60-510 LITERATURE REVIEW AND SURVEY

Social Network Crawling and Properties Analyzing Using Different Sampling Methods

School of Computer Science

Student Name: Hao Wang

Course: 60-510

Instructor: Dr. Richard Frost

Supervisor: Dr. Jianguo Lu

ABSTRACT

Today's social networking services have tens of millions of users, and are growing fast. Also the properties of online social networks are of great interest to general public as well as IT professionals. Often the raw data is not available and the summaries released by the service providers are sketchy. Thus good crawling strategies and sampling methods are needed to reveal the hidden properties of the underlying data. In this survey, a review is given of research on the complete process from crawling raw data, getting samples to showing properties of online social networks. The properties identified include the macroscopic view of the OSN such as the population of the network, degree distribution, clustering coefficients and assortativity. More specially, the survey provides 1) different crawling strategies to the social networks, 2) the sampling methods applied in different situations, 3) the general properties of the OSNs.

Keywords: Online Social Networks, Crawling Strategy, Sampling Method, Topological Property

Contents

1. INTRODUCTION

The increasing popularity of online Social Networks (OSNs) is witnessed by the huge number of users acquired in a short amount of time: some social networking services now have gathered hundreds of millions of users, e.g. Facebook, MySpace, Twitter, etc. The growing accessibility of the Internet, through several media, gives to most of the users a 24/7 online presence and encourages them to build a solid online interconnection of relationships. With the knowledge of the topology of the network, users can promote their status on the network, so that their information can diffuse more effectively.

This survey is a comprehensive study of research on comparing different crawling strategies in terms of accuracy and efficiency, applying different sampling methods to the networks based on specific purposes and revealing the topological properties of OSNs.

All the papers were identified using Google Scholar. The ten papers are published in journals, conferences and workshop. Seven papers are from Conferences (Ahn et al. [2007], Paolillo et al. [2008], Gjoka et al. [2010], Kwak et al. [2010], Ye et al. [2010], Catanese et al. [2011], Katzir et al. [2011]), two from Journals (Kwak et al. [2006], Lee et al. [2006]) and only one from Workshop (Java et al. [2007]).

In this survey, a review is given of research on a complete process of exploring the general properties from online social networks with respect to crawling, sampling and analyzing. It is based on a detailed review of papers on related topics mentioned above and the survey will summarize the general ideas of those works, including the problem noted, the claims that authors made to improve the performance, the analysis methods used and the experimental results. Section 2 introduces research on how to apply different crawling methods towards social networks with different properties and purposes. When crawling a social network with small size, we have confidence to get complete data from the network, but when applying to a large network with millions of users, we will combine sampling methods with crawling strategies. In section 3, three important sampling methods are discussed and show their advantages in accuracy and efficiency. Section 4 groups the papers which analyze the topological properties (degree distribution, clustering coefficient and assortativity) of OSNs. Section 5 includes the concluding comments and future open research topics and future work in OSNs.

2. DEFINITIONS AND BACKGROUND

2.1 SOCIAL NETWORK

A social network is a social structure made up of a set of actors (such as individuals or organizations) and the dyadic ties between these actors (such as relationships, friendships, connections, or interactions). A social network perspective is employed to model the structure of a social group, how this structure influences other variables, or how structures change over time. The study of these structures uses methods in social network analysis to identify influential nodes, local and global structures, and network dynamics. More specifically, Online Social Networks, e.g. Facebook, MySpace, Twitter, etc. are studied in the field of statistics and graph theory.

Scale-free Networks: A scale-free network is a network whose degree distribution follows a power law.

Directed Graph: A directed graph is graph, where the edges have a direction associated with them. Twitter is a directed graph.

Undirected Graph: A graph for which the relations between pairs of vertices are symmetric, so that each edge has no directional character (as opposed to a directed graph).

2.2 WEB CRAWLER

Network crawling also named as spidering. Many sites, in particular search engines, use spidering as a means of providing up-to-date data. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches.

