A measure of Online Social Networks

(Invited Paper)

Balachander Krishnamurthy
AT&T Labs–Research
http://www.research.att.com/~bala/papers

***Abstract*—Online Social Networks (OSN) command a user base of about half a billion users on the Internet. Although the traffic contribution in bytes by OSNs is significantly less than earlier applications responsible for dramatic increases on the Internet (such as peer-to-peer networks), OSNs have already had a profound impact on the Internet. The organic growth in the sheer volume of users, the range and diversity of applications run using OSNs as a distribution platform, and the wide range of new technologies underpinning their growth, all portend an enduring effect as well. While there are similarities to earlier phenomena, there are numerous differences due to some properties unique to OSNs. This paper enumerates interesting properties, the methodologies used to study them, and the challenges faced by researchers in measuring OSNs. A few results from recent studies my colleagues and I have been involved in are also presented.**

I. Introduction

The World Wide Web rose to prominence in the early 1990s. The convergence of the research and development community towards a message exchange protocol (HTTP), a naming infrastructure (URI), and a document markup language (HTML), followed by a popular graphical interface (Mosaic) resulted in millions of users accessing the Internet for the first time. By the early 21st century, the Web had become the number one application on the Internet. We are witnessing a similar phenomenon with the rise of Online Social Networks (OSN). In some sense a OSN is non-novel—a community bulletin board similar to early Usenet newsgroups—except the central entity is not a newsgroup topic but the user herself. The user creates virtually all the content and is responsible for most of the traffic on any OSN. The site owners, be they *Myspace* or *Facebook* are parsimonious in their contribution. Other than providing a distribution platform (an ability to reach a large number of users) and a few internal applications and pointers to many external ones, OSNs generally tend to stay out of the way. The OSN user uploads content in various formats, seeks out *friends* and interacts with them in different ways. In the WWW—a client/server system, the server owners control the content and manner of delivery with the clients being largely passive readers. The OSN world is a bit closer to the peer-to-peer (P2P) model. In traditional P2P systems content is *all* that matters: people want to "borrow" the bits of a song or a movie and do not really care which peer they download it from, as long as it is quick and clean. On OSNs the *users* are the focal objects and virtually all communications are between users and applications triggered by them. Understanding the role played by users is key to understanding the potential impact on the network due to OSNs. The nature of the content, size distributions, frequency of communication, inter-arrival time of requests are all different from the Web and P2P systems.

What is an Online Social Network? An OSN is a network consisting of real users who communicate with each other in an online setting in diverse ways. The set of participants in an OSN grows (and falls in some cases) over time; for example, Facebook has been adding 250,000 users *daily* for many months and has crossed 100 Million users since inception in August 2004. Users can solicit others to join and real world friends and acquaintances create sub-communities online. Relationships can be fragile or solid similar to the physical world and the types of OSNs can vary with the nature of social connections. Professionals, seniors, writers, students, just to name a few groups, have their own OSNs. Users can and do participate in more than one OSN but a significant fraction of their time is often spent in a single OSN. In the physical world we have local and distant friends, and random acquaintances; use different means—telephone, email, face-to-face, text messaging—to communicate with them. Inside an OSN, a user is likewise capable of using email, instant messaging, bulletin board writing etc. The range and diversity of communication styles available in OSNs run the gamut and many OSNs have similar and overlapping features. As yet, there are no official standards for OSNs: no broadly agreed-upon open APIs[1] or common languages.

What are the technical aspects that have driven the rapid growth of OSNs? OSNs became popular contemporaneously with the rise of the Web 2.0 phenomenon that ushered in several new concepts. Web 2.0 has significantly more content creators unlike the original Web 1.0. The essential difference between Web 1.0 and Web 2.0 can be seen along a few axes: technological, sociological, and structural [10]. Scripting and presentation technologies used to render the site and allow user interaction consist primarily of mashups and the open standards-based Ajax (asynchronous Javascript and XML). Ajax helps integrate Web page presentation, interactive data exchange between client and server, and asynchronous update of server response. Ajax's API allows large scale construction of code snippets to send data between a client and a Web server, often in XML format, but can be HTML, text, or customized formats. The sociological aspects deal with the notions of friends and groups, along with related issues such as their anonymity and privacy. Social aspect of OSNs provided the basis for their dramatic growth by virally drawing in a large

[1] OpenSocial and FBP notwithstanding

number of users in a short time. The *social graph* induced by the users (nodes) and links to their friends (edges) is at the heart of an OSN. The structural axis deals with the purpose of the site—enabling locating, linkage, and communication between friends and communities. The substrate of an OSN had to scale in order to keep up with the explosive growth of the social graph. Many OSNs have adopted virtually all the technical advances in Web 2.0. New features (like external applications) and new content types (such as videos) have forced the large OSNs to be very well provisioned to handle the sudden increase in number of network connections and the traffic that flows through them.

Why should networking researchers care about OSNs? To start, over half a *billion* users are members on various OSNs. That is nearly a tenth of the world population use OSNs. Although the volume in bytes exchanged on OSNs is still a small fraction of overall Internet traffic (as compared to, say, on P2P networks), there are clear indications that this will rise. The reason for this is not just the large number of users, but the overlay network induced by the popular external applications that use the distribution platform provided by OSNs to grow virally. Each application generates additional traffic between existing users and raises the probability of new users joining the OSN to interact with the rich and growing set of applications. Facebook alone already has over 40,000 user-contributed applications written using its FBP API. Provisioning for viral growth may be feasible within the OSN in a manner similar to how some popular Web sites have handled flash crowds: buying bandwidth and ensuring scalable server farms. However, the load on sections of the overall Internet could grow dramatically due to independent decisions made by a few OSNs (e.g., allowing uploads of videos by 100 million users or opening up their APIs to external developers). As soon as the micro-blogging OSN *Twitter* [22] opened up its API, the traffic on Twitter increased by a factor of *twenty*. The breadth of communication possibilities, with input to and output from OSNs increasingly diversifying, implies that anytime-anywhere-anyway communication is becoming a reality. The open-API model broadens choices to users and each change causes a new upsurge in the diversity of uses, number of users, and thus traffic volume. The concurrent explosive growth in worldwide cellular penetration (over 3 billion users) is likely to hasten the large-scale adoption of mobile-OSNs. Managing traffic growth due to OSN from a network infrastructure point of view is thus essential. Unlike the Web and P2P where content drove the traffic, OSN traffic growth is heavily dependent on what applications may become popular with users; i.e., the need for recognizing the centrality of the role of users is crucial.

