

# Representation of Online Social Networks

Samantha Lam

s0790004

HCI Literature Review/Report

Due: Friday, 7th December 2007

## Abstract

Since the foundation and continual growth of MySpace in 2003, many new social networking services, such as Facebook, Bebo etc., have been developed and grown at a similar rate. The underlying structure of these networks are necessary to evaluate current systems, to design future similar systems and to understand the impact of them on the Internet.[18] We look at recent concrete analyses of such networks which use data retrieved from web-crawling and discuss what the most important features of the graphs obtained from these networks would be if we were to design an interface for extracting information about a particular social network.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	History and Motivation . . . . .	2
<b>2</b>	<b>Social Network Analysis</b>	<b>2</b>
2.1	Methods of Analyses . . . . .	2
2.2	Drawbacks . . . . .	7
<b>3</b>	<b>Discussions</b>	<b>8</b>
3.1	Summary . . . . .	8
3.2	Graphical Representation . . . . .	9
<b>4</b>	<b>Bibliography</b>	<b>10</b>

# 1 Introduction

## 1.1 History and Motivation

Early social networking websites included Classmates.com (1995), focusing on ties with former school mates, and SixDegrees.com (1997-2001), focusing on indirect ties. By 2005, one social networking service MySpace, was reportedly getting more page views than Google, with Facebook, a competitor, rapidly growing in size (in fact, at an exponential rate and has recently overtaken MySpace in terms of web traffic[1]). In 2007, Facebook began allowing externally-developed add-on applications, with some applications enabling the graphing of a user's own social network - thus linking social networks and social networking.[21] One of the applications of this type that is most akin to the nature of this report called the 'Friend Wheel' developed by Thomas Fletcher displays a 'wheel' with nodes representing people in one's network and lines (or edges in graph theory terms) connecting those who are 'friends'. However, there is a limit of 400 nodes due to the speed of the algorithm and the graphical representation appearing cluttered.[6] Thus, amongst our design for suitable criteria in characterising online social networks, finding a suitable representation for such large graphs of networks is also an issue.

## 2 Social Network Analysis

### 2.1 Methods of Analyses

We begin with a brief discussion from Jamali and Abolhassani's paper "Different Aspects of Social Network Analysis"[12] which presents some important properties of social network in general. In their terms a social network 'is a set of people (or organizations or other social entities) connected by a set of social relationships, such as friendship, co-working or information exchange.' They begin by giving three conventional social network models such as using graphs and matrices to represent social relations and statistical ones for analysis. However, there are associated problems - in the case of statistical models, problems such as degeneracy analysed by [8] and scalability as discussed in [10] and [19]. Whereas for both the graph and matrices representation, they have the problem of decreasing readability with increas-

ing number of nodes and connections. We will discuss in detail the possible benefits of a graph representation despite this disadvantage later.

Most relevant are the properties and substructures in social networks that Jamali and Abolhassani discuss in terms of actors as nodes and the edges as a certain relationship between two actors who are joined by them. The following table succinctly characterises what we consider to be the most important aspects in social networks because of the frequency of these definitions([12], [9]) found in the literature (in brackets are their common names outside of Jamali and Abolhassani’s article):

### **Some Properties of Social networks**

<b>Name</b>	<b>Definition/Description</b>
Degree (of a node)	Number of ties for an actor
Closeness (path length)	Lengths of paths to other actors
Betweenness	Lying between each other pairs of actors
Clique (hub/clusters)	Actors who have all possible ties among themselves
N-Clique (hub/clusters with diameter N)	Actors are connected to every member of the group at a maximum distance of N
Component (core)	Parts of graph that are connected within but disconnected with other components
Cut Points	Nodes which if removed, the structure becomes divided into un-connected systems
Block	The divisions into which cutpoints divide a graph
Lambda Set	Set of actors who if disconnected, would most greatly disrupt the flow among all of the actors

With these definitions in mind, we examine what research has been done towards the analysis of specific online social networks. A large-scale analysis of this type was done in 2006 by Mislove et al.[18] where they used data from popular social networks Flickr, YouTube, LiveJournal and Orkut. They obtained the data by using automated scripts on a cluster of 58 machines to crawl the social network graphs of said sites. This collection of data is representative of late 2006/early 2007 and although they are not complete (which is impossible, since online social networks are continuously growing), Mislove et al. give clear arguments using statistical analyses for why their

data is representative of the networks. However, awareness is also made with regard to their data for Orkut. The crawling Mislove et al. did for the other three networks was through those networks' API whereas with Orkut they used HTML screen-scraping using a breadth-first search which has associated problems of over- and under-sampling particular nodes (see [4] and [16]). Furthermore, some of the results from the Orkut data which conflicted with the other three networks reflected this under-representation.