In our research, any OSNs server can be regarded as a database, the crawlers will automatically construct queries with pre-defined keywords and post the queries to the server and collect the raw data replied by the OSN database server.

There are several different approaches to accessing a social network data, including:

- By queries: Each social network provides searchable interface, either by providing an API or an HTML form. Either way queries are accepted or matched pages are returned.

- By crawling using API: New account or blogger data can be obtained by following the links provided in the current page. For instance, some API can provide links to the followers. What the data can be retrieved is dependent on the API provided.
- Crawling HTML pages and screen scraping: this approach is not restricted by the API interface (the methods can be used, and the data returned) and the daily quota for each machine is more flexible.
- By account ID: In some social network sites such as Facebook, a user can be accessed directly using the account ID. If the account ID is randomly generated, we can obtain a random sample of the bloggers.

2.3 SAMPLING

An OSN often has hundreds of millions of users and billions of messages. The torrent of data is too large to process, yet too critical to ignore. Therefore sampling the network is an obvious choice.

The design of a sampling method depends on how we can access the data and what is the goal of the sampling. In general, there are following methods:

- Random node sampling: nodes are selected independent of the edges. Each node is sampled with certain probability and an edges is included if both ends of the edge reside in the sampled nodes.
- Random edge sampling: edges are selected independent of the nodes.
- Random walk sampling: nodes are selected by following the edges.

2.4 TOPOLOGICAL PROPERTIES OF OSNS

2.4.1 DEGREE DISTRIBUTION

The web, blogosphere, online social networks and human contact networks all belong to a class of "scale-free networks" and exhibit a "small world phenomenon". It has been shown that many properties including the degree distributions on the Web follow a power law distribution.

2.4.2 CLUSTERING COEFFICIENT

In social networks, it is likely that two friends of a user are also friends of each other. The intensity of this phenomenon can be captured by the clustering coefficient C_v of a node v,

defined as the relative number of connections between the nearest neighbors of v. The clustering coefficient of a network is the average C over all nodes.

2.4.3 ASSORTATIVITY

Depending on the type of complex network, nodes may tend to connect to similar or different nodes. For example, in many social networks high degree nodes tend to connect to other high degree nodes. Such networks are called assortative.

3. CRAWLING STRATEGIES

This section reviews the crawling strategies when collecting the raw data from OSNs' database server. In the related surveyed papers, the researchers argued that different crawling strategies may be applied to the different OSNs based on the properties of the data and purposed of the crawl. E.g. we should use HTML crawling method when the service provider put rigorous limitations on API crawling. We should use API crawling method when we want to collect structured data. We should use ID crawling method when the IDs of the network are normative.

The related papers are listed in the ascending order of publication year in Table 1.

Year	Authors	Title	Papers referred to	Major contribution
2008	Paolillo et al.	Structure and network in the YouTube core	none	First paper to analyze the structure of YouTube commenting graph using API.
2011	Catanese et al.	Crawling Facebook for social network analysis purposes	Ahn et al. 2007	First paper to find BFS crawling strategy introduces bias into graph study.

Table 1: Papers involve crawling strategies

3.1 CRAWLING BY API

The API approach is to use the API provided by the vendor. The advantage is that both the code and the returns are more structured. The disadvantage is that it is more restrictive, may be slower, and limited to its daily and IP quota.

Since being launched in December 2005, YouTube has unexpectedly emerged as a major player in video distribution. In Paolillo et al. [2008], the researchers posed two problems:

a) What are the characters of YouTube from a systematic perspective?

b) How and why do people use YouTube (regard YouTube as a graph, and how the people connect with each other)?

As the research went deeper, the authors recognized "friending" was not the only relationship that structured YouTube interaction. They also regarded "commenting" as another important social affordance.

In order to support their ideas, the researchers designed an experiment with four steps:

Step 1: They crawled the data from YouTube using API.

