A Manifesto for Modeling and Measurement in Social Media

Graham Cormode Balachander Krishnamurthy Walter Willinger
AT&T Labs-Research

July 7, 2010

Abstract

Online Social Networks (OSNs) have been the subject of a great deal of study in recent years. The majority of this study has used simple models, such as node-and-edge graphs, to describe the data. In this paper, we argue that such models, which necessarily limit the structures that can be described and omit temporal information, are insufficient to describe and study OSNs. Instead, we propose that a richer class of Entity Interaction Network models should be adopted. We outline a checklist of features that can help build such a model, and apply it to three popular networks (Twitter, Facebook and YouTube) to highlight important features. We also discuss important considerations for the collection, validation and sharing of OSN data.

1 Introduction

Online Social Network (OSN) research to date has been dominated by a tendency to abstract a given OSN as a static graph. The appeal of modeling all OSNs uniformly with nodes to represent generic users and links to connect those that are "acquaintances" (covering a range of meanings from real-world friendship to mere interest in their comments and links) is understandable. However, not all OSNs are equal—indeed, not all networks studied under the guise of OSNs are truly "social networks"—and ignoring the full array of features of real-world OSNs is detrimental to understanding these systems.

The allure of a "one size fits all" model is that it brings the study of OSNs squarely into the realm of graph theory, with ready access to a rich set of available techniques and models. It also makes OSNs a prime application for the popular new field of network science, which encourages comparisons of networks from very different domains via well-known metrics such as node degree distribution, clustering coefficient, network diameter, etc. Observed similarities in these metrics across different networks are used to argue for the existence of universal features of networks.

Nevertheless, squeezing increasingly richly structured OSNs into the restrictive framework provided by static graphs is no longer tenable. It limits our capacity to pose and answer challenging questions that arise

in the context of real-world OSNs. Treating an OSN as a collection of generic nodes and links reduces all questions to counting hops and testing connectivity. Links between objects of fundamentally different types must either be forcibly equated or dropped. OSN-specific "details" such as its overall purpose, the functionalities it provides, or the features and mechanisms explicitly designed to support the provided functionalities have simply no room in a connectivity-only description.

Furthermore, OSNs are inherently dynamic constructs: their structure is continuously evolving over time as new objects and connections are added or (more subtly) become stale and irrelevant. Therefore, it is vital to include temporal information about *activities*, such as object creation time and time of actions. Such continuous change or dynamism is an overarching characteristic of real-world OSNs, impacting every aspect of their study, from measurement to analysis and modeling and model validation. Such systems with many types of nodes and interactions are perhaps better thought of as *Entity Interaction Networks (EINs)*, to emphasize the rich structure.

Trying to attach "meaning" to nodes and links of a given OSN requires first and foremost domain knowledge—what is the overall purpose or functionality of the system at hand, and what are the main features and mechanisms that the system supports to provide it. By "purpose" or "functionality" we do not require a formal mathematical definition but rather a verbal description that captures the essence of how the system is used. An exhaustive enumeration of features supported by a system is unnecessary—capturing those features used by the majority of active users will suffice.

For example, YouTube's network consists of videos as first-class objects, users as second-class objects, and multiple kinds of links amongst these objects. The system allows users to benefit from and contribute to the main purpose of sharing videos, via uploading, commenting, rating etc. Although users may "befriend" each other, characterizing YouTube as an OSN and studying the (simple) graph of user friendships in isolation is neither informative nor relevant as it completely ignores the main purpose of YouTube and its popularity. Instead, we claim it is better modeled by the broader class of EINs.

Facebook is an example where social features are paramount and users are the most important object. At first glance, it might appear that it is well-suited to a simple static graph model with friendship links connecting users. But on reflection, this model also fails to capture the true nature of the system. The core functionalities of the system allow multiple ways for users to interact: via private messages, status updates, wall postings etc. The type and frequency of such interactions are much more nuanced than simply recording that two individuals once indicated mutual friendship. On top of this, the growth of third-party applications has led to a feature explosion, exposing many more ways for interactions to occur. Other social networks have different mixes of YouTube-like content-sharing (e.g. Flickr) and Facebook-like socialization (e.g. MySpace).

What about Twitter, the minimalist site based on micro-content sharing—surely this must fit the classic OSN model of nodes and edges? But even here, the usage of the service has evolved more complex struc-

tures: follower/following relationships, targeted replies, hashtags to group tweets, re-tweeting and more. The disparate modes of access (web, various smartphone apps, SMS) further complicate the model. The integration of Twitter into other services, such as Facebook, further indicates that to get a full picture of the dynamic social ecosystem, we need a framework that goes beyond a static graph.