Looking back, we see that networking researchers' contribution to the P2P revolution was minimal; popular client programs (like eDonkey and BitTorrent) induced dramatic traffic growth on the Internet. There was little attempt to standardize and academic contributions were too little and too late. Earlier, with the World Wide Web, which evolved more systematically, there were considerable delays and difficulties in standardizing the HTTP/1.1 version of the protocol. An early understanding of OSNs is thus imperative for networking researchers who are often removed from any specific application consideration. A key goal of this paper is to impart a broad idea of what OSNs are and some of the key challenges faced by researchers in measuring their properties of interest.

There are several important aspects of OSNs that are *not* discussed here, including information propagation, graph models, recommendations, and advertising. Likewise, the paper steers clears of any quantitative results, presenting trends instead. Snapshots of results are in the cited works and have limited shelf life in a rapidly changing field.

Section II presents a quick overview of a typical OSN session and distinguishes it from Web and P2P sessions. Section III enumerates properties of interest of OSNs. Section IV examines various challenges involved in measuring these properties. Section V discusses a few OSN-related studies in which my collaborators and I have been involved. Section VI examines related work followed by a few conclusive speculations on the future of this field.

II. A TYPICAL OSN SESSION

Figure 1 shows a typical OSN session to aid in the understanding of the complexities of OSN and potential difficulties in measuring and analyzing OSN traffic. OSNs differ in their interface requirements; some OSNs do not require users to log in while others do. Even OSNs that require a login differ in their choice of protocol; some require HTTPS (e.g., Facebook, Flickr, Hi5, Imeem, LinkedIn) while others use simple HTTP (e.g., Digg, Livejournal, Myspace). So while we discuss a "typical" session, it is important to note that the underlying set of interactions can and do vary across OSNs.

Figure 1 shows some participating entities and traffic paths in a user's interaction with an OSN. An OSN has several internal applications that access its internal database to present updates, lists of friends, output from various communication streams (e.g., the Facebook "Wall"), and advertisements. There are many third party applications that use the OSN's distribution platform—such as multi-user games, content rating, etc. These applications need credentials from the OSN for users to interact with their friends. The applications themselves run on the external developer's servers or on outsourced infrastructure (e.g., Amazon's Elastic Compute Cloud–EC2[2]).

In Step 1, the user logs in to the OSN (via HTTP or HTTPs). Until logging out in Step 5, the user communicates with the OSN and various external applications. Step 2 shows user communicating with internal applications (typically over HTTP) and facing typical latencies of interaction with any busy Web site. Some OSNs outsource portion of their content to CDNs (Content Distribution Networks, which are not shown in the Figure) and display advertisements as part of the output presented to the user. Step 3 shows an entirely different class of communication—with external third party application developers. Bi-directional communication between the user

[2]http://aws.amazon.com/ec2

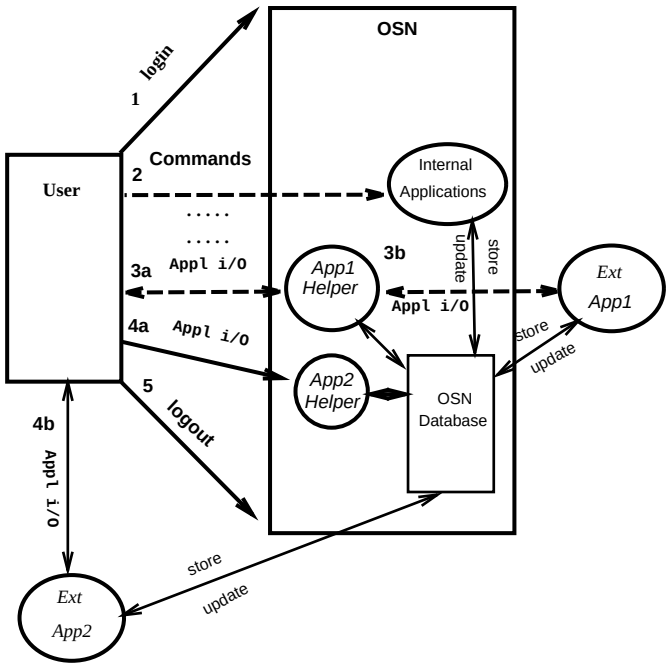

Fig. 1. Interactions between users, OSN, and external applications

and the third-party applications are routed through the OSN, as shown in Steps 3a and 3b. However, some OSNs do allow some direct communication between the user and the external application (steps 4a and 4b). The OSN database, depending on the privacy settings associated with the data elements, acts as the central repository of information and is accessed by both internal and external applications. More sophisticated options for interactions with external applications are now being created (e.g., Facebook Connect and MySpace's Data Availability) that allow external Web applications to access the internal social context of an OSN user on an opt-in basis.

Figure 1 shows a relatively simple OSN session; a real session includes several other potential interactions (such as communication between friends, different groups/networks), and numerous other external applications with possible simultaneous asynchronous communication. Each external application server (such as Ext App1 or Ext App2) interacts simultaneously with a handful or millions of users and numerous objects internal to the OSN (such as authentication and communication modules). The figure shows the typical *paths* of communication between the entities. There is complexity involved in tracking the overall network-level activities and in teasing apart the contributions of various entities. For example, a user may suspect delays at the external application server even if the bottleneck were inside the OSN and vice-versa. Likewise, delays in accessing and updating the OSN databases cannot be properly attributed. We will examine the impact of these complexities in the section on measurement challenges (Section IV).

