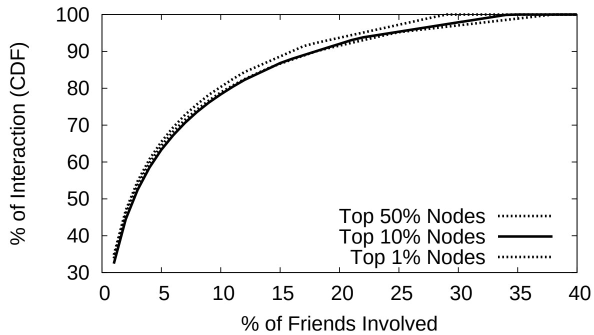

Figure 5. Normalized Wall post distribution of the users with top total Wall interaction.

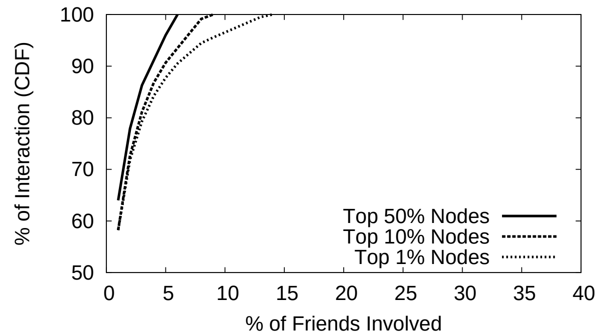

Figure 6. Normalized photo comments distribution of the users with top total photo interaction.

derive from this graph the distribution of Facebook users' "profile age," the time they have been on Facebook. We see that an overwhelming majority ($>$80%) of profiles are "young profiles" that joined Facebook after it went public in 2006.

4.2 User Interaction Analysis

The goal of our analysis of Facebook user interactions is to understand how many social links are actually indicative of active interactions between the connected users. Delving into this issue raises several specific questions that we will address here. First, is the level of interactions even across the user population, or is it heavily skewed towards a few highly-active users? Second, is the distribution of a user's interactions across its friends affected by how active the user is? And finally, how does the interaction of users change over their lifetime, and do interactions exhibit any periodic patterns over time? We punctuate our analysis of user interactions on Facebook by looking at short-timescale, fine-grained measurements from our Mini-Feed data collected from the San Francisco regional network.

Interaction Distribution Among Friends. We first examine the difference in size between interaction graphs and social graphs for users in our data set. We compute for each user a distribution of the user's interaction events across the user's social links. We then select several points from each distribution (70%, 90%, 100%) and aggregate across all users the percentage of friends these events involved. The result is a cumulative fraction function plotted in Figure 4. This is essentially a CDF showing corresponding points from each user's CDF. We see that for the vast majority of users ($\sim$ 90%), 20% of their friends account for 70% of all interactions. The 100% fraction line shows that nearly all users can attribute all of their interactions to only 60% of their friends. This proves that for most users, the large majority of interactions occur only across a small subset of their social links. This confirms our original hypothesis, that only a subset of social links actually represent interactive relationships.

We also want to understand if user interaction patterns are dependent on specific applications, and how interaction patterns vary between power users and less active users. Figures 5 and 6 organize users into user groups of Top 50%, Top 10% and Top 1% by their total level of activity, and show the distribution of incoming Wall posts and photo comments among friends for users within each group. The distribution of Wall posts in Figure 5 shows that the same distribution holds across all Wall users regardless of their overall activity level. In contrast, distribution of photo comments in Figure 6 varies significantly. The most active users only receive photo comments from a small segment ($<$15%) of their friends, while the majority of users receive comments from a third as many ($\sim$5%) of their friends.

The low percentage of friends that comment on photos is notable because photo comments generally occur when friends are tagged in the same picture, implying a level of physical proximity in addition to social closeness. In our data set, 57% of users self-identify with the photo albums they upload by tagging themselves in one or more photos. This fact lends credence to our argument that photo tags accurately capture real life social situations. The photo comment results indicate that users, even highly social ones, show significant skew towards interacting with, and sharing physical proximity with a small subset of their friends.

Distribution of Total Interactions. Next, we wanted to look at how interaction activity was spread out across different kinds of Facebook users. We plot Figure 7 to further understand the contribution of highly interactive users to the overall interaction in the network. For both Wall posts and photo comments, we plot the contribution of different users sorted by each user's interaction in that application. We see that the top 1% of the most active Wall post users account for 20% of all Wall posts and the top 1% of photo comment users account for nearly 40% of all photo comments. Clearly, the bulk of all Facebook interactive events are generated by a small, highly active subset of users, while a majority of users are significantly less active. This result lends credence to our assertion that not all social links are equally useful

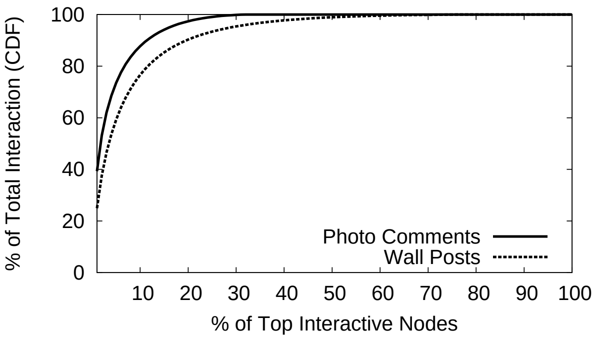

Figure 7. The contribution of different users to total interactions in Facebook.