Understanding the overall purpose of an OSN and the main features it provides facilitates assigning "meaning" to its nodes and links. These nodes and links should have "types". Consider Flickr: nodes are of type "photo" or "user", links can be of type "friend" (one user declares a "friendship" with another) or "interaction" (a directed link from a user to a photograph). Interaction links are further subtyped, based on whether a user posted the photo (a one-to-many relation), or marked themself as a fan of the photo (a many-to-many relation). Both links and nodes have attributes: in Flickr, a photo has a time of posting, resolution, properties of the camera used, a list of comments; link properties include timestamps indicating time of creation (and possibly deletion) of the link.

The resulting entity interaction network makes it explicit that users of Flickr interact with one another in multiple ways: through user-generated content; i.e., photos, and also by declaring friendship relations. When annotated with temporal data, all features evolve over time. With this rich entity interaction network in place, it is now possible to investigate its dynamism from a number of different angles. Questions of how, say, photos become popular in Flickr, and whether the friendship information influences this popularity, can now be studied in great detail,

The richness and dynamism of the resulting entity interaction networks reflect and capture many of the new features common in Web 2.0 (e.g., purpose and layout of a site, enabling user interactions, support for friendship and group formation). These evolving network structures generalize the notion of OSNs in the sense that they encompass online "non-social" networks such as YouTube or Flickr as well as online "social" networks like Facebook or MySpace and everything in between (e.g., Twitter). They enable different ways of "slicing" a given structure (i.e., specifying particular node or link types and/or attributes) to account for the different functionalities of the system and their evolutionary nature.

The static graph representations that have been the focus of conventional OSN research to date can be recovered by appropriately collapsing nodes and links of different types and/or attributes and ignoring all timing information. In this sense, methods from graph theory or network science are still applicable, but with this more complex entity interaction network framework in place, it is now possible to ask more strongly whether studying certain "slices" (either as static or dynamic objects) is even meaningful. Since nodes correspond to particular OSN-specific entities (e.g., user, photo, video, object for sale, artifacts within the network such as comments or ratings), it is largely the task of the researcher to determine how feature-rich the resulting entity interaction network has to be so the set of problems and questions of interest can be properly studied and resolved.

Different problems are likely to require different abstractions. While there is generally no "right model"

for any given set of problems, OSN-specific domain knowledge is likely to serve as the ultimate guide for determining the levels of detail needed for a particular study. An appropriate comparison is to database design: modeling a system to form a schema requires identifying the entities and their attributes, and the links (and link attributes) that connect them. There may be several different valid solutions, but a general set of principles guide the modeling process.

Our goal in this paper is to lay out the key properties of a network that should be made explicit in a model, and the measurement choices that affect the collection of data in this model. We then outline the effect of this exercise for three popular networks (Facebook, Twitter and YouTube), and discuss how prior studies have fared in their modeling. This leads us to present recommendations for both the measurement community and the sites themselves to improve our ability to model and measure OSNs effectively.

2 Modeling and Measurement Methods

Compared to abstracting a given OSN to a simple static graph, representing it in a meaningful manner as an Entity Interaction Network model to account for its purpose and its usage looks daunting. In this section, we provide some general guidelines to facilitate this process and address different aspects of OSN modeling and measurement.

2.1 Modeling

A simple graph G, may be represented as a set of nodes V and edges E. Entity Interaction Networks may better be represented as multigraphs or hypergraphs, H, which consist of collections of nodes (one set of nodes for each type), and edges connecting pairs of nodes. While deciding on the types of nodes or edges or on which node/edge attributes to include when building the entity interaction network for an OSN, the following is a (partial) list of questions that can serve as useful guide.

Node Properties. Concerning the entities which are represented within the network and which may be abstracted as "nodes":

- How is a new node created (by user action, by site)?
- How are various node properties set (by user, site, 3rd party data, from a digital object's properties)?
- What determines the life time of a node (creation time, deletion time)?
- Where does a node "live" in the network (i.e., where can it be observed/measured)?
- How can a node be found by a user (via search, links from other objects)?

These questions necessarily relate to the view presented to the measurer: if the user has access to additional properties of a node, or the OSN retains nodes after deletion, this does not alter the properties if these are not visible to third-parties.

Edge Properties. Concerning the connections which can be formed between two entities within the network, and which may be abstracted as "edges":

- What types of entities are linked by a particular edge?
- Are the edges directed or undirected? Can directed edges be reciprocated (a matching edge in the reverse direction added)?
- Who can create an edge from entity A to entity B: the owner of A? the owner of B? either? other users in the site? by the site itself?
- Whose consent is required for the creation of a proposed edge, if any? (the owner of the entity receiving the link? the site?)
- Can edges be deleted, and if so, by whom?
- What determines the lifetime/shelf-life of an edge?
- Are links between certain entity types many-many, many-one, one-one?

Where groups may be formed (such as a group dedicated to a common interest), these may either be represented as a node for each group, with an edge from an individual to the group node to indicate membership or as an additional node attribute.

