Beyond Social Graphs: User Interactions in Online Social Networks and their Implications

CHRISTO WILSON, ALESSANDRA SALA, KRISHNA P. N. PUTTASWAMY,
and BEN Y. ZHAO, University of California Santa Barbara

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 article, 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 present in their social graph counterparts. This means that these graphs have fewer "supernodes" with extremely high degree, and overall graph diameter increases significantly as a result. To quantify the impact of our observations, we use both types of graphs to validate several well-known social-based applications that rely on graph properties to infuse new functionality into Internet applications, including Reliable Email (RE), SybilGuard, and the weighted cascade influence maximization algorithm. The results reveal new insights into each of these systems, and confirm our hypothesis that to obtain realistic and accurate results, ongoing research on social network applications 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; J.4 [**Computer Applications**]: Social and behavioral sciences

General Terms: Measurement, Performance

Additional Key Words and Phrases: Social networks, interaction graphs, Facebook

ACM Reference Format:
Wilson, C., Sala, A., Puttaswamy, K. P. N., and Zhao, B. Y. 2012. Beyond social graphs: User interactions in online social networks and their implications. ACM Trans. Web 6, 4, Article 17 (November 2012), 31 pages.
DOI = 10.1145/2382616.2382620 http://doi.acm.org/10.1145/2382616.2382620

The preliminary version of this article appeared in *Proceedings of the 4th European Conference on Computer Systems (EuroSys09)* [Wilson et al. 2009].

This material is based in part upon work supported by the National Science Foundation under grants IIS-847925, IIS-0916307, 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.

Authors' address: C. Wilson (corresponding author), A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao, Department of Computer Science, University of California at Santa Barbara, Santa Barbara, CA 93106; email: bowlin@cs.ucsb.edu.

DOI 10.1145/2382616.2382620 http://doi.acm.org/10.1145/2382616.2382620

1. INTRODUCTION

Online Social Networks (OSNs) are popular tools for communication, interaction, and information sharing on the Internet. 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 applications, including spam email mitigation [Garriss et al. 2006], Internet search [Mislove et al. 2006], and defense against Sybil attacks [Yu et al. 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 and Boyd 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 article, we address this question through a detailed study of user interaction events in Facebook, the most popular social network in the world with over 800 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 social network, method of interaction, and types of users involved.

We make four key contributions through our study. First, at the original time of publication this article presented the first large-scale study of the Facebook social network [Wilson et al. 2009]. Unlike Orkut, YouTube, or Flickr, Facebook's strong focus on user privacy has generally prevented researchers from "crawling" their social network for user profiles. We present detailed analysis of our dataset 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 social relationships based on interactions between users. 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 graph diameters, lower clustering coefficients, and higher assortativity.

Third, in Section 6 we examine the impact of using different graph models in evaluating socially enhanced applications. We conduct simulated experiments of the Reliable Email [Garriss et al. 2006] and SybilGuard [Yu et al. 2006] systems on both social and interaction graphs derived from our Facebook data. We also evaluate the efficacy of the weighted cascade influence maximization algorithm [Chen et al. 2009] on our two graph models. Our results demonstrate that differences in the two graph models translate into significantly different application performance results.

Lastly, in Section 7 we compare and contrast data from two separate Facebook crawls conducted in 2008 and 2009. Our results show how the Facebook user base has grown since our initial Facebook study was published, and the effect this has on the social graph. We also use interaction data from 2009 to validate our conclusions that were originally derived from 2008 data. We show that although Facebook continues to evolve, the overall trends in user interactions do not vary with time.

2. THE FACEBOOK SOCIAL NETWORK

Before describing our dataset and the results of our analysis, we first provide background information on Facebook's social network. With over 800 million active users (as of winter 2011), 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 undirected 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 (which now exists in an evolved form as the News-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 friends. Other events include new Wall posts, photo uploads and comments, profile updates, and status changes. The Mini-Feed is ordered chronologically, and only displays (at most) the user's 100 most recent actions.

Originally, Facebook was designed around the concept of "networks" that organized users into membership-based groups. Although Facebook ceased being structured this way in the summer of 2009, this was the model during 2008 when our crawls were conducted. Each network represents an educational institution (university or high school), a company or organization (called work networks), or a geographic (regional network) location. Facebook authenticated membership in college and work networks by verifying that users had a valid email address from the associated educational or corporate domain. Users authenticated membership in high school networks through confirmation by an existing member. In contrast, no authentication was required for regional networks. Users could belong to multiple school and work networks, but only one regional network, which they could change twice every sixty days.

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

3. DATASET AND COLLECTION METHODOLOGY

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

3.1. Definitions

In this article, we model each social network as an undirected graph $G = (V, E)$. The set of nodes V corresponds to users on the social network. We use the term "node" and "user" interchangeably in this article. The set of edges E corresponds to *social links* between users on the social graph. On Facebook, users explicitly create undirected social links by "friending" each other. We say that two users have a *relationship* if they are connected by an edge.

Much of the analysis in this article focuses on *user interactions*. We define an interaction as an explicit message $i_{u,v}$ that is generated by one user u and directed at a second user v where $u, v \in V$. The set of interactions I is a multiset, for example, interactions between pairs of users can occur multiple times. $I \subseteq E$, meaning each interaction $i_{u,v} \in I$ can exist if and only if edge $e_{u,v} \in E$.

On social networks like Facebook, interactions correspond to explicit events like writing on a friend's wall, or commenting on a friend's photo. For the remainder of this article, we will refer to Wall posts and photo comments collectively as "interactions." For example, a user who writes on a friend's wall and comments on one of that friend's photos has just interacted with that friend twice. In this article, we only consider explicit interactions, as opposed to *latent interactions*, which refers to profile browsing behavior.

3.2. Data Collection Process

As we mentioned, Facebook used to be divided into networks that represented schools, institutions, and geographic regions. Membership in regional networks was unauthenticated and open to all users. Since the majority of Facebook users belonged to at least one regional network, and most users do not modify their default privacy settings, a large portion of Facebook's user profiles could 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. Networks and their size statistics have since been removed from Facebook.

While other studies of social networks rely on statistical sampling techniques [Mislove et al. 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 dataset 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 a subset of these networks and their key characteristics in Table I. 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, multithreaded crawler using Python with support for Remote Method Invocation (RMI) [Boe and Wilson 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 regional network. Two dual-core Xeon servers were generally able to complete each crawl in under 24 hours, while averaging

Table I. Statistics for the Ten Largest and Two Smallest Regional Networks in Our Dataset

Network	Users Crawled (%)	Links (%)
London, UK	1,241K (50.8)	30,743K (26.5)
Australia	1,215K (61.3)	27,261K (36.0)
Turkey	1,030K (55.5)	17,739K (35.2)
France	728K (59.3)	11,227K (34.6)
Toronto, ON	483K (41.9)	11,829K (21.9)
Sweden	575K (68.3)	17,290K (44.8)
New York, NY	378K (45.0)	7,233K (15.7)
Colombia	565K (71.7)	10,242K (31.7)
Manchester, UK	395K (55.5)	11,124K (35.2)
Vancouver, BC	314K (45.1)	8,240K (25.3)
Egypt	246K (57.8)	3,236K (25.5)
San Francisco	172K	2,911K
Total:	10,697K (56.3)	240,265K (29.4)
Orkut [Mislove et al. 2007]	3,072K (11.3)	223,534K
Flickr [Mislove et al. 2007]	1,846K (26.9)	22,613K

roughly 10MB/s of download traffic. Our completed dataset is approximately 500GB in size, and includes full profiles of more than 10 million Facebook users.

3.3. Dataset Completeness and Limitations

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 et al. 2007]. Since social links on Facebook are undirected, breadth-first crawling of social links should be able to generate complete coverage of the large connected component, assuming that at least one of the initial seeds of the crawl is linked to the connected component. 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 part of the connected component.

To validate our data collection procedure and ensure that our crawls are reaching every available user in the connected component, 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 ∼169K total (a difference of only 0.1%). The 242 variable users display uniformly low node degrees of 2 or less, indicating that they are either new accounts that were added during our crawl, or outliers to the connected component 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 connected component in each regional network within a negligibly small margin of error. This testing procedure is the same one used in Mislove et al. [2007] to verify their crawling methodology.

In this study we are limited by Facebook's privacy settings. Our dataset only includes users with public profiles, and we only collect public interactions. On Facebook public interactions include Wall posts and photo comments, while private interactions are direct messages between users and "pokes." As discussed in Section 4.1, our crawl covers the majority of users in each regional network, and therefore we believe it is representative of the overall Facebook population. Our results from the crawled Facebook graph are extremely similar to those calculated using the entire Facebook graph (see Section 7) [Ugander et al. 2011]. Similarly, other studies of interactions on

Facebook that have had access to private interactions have reached similar conclusions to our study [Backstrom et al. 2011; Golder et al. 2007].

3.4. 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.

