Measurement and Analysis of Online Social Networks

Alan Mislove
MPI for Software Systems
Campus E1 4
Saarbrücken 66123, Germany

Massimiliano Marcon
MPI for Software Systems
Campus E1 4
Saarbrücken 66123, Germany

Krishna P. Gummadi
MPI for Software Systems
Campus E1 4
Saarbrücken 66123, Germany

Peter Druschel
MPI for Software Systems
Campus E1 4
Saarbrücken 66123, Germany

Bobby Bhattacharjee
Computer Science Department
University of Maryland
College Park, MD 20742

ABSTRACT

Online social networking sites like Orkut, YouTube, and Flickr are among the most popular sites on the Internet. Users of these sites form a social network, which provides a powerful means of sharing, organizing, and finding content and contacts. The popularity of these sites provides an opportunity to study the characteristics of online social network graphs at large scale. Understanding these graphs is important, both to improve current systems and to design new applications of online social networks.

This paper presents a large-scale measurement study and analysis of the structure of multiple online social networks. We examine data gathered from four popular online social networks: Flickr, YouTube, LiveJournal, and Orkut. We crawled the publicly accessible user links on each site, obtaining a large portion of each social network's graph. Our data set contains over 11.3 million users and 328 million links. We believe that this is the first study to examine multiple online social networks at scale.

Our results confirm the power-law, small-world, and scale-free properties of online social networks. We observe that the indegree of user nodes tends to match the outdegree; that the networks contain a densely connected core of high-degree nodes; and that this core links small groups of strongly clustered, low-degree nodes at the fringes of the network. Finally, we discuss the implications of these structural properties for the design of social network based systems.

Categories and Subject Descriptors

H.5.m [**Information Interfaces and Presentation**]: Miscellaneous; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

General Terms

Measurement

IMC'07, October 24-26, 2007, San Diego, California, USA.

Keywords

Social networks, measurement, analysis

1. INTRODUCTION

The Internet has spawned different types of information sharing systems, including the Web. Recently, *online social networks* have gained significant popularity and are now among the most popular sites on the Web [40]. For example, MySpace (over 190 million users[1]), Orkut (over 62 million), LinkedIn (over 11 million), and LiveJournal (over 5.5 million) are popular sites built on social networks.

Unlike the Web, which is largely organized around content, online social networks are organized around users. Participating users join a network, publish their profile and any content, and create links to any other users with whom they associate. The resulting social network provides a basis for maintaining social relationships, for finding users with similar interests, and for locating content and knowledge that has been contributed or endorsed by other users.

An in-depth understanding of the graph structure of online social networks is necessary to evaluate current systems, to design future online social network based systems, and to understand the impact of online social networks on the Internet. For example, understanding the structure of online social networks might lead to algorithms that can detect trusted or influential users, much like the study of the Web graph led to the discovery of algorithms for finding authoritative sources in the Web [21]. Moreover, recent work has proposed the use of social networks to mitigate email spam [17], to improve Internet search [35], and to defend against Sybil attacks [55]. However, these systems have not yet been evaluated on real social networks at scale, and little is known to date on how to synthesize realistic social network graphs.

In this paper, we present a large-scale (11.3 million users, 328 million links) measurement study and analysis of the structure of four popular online social networks: Flickr, YouTube, LiveJournal, and Orkut. Data gathered from multiple sites enables us to identify common structural properties of online social networks. We believe that ours is the first study to examine multiple online social networks at scale. We obtained our data by crawling publicly accessible information on these sites, and we make the data available

[1]Number of distinct identities as reported by the respective sites in July 2007.

to the research community. In contrast, previous studies have generally relied on propreitary data obtained from the operators of a single large network [4].

In addition to validating the power-law, small-world and scale-free properties previously observed in offline social networks, we provide insights into online social network structures. We observe a high degree of reciprocity in directed user links, leading to a strong correlation between user indegree and outdegree. This differs from content graphs like the graph formed by Web hyperlinks, where the popular pages (*authorities*) and the pages with many references (*hubs*) are distinct. We find that online social networks contain a large, strongly connected core of high-degree nodes, surrounded by many small clusters of low-degree nodes. This suggests that high-degree nodes in the core are critical for the connectivity and the flow of information in these networks.

The focus of our work is the social network users within the sites we study. More specifically, we study the properties of the large weakly connected component[2] (WCC) in the user graphs of four popular sites. We do not attempt to study the entire user community (which would include users who do not use the social networking features), information flow, workload, or evolution of online social networking sites. While these topics are important, they are beyond the scope of this paper.

The rest of this paper is organized as follows: We provide additional background on social networks in Section 2 and detail related work in Section 3. We describe our methodology for crawling these networks, and its limitations, in Section 4. We examine structural properties of the networks in Section 5, and discuss the implications in Section 6. Finally, we conclude in Section 7.

2. BACKGROUND AND MOTIVATION

We begin with a brief overview of online social networks. We then describe a simple experiment we conducted to estimate how often the links between users are used to locate content in a social networking site like Flickr. Finally, we discuss the importance of understanding the structure of online social networks.

Online social networks have existed since the beginning of the Internet. For instance, the graph formed by email users who exchange messages with each other forms an online social network. However, it has been difficult to study this network at large scale due to its distributed nature.

Popular online social networking sites like Flickr, YouTube, Orkut, and LiveJournal rely on an explicit user graph to organize, locate, and share content as well as contacts. In many of these sites, links between users are public and can be crawled automatically to capture and study a large fraction of the connected user graph. These sites present an opportunity to measure and study online social networks at a large scale.

2.1 Online social networking sites

Online social networking sites are usually run by individual corporations (e.g. Google and Yahoo!), and are accessible via the Web.

[2]A weakly connected component in a directed graph is a set of nodes where each node in the set has a path to every other node in the set if all links are viewed as undirected.

Users. To participate fully in an online social network, users must register with a site, possibly under a pseudonym.[3] Some sites allow browsing of public data without explicit sign-up. Users may volunteer information about themselves (e.g., their birthday, place of residence, or interests), which is added to the user's *profile*.

Links. The social network is composed of user accounts and links between users. Some sites (e.g. Flickr, LiveJournal) allow users to link to any other user, without consent from the link target. Other sites (e.g. Orkut, LinkedIn) require consent from both the creator and target before a link is created connecting these users.

Users form links for one of several reasons. The nodes connected by a link can be real-world acquaintances, online acquaintances, or business contacts; they can share an interest; or they can be interested in each other's contributed content. Some users even see the acquisition of many links as a goal in itself [14]. User links in social networks can serve the purpose of both hyperlinks and bookmarks in the Web.

A user's links, along with her profile, are visible to those who visit the user's account. Thus, users are able to explore the social network by following user-to-user links, browsing the profile information and any contributed content of visited users as they go. Certain sites, such as LinkedIn, only allow a user to browse other user accounts within her neighborhood (i.e. a user can only view other users that are within two hops in the social network); other sites, including the ones we study, allow users to view any other user account in the system.

Groups. Most sites enable users to create and join special interest *groups*. Users can post messages to groups and upload shared content to the group. Certain groups are moderated; admission to such a group and postings to a group are controlled by a user designated as the group's moderator. Other groups are unrestricted, allowing any member to join and post messages or content.

2.1.1 Is the social network used in locating content?

Of the four popular social networking sites we study in this paper, only Orkut is a "pure" social networking site, in the sense that the primary purpose of the site is finding and connecting to new users. Others are intended primarily for publishing, organizing, and locating content; Flickr, YouTube, and LiveJournal are used for sharing photographs, videos, and blogs, respectively.

To investigate the role played by the social network in organizing and locating content, we conducted a simple measurement of how users browse the Flickr system. We analyzed the HTTP requests going to the `flickr.com` domain from a 55-day HTTP trace taken at the border routers of the Technical University of Munich between August 17th, 2006 and October 11th, 2006. We found 22,215 photo views from at least 1,056 distinct users. For each of these views, we examined the browser's click stream to determine what action led the user to a given photo.

We found that 17,897 of the views (80.6%) resulted either from following links in the Flickr user graph or were

[3]In the rest of this paper, we use the term "user" to denote a single unique identity in a social network. A person may create multiple identities, and may even create links between these identities. We consider each of these identities as separate users.

additional views within a visited user's collection. In other words, in 80.6% of the views, the user network was involved in browsing content. We count these views as being influenced by the social network. Focusing on the remaining views, 1,418 (6.3%) views were the result of using the Flickr photo search facilities. The remaining 2,900 (13.1%) views were the result of a link from an external source, such as links from an external site or links received via email. Neither of the latter sets of views involved the social network.

Our experiment suggests that the social network in Flickr plays an important role in locating content. Four out of five photos were located by traversing the social network links.