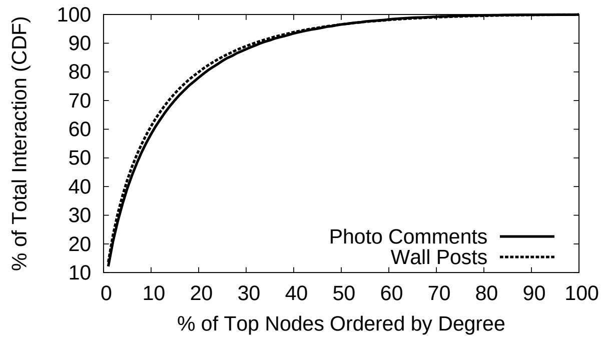

Figure 8. Plot of top % of users ordered by social degree and the interaction contributed by them.

when analyzing social networks, since only a small fraction of users are actively engaged with the network. This also identifies a core set of "power users" of Facebook, who could be identified to leverage their active opinions, ad-clicks, and web usage patterns.

Our next step is to quantify the correlation between users with high social degree and user activity. Figure 8 shows that there is a strong correlation between the two: half of all interactions are generated by the 10% most well-connected users. Nearly all interactions can be attributed to only the top 50% of users. This result confirms that a correlation between social degree and interactivity does exist, which is an important first step to validating our formulation of interaction graphs in Section 5.

Interaction Distribution Across User Lifetime. There is recent speculation that the popularity of social networks is in decline [Sweney 2008, Worthen 2008], perhaps due to the initial novelty of these sites wearing off. This potentially impacts our proposed use of interaction data to augment social graphs: if user activity wanes, then its relevance for assessing social link quality may drop as the information becomes less timely and relevant. Using our records of user interactions over time, we study the gradual growth or decline in interaction events after users join Facebook.

Figure 9 shows users' average number of interactions at different points in their lifetime. We divide the users in the 22 regional networks into 2 groups: the 10% oldest and the 10% newest users. Both user groups show very high average interaction rates in their first days in Facebook, supporting the hypothesis that users are most active when they first join. For the 10% oldest users (average lifetime of 20 months), we see a net increase in interaction rates over time, which we attribute to the "network effect" caused by more friends joining the network over time (see Figure 3). Newer users (average lifetime of 3 weeks) show a different trend, where interactions drop to nearly nothing as the initial novelty of the site wears off. There are two possible interpretations of this. One view is that the oldest users were the original users who participated in Facebook's growth, and therefore are self-selected to users highly interested in social networks (and Facebook in particular). An alternative interpretation is that many of those users who lose interest in Facebook over time closed their accounts, leaving only active Facebook users from that time period.

4.3 Mini-feed Analysis

Two perspectives are missing from our Wall and photo user interaction data. First, these application events do not tell us about the formation of new friend links, one of the dominant activities for Facebook users. In addition, our data set does not describe user interactions in other applications outside of Wall and photos. To rectify this, we perform crawls of user "Mini-Feeds," a continually refreshed list of all[1] user events, including "friend add" events and activity in other applications.

Figure 10 shows the percentage of user Mini-Feed actions each day broken down by category. The most numerous event type is the formation of new social links (adding friends), which accounts for $\sim$45% of daily events. Comment activity, which encompasses both Wall posts and photo comments, only accounts for $\sim$10% of daily activity. Application platform events, which includes events generated from all other applications, only accounts for slightly more than 10%. Clearly, the majority of Facebook events are formation of new friend links, which seems to indicate that the social graph is growing at a faster rate than users are able to communicate with one another. This lends further credence to our argument that average users do not interact with most of the their "Facebook friends."

5. Analysis of Interaction Graphs

Using data from our Facebook crawls, we show in Section 4 that not all social links represent active social relationships. The distribution of each user's interactions is skewed heavily towards a fraction of his or her friends. In addition, interactions across the entirety of Facebook are themselves concen-

[1] Events can be manually deleted by the owner, or suppressed through explicit changes to privacy settings.

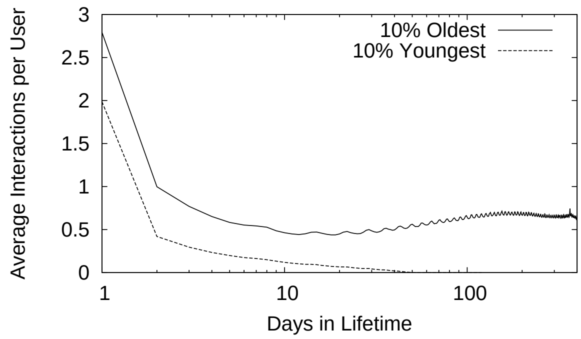

Figure 9. Average number of interactions per day for old and new Facebook users.