Step 2: They extracted author groups and associated keywords from the raw data.

Step 3: They reconstructed a complete graph by linking the edges between node pairs.

Step 4: They analyzed the structure of YouTube and drew some conclusions.

By following the steps above, Paolillo et al. achieved some goals as they wished:

a) They used YouTube REST API to get 857456 references in the process of crawling.

b) They extracted author groups and associated keywords from the raw data.

c) They achieved YouTube friends' degree distribution.

Compared with other papers, this paper made two significant contributions: one is that they had found YouTube appear to have a social core among authors, it functions like other social networking sites, both in terms of degree distribution and internal structure. Another is they had analyzed the structure of YouTube commenting graph and reported that it follows social affordance.

3.2 CRAWLING BY IDS (MIXED WITH API)

For ID sampling, researchers use the hybrid approach that url probes are first sent to see whether an ID exists. Thereafter valid IDs are collected and feed into the Html or API to collect the details of each account.

HTML approach is more flexible. But the return results are in html format and needs to extract the relevant data from those html pages. In addition, they may require login

information and sometimes other verification methods such as answering simple math question.

As mentioned at the beginning of section 3, when the IDs of the network are normative, for instance, they are all 10-digits numbers, we should use ID crawling.

Catanese et al. [2011] study Facebook whose ID is a 16-digits number which can uniquely identify a user. Thus ID crawling method was applied into the research. However, because of the privacy limits of Facebook, ID crawling method could not be applied individually. The authors combined ID crawling with API crawling method.

After getting the data from Facebook, the researchers posed two problems: 1) is it possible to study OSNs' properties without having access to their complete dataset? 2) What is the difference of the Facebook graph when applying different sampling methodologies?

First of all, the researchers review some previous works and they refer to Ahn et al. [2007]. They state that the graph studied in this paper was small and it was not representative.

Then they proposed new ideas to tackle the problem: 1) they introduced a new concept of data cleaning which is an important step in a sampling task, 2) they used large component as a strong connected graph to study the properties of Facebook.

Through the experiment, the authors claim to have proved that uniform dataset was more accurate than BFS dataset. Moreover, even if BFS introduces bias into graphs, the uniform datasets can explain the network perfectly.

4. SAMPLING METHODS

This section reviews different kinds of sampling methods, and shows the pros and cons of each sampling algorithm.

Classification	Year	Authors	Title	Papers referred to	Major contribution
All covered	2006	Lee et al.	Statistical properties of sampled networks	none	First paper to combine OSNs' analysis with statistics.
Random Walk Sampling	2006	Kwak et al.	Impact of snowball sampling ratios on network characteristics estimation: A case study of cyworld	Lee et al. 2006	A new approach to evaluate snowball sampling method.
	2010	Gjoka et al.	Walking in facebook: A case study of unbiased sampling of osns	Henzinger et al. 2000 Lee et al. 2006 Ahn et al. 2007	First paper to introduce convergence diagnostics into sampling social networks.
Random Node Sampling	2010	Ye et al.	Crawling online social graphs	Lee et al. 2006 Ahn et al. 2007	First paper to introduce quantification into evaluating crawling bias.
Random Edge Sampling	2011	Katzir et al.	Estimating sizes of social networks via biased sampling	Lee et al. 2006	A new algorithm outperforms uniform sampling.

Table 2: paper involve sampling methods

In table 2, we can see that the paper Lee et al. [2006] gives fundamental study of three mainstreamed sampling methods from the perspective view of statistics, all the rest papers follow its ideas and introduce some other techniques into the papers to improve the sampling algorithms.

In Lee et al. [2006], the researchers wanted to solve the problem of how to sample large complex networks to reveal the properties. They state that they introduced three kinds of sampling methods called node sampling, link sampling, and snowball sampling. And they used these sampling methods to calculate the degree distribution, the assortativity and the clustering coefficient.

To reveal the properties of the network, the authors designed an experiment with three steps:

Step 1: Crawling raw data from PIN, Internet AS and arxir.org.