2.2 Why study social networks?

Online social networks are already at the heart of some very popular Web sites. As the technology matures, more applications are likely to emerge. It is also likely that social networking will play an important role in future personal and commercial online interaction, as well as the location and organization of information and knowledge. Examples include browser plug-ins to discover information viewed by friends [39, 50], and social network based, cooperative Web search tools [35]. Even major Web search companies are deploying services that leverage social networks, like Yahoo!'s MyWeb 2.0 [54] and Google Co-op [19].

Below, we outline a few of the ways is which an understanding of the structure of online social networks can benefit the design of new systems and help us understand the impact of online social networks on the future Internet. Additionally, we speculate how our data might be of interest to researchers in other disciplines.

2.2.1 Shared interest and trust

Adjacent users in a social network tend to trust each other. A number of research systems have been proposed to exploit this trust. SybilGuard [55] uses a social network to detect Sybil attacks in distributed systems, leveraging the fact that Sybil users will not be able to create many trust links to non-Sybil users. Re [17] exploits the trust between email users to aid spam classification by whitelisting messages from friends and friends-of-friends. We believe that a deeper understanding of the underlying topology is an essential first step in the design and analysis of robust trust and reputation metrics for these systems.

Adjacent users in a social network also tend to have common interests. Users browse neighboring regions of their social network because they are likely to find content that is of interest to them. Systems such as Yahoo! My Web [54], Google Co-op [19], and PeerSpective [35] use social networks to rank Internet search results relative to the interests of a user's neighborhood in the social network. These systems observe content viewed and search results clicked on by members of a social network in order to better rank the results of the user's future searches.

Understanding the structure of online social networks, as well as the processes that shape them, is important for these applications. It would be useful to have efficient algorithms to infer the actual degree of shared interest between two users, or the reliability of a user (as perceived by other users). With respect to security, it is important to understand the robustness of such networks to deliberate attempts of manipulation. These topics are beyond the scope of this paper; however, a fundamental understanding of online social network structure is likely to be a necessary first step in these directions.

2.2.2 Impact on future Internet

The social networks we study in this paper exist in the databases of online social networking sites. However, other online social networks are implemented as overlay networks. For instance, the graph formed by people who exchange email, or the graph formed by Skype [49] users who include each other in their contact lists can be viewed as another social network on top of the Internet. If future distributed online social networks are popular and bandwidth-intensive, they can have a significant impact on Internet traffic, just as current peer-to-peer content distribution networks do. Understanding the structure of online social networks is not only critical to understanding the robustness and security of distributed online social networks, but also to understanding their impact on the future Internet.

2.2.3 Impact on other disciplines

Additionally, our work has relevance beyond computer science. To social scientists, online social networks offer an unprecedented opportunity to study social networks at a large scale. Sociologists can examine our data to test existing theories about offline social networks, as well as to look for new forms of behavior in online social networks.

Studying the structure of online social networks may help improve the understanding of online campaigning and viral marketing. Political campaigns have realized the importance of blogs in elections [47]. Similarly, marketing experts are experimenting with paid viral marketing [44] to better promote products and companies. Regardless of one's stance on these phenomena, a better understanding of the structure of social networks is likely to improve our understanding of the opportunities, limitations, and threats associated with these ideas.

3. RELATED WORK

In this section we describe studies of social networks, information networks, as well as work on complex network theory.

3.1 Social networks

Sociologists have studied many of the properties of social networks. Milgram [34] shows that the average path length between two Americans is 6 hops, and Pool and Kochen [46] provide an analysis of the small-world effect. The influential paper by Granovetter [20] argues that a social network can be partitioned into 'strong' and 'weak' ties, and that the strong ties are tightly clustered. For an overview of social network analysis techniques, we refer the reader to the book by Wasserman and Faust [51].

As online social networks are gaining popularity, sociologists and computer scientists are beginning to investigate their properties. Adamic et al. [3] study an early online social network at Stanford University, and find that the network exhibits small-world behavior, as well as significant local clustering. Liben-Nowell et al. [32] find a strong correlation between friendship and geographic location in social networks by using data from LiveJournal. Kumar et al. [26] examine two online social networks and find that both possess a large strongly connected component. Girvan and Newman observe that users in online social networks

tend to form tightly knit groups [18]. Backstrom et al. [8] examine snapshots of group membership in LiveJournal, and present models for the growth of user groups over time. We were able to verify these properties on a much larger scale.

In recent work, Ahn et al. [4] analyze complete data from a large South Korean social networking site (Cyworld), along with data from small sample crawls of MySpace and Orkut. The authors obtained data directly from CyWorld operators, and the volume of available data allows the authors to conduct an in-depth study of that site using some of the same metrics that we use in this paper. The comparison with different networks, on the other hand, is limited by the small crawled data samples of MySpace and Orkut. Our study is largely complementary: the data available to us for any one site is less detailed, but we are able to compare large crawled data sets from multiple sites.

3.2 Information networks

A long thread of research examines the structure of complex networks like the Web and the Internet. A prominent study of the Web link structure [12] shows that the Web has a "bow-tie" shape, consisting of a single large strongly connected component[4] (SCC), and other groups of nodes that can either reach the SCC or can be reached from the SCC. We show that online social networks have a similar large component, but that its relative size is much larger than that of the Web's SCC. Faloutsos et al. [16] show that the degree distribution of the Internet follows a power-law, and Siganos et al. demonstrate that the high-level structure of the Internet resembles a "jellyfish" [48].

Kleinberg [24] demonstrates that high-degree pages in the Web can be identified by their function as either hubs (containing useful references on a subject) or authorities (containing relevant information on a subject). Kleinberg also presents an algorithm [21] for inferring which pages function as hubs and which as authorities. The well-known PageRank algorithm [43] uses the Web structure to determine pages that contain authoritative information.

3.3 Complex network theory

There has been much theoretical work on the properties of various classes of complex graphs.

Random networks have been extensively studied, starting with the seminal paper by Erdös and Réyni [15]. These graphs are usually constructed by randomly adding links to a static set of nodes. Researchers have shown that random graphs tend to have very short paths between any two nodes [25]. More recent work on random graphs has provided mechanisms to construct graphs with specified degree distributions [36] and has characterized the size of the large connected component [37].

Power-law networks are networks where the probability that a node has degree k is proportional to $k^{-\gamma}$, for large k and $\gamma > 1$. The parameter γ is called the *power-law coefficient*. Researchers have shown that many real-world networks are power-law networks, including Internet topologies [16], the Web [9, 27], social networks [3], neural networks [11], and power grids [45].

Scale-free networks are a class of power-law networks where the high-degree nodes tend to be connected to other high-degree nodes. Scale-free graphs are discussed in detail by Li et al. [31], and they propose a metric to measure the scale-freeness of graphs. Expectedly, the social networks we study display power-law distributions; by Li's measure, these networks show scale-free properties as well.

Small-world networks have a small diameter and exhibit high clustering. Studies have shown that the Web [5,12], scientific collaboration on research papers [41], film actors [6], and general social networks [3] have small-world properties. Kleinberg [23] proposes a model to explain the small-world phenomenon in offline social networks, and also examines navigability in these networks [22]. The online social networks examined in this paper have small-world properties much like their offline counterparts.

[4]A strongly connected component in a graph is a set of nodes where each node in the set has a path to every other node in the set.

4. MEASUREMENT METHODOLOGY

We now describe the data presented in this paper and the methodology we used to collect it. We were not able to obtain data directly from the respective site operators. Most sites are hesitant to provide even anonymized data, and signing non-disclosure agreements to obtain data from multiple competing sites may not be feasible or desirable. Instead, we chose to crawl the user graphs by accessing the public web interface provided by the sites. This methodology gives us access to large data sets from multiple sites.

Since the focus of this paper is to investigate the structure of online social networks, we focus on the large weakly connected component (WCC) of the corresponding graphs in the rest of this paper. As we show later in this section, the large WCC is structurally the most "interesting" part of the network. The nodes not included in the WCC tend to be either part of very small, isolated clusters or are not connected to other users at all.

4.1 Challenges in crawling large graphs

Crawling large, complex graphs presents unique challenges. In this section, we describe our general approach before discussing the details of how we crawled each network.

4.1.1 Crawling the entire connected component

The primary challenge in crawling large graphs is covering the entire connected component. At each step, one can generally only obtain the set of links into or out of a specified node. In the case of online social networks, crawling the graph efficiently is important since the graphs are large and highly dynamic. Common algorithms for crawling graphs include breadth-first search (BFS) and depth-first search.