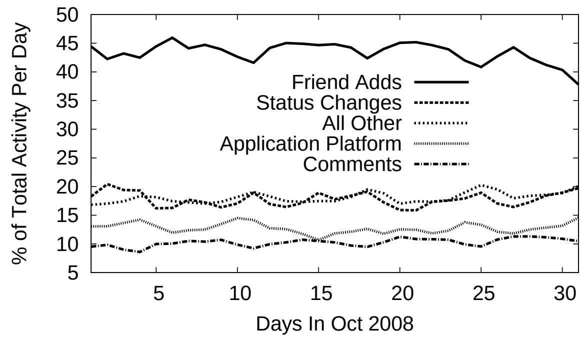

Figure 10. Distribution of user actions in October from the Mini-Feed.

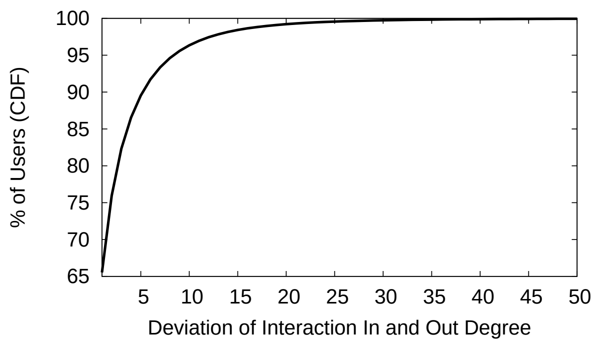

Figure 11. Deviations in pairwise interaction patterns on Facebook.

trated within a subset of Facebook users. These results imply that social links, and the social graphs they form, are not accurate indicators of social relationships between users. This has profound implications on the emerging class of applications that leverage social graphs.

We propose a new model that more accurately represents social relationships between users by taking into account real user interactions. We call this new model an *interaction graph*. We begin this section by formally defining interaction graphs. Next, we implement them on our Facebook data set and explore how the time variant nature of user interactions affects the composition of interaction graphs. Finally, we analyze the salient properties of interaction graphs and compare them to those of the Facebook social graph.

5.1 Definition of Interaction Graphs

To better differentiate between users' active friends and those they merely associate with by name, we introduce the concept of an *Interaction Graph*. An interaction graph is parameterized by an two constants n and t, where n defines a minimum number of interaction events, and t stipulates a window of time during which interactions must have occurred. Taken together, n and t delineate an interaction rate threshold. This leads us to define an interaction graph as the subset of the social graph where for each link, interactivity between the link's endpoints is greater than the rate stipulated by n and t. A user's *Interaction Degree* is the number of friends who interact with the user at a rate greater than the parameterized minimum.

Since a single interaction can be viewed as unidirectional, interaction graphs can contain both directed and undirected edges. It is reasonable to represent interactions in an undirected graph, however, if it can be shown that, for a given data set, per-user interaction in- and out-degrees are similar in value. We discuss this issue in greater detail as it applies to our Facebook data in Section 5.2.

Our formulation of interaction graphs use an unweighted graph. It is feasible, however, to reparameterize the interaction graph such that the interaction threshold no longer functions as a culling value, but instead imparts a weight to each edge in the interaction graph. We do not attempt to derive a weight scheme for interaction graphs analyzed in this paper, but leave exploration of this facet of interaction graphs to future work.

An implicit assumption underlying our formulation of interaction graphs is that the majority of user interaction events occur across social links. Facebook only allows social friends to post Wall and photo comments, thus this assumption holds true for our data set. However, it is conceivable to envision other social networks that do not share these restrictions. In this case it might be beneficial not to define interaction graphs as a subset of the social graph, but instead a wholly new graph based solely on interaction data.

5.2 Interaction Graphs on Facebook

To reasonably model directed Facebook interaction events as an undirected interaction graph, we must first demonstrate that pairwise sets of social friends perform reciprocal interactions with each other. Intuitively, this means that if x writes on y's Wall, y will respond in kind, thus satisfying our conditions for an undirected link. Evaluating each user's incoming and outgoing interactions is challenging, because Facebook data only records incoming events for a specific user, *i.e.* the event x writes on y's Wall is only recorded on y's Wall, not x. Since we are limited to users within specific regional net-

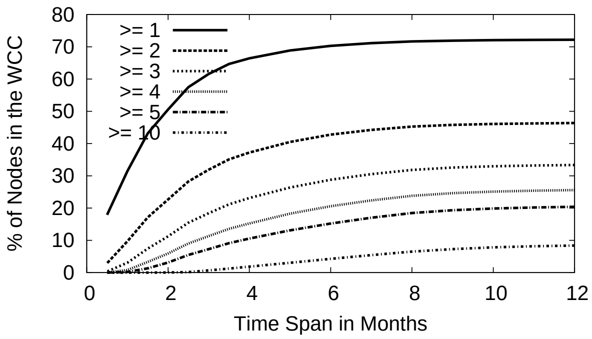

Figure 12. Percentage of nodes remaining in interaction graphs WCC as n and t vary.