Even at this simple level we can see some of the key differences between an OSN session and a typical Web or P2P session. There are hardly any external applications in most Web sessions. In a P2P exchange the number of entities (peers) involved can be in the hundreds but the set of actions is quite limited: once a peer is chosen, bytes are up- or down-loaded. Most peers are rarely involved in short sequential interactions; something typical between a user and an OSN. All the entities in an OSN session are generally highly available unlike the autonomous peers in P2P sessions. The degree of centralization in OSN is roughly between that of Web and P2P: applications are registered with an OSN, interactions are initiated through it but could continue independently. In the Web, all interaction is typically with the central content delivering authority (the Web server) while in P2P, the set of interacting peers are numerous, independent, and lacking any social connection to other peers. The mid-level centrality in an OSN session may imply some performance guarantees that again lie somewhere between a Web (more) and a P2P session (none).

OSNs evolved by starting out as typical Web sites. However, the dramatic increase in number of users, volume of data stored on their behalf, proliferation of external applications, advertisements, new features, etc. has caused popular OSNs to move towards a more distributed architecture. OSNs now use CDNs, large data centers, as well as advertisement networks similar to popular Web sites. As yet I am not aware of any published work examining the internal set up and architecture of a large OSN. The nature and use of OSNs do not present technical roadblocks to a much more decentralized architecture similar to P2P systems. However, the necessary trade-off in control of user information and associated commercial considerations will retard such an evolution.

III. OSN PROPERTIES OF INTEREST

We enumerate properties of interest of OSNs to provide insights at a macro and micro level and to examine their impact on the network. Table I lists the axes along which we can examine OSN properties: starting with high-level characterization and then moving to interactions with OSNs and intra-OSN issues. The third axis goes lower in the stack to examine network level traffic issues and the final axis examines social issues. Many properties are similar to other applications—Web and P2P, but some are unique to OSNs.

Basic characteristics: Characterizing a new application is often the first step undertaken in studies. Static properties characterizing an OSN represent a snapshot of the social graph at the time of the study. These include the number of users, distribution of friend counts, range of personal attributes, modes of communication opportunities, sub-communities within the communication graph, range and diversity of content associated with each node (e.g., object size and content types), ambient properties (geographical location and cultural attributes), forward and backward link structures enabling the graph to be traversed by users and crawled by programs, etc.

Static properties are generic across most OSNs and gives us a way to compare OSNs. For example, the number of users in an OSN is often the most commonly cited property. The difference in content types and frequency of updates hints

Basic characteristics	Dynamic interaction	Network traffic specific	Social
Number of users	Inter-communication frequency	Protocol usage	Anonymity
Friend count distributions	Session duration	Induced overlay network	Privacy
Personal attributes	Diurnal properties	Byte-fraction distributions	
Communication options	Rate of change	Signature of individual OSNs	
Sub-communities	Popularity growth	Signature of intra-OSN functions	
Content diversity	External applications		
Ambient properties	Sub-session features		
Friendship link structure			

TABLE I
OSN PROPERTIES OF INTEREST

at the demographics of a OSN; teenagers are more likely to update their pages with higher frequency than OSNs with older populations. Personal attributes are often captured in a profile at the time of account creation with aperiodic updates, and are strongly tied to the issue of privacy.

Different communication options enable a range of interaction opportunities in an OSN, reflecting its technical currency and sophistication. Writing on a group bulletin board possibly filtered to be visible only to a subset of friends, sending an Instant Message within the OSN, and automatic generation of update streams ('feeds'), are features in popular OSNs.

Numerous overlay networks can be formed in OSNs via sub-communities: school and work-related networks, geographical networks, or groups based on specific interests. These overlay networks help in discovering other users. OSNs differ in the set of data formats in which user content can be uploaded and hint at potential traffic volume (some OSNs allow video content to be uploaded while many do not).

Ambient properties capture mostly non-technical aspects of the OSN. Issues such as the presence of users from certain geographical regions, use of particular languages, and cultural norms can impact other properties. For example, Twitter is very popular in Japan leading to a large amount of Kanji characters seen in Twitter messages; such messages are likely to be exchanged only between Japanese users. The density of interconnection within an OSN yields clues about the participants and their closeness as a community. Knowing the friendship link structure is key to obtaining a coarse-level understanding of the social graph.

Although we have simply enumerated the static list of properties above, each of the properties can have interesting sub-properties. For example, the graph's diameter, ratio of node to edges, presence of distinct components are all of considerable interest. All these properties provide hints about the macro structure of the social graph and point out unusual aspects (e.g., the presence of a particular strongly connected component may be indicative of a special sub-community). Such properties have been examined in other Internet applications (e.g., the BowTie structure in the Web [7]). Detecting backward links helps us understand outliers like high volume communicators or spammers in the network.

Crawling a social graph is crucial for any characterization analysis. The static connectivity details representing the friendship structure should not be lost during any anonymization of the static social graph—required to make data available to researchers and preserve OSN users' privacy. It is not easy to reconstruct the large scale connectivity by cobbling together smaller chunks of the OSN—a sparsely connected graph will not have a straightforward connectivity pattern.

Dynamic interaction: Dynamic properties include temporal aspects related to communication (inter-communication frequency, diurnal effects etc.), rate of change of connectivity and manner of change (e.g., appearances of articulation points in the graph), popularity of nodes (number of people who access a particular node), the amount and nature of information exchanged between nodes and within subsets of the network. The amount of time spent interacting with the OSN and between users can help us characterize both the popularity of the OSN and the richness of communication afforded by the features available in the OSN. The amount of (clock) time spent on some of the popular OSNs on a daily basis is significantly higher than any individual Web site. The time of interaction (protocol-level time) with the OSN is however not that different from other Web sites.

Rate of change of contents in an OSN is different than content owner controlled Web sites. Popular news Web sites like `nytimes.com` or `cnn.com`, that are centrally administered and deal with timely information dissemination, have a higher rate of change than individually updated pages on an OSN. But many other categories of Web sites have pages that tend to change infrequently. Interaction with friends is one of the primary activity on an OSN—users are thus more active. While there are differences within OSNs, the rate of change on OSNs is generally much higher than many traditional Web sites. Given the sparseness of OSN connections many pages will only be accessed by a handful of people on a frequent basis. Such differences might argue for a different way to approach issues related to use of CDNs for OSN content.