Often, crawling an entire connected component is not feasible, and one must resort to using samples of the graph. Crawling only a subset of a graph by ending a BFS early (called the *snowball method*) is known to produce a biased sample of nodes [29]. In particular, partial BFS crawls are likely to overestimate node degree and underestimate the level of symmetry [10]. In social network graphs, collecting samples via the snowball method has been shown to underestimate the power-law coefficient, but to more closely match other metrics, including the overall clustering coefficient [29].

Some previous studies of social networks have used small graph samples. Example studies have used samples of 0.3% of Orkut users [4], less than 1% of LiveJournal communities [8], and 0.08% of MySpace users [4]. In this paper, we obtain and study much larger samples of the user graphs.

	Flickr	LiveJournal	Orkut	YouTube
Number of users	1,846,198	5,284,457	3,072,441	1,157,827
Estimated fraction of user population crawled	26.9%	95.4%	11.3%	unknown
Dates of crawl	Jan 9, 2007	Dec 9 - 11, 2006	Oct 3 - Nov 11, 2006	Jan 15, 2007
Number of friend links	22,613,981	77,402,652	223,534,301	4,945,382
Average number of friends per user	12.24	16.97	106.1	4.29
Fraction of links symmetric	62.0%	73.5%	100.0%	79.1%
Number of user groups	103,648	7,489,073	8,730,859	30,087
Average number of groups memberships per user	4.62	21.25	106.44	0.25

Table 1: High-level statistics of our social networking site crawls.

4.1.2 *Using only forward links*

Crawling directed graphs, as opposed to undirected graphs, presents additional challenges. In particular, many graphs can only be crawled by following links in the forward direction (i.e., one cannot easily determine the set of nodes which point *into* a given node). Using only forward links does not necessarily crawl an entire WCC; instead, it explores the connected component reachable from the set of seed users. This limitation is typical for studies that crawl online networks, including measurement studies of the Web [12].

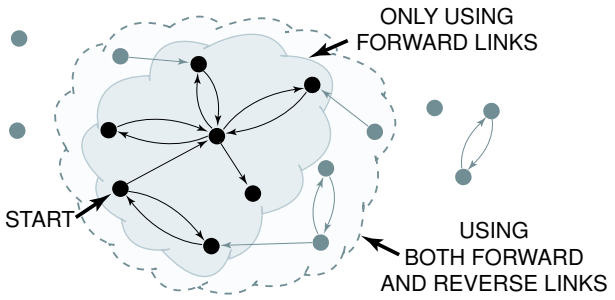


Figure 1: Users reached by crawling different link types. If only forward links are used, we can reach only the inner cloud (shaded cloud); using both forward and reverse links crawls the entire WCC (dashed cloud).

Figure 1 shows an example of a directed graph crawl. The users reached by following only forward links are shown in the shaded cloud, and those reached using both forward and reverse links are shown in the dashed cloud. Using both forward and reverse links allows us to crawl the entire WCC, while using only forward links results in a subset of the WCC.

4.2 Crawling social networks

We now discuss our methodology for crawling each of the networks we crawled, its limitations, and high-level statistics of the resulting data sets. Using automated scripts on a cluster of 58 machines, we crawled the social network graphs of Flickr, LiveJournal, Orkut, and YouTube. High-level statistics of the resulting data sets are presented in Table 1.

We chose these four sites because they are among the most popular social networking sites and they allow us to view the links out of any user in the network. In each step of our crawls, we retrieved the list of friends for a user we had not yet visited and added the retrieved users to the list of users to visit. We continued until we exhausted the list. This corresponds to a BFS of the social network graphs.

4.2.1 *Flickr*

Flickr (`www.flickr.com`) is a photo-sharing site based on a social network. The Flickr data presented in this paper is from a crawl of the large WCC conducted on January 9th, 2007, and contains over 1.8 million users and 22 million links. Flickr exports an API for third-party developers, and we used this API to conduct the crawl. We also obtained group membership information via Flickr's API.[5]

Flickr only allows us to query for forward links. Therefore we were unable to crawl the entire large WCC. To estimate the fraction of users who are part of the WCC but missing in our crawl, we performed the following experiment. We used the fact that the vast majority of Flickr user identifiers take the form of *[randomly selected 8 digit number]@N00*. We generated 100,000 random user identifiers of this form (from a possible pool of 90 million) and found that 6,902 (6.90%) of these were assigned usernames. These 6,902 nodes form a random sample of Flickr users.

Among these 6,902 users, 1,859 users (26.9%) had been discovered during our crawl. Focusing on the 5,043 users *not* previously discovered by our crawl, we conducted a BFS starting at each user to determine whether or not they could reach our set of previously crawled users. We found that only 250 (5.0%) of the missed users could reach our crawled set and were definitively in the WCC. While we cannot conclusively say that the remaining 4,793 (95.0%) missed users are not attached to the WCC (there could be some other user who points to them and to the WCC), the fact that 89.7% of these have no forward links suggests that many are not connected at all.

Finally, to explore how the remaining missing nodes are connected, we crawled the social network using these missing users as seeds, and compared the results with our initial crawl. We found only 11,468 new nodes that were not in the connected component of 1.8 million nodes discovered in the original crawl. Of these new nodes, 5,142 (44.8%) were singleton nodes with no forward links, 3,370 (29.3%) had one link, 620 (5.4%) had two or three links, and 2,336 (20.3%) had four or more links. Thus, the nodes missing from our crawls tend to have low degree and are connected only to small clusters that are not reachable from the large connected component we crawled.

Thus, we believe that our crawl of the large WCC, although not complete, covers a large fraction of the users who are part of the WCC. Further, our experience with the randomly generated Flickr user identifiers indicates that (at least for Flickr), the nodes not in the largest WCC do not form large subgraphs.

[5]Flickr allows users to form private groups, which do not appear in the user's profile list. We were unable to determine any information about the membership of such groups.

4.2.2 LiveJournal

LiveJournal (www.livejournal.com) is a popular blogging site whose users form a social network. The LiveJournal data set considered in this paper is the largest we examine: it contains over 5.2 million users and 72 million links. Due to its size, the LiveJournal crawl took several days, from December 9-11, 2006. LiveJournal offers an API that allows us to query for both forward and reverse links. We followed both link types to crawl the entire large WCC. We also obtained group membership information via LiveJournal's API.[6]

To estimate the fraction of the LiveJournal network covered by our crawl, we used a feature of LiveJournal[7] that returns random users to select a list of 5,000 random LiveJournal users. We then checked how many of these random users our crawl had already covered. We found that we had already crawled 4,773 (95.4%) of these users, showing that our LiveJournal crawl covers the vast majority of the LiveJournal population. Finally, we started another crawl from the previously unknown 227 users to determine how many additional users could be discovered. This technique found only 73 additional users. These results suggest that our LiveJournal crawl covers almost the entire LiveJournal user population, and that the users not included in our crawl are part of small, isolated clusters.

Using the entire WCC from LiveJournal, we calculated the fraction of the WCC that is not reachable by using only forward links (as we did for the Flickr and YouTube crawls). We found that of the 5,284,457 nodes in the discovered weakly connected component, only 404,134 (7.64%) would have been missed had we followed only forward links. Finally, we examined the 404,134 users who would have been missed to see how well these users were connected. We found that 201,694 (49.9%) of these users had a single forward link, 86,561 (21.1%) had two or three links, and 78,463 (19.4%) of the users had four or more forward links. Since, as we will show later, Flickr and YouTube share many characteristics with LiveJournal, this result suggests that the users that are missing in our Flickr and YouTube crawls tend to be small in number and have relatively small outdegree.

4.2.3 Orkut

The next site we examined is Orkut (www.orkut.com), a social networking site run by Google. Orkut is a "pure" social network, as the sole purpose of the site is social networking, and no content is being shared. In Orkut, links are undirected and link creation requires consent from the target. Since, at the time of the crawl, new users had to be invited by an existing user to join the system, the Orkut graph forms a single SCC by definition.

The Orkut data considered in the paper was collected during a crawl performed between October 3rd and November 11th, 2006. Because Orkut does not export an API, we had to resort to HTML screen-scraping to conduct our crawl, which requires more bandwidth. We obtained group information in a similar manner. Furthermore, Orkut limits the rate at which a single IP address can download information and requires a logged-in account to browse the network. As a result, it took more than a month to crawl a subset of 3,072,441 users, at which point we stopped. This subset corresponds to 11.3% of Orkut's user population of about 27 million users at the time of the crawl. The Orkut data considered in this paper, therefore, is limited to this connected component and disregards all links from this component to other, uncrawled users.

Because our Orkut data set contains only a sample of the entire Orkut network, there are two potential concerns with the representativeness of the data. The first question is how the 11.3% subset of the network we gathered would compare to a different 11.3% subset gathered in the same way. In other words, are the properties of our sample representative of other samples of similar size? The second question is how the properties of our sample compare to the properties of the network as a whole.

