Tracing the Birth of an OSN: Social Graph and Profile Analysis in Google+

Doris Schiöberg[*] **Fabian Schneider**[†] **Harald Schiöberg**[*]
Stefan Schmid[*] **Steve Uhlig**[‡] **Anja Feldmann**[*]

[*] TU Berlin & Telekom Innovation Laboratories
{doris,harald,stefan,anja}@net.t-labs.tu-berlin.de
[†] NEC Laboratories Europe fabian@ieee.org
[‡] Queen Mary University of London steve@eecs.qmul.ac.uk

ABSTRACT

This paper leverages the unique opportunity of Google launching the Google+ OSN. Through multiple crawls of the Google+ OSN, before and after the official public release of the network, our results provide insights into the social graph dynamics of the birth of an OSN. Our findings underline the impact of peculiar aspects of Google+ such as (a) Google's large initial user base taken over from other Google products and (b) Google+'s provision for asymmetric friendships, on its graph structure, especially in light of previously studied OSN graphs. In addition, we study the geographic distribution of the users and links of Google+, and correlate the social graph with additional information available from the public profiles.

Author Keywords

Online Social Networks, Measurements, Google+, Dynamic Graphs, Asymmetric and symmetric Links, User Profiles.

ACM Classification Keywords

H.4 Information Systems Applications: Miscellaneous

General Terms

Measurement

INTRODUCTION

Online Social Networks (OSNs) such as Flickr, Twitter, and Facebook have become popular within the recent years. OSNs allow users to form online communities among people with common interests, activities, backgrounds, and/or friendships. A variety of studies of OSNs have focused on understanding the relationships between users by studying the graph properties of the online communities, e. g., [1, 4, 6, 7, 17, 18]. Works like these mostly use crawls of the OSNs social graph or the users' profiles to gain insights about the OSN's user behavior, social dependencies, dynamics, and changes over time. Understanding such properties is not only crucial for understanding how the Internet is changing the society and inter-personal relationships but also for designing the network itself [12]. In an OSN most content is generated by its users. User generated content is a new class of content and its volume, e. g., from YouTube [5], can put a significant burden on the infrastructure which may require specialized content distribution networks, e. g., Akamai [12].

In this paper, we focus on the newest major player in the OSN ecosystem: Google+. Launched end of June 2011, Google+ is Google's newest attempt to establish an OSN. Until late September 2011, Google+ could only be joined by invitation. Since then, it is possible to freely join. We report our observations about its growth from early September until late October 2011. We base our observations on 16 crawls of the Google+ OSN over a two month period. The first data set is from September 2 when Google+ had 19 million users. The latest data set is from October 20. So we have data for a time span of more than a half month before and one month after Google+ was accessible without restrictions. Thus our data provides us with the unique opportunity to witness the initial growth of a new, large OSN.

Google+ differs from other social networks as it is supported by one of the major players in the Internet – Google, which offers much more than just an OSN service. Google builds upon a large user base, e. g., from its mail and document services, for advertising the new network and thus has a huge growth potential. This makes Google+ unique. All other OSNs had to start from scratch with zero users. The social graph of Google+ still shows a lot of similarities with classical OSNs, e.g., Twitter and Flickr. Moreover, Google+ occupies an interesting position between classical friendship networks (à la Facebook) and blogger or fan networks (à la Twitter). In Google+, users can add "friends" to different circles to receive their information (à la Twitter) but also form symmetric relationships (à la Facebook). For an overview of the Google+ features see Section .

Our unique data set traces the birth of a novel and large OSN and the following period of rapid growth. Indeed, during these six weeks, the user base doubled. The repeated crawls allow us to study the multi-million user network over time;

WebSci 2012, June 22–24, 2012, Evanston, Illinois, USA.

this sheds light onto the user dynamics of the OSN. Our crawls cover almost the entire (public) Google+ network by utilizing Google's site-maps – a feature intended for search engines. On October 13, Google+ announced it surpassed 40 million users, and our data from October 15 shows around 38 million users. Thus we are confident that we cover at least 95 % of the Google+ network. Finally, in addition to the network graph our crawls include the public profile information provided by the user which contain information about the users locations, jobs, education, etc. This allows us to combine the social graph with, e. g., user locations and study the resulting geographic social graph. We summarize our findings as follows:

- Looking at the in/out-degree structure of Google+, our analysis reveals that Google+ cannot be classified as particularly asymmetric ("Twitter-like"), but it is also not as symmetric as, e. g., Flickr that displays a higher correlation between their in- and out-degrees. This makes Google+ structurally different from other OSNs.
- The transition period after Google+ became public exhibits a significant growth. At this time there is also a decrease of the median and mean out-degree, due to the presence of a larger number of lonely nodes and weakly connected components in the graph compared to the period before the transition.
- Our analysis of the public user profiles shows that Google+ users span all regions of the world, and that they have a clear bias towards a highly-educated audience, e. g., college students or IT professionals. While more than $50\,\%$ of the users (sharing this info) are within the same timezone, more than $50\,\%$ of the social links are separated by a distance of more than 1000km. It also turns out that directed links typically span geographically larger distances than symmetric links (i. e., users including each other mutually in their profile). Further, we see that asymmetric links tend to have an east-west-direction, e.g., people from Europe tend to follow people in the USA.
- Overall, the probability that a user shares a specific type of information in its profile is low, e. g., 25 % or less. On the other hand, users who share their out-bound neighbors have a high probability of giving away pieces of their profile information. Furthermore, the conditional availability of different pieces of profile information is highly dependent on the nature of the information. For example, when a user shares its school information (or employer resp.), then it tends to share its major (or job description resp.).

The remainder of this paper is organized as follows. Section Section provides the necessary background information on Google+. Section describes how we crawled the network and summarizes the data sets. While Section presents our analysis of the social graph, Section shows insights obtained for publicly available profile data. We review related literature in Section . Section concludes the paper.

BACKGROUND

In this section we introduce the Google+ platform with its features and define some terminology used in this paper.

Background on Google+

Google launched Google+ on June 28, 2011 as a so called beta-test. An invitation was necessary to join. This mode of operation continued until September 20th, 2011 when Google+ became public.