The node and edge popularity in OSNs does change with time as users gain more friends and interact more frequently with a subset of their friends. Depending on the OSN there may be different kinds of communication between nodes on an OSN. One can visualize an overlay network formed on a per-external application basis that maps the set of users who participate in a particular application. For example, the collection of networks of Scrabulous players in Facebook may be an indication both in aggregate of the popularity of the particular application but could also indicate the depth of

connectedness between friends who are present in multiple such application overlays. A set of friends who interact with each other through multiple applications may be an indicator of the closeness of their friendship.

Growth in the addition of new members, recommendations of books, and similar cascades have been studied [12]. The viral nature of external applications is a novel phenomenon in OSNs. Some applications suddenly explode in popularity leading to a large number of downloads followed by traffic between the users, OSN, and the application. The potential partitioning of traffic in an OSN to be simply a union of communications between sets of friends is tempered by the realization that a growing fraction of traffic flow in an OSN is between users and external applications. Not all of the latter traffic has to flow through the OSN.

The temporal distribution of communication can help identify affinity groups. However, a detailed knowledge of functions internal to the OSN have to be known to extract sub- or intra-session features. The diversity of actions possible entirely *within* an OSN and those with external applications have to be individually teased apart. Separately, the increasingly popularity of Ajax requires examination of sub-session interactions at narrower time scales. Ajax is used for dynamic layout and reformatting of a Web page, requesting small portions of a Web page and reloading it quickly, and interacting on demand with the server.

Network traffic specific: The choice of protocols used and their extent of use is of interest. While there is some diversity across OSNs, most OSNs tend to use HTTP (and thus TCP for transport). Given the connection-oriented nature of communication, this is to be expected. The induced overlay network formed as a result of communication between sets of friends inside the OSN and with external applications is a novel aspect of OSNs that has not yet been studied in any depth. There are various difficulties in exploring this key property. Presently, the byte fraction due to OSN interactions are relatively small but steadily increasing. The actual set of operations that take place within an OSN—termed the 'signature' of an OSN—is an in-depth exploration of the details of a user's micro-interactions with the OSN. Depending on the OSN, the set of popular actions and the resulting network flow-level patterns will be different. Constructing a signature will allow us to reverse engineer network-level traces should they become available. Likewise, we may be able to identify internal functions of an OSN; Section V-A explores these aspects further.

Social issues: Social issues have been studied extensively in offline social networks. We examine two key social issues related to OSNs: anonymity and privacy. Both are of considerable importance given the penchant for broader disclosure by individual users on OSNs (as opposed to any other Internet application) and the potential for wide dissemination of such data. At a high level anonymity implies the absence of identity [2] or prevention of linking identity to actions, while privacy relates to specific attributes of individual users.

Anonymity is an antithetical thought in OSNs where one of the key purposes of joining is to share information voluntarily by users. It is probably fair to say that a vast majority of OSN users are thus willing to give up some degree of their anonymity to at least a small subset of selected users on the OSN[3] OSNs are typically reluctant to open up their networks to anyone (including researchers) who may be interested in characterizing its properties. Thus, from a research perspective it would be useful to have portions of the social graph available in an anonymized form. This is especially of interest when crawling an OSN is difficult. However, any anonymization has to preserve certain properties so that the modified social graph remains useful for querying. At the same time there should be analytical guarantees that the anonymized graph cannot be reverse engineered by adversaries.

Privacy has increasingly become a focal point of discussion in OSNs as concerns have arisen about identity theft and other abuses of personal data. Many OSNs early on provided options to their users to limit who can access different portions of their data. The default privacy settings and the set of privacy bits that are actually changed by users over time are of interest [11]. The concern of private information leaking to external applications and the risk of linking external information about the user has made privacy in OSNs a contentious topic.

IV. Measurement challenges in OSNs

Section III outlined various properties of interests of OSNs; we now examine some of the key measurement challenges and the difficulty in drawing inferences based on measurements.

Characterization challenges: As mentioned in Section III, researchers often first attempt to characterize a new application by gathering large amounts of data. The challenges in crawling OSNs are distinct from traditional Web or Peer-to-Peer crawling. OSN Crawlers must parse and extract a wide variety of links: navigation, friend, group etc., handle Javascript and asynchronous interactions by simulating user clicks. As pointed out in [10], the community needs general purpose tools that can be customized to crawl and parse a particular OSN site. Such tools will expose commonalities across OSNs and highlight generic technical issues that will help future measurers in OSNs. The controlled structure of OSNs together with their economic and privacy concerns, distinguishes the access issues from that of Web sites. Web sites often benefit from being crawled by search engines, as traffic can be directed towards them. OSNs do not have a similar need and control access to the social graph data. Restrictions on data gathering are common and often enforced by rate limiting the number of permitted requests within a specific time period (e.g., a few thousand requests a day). Researchers tend to circumvent by obtaining permission directly from the OSN authorities or rely on a broad-based measurement infrastructure such as PlanetLab. Using multiple client sites can help with getting a larger sub-graph but could violate the spirit of the restrictions of the OSN.

[3]We discount the relatively small number of fake accounts, as the effort needed to form a friends circle is harder due to the anonymity.