For their analysis of the network structure they look at eight characteristics of the networks - link symmetries, power-law node degrees, correlation of in and outdegrees, path lengths and diameter, link degree correlations, densely connected cores, tightly clustered fringes and groups. They also use the internet (used interchangeably with the Web) as a means of comparison because of its obvious similarities (online social networks being a subset of it).

Link symmetry is the implication that if page A has a link to page B then page B also has a link to page A. With the exception of Orkut the networks are directed since users may link to any other user they wish. So in the case of the other three networks, Mislove et al. found that these networks have a high degree of symmetry. Regardless of the cause of symmetry, it is known that this property affects the overall network structure, e.g. the overall connectivity of the network increases and its diameter increases. Thus, knowing that a network is highly symmetric already gives one an idea of what to expect in terms of values of diameter etc.

Next, Mislove et al. found that the degree distributions of the online social network's (barr Orkut) follow a power-law, and that the power-law coefficients for both in-degree and out-degree are similar. Also, nodes with high in-degree also tend to have high out-degree. *Power-law* networks are networks where the probability that a node has degree  $k$  is proportional to  $k^{-\gamma}$ , for large  $k$  and  $\gamma > 1$ . [18] It has been shown that many real-world networks such as the internet exhibits a power-law distribution [3],[15] as well as offline social networks [2] and so observing this distribution in online social networks is a correlatable indicator. However, the correlation between the in and out degree found in online social network's doesn't correspond to the structure of the internet in which 5% of the Web nodes account for 75% of all incoming links but only 25% of all outgoing links, whereas for the online social networks over 50% of nodes have an in-degree within

20% of their out-degree. So the combination of these two factors (degree distribution and correlations of in & out degree) can help distinguish a social network from the Web.

Mislove et al. also found that all four online social networks appear to be composed of a large number of highly connected clusters consisting of low-degree nodes and that these clusters connect to each other through a small number of high-degree nodes, i.e. short path lengths and diameters. An observation of high-degree nodes tending to connect to other high-degree nodes also suggested a formation of a ‘core’ in the networks (found in all but YouTube). Moreover, they found that each network have a densely connected core which is held together by  $\sim 10\%$  of the nodes with highest degree (from the formation of groups which was found to exist in the way such that low-degree nodes tend to be part of very few groups, while high-degree nodes tend to be members of multiple groups[18]).

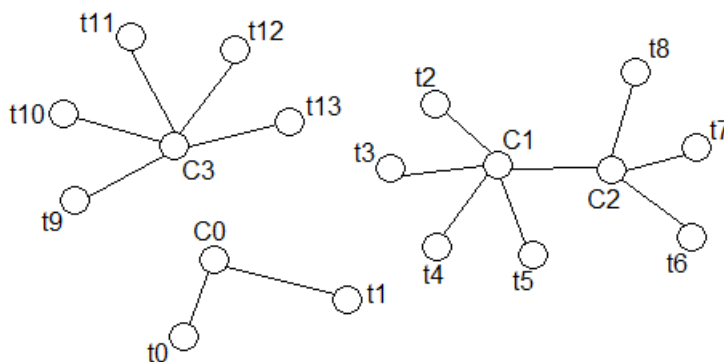
Overall, the findings from Mislove et al. reveal that online social networks have a rich, distinct structure from the internet with similarities to ‘real’ social networks. In particular, the properties with which they analysed the networks may be calculated on an interface where one inputs the data of a certain network and a deduction of whether it is an online social network is made may be possible. However, this shall be elaborated in the Summary/Discussions section. Furthermore, there is an incompleteness associated with this study - the aspect of time is not taken into account, lest not fully. Their incorporation of time was done simply by recollecting and recalculating the data five months after their initial study which confirms a robustness of their results, however, this doesn’t give any insight into the growth of an online social network which is a phenomenon in itself. So, in addition to Mislove et al.’s properties, we look at incorporating the notion of time as part of an online social network’s features.

Kumar et al.’s paper ‘Structure and Evolution of Online Social Networks[14] study the evolution of Flickr and Yahoo! 360°. Their data involved the Flickr timegraph consisting of  $\sim$ one million nodes and  $\sim$ eight million directed edges and the Yahoo! 360° timegraph consisting of  $\sim$ five million nodes and  $\sim$ seven million directed edges (approx.  $\frac{1}{3}$  of users of Flickr and  $\frac{1}{10}$  of Yahoo! 360°). In both cases, Kumar et al. discarded the initial segment of the timegraphs to filter out pre-launch bias (such as initial limitations to internal users). They also similarly observe the frequent existence

of link symmetry, i.e. reciprocal friendship, and so analysed the networks' graphs as undirected.