Google+ integrates other Google services such as *Google Profiles*, *YouTube*, *Picasa*, *Google Talk*, and *Google Mail*. When a user joins Google+, an existing profile from *Google Profiles* is used as the starting profile. (A profile is a collection of personal data.) Moreover, we observed that a user who had a *Google Mail* account for long enough, will get his Google+ profile pre-filled with various information extracted from *Google Mail*. Upon sign-up the user is asked whether she wants to keep the pre-filled information. Further information is used, e.g., all *GTalk* contacts show up in Google+ and if a Google+-user in ones circle has a *GTalk*-account, that account is automatically added to ones *GTalk*-contacts list. Today, the Google+ account is automatically created when signing up for any Google service, e.g., GMail.[1]

A profile can contain any personal and/or professional information, such as employment, education, relationship status, or gender. If a user fills in the field *Places lived*, Google adds a marker on a small embedded Google map. The URL of that map contains GPS coordinates of the places entered by the user.

Users are encouraged to form social relationships through the means of *circles*. A circle is a named set of other Google+ users. A user can have multiple circles, e. g., there can be a circle for close friends, one for colleagues from work, and another for popular bloggers. In graph-theoretical terms, the relationship of a user x having a user y in one of its circles can be represented as a directed social edge (x, y); if user y also includes user x in one of her circles, the relationship $\{x, y\}$ is called symmetric. We will later use the fact that the Google+ graph is directed to better understand how its structure compares to other popular OSNs. Internally each user is identified by a 21-digit number, which seems to be randomly generated. We call this number user ID or UID.

The paramount method of communication in Google+ is *sharing a post*. Users can write a piece of text and share it. They can also share photos, links, videos, etc.

Each piece of information, including posts, can have different *privacy settings* that define the visibility of this information. The different options are: world, friends of friends, all circles, certain named circles, a set of individual users, or private. Note that only public (i. e., "world" readable) information can be obtained through our unauthenticated crawls.

In Google+ the privacy settings for most of the profile elements is set to *public* by default. Examples include a users circles (i. e., the social graph of a user), employment, education, and places lived (i. e., locations). Exceptions include information on "relationship" and "looking for" which are

[1]For further info on how much *Google* combines data from all their services we want to refer to their recent privacy statement.

by default visible to *extended circles*, i. e., circled users and users in the circled users circles (in Facebook this would be friends of friends). Another set is available only to directly circled users, e. g., users phone number and possibility to receive personal messages. Note that all information that we crawl and analyze is public by default. However, Google+ is a system that is under permanent construction, so default settings can change over time.

Terminology

In this paper we use the following terminology:

Social Graph: The *social graph* consists of the set of users (nodes) in Google+ and their relationships (directed edges) to other users expressed through the circles.

Node: Each user in Google+ is a social graph *node*/vertices.

Edge: A (directed) *edge* $A \rightarrow B$ represents the fact that user A included user B in one of his circles. In case user B also included user A in his circles the graph contains another directed edge $B \rightarrow A$.

Links (asymmetric/symmetric): The social relation between two nodes in Google+ is called a *link*. Links can either be *asymmetric* and consist of one directed edge or they can be *symmetric*—in case they mutually circle each other—and consists of two directed edges forming a zweieck[2].

Out-going, in-coming: The *out-going* edges of a node are those directed edges which start at this node, pointing to the members of this user's circles. The *in-coming* edges of a node are the edges that end at that node, that is somebody else has "circled" the user.

Out-degree: The out-degree is the number of (directed) edges that start at a certain node.

In-degree: The in-degree is the number of (directed) edges that end at a certain node.

Neighbor: Two nodes are neighbors if they are connected by an edge, no matter which direction that edge has.

Profile: A profile is the set of personal data a user reveals about herself. It contains the total number of in- and outgoing edges, the place(s) the user lives, the employer, etc.

CRAWLING GOOGLE+

To crawl Google+, we take advantage of the set of publicly accessible static site-map files hosted by Google. These site-map files contain a large portion of the UIDs, and are up to date on a timescale of a few days. From a UID, one can easily construct publicly accessible URLs, from which the user's profile data as well as his friends can be downloaded in a JSON-like format. These JSON files are designed for use by the AJAX framework of the website, and hence are always up to date.

Crawling Methodology

On all Google domains the `robots.txt` file points to a standard site-map file—`profiles-sitemap.xml`. This XML file in turn points to a large set of `sitemap-0*.txt` files, each containing the URLs of 5,000 user profiles. The UID is part of this URL. By observing the timestamps in the XML file, we found that the whole set of site-map files is updated irregularly, but roughly every other day. We notice that the content of the site-map files changes in the sense that some UIDs can disappear. Following up on these disappearing UIDs, we find that most of these accounts are not deactivated. Indeed, while some UIDs can disappear in the site-map files, we still find them by crawling the other users' site-maps as explained below.

The first step of each crawl is the download of the *initial UIDs* present in the site-map files. In the second step, we download for each user the JSON objects describing the users "has in circles" (outbound social graph edges), "is in circles" (inbound edges) and the user's profile data which includes, e. g., the users location information. Google limits the number of download-able edges to $10,000$ for either direction. This is not an issue for the outgoing edges, as Google+ allows each user to "circle" only $5,000$ other users. To clarify, every user can, technically, have only $5,000$ outgoing edges, while she can have as many in-coming edges as there are users in Google+. For the in-coming edges we are simply limited to "see" only $10,000$ when crawling. The edges we might miss when crawling can be inferred from the outgoing edges of other users. This second phase yields the public outbound edges for 76 %[3] and inbound edges for 52 % of the users (see Section). The reason we see only a certain percentage of a user's in-coming or outgoing edges is that users can hide this information. This *first iteration* is based on the *initial UIDs* from the slightly out-of-date site-map files, and hence it is likely that some of the edges refer to unknown UIDs. For example, the crawl from October 20 had $4,010,931$ missing UIDs after the first iteration. We then include these newly found UIDs into our set, and re-iterate until no further UIDs are found. This led to $283,839$ missing UIDs in the second iteration, $16,666$ in the third, $1,350$ in the fourth, 110 in the fifth, 14 in the sixth, and finally 1 missing UID in the seventh iteration.

We are aware that Google provides the Google+ API, which offers a more direct way for obtaining specific profile information. We chose to not use it for two reasons. First, the Google+ API was only released on September 16 while our first crawls were performed as early as September 2. Second, in order to use the API, a key is required and, depending on the application, a limit of $50,000$ or $100,000$ queries per day is enforced. With our approach, we query more than 35 million profiles in a single crawl. For crawling Google+, we dedicated a 16-core AMD Opteron 6168 with 64GB of RAM and 6 disks of raid-0 storage. Our current crawler is written in python. For performance reasons, we run 400 crawl processes in parallel.

Data Sets

We base our analysis on our own Google+ crawls and public data from Twitter and Flickr.