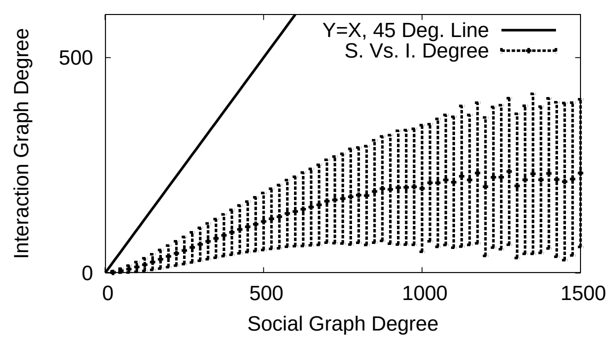

Figure 13. Comparison of the Facebook Social Graph degree and Interaction Graph degree.

works who have not modified their default privacy settings, we do not have access to 100% of the user population. This means we cannot match up all directed interaction events across users. A simple alternative is to examine only users whose friends are also completely contained in our user population. Unfortunately, the high degree of social connectivity in Facebook meant this applied to only about 400K users (4%) in our dataset.

A more reasonable way to study interaction reciprocation on Facebook is to only sample interactions that occur over social links that connect two users in our user population, *i.e.* ignore interactions with users outside our data set. Rather than filtering on users as in the previous approach, this performs filtering on individual social links. Assuming that user interactions do not change significantly due to user privacy settings and geolocation, these sampled results should be representative.

After this sampling, Figure 11 shows the length of the set resulting from the symmetric set difference of each user's incoming and outgoing interaction partners plotted as a CDF. We refer to this metric as *deviation*. Intuitively, the deviation for each user counts the number of directed interactions that were not reciprocated with a direct reply, thus forming a solely directed interaction link. For 65% of the users, all interactions are reciprocated, meaning that all of these interactions can be modeled as undirected links. Based on these results, we believe it is acceptable to model interaction graphs on Facebook using undirected edges, since this model suits the interactivity patterns of the majority of users.

We now discuss the interaction rate parameters n and t. The simplest formulation of these parameters is to consider all interactions over the entire lifetime of Facebook ($t = 2004$ to the present, $n \geq 1$). We will refer to the interaction graph corresponding to this parameterization as the *full interaction graph*. We also consider additional interaction graphs that restrict t and increase n beyond 1. This allows time and rate thresholds to be applied to generate interaction graphs appropriate for specific applications that have heterogeneous definitions of interactivity.

Figure 12 shows the size of the weakly-connected components for interaction graphs as t and n change. This figure is based on data for the year 2007, i.e. 2 months refers to interactions occurring between November 1 and December 31, 2007. As expected, larger t and lower n are less restrictive on links, therefore allowing for more nodes to remain connected. Based on Figure 12, we chose several key interaction graphs for further study, including those with $n \geq 1$ at the 1 year, 6 months, and 2 months time periods. These three graphs each contain WCCs that contain a majority of all nodes, and are amenable to graph analysis. For the remainder of this paper, we will only consider interaction graphs for which $n \geq 1$.

5.3 Comparison of Social and Interaction Graphs

We now take a closer look at interaction graphs and compare them to full social graphs. We look at graph connectivity and examine properties for power-law networks, small-world clustering, and scale-free networks.

Social vs. Interaction Degree. Figure 13 displays the correlation between social degree and interaction degree for the full interaction graph. The error bars indicate the standard deviation for each plotted point. Even with this "least-restricted" interaction graph, it is clear that interaction degree does not scale equally with social degree. If all Facebook users interacted with each of their friends at least once then this plot would follow a 45 degree line. This is not the case, confirming once again the disparity between friend relationships and active, social relationships.

Interaction Degree Analysis. Figure 15 plots the degree CDFs of the four interaction graphs and the Facebook social graph. The interaction graphs exhibit a larger percentage of users with zero friends, and reach 100% degree coverage more rapidly than the social graph. This is explained by the uneven distribution of interactions between users' friends. Referring back to Figure 4, we showed that interactions are skewed towards a fraction of each user's friends. This means many links are removed from the social graph during conversion into an interaction graph. This means many weakly