There are numerous challenges in gathering representative data, and results of the measurements have a limited shelf-life. First, crawling in an OSN can be blocked by OSNs through request count restrictions; and numerous accounts may be needed to get information in different sub-communities. Yet early attempts have been made using the open API of some OSNs to crawl them [19]. Given the extremely sparse connectivity in the OSN graph, the set of entry points for crawling have to be carefully chosen before claims of representativeness can be made. While repeated data capture starting in multiple random locations is one way to improve representativeness, parts of the graph may be inaccessible. The difficulty of obtaining a reasonable sample of users remains problematic, as we will see in the Twitter case study (Section V-B). The risk of missing one or more sub-populations can have a significant impact on observations related to personal attributes and or communication options. For example, the popularity of a particular technology in one culture (e.g., cellphone among users in Italy) may have to be taken into account in order to identify the reasons for some significant deviations from the norm. Ultimately, the only real way to obtain good quality measurements in the presence of constraints imposed by OSNs, is statistically valid sampling. There needs to be a large enough longitudinal sample that can withstand the variance in OSN characterization. The dynamics of OSNs have yet to be understood well enough for us to draw any long term inferences. Most of the current work consists of one-time (or a handful of) static snapshots that do not lend themselves for any deep inferences and lack the ability to predict direction of evolution of the OSN properties.

To study sub-communities one may have to become members of various regional networks and deal with limitations on the frequency of switching membership between them (e.g., twice every 60 days on Facebook). Multiple accounts may be needed to circumvent this limitation.

Examining content diversity is not that difficult as most OSNs tend to have just a few types of content: text, audio, static images, and occasionally video. However, obtaining specific fractions of each content type relies on the representativeness of the data snippet gathered. External indications of popularity of certain features may be used as a hint; some OSNs are well known for a particular kind of interaction and some OSNs even have significant limits on content diversity. Twitter, for example has only one kind of content that can ever be exchanged between users: a short (140 character) message. However, this is an exception.

Measuring some ambient properties is relatively easier as no new techniques need to be invented to handle geographical issues. Cultural differences can play a significant role in differences in properties of OSNs that are confined largely to a region. Some studies have already reported on cultural differences between OSNs that are specific to certain regions of the world (such as Orkut, popular in just a handful of countries and the large Korean OSN CyWorld [1]).

It is generally easier to obtain a handle on the overall link structure of an OSN; most users have a small number of friends, modulo a few outliers. Some OSNs require bi-directional acceptance before a link is allowed (i.e., one way friendship is not permitted) but there are exceptions to this (e.g., one can have many followers in Twitter and follow no one). Most OSNs display the friend count and statistically valid sampling can aid in obtaining coarse-level link count distributions. The dynamic nature of OSNs may require such data gathering to be repeated frequently.

What is important to note is that the various properties associated with the user are much more important than the traditional connectivity information. Users are the central objects in OSNs and thus any attributes measured need to be relevant to the user experience. A deeper understanding of the semantics of the interaction and cultural issues need to be factored in before attempting to draw conclusions about statistical properties of OSNs. Additionally, lessons learned from one cannot be trivially applied to other OSNs. The lack of a generic API across all OSNs further worsens the problem.

Dynamic interaction challenges: We next examine challenges in dealing with the dynamic interaction properties. The set of features in an OSN change often enough, necessitating more frequent macro measurements. As one of our studies (Twitter, in Section V) showed, dramatic increases in traffic can occur when an OSN opens up its API. The ability to write external applications that can be linked using the API allows new classes of *uses* and thus a new class of *users*, leading to the traffic explosion. Similarly, when some externally constructed applications spread virally, or an entirely new class of users join the OSN, frequency of interaction can change significantly. New users may download suddenly popular applications and existing users may start participating in large numbers. Such an increase may lead to patterns differing from traditional diurnal effects. A time bound OSN game that is going to expire shortly may trigger a flood of interaction during the last minutes, significantly altering session duration and frequency of communication. Such issues rarely arise on the Web but may have some parallels to spikes during downloads of new versions of popular Operating System kernels on Peer-to-Peer networks.

Popularity of individual nodes can change significantly due to external events: an article in popular press may lead to a large number of friendship requests. A program masquerading as a user may suddenly generate significant traffic to its followers.

Examining similarities across OSNs for common functions (listing sets of friends or communication between friends) via passive packet traces requires in-depth examination aided by traces of active interactions. Examining traffic interaction *inside* an OSN is harder due to the often opaque nature of its interface. Even if an OSN provides an open API, there is little indication of how internal functions operate.

Session times are macro-features obtained by examining packet traces or logs. The definition of an OSN session is tricky. Just as the "think time" issue in Web sessions (time spent reading the current Web page before accessing the next), users may have multiple tabs open on their screen and

switch between OSN sessions and other activities. In early experiments (Section V-A) we have run into the problem of automatically identifying session durations when an explicit beginning or end is not detected, leading to reliance on timeouts to bound session-related activities. Features of actions inside a session are harder to track without detailed packet traces and an understanding of the specifics of the OSN functions. A detailed temporal understanding of a user's interaction implies the ability to tease apart individual interactions such as writing on a shared board, sending an Instant Message to another user, or interacting with internal functions of the OSN such as updating one's settings.

External applications in OSNs present some distinct challenges. Over 40,000 external applications that have had a collective installation count of over a billion, are used over 34 million times daily on Facebook *alone*. Although applications are constructed using the API provided by the OSN, their interaction with the users can vary. The external applications are hosted in the application developer's machines or a computing cloud. Users may communicate with some applications exclusively via the OSN while some applications may use the OSN just for initial invocation and some other user interface aspects. Performance can be affected by delays at various stages: the user's browser rendering the OSN page during normal interactions, delays internal to the OSN, and those introduced by external applications. Multiple third party servers such as advertisement servers and image holding sites may also be involved. Separating and tracking the fraction of traffic that flows through the OSN from what is exchanged between users and external servers is necessary to understand the overall traffic dynamics induced by OSNs.

Measuring HTTP traffic on OSNs have to take into account interactions due to Ajax [10]. The precision related to measuring click counts, page views, and popularity in regular Web sites are harder in the presence of multiple asynchronous transfers for small updates to a Web page. Without an explicit 'click' a user can scroll and zoom in/out of interactive maps, leave browser tabs open in the background and scan the page later for new messages, status updates, etc. The updates are triggered either by HTTP requests or Javascript calls handled locally at the client end, avoiding a round trip to the server with significantly smaller typical response sizes. Internal to an OSN session, Ajax may be used for updating profile information or shared writable structures and status updates of friends (e.g., the Facebook"Wall" and "Minifeed"), and during common interactions.