Although 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 his/her 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 datasets 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 October 2008 to ensure that we built 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 regional networks for Mini-Feed data was not feasible, so we focused solely on the relatively small San Francisco network (172K users). We use a log integration algorithm similar to the one used by Jiang et al. [2010] to reconstruct each user’s month-long interaction record from our daily crawl results.

4. ANALYSIS OF SOCIAL GRAPHS

In this section, we present high-level measurement and analysis results on our Facebook dataset. 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 graphs. 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 the San Francisco network, paying special attention to social 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 restrictive privacy policies or disconnection from the connected component of the graph. Our complete dataset includes about 818 million social links and 24 million

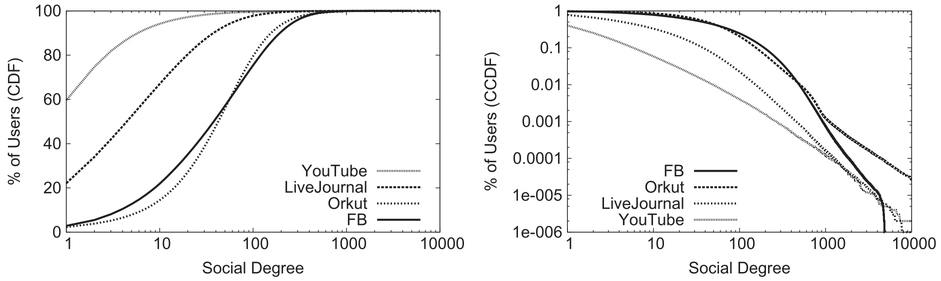

Fig. 1. Comparing the degree distribution of Facebook to Orkut, YouTube, and LiveJournal [Mislove et al. 2007]. Both CDF and CCDF distributions are shown.

interaction events. Table I lists statistics on the ten largest and the two smallest regional networks that we crawled, as well as the totals for our entire dataset.

4.1.1. 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 et al. 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.

Figure 1 shows the Cumulative Distribution Function (CDF) and Complementary Cumulative Distribution Function (CCDF) of social degrees for four social networks. Prior work has shown that many measured large graphs, including social networks, exhibit a power-law degree distribution [Barabasi and Albert 1999]. However, Facebook does not follow a pure power law: as shown in the CCDF of Figure 1, the degree distribution is not a straight line in a log-log plot. Facebook's degree distribution is most similar to Orkut, which is also not pure power law [Mislove et al. 2007]. We use the method from Mislove et al. [2007] (which is a modification of the method from Clauset et al. [2009]) to find the power-law exponent for the body of the Facebook degree distribution. We calculate that Facebook has an alpha value of 1.25, with fitting error of 0.31. This is significantly lower than the alpha value derived for YouTube (alpha = 1.63, fitting error = 0.13), which does demonstrate a very clear power-law degree distribution.

4.1.2. Graph Distances. To evaluate graph distance properties such as radius, diameter, and average path length, we construct a social graph for each crawled regional network. Some of the social links in our dataset were not crawled, 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 I, 29% of all social links observed 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 II. 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

Table II. Social Graph Measurements for the Ten Largest and Two Smallest Regional Networks in our Dataset

Network	Rad.	Diam.	PathLen.	C. Coef.	Assort.
London, UK	11	15	5.09	0.170	0.25
Australia	10	14	5.13	0.175	0.17
Turkey	13	17	5.10	0.133	0.06
France	10	13	5.21	0.172	0.12
Toronto, ON	10	13	4.53	0.158	0.23
Sweden	8	11	4.55	0.158	0.19
New York, NY	11	14	4.80	0.146	0.19
Colombia	9	12	4.94	0.136	0.09
Manchester, UK	11	15	4.79	0.195	0.21
Vancouver, BC	9	14	4.71	0.170	0.23
Egypt	9	12	4.88	0.167	0.01
San Francisco	9	14	4.8	0.194	0.18
Average [Std. Dev.]:	9.8 [1.34]	13.4 [1.84]	4.8 [0.41]	0.164	0.17 [0.07]
Orkut [Mislove et al. 2007]	6	9	4.25	0.171	0.072
Flickr [Mislove et al. 2007]	13	27	5.67	0.313	0.202

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 graphs, such as the World Wide Web [Broder et al. 2000], but similar to the values presented for other social networks [Mislove et al. 2007].

4.1.3. Clustering Coefficient. Clustering coefficient is a measure to determine whether social graphs conform to the small-world principle [Watts and Strogatz 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. The clustering coefficient for an entire graph is the mean of all clustering coefficients for individual nodes.

Table II 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 et al. 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 graph diameters in our data, is a strong indication that Facebook is a small-world graph [Watts and Strogatz 1998].

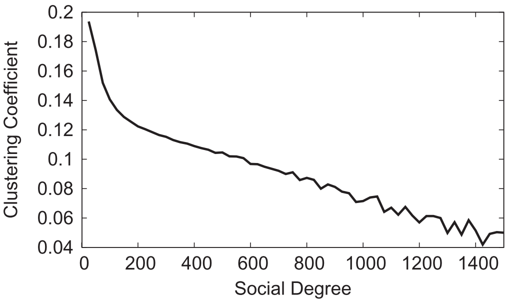

Fig. 2. Clustering coefficient of Facebook users as a function of social degree.

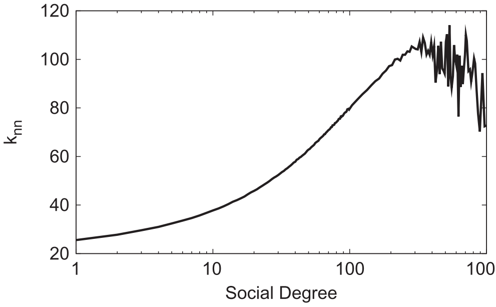

Fig. 3. k_{nn} of Facebook users as a function of social degree.

4.1.4. Joint Degree Distribution and Assortativity. The Joint Degree Distribution (JDD) of a graph describes the likelihood of nodes of different degrees connecting to one another. JDD is approximated on large graphs by the degree correlation function k_{nn}. For undirected graphs k_{nn} is defined as the average degree of all nodes connected to nodes of a given degree. Figure 3 shows how average k_{nn} varies with social degree on Facebook. Low-degree nodes tend to connect to other low-degree nodes, while the reverse is true for high-degree nodes.

Closely related to JDD, 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 II, are uniformly positive, implying that connections between high-degree nodes in our graphs are numerous. Our assortativity coefficient values closely resemble those for other large social networks [Mislove et al. 2007; Newman 2003].

Our k_{nn} and assortativity results both indicate the presence of a well-connected "core" of high-degree nodes in our Facebook graphs. These nodes form the backbone of small-world graphs, enabling the highly clustered nodes at the edge of the graph (see Figure 2) to achieve low average path lengths to all other nodes.

4.1.5. Network Core Analysis. Previous studies of large power-law graphs have shown that the densely connected core of high-degree nodes are necessary to hold the graph together [Mislove et al. 2007]. When these nodes are removed the graph fractures, that is, the nodes no longer form a single, large, connected component [Broder et al. 2000].

We analyze the core of our Facebook regional networks by ordering all nodes by degree, iteratively removing the highest-degree nodes, and measuring the size of the resulting connected component. The percentage of nodes remaining in the connected component at a given iteration are measured relative to the current size of the graph (i.e., with supernodes removed), not the size of the original, unmodified graph. Figure 4 depicts the results. Flickr quickly breaks apart as the core is removed. As shown in Table II, the radius and diameter of Flickr are significantly larger than Facebook and Orkut, but the average path length is essentially equal. This indicates that the core of the Flickr graph is systemically important: as it gets removed, the outlier nodes that are responsible for the large graph diameter quickly disconnect from the connected component.

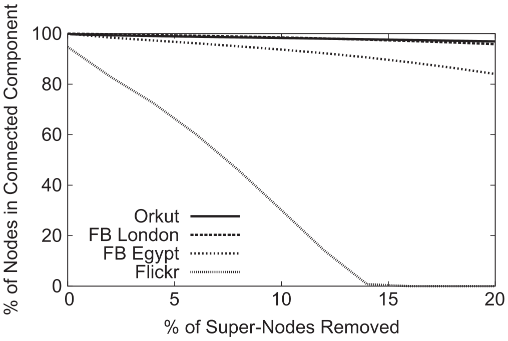

Fig. 4. Percentage of nodes remaining in the connected component as supernodes are removed from various social graphs.

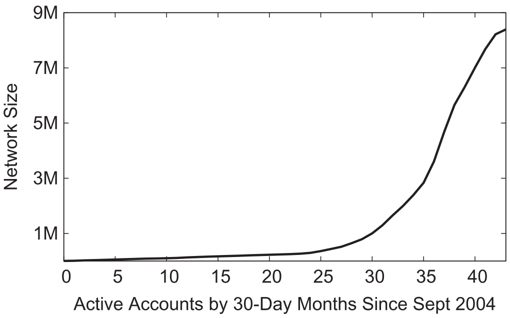

Fig. 5. The growth of users in our sample set, starting in September 2004.