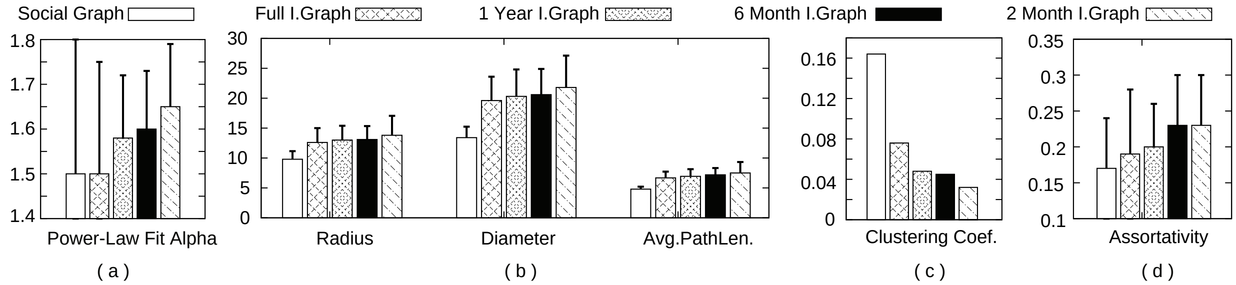

Figure 14. Graph measurements for four interaction graphs compared to the entire Facebook social network.

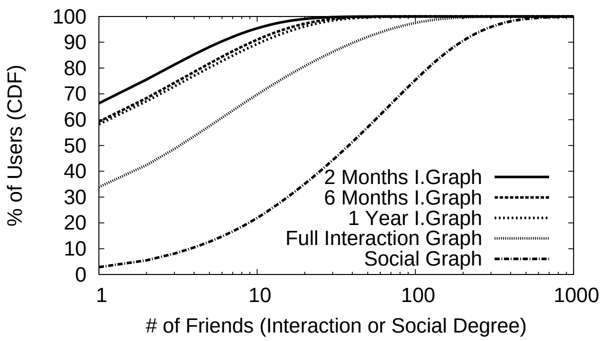

Figure 15. CDF of node degrees for the interaction graphs.

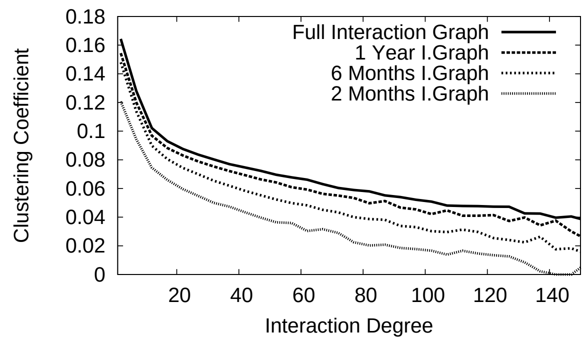

Figure 16. Clustering coefficient of interaction graphs as a function of interaction degree.

connected users in the social graph have zero interaction degree, while highly connected users in the social graph are significantly less connected in the interaction graph.

Despite these differences, the interaction graphs still exhibit power-law scaling. Figure 14 (a) shows the alpha values for the four interaction graphs compared to the social network. The error bars above the histogram are the fitting error of the estimator [Clauset 2007]. The fitting error for the interaction graphs are lower than that for the social graph, indicating that the interaction graphs exhibit more precise power-law scaling. As the link structure of the interaction graphs gets restricted, alpha rises, corresponding to an increased slope in the fitting line. This property is visualized in Figure 15 as a lower number of high degree nodes in the most constrained interaction graphs. These results are further validated by studies on LiveJournal that have uncovered degree distribution and power-law scaling characteristics very similar to those depicted here for Facebook interaction graphs [Mislove 2007].

Interaction Graph Analysis. Figure 14 (b) shows the average radius, diameter, and path lengths for all of the interaction graphs, as well as for the social network. These measures all display the same upward trend as the interaction graphs become more restricted. This makes intuitive sense: as the average number of links per node and the number of high-degree "super-nodes" decreases (see Figure 15) the overall level of connectivity in the graph drops. This causes average path lengths to rise, affecting all three of the measures presented in Figure 14 (b).

Clustering Coefficient Measurements. Besides average path length, another metric intrinsically linked to node connectivity is the clustering coefficient. Figure 14 (c) shows that average clustering coefficient drops as interaction graphs become more restricted. This is another ramification of link removal, as fewer links leads to less clustering between nodes. Figure 16 depicts average clustering coefficients as a function of interaction degree. As with the Facebook social graph, there is more clustering among nodes with lower degrees. However, the overall amount of clustering is reduced by over 50% across all interaction graphs.

Taken together, the reduced clustering coefficients and the higher path lengths that characterize Facebook interaction graphs indicates that they exhibit significantly less small-world clustering. In order for the interaction graphs to cease being small-world, the average clustering coefficient would have to approach levels exhibited by a random graph with an equal number of nodes and edges. This number can be estimated by calculating K/N, where K is average node degree and N is the total number of nodes [Watts 1998]. For the Facebook social graph, $K = 76.54$. We can estimate from this that an equivalent random graph would have an average clustering coefficient of $7.15 * 10^{-6}$. K is smaller for our interaction graphs, therefore the estimated clustering coefficient for equivalent random graphs will be smaller as well. These estimated figures are orders of magnitude smaller than the actual clustering coefficients observed in our social and interaction graphs, thus confirming that they both remain small-world.