Network traffic specific challenges: As only a few protocols are used in an OSN, modeling traffic is easier. The capabilities provided in each OSN often overlap and identifying them once might suffice. However, the popular network flow level data capture will not suffice to understanding the intra-OSN semantics. Passive packet traces combined with real-time explicit user actions is needed to see the efficiency of usage of any protocols. When it comes to external applications, measurement is virtually impossible without the ability to monitor at the external application server. Simply gathering packet traces at a few links will not suffice to gather a reasonable signature of the overlay traffic since the popularity of third party applications can often be spread geographically. Beyond the venue difficulty, as outlined in Section II (see Figure 1), different portions of external application related traffic may flow either entirely through the OSN or some portions may bypass it. If the fraction of such direct traffic between user and third party servers is high, measurements at an OSN will be an underestimate.

A simpler property to measure, that of byte-fraction distribution, can still present challenges if there is a policy change. For example, recently MySpace allowed its 150 Million users to upload videos instead of just audio and static images. Such a policy shift can radically alter the mix of content type and byte volume distributions and overall traffic ratios. Predicting such policy changes is hard.

Sub-session times are even harder to measure without detailed packet traces *combined* with a deep understanding of the actual semantics of the OSN's internal functions. Enumerating the set of popular actions inside an OSN is difficult without first generating individual signatures of possible actions. Section V-A details our initial attempts at reverse engineering intra-OSN communication.

Social issue challenges: An issue well known in the database community is the merging of external publicly available information with anonymized data in order to extract hidden connections and to deanonymize the graph. Thus, the question of identity being established or narrowed by merging external data is critical in OSN anonymization. For example, there are several ambient parameters to consider: Is there a linkage between physical geographical distance and friends on an OSN? On campus networks it is very likely that a significant fraction of friends are 'local'. This tends to diverge a bit in regional networks and high school or college networks. Another ambient parameter is the connection between the use of popular external applications and the differing strength of connection between friends. Close friends are more likely to have similar interests and notify each other about external applications and participate more often in them. The frequency of communications and choice of manner of communication can be an additional indicator. For example, it is a known sociological factor on OSNs that the younger demographic uses email almost exclusively with older members and text and instant messaging with other younger members. The bandwidth usage between edges in a clique can thus be an indicator of differences in communication. The presence of potential cliques in the graph are of interest.

Available properties that would deanonymize the social graph are relatively few. Path length (diameter of the graph) is not a concern unless we can say how that would lead to re-identification. Breaking a large graph (such as a typical OSN) into cliques will still likely give *k-anonymity* [21] (a level of obscurity attained by ensuring indistinguishability of a released item of data among k different items) with a very large k for a given clique. The logical overlay network (e.g., application based links) could be a source of leakage. Note that

many of the issues raised in the dynamic interaction category cannot be answered via just the static graph. Thus, if only details about the static graph are made available to researchers, then the privacy aspect of the graph is higher while its utility is lower.

Many OSNs require users to log in before providing access to any information regarding internal settings; this raises the need for obtaining multiple accounts on different OSNs. Gathering privacy related data in OSNs faces the familiar problem of representative data gathering. There may well be cultural differences reflected in the levels of concern about privacy and such concerns may change over time. A broad-based longitudinal data gathering is thus essential. OSNs periodically change their policies regarding privacy settings. The potential for privacy to leak as a result of combination of data about the user is the hardest measurement challenge—personally identifying information about a user does not have to be explicitly present in an OSN. It may be possible to narrow down the attributes to a small set of users and then associate information to identify a specific user. Obtaining all sources of diffusion of personal information can be hard and thus an effective metric for privacy will remain elusive.

V. OSN STUDIES

Short glimpses of early studies that I have undertaken with my colleagues exploring OSN properties at various levels are now provided. There are several other interesting pieces of early works that have been carried out by others (see the proceedings of Workshop on Online Social Networks [23]). The first study examined packet trace gathering focusing on session reconstructions based on network-level characterization. The second study is a characterization of a popular micro-OSN (Twitter) to examine properties such as traffic volume, node popularity, diurnal nature, access patterns, geographic spread of users etc. The final study explores the role of privacy in various OSNs.

A. Sniffing OSN traffic

Packet traces have been captured by the measurement community for numerous applications. Based on where and how traces are captured, they can provide a detailed view of bi-directional traffic with attributes like timestamps, source/destination addresses, packet headers and even payloads. The challenges are well understood, mainly dealing with accurate capture of high volume traces in high-speed links.[4] Mapping the low-level traces io higher level connections has been done for other applications via generic tools.

We now examine unique challenges in dealing with OSN traffic. Assuming that all ingress and egress traffic goes through a single link monitored without any loss, we can make concrete statements about OSN usage patterns of the users behind the link. The duration of data capture will have to be sufficiently long to draw any meaningful inferences about OSN usage pattern, since a typical user spends only a few minutes a day on OSNs. The volume of data is sufficiently low to allow gathering of all interactions; however this may change if byte-heavy data formats, such as video, become a key part of data uploaded by users. The ability to gather full header and payload makes rich inferences feasible.

A typical approach is to either target one or more OSNs that are of interest, and identify the set of destination IP addresses that comprise the OSNs. For example, a single OSN, such as Facebook, may have a dozen IP addresses that cover the main Web site (`www.facebook.com`), the various support sites including any CDNs. If we are interested in knowing even a subset of external applications that use the OSN site as a distribution platform, then the number of IPs to track can grow arbitrarily large (such applications run on servers hosted by the application creators). So even the simple notion of tracking all actions related to a single OSN can be quite complex. Identifying the complete set of IP addresses is not a one-time task however, as there can be evolution within the OSN as a result of new features or new applications that emerge almost daily.

To identify IP addresses, we used reverse DNS lookup mechanisms and public databases. To bootstrap we generated interactions with the OSN with the traffic being monitored. The set of destinations accessed, the various protocols used (e.g., *https*, *http*), interactions with third-party sites (such as advertisement sites), could all be tracked. Uninteresting destination IP addresses were eliminated during the subsequent passive data gathering. Such an active injection of traffic combined with passive analysis yields a broader set of destination addresses and better identify intra-OSN actions.