Conversely, Facebook networks are highly resilient to supernode removal. Even when 20% of supernodes are removed, the graph still retains 96% of nodes for Facebook London (our largest region), and 85% for Egypt (our smallest). The rest of our Facebook regions display characteristics in-between these two. These results remain the same even if nodes are ordered using different importance metrics, such as distance-centrality or clustering coefficient.

Figure 4 demonstrates the importance of graph density in social networks. In denser graphs like Facebook and Orkut, the systemic importance of supernodes decreases. There exist so many connections between disparate, low-to-average-degree users that even in the total absence of high-degree nodes, the graph does not partition.

4.1.6. 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 5 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 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?

4.2.1. Interaction Distribution among Friends. We first examine the difference in size between each user's entire friend list and the subset they actually interact with. 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

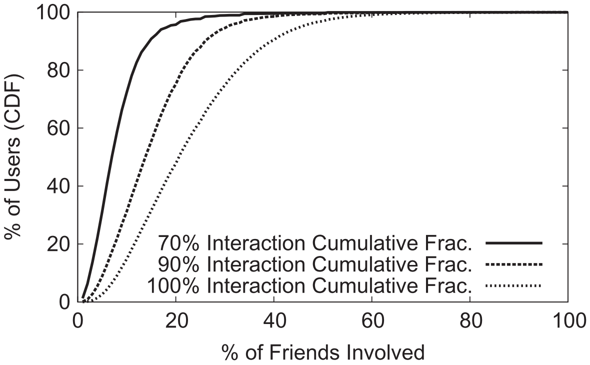

Fig. 6. The distribution of users' interaction among their friends, for different % of users' interactions.

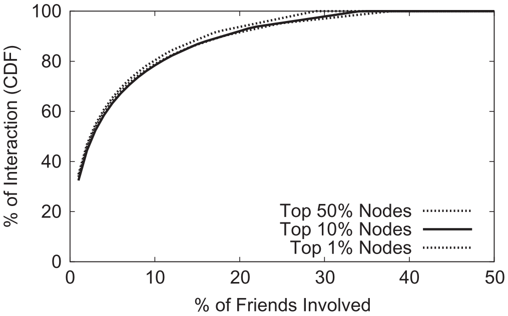

Fig. 7. Normalized Wall post distribution of the users with top total Wall interaction.

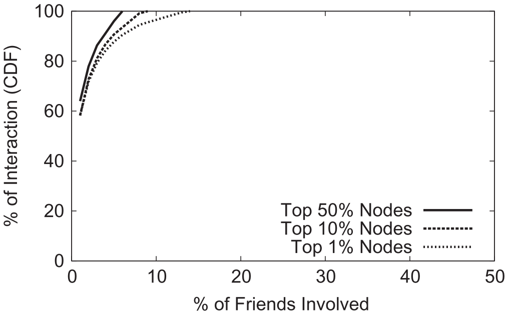

Fig. 8. Normalized photo comments distribution of the users with top total photo interaction.

result is a cumulative fraction function plotted in Figure 6. This is essentially a CDF showing corresponding points from each user's CDF. We see that for the vast majority of users (∼ 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 result allows us to answer our original question: are social links valid indicators of real user interaction? The answer is no, 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 7 and 8 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 7 shows that the same distribution holds across all Wall users regardless of their overall activity level. In contrast, distribution of photo comments in Figure 8 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 (∼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 dataset, 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

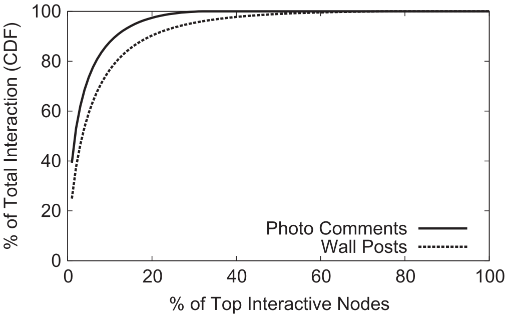

Fig. 9. The contribution of different users to total interactions in Facebook.

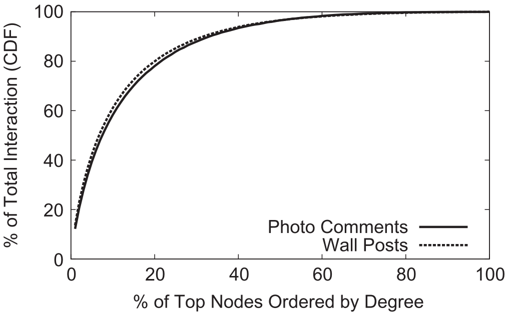

Fig. 10. Plot of top % of users ordered by social degree and the interaction contributed by them.

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. Recent studies that focus on location-based OSNs (e.g., Foursquare, Gowalla, etc.) further examine the correlations between online friendship and geographic proximity [Chang and Sun 2011; Scellato et al. 2010].

4.2.2. 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 9 to further understand the contribution of highly interactive users to the overall interaction in the social 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 when analyzing social networks, since only a small fraction of users are actively engaged with the social 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 10 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.

4.2.3. 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 11 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"

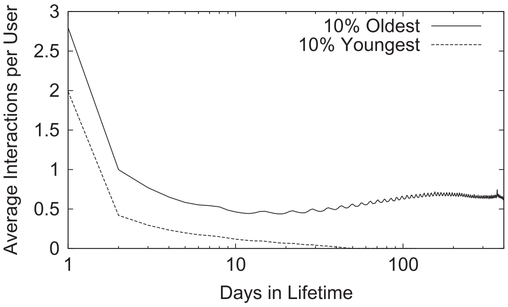

Fig. 11. Average number of interactions per day for old and new Facebook users.

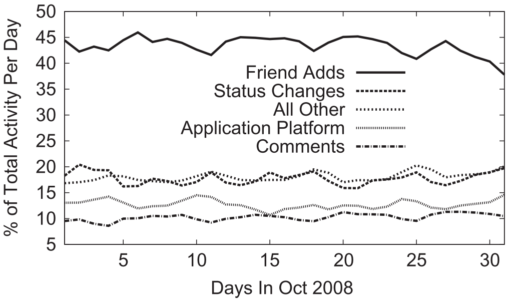

Fig. 12. Distribution of user actions in October from the Mini-Feed.

caused by more friends joining the social network over time (see Figure 5). 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. Additionally, our dataset does not capture 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 12 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 ∼45% of daily events. Comment activity, which encompasses both Wall posts and photo comments, only accounts for ∼10% of daily activity. Application platform events, which include events generated from all other applications, account 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. 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 concentrated 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.

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

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 dataset 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*. We define an interaction graph as an undirected graph $G'(n,t) = (V, I)$. A social graph $G = (V, E)$ and interaction graph G' share the same set of vertices V. However, G uses edge set E (the social links between users) while G' uses edge set I (the interactions between users). Recall that $I \subseteq E$. Although I is a multiset, G' is *not* a multigraph: duplicate edges are simply filtered out.

An interaction graph is parameterized by two constants n and t. These constants filter out edges from the set of interactions I. n defines a minimum number of interaction events for admitting each edge $i_{u,v}$, such that $|i_{u,v}| \geq n$. For example, if $n = 2$, then an edge between users u and v will only exist in the interaction graph if there are 2 or more total Wall posts and photo comments between u and v. t stipulates a window of time during which interactions must have occurred. Taken together, n and t delineate an interaction rate threshold. Intuitively, an interaction graph is the subset of the social graph where for each edge, interactivity between the edge's endpoints is greater than or equal to the rate stipulated by n and t. A user's *interaction degree* is the number of nodes adjacent to u in G'. The equivalent metric on G is a node's degree, or $deg(u)$.

Interaction graphs differ from *inference graphs* because interactions between users are explicit. Recall that, in this article, the interactions we are analyzing come from Wall posts and photo comments on Facebook. These interactions record actual events (with a source, a destination, and a timestamp) generated by users, without any ambiguity. In contrast, inference graphs model the edge relationships between nodes using the concept of "similarity," for example, nodes that share similar metadata attributes should be connected via an edge [Vert and Yamanishi 2004].

We define interaction graphs as undirected for two reasons. First, making interaction graphs undirected allows us to directly compare them to social graphs using the same graph metrics. Second, in Section 6, we compare the performance of social applications on social graphs and interaction graphs. Social applications are designed for undirected graphs, and hence we must define interaction graphs the same way.