Their first interesting finding is the answer to how the density of online social networks behave over time; non-monotonically. Results show that there are three stages: an initial upward trend leading to a peak, followed by a dip, and the final gradual steady increase.[14] Furthermore, according to Kumar et al. this predicament has not been found before in real social networks. The power-law distribution is again observed in the networks tested which confirms this distribution as a trait of online social networks. The paper next focusses on the dynamics of component formation and the evolution of them in the networks in terms of the *singletons*, *middle region* and *giant components* which are nodes with zero-degree, small hubs which are connected locally but not with the network at large and the large core which are connected to a large fraction of the entire network through paths in the social network, respectively.

Kumar et al. find that the component merges on arrival of new nodes and edges are of effectively two types: either singletons merging with current non-giant components or giant, non-giant components and singletons merging with the giant component. More surprisingly, this implies that it is rare for two non-giant components to merge as one non-giant component over time. For further analysis, Kumar et al. also give a definition of a star of which we shall describe in intuitive terms with the following diagram rather than the formal definition which can be found in [14].



Each component seen is a star,  $C_i, i = 0, 1, 2, 3$  are the centres and  $T_i, i = 1, \dots, 13$  are the twinkles. Based on this approximate definition of a star, they find that 92.8% and 88.7% of the middle regions of the Flickr and Yahoo! 360° graphs are composed of stars and that a large fraction of these

stars are more than 10 weeks old suggesting that these stars don't get absorbed into the giant component as quick as they form. Further analyses and the overall result show that online social networks (or at least, in the case of Flickr and Yahoo! 360°) often contain more than half their mass outside the giant component, and the structure outside the giant component is largely characterised by stars[14]. The creation of stars emulate the dynamics of invitation in an online social network and thus is of interest for calculation and representation to provide information about the dynamics of an online social network.

Stars represent a structure of the middle region network, but give little insight into the giant component itself. Kumar et al. thus analyses the giant component in terms of its average and effective diameter and how they behave over time. The average diameter is defined as the length of the shortest path between a *random pair* of nodes whereas the effective diameter is the 90-th percentile of the shortest path lengths between *all pairs* of nodes. The results of their findings is that for both of the online social networks, the behaviour is that there are 3 stages akin to the density; the first is a flat diameter, next is an increase in diameter when the edge density drops and finally, on reaching a peak, the diameter decreases. However, it is known that models of networkgrowth based on preferential attachment do not exhibit this property[5] and notably, the Web is a model of this type.

## 2.2 Drawbacks

So far we have found and discussed several major properties of online social networks that appear promising as representable criteria. However there may be associated issues with the availability of such data. First, as highlighted in [12], the source of the data presented - to quote: 'the information is under the control of the database owner who has an interest in keeping the information bound to the site...and centralized systems do not allow users to control the information they provide on their own terms'. The data collected in our discussed papers are (presumably) not directly from the databases of the online social networks and even if so, there still is the possibility of the database owner manipulating or withdrawing some data.

Secondly, having social network aspects available may lead to susceptibility for exploitation. As discussed in [18], again to quote: 'In a social network, the underlying user graph can potentially be used as a means to

infer some level of trust in the unknown user[17]'. Unmoderated analyses of data collection may assist in the exploitation of such trust patterns.

## 3 Discussions

### 3.1 Summary

The most important of aspects of online social networks have been presented through two of the most recent and large scale studies of such networks. Much of the criteria derived from these two papers have come from 'offline' social networks and some of their important distinguishing features have been confirmed. In particular, the fact that data for these online social networks are more readily available than data from 'real' social networks gives rise to the possibility of large-scale analysis. Additionally, given the growth (and popularity) in number of online social networks, it may be a useful tool to be able to 'see'/characterise what stage of evolution a new online social network may be as a method of determining whether it will progress.

Thus, the interface design proposed would be a software package which allows one to input suitable data for analysis and display values for the link symmetry, degree distribution, density of the core and existence of groups with the explanations of each these terms (possibly in terms proposed in [13]). An explanation which relates to a 'control' version of an online social network and the Web may also be incorporated for comparison purposes. Furthermore, evolution of the density of the network and its diameter may be implemented as an interactive evolving graph over time with the possibility of presentation either together (scaled appropriately) or separately. Additionally, graphical representations of the giant component and middle region (highlighting the stars) would be indicative of the online social network.