With the set of destination IP addresses identified, the sniffer simply gathers bi-directional traffic associated with them. Our sniffer was in front of large collections of users. The traffic was mapped from low-level packet traces to higher-level application-specific actions using traditional tools that reconstruct HTTP request-response streams. An OSN session could be identified if the OSN required the users to explicitly login and logout, else we used simple timeouts. Next, signatures were generated on a per-OSN basis to map the HTTP request-response streams into records that map to individual OSN action sequences. Once the individual intra-OSN sessions are bracketed, we can infer both macro-level characteristics to compare OSNs, and micro-characteristics to examine what kind of actions are typically carried out within an OSN. The actively injected stream can be of significant help in identifying common action sequences within each OSN and improve the signatures. The use of Ajax for generating updates in the middle of a user's session must also be tracked.

Using a set of packet traces gathered in multiple geographical locations, our (ongoing) study showed multiple servers involved within a single OSN with considerably more complex interactions than originally expected. Even identifying a single user's session was complicated due to the difficulty in constructing signatures: each OSN differs enough in the way in which they maintain session information associated with a user. Additionally, some users have multiple simultaneous

[4]For details on difficulties with gathering packet traces see Chapters 4 (general issues) and 7 (application level trace collection) of [18].

sessions from the same IP address; thus complicating the notion of session duration.

B. Micro-OSNs

Micro Online Social Networks are distinguished by the brevity of the content exchanged. A prime example is Twitter, a popular OSN, which uses Short Message Service (SMS[5], a store and forward best effort delivery system for text messages). YouTube videos, in contrast, are significantly larger (order of megabytes) while 'tweets' – a status update or message in Twitter – are limited to 140 characters (an SMS limitation). Other micro-OSN examples include *qik* for streaming video from cell phones, Dodgeball (`http://www.dodgeball.com`) which lets users update their status along with fine-grained geographical information, GyPSii (`http://www.gypsii.com`) aimed at the mobile market that combines geo-location of users with image uploading, and Bliin (`http://www.bliin.com`).

Twitter functions as a publish-subscribe mechanism. Micro-OSNs like Twitter deliver data to interested users over multiple delivery channels. A user can generate tweets via the Web, SMS, Instant Message, tailored applications in OSNs like Facebook, or through literally dozens of customized applications written to interact with Twitter. Tweets can also be received via many of the above means. Twitter has already been used in diverse settings: helping people communicate during riots and large-scale fires, traffic updates etc. We present a brief precise of characterizing and analyzing Twitter next (see [6] for details).

Consider someone interested in monitoring Twitter traffic to obtain a representative sample of the set of users and their interaction. The limited "timeline" (random recent updates) provided by Twitter itself is an insufficient biased sample, consisting of updates of active Twitter users. If we were to start in some node in the Twitter graph and iteratively fetch information about a set of followers and friends, we will obtain a portion of the *static* graph, including many users who may not have been active for a very long time. We carried out multiple crawls to capture static and dynamic snapshots. Our study shows the difference in demographics obtained: a distinct community of Twitter users (in Japan) were poorly represented in the static crawl because they tended to have a disproportionate number of friends who tweeted using Kanji and thus lacked followers in the large English-only set of Twitter users. An examination of the types of users ferreted out the presence of *broadcasters*: software programs (not actual users) that have a large number of followers; operated by newspaper sites (e.g., New York Times) and radio stations, generating headline messages and song playlists. The assignment of user IDs was not sequential and had large jumps in the middle: inferences (e.g., on user counts) will be skewed if we ignore such outliers or inflection points. Policy changes will continue to affect the numbers–recently Twitter limited the number of users one could follow to 2000 and curtailed tweet deliveries to cellphones in all but three countries [17].

C. Privacy

The large number of users on OSNs combined with the extent and nature of information posted online have raised privacy concerns significantly. We carried out a study [5] exploring the range of privacy settings available on OSNs and the roles of OSNs and third party aggregators. Most users are not aware of who has access to their private information and more importantly whether there is a real need for such unfettered access.

A privacy bit is a unit of personal information about the user, such as age, data of birth, list of friends etc. Some bits are more private and the degree of their importance vary across users. An OSN typically assigns its own weight to the various privacy bits and offers different degrees of control over them. The private information is visible to the OSN itself but some bits are shared with external applications downloaded by the user. We identified the various bits of private information currently being shared, with whom, and if users could do anything to *prevent* such sharing. We also identified different bits of information that are shared with external data aggregators and advertisement sites with or without the user's knowledge or consent. Our study showed that most OSNs have somewhat similar notions of privacy bits and the bits could be grouped into a few classes such as thumbnail, greater profile, list of friends, user generated content, comment etc. We examined the default settings and the ability to change them in several popular OSNs[6]. An interesting observation was that MySpace allowed all users (even those who have never had an account in that OSN) to see all the privacy bits by default! Facebook was slightly more restrictive and other OSNs were in between. Many privacy bits are controlled by a single setting and the default settings are quite permissive. Most users rarely changed the default settings in many OSNs. Even though only a few thousand random IDs were examined, the sample was statistically valid. On Twitter we had access to nearly 10% of the user base and 99% of them had not changed the default privacy setting. On Facebook we examined a large number of "regional" networks (i.e., geographical communities) inside the US and in numerous cities worldwide representing different population sizes. There was a strong *negative* correlation across the population sizes in the extent to which trust was shown in the form of having their profiles and list of friends visible to everyone—users in smaller regions were more trusting. This observation held for regional networks in the US as well as across a wide variety of cultures worldwide.

Examination of third party advertisers and data aggregators showed the same disturbing trends of a few well known large aggregators learning about OSN user's access. These aggregators (such as doubleclick.net, atdmt.com, googlesyndication.com, yieldmanager.com) are the same ones that gather

[5]http://en.wikipedia.org/wiki/Short_Message_Service

[6]MySpace, Facebook, Imeem, Bebo, Orkut, Friendster, Hi5, Xanga, Twitter

information about user's movements on the World Wide Web. The ability to correlate information across different points on the Internet remains a major concern for privacy.