Although interactions are inherently directed, it is reasonable to represent them as undirected if it can be shown that, for a given dataset, 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 to reparameterize the interaction graph such that the interaction thresholds n and t no longer cull links, but instead impart a weight to each edge. We do not attempt to derive a weight scheme for interaction graphs analyzed in this article, 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, that is, $I \subseteq E$. Facebook only allows friends to post Wall and photo comments, thus this assumption holds true

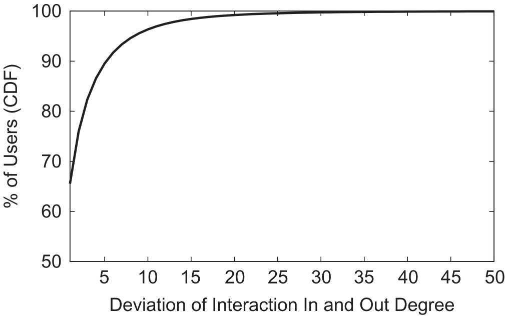

Fig. 13. Deviations in pairwise interaction patterns on Facebook.

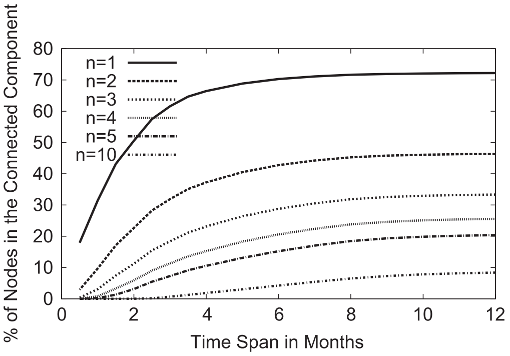

Fig. 14. Percentage of nodes remaining in interaction graph connected components as n and t vary.

for our dataset. However, it is conceivable to envision other social networks that do not share these restrictions. In this case it may 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 a writes on b's Wall, b 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, that is, the event a writes on b's Wall is only recorded on b's Wall, not a. Since we are limited to users within specific regional networks 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, that is, ignore interactions with users outside our dataset. 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 13 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. Unfortunately, this model overestimates the number of edges supernodes will have in the interaction graph, since the deviation for celebrities is high due to practical constraints. However, these high-degree nodes account for $< 1\%$

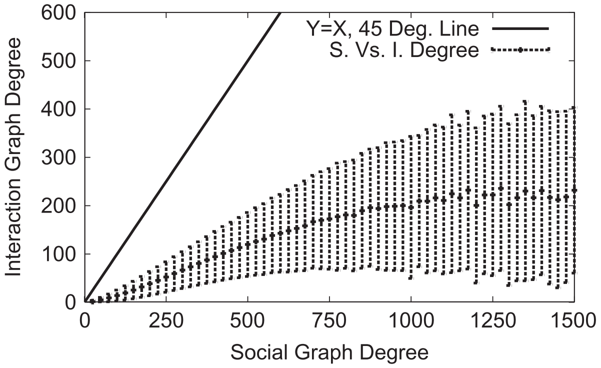

Fig. 15. Comparison of the Facebook social graph degree and interaction graph degree.

of the Facebook population, and they are limited to 5,000 friends maximum, so the effects of this overestimation are minimal.

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 = 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 14 shows the size of the connected components for interaction graphs as t and n change. Intuitively, higher n filters out more edges since user pairs need to have a greater number of pairwise interactions. Similarly, smaller t also results in fewer edges, since the span of time during which interactions must occur is tighter. This figure is based on data for the year 2007, that is, 2 months refers to interactions occurring between November 1 and December 31, 2007.

As expected, lower n and larger t are less restrictive on links, therefore allowing for more nodes to remain connected. Based on Figure 14, we choose 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 include connected components that contain a majority of all nodes, and are amenable to graph analysis. For the remainder of this article 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 graphs, small-world clustering, and scale-free graphs.

5.3.1. Social vs. Interaction Degree. Figure 15 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.

5.3.2. Interaction Degree Analysis. Figure 16 plots the degree CDFs of the four interaction graphs and the Facebook social graph. The interaction graphs exhibit a larger

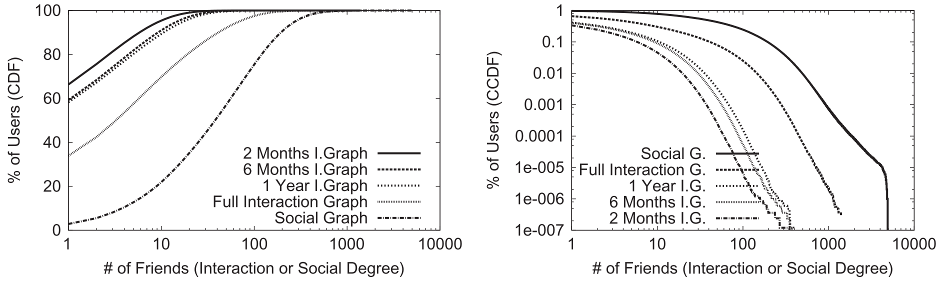

Fig. 16. Comparison of degree distributions for interaction graphs and the full social graph. Both CDF and CCDF distributions are shown.

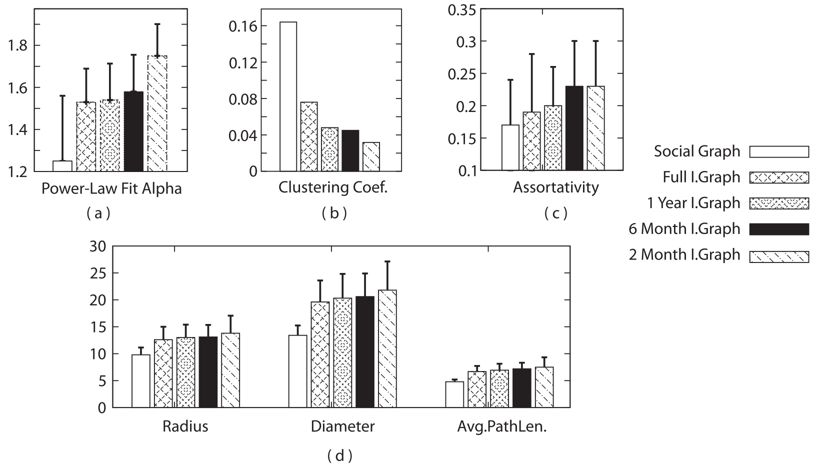

Fig. 17. Graph measurements for four interaction graphs compared to the entire Facebook social network.

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 6, 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 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 17(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 et al. 2009]. The fitting error for the interaction graphs is 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 16 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 et al. 2007].

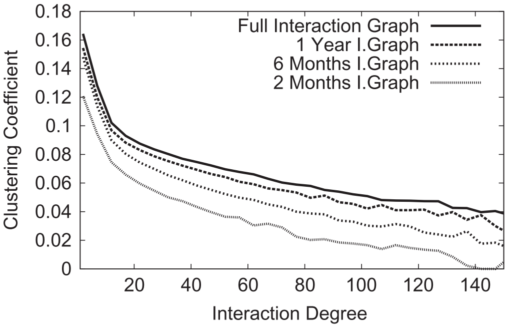

Fig. 18. Clustering coefficient of interaction graphs as a function of interaction degree.

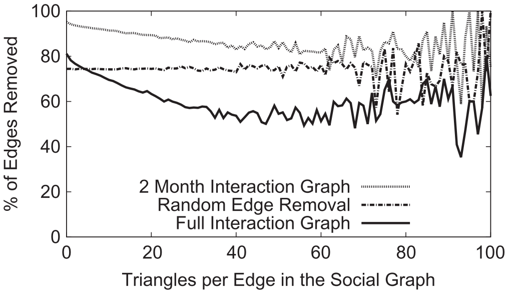

Fig. 19. Percent of edges cut in different interaction graphs versus the original social graph.

5.3.3. Interaction Graph Analysis. Figure 17(d) 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 "supernodes" decreases (see Figure 16) 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 17(d).

5.3.4. Clustering Coefficient Measurements. Besides average path length, another metric intrinsically linked to node connectivity is the clustering coefficient. Figure 17(b) shows that average clustering coefficient drops as interaction graphs become more restricted. Figure 18 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.

To understand why clustering coefficient drops in the interaction graphs, we examine which edges are removed from the social graph. Figure 19 depicts the percent of edges removed from different interaction graphs. Edges are grouped together based on how many triangles they are part of in the original social graph. For example, edges that are not part of any triangles fall into the 0 bucket. In contrast, an edge that forms one side of twenty unique triangles will fall into the 20 bucket. Edges that complete many triangles are more systemically important for increasing average clustering. The "Random Edge Removal" line acts as an experimental control: in this scenario, x edges are randomly removed from the social graph, where x is the difference in edges between the social graph and the full interaction graph.

Figure 19 reveals that edge importance (in terms of triangles) does correlate with how likely that edge is to be retained in the interaction graphs. Low importance edges (e.g., 0 or 1 triangles) are about 20% more likely to be removed than high importance edges. This contrasts with random removal, where edges of all types are equally likely to be removed (all lines are jagged when the number of triangles >60 because such high importance edges are rare). This result indicates that edges which complete many triangles are more likely to correspond to active social relationships.