The conclusion that Facebook interaction graphs exhibit less small-world behavior than the Facebook social graph has important implications for all social applications that rely on this property of social networks in order to function, as we will show in Section 6.

Assortativity Measurements. Figure 14 (d) shows the relative assortativity coefficients for all social and interaction graphs. Assortativity measures the likelihood of nodes to link to other nodes of similar degree. Since interaction graphs restrict the number of links high degree nodes have, this causes the degree distribution of interaction graphs to become more homogeneous. This is reflected by the assortativity coefficient, which rises commensurately as the interaction graphs grow more restricted.

6. Applying Interaction Graphs

When social graphs are used to drive simulations of socially-enhanced applications, changes in user connectivity patterns can produce significantly different results for the evaluated application. Given the lack of publicly available social network topological datasets, many current proposals either use statistical models of social networks based on prior measurement studies [Yu 2006, Watts 1998, Marti 2004], or bootstrap social networks using traces of emails [Garriss 2006].

The hypothesis of our work is that validation of socially-enhanced applications require a model that takes interactions between users into account. To validate how much impact the choice of user model can make on socially enhanced applications, we implement simulations of two well-known socially-enhanced distributed systems [Yu 2006, Garriss 2006], and compare the effectiveness of each system on real social graphs, and real interaction graphs derived from our Facebook measurements.

6.1 RE: Reliable Email

"RE" [Garriss 2006] is a white-listing system for email based on social links that allows emails between friends and Friends-of-Friends (FoFs) to bypass standard spam filters. Socially-connected users provide secure attestations for each others' email messages while keeping users' contacts private. RE works automatically based on social connectivity data: no per sender or per email classification is requested from users.

Expected Impact The presence of small-world clustering and scale-free behavior in social graphs translate directly into short average path lengths between nodes. For RE, this means that the set of friends and FoFs that will be white-listed for any given user is very large. In this situation, a single user who sends out spam email is likely to be able to successfully target a very large group of recipients via the social network. Keep in mind that a spammer in this context could be an openly malicious, rogue user, or a legitimate user whose account has been compromised. In contrast, RE that leverages interaction graphs should not experience as high a proliferation of spam, given an equal number of spammers. The reduced presence of small-world clustering in interaction graphs, coupled with lower average node degrees, causes average path lengths to grow as compared to social networks (see Figure 14 (b)). This should have a damping effect on the size of friend and FoF populations, and consequently limit spam penetration.

Results. We present experimental evaluation of RE here. For social graph and interaction graphs, we randomly choose a percentage of nodes to act as spammers. In the RE system, all friends and FoFs of the spammer will automatically receive the spam due to white-listing. All experiments were repeated ten times and the results averaged.

This experiment leads to Figure 17, which plots the percentage of users in each graph receiving spam versus the percentage of users who are spamming. On the social network spam penetration quickly reaches 90% of users, covering the majority of users in the WCC. In contrast spam penetration is reduced by 40% over the social graph when the number of spammers is low, and 20% when the number of spammers is high when RE is run on the interaction graphs.

6.2 SybilGuard

A Sybil attack [Douceur 2002] occurs when a single attacker creates a large number of online identities, which when colluding together, allows the attacker to gain significant advantage in a distributed system. Sybil identities can work together to distort reputation values, out-vote legitimate nodes in consensus systems, or corrupt data in distributed storage systems.

SybilGuard [Yu 2006; 2008][2] proposes using social network structure to detect Sybil identities in an online community to protect distributed applications. It relies on the fact that it is difficult to make multiple social connections between Sybil identities and legitimate users. The result is that Sybil identities form a well-connected subgraph that has only a limited number of connection edges (called *attack edges*) to the legitimate network.

Each node in the social network creates a persistent routing table that maps each incoming edge to an outgoing edge in an unique one-to-one mapping. To determine whether to accept a "suspect" node *s* as a real user, a "verifier" node *v* creates a "random route" of *w* hops, where a random route is a deterministic route formed by following the stored routing table entries at *w* consecutive nodes. A similar *w* hop random route is initiated at *s*, and *v* accepts *s* if the two random routes intersect. Note that as *w* increases, the number of Sybils is the network allowed under the SybilGuard protocol also increase. Thus, it is beneficial for *w* to be small.

[2] Although SybilLimit is an advanced proposal, SybilGuard is a simple version that we believe is sufficient for our purpose.

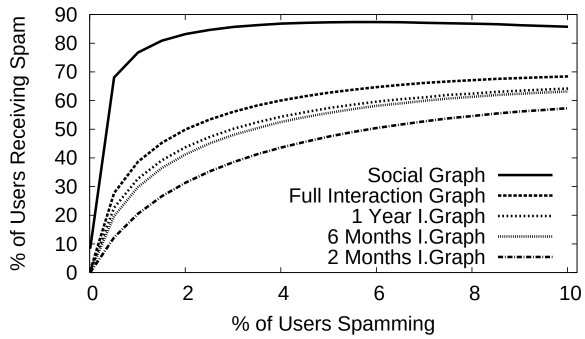

Figure 17. Spam penetration as the number of spammers is varied for the Reliable Email [Garriss 2006] system.