VI. Other related work

Many studies have examined individual OSNs (e.g., YouTube, LiveJournal, MySpace [4], [14], [15], [16]). A study of Flickr and Yahoo! 360 networks [16] explored path properties (such as diameter), density (ratio of undirected edges to nodes) change over time, and presence of a single giant component. A more recent study of Flickr's growth [20] examined its symmetry and the adherence to the preferential attachment property and pointed out clustering at a local level. Issues related to finding "backlinks" when the inward-pointing nodes have few incoming links themselves has been studied in the context of OSNs [19], which also measured degree distribution, clustering coefficient, and connected components of OSNs. YouTube has been studied more broadly for number of views and rankings, popularity time [8], [25], access patterns [9], and degree and cluster coefficient of the embedded network [19]. Rather than fetching large content (like videos), simple statistics on them can be reported via indexes. However this means that analysis of bit-rate choices or other encoding features are hard [9]; this study also showed that video clips on YouTube were longer than the ones found in the general Web and uploaded at higher bitrates. Similarly [25] showed that local and global popularity of video clips are significantly different by studying popularity of YouTube in campus environments, adding support for local caching.

There are several papers on OSN anonymity in general [3], [13], [24] with the focus on examining identity leakage due to attacks or data being published. A close related concept is that of re-identification [18] that lets anonymous data to be linked with actual identities by combining external data.

Popular extensions in the Firefox browser allow for anonymized access and the new features in the Internet Explorer browser 8.0 version such as InPrivate browsing and InPrivate blocking of certain JavaScripts may be a harbinger of things to come in the world of OSNs as well. Popular OSNs like Facebook have recently revamped their privacy setting but only time will tell if this leads to a focus on the part of the users on this problem.

VII. Conclusion

The key properties of interest related to OSNs and a set of challenges faced in measuring them have been outlined. The large number of users and external applications, and the potential for an explosion in traffic merits a closer examination of OSNs. Initial studies characterizing and measuring OSNs brought out similarities to P2P and Web and some novel challenges. The distribution platform provided by OSNs and the increasing migration of rich social connections to their online counterpart are introducing new challenges such as privacy concerns. Architectural changes are also likely to take place as OSNs may move from being largely centralized to a more distributed set up.

VIII. Acknowledgment

I'd like to thank my collaborators in various OSN projects: Martin Arlitt, Chen-Nee Chuah, Graham Cormode, Anja Feldman, Atif Nazir, Fabian Schneider, Phillipa Gill, Walter Willinger, and Craig Wills. Many of them also gave comments on this paper along with Virgilio Almeida, Zihui Ge, Lee Humphreys, K K Ramakrishnan, Rick Schlichting, and Kobus van der Merwe–my thanks to all.

References

[1] Ahn et al. Analysis of Topological Characteristics of Huge Online Social Networking Services. In *WWW 2007*, May 2007.

[2] Anonymity and privacy in electronic services. https://www.cosic.esat.kuleuven.ac.be/apes/.

[3] L. Backstrom, D. Huttenlocher, and J. Kleinberg. Wherefore art thou R3579X? In *Proceedings of the WWW*, 2007.

[4] Backstrom et al. Group formation in large social networks: membership, growth, and evolution. In *KDD*, 2006.

[5] Balachander Krishnamurthy and Craig Wills. Characterizing Privacy in Online Social Networks. In *SIGCOMM Worskhop on Online Social Networks*, August 2008.

[6] Balachander Krishnamurthy and Phillipa Gill and Martin Arlitt. A few chirps about Twitter. In *ACM SIGCOMM Worskhop on Online Social Networks*, August 2008.

[7] Broder et al. Graph structure in the Web. In *WWW*, 1999.

[8] Cha et al. I tube, you tube, everybody tubes. In *IMC*, 2007.

[9] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. YouTube traffic characterization: A view from the edge. In *IMC*, 2007.

[10] Graham Cormode and Balachander Krishnamurthy. Key differences between Web 1.0 and Web 2.0. *First Monday*, 13(6), June 2008.

[11] R. Gross and A. Acquisti. Information revelation and privacy in online social networks (the Facebook case). In *Proceedings of the ACM Workshop on Privacy in the Electronic Society*, November 2005.

[12] J. Kleinberg. Challenges in social network data: Processes, privacy and paradoxes, 2007. Invited Talk, ACM KDD07.

[13] A. Korolova, R. Motwani, S. Nabar, and Y. Xu. Link privacy in social networks. In *ACM 17th Conference on Information and Knowledge*, 2008. http://www.stanford.edu/~korolova/link_privacy_CIKM08.pdf.

[14] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. Structure and evolution of blogspace. *CACM*, 47(12):35–39, 2004.

[15] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *WWW*, 2005.

[16] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *KDD*, 2006.

[17] B. Kunz. The trouble with Twitter. http://www.businessweek.com/print/technology/content/aug2008/tc20080815_597307.htm. BusinessWeek, Aug 18, 2008.

[18] Mark Crovella and Balachander Krishnamurthy. *Internet Measurement: Infrastructure, Traffic, and Applications*. John Wiley & Sons, 2006.

[19] Mislove et al. Measurement and analysis of online social networks. In *IMC*, 2007.

[20] Mislove et al. Growth of the Flickr social network. In *SIGCOMM Worskhop on Online Social Networks*, August 2008.

[21] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based systems*, 10(5):557–570, 2002.

[22] Twitter: What are you doing? http://www.twitter.com.

[23] Workshop on Online Social Networks–2008. http://www.acm.org/sigcomm/sigcomm08/wosn.

[24] B. Zhou and J. Pei. Preserving privacy in social networks against neighborhood attacks. In *ICDE*, 2008.

[25] Zink et al. Watch global, cache local: Youtube network traces at a campus network - measurements and implications. In *IEEE MMCN*, 2008.