Google+ Data We obtain the Google+ social graph in two data sets: the lists of in-coming edges and the lists of outgoing edges, for each node. If a user marks the visibility of these items as non-public, the list of in-coming or out-going

[2] Zweieck: Pair of diedges (u,v) and (v,u). See Wikipedia on "Glossary of graph theory": `http://bit.ly/JSRJsl`

[3] All numbers in this paragraph are taken from the Oct. 20th crawl.

neighbors is empty. We merge these two lists to form the adjacency list of each node. In this process, we repeatedly find edges that are listed in one set but not in the other set. This is either due to visibility restrictions or to the 10,000 edge limit mentioned above. We add the edge to the adjacency list if we find it in at least one set. Missing UIDs discovered during this process are then crawled in the next iteration. Only the October crawls used this multi-iteration approach. Hence, the September data contains some unresolved UIDs and may miss some profile data.

In this paper, we use the data set from October 20 unless otherwise mentioned, e. g., when discussing general observations. For our analysis of how Google+ evolved over time, we use the following data sets: September 2 (only the list of UIDs available), 7, 12, 16, 19, 20, 23, 25, 26, 28, 29, 30; October 4, 6, 10, and 20.

In addition to the social network edges, we also download the publicly visible *profile* of each identified UID. To preserve storage space, we parse the profile online and store only the interesting parts of the profile information on disk. In particular, we extract the set of locations (up to seven) the user has been living in, the college and the major, the work place (company) and job title as well as the longitude and latitude of the locations. This gives us a complete profile dump that corresponds to the October 20 trace, which we use later in the paper. Note, that we obtained only data that is publicly available. While crawling the data we never use any login-method that could reveal more than is configured as publicly visible.

Flickr and Twitter In addition to our own crawls, we obtained two social graph data sets for Twitter and Flickr. Meeyoung Cha kindly provided access to the same data that was used in [4] and [6]. Please refer to those works for further details about how the data was gathered. Note that these data sets have been collected long after the public release of the respective OSN. Also the data sets have been collected few years ago. They represent a stable stage of the named OSN. Despite these differences to our data we still believe them to be a valuable comparison point, esp. in face of the lack of any other comparable data sets. As we will show in the following sections, Google+ seems to be quite similar to them in many properties.

Crawling Restrictions and Timeline

We were able to crawl almost complete data sets from early September until end of October. In November we discovered that Google deployed mechanisms to prevent our crawling. It is possible that we triggered request limits given that we crawled from one single public IP address. During the writing of the paper we are in the process of releasing a new version of our crawler, and plan to present results and the methodology of the new crawls in a follow-up paper.

GOOGLE+ OSN GRAPH

This section reports on our main results from crawling Google+. In this paper, we go beyond previous work, as we are able to observe the dynamics of an OSN graph during the transition of the OSN from beta test status to fully operational. In the following, we focus on the topological aspects of the network. For the discussion of profiles data see Section .

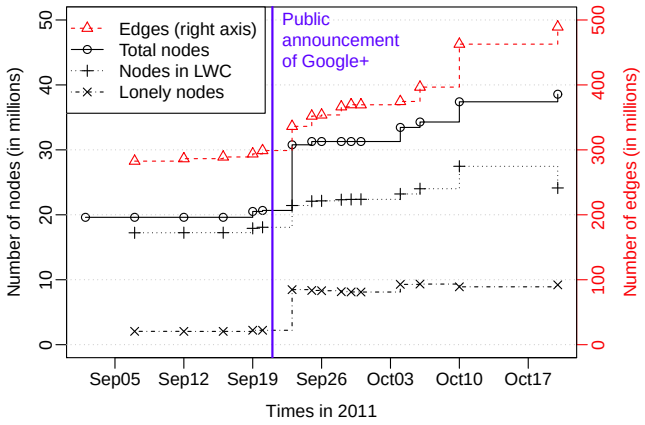

Figure 1. Growth of Google+: number of users over time.

Growth of the Network

The natural first step to understanding the evolution of the OSN is to examine how the number of nodes and directed edges in the OSN evolved across the crawls. Accordingly, Figure 1, plots for all crawls the number of nodes (left axis) and edges (right axis) across time. In addition, we added a vertical line for the public announcement of Google+ on September 20th. Until this date, we do not observe rapid growth–neither in the number of nodes nor edges. We speculate that the service was not that "hot" anymore as reflected by comments in discussions and by people in our circles announcing that they go back to Facebook. After the public announcement of Google+, we see a period of rapid growth until October 10th. After October 10th, the number of newcomers has been limited, indicating that the "hype" may have slowed down, again.

To understand the relationships between users, we identify weak components in the OSN graph. A weak component is a maximal subgraph which would be connected if one ignores the directionality of the edges. Figure 1 also plots the number of nodes in the largest weak component (LWC) and the total number of nodes that are isolated (lonely nodes), i. e., have no link to any other node. We see that until the public announcement, the number of lonely nodes is relatively small and that more than 87 % of the nodes are in the LWC. The number of nodes in local islands, e. g., smaller clusters of nodes, is relatively small. At the same time, there are many weak components, implying that most of these components are small. After the public announcement, the number of lonely nodes increases substantially, i. e., roughly 4-fold, while the size of the LWC increases continuously with only one smaller jump. This is also reflected in the jump in the total number of edges in the OSN graph.

Node Degree Distribution

One of the most popular ways of characterizing OSN graphs is the node degree distribution. In this context, the out-

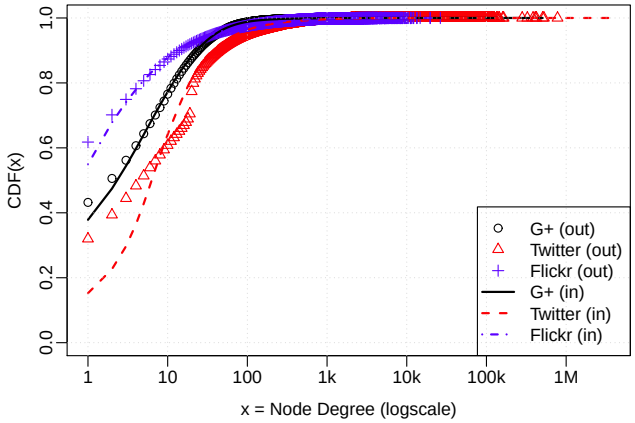

Figure 2. CDF (focus on body) of degree distributions of Google+, Twitter and Flickr.