2.2 Measurement

The methods used to collect data about OSNs can have considerable impact on the conclusions reached about properties of the networks. To properly place OSN studies in context it is necessary to address the following issues, and be explicit about the choices made in collecting the measurements:

Data Collection Technique. In an ideal world, the OSN would release full and accurate data to researchers directly. However, due to privacy concerns and competitive reasons, OSNs have strong motivation to avoid this. There are some examples of anonymized datasets being released—the Netflix prize dataset being the best known—but these may not capture the necessary data to answer a given question. So instead, researchers mostly collect their own data. There are three general approaches used to collect data on social networks, each with its own sets of limitations:
API driven: the API (Application Programming Interface) provided by the OSN is used to query the entities, properties and relationships. However, this requires us to assume that the answers to queries via the API are

up-to-date and accurate. Further, it is important to understand how random an API call is which provides a "random" or "recent" example of a particular type. Does the site impose a limit on the number of API calls per user per day? The order in which answers are supplied together with a limit on the answers may end up in yielding a highly biased sample and restrict the scope of a measurement study.

Scraping based: the measurer directly accesses the site via a web client, which imitates the actions of a user to capture HTML which is parsed with a hand-crafted site-specific parser. This is necessarily more arduous than API-based methods, and may still run into bandwidth limitations imposed by the site (many sites actively identify and block attempts to scrape). The scraper also has to contend with site redesigns which break the parser: these tend to occur with greater frequency and with less notice than changes to the API.

Passive network measurement: the measurer sniffs network traffic (say, at the edge of a campus or enterprise network), and sifts out and parses requests to and from the OSN of interest. In some ways, this provides the most honest view of the network in use, in that it can capture properties of the network as its users experience it. However, there are significant privacy issues around sniffing and parsing individual's activities within an OSN. The plethora of access modalities for modern OSNs (direct web based, mobile web, mobile app, external apps using API) mean that it is hard to capture all accesses from any meaningful subpopulation of users.

Sampling Methodology. Given that the most popular OSNs boast hundreds of millions of entities, it is not feasible to gather complete information on all properties and activity within the network. Necessarily, any measurement study yields only a sample of the full data, and it is important to make explicit the description of what has been sampled, to materialize the biases and systematic errors. The most common sampling methods are based on a simple graph view of the network, and each has its limitations. A limitation that is common to all existing sampling methods is that they are largely unable to deal with the dynamism of real-world OSNs—the system changes as it is sampled, causing biases and errors that remain ill understood.

Random node identification. A random set of nodes are identified and their properties and associated links are collected. Truly random sampling requires detailed knowledge of the space of node identifiers, and the ability to access arbitrary nodes given their identifier; or the existence of an API call or website function to return a random node (often these are far from "random"). Given the large size of OSNs, this approach typically yields many isolated nodes, making it unsuitable for studying connectivity, reachability, or distance-based questions.

Exhaustive crawling within a defined boundary. All nodes which satisfy a given property are retrieved, such as all members of a particular group, or all individuals who identify themselves with a particular institution or locale. This assumes that there is the functionality to identify all such nodes and collect them and their interactions, which varies across OSNs. Caution is also needed in extrapolating findings from a particular

demographic group in a particular OSN to wider populations, and in claiming that all individuals from a particular population have been included.

Breadth first search from one or a few seed nodes. Since OSNs typically make available information about links out of a given node, a common practice is to start with a few seed nodes and perform exploration of their neighborhoods, often in a breadth first manner. After some number of nodes have been reached, or some other condition is met, the search is halted. Clearly, this approach is strongly influenced by the choice of initial nodes. It tends to over-represent large connected components relative to islands (smaller connected components) and miss nodes with only outgoing links. These biases also impact the observed connectivity, hop distances between nodes, etc.

Quality of Data Collected. Measurement efforts in OSNs have tended to fixate on quantity rather than quality of data. Certainly, a larger sample gives one greater confidence in statistical measures derived, but there is a danger in assuming that bigger is better. The earliest data sets from sociology. although minuscule by today's standards, were hand-compiled and carefully curated: each node and edge was based on detailed information (interview or questionnaire). Collecting huge volumes of OSN data means that necessarily there is no detailed examination of any portion, and the quality of the obtained data is often unknown. Some problems which can arise in the unexamined data set include:

Dormant entities: many users set up an account to try a service, then abandon it. Simply counting all entities will inflate the number of active users, and skew statistics on usage patterns, interaction density etc. Similar problems arise from a high volume of fake entities (e.g. pets with OSN accounts), or users who set up multiple accounts for spamming or manipulation purposes. In a different guise, the same phenomenon affects other entity types. A news story about a years-old event, or a videoblog entry for months back is effectively "dormant" for most purposes: although a crawl of the network will recover the entity, it plays no significant role in the current state of the network. Hence, these dormant entities would also distort properties of the network.