However, in absolute terms, *all* edges are >50% likely to be removed in the interaction graphs, irrespective of importance. Thus, although interaction graphs retain a higher percentage of important edges than random chance predicts, a large number of triangles are still being severed. This problem is particularly acute for edges that complete many triangles, since these edges are more vital for high clustering coefficients, and much more rare.

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 and Strogatz 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.

5.3.5. Assortativity Measurements . Figure 17(c) 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 [Marti et al. 2004; Watts and Strogatz 1998; Yu et al. 2006], or bootstrap social networks using traces of emails [Garriss et al. 2006].

The hypothesis of our work is that validation of socially enhanced applications requires 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 three well-known socially enhanced distributed systems [Chen et al. 2009; Garriss et al. 2006; Yu et al. 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 et al. 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. The key advantage of RE is that it works automatically based on social connectivity data: users do not have to take the time to manually create white-lists of authorized senders.

6.1.1. Expected Impact. The presence of small-world clustering and scale-free behavior in social graphs translates 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 each given

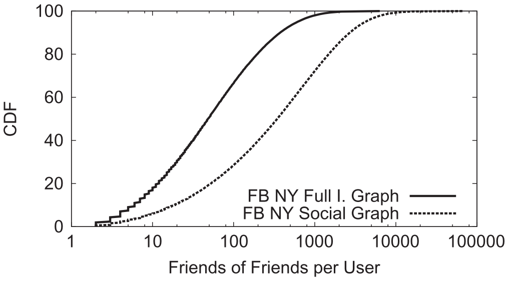

Fig. 20. Friends-of-friends per user in the Facebook New York social and full interaction graphs.

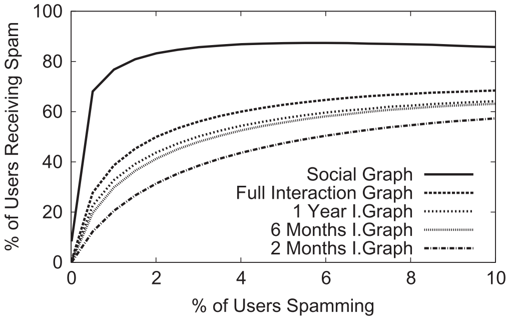

Fig. 21. Spam penetration as the number of spammers is varied for the Reliable Email system.

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.

In contrast, RE that leverages interaction graphs should not experience as high a proliferation of spam, given an equal number of spammers. Figure 20 shows the number of friends-of-friends per user in the Facebook New York social graph and full interaction graph. The size of each user's friend-of-friend set is reduced by about an order of magnitude in the interaction graph. Similar results are present in the other Facebook regional graphs. The reduced size of the friend-of-friend set should limit the dissemination potential for spam, while still maintaining the key advantage of RE, that is, users do not need to manually enumerate white-lists of senders.

6.1.2. 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 21, 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 the users in the connected component. In contrast, spam penetration when RE is run on the interaction graph is reduced by 40% over the social graph when the number of spammers is low, and 20% when the number of spammers is high.

The secondary benefit of using RE on interaction graphs is that spammers are naturally excluded from the graph. Intuitively, honest users are unlikely to interact with spammers. Hence the interaction graph should have few edges connection spammers and honest users, which prevents spammers from being white-listed by RE. This phenomena has been observed by prior work that examined social e-commerce marketplaces: scammers were almost totally excluded from the interaction graph [Swamynathan et al. 2008].

Unfortunately, the Achilles' heel of RE is that friend's accounts can be compromised by attackers. Spam sent from compromised accounts will successfully disseminate due to RE's white-listing of friendly accounts. This shortcoming is equally damaging when using RE on social and interaction graphs.

6.2. SybilGuard

A Sybil attack [Douceur 2002] occurs when a single attacker creates a large number of online identities that can collude together and grant the attacker 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 et al. 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 connections (called *attack edges*) to the honest nodes in the graph.

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 v as a real user, a "verifier" node u initiates n random walks of length w on the social graph. Node v also initiates n random walks of length w. u accepts v if some predefined percentage of the random walks intersect. w is the most important parameter in SybilGuard: as w grows, the number of Sybils that will be erroneously accepted grows. Thus, it is beneficial for w to be small.

6.2.1. 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 friend pairs do not even interact, it seems unreasonable to assume that most friend pairs have the requisite level of trust to exchange secure keys.

Instead, we expect that the interaction graph is a closer approximation to the representation of trusted links that SybilGuard would observe in reality. Unfortunately, under these conditions, we expect the effectiveness of SybilGuard to *decrease*. SybilGuard's functionality is dependent on the fast mixing behavior of graphs. Mohaisen et al. [2010] provide an overview of the mixing behavior of "trusted" social networks (e.g., physics coauthorship, Enron emails), "untrusted" social networks (e.g., Facebook), and interaction graphs. Their results confirm that the mixing properties of interaction graphs closely resemble trusted social networks, and that both are slow mixing.

6.2.2. Results. For our experiments, we implement the SybilGuard algorithm on both our social graph and interaction graphs and measure the percentage of random walks 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 supernodes in the interaction graph means that random walks are less likely to connect. Figure 22 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.

A major factor affecting the performance of the SybilGuard algorithm is the prevalence 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 III shows the total number of self-loops encountered during all experimental runs on each graph. The drop in efficacy observed in Figure 22 is directly correlated

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

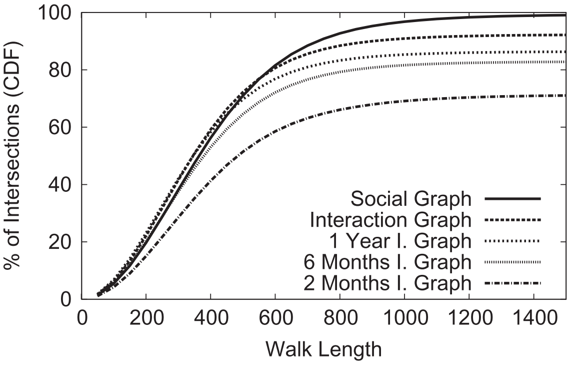

Fig. 22. Percentage of path intersections for SybilGuard as random walk length increases.

Table III. Self-Looping Statistics for SybilGuard

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)

to the increase in self-looping from 3.8% on the social graph to an upwards of 20% on interactions graphs.

6.3. Influence Maximization

The capability to model and predict the spread of information through social networks has many real-world applications. These range from combating the spread of disease to generating effective word-of-mouth marketing campaigns. An important problem in this area is *influence maximization*: locating the most influential users who will maximize the spread of information through the social network.

Previous works have designed algorithms that use statistical methods to model information dissemination over social links [Chen et al. 2009; Kempe et al. 2003]. One such model is the *weighted cascade model*. In this model, each user u that is "activated" by receiving or producing some new information has a chance to activate his/her friend v with probability $1/deg(v)$. This process is repeated for all of u's friends. The Mixed-GreedyWC algorithm implements the weighted cascade model and calculates, for a given social network topology, the most influential users (called "seeds") and the set of nodes influenced by them [Chen et al. 2009].

6.3.1. Expected Impact. The weighted cascade model assumes that, for a given node, the probability of activating each neighbor is proportional to their degree. However, as we have demonstrated in this work, not all social links are equally important. A node is more likely to be influenced by users it interacts with, as opposed to familiar strangers. Thus, we propose running the weighted cascade model on interaction graphs, as the interaction graph prunes out edges that are unlikely to ever be activated in reality. The reduction in average node degrees will increase the activation probability of the remaining links in the graph. However, the overall reach of each node will be reduced, thus constraining the spread of information as compared to the full social graph.

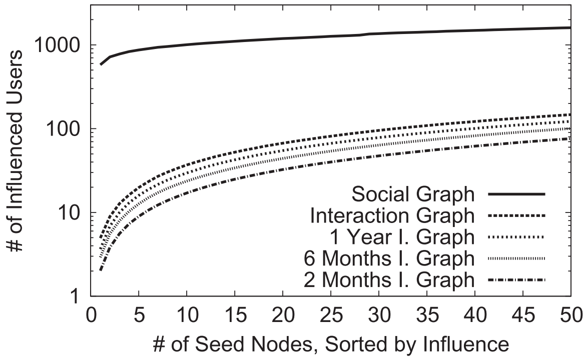

Fig. 23. Number of users influenced by a given number of seed nodes across social and interaction graphs. Seed nodes are sorted from most to least influential.