To explore the first of these concerns, we conducted five separate, small crawls of Orkut starting from random locations. Our random starting locations were chosen using Maximum-Degree random sampling [7] configured with a path length of 100,000 hops. Each of the five crawls was configured to cover 80,000 nodes in the same manner as our single, large crawl. We then examined how similar the properties of the resulting samples were to each other.

We found that the properties of the five smaller crawls were similar, even though these crawls covered only 0.26% of the network. For example, we found that the clustering coefficient of these crawls had an average of 0.284 with a standard deviation of 0.040. Similarly, we found that the scale-free metric had an average of 0.550 with a standard deviation of 0.083 (both of these metrics are discussed in more detail in the following section). Thus, we believe that the properties of our 11.3% sample of the network are likely to be similar to other crawls of similar size that are done in the same manner.

However, we caution the reader to be mindful of the second concern when extrapolating the results from our crawl to the entire Orkut network. Partial BFS crawls are known to over-sample high-degree nodes, and under-sample low-degree nodes [29]. This has been shown to overestimate the average node degree and to underestimate the level of symmetry [10]. Thus, our results may not be representative of the Orkut network as a whole.

4.2.4 YouTube

YouTube (www.youtube.com) is a popular video-sharing site that includes a social network. The YouTube data we present was obtained on January 15th, 2007 and consists of over 1.1 million users and 4.9 million links. Similar to Flickr, YouTube exports an API, and we used this feature to conduct our crawls.

YouTube allows links to be queried only in the forward direction, similar to Flickr. Unfortunately, YouTube's user identifiers do not follow a standard format,[8] and we were therefore unable to create a random sample of YouTube users. Also, YouTube does not export group information via the API. Instead, we obtained group membership information by screen-scraping the HTML pages attached to user profiles.

Because we were unable to crawl reverse links or estimate the size of the user population in YouTube, we advise the reader to be cautious in extrapolating the YouTube results

[6] We inferred groups in LiveJournal by crawling the *interests* specified by users.

[7] http://www.livejournal.com/random.bml

[8] YouTube's user identifiers are user-specified strings.

to the entire YouTube population, as we do not know the number of users who do not participate in the social network.

4.2.5 Summary

Our results indicate that

- The Flickr and YouTube data sets may not contain some of the nodes in the large WCC, but this fraction is likely to be very small.
- The LiveJournal data set covers almost the complete population of LiveJournal, and contains the entire large WCC.
- The Orkut data set represents a modest portion of the network, and is subject to the sampling bias resulting from a partial BFS crawl.

Moreover, the results also indicate that the vast majority of missed nodes in Flickr, LiveJournal, and YouTube have low degree and are likely to be part of small, isolated clusters.

Based on the number of users published by the sites at the time of the crawl, we estimate the fraction of nodes our crawls cover as 1.8 million out of 6.8 million (26.9%) for Flickr, 5.2 million of 5.5 million (95.4%) for LiveJournal, and 3.0 million out of 27 million (11.3%) for Orkut. Unfortunately, we do not know the number of accounts in YouTube. Thus, we were unable to estimate the fraction of the population that our 1.1 million crawled YouTube users represent.

All of the data sets considered in this paper are available to the research community. The data has been anonymized in order to ensure the privacy of the social network users. A detailed description of the data format and downloading instructions are available at
`http://socialnetworks.mpi-sws.mpg.de`

4.3 High-level statistics

Table 1 presents the high-level statistics of the data we gathered. The crawled network sizes vary by almost a factor of five (1.1 million users in YouTube vs. 5.2 million in LiveJournal), and the number of links varies by almost two orders of magnitude (4.9 million in YouTube versus 223 million in Orkut). Similarly, other metrics such as the average number of friend links per node and user participation in shared interest groups also vary by two to three orders of magnitude. Our analysis later will show that despite these differences, these graphs share a surprisingly large number of key structural properties.

4.4 Web graph analysis

The Web is one of the most well-studied online networks, and our study shares much of its methodology with previous studies of the Web. It is natural to compare the structure of online social networks to the structure of the Web. However, we are well aware that the user graph in social networks is fundamentally different from the Web graph; our comparisons serve more to provide a point of reference for our results than to point out (expected) differences.

In order to compare the structure of online social networks with that of the Web, we cite previous studies of the Web structure where possible. We also performed some of our own analysis, using the data collected by the Stanford WebBase Project [1] during their crawl of December 2003. We selected 8.6 million pages and 132 million hyperlinks collected from over 3,900 Web sites contained in the crawl.

5. ANALYSIS OF NETWORK STRUCTURE

In this section, we characterize the structural properties of the four networks we measured. We compare the networks to each other, and we compare their properties with those previously observed for the Web.

5.1 Link symmetry

The fact that links are directed can be useful for locating content in information networks. For example, in the Web graph, search algorithms such as PageRank [43] consider a directed link from a source to a destination as an endorsement of the destination by the source, but not vice-versa. For instance, numerous Web pages point to sites like `cnn.com` or `nytimes.com`, but very few pages receive pointers back from these sites. Search engines leverage this to identify reputed sources of information, since pages with high indegree tend to be authorities [21].

With the exception of Orkut, links in the social networks we studied are directed and users may therefore link to any other user they wish. The target of the link may reciprocate by placing a link pointing back at the source. Our analysis of the level of symmetry in social networks, shown in Table 1, reveals that all three social networks with directed links (Flickr, LiveJournal, and YouTube) have a significant degree of symmetry. Their high level of symmetry is consistent with that of offline social networks [20]. Furthermore, social networking sites inform users of new incoming links, which may also contribute to the high level of symmetry.

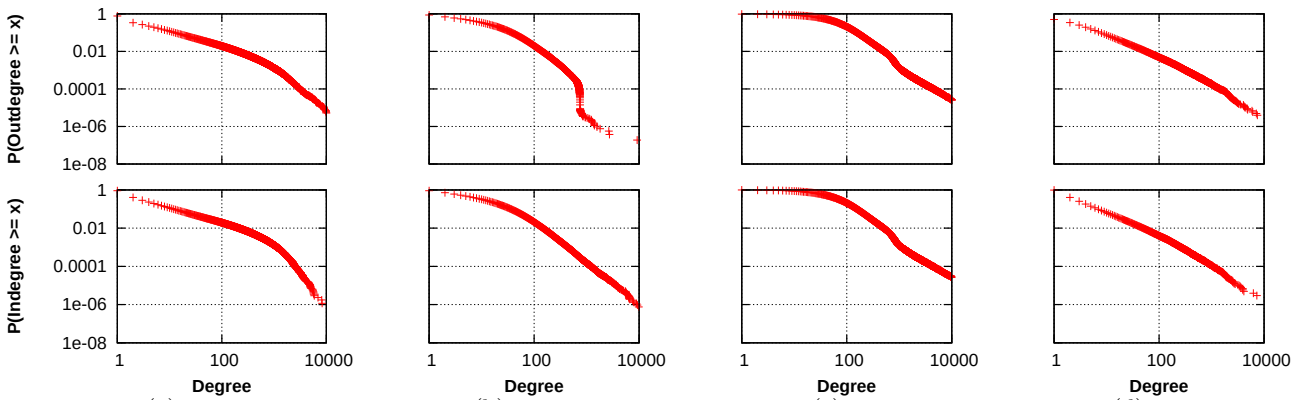
Independent of the causes, the symmetric nature of social links affects the network structure. For example, symmetry increases the overall connectivity of the network and reduces its diameter. Symmetry can also make it harder to identify reputable sources of information just by analyzing the network structure, because reputed sources tend to dilute their importance when pointing back to arbitrary users who link to them.

5.2 Power-law node degrees

We begin to examine the graph structure by considering the node degree distribution. As discussed in Section 3, the degree distributions of many complex networks, including offline social networks, have been shown to conform to power-laws. Thus, it may not be surprising that social networks also exhibit power-law degree distributions. However, as our analysis shows, the degree distributions in social networks differ from that of other power-law networks in several ways.

Figure 2 shows the outdegree and indegree complementary cumulative distribution function (CCDF) for each measured social network. All of the networks show behavior consistent with a power-law network; the majority of the nodes have small degree, and a few nodes have significantly higher degree. To test how well the degree distributions are modeled by a power-law, we calculated the best power-law fit using the maximum likelihood method [13]. Table 2 shows the estimated power-law coefficient along with the Kolmogorov-Smirnov goodness-of-fit metric [13]. While the best power-law coefficients approximate the distributions very well for Flickr, LiveJournal, and YouTube, the Orkut data deviates significantly.