Design vs. usage: the features of an OSN are designed by the site operators for a particular purpose and with a particular mode of use in mind. However, with millions of users, it is inevitable that these features will be used in ways that are unanticipated. For example, many of Twitter's "core features" (hash tags, retweets) are the formalization of previously unsupported conventions adopted organically by Twitter users. A common feature of OSNs is to allow the user to specify a (personal) website along with their other contact information–but younger users find the concept of a personal webpage to be alien, so instead populate this with a favorite webpage of a popstar. Naive statistics about this field would be meaningless without appreciation of this fact. Perhaps most relevant is that while many OSNs use the common terminology of "friend" to allow individuals to link up to other entities, the semantics of these links vary wildly across OSNs, and across individuals. Some use this feature to bookmark interesting entities, while others only for

their "true" friends. In MySpace, the notion is particularly blurred, since entities representing bands and movies are not distinguished from entities representing individuals, and the common notion of "friend" is used to connect entities from these disparate categories. Certainly, it is not meaningful to compare statistics on distribution and variation in patterns of friend links across OSNs given the wide variation in semantics attached to this term.

Site Redesigns. OSNs frequently radically change the way they appear to their users, to reflect current usage and design trends. Often this is more than just cosmetic: new features may be added, or unpopular ones dropped; informal conventions may be adopted as supported features. Longitudinal studies which span such redesigns, or which compare snapshots from before and after, should try to use awareness of the structural changes. In the extreme case, a redesign may dramatically alter the features, causing certain entity and link types to be added or removed.

2.3 Ambient Factors

Access Methods. While the first generation of OSNs could be accessed in only one way (via a web interface), the current generation has an increasingly disparate set of access methodologies. The rapid growth in mobile networked devices (cell phones and smartphones) has prompted OSNs to enable mobile access to their services. In particular, this means alternate views of the website which present information differently for small-screened devices with lower bandwidth. Additionally, numerous applications (either created by the OSN or by third-parties) use API functionality to access the network on platforms such as iPhone.

This has substantial impact on measurement: since the user experience varies significantly depending on the access modality, studies should take these factors into account. In some networks, looking only at website visits would vastly underestimate the amount of activity in the network, since the majority of users engage via other avenues[1]. Some bandwidth-intensive activities, such as uploading video, are only permitted (or feasible) from a broadband connection, so this knowledge is needed to understand differences in user behavior.

Lifetime of Data. As we have emphasised the importance of modeling the dynamic nature of data, so it is important to characterize the relevant lifetime of particular objects within the OSN system. Prior work has characterized the typical useful life of a webpage in terms of years, while life of a blog post is measured in weeks. In contrast, most "status updates" in OSNs have a useful life of a matter of days or hours. Therefore, statistics based on "all time" activity (total number of status updates, say) do not show a useful picture: instead, rolling statistics (based on a sliding window or exponential decay) are needed, where the window-size/half-life depends on the relevant useful lifetime.

[1] See e.g. `http://mashable.com/2010/02/10/twitter-tweet-volume/`

Cross-OSN activity As more OSNs open up, and allow information flow into and out of the network, it is no longer sufficient to look at a single OSN in isolation. For example, many Facebook users update their FB status via Twitter; other services (such as friendfeed) aim to aggregate information from multiple networks. Further, a new generation of OSNs are built on top of existing OSN systems: the underlying OSN provides identity management and other functionality, while the overlay OSN provides custom additional functionality for a particular interest or purpose.

It is therefore important to recognize this richer environment. Modeling it accurately is virtually impossible in the naive graph model, but can be easily incorporated in the EIN setting: different node types represent entities from different sites. However, identifying the same individual within different networks remains a research problem. Some individuals explicitly list their identities across multiple OSNs, but relying on this information alone skews towards those who particularly want to make these connections known (i.e. those who are actively trying to recruit "followers").

2.4 Skewness of Features

Early OSN studies, while rarely explicitly bringing dynamism into their data models, nevertheless observed high variability and skewness in properties of the networks. These properties have implications for measuring, modeling and analyzing OSNs.

Skewness in network structure. OSNs are commonly described as a "core" of relatively highly connected users and "islands" of varying sizes (including a large number of isolated nodes) made up of few loosely connected users. With temporal features available, it is possible to ask whether this core-islands decomposition holds constant and how islands evolve over time— whether they split off from the core, merge into it, or remain isolated from it. Any such core-islands decomposition of OSNs has great potential for efficient and effective measurements by restricting attention to the core portion of the OSN structure. However, it is not appropriate to ignore the islands when they represent a large fraction of users (as in YouTube under the friends relation).

Skewness in user activity. In addition to the core-islands property of the structures of OSNs, often a few users in the core of the network are responsible for a large portion of the overall interactions. If or how this highly skewed nature of user activity holds over time (do the "power users" change over time, and at what rate? what determines whether a new users will join this group?) remains a topic for further study. The measurement implications of validated high skewness in user activity in evolving OSNs are promising and far-reaching.

Skewness in traffic. In cases where user interactions produce measurable network traffic between users, another observed skewness feature shows that a large portion of OSN traffic is due to a small percentage of highly active users of the system. How users become such "heavy hitters" or move from highly active to essentially insignificant as far as traffic is concerned looms as an interesting open problem. The potential for exploiting this skewness property for measurement, modeling and analysis of global traffic aspects of OSNs is clearly very high.