6.3.2. Results. Figure 23 shows the results of running the MixedGreedyWC algorithm on our social and interaction graphs. Each line is an average over 12 of our Facebook regional graphs, since the largest regions are too big to be processed by the software. For all graph types, the total number of influenced users grows as the number of highly influential seed nodes is increased. However, the relative number of users reached is an order of magnitude lower for the interaction graphs, as compared to the full social graph. This is due to a combination of factors. Firstly, there are fewer total nodes in the interaction graphs, since nodes that do not interact at all have zero links. This shrinks the pool of potential targets. More importantly, the interaction graphs also have drastically fewer edges than the full social graph (average node degree drops from ∼77 on the full graph to 1 for the time-constrained interaction graphs). This has the effect of limiting the potential reach of seed nodes. These results agree with other published comparisons of the weighted cascade model on social and interaction graphs [Jiang et al. 2010].

We compared the 50 seed nodes chosen by the MixedGreedyWC algorithm on the 12 social graphs versus the corresponding full (i.e., nontime constrained) interaction graphs. Only 11% of the seeds from the social graphs are also selected on the full interaction graphs. This demonstrates that the optimal seed selection changes depending on the type of graph being examined. This result agrees with Figure 10, which shows that the highest-degree nodes on the social graph are not necessarily the most interactive.

In terms of practical impact, these results indicate that researchers examining information dissemination and influence maximization should take care when performing experiments. Assuming uniform information spread along all social links can lead to overestimation of information dissemination, as well as leading to the selection of influential seeds that may not be optimal on more constrained graph topologies.

7. FACEBOOK OVER TIME

Since our initial work in this area, Facebook has continued to grow and mature. The user base has grown exponentially since our original data was collected in 2008, recently reaching the 800 million user milestone. The site itself has also gone through significant architectural changes, such as the shift towards a Twitter-like, News-Feed-centric interface. All of these changes beget the question: do the social graph and interaction characteristics observed in Facebook 2008 continue to hold true?

In this section, we address this question by performing a comparative analysis between data gathered from Facebook in 2008 and 2009. We also compare our results

Table IV. Statistics for 2009 Facebook Regional Networks Compared to 2008

Network	Nodes (%)	Links (%)	Rad.	Diam.	Avg.PathLen.
London	1,690K (36)	46,169K (50)	11/10	15/15	5.09/4.99
New York	905K (139)	21,230K (194)	11/11	14/15	4.80/4.77
Sweden	651K (13)	23,213K (34)	8/8	11/12	4.55/4.36
Los Angeles	603K (119)	15,352K (263)	12/10	16/15	5.14/4.54
Mexico	598K (90)	9,104K (39)	9/9	13/15	4.89/5.22
Egypt	298K (21)	5,047K (56)	9/8	12/13	4.88/4.58
Total	4,745K (57)	120,115K (73)	10/9	13/13	4.8/4.74
Facebook 2011*	721M	68.7B	N/A	≥11	4.7

(%) are percent increases from 2008 to 2009. Other values are presented as 2008/2009. The final row (*) shows the values for the entire Facebook social graph in 2011 from Ugander et al. [2011].

to those from a recent study of the Facebook social graph conducted in 2011 [Ugander et al. 2011]. Our results show that while the rapid growth of Facebook's user population has added weight to the long tail of the social graph, the overall trends of interactions on Facebook remain the same.

7.1. Description of Collected Data and Methodology

In order to validate our conclusions drawn from Facebook 2008 data, we crawled additional data from Facebook in 2009. We crawled 6 regional networks between April and June of 2009, just over one year after our original crawls. This resulted in data on 4.7 million users with 120 million friend links (see Table IV), at a time when Facebook's total population was ∼200 million [Zuckerberg 2009]. Our crawl methodology remained the same as for the 2008 crawls: 50 random users were chosen to seed the crawlers BFS of each region. These crawls were conducted before Facebook deprecated the networks feature in summer of 2009.

Between 2008 and 2009, the Facebook site went through significant architectural and usability changes, the most significant of which was the move to a News-Feed-centric profile layout. These changes impact the type and amount of per-user information accessible to our crawlers. In 2008, each user's profile page was composed of different applications such as photos, Wall, and events, each of which inhabited a separate area of the page. The data contained in each application domain was completely separate from the data in other applications. This partitioning made it straightforward to completely crawl application-specific data. In 2009, Facebook began moving towards its current architecture, which is centered around the News-Feed. Each users' News-Feed aggregates all of their status updates, as well as all incoming interactions from friends.

We gathered interaction data from each crawled user by downloading their News-Feed histories going back to January 1st, 2008. This gives us a complete 1.5 year record of incoming interactions and status updates for each user. Limitations stemming from Facebook's back-end architecture made it impractical to crawl older Feed data. However, because Facebook's population more than doubled between 2008 and 2009, this 1.5 year history encompasses the full lifetime of the majority of Facebook accounts. In total, we gathered 244 million interactions between Facebook users.

Each Feed item is characterized by a sender and receiver, a timestamp, an application descriptor, and an application-specific data payload. The application descriptor either refers to one of the built-in Facebook applications, such as Wall, photos, or events, or to a third-party application (referenced by a unique application ID). Although Facebook supported "likes" and comments on Feed items during the time of our crawls, security measures prevented us from reliably gathering these interactions. Thus, our

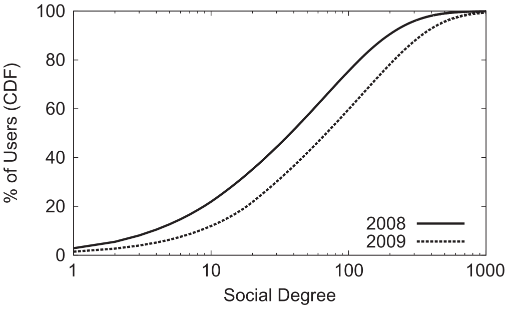

Fig. 24. Comparing social degree in Facebook 2008 to 2009.

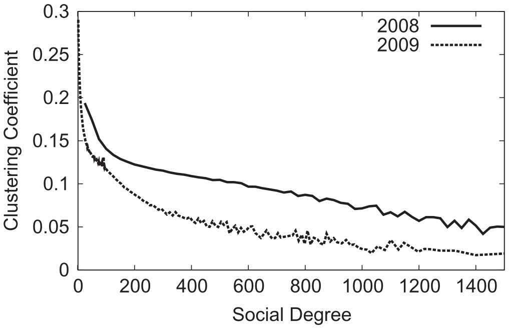

Fig. 25. Clustering coefficient of Facebook users in 2008 and 2009.

2009 interaction data should be viewed as a lower bound on the total number of interactions on Facebook at the time.

To maintain compatibility with our 2008 study, our 2009 interaction analysis focuses on Feed items from the Wall application. Facebook lumps all text comments into this umbrella category, meaning the Wall application includes users' traditional Wall posts, as well as comments left on photos, events, videos, notes, etc. Items from the Wall application account for over 85% of all interactions in our 2009 dataset.

7.2. Social Graph Analysis

We begin our comparison between Facebook 2008 and 2009 by focusing on the overall social graph. Table IV shows the number of nodes and edges in each of our 2009 regional networks, as well as the percent increase in size of each compared to 2008. Overall, the user base of the regions grew by over 57%. More significantly, the number of edges grew by 73%, outstripping the growth of the user population. This results in higher average social degrees for users in 2009, which in turn causes the radius and average path lengths for 2009 social graphs to decrease slightly. Between 2009 and 2011 Facebook's user population grew by an additional order of magnitude to 721 million, but the average path length and diameter of the graph stayed relatively constant. This indicates that the Facebook graph may have reached an equilibrium point by 2009.

Figure 24 depicts the social degree CDF for Facebook 2008 and 2009. The 2009 graph shifts to the right of 2008, reflecting the increase in average social degree during this time period. The two lines reconverge around the 900 friend mark, indicating that the additional links fueling this growth are not concentrated among supernodes. On the contrary, Facebook's hard limit of 5000 friends ensures that additional edges are formed between lower-degree users. The power-law coefficient for Facebook 2009 is 1.21 with a fitting error of 0.34157, which is slightly lower than the alpha value of 1.25 observed for 2008.

As a node's degree increases, its clustering coefficient usually drops commensurately, since the likelihood of forming complete three-person friend cliques is reduced. However, as shown in Figure 25, even users of the same degree have lower clustering coefficients in 2009 than users in 2008. Although the overall trend remains the same, that is, lower-degree nodes demonstrate more local clustering, the drop in 2009 reflects changing dynamics in Facebook. As the overall population of Facebook grows, user's friend bases are diversifying such that the likelihood of sharing mutual acquaintances with your friends is reduced. The average clustering coefficient for users with degree = 100 is 0.14 in Facebook 2011 [Ugander et al. 2011], which is in-between the values for 2007 and 2009.

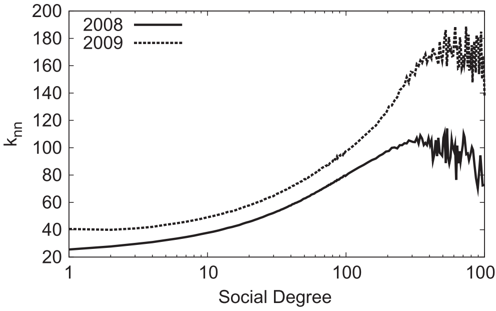

Fig. 26. k_{nn} of Facebook users in 2008 and 2009.