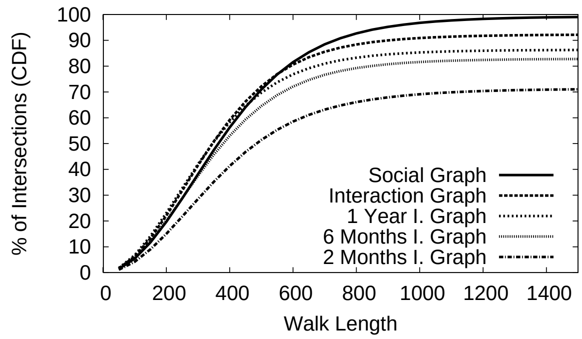

Figure 18. Percentage of path intersections for SybilGuard [Yu 2006] as random walk length increases.

Expected Impact The success of SybilGuard relies on the premise that Sybil identities cannot easily establish trusted social relationships with legitimate users, and hence have few "attack edges" in the social network. In particular, SybilGuard requires connected users to exchange encryption keys. We believe that typical social connections in social graphs do not represent this level of trust. Given our results that demonstrate most Facebook friends pairs do not even interact, it seems unreasonable to assume that most friend pairs have the requisit level of trust to exchange secure keys. Instead, we expect that our interaction graph is a closer approximation to the representation of trusted links that SybilGuard would observe in reality.

Results For our experiments, we implement the SybilGuard algorithm on both our social graph and interaction graphs and measure the percentage of paths that successfully intersect as w increases. For each graph and each value of w we chose 25000 random pairs of nodes to perform intersection tests on.

The reduction of highly connected super nodes in the interaction graph means that random walks (and random routes) are less likely to connect. Figure 18 shows that for the Facebook social graph, the probability for all paths to intersect approaches 100% at $w = 1200$. For interaction graphs, the percentage of intersecting paths never reaches 100% since a large fraction of random walks never intersect. SybilGuard, as a result, is less effective on a graph that models user trust (interaction graph) than on a normal social graph.

Graph	Total Loops (%)
Social	951 (3.8)
Full Interaction	3196 (12.8)
1 Year I.Graph	4726 (18.9)
6 Month I.Graph	4953 (19.8)
2 Month I.Graph	5782 (23.1)

Table 2. Self-Looping Statistics for SybilGuard

A major factor affecting the performance of the SybilGuard algorithm is the prevelence of self-loops in the random walks. Any walk that returns to the origin point before going w steps is useless for the purposes of performing intersection tests. Table 2 shows the total number of self-loops encountered during all experimental runs on each graph. The drop in efficacy observed in Figure 18 is directly correlated to the increase in self-looping from 3.8% on the social graph to an upwards of 20% on interactions graphs.

7. Related Work

The body of research geared towards real-world social webs and physical networks has only recently begun to be leveraged to understand online social networks. One of the original papers to study the emerging social network phenomena focused on the Club Nexus website of Stanford University [Adamic 2003]. More recently traces from CyWorld, MySpace and Orkut have been profiled [Ahn 2007], as have YouTube, Flickr, LiveJournal, and (again) Orkut [Mislove 2007]. Yet another study focused on profiling social network evolution on Flickr and Yahoo! 360 [Kumar 2006]. Finally, a recent measurement study analyzed the growth of Flickr social network using a three month crawl data [Mislove 2008]. These studies confirm that online social networks obey power-law scaling characteristics [Barabasi 1999] and exhibit high clustering coefficients, firmly establishing them as small-world networks [Amaral 2000].

Recent studies analyzed the online communication patterns among the users in a large IM trace [Leskovec 2008], and in an online social network [Chun 2008]. The IM study [Leskovec 2008] also reported a relatively higher value of average path length for the graph formed from user interactions. However, the IM interaction graph is more resilient to node removal than the interaction graphs in Facebook, as indicated by our assortativity values. Like our study, the CyWorld interaction study [Chun 2008] showed that CyWorld user interactions are bi-directional. User interaction behavior differs significantly from our study, however. CyWorld users with less than 200 friends interact only with a small subset of friends and users with more than 200 friends interact evenly. In addition, both activity and social graphs are similar in CyWorld and exhibit multi-scaling behavior.

This multi-scaling is unique to CyWorld; all other social networks analyzed so far, including Facebook, exhibit simple power-law connectivity scaling [Mislove 2007, Ahn 2007, Leskovec 2008].

8. Conclusion

This paper aims to answer the question: *Are social links valid indicators of real user interaction?* To do this, we gathered extensive data from crawls of the Facebook social network, including social and interaction statistics on more than 10 million users. We show that interaction activity on Facebook is significantly skewed towards a small portion of each user's social links. This finding casts doubt on the assumption that all social links imply equally meaningful friend relationships.