3 Case Studies

3.1 Facebook

Early studies of Facebook (when still restricted to college users) focused on the structure of social networks of students at particular US universities. They examined the graphs of Facebook friendships, where nodes are users/students, and links represent reciprocated "friendship" relationships and are hence undirected. [Traud et al., 2009] use such graphs as well as profile-based information to infer the community structure of the social networks of students at different universities. The appeal of such single-institution datasets, especially when augmented with factual demographic student information provided by the university, for social scientists is well articulated in [Lewis et al., 2008].

A particular popular topic that has benefited from these new datasets is "tie strength" as a measure for relationships, where "strong" ties are among trusted user and "weak" ties are relationships among mere acquaintances [Granovetter, 1983]. For example, recent studies try to impart meaning to Facebook "friendship" links in campus-wide social networks by examining user interactions; that is, using the inferred "interaction graph" to quantify user interactions. Facebook allows many different interactions types, implying a potentially complex EIN model: a user can interact with another by posting on their wall, sending a private message, initiating an IM chat, or making a comment on a photo or status update. More subtly, they can 'tag' the user in a photo or reference them in a status update. While [Wilson et al., 2009] rely on static interaction graphs inferred from collected data about all users' wall posts and photo comments, [Viswanath et al., 2009] examine the dynamism of user interactions based on wall postings only.

More recent studies considered Facebook-wide questions such as measurement, structure, applications, traffic and performance. [Gjoka et al., 2010] compare crawling and sampling methods for obtaining the social network of Facebook users and characterize the resulting static graph structure using well-known metrics such as degree distribution, assortativity, and clustering. A particularly interesting aspect of Facebook is the popularity and user reach of Facebook applications. Most applications have been developed by third-party developers since Facebook opened their application platform in May 2007. While [Gjocka et al., 2008] base their application characterization on profile-based per-user application information, [Nazir et al., 2008, Nazir et al., 2009] perform an in-depth study of their own applications which pro-

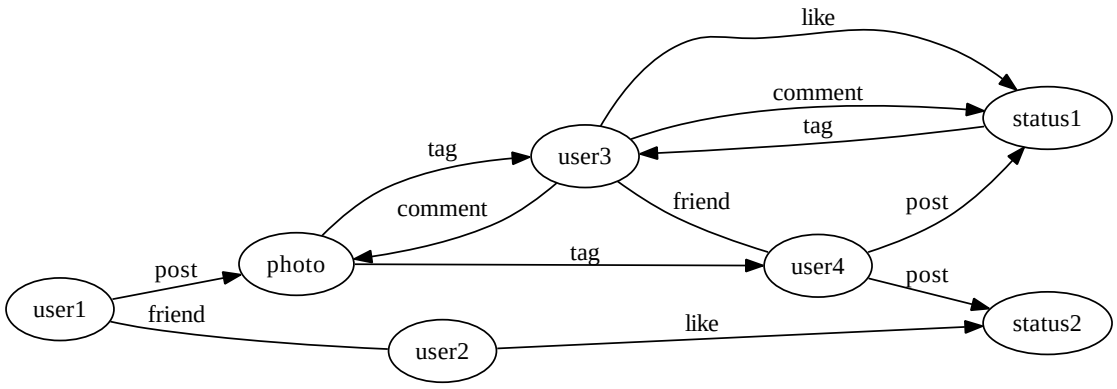

Figure 1: Example EIN for Facebook

vide them with a wealth of data that is in general only available to the application developers (and Facebook). This data allows them, for example, to characterize application workload and performance. The question of which features of Facebook (not accounting for external applications) are popular with the users and capture their attention is studied in [Schneider et al., 2009] using packet traces from large user populations collected at different vantage points within large ISPs.

The collection of these studies are based on datasets obtained by crawling or sampling Facebook and exploiting the presence of an API to extract the desired information[2]. With this API, social context can be extracted or added to an application by utilizing profile, friend, page, group, photo, and event data. A limitation of the current API model is that it is designed primarily to allow development of third party applications and for integration with other sites. Calls to the API execute with the permissions of the user on behalf of which the call is made. This affects the view of the network available to the researcher (perhaps necessarily so, given the privacy settings of other users). Nevertheless, given the richness of API functionalities provided, measurement of Facebook is still in its infancy. In particular, there is high potential to instantiate a rich dynamic entity interaction network with Facebook data. Many critical entities and interactions (with associated timing information) are now measurable quantities thanks to the various API functionalities.

Figure 1 shows a small example EIN for facebook, connecting four users. All of the edges (and nodes) may have timestamps. The example shows that the interactions between users can grow moderately complex. For example, user4 posts a status update, which tags user3, who responds by "like"ing the update, and posting a comment about it. Such interactions are hard to encode in a meaningful way in a simple graph, where nodes and edges do not have types.