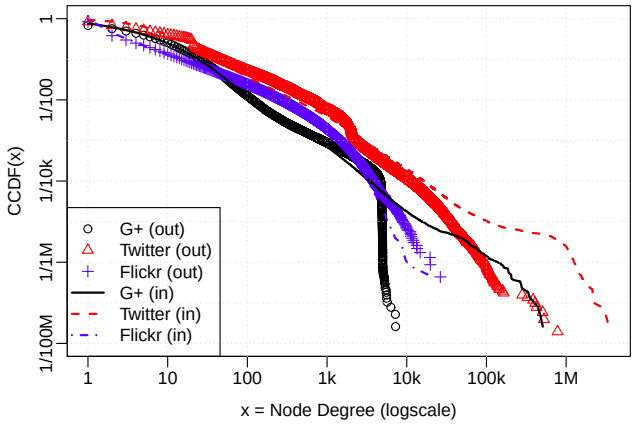

Figure 3. CCDF (focus on tail) of degree distributions of Google+, Twitter and Flickr.

degree of node/user x denotes the number of other users a given user x is connected to (in Google+ terminology: the total number of users across all of the user's circles). Accordingly, the in-degree corresponds to how many users "follow" user x, i. e., have x in their circles. Put differently, a user with a large out-degree is interest*ed* in many other users, and a user with a large in-degree is interest*ing* for many users. Since a Google+ user x can connect to another user y without user y connecting to x, the Google+ graph is *asymmetric*. The resulting graph may therefore significantly differ from symmetric OSN graphs, e. g., from Facebook.

Figure 2 shows the cumulative distribution function (CDF) and Figure 3 shows the complementary cumulative distribution function (CCDF) of both the out-degree (black circles) as well as the in-degree of Google+ users (black lines) for the 20th of October 2011. As can be expected from past OSN analyzes, most nodes have a relatively small out- and in-degree. From the CDF, we see that the in-degrees are only slightly lower than the out-degrees for small node degrees. However, there is a limit to the out-degrees as is apparent from the CCDF. The fact that Google limits the number of outgoing edges to $5,000$ explains the apparent drop-off in the CCDF for the out-degree. Note that our current crawling methodology is able to identify the first $10,000$ incoming edges for each Google+ user. However, less than $1,000$ nodes have an in-degree larger than $10,000$, and reversing the edges allows us to find most of the remaining ones. Thus, while we may not have captured all incoming edges for users with more than $10,000$ followers, we see them in the plot. In principle, the tail of the in-degree distribution is consistent with a heavy-tailed distribution, as expected. The tail of the out-degree distribution is consistent to a heavy-tailed distribution, with a cutoff.

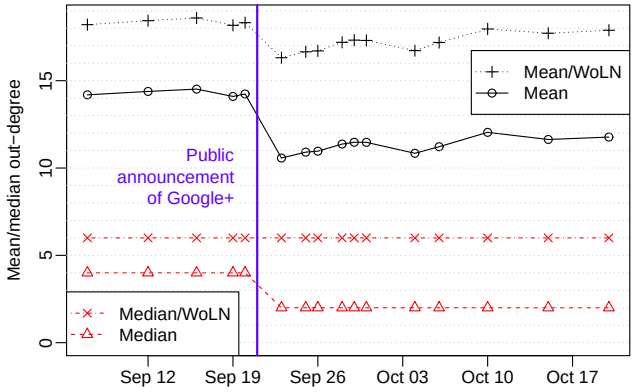

Figure 4. Evolution of the mean and median out-degree across the crawls. (WoLN denotes mean/median without considering lonely nodes.)

In addition, we added the out-degree and in-degree distributions for Flickr [6] (plusses) and Twitter [4](triangles), two popular OSNs with asymmetric relationships. We obtained these data sets from the authors of [4, 6]. From their CDF, we see that the Google+ falls in-between Twitter and Flickr. The average degrees of Google+ are higher than Twitter but lower than Flickr. The same observation holds for the tails (see CCDF). Indeed, Twitter shows a similar difference between its out-degree and in-degree distribution, despite the lack of enforced limit in the degree in Twitter. Flickr does not show such effects. Thus, in the body of the distribution Google+ seems to be closer to Flickr and in the tail to Twitter.

Figure 4 shows how the median and mean out-degree evolved during the observation period. The lines marked with WoLN show the mean/median without considering lonely nodes. First of all, the median is significantly smaller than the mean, which is consistent with the skewed heavy-tailed distribution. We find that the average out-degree decreased significantly after the public announcement of Google+ and is then only slowly increasing, again. This holds for both cases, considering lonely nodes or not. Yet, when leaving out lonely nodes (WoLN) the median does not change after the public announcement. This is consistent with our earlier observation of a limited increase in the size of the largest weak component, but a huge increase in the number of lonely nodes. For Google+ we observe an average out-degree between *10* to *15*. Mislove et al. [17] report similar node-degrees for Flickr and LiveJournal. They also studied

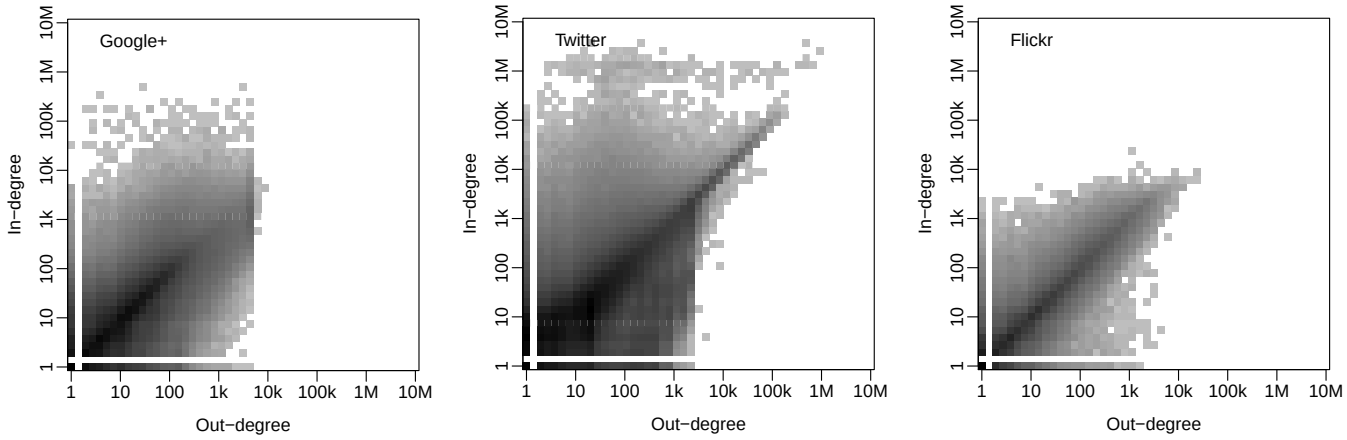

Figure 5. Correlation of in- and out-degree of Google+, Twitter and Flickr (sampled for Google+ and Twitter). The intensity shows the logarithmic frequency of a given combination of in- and out-degree (darker is more frequent).