Two factors contribute to this deviation. First, our Orkut



(a) Flickr (b) LiveJournal (c) Orkut (d) YouTube

Figure 2: Log-log plot of outdegree (top) and indegree (bottom) complementary cumulative distribution functions (CCDF). All social networks show properties consistent with power-law networks.

Network	Outdegree α	Outdegree D	Indegree α	Indegree D
Web [12]	2.67	-	2.09	-
Flickr	1.74	0.0575	1.78	0.0278
LiveJournal	1.59	0.0783	1.65	0.1037
Orkut	1.50	0.6319	1.50	0.6203
YouTube	1.63	0.1314	1.99	0.0094

Table 2: Power-law coefficient estimates (α) and corresponding Kolmogorov-Smirnov goodness-of-fit metrics (D). The Flickr, LiveJournal, and YouTube networks are well approximated by a power-law.

crawl reached only 11.3% of the network — partial BFS crawls tend to undersample nodes with lower degree, which can explain the flat head of the distribution [29]. Second, both LiveJournal and Orkut artificially cap a user's number of outgoing links,[9] which leads to a distortion in the distribution for high degrees.

Additionally, we tested the stability of the power-law coefficient estimates by running the maximum likelihood estimator over varyingly sized subsamples of our data [53]. We found that the estimates of the power-law coefficient were remarkably stable; the estimates varied by less than 6% from those provided in Table 2 when we considered as few as 1,000 data points.

Table 2 also shows a difference between the structure of social networks and that of previously observed networks. In the Web, for example, the indegree and outdegree power-law exponents have been shown to differ significantly, while the power-law exponents for the indegree and outdegree distributions in each of our social networks are very similar. This implies that in online social networks, the distribution of outgoing links is similar to that of incoming links, while in the Web, the incoming links are significantly more concentrated on a few high-degree nodes than the outgoing links.

Focusing on this difference, Figure 3 shows the distribution of incoming and outgoing links over nodes in the Web

[9]Orkut caps the outdegree at 1,000, and LiveJournal at 750. Both of these caps were instituted after the networks were established, and some users therefore exceed the caps. Also, Flickr has since instituted a cap of 3,000 *non-reciprocal* links; however, the data shown here was collected before this cap was established.

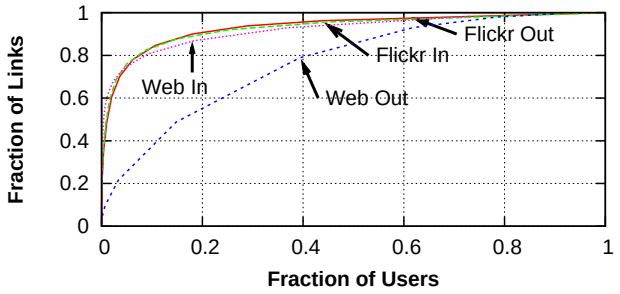


Figure 3: Plot of the distribution of links across nodes. Social networks show similar distributions for outgoing and incoming links, whereas the Web links shows different distributions.

and Flickr graphs.[10] The difference is readily apparent: 5% of the Web nodes account for 75% of all incoming links, but for only 25% of all outgoing links. In all social networks we considered, the distributions of incoming and outgoing links across the nodes are very similar. We now examine this phenomenon in more detail.

5.3 Correlation of indegree and outdegree

Studies of the indegree and outdegree distributions in the Web graph helped researchers find better ways to find relevant information in the Web. In the Web, the population of pages that are *active* (i.e., have high outdegree) is not the same as the population of pages that are *popular* (i.e., have high indegree) [21]. For example, many Web pages of individual users actively point to a few popular pages like `wikipedia.org` or `cnn.com`. Web search techniques are very effective at separating a very small set of popular pages from a much larger set of active pages.

In social networks, the nodes with very high outdegree also tend to have very high indegree. In our study, for each network, the top 1% of nodes ranked by indegree has a more than 65% overlap with the top 1% of nodes ranked by outdegree. The corresponding overlap in the Web is less than 20%. Hence, active users (i.e., those who create many links)

[10]The Flickr topology is representative of all four networks; we omitted the others in the plot for readability.

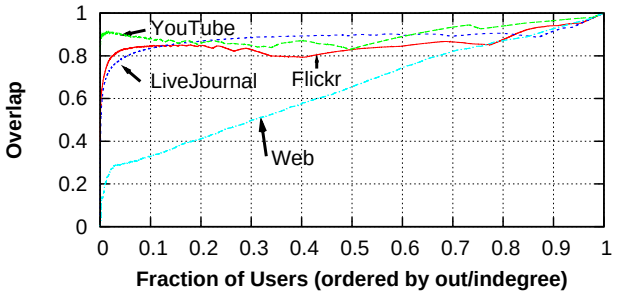


Figure 4: Plot of the overlap between top $x\%$ of nodes ranked by outdegree and indegree. The high-indegree and high-outdegree nodes are often the same in social networks, but not in the Web.

in social networks also tend to be popular (i.e., they are the target of many links). Figure 4 shows the extent of the overlap between the top $x\%$ of nodes ranked by indegree and outdegree.

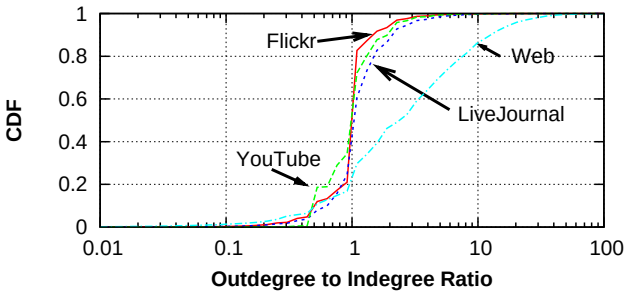


Figure 5: CDF of outdegree to indegree ratio. Social networks show much stronger correlation between indegree and outdegree than the Web.

Next, we compared the indegree and outdegree of individual nodes in the social networks. Figure 5 plots the cumulative distributions of the outdegree-to-indegree ratio for the four social networks and the Web. The social networks show a remarkable correspondence between indegree and outdegree; for all networks, over 50% of nodes have an indegree within 20% of their outdegree. The distribution for the Web is markedly different; most nodes have considerably higher outdegree than indegree, while a small fraction of nodes have significantly higher indegree than outdegree.

The high correlation between indegree and outdegree in social networks can be explained by the high number of symmetric links. The high symmetry may be due to the tendency of users to reciprocate links from other users who point to them. This process would result in active users (who place many outgoing links) automatically receiving many incoming links, and lead to the distributions we have observed.

5.4 Path lengths and diameter

Next, we look at the properties of shortest paths between users. Table 3 shows the average path lengths, diameters, and radii[11] for the four social networks. In absolute terms, the path lengths and diameters for all four social networks

Network	Avg. Path Len.	Radius	Diameter
Web [12]	16.12	475	905
Flickr	5.67	13	27
LiveJournal	5.88	12	20
Orkut	4.25	6	9
YouTube	5.10	13	21

Table 3: Average path length, radius, and diameter of the studied networks. The path length between random nodes is very short in social networks.

are remarkably short. Interestingly, despite being comparable in size to the Web graph we considered, the social networks have significantly shorter average path lengths and diameters. This property may again result from the high degree of reciprocity within the social networks. Incidentally, Broder et al. [12] noted that if the Web were treated as an undirected graph, the average path length would drop from 16.12 to 7.

5.5 Link degree correlations

To further explore the difference in network structure between online social networks and previously observed networks, we examine which users tend to connect to each other. In particular, we focus on the *joint degree distribution* (JDD), or how often nodes of different degrees connect to each other. This property is also referred to as the 2K-distribution [33] or the mixing patterns [42].

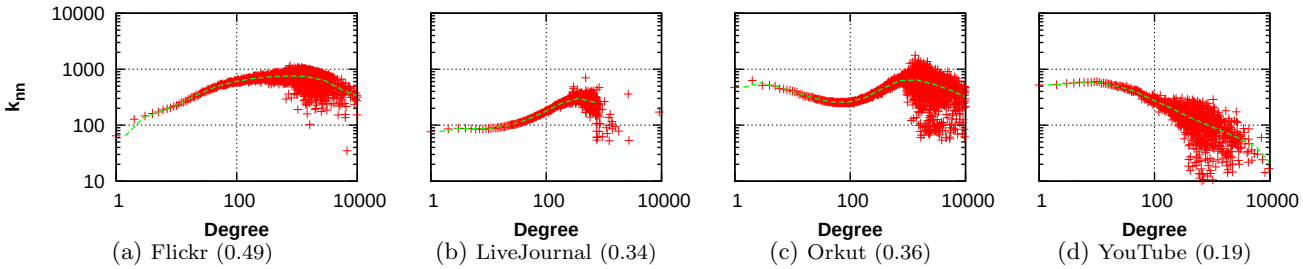
5.5.1 Joint degree distribution

The JDD provides many insights into the structural properties of networks. For example, networks where high-degree nodes tend to connect to other high-degree nodes are more likely to be subject to epidemics, as a single infected high-degree node will quickly infect other high-degree nodes. On the other hand, networks where high-degree nodes tend to connect to low-degree nodes show the opposite behavior; a single infected high-degree node will not spread an epidemic very far.