[2]See `http://wiki.developers.facebook.com/index.php/API` for documentation on the Facebook API

3.2 Twitter

Twitter, a very popular OSN with over 20 Million users generating over 50 million tweets daily, has a very simple premise: users update their status by generating a (upto) 140 character message, called a tweet. A set of users "follow" them (subscribe to their tweets) and the user can likewise follow other users; the two sets do not have to overlap. For a simple OSN like Twitter, it has proven to be quite popular. There are numerous interfaces through which the tweets can be sent to Twitter and many ways to receive others tweets. A few additional features are built on top of this simple premise: users who approve (or disapprove) of another user's tweet can 'retweet' them or 'favorite' them (with annotations) providing a measure of popularity for the tweets. Such APIs exist to forward the tweets from various devices and programs to Twitter.

Similar to other OSNs, the Twitter "graph" consisted of users functioning as nodes with two kinds of edges: to their followers and the ones they followed. A key difference in Twitter is that there are actually two central objects: the user and the tweet itself. Users decide who could follow them and a class of celebrity users can and do have a large number of followers while they follow a significantly smaller number. Media outlets (such as New York Times) can be Twitter 'users' as can be programs (Remember the Milk) that serve as reminders for users–these also have a very large number of followers but do not follow anyone.

The first studies in Twitter were naturally characterization in nature and began to appear in 2007-2008 shortly after the popularity of Twitter began to be noticed. These studies [Krishnamurthy et al., 2008, Java et al., 2007], beyond introducing the ways by which users can use Twitter, focused on the key properties of both first class objects. The set of questions related to users included the relationship between the counts of followers and following (thus identifying celebrities and broadcasters), the number of tweets, and their geographical distribution. Questions related to tweets themselves included the time of day when they were generated, the preferred manner in which tweets were generated,

The next set of papers examined the tweets themselves to see if linguistic and semantic analysis on them could help in automatically classify them. For example, tweets could be grouped into information related updates (indication of some news event), or as spam [Pear Analytics, 2009]. Attempts to construct "conversations" by grouping related tweets was mixed with identifying trend setting users [Cheong and Lee, 2009] on a specific topic. A more classical sociological approach was taken by creating categories and explicitly looking in tweets for the presence of personally identifiable information [Humphrey et al., 2010].

The interface used to tweet varies with the users and thus passively measuring traffic is much harder. The various characterizations dealing with users and their friends and followers require crawling the Twitter universe. The API provided by Twitter (with rate limitation) has to be used with care. For example, the API returns the list of friends of a user with the most recently added friends, which could bias analysis. The recently available Twitter stream API data only makes a sample of the tweets available. As a privacy measure, Twitter allows users to protect their tweets; the fraction of private tweets is hard to measure.

The set of properties and interests between these two first order objects have a few similarities and

numerous differences. For example, a collection of tweets on a popular topic can be grouped together (with Twitters hashtag feature) so other users could isolate this collection quickly. Yet, there is no direct notion of grouping an arbitrary set of users together. If one wants to construct a list of users who primarily followed celebrities, then a complex set of operations has to be performed. Questions of interest regarding both objects are also of interest: Do users who have a large number of followers generate more or fewer tweets than others? Do users who passively follow a large number of people generate few tweets? Celebrity users who follow few users are considered a valuable presence on Twitter by the company and the users. However, they need to be distinguished from spammers.

3.3 YouTube

The video sharing site YouTube incorporates many social features. Users who set up accounts can upload video content, rate and comment on videos, add users as "friends" or subscribe to the videos produced by users or channels. In IMC'07, all papers in the session on social networking addressed YouTube as a topic of study [Cha et al., 2007, Gill et al., 2007, Mislove et al., 2007]. Yet, one can argue that YouTube is not an OSN in the strong sense of the term: the social features are decidedly secondary to the primary usage, of uploading and viewing videos. Social features like adding friends and subscriptions are by no means core to the user experience. Hence, the video is a first class object in the YouTube ecosystem, while the user is at best a second class object.

The main node type in the YouTube EIN is the video, which has many potential attributes: those determined by the site (the duration, resolution, and bitrate); those determined by the uploader (a summary, category, and tags); and those determined by the actions of other users (an average rating, and number of views). A video may be deleted by the uploader, or by the site (e.g. for copyright or terms of service violations), but otherwise remains "forever". However, measurement studies indicate that the number of views mostly decays with time [Cha et al., 2007, Cheng et al., 2008, Gill et al., 2007]. User accounts provide space for the user to list age, sex, location, and other demographics and interests. The system records when the user joined and last visited.

Between these objects, multiple distinct types of links can exist: a comment is implicitly a link between a user and a video. Although comments remain associated with a video indefinitely, the fact that most recent comments are shown first, and the perception that many comments have low value, argues that they have a very short useful lifetime, and in some cases are "write-only". Between users, there are both friend links and subscription links. The key distinction is that friend links must be reciprocal (and can be removed by either user), whereas subscriptions do not need the agreement of the target. Since friend links have little impact on the user experience of the site, they seem to serve little purpose other than to act as bookmarks for a user's acquaintances. The system also provides various "related" links from one video to another, in various forms (as a sidebar, and after the video has completed). It is unclear how these links are chosen (or how they vary

over time): clearly, user behavior and keyword similarity play some part in setting them.