Table 1. In/Out-degree correlation for OSNs.

Graph	Nodes (in Mio.)	Degree Correlation
Google+	38.5	0.11606
Twitter	51.2	0.24532
Flickr	2.3	0.75584

Orkut, an earlier OSN from Google, which is closer to Facebook and enforces symmetric friendships. For Orcut they observed an average node-degree of *106*. One more indication that Google+ is not just another Facebook.

Degree Correlation

We re-examine the in-degree to out-degree relationship, to better understand the symmetry in the Google+ network, compared to other OSNs. As mentioned earlier, there is a major difference in the tail of the in-degree and out-degree distributions for Google+, but only a limited difference in the body. Accordingly, we investigate the correlation between the number of in-coming edges to the number of out-going edges, i. e., is a user interested in many users also interesting for other users?

Table 1 shows the overall correlation for out-degree and in-degree for Google+, Twitter, and Flickr. Interestingly, the correlation is significantly higher for Flickr than for Twitter and Google+. Still the correlation for Twitter is higher than for Google+.

To further investigate the degree asymmetry, in Figure 5, we plot a heat map of the two-dimensional histogram of in-degree vs. out-degree. A darker gray shade in the graph corresponds to a larger fraction of nodes with this in-degree/out-degree combination. Google+ users can only circle at most 5,000 users, hence the out-degree is limited by the system to 5,000. However, we observe that some users have a slightly higher out-degree. It is unclear why Google's technical restrictions do not apply to them. We notice that most of the darker color is close to the diagonal—indicating symmetry. However, this symmetry is much more pronounced for Flickr than the other two OSNs. Twitter exhibits several outliers with very large in-degree but relatively small out-degree. Google+ shows the same effect, even though in a less extreme manner. Most nodes fall within the first quadrant of the graphs, indicating that some users have larger in-degree than out-degree and vice versa. The magnitude of this phenomenon is more pronounced for Google+ and Twitter than for Flickr, explaining the lower overall correlation. However, there is more data on the diagonal for Twitter than Google+, again explaining the difference in the overall correlation.

Additionally, manual inspection of the Google+ data gives us the impression that VIPs with a huge set of followers tend to reveal their in-degree more compared to their out-degree. On the contrary, privacy-aware users seem to first publish their out-degree if they reveal anything at all. We conjecture that people who have high public visibility tend to hide their private life (who they like) but want to expose how many followers they have.[4]

ANALYSIS OF PROFILE INFORMATION

As quoted from `https://profiles.google.com`, Google profiles provides the following service to its users: "Decide what the world sees when it searches for you. Display the information you care about and make it easy for visitors to get to know you." One of the main features of Google+ is the ability for users to control which other users can access which part of their profile, including the ability to use circles. While in principle Google profiles are an independent service from Google+, Google+ imports the information from the Google profile service if it exists. When not changed or deleted later it remains in the Google+ profile. Note that Google ensures that private information cannot be seen from outside. In this section we examine how many users have publicly available profiles and which parts of the information is public. Moreover, we analyze the information that is exposed by Google+ users. We especially focus

[4]To learn who are the top Google+ ers, we used `http://socialstatistics.com`, and verified whether the numbers listed there correspond to the ones we collected.

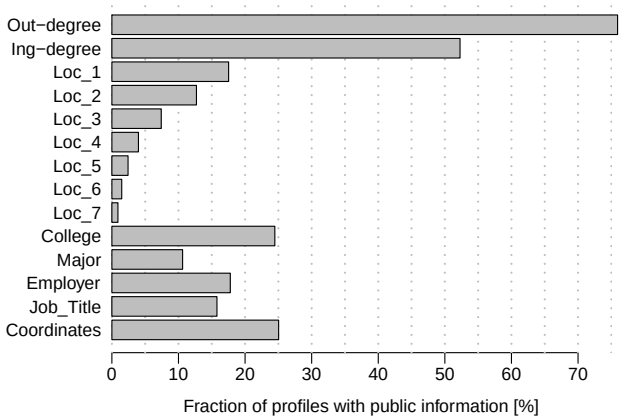

Figure 6. Bar plot showing the fraction of profiles which provide a certain information (y-axis) publicly. The remainder of profiles either did not enter the information or decided to keep it private.

on the (geo-)location information and correlate that with the social graph. All results reported in this section are derived from the crawl on October 20.

Publicly Available Profile Information

Figure 6 shows a bar plot of the fraction of profiles which allow access to different types of profile information. We consider (from top to bottom in the plot): The list of friends in circles, and thus the *out-degree*; The list of friends who have the observed node in their circles and the *in-degree*; The "current" location (*Loc_1*) which corresponds to the location marked blue (though selection by the user) on the user's location map; Additional locations (*Loc_2–Loc_7*), which can for example include the place of birth, the college town, or a previous home location; The school or *College* of the user; And the *Major* or main subject; The company or *employer*; And the occupation or *job title*; As *Coordinates*, we consider if at least one location could be extracted.

While 75 %/50 % of the profiles share out-degree/in-degree, other profile information is publicly shared by no more than 25 % of the Google+ users. The same number shares at least one pair of coordinates. This indicates that users consider this information as more sensitive as other pieces, or never bothered to enter this information at all. For a large fraction of users with coordinates, we find multiple locations. Indeed, some users reveal up to seven possible locations and a large fraction of them indicate their current location. We find that roughly the same fraction of profiles contain sharable coordinates as college information. On the other hand, fewer users also supply their employer, job title, and/or major. From the fraction of users who declare college information, we can infer a bias of Google+ users towards a highly educated sample of the world population. Furthermore, the job descriptions and employers provided indicate that Google+ users are particularly biased towards an IT-educated audience, such as engineers and programmers. This might be related to the fact that Google+ was limited to an invited audience in the beginning. This audience consisted mainly of IT-related people.

Table 2. Availability of B under the condition of A about sharing profile information. Unconditional availability is in the diagonal. All numbers in % of all profiles.