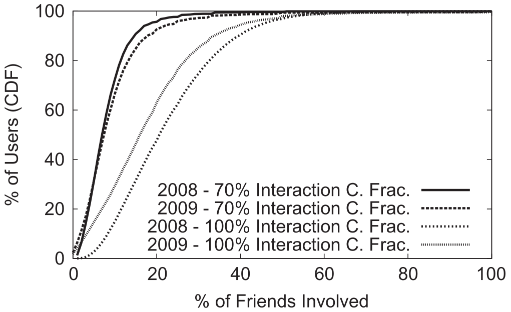

Fig. 27. The distribution of users' interaction among their friends, for users in 2008 and 2009.

As Figure 26 demonstrates, k_{nn} values in Facebook 2009 are generally higher than in 2008, with the peak values for supernodes being almost twice as high. In this case, the higher average node degrees for 2009 shown in Figure 24 translate to increased k_{nn} values: higher node degrees on average cause the average degree of all friends for each given node to also increase commensurately. The large increases for high-degree nodes indicates increasing homogeneity for this class of users, that is, supernodes in 2009 are more likely to be friends with other supernodes than 2008.

The assortativity for Facebook in 2007, 2009, and 2011 is 0.17, 0.21, and 0.22, respectively [Ugander et al. 2011]. Once again, it appears that Facebook reached an equilibrium point in the structure of the social graph around 2009.

7.3. Interaction Analysis

At this point we have examined how the social graph of Facebook changed in the one-year period between spring of 2008 and 2009. The next set of comparative tests examines how visible interactions between users have changed during this period.

The first question we revisit is how interactions are distributed among each user's friends. Figure 27 depicts the results when considering 70% and 100% of total interactions. The trends for each range are similar for 2008 and 2009, and thus the high-level conclusion of this figure remains the same as our original discussion in Section 4: users do not interact with the majority of their friends. The 100% lines are slightly divergent, indicating that users in 2009 interact with less of their friends than in 2008. This result is consistent with our observation of increased average node degrees: even though each user's friend base keeps accumulating, the amount of time he/she has to dedicate towards social interactions remains limited and fixed.

Figure 28 demonstrates that the top interactive users on Facebook contribute less towards total interactions in 2009 than in 2008. This can be attributed to the exponential growth in the Facebook user population during this time, which results in many more users interacting overall. Even though these casual users may only interact seldomly, taken in aggregate their numbers are large enough to dwarf the output of the most interactive users. This observation also holds true in Figure 29, which plots the contributions of the highest-degree nodes to total interactions.

In summation, we observe that the high-level conclusions we have drawn about interactions on Facebook in 2008 also hold true in 2009. Specifically, we observe that: (1) interactions are confined to a subset of each user's friends (Figure 27), (2) interactions are skewed towards a highly active subset of the population (Figure 28), and (3) supernodes are not necessarily the most interactive users on Facebook (Figure 29). However, the massive population growth on Facebook between 2008 and 2009 does have an effect on interaction patterns. The overall increase in average node degrees

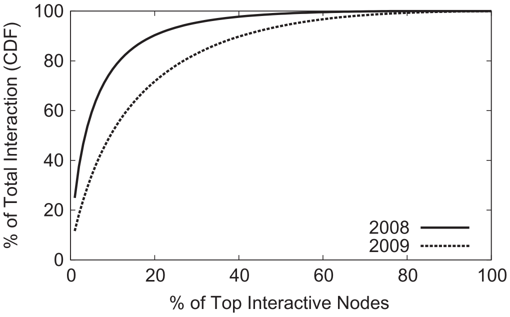

Fig. 28. The contribution of the most interactive users to total interactions in 2008 and 2009.

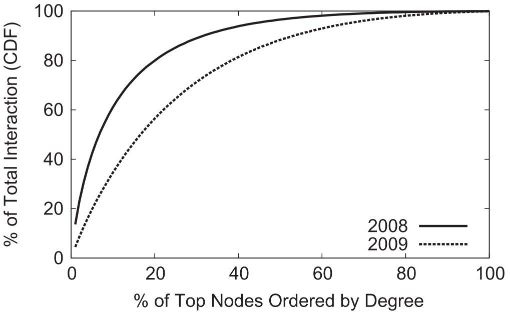

Fig. 29. The contribution of the most well-connected nodes to total interactions in 2008 and 2009.

means users end up interacting with even less of their "friends." Similarly, the increase in normal users relative to supernodes reduces the impact of high-degree and high-interaction nodes on total interactions. These results are to be expected, given the changes in the social graph over the measurement period.

8. RELATED WORK

The body of research geared towards real-world social webs and physical networks has only begun to be leveraged to understand online social networks within the last few years. The original papers in this area focused primarily on static, structural characteristics of OSN graphs. One of first was a study focusing on the Club Nexus Web site of Stanford University [Adamic et al. 2003]. Since then, traces from CyWorld, MySpace, and Orkut [Ahn et al. 2007] have been analyzed, as well as YouTube, Flickr, and LiveJournal [Mislove et al. 2007]. Other studies have focused on social graph evolution over time using traces from Flickr and Yahoo! 360 [Kumar et al. 2006; Mislove et al. 2008]. These studies confirm that online social networks obey power-law scaling characteristics [Barabasi and Albert 1999] and exhibit high clustering coefficients, firmly establishing them as small-world graphs [Amaral et al. 2000].

More recently, OSN studies have shifted focus to analyzing interactions between users, rather than static graph topologies alone. Prior to our work, there were two studies in this vein, one focusing on the online communication patterns among users in a large IM trace [Leskovec and Horvitz 2008], and one targeting users of the CyWorld OSN [Chun et al. 2008]. Like our study, the CyWorld interaction study [Chun et al. 2008] showed that CyWorld user interactions are reciprocal. 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 multiscaling behavior. This multiscaling is unique to CyWorld; all other social networks analyzed so far, including Facebook, exhibit simple power-law connectivity scaling [Ahn et al. 2007; Leskovec and Horvitz 2008; Mislove et al. 2007].

Since our initial work was published, numerous other studies have been published that analyze visible interactions between OSN users. Two papers have leveraged Flickr data to study user interactions and the dynamics of photo popularity over time [Cha et al. 2009; Valafar et al. 2009]. Twitter has rapidly grown in popularity in recent years, and its relatively open data access policies have made it the target of several graph topological and user interaction studies [Cha et al. 2010; Kwak et al. 2010; Yang and Counts 2010]. Facebook has also been the subject of additional studies focused on time-varying dynamics in user interactions [Viswanath et al. 2009]. The general consensus from this growing body of research reinforces the findings of this article: only a fraction of all friendship links represent active connections between users, interactions are not spread evenly over users' friends, and the relative interactivity of different links varies with time. These observations have led to the formulation of models that attempt to predict pairwise tie-strength between OSN and email users using interaction events [Choudhury et al. 2010; Xiang et al. 2010].

The study of visible interactions on OSNs has naturally led to work focused on nonvisible, or latent, user interactions. Latent interactions are characterized by things like profile browsing or photo viewing, where (usually) no explicit trace of the behavior remains, except for perhaps in the OSN provider's private HTTP logs. This lack of available data makes latent interactions a challenge to study. Latent interactions on Facebook, MySpace, LinkedIn, Hi5, and StudiVZ were studied by using anonymized HTTP traces captured at ISP level [Schneider et al. 2009]. Similarly, latent interactions on Orkut, MySpace, Hi5, and LinkedIn were analyzed using HTTP session data collected at a Brazilian social network aggregator site [Benevenuto et al. 2009]. Latent interactions on the Chinese OSN Renren were collected and analyzed by leveraging features unique to Renren that display recent visitors on each user's profile [Jiang et al. 2010]. Finally, visible and latent interactions on Facebook have also been characterized [Backstrom et al. 2011]. These studies demonstrate that latent interactions exhibit different characteristics than visible interactions: more friend links tend to be active for browsing behavior, and latent interactions are nonreciprocal. These results suggest that social application developers need to evaluate what types of interactions are important to their applications (visible or latent) before settling on an appropriate evaluation model.

9. CONCLUSION

This article 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 in 2008, 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 reveals 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.

We conduct experiments to evaluate the effects of interaction graphs on three well-known social applications. The performance of RE [Garriss et al. 2006] improves with the use of interaction graphs, as the streamlined link structure helps control spam proliferation. In the case of SybilGuard [Yu et al. 2006], the system becomes less able to effectively classify nodes once its assumptions about graph structure are violated. Furthermore, we observe that while computing influence maximization [Chen et al. 2009], the selection of influential nodes and their effective range of influence changes when interactivity is taken into account. 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.