The JDD is approximated by the degree correlation function k_{nn}, which is a mapping between outdegree and the average indegree of all nodes connected to nodes of that outdegree. Clearly, an increasing k_{nn} indicates a tendency of higher-degree nodes to connect to other high-degree nodes; a decreasing k_{nn} represents the opposite trend. Figure 6 shows a plot of k_{nn} for the four networks we studied.

The trend for high-degree nodes to connect to other high-degree nodes can be observed in all networks except YouTube (the unexpected bump at the head of the Orkut curve is likely due to the undersampling of users). This suggests that the high-degree nodes in social networks tend to connect to other high-degree nodes, forming a "core" of the network. Anecdotally, we believe that the different behavior seen in YouTube is due its more "celebrity"-driven nature; there are a few extremely popular users on YouTube to whom many unpopular users connect.

[11]The *eccentricity* of a node v is the maximal shortest path distance between v and any other node. The radius of a graph is then the minimum eccentricity across all vertices, and the diameter is the maximum eccentricity across all vertices. Due to the computational complexity associated with determining the actual radius and diameter, the numbers presented here are from determining the eccentricity of 10,000 random nodes in each network. Therefore, the diameter should be viewed as a lower bound, and the radius as an upper bound.



(a) Flickr (0.49) (b) LiveJournal (0.34) (c) Orkut (0.36) (d) YouTube (0.19)

Figure 6: Log-log plot of the outdegree versus the average indegree of friends. The scale-free metrics, included in the legend, suggest the presence of a well-connected core.

To quantitatively explore this phenomenon, we next examine two metrics based on the joint degree distribution: the scale-free metric s and the assortativity r.

5.5.2 *Scale-free behavior*

The scale-free metric s [31] is a value calculated directly from the joint degree distribution of a graph. The scale-free metric ranges between 0 and 1, and measures the extent to which the graph has a hub-like core. A high scale-free metric means that high-degree nodes tend to connect to other high-degree nodes, while a low scale-free metric means that high-degree nodes tend to connect to low-degree nodes.

The scale-free metric of the networks are shown in the legend of Figure 6. All of the networks with the exception of YouTube show a significant s, indicating that high-degree nodes tend to connect to other high-degree nodes, and low-degree nodes tend to connect to low-degree nodes.

5.5.3 *Assortativity*

The scale-free metric is related to the assortativity coefficient r, which is a measure of the likelihood for nodes to connect to other nodes with similar degrees. The assortativity coefficient ranges between -1 and 1; a high assortativity coefficient means that nodes tend to connect to nodes of similar degree, while a negative coefficient means that nodes likely connect to nodes with very different degree from their own Recent work has suggested that the scale-free metric is more suitable for comparing the structure of different graphs [30], as it takes into account the possible configurations of networks with properties including connectedness and no self-loops. However, for completeness, we calculated the assortativity coefficients for each of the networks, and found 0.202 for Flickr, 0.179 for LiveJournal, 0.072 for Orkut, and -0.033 for YouTube.

The assortativity shows yet another difference between the social networks and other previously observed power-law networks. For example, the Web and the Internet have both been shown to have negative assortativity coefficients of -0.067 and -0.189, respectively [42]. On the other hand, many scientific coauthorship networks, a different type of social network, have been shown to have positive r [42].

Taken together, the significant scale-free metric and the positive assortativity coefficient suggests that there exists a tightly-connected "core" of the high-degree nodes which connect to each other, with the lower-degree nodes on the fringes of the network. In the next few sections, we explore the properties of these two components of the graph in detail.

5.6 Densely connected core

We loosely define a *core* of a network as any (minimal) set of nodes that satisfies two properties: First, the core must be necessary for the connectivity of the network (i.e., removing the core breaks the remainder of the nodes into many small, disconnected clusters). Second, the core must be strongly connected with a relatively small diameter. Thus, a "core" is a small group of well-connected group of nodes that is necessary to keep the remainder of the network connected.

To more closely explore the core of the network, we use an approximation previously used in Web graph analysis [12]. Specifically, we remove increasing numbers of the highest degree nodes and analyze the connectivity of the remaining graph.[12] We calculate the size of the largest remaining SCC, which is the largest set of users who can mutually reach each other.

As we remove the highest degree nodes, the largest SCC begins to split into smaller-sized SCCs. Figure 7 shows the composition of the splits as we remove between 0.01% and 10% of the highest-degree nodes in Flickr. The corresponding graphs for the other social networks look similar, and we omit them for lack of space. Once we remove 10% of the highest indegree nodes,[13] the largest SCC partitions into millions of very small SCCs consisting of only a handful of nodes.

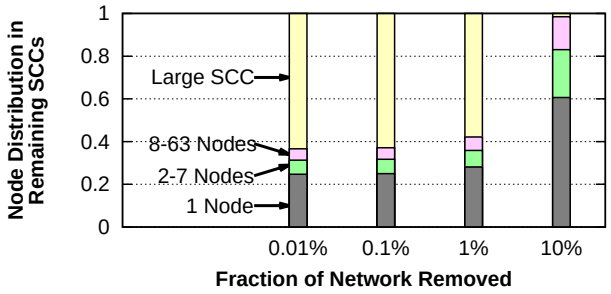


Figure 7: Breakdown of network into SCCs when high-degree nodes are removed, grouped by SCC size.

To understand how much the network core contributes towards the small path lengths, we analyzed the path lengths

[12]The large size of the graphs we study makes a cut set analysis computationally infeasible.

[13]We obtained the same results using both indegree and outdegree.

of subgraphs containing only the highest-degree nodes. Figure 8 shows how path lengths increase as we generate larger subgraphs of the core by progressively including nodes ordered inversely by their degree. The average path length increases sub-logarithmically with the size of the core. In Flickr, for example, the overall average path length is 5.67, of which 3.5 hops involve the 10% of nodes in the core with the highest degrees. This suggests that the high-degree core nodes in these networks are all within roughly four hops of each other, while the rest of the nodes, which constitute the majority of the network, are at most a few hops away from the core nodes.

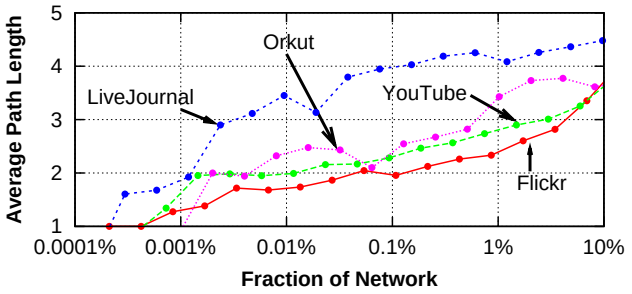


Figure 8: Average path length among the most well-connected nodes. The path length increases sub-logarithmically.

Thus, the graphs we study have a densely connected *core* comprising of between 1% and 10% of the highest degree nodes, such that removing this core completely disconnects the graph.

The structure of social networks, with its high dependence on few highly connected nodes, may have implications for information flow, for trust relationships, and for the vulnerability of these networks to deliberate manipulation. The small diameter and path lengths of social networks are likely to impact the design of techniques for finding paths in such networks, for instance, to check how closely related a given pair of nodes is in the network. Such techniques have applications, for instance, in social networks used to verify the trustworthiness or relevance of received information [17].

5.7 Tightly clustered fringe

Next, we consider the graph properties at the scale of local neighborhoods outside of the core. We first examine clustering, which quantifies how densely the neighborhood of a node is connected.

The *clustering coefficient* of a node with N neighbors is defined as the number of directed links that exist between the node's N neighbors, divided by the number of possible directed links that could exist between the node's neighbors ($N(N-1)$). The clustering coefficient of a graph is the average clustering coefficient of all its nodes, and we denote it as C.

Table 4 shows the clustering coefficients for all four social networks. For comparison, we show the ratio of the observed clustering coefficient to that of Erdös-Réyni (ER) random graphs [15] and random power-law graphs constructed with preferential attachment [9], with the same number of nodes and links. ER graphs have no link bias towards local nodes. Hence, they provide a point of reference for the degree of local clustering in the social networks. Graphs constructed

Network	C	Ratio to Random Graphs	
		Erdös-Rényi	Power-Law
Web [2]	0.081	7.71	-
Flickr	0.313	47,200	25.2
LiveJournal	0.330	119,000	17.8
Orkut	0.171	7,240	5.27
YouTube	0.136	36,900	69.4

Table 4: The observed clustering coefficient, and ratio to random Erdös-Réyni graphs as well as random power-law graphs.

using preferential attachment also have no locality bias, as preferential attachment is a global process, and they provide a point of reference to the clustering in a graph with a similar degree distribution.