↓B /A→	Outdeg	Indeg	School	Major	Empl.	Job	Loc
Out	**75.9**	68.8	27.4	13.0	20.0	17.8	26.9
In	100.0	**52.28**	22.1	14.8	17.1	15.4	18.4
School	85.0	47.3	**24.47**	43.1	53.2	47.3	65.9
Major	92.8	73.0	99.2	**10.62**	53.6	49.0	48.5
Empl.	85.5	50.1	73.2	32.0	**17.78**	80.5	67.0
Job	85.6	51.1	73.3	33.0	90.6	**15.78**	69.6
Loc	81.6	38.4	64.4	20.6	47.5	43.8	**25.04**

From these numbers alone, we do not see how users share different parts of their profile information, e. g., his job title as well as his employer. Therefore, we next study the conditional probabilities of sharing information. Table 2 summarizes the conditional probability of making a particular information public. We find that if a user reveals some data at all it is more likely that this user also reveals more information, e. g., if someone makes her job description public there is a 90 % chance that she also gives the name of the employer, where giving the name of the employer gives only a 80 % chance that the job description is given. We can see that the more general an information is, the more people are willing to give it. The more detailed it gets, the less likely people are to give it away, e. g., if the school is given not even half of the users want to tell their major.

Publicly Available User Locations

We begin by examining the user coordinates (i. e., Loc_1 or if that is not available Loc_2). Figure 7 show a world map of the user locations. We observe that Google+ users widely sample locations around the world. Note that in Figure 7 we only plotted the locations of users. We did not add borders or coastal lines. Yet, they show up anyway through the distribuion of users across the world. From the density of Google+ users on the map, we see that most users are in the US or in Europe. Within Asia, most users are either in India, Japan, or the other IT savvy regions. There are not as many users in Africa and Australia as in South America. Overall, we see that the map reflects the population density as well as the IT activity in different parts of world. [16] did a study on Twitter user locations and interactions across the USA. Since we look at all users and do not zoom that far into the USA, the studies are not directly comparable. Yet, we want to point out one finding that holds for us as well as for all such investigations: "users may lie about their location, or may list an out-of-datelocation". Profile data of OSN users is always limited in terms of validity and completness.

While the exact location of users is relevant provisioning of OSN infrastructure, when modeling communication and session patterns of OSN users it is more important to know how users distribute across timezones. Accordingly, we compute the time zone of each user and plot their distribution in Figure 8. We confirm our earlier findings that most users are either located in the US or in Europe, with the US east coast and the central European time zones dominating. This matches the expectations about access to the Internet and the

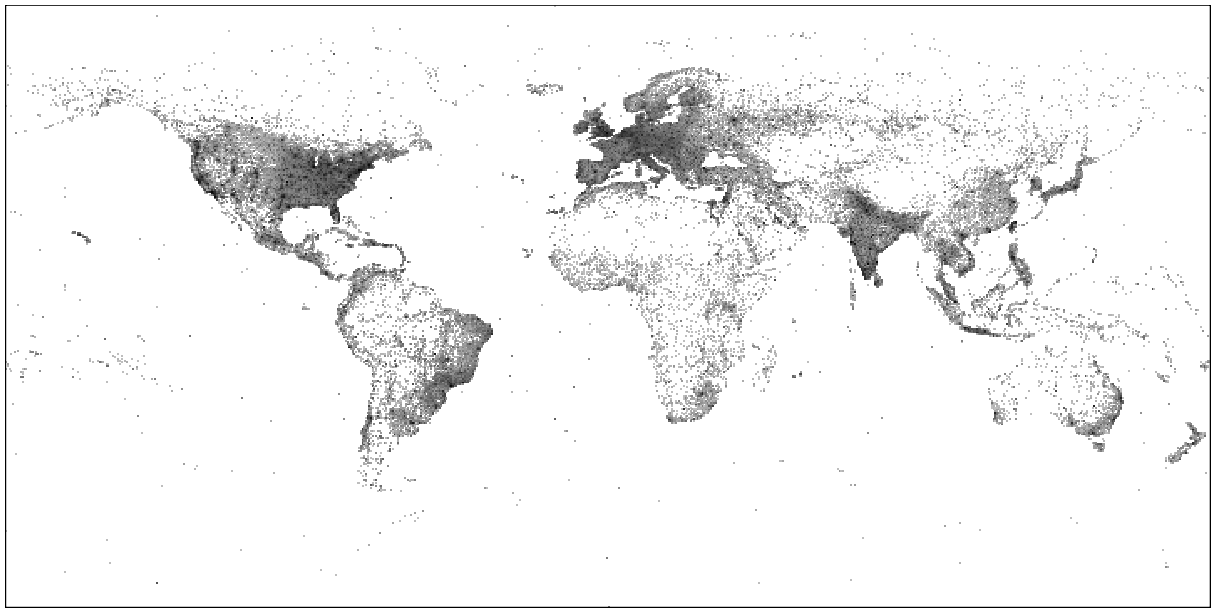

Figure 7. Locations of Google+ users. Each dot represents an area of 0.5 square degree latitude $\times$ longitude. The intensity (grey value) of the dot shows the number of users in that area in log-scale (black is maximum).

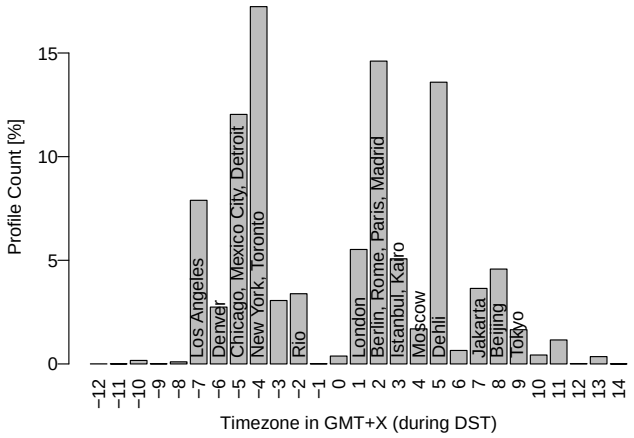

Figure 8. Distribution of profiles across time-zones.

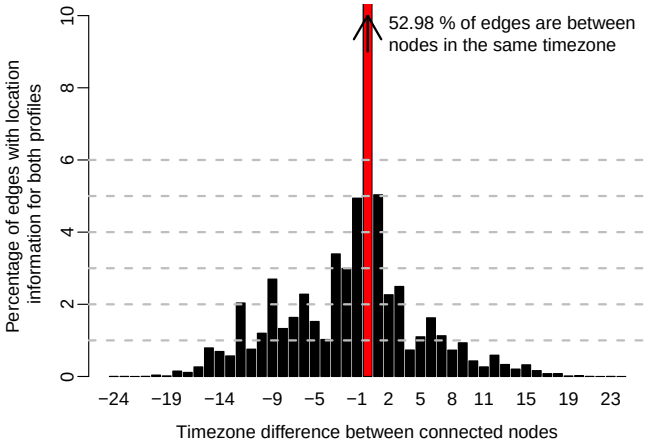

Figure 9. Distribution of timezone differences of connected profiles.