Step 2: Using three sampling methods to sample these networks.

Step 3: Applying the samples to calculate the degree distribution, the assortativity and the clustering coefficient.

Through the experiment, the authors claimed to have obtained three tables representing the degree distribution, the assortativity and the clustering coefficient of the sampled networks. And these properties provide criteria of specific quantities are supposed to be investigated by the sampling.

After analyzing the results of the experiment, the researchers state that they had made two contributions:

1. They combined the analysis of the network with statistics and showed the sampling methods in quantification.

2. They provided mathematical analysis to explain some characteristics patterns in changes of the quantities based on the properties of different sampling methods.

In the following sub-sections, one or two papers are reviewed to show the use of each sampling method.

4.1 RANDOM WALK SAMPLING

There are many variants of the sampling method based on following the edges. The overall idea is to explore the next node in the vicinity of the current node. The variations come when we decide how many nodes to select in the next step, how to choose the next step with same probability or different probability depending on some measurement, and what we can do when the walk is stuck in a dead end and loop.

It is known that snowball sampling is more efficient than other sampling methods. However, the disadvantage of this method is also obvious for it is not easy to determine how large the dataset could be a representative sample. So in Kwak et al. [2006], the researchers want to know how the snowball sampling method performs in estimating topological characteristics of social networks and how the sampling ratios affect the estimating results.

At the beginning of the paper, the authors reviewed some previous works and pointed out that in Lee et al. [2006], the size of the networks they have analyzed, including the Internet at the autonomous systems (AS) level, is two to three orders of magnitude smaller than today's commercially available SNSs. Thus it is necessary to apply sampling method in analyzing the OSNs.

In order to reveal the relation between sample properties with the estimating results, the authors introduced a factor called snowball sampling ratio and compare the complete topology (including degree distribution, clustering coefficient and assortativity) of Cyworld (an OSN in South Korea) to sampled networks by varying sampling ratios.

After getting the complete graph of Cyworld and analyzing different properties of Cyworld including degree distribution, clustering coefficient and assortativity, they showed how those topological properties change while varying sampling ratios.

Then Kwak et al. summarized three achievements from the results: firstly, they had obtained a figure showing degree distributions from sampled networks of Cyworld under different snowball sample ratios. Secondly, they had obtained a figure showing degree correlation from complete network. Thirdly, they said that the sampled networks from snowball sampling were more likely to be clustered than the complete network.

The contribution of this paper is lying in the methodological approach in the evaluation of the snowball sampling methods with Cyworld.

Since there are many different node selection algorithms, the random walk sampling method branches into many specific algorithms. In Gjoka et al. [2010], the authors proposed Metropolis Hasting Random Walk sampling method to solve three problems:

a) What are the differences between four candidate crawling methods (BFS, RW, RWRW and MHRW)?

b) When applying convergence diagnostics to analyze the properties, how to control the iterations to get accurate results in terms of bias at the beginning of the process?

c) Which random walk sampling method is the most representative one compared to the ground-truth?

Before starting the experiments, the authors reviewed some previous works and referred to Henzinger et al. [2000], Lee et al. [2006], Ahn et al. [2007]. Compared with these previous works, the authors state that Krishnamurthy et al. [2008] only examined the usage of privacy settings which could not be a comprehensive analysis. To overcome this shortcoming, this paper has additional privacy settings and the one-hop neighborhood for every node, which allows researchers to analyze user properties conditioned on their privacy awareness.

In addition, the authors introduce some new ideas in this paper. One is that they compare several candidate graph-crawling techniques in terms of sampling bias and efficiency. Also, they introduce ID sampling as the ground-truth of those sampling methods.

Another one is that they introduce the use of formal convergence diagnostics (namely Geweke and Gelman-Rubin) to assess sample quality in an online fashion.