We introduce the *interaction graph* as a more accurate representation of meaningful peer connectivity on social networks. Analysis of interaction graphs derived from our Facebook data reveal different characteristics than the corresponding social graph. Most notably, interaction graphs exhibit an absence of small-world clustering. We also observe much lower average node degrees in the interaction graph as compared to the Facebook social graph. This confirms the intuition that human interactions are limited by constraints such as time, and brings into question the practice of evaluating social networks in distributed systems directly using social connectivity graphs.

Our study concludes with experiments to evaluate the effects of interaction graphs on two well known social applications. The performance of RE [Garriss 2006] improves with the use of interaction graphs, as the streamlined link structure helps control spam proliferation. In the case of Sybilguard [Yu 2006], the system becomes less able to effectively classify nodes once its assumptions about graph structure are violated. These experiments strongly suggest that social-based applications should be designed with interactions graphs in mind, so that they reflect real user activity rather than social linkage alone.

Acknowledgments

We thank the anonymous reviewers and our shepherd Rodrigo Rodrigues for numerous suggestions to improve the paper. We also thank Alan Mislove and colleagues at MPI-SWS for sharing their data sets on YouTube, LiveJournal and Orkut. This material is based in part upon work supported by the National Science Foundation under grant IIS-847925 and CAREER CNS-0546216. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[Adamic 2003] Lada A. Adamic, Orkut Buyukkokten, and Eytan Adar. A social network caught in the web. *First Monday*, 8(6), 2003.

[Ahn 2007] Yong-Yeol Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *Proc. of the International World Wide Web Conference*, 2007.

[Amaral 2000] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of small-world networks. In *Proc. of National Academy of Sciences*, pages 11149–11152, 2000.

[Barabasi 1999] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286, 1999.

[Boe 2008] Bryce Boe and Christo Wilson. crawl-e: Highly distributed web crawling framework written in python. Google Code, 2008.

[Broder 2000] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web: Experiments and models. In *Proc. of the International World Wide Web Conference*, 2000.

[Chun 2008] H. Chun, H. Kwak, Y.H. Eom, Y.Y. Ahn, S. Moon, and H. Jeong. Comparison of online social relations in volume vs interaction: a case study of cyworld. In *Proc. of Internet Measurement Conference*, 2008.

[Clauset 2007] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *Journal of Computer-Mediated Communication*, June 2007.

[Donath 2004] J. Donath and D. Boyd. Public displays of connection. *BT Technology Journal*, 22(4), 2004.

[Douceur 2002] John R. Douceur. The Sybil attack. In *Proc. of IPTPS*, 2002.

[Facebook 2008] Facebook. Statistics. facebook.com, March 2008.

[Garriss 2006] S. Garriss, M. Kaminsky, M. J. Freedman, B. Karp, D. Mazi'eres, and H. Yu. Re: Reliable email. In *Proc. of NSDI*, San Jose, CA, May 2006.

[Kumar 2006] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *Proc. of KDD*, pages 611–617, 2006.

[Leskovec 2008] Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *Proc. of the International World Wide Web Conference*, Beijing, China, April 2008.

[Marti 2004] Sergio Marti, Prasanna Ganesan, and Hector Garcia-Molina. DHT routing using social links. In *Proc. of IPTPS*, San Diego, CA, February 2004.

[Milgram 1967] Stanley Milgram. The small world problem. *Psychology Today*, 6:62–67, 1967.

[Milgram 1977] Stanley Milgram. *The familiar stranger: an aspect of urban anonymity*. Addison-Wesley, 1977.

[Mislove 2006] A. Mislove, K. P. Gummadi, and P. Druschel. Exploiting social networks for internet search. In *Proc. of HotNets*, Irvine, CA, November 2006.

[Mislove 2008] Alan Mislove, Hema Swetha Koppula, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Growth of the flickr social network. In *Proc. of WOSN*, Seattle, WA, August 2008.

[Mislove 2007] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proc. of Internet Measurement Conference*, October 2007.

[Newman 2003] M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, 67, 2003.

[Sweney 2008] Mark Sweney. Facebook sees first dip in uk users. guardian.co.uk, February 2008.

[Watts 1998] Duncan J. Watts and Steven Strogatz. Collective dynamics of 'small-world' networks. *Nature*, (393):440–442, 1998.

[Worthen 2008] Ben Worthen. Bill Gates quits facebook. Wall Street Journal Online, Feb. 2008.

[Yu 2008] Haifeng Yu, Phillip B. Gibbons, Michael Kaminsky, and Feng Xiao. Sybillimit: A near-optimal social network defense against sybil attacks. In *Proc. of IEEE Security & Privacy*, 2008.

[Yu 2006] Haifeng Yu, Michael Kaminsky, Phillip B. Gibbons, and Abraham Flaxman. Sybilguard: defending against sybil attacks via social networks. In *Proc. of SIGCOMM*, 2006.