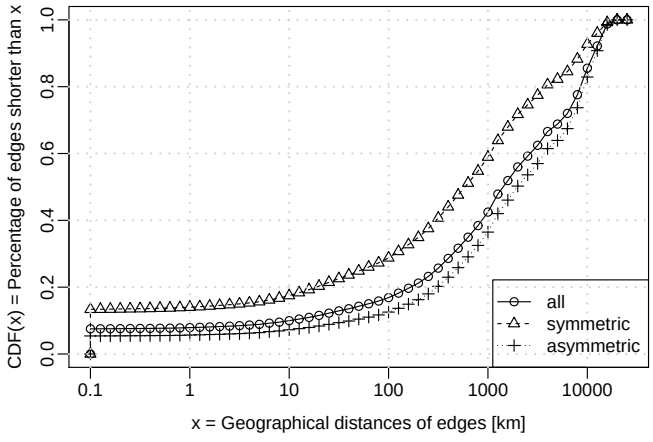

Figure 10. Geographical distance (in meters) of two connected users, differentiating between asymmetric and symmetric links. The distance distribution over all links is denoted by *all*.

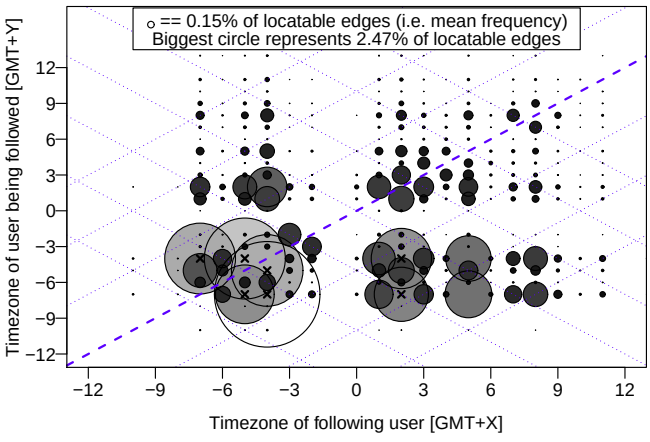

Figure 11. Frequency of edges per source/destination timezone pair. Bigger/lighter circles represent higher frequencies. We omitted circles when source and destination timezone are the same to increase readability. Timezone pairs with small frequency (less than median, i. e., 0.005 %) are also omitted.

technical expertise of the user population. However, quite a number of users are also in Asian time zones, such as those of Jakarta, Beijing, and Tokyo.

In the following we consider those links in the OSN graph, which connect two users for which both share a coordinate publicly . For each of those links we compute the time difference between the locations of the two involved users. Figure 9 shows the histogram of these differences. Note, that we cut the bar at $x = 0$ at 10 % although it reaches 52 %, for improved readability. A significant fraction of the links are within the same time zone or have a limited time difference. However, some links correspond to a large time difference. Notably, the timezone differences are skewed to the left, which indicates that people living east more often circle people living in the west. Particularly we find that American users are significantly more often circled from abroad, than they circle foreign users. It is also interesting to note the difference between asymmetric links (links between two users where only one user circles the other) and symmetric links where the two users circle each other mutually.

Figure 10 shows that around 50% of the Google+ neighbors are less than 1000km away from each other (see *all*, circels). Despite the many social neighbors who live in adjacent time zones, a significant fraction of these neighbors are separated by long distances, e. g., larger than $10,000$km. This might be an artifact of the micro-blogging features of Google+ that results in adding famous people who are really far away in a user's circles. Figure 10 shows that asymmetric links (plusses) are typically longer. We conjecture that symmetric links (triangles) are more likely to represent friendships which are more local in nature, while asymmetric links tend to describe (more global) follower-relationships. Scellato et al. [19] studied geo-social metrics for OSNs. Their results complement ours. They find that social links tend to stay local, whereas news- and file-sharing leads to longer geographical distances. They emphazise that the type of OSN service impacts the geographical distances of its links.

In Figure 11 we plot the frequency of link timezone pairs. Each timezone combination is represented by a circle, whose diameter/color determines the frequency of this combination. For the sake of readability we omitted circles on the diagonal (i. e., same timezone links) and timezone pairs with very low frequency. To give an example: The biggest circle (at $x, y = -4, -7$) represents for example links from New York City to Los Angeles (east coast to westcoast in general). We also see many Europeans folling users in the USA ($x = [1, 3], y = [-4, -7]$) and some asians follow US users as well. From Figure 11 we can see that there is indeed a trend in the directions of asymmetric links. In general they have an east-to-west tendency, as can be observed from higher number of bigger circles below the diagonal. Note, that Figure 11 is based on time zones. Therefore, a user in time zone UTC$+1$ might be located in Europe or in Africa. We also categorized users based on their coordinates on a per continent-level and studied sources and destinations of the links. The results (not shown) reveal that indeed a lot of links are directed to the US, e.g., twice as many edges start in Europe and end in North America compared to the opposite. For Asia, almost three times as many links end in North America compared to the opposite direction. Finally, many more links go from Asia to Europe than from Europe to Asia.

RELATED WORK

Researchers have been fascinated by the complex structure and organic growth of the Internet and the networks overlaying it (e. g., the WWW or peer-to-peer networks) ever since.

(Online) social networks are a particularly interesting type of networks as they reflect individual and collective human interactions [2] at a large scale and over time. For a (historic) overview, we refer to [3]. Researchers have investigated, e. g., algorithmic implications on the spread of information or routing [10], and have developed methods for predicting the creation of new links [13, 22]. Most of these works are inspired by empirical phenomena and insights from experiments, or extensive measurements. For example, the *small-world phenomenon*—the principle that people are all linked by short chains of acquaintances—has been a folklore and subject to anecdotal evidence until the pioneering experimental work of Stanley Milgram [14] in the 1960's. Milgram's quantitative results led to refined models, most prominently the Watts and Strogatz model [21], providing evidence of the natural and technological universality of this phenomenon, which also includes the World Wide Web.