By using these new ideas, Gjoka et al. conducted an experiment with four steps:

Step 1: The authors crawled the data from Facebook using 5 different crawling strategies—BFS, Random Walk, Re-Weighted Random Walk, Metropolis-Hastings Random Walk and ID Rejection (ground-truth).

Step 2: They used two online convergence diagnostics to adjust the dataset omitting the first several crawling iterations.

Step 3: They extracted the data of interest and structured them, the data of interest include name, userID, friend list, networks, privacy settings, profiles, ego networks.

Step 4: They analyzed those data in terms of degree distribution, regional networks, userID space, clustering coefficient, privacy awareness and assortativity.

After the experiment, the authors achieved several results which include:

a) The authors state that they developed a framework for unbiased sampling of users in an OSN by crawling the social graph, and they provided recommendations for its implementation in practice.

b) They state that they summarized the data by calculating the Pearson correlation coefficient, or assortativity coefficient which is $r = 0.233$.

c) They state that RWRW and MHRW achieved best in terms of unbiased sampling and properties analysis.

d) They state that they found RWRW to be the most efficient in practice while MHRW has the advantage of providing a ready-to-use sample.

e) They also state that they obtained first unbiased sample of Facebook users, which they used it to characterize several key user and structural properties of Facebook.

This paper showed a comprehensive comparison between several kinds of Random Walk sampling methods with uniform sampling which played the role as benchmark.

The contributions made in this paper include: 1) the introduction of the use of formal online convergence diagnostics. In addition, the researcher performed an offline comparison of all crawling methods against the ground truth, 2) they also provided guidelines for implementing high performance crawlers for sampling OSNs.

4.2 RANDOM NODE SAMPLING

In this sampling method nodes are selected directly. This is in contrast to the crawling/walking sampling where nodes are obtained by following some links. There are a few strategies to select the nodes which depends the OSN access methods available. For instance, if ID sampling is possible, the nodes can be selected randomly with uniform distribution. If nodes can be found by search the key words, nodes are selected by pagerank if they are ranked by pagerank algorithm, or by time if they are ranked by freshness.

There are choices as for what the links are in the sample. One approach is to expand the graph one step further by selecting all the neighbors. In the case of a popular user who has millions of fans, collecting all of them impose a practical problem - the sample size will be too large, and all the sample nodes are concentrated on this particular spot, therefore may be biased.

It is always unclear that how to make a sample representative. As a consequence, many researchers are intent on crawling the complete graph from the database. So what is the difference between analyzing sampled sub-graph and complete graph of the network?

In order to tackle this problem, Ye et al. [2010] proposed a random node sampling method and applied this method to four different social networks. They improved some factors including choice of seed, node selection algorithm and sample size to evaluate the

OSN graph crawling problem. In addition, they created a strategy to omit the "black hole" users (the users who set their profiles invisible) to evaluate sampling accuracy.

By analyzing a figure showing node and edge coverage under the different node selection algorithms, they found that the mean degree reported by each crawler follows the order of hypothetical greedy > greedy > lottery > BFS. Based on this result, Ye et al. [2010] formalized how to evaluate crawling bias and what parameters needed to be considered.

4.3 RANDOM EDGE SAMPLING

In some social networks, "relations" which are regarded as edges can be easily extracted from the dataset. These edges connect the users on their ends and all together form the complete graph of the network.

This method may be effective for affiliation graph or message graph. For instance, random messages are selected first, where each message connects users who repost or comment on the message.

Katzir et al. [2011] compared biased sampling and unbiased sampling method with respect to the OSNs' size estimation. The authors pointed out that in Gjoka et al. [2010], the mixing rate of MH random walk can be significantly worse than that of the original graph, and so, it is unclear when it is expected to outperforms rejection sampling, i.e., require fewer random walk steps. They also stated that Yong-Yeol et al. [2007] said the total number of users (or users in a certain demographic) seems to be one of the most crucial factors in deriving the worth and overall performance of social networks.