Three initial studies of YouTube addressed disparate aspects of the ecosystem. [Cha et al., 2007] studied video viewing distributions as a function of category and age of video, as well as duplication and source of video (whether the uploader was not the copyright holder). Their measurements crawled all videos in particular categories. [Gill et al., 2007] captured data by recording all YouTube related traffic observed on a campus network, and correlated this with data on popular videos from the site. Clearly, one cannot extrapolate usage from one campus to obtain global behavior; more seriously, the spread of mobile devices can radically change the interaction patterns with a site like YouTube (e.g. the iPhone, for which YouTube transcoded videos to h.264 format since the OS does not support flash video). [Mislove et al., 2007] considered YouTube in conjunction with other sites (Orkut, Flickr, LiveJournal) and adopted a uniform simple graph model of users and friend links to study degree and path length distributions. Although the conclusion was that the graph appeared to posses power-law and small-world characteristics, it is unclear how significant this is, due to the relative unimportance of friend links in YouTube.

Subsequent studies have followed similar approaches. [Zink et al., 2009] also studied YouTube traffic on a campus network, to identify the potential for caching in the network. [Cheng et al., 2008] made a breadth-first crawl of the simple graph formed by videos and related links. Measurements were also made of friend links, observing that the majority of users do not have such links. So while these studies looked at various subgraphs of the interaction data, each modeled as a simple graph, no study has yet taken in the "big picture" of YouTube.

Previous studies have focused on crawling and packet sniffing to make measurements. However, the API now allows measurement of many parameters of the site. Some aspects of the site can only be measured by longitudinal studies: monitoring a video's views over time, for example, requires repeated measurements. Other aspects are outside the scope of the API: determining the bitrate requires analysis of the video object, and determining copyright issues may require human judgment. It does not currently seem feasible to measure the breakdown of viewing patterns across different countries, between mobile/static clients, or whether viewing was on-site or embedded in another site.

4 Discussion

4.1 Mobile Social Networks

As new networks emerge, we must ask whether the models and measurement techniques discussed are adequate to capture information about them. A case in point is the arrival of "mobile social networks", which are as yet unstudied by the research community. These networks rely on location-aware mobile devices, where the user's current location forms an integral part of the functionality: they allow friends to track each other's locations (and meet up), and play location-aware games. The EIN model can certainly

be extended to include the location of entities at a given time to be modeled and recorded. However, it may require new insights to specify how best to use this information to answer questions, such as how user movements change over time, to what extent do friends use the mobile social network information to arrange a meetup, and so on.

4.2 Afterlife of data

Researchers who have put in the time and effort to collect social network data for a particular project are often able to share it with others for new studies. If the latter differs from the purpose for which the original data was collected, it is critical to explain why the data can be used for a different purpose. In this case, it is doubly important to document the modeling performed, and the details of the collection method. Otherwise, it is too easy for the recipients of the data to draw erroneous conclusions, or find structures which are merely a facet of the collection method.

Because of the personal nature of information in social networks, some form of anonymization may be applied to the data. However, anonymization is still an active research area, and there are no standards for the process. Putting aside concerns about the possibility of breaking privacy and reidentifying individuals, it is important to be explicit about the anonymization method used. Most anonymization introduces noise, and possibly modifies the link structure of the network. Therefore, researchers have to be sure that their conclusions represent the true state of the network, and are not artefacts of the anonymization process.

4.3 Validation and the EIN framework

When modeling OSNs as simple static graphs, model validation typically reduces to showing that the proposed model matches a certain statistic (e.g., node degree distribution) of the available data. See [Krishnamurthy and Willinge for a critique of this process. It should be obvious that in the framework of EINs, model validation has to mean more, and this alone makes model validation for EINs an interesting problem in its own right.

The complexity of EINs as expressed by the modelers' choices of node and link types and attributes will require an approach to model validation that moves beyond matching individual statistics. As argued in [Krishnamurthy and Willinger, 2008], the model should identify new OSN-specific quantities or attributes that played no role whatsoever in the original definition of the model, are measurable, and whose properties can be predicted by the model. If the newly measured data agrees roughly with the model prediction and does so for a number of relevant and informative quantities, we view the model to be a valid representation of the OSN at hand as well as for the purpose for which it is intended to be used.

Note that such a first-principles approach to model validation directly impacts the processes of OSN measurement and modeling. By identifying and measuring features or attributes omitted in the initial EIN framework, we augment our above-mentioned ad-hoc recipe for defining an OSN-specific EIN with a more

quantitative and purposeful algorithm.