Finally, we reexamine our conclusions about interactions on Facebook using additional data crawled from Facebook in 2009. We demonstrate that although Facebook's user population continues to grow exponentially, interaction patterns between users remain largely the same. This shows that our conclusions generalize over time, and are likely to remain applicable for the foreseeable future.

To support the efforts of the social network research community, we make available a selection of anonymized social and interaction graphs from our Facebook dataset. Details on the available graphs, as well as instructions for requesting access to the data, are available on our lab Web site[3].

ACKNOWLEDGMENTS

We thank Alan Mislove and colleagues at MPI-SWS for sharing their datasets on YouTube, LiveJournal and Orkut.

REFERENCES

ADAMIC, L. A., BUYUKKOKTEN, O., AND ADAR, E. 2003. A social network caught in the web. *First Monday 8*, 6.

AHN, Y.-Y., HAN, S., KWAK, H., MOON, S., AND JEONG, H. 2007. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the International Conference on the World Wide Web (WWW'07)*.

AMARAL, L. A. N., SCALA, A., BARTHELEMY, M., AND STANLEY, H. E. 2000. Classes of small-world networks. *Proc. Nat. Acad. Sci.* 11149–11152.

BACKSTROM, L., BAKSHY, E., KLEINBERG, J. M., LENTO, T. M., AND ROSENN, I. 2011. Center of attention: How facebook users allocate attention across friends. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'11)*.

BARABASI, A.-L. AND ALBERT, R. 1999. Emergence of scaling in random networks. *Science 286*.

BENEVENUTO, F., RODRIGUES, T., CHA, M., AND ALMEIDA, V. 2009. Characterizing user behavior in online social networks. In *Proceedings of the Internet Measurement Conference (IMC'09)*.

BOE, B. AND WILSON, C. 2008. Crawl-E: Highly distributed web crawling framework written in python. Google Code.

BRODER, A., KUMAR, R., MAGHOUL, F., RAGHAVAN, P., RAJAGOPALAN, S., STATA, R., TOMKINS, A., AND WIENER, J. 2000. Graph structure in the web: Experiments and models. In *Proceedings of the International Conference on the World Wide Web (WWW'00)*.

CHA, M., MISLOVE, A., AND GUMMADI, K. 2009. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the International Conference on the World Wide Web (WWW'09)*.

CHA, M., HADDADI, H., BENEVENUTO, F., AND GUMMADI, K. P. 2010. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the International Conference on the Weblogs and Social Media (ICWSM'10)*.

CHANG, J. AND SUN, E. 2011. Location3: How users share and respond to location-based data on social networking sites. In *Proceedings of the International Conference on the Weblogs and Social Media (ICWSM'11)*.

CHEN, W., WANG, Y., AND YANG, S. 2009. Efficient influence maximization in social networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*.

CHOUDHURY, M. D., MASON, W. A., HOFMAN, J. M., AND WATTS, D. 2010. Inferring relevant social networks from interpersonal communication. In *Proceedings of the International Conference on the World Wide Web (WWW'10)*.

CHUN, H., KWAK, H., EOM, Y., AHN, Y., MOON, S., AND JEONG, H. 2008. Comparison of online social relations in volume vs interaction: A case study of cyworld. In *Proceedings of the Internet Measurement Conference (IMC'08)*.

CLAUSET, A., SHALIZI, C. R., AND NEWMAN, M. E. J. 2009. Power-Law distributions in empirical data. *SIAM Rev.* To appear.

DONATH, J. AND BOYD, D. 2004. Public displays of connection. *BT Tech. J. 22*, 4.

DOUCEUR, J. R. 2002. The Sybil attack. In *Proceedings of the International Workshop on Peer-to-Peer Systems (IPTPS'02)*.

[3] http://current.cs.ucsb.edu/facebook/

Facebook. 2008. Statistics. facebook.com.

Garriss, S., Kaminsky, M., Freedman, M. J., Karp, B., Mazi‘eres, D., and Yu, H. 2006. Re: Reliable email. In *Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI'06)*.

Golder, S. A., Wilkinson, D., and Huberman, B. A. 2007. Rhythms of social interaction: Messaging within a massive online network. In *Proceedings of the Conference on Communities and Technologies*.

Jiang, J., Wilson, C., Wang, X., Huang, P., Sha, W., Dai, Y., and Zhao, B. Y. 2010. Understanding latent interactions in online social networks. In *Proceedings of the Internet Measurement Conference (IMC'10)*.

Kempe, D., Kleinberg, J. M., and Tardos, E. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03)*.

Kumar, R., Novak, J., and Tomkins, A. 2006. Structure and evolution of online social networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*.

Kwak, H., Lee, C., Park, H., and Moon, S. B. 2010. What is twitter, a social network or a news media? In *Proceedings of the International Conference on the World Wide Web (WWW'10)*.

Leskovec, J. and Horvitz, E. 2008. Planetary-scale views on a large instant-messaging network. In *Proceedings of the International Conference on the World Wide Web (WWW'08)*.

Marti, S., Ganesan, P., and Garcia-Molina, H. 2004. DHT routing using social links. In *Proceedings of the International Workshop on Peer-to-Peer Systems (IPTPS'04)*.

Milgram, S. 1967. The small world problem. *Psychol. Today 6*, 62–67.

Milgram, S. 1977. *The Familiar Stranger: An Aspect of Urban Anonymity*. Addison-Wesley.

Mislove, A., Gummadi, K. P., and Druschel, P. 2006. Exploiting social networks for internet search. In *Proceedings of the ACM Workshop on Hot Topics in Networks (HotNets'06)*.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. 2007. Measurement and analysis of online social networks. In *Proceedings of the Internet Measurement Conference (IMC'07)*.

Mislove, A., Koppula, H. S., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. 2008. Growth of the flickr social network. In *Proceedings of the ACM Workshop on Online Networks (WOSN'08)*.

Mohaisen, A., Yun, A., and Kim, Y. 2010. Measuring the mixing time of social graphs. In *Proceedings of the Internet Measurement Conference (IMC'10)*.

Newman, M. E. J. 2003. Mixing patterns in networks. *Phys. Rev. E 67*.

Scellato, S., Mascolo, C., Musolesi, M., and Latora, V. 2010. Distance matters: Geo-social metrics for online social networks. In *Proceedings of the ACM Workshop on Online Networks (WOSN'10)*.

Schneider, F., Feldmann, A., Krishnamurthy, B., and Willinger, W. 2009. Understanding online social network usage from a network perspective. In *Proceedings of the Internet Measurement Conference (IMC'09)*.

Swamynathan, G., Wilson, C., Boe, B., Almeroth, K. C., and Zhao, B. Y. 2008. Can social networks improve e-commerce: A study on social marketplaces. In *Proceedings of the ACM Workshop on Online Networks (WOSN'08)*.

Sweney, M. 2008. Facebook sees first dip in uk users. guardian.co.uk.

Ugander, J., Karrer, B., Backstrom, L., and Marlow, C. 2011. The anatomy of the facebook social graph. Arxiv online pre-print, abs/1111.4503v1.

Valafar, M., Rejaie, R., and Willinger, W. 2009. Beyond friendship graphs: A study of user interactions in flickr. In *Proceedings of the ACM Workshop on Online Networks (WOSN'09)*.

Vert, J.-P. and Yamanishi, Y. 2004. Supervised graph inference. In *Proceedings of the Annual Conference on Neural Information Procesing Systems (NIPS'04)*.

Viswanath, B., Mislove, A., Cha, M., and Gummadi, K. P. 2009. On the evolution of user interaction in facebook. In *Proceedings of the ACM Workshop on Online Networks (WOSN'09)*.

Watts, D. J. and Strogatz, S. 1998. Collective dynamics of ‘small-world’ networks. *Nature* 393, 440–442.

Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P. N., and Zhao, B. Y. 2009. User interactions in social networks and their implications. In *Proceedings of the European Conference on Computer Systems (EuroSys'09)*.

Worthen, B. 2008. Bill Gates quits facebook. *Wall St. J. Online*.

Xiang, R., Neville, J., and Rogati, M. 2010. Modeling relationship strength in online social networks. In *Proceedings of the International Conference on the World Wide Web (WWW'10)*.

YANG, J. AND COUNTS, S. 2010. Predicting the speed, scale, and range of information diffusion in twitter. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'10)*.

YU, H., KAMINSKY, M., GIBBONS, P. B., AND FLAXMAN, A. 2006. Sybilguard: Defending against Sybil attacks via social networks. In *Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures and Protocols for Computer Comminications*.

YU, H., GIBBONS, P. B., KAMINSKY, M., AND XIAO, F. 2008. Sybillimit: A near-optimal social network defense against Sybil attacks. In *Proceedings of the IEEE Conference on Security and Privacy*.

ZUCKERBERG, M. 2009. 200 million strong. The Facebook Blog.

Received November 2010; revised March 2012; accepted July 2012