A large number of empirical studies of OSNs have been conducted already. A complete overview is beyond the scope of this paper. Many results about the topological and sociological character of OSNs are due to Mislove and his collaborators. In [17], a large-scale measurement study is conducted of the topological structure of Flickr (see also the related growth study [15]), YouTube, LiveJournal, and Orkut, confirming the power-law, small-world, and scale-free properties of OSNs. [16] provides a demographic perspective by investigating the representativeness of Twitter users. In [18], the authors find evidence that users with common attributes are more likely to be friends and often form dense communities. Ahn et al. [1] study the growth patterns and topological (degree-based) evolution of OSNs (Cyworld, MySpace, and Orkut) and compare their results with the ones in real-life social networks. Cha et al. [4] compare three topological measures of influence (in-degree, re-tweets, and mentions) based on a large crawl of the Twitter OSN. Scellato et al. [19] analyze the annotated geo-location graphs of BrightKite, FourSquare, LiveJournal and Twitter, based on snowball sampling crawls. Gjoka et al. [8] study parallel relations between OSN users, by conducting multi-graph measurements of Last.fm.

An overview of alternative crawling approaches is discussed by Cormode et al. [7], who also provide a checklist for crawling (see also [9]) and argue that OSN studies must go beyond simple node-link models to include, e. g., time aspects. The authors apply their model to Twitter, Facebook and YouTube. For example, one approach to get a fast overview of the graph is to crawl from a set of sampled nodes (and e. g.,

perform random walks from there [8]). For instance, Cha et al. [6] study the Flickr graph, by crawling the graph from one node chosen randomly as a seed and following all links in the forward direction (snowball sampling), i. e., performing a breadth first search. This way, they obtain a single weakly connected component and study the in-degree and out-degree distribution of their sample of the resulting sampled Flickr graph. However, the resulting graph can depend heavily on the chosen start node and will find only one connected component, providing a limited and biased view of the network [11].

CONCLUSION

Google+ occupies an interesting position in the OSN space, between classic "friendship networks" such as Facebook where users typically have symmetrical relationships, and more asymmetric, "social media" / (micro-)blogging networks such as Twitter. [20]. Our analysis shows that Google+ users span all regions of the world, and have a clear bias towards a highly-educated audience, e. g., college students or IT professionals, making this OSN distinctive. Our analysis of the topological structure of Google+ reveals that it has a relatively symmetric in/out-degree structure for smaller node-degrees, but cannot clearly be classified as asymmetric (micro-blog) or symmetric (OSN for friendships), which makes it an interesting object to study. During the transition of the network just after its public announcement, we observe a decrease of the median and mean out-degree due to the presence of a larger number of weakly connected components. Note that the network is still rapidly evolving and it will be interesting to follow its developments in the future, especially since Google started to include more and more other Google services and when they will open it for third party applications.

We understand our work as a first effort to shed light onto the initial structure and evolution of Google+. We will continue crawling the publicly available Google+ data in the future to study long-term trends, and also aim to increase the time resolution of the crawls such that individual interactions and their dependencies and causalities can be studied.

ACKNOWLEDGEMENTS

We want to thank Meeyoung Cha (KAIST) and her co-authors who gave us access to the Twitter [4] and Flickr [6] data sets. We also want to thank Balachander Krishnamurthy from AT&T Labs-Research for a very fruitful discussion.

REFERENCES

1. Ahn, Y.-Y., Han, S., Kwak, H., Moon, S., and Jeong, H. Analysis of topological characteristics of huge online social networking services. In *Proc. of the* 16^{th} *International Conference on the World Wide Web (WWW)* (2007).
2. Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. Group formation in large social networks: membership, growth, and evolution. In *Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (2006), 44–54.
3. Boyd, D. M., and Ellison, N. B. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication* (2007).
4. Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of the* 4^{th} *International AAAI Conference on Weblogs and Social Media (ICWSM)* (2010).
5. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., and Moon, S. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *ACM Internet Measurement Conference* (October 2007).
6. Cha, M., Mislove, A., and Gummadi, K. P. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the* 18^{th} *international conference on World wide web (WWW)* (2009).
7. Cormode, G., Krishnamurthy, B., and Willinger, W. A manifesto for modeling and measurement in social media. *First Monday [Online] 15*, 9 (2010).
8. Gjoka, M., Butts, C. T., Kurant, M., and Markopoulou, A. Multigraph sampling of online social networks. *IEEE J. Sel. Areas Commun. on Measurement of Internet Topologies* (2011).
9. Gjoka, M., Kurant, M., Butts, C. T., and Markopoulou, A. Practical recommendations on crawling online social networks. *IEEE J. Sel. Areas Commun. on Measurement of Internet Topologies* (2011).
10. Kleinberg, J. The small-world phenomenon: an algorithm perspective. In *Proceedings of the thirty-second annual ACM Symposium on the Theory of Computing (STOC)* (2000).
11. Lee, S. H., Kim, P.-J., and Jeong, H. Statistical properties of sampled networks. *Phys. Rev. E 73* (Jan 2006).
12. Leighton, T. Improving Performance on the Internet. *Commun. of the ACM* (2009).
13. Liben-Nowell, D., and Kleinberg, J. The link prediction problem for social networks. In *Proc. 12th International Conference on Information and Knowledge Management (CIKM)* (2003).
14. Milgram, S. The small world problem. *Psychology Today 61*, 1 (1967), 9340–9346.
15. Mislove, A., Koppula, H. S., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. Growth of the flickr social network. In *Proc. Workshop on Online Social Networks (WOSN)* (2008), 25–30.
16. Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J. N. Understanding the demographics of twitter users. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)* (2011).
17. Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. Measurement and analysis of online social networks. In *Proc. of* 5^{th} *ACM/USENIX Internet Measurement Conference (IMC)* (2007).
18. Mislove, A., Viswanath, B., Gummadi, K., and Druschel, P. You are who you know: inferring user profiles in online social networks. In *WSDM* (2010).
19. Scellato, S., Mascolo, C., Musolesi, M., and Latora, V. Distance matters: geo-social metrics for online social networks. In *Proceedings of the* 3^{rd} *Workshop on Online social networks (WOSN)* (2010).
20. Spiegel Online. Wem google+ wirklich konkurrenz macht. In *Issue of July 6* (2011).
21. Watts, D., and Strogatz, S. The small world problem. *Collective Dynamics of Small-World Networks 393* (1998), 440–442.
22. Yang, S. H., Long, B., Smola, A., Sadagopan, N., Zheng, Z., and Zha, H. Like like alike: joint friendship and interest propagation in social networks. In *Proc. 20th International Conference on World Wide Web (WWW)* (2011), 537–546.