The clustering coefficients of social networks are between three and five orders of magnitude larger than their corresponding random graphs, and about one order of magnitude larger than random power-law graphs. This unusually high clustering coefficient suggests the presence of strong local clustering, and has a natural explanation in social networks: people tend to be introduced to other people via mutual friends, increasing the probability that two friends of a single user are also friends.

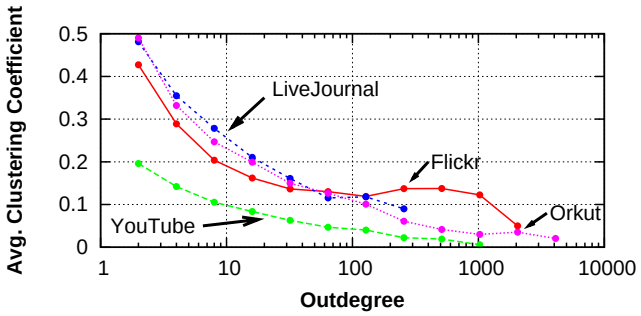


Figure 9: Clustering coefficient of users with different outdegrees. The users with few "friends" are tightly clustered.

Figure 9 shows how the clustering coefficients of nodes vary with node outdegree. The clustering coefficient is higher for nodes of low degree, suggesting that there is significant clustering among low-degree nodes. This clustering and the small diameter of these networks qualifies these graphs as small-world networks [52], and further indicates that the graph has scale-free properties.

5.8 Groups

In many online social networks, users with shared interests may create and join groups. Table 5 shows the high-level statistics of user groups in the four networks we study. Participation in user groups varies significantly across the different networks: only 8% of YouTube users but 61% of LiveJournal users declare group affiliations. Once again, the group sizes follow a power-law distribution, in which the vast majority have only a few users each.

Note that users in a group need not necessarily link to each other in the social network graph. As it turns out, however, user groups represent tightly clustered communities of users in the social network. This can be seen from the average group clustering coefficients of group members,

Network	Groups	Usage	Avg. Size	Avg. C
Flickr	103,648	21%	82	0.47
LiveJournal	7,489,073	61%	15	0.81
Orkut	8,730,859	13%	37	0.52
YouTube	30,087	8%	10	0.34

Table 5: Table of the high-level properties of network groups including the fraction of users which use group features, average group size, and average group clustering coefficient.

shown in Table 5.[14] These coefficients are higher than those of the corresponding network graph as a whole (shown in Table 4). Further, the members of smaller user groups tend to be more clustered than those of larger groups. Figure 10 shows this by plotting the average group clustering coefficient for groups of different sizes in the four observed networks. In fact, many of the small groups in these networks are cliques.

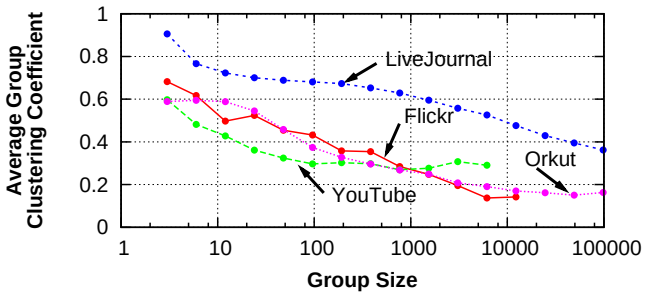


Figure 10: Plot of group size and average group clustering coefficient. Many small groups are almost cliques.

Finally, Figure 11 shows how user participation in groups varies with outdegree. Low-degree nodes tend to be part of very few communities, while high-degree nodes tend to be members of multiple groups. This implies a correlation between the link creation activity and the group participation. There is a sharp decline in group participation for Orkut users with over 500 links, which is inconsistent with the behavior of the other networks. This result may be an artifact of our partial crawl of the Orkut network and the resulting biased user sample.

In general, our observations suggest a global social network structure that is comprised of a large number of small, tightly clustered local user communities held together by nodes of high degree. This structure is likely to significantly impact techniques, algorithms and applications of social networks.

5.9 Summary

We end this section with a brief summary of important structural properties of social networks which we observed in our data.

- The degree distributions in social networks follow a power-law, and the power-law coefficients for both indegree and outdegree are similar. Nodes with high indegree also tend to have high outdegree.
- Social networks appear to be composed of a large number of highly connected clusters consisting of relatively low-degree nodes. These clusters connect to each other via a relatively small number of high-degree nodes. As a consequence, the clustering coefficient is inversely proportional to node degree.
- The networks each contain a large, densely connected core. Overall, the network is held together by about 10% of the nodes with highest degree. As a result, path lengths are short, but almost all shortest paths of sufficient length traverse the highly connected core.

[14]We define the *group clustering coefficient* of a group G as the clustering coefficient of the subgraph of the network consisting of only the users who are members of G.

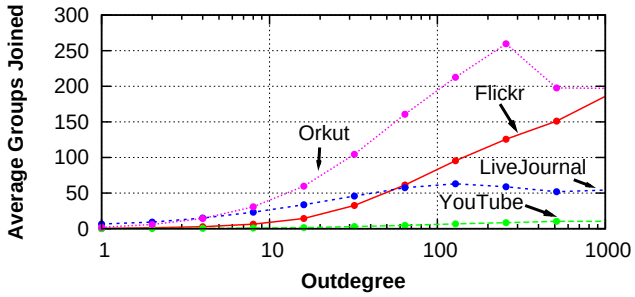


Figure 11: Outdegree versus average number of groups joined by users. Users with more links tend to be members of many groups.

6. DISCUSSION

In this section, we discuss some implications of our findings. Our measurements indicate that online social networks have a high degree of reciprocity, a tight core that consists of high-degree nodes, and a strong positive correlation in link degrees for connected users. What do these findings *mean* for developers? Alternately, how should applications for social networks be designed to take advantage of these properties? Do these properties reveal straightforward attacks on the social structure? Finally, does it make sense to "optimize" algorithms and applications based upon our findings, since these networks are still growing rapidly and any property we assert now may soon change?

While our findings are likely applicable to many different applications, we concentrate on their effect on information dissemination, search, and trust inference.

6.1 Information dissemination and search

Social networks are already used as a means for rapidly disseminating information, as witnessed by the popularity of "hot" videos on YouTube. The existence of a small, well-connected core implies that information seeded via a core node will rapidly spread through the entire network. This is both a strength and a weakness, as spam or viruses could be disseminated this way, as well as important information.

Similarly, searches that proceed along social network links will quickly reach the core. This suggests that simple unstructured search algorithms could be designed if the core users were to store some state about other users. In effect, the users in the core represent "supernodes" in a two-level hierarchy, similar to existing search protocols for unstructured networks, such as Gnutella.

6.2 Trust

Social networking sites are the portals of entry into the Internet for many millions of users, and they are being used both for advertisement as well as for the ensuing commerce. Many of these applications, ranging from mail to auctions, implicitly rely on some form of *trust*. For example, when a user accepts email from an unknown user, she is trusting the other party not to send spam. When a user selects a winning bidder in an auction, she is trusting the other party to pay the winning amount, and the winning user is trusting the seller to produce the auctioned item.

In a social network, the underlying user graph can potentially be used as a means to infer some level of trust in an unknown user [28], to check the validity of a public key certificate [38], and to classify potential spam [17]. In all of these, trust is computed as a function of the path between the source and target user.

Our findings have interesting implications for trust inference algorithms. The tight core coupled with link reciprocity implies that users in the core appear on a large number of short paths. Thus, if malicious users are able to penetrate the core, they can skew many trust paths (or appear highly trustworthy to a large fraction of the network). However, these two properties also lead to small path lengths and many disjoint paths, so the trust inference algorithms should be adjusted to account for this observation. In particular, given our data, an unknown user should be highly trusted only if multiple short disjoint paths to the user can be discovered.

The correlation in link degrees implies that users in the fringe will not be highly trusted unless they form direct links to other users. The "social" aspect of these networks is self-reinforcing: in order to be trusted, one must make many "friends", and create many links that will slowly pull the user into the core.

6.3 Temporal invariance

One possible criticism of our study is the snapshot character of our data, which does not account for change over time. To explore this, we repeated the entire crawl for both Flickr and YouTube on May 7th, 2007, and recomputed the complete statistics on the new data set. Both of the networks showed rapid growth over this five month time period, with Flickr growing by 38% and YouTube by 83%.