4.4 Recommendations for OSNs

Presently, researchers try to get as much information out of OSNs as possible, while OSNs try to limit the crawling activities by researchers through rate limiting the number of queries or controlling the number of OSN features accessible via the OSN-specific APIs. A key step moving from this antagonistic scenario to a more collaborative relationship is to increase the "economic value" of individual measurements. For example, by ensuring that a given API gives access to essentially all relevant OSN features (e.g., friends, activities, traffic, timing information), OSNs would likely experience a drastic *drop* in queries because the semantic context of each obtained measurement would be higher and yield the desired information. By making it easier to get the desired information, OSNs could reduce the number of "uneconomical" queries and spend less effort trying to block measurement efforts.

The designers of Internet protocols made the mistake of failing to build-in measurement as an integral part of system design. The first generation of OSNs have followed this example, and do not appreciate that enabling effective measurement can improve the understanding and use of their product. However, the power of opening up the API to third-parties for application building has been realized by the current generation of OSN designers. Since the API is not set in stone, there is hope to persuade OSNs to add more support for measurements into the next-generation APIs. Whether or not this development will create a win-win situation for researchers and OSNs alike remains to be seen.

References

[Cha et al., 2007] Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., and Moon, S. B. (2007). I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Internet Measurement Comference*.

[Cheng et al., 2008] Cheng, X., Dale, C., and Liu, J. (2008). Statistics and social network of YouTube videos. In *International Workshop on Quality of Service*.

[Cheong and Lee, 2009] Cheong, M. and Lee, V. (2009). Integrating web-based intelligence retrieval and decision-making from the Twitter trends knowledge base. In *ACM workshop on Social web search and mining*.

[Gill et al., 2007] Gill, P., Arlitt, M. F., Li, Z., and Mahanti, A. (2007). YouTube traffic characterization: a view from the edge. In *Internet Measurement Comference*, pages 15–28.

[Gjocka et al., 2008] Gjocka, M., Sirivianos, M., Markopoulou, A., and Yang, X. (2008). Poking facebook: Characterizationof OSN applications. In *Workshop on Online Social Networks*.

[Gjoka et al., 2010] Gjoka, M., Kurant, M., Butts, C., and Markopoulou, A. (2010). A case study of unbiased sampling of OSNs. In *Infocom*.

[Granovetter, 1983] Granovetter, M. S. (1983). The strength of weak ties: A network theory revisited. *Sociological Theory*, 1:201–233.

[Humphrey et al., 2010] Humphrey, L., Gill, P., and Krishnamurthy, B. (2010). How much is too much? privacy issues on Twitter. In *International Communication Association Conference*.

[Java et al., 2007] Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we Twitter: Understanding microblogging usage and communities. In *KDD*.

[Krishnamurthy et al., 2008] Krishnamurthy, B., Gill, P., and Arlitt, M. (2008). A few chirps about Twitter. In *ACM SIGCOMM Workshop on Online Social Networks*.

[Krishnamurthy and Willinger, 2008] Krishnamurthy, B. and Willinger, W. (2008). What are our standards for validation of measurement-based networking research? In *HotMETRICS Workshop*.

[Lewis et al., 2008] Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., and Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using facebook.com. *Social Networks*, 30:330–342.

[Mislove et al., 2007] Mislove, A., Marcon, M., Gummadi, P. K., Druschel, P., and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Internet Measurement Comference*, pages 29–42.

[Nazir et al., 2009] Nazir, A., Raza, A., Gupta, D., Cuhuah, C.-N., and Krishnamurthy, B. (2009). Network level footprints of Facebook applications. In *Internet Measurement Conference (IMC)*.

[Nazir et al., 2008] Nazir, A., Raza, S., and Chuah, C.-N. (2008). Unveiling Facebook: A measurement study of social network based applications. In *Internet Measurement Conference (IMC)*.

[Pear Analytics, 2009] Pear Analytics (2009). Twitter study. `http://www.slideshare.net/stephendann/twitter-analytics`.

[Schneider et al., 2009] Schneider, F., Feldmann, A., Krishnamurthy, B., and Willinger, W. (2009). Understanding online social network usage from a network perspective. In *Internet Measurement Conference (IMC)*.

[Traud et al., 2009] Traud, A. L., Kelsic, E. D., Mucha, P. J., and Porter, M. A. (2009). Community structure in online collegiate social networks. In *APS March Meeting*.

[Viswanath et al., 2009] Viswanath, B., Mislove, A., Cha, M., and Gummadi, K. (2009). On the evolusion of user interaction in facebook. In *Workshop on Online Social Networks*.

[Wilson et al., 2009] Wilson, C., Boe, B., Sala, A., Puttaswamy, K., and Zhao, B. (2009). User interactions in social networks and their implication. In *EuroSys'09*.

[Zink et al., 2009] Zink, M., Suh, K., Gu, Y., and Kurose, J. (2009). Characteristics of YouTube network traffic at a campus network - measurements, models, and implications. *Computer Networks*, 53(4):501–514.