To summarise, there are consistent theories about social networks specifically, and the correlation of these theories with online social networks appears to prove promising as displayed throughout this report. Due to the availability of data for analysis from the Web, it would be useful to be able to analyse this data in a more standard format. However, before this can be done, more research needs to be done in the area of graphical representation which we shall briefly mention.

## 3.2 Graphical Representation

There has been significant research in the area of social network visualisation, however, it is beyond the scope of this report to discuss in detail. We can only provide informative resources, as a result. In particular, Freeman's review in the history of image use in social network visualisation gives a thorough overview [7]. Similarly, Huang et al.'s 'How People Read Sociograms: A Questionnaire Study' [11] would be useful when deciding on a particular graphical representation. The only suggestion one can make here in the context of the online social network graphs is the possibility of displaying specific clusters of the network as opposed to whole network as a naïve way of overcoming the 'cluttered' graph representation problem.

## 4 Bibliography

### References

- [1] <http://www.alexametric.com> Traffic History Graph for facebook.com
- [2] Adamic LA, Buyukkokten O, Adar E, “A social network caught in the Web”, *First Monday*, vol. 8, no. 6, 2003.
- [3] Barabási A-L, Albert R, “Emergence of Scaling in Random Networks”, *Science*, vol. 286, no. 5439, p. 509-512, 1999.
- [4] Becchetti L, Castillo C, Donato D, Fazzone A, “A Comparison of Sampling Techniques for Web Graph Characterization”, *in Proceedings of the Workshop on Link Analysis (LinkKDD '06)*, Philadelphia, PA, USA, August 2006.
- [5] Bollobas B, Riordan O, “Mathematical results on scale-free random graphs”, p. 1-37, Wiley-WCH, 2002.
- [6] <http://thomas-fletcher.com/friendwheel/faq.php>
- [7] Freeman LC, “Visualizing Social Networks”, Carnegie Mellon, *Journal of Social Structure*, Visualizing Social Networks 2005.
- [8] Handcock M, “Assessing degeneracy in statistical models of social networks”, Working Paper no. 39, Center for Statistics and the Social Sciences, University of Washington, 2003.
- [9] Hanneman RA, Riddle M, *Introduction to social network methods*. Riverside, University of California, USA, Riverside ( published in digital form at <http://faculty.ucr.edu/hanneman/> ),2005.
- [10] Hoff P, Raftery A, Handcock M, “Latent space approaches to social network analysis”, *Journal of the American Statistical Association*, vol. 97, p. 1090-1098, 2002.
- [11] Huang W, Hong S-H, Eades P, “How People Read Sociograms: A Questionnaire Study”, *ACM International Conference Proceeding Series*, vol. 243, p. 199-206, 2006.

- [12] Jamali M, Abolhassani H, “Different Aspects of Social Network Analysis”, *WI '06*, IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings), p. 66-72, 2006.
- [13] Khan JI, Shaikh, S, “Relationship Algebra for Computing in Social Networks and Social Network Based Applications”, *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, p. 113-116, 2006.
- [14] Kumar R, Novak J, Tomkins A, “Structure and Evolution of Online Social Networks”, *KDD '06*, Philadelphia, Pennsylvania, USA, August 2006.
- [15] Kumar R, Raghaven P, Rajagopalan S, Tomkins A, “Trawling the web for Emerging Cyber-Communities”, *Computer Networks*, vol. 31, p. 1481-1493, 1999.
- [16] Lee SH, Kim P-J, Jeong H, “Statistical properties of sampled networks”, *Physical Review E*, 73, 016102, 2006.
- [17] Lee S, Sherwood R, Bhattacharjee B, “Cooperative peer groups in NICE”, in *Proceedings of the Conference on Computer Communications (INFOCOM '03)*, San Francisco, CA, March 2003.
- [18] Mislove A, Marcon M, Gummadi K, Druschel P, Bhattacharjee B, “Measurement and Analysis of Online Social Networks”, *IMC '07*, San Diego, California, USA, October 2007.
- [19] Smyth P, “Statistical modeling of graph and network data”, in *Proceedings of IJCAI Workshop on Learning Statistical Models from Relational Data*, Acapulco, Mexico, August 2003.
- [20] Watts DJ, Strogatz SH, “Collective dynamics of ‘small-world’ networks”, *Nature*, vol. 393, no. 6684, p. 440-442, 1998.
- [21] [http://en.wikipedia.org/wiki/Social\\_Networking\\_Software#History\\_of\\_social\\_network\\_services](http://en.wikipedia.org/wiki/Social_Networking_Software#History_of_social_network_services)