However, the salient observations in our original data are still valid; for Flickr, most of the updated results are indistinguishable from the results presented. YouTube showed a difference due to a policy change between our original and new crawls: YouTube switched from directed links to a two-phase symmetric link creation process. Thus, in the new YouTube crawl, we observe a much higher level of symmetry and a correspondingly larger SCC. However, many of the other metrics, such as the assortativity, clustering coefficient, and average path length are similar.

This experiment gives us some assurance that our structural observations are not incidental to the stage of growth at which we sampled the network. Our data indicates that, even though the networks are growing rapidly, their basic structure is not changing drastically.

7. CONCLUSIONS

We have presented an analysis of the structural properties of online social networks using data sets collected from four popular sites. Our data shows that social networks are structurally different from previously studied networks, in particular the Web. Social networks have a much higher fraction of symmetric links and also exhibit much higher levels of local clustering. We have outlined how these properties may affect algorithms and applications designed for social networks.

Much work still remains. We have focused exclusively on the user graph of social networking sites; many of these sites allow users to host content, which in turn can be linked to other users and content. Establishing the structure and dynamics of the content graph is an open problem, the solution to which will enable us to understand how content is introduced in these systems, how data gains popularity, how users interact with popular versus personal data, and so on.

Acknowledgments

We thank the anonymous reviewers, our shepherd Yin Zhang, and Walter Willinger for their helpful comments. We would also like to thank Anja Feldmann and Nils Kammenhuber for their assistance with the TU Munich trace. This research was supported in part by US National Science Foundation grant ANI-0225660.

8. REFERENCES

[1] Stanford WebBase Project. `http://www-diglib.stanford.edu/~testbed/doc2/WebBase`.

[2] L. A. Adamic. The Small World Web. In *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries (ECDL'99)*, Paris, France, Sep 1999.

[3] L. A. Adamic, O. Buyukkokten, and E. Adar. A social network caught in the Web. *First Monday*, 8(6), 2003.

[4] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of Topological Characteristics of Huge Online Social Networking Services. In *Proceedings of the 16th international conference on World Wide Web (WWW'07)*, Banff, Canada, May 2007.

[5] R. Albert, H. Jeong, and A.-L. Bárabási. The Diameter of the World Wide Web. *Nature*, 401:130, 1999.

[6] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences (PNAS)*, 97:11149–11152, 2000.

[7] A. Awan, R. A. Ferreira, S. Jagannathan, and A. Grama. Distributed uniform sampling in real-world networks. Technical Report CSD-TR-04-029, Purdue University, 2004.

[8] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group Formation in Large Social Networks: Membership, Growth, and Evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*, Philadelphia, PA, Aug 2006.

[9] A.-L. Bárabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286:509–512, 1999.

[10] L. Becchetti, C. Castillo, D. Donato, and A. Fazzone. A Comparison of Sampling Techniques for Web Graph Characterization. In *Proceedings of the Workshop on Link Analysis (LinkKDD'06)*, Philadelphia, PA, Aug 2006.

[11] V. Braitenberg and A. Schüz. *Anatomy of a Cortex: Statistics and Geometry*. Springer-Verlag, Berlin, 1991.

[12] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph Structure in the Web: Experiments and

Models. In *Proceedings of the 9th International World Wide Web Conference (WWW'00)*, Amsterdam, May 2000.

[13] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data, Jun 2007. `http://arxiv.org/abs/0706.1062v1`.

[14] d. boyd. Friends, Friendsters, and Top 8: Writing community into being on social network sites. *First Monday*, 11(12), 2006.

[15] P. Erdös and A. Rényi. On Random Graphs I. *Publicationes Mathematicae Debrecen*, 5:290–297, 1959.

[16] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On Power-Law Relationships of the Internet Topology. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM'99)*, Cambridge, MA, Aug 1999.

[17] S. Garriss, M. Kaminsky, M. J. Freedman, B. Karp, D. Mazières, and H. Yu. Re: Reliable Email. In *Proceedings of the 3rd Symposium on Networked Systems Design and Implementation (NSDI'06)*, San Jose, CA, May 2006.

[18] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences (PNAS)*, 99:7821–7826, 2002.

[19] Google Co-op. `http://www.google.com/coop/`.

[20] M. Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78(6), 1973.

[21] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46:604–632, 1999.

[22] J. Kleinberg. Navigation in a Small World. *Nature*, 406:845–845, 2000.

[23] J. Kleinberg. The Small-World Phenomenon: An Algorithmic Perspective. In *Proceedings of the 32nd ACM Symposium on Theory of Computing (STOC'00)*, Portland, OR, May 2000.

[24] J. Kleinberg and S. Lawrence. The Structure of the Web. *Science*, 294:1849–1850, 2001.

[25] J. M. Kleinberg and R. Rubinfeld. Short paths in expander graphs. In *IEEE Symposium on Foundations of Computer Science (FOCS'96)*, Burlington, VT, Oct 1996.

[26] R. Kumar, J. Novak, and A. Tomkins. Structure and Evolution of Online Social Networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*, Philadelphia, PA, Aug 2006.

[27] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for Emerging Cyber-Communities. *Computer Networks*, 31:1481–1493, 1999.

[28] S. Lee, R. Sherwood, and B. Bhattacharjee. Cooperative peer groups in NICE. In *Proceedings of the Conference on Computer Communications (INFOCOM'03)*, San Francisco, CA, Mar 2003.

[29] S. H. Lee, P.-J. Kim, and H. Jeong. Statistical properties of sampled networks. *Physical Review E*, 73, 2006.

[30] L. Li and D. Alderson. Diversity of graphs with highly variable connectivity. *Physics Review E*, 75, 2007.

[31] L. Li, D. Alderson, J. C. Doyle, and W. Willinger. Towards a Theory of Scale-Free Graphs: Definitions, Properties, and Implications. *Internet Mathematics*, 2(4):431–523, 2006.

[32] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic Routing in Social Networks. *Proceedings of the National Academy of Sciences (PNAS)*, 102(33):11623–11628, 2005.

[33] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat. Systematic Topology Analysis and Generation Using Degree Correlations. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM'06)*, Pisa, Italy, August 2006.

[34] S. Milgram. The small world problem. *Psychology Today*, 2(60), 1967.

[35] A. Mislove, K. P. Gummadi, and P. Druschel. Exploiting social networks for Internet search. In *Proceedings of the 5th Workshop on Hot Topics in Networks (HotNets-V)*, Irvine, CA, Nov 2006.

[36] M. Molloy and B. Reed. A critical point for random graphs with a given degree distribution. *Random Structures and Algorithms*, 6, 1995.

[37] M. Molloy and B. Reed. The size of the giant component of a random graph with a given degree sequence. *Combinatorics, Probability and Computing*, 7, 1998.

[38] R. Morselli, B. Bhattacharjee, J. Katz, and M. A. Marsh. Keychains: A Decentralized Public-Key Infrastructure. Technical Report CS-TR-4788, University of Maryland, 2006.

[39] MozillaCoop. `http://www.mozilla.com`.

[40] MySpace is the number one website in the U.S. according to Hitwise. HitWise Press Release, July, 11, 2006. `http://www.hitwise.com/press-center/hitwiseHS2004/social-networking-june-2006.php`.

[41] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences (PNAS)*, 98:409–415, 2001.

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

[43] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford University, 1998.

[44] PayPerPost. `http://www.payperpost.com`.

[45] A. G. Phadke and J. S. Thorp. *Computer relaying for power systems*. John Wiley & Sons, Inc., New York, NY, USA, 1988.

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

[47] D. Rezner. The Power and Politics of Weblogs. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW'04)*, Chicago, IL, Nov 2004.

[48] G. Siganos, S. L. Tauro, and M. Faloutsos. Jellyfish: A Conceptual Model for the AS Internet Topology. *Journal of Communications and Networks*, 8(3):339–350, 2006.

[49] Skype. `http://www.skype.com`.

[50] StumbleUpon. `http://www.stumbleupon.com`.

[51] S. Wasserman and K. Faust. *Social Networks Analysis: Methods and Applications*. Cambridge University Press, Cambridge, UK, 1994.

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

[53] W. Willinger, D. Alderson, and L. Li. A pragmatic approach to dealing with high-variability in network measurements. In *Proceedings of the 2nd ACM/Usenix Internet Measurement Conference (IMC'04)*, Taormina, Italy, Oct 2004.

[54] Yahoo! MyWeb. `http://myweb2.search.yahoo.com`.

[55] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman. SybilGuard: Defending against Sybil attacks via social networks. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM'06)*, Pisa, Italy, August 2006.