To make the results more reliable, they created a new algorithm called COLLISION COUNTING (the algorithm divides the graph into sub-graphs based on the pre-defined value) to compute sub-graph size and then integrated the proportions together.

They used their new algorithm to test in three real-world networks --- synthetic network, Digital Bibliography and Library Project (DBLP) and Internet Movie Database (IMDB).

The results showed both analytically and experimentally that, for social-networks and other small world graphs, COLLISION COUNTING algorithm considerably outperforms uniformly sampling method.

5. PROPERTIES ANALYSIS

This section describes some of the topological properties of the Social Network including degree distribution, clustering coefficient and assortativity.

All the papers reviewed in this section discuss these three basic properties.

Year	Authors	Title	Papers referred to	Major contribution
2007	Ahn et al.	Analysis of topological characteristics of huge online social networking services	Lee et al. 2006	First paper to claim that heterogeneous types of users were the force behind the behavior of properties analyzing.
2007	Java et al.	Why we twitter: understanding microblogging usage and communities	none	First paper to show the intentions of OSNs user.
2010	Kwak et al.	What is Twitter, a social network or a news media?	Ahn et al. 2007 Java et al. 2007	First paper to classify the trending topics based on the active period and the tweets.

Table 3: papers involve in properties analysis

5.1 DEGREE DISTRIBUTION AND CLUSTERING COEFFICIENT

When social networks were launched, their size was three or more orders of magnitude smaller than today's commercially available SNSs. At that time, the researchers had the privilege to study the properties of the OSNs for it was easy to crawl all nodes and links which seems impossible now.

It is known that all the human contact networks including online social networks belong to a class "scale-free networks" and exhibit a "small world phenomenon". But different OSNs always present different topological properties. In Ahn et al. [2007], the authors studied three OSNs – Cyworld, orkut and MySpace and found that the degree distribution of these social networks showed heterogeneous user phenomenon.

Ahn et al. pointed out that in Lee et al. [2006] snowball sampling is very difficult to get a power-law degree distribution from a network without the power-law decaying degree distribution. Based on this, the researchers raised two ideas: 1) they created a simple functional form with the combination of power-law and exponential function and the degree distribution were remarkably well fitted by this form, 2) they extracted the testimonial network from the massive dataset, and defined it as the real-life human societies since not all online friends left testimonials on their friends' front pages. This could be regarded as close friendship.

To support the ideas, the authors conduct an experiment with four steps:

Step 1: They crawled the entire data from Cyworld and partial data from MySpace and orkut.

Step 2: Using the method of maximum likelihood, the authors estimated the degree distribution, clustering coefficient and assortativity. Moreover, they revealed the topological structure of Cyworld by analyzing the data.

Step 3: They showed the evolution of the online social network based on the basic properties including degree distribution, clustering coefficient, degree correlation and average path length.

Step 4: They summarized evaluation of snowball sampling method.

The degree distribution figure drawn from the data of Cyworld divides the CCDF into two regions: a rapid decaying ($\gamma \sim 5$) region and a heavy tailed ($\gamma \sim 2$) region. This behavior has not been reported previously about any SNS topologies. The multi-scaling behavior observed in Cyworld suggests that Cyworld consists of two different types of networks, i.e., two types of users. This proves that the heterogeneous types of users were the force behind the behavior of properties analyzing.

Compared with analysis above, the constituent of Twitter seems much purer. In Java et al. [2007], the researchers were trying to reveal the topological properties of Twitter and what the intentions of people in their daily activities. They regarded the network as the composition of different communities or clusters and this helps them to find the interests lies in the Twitter users.

The authors of this paper used the HITS algorithm to find hubs so that they could have detected the user intention based on the clustering. Then they designed an experiment with three steps:

Step 1: They used developer API to fetch the social network of all users.

Step 2: They used the data to construct a directed graph G (V, E), where V represents a set of users and E represents the set of "friend" relations.

Step 3: They showed the evolution of the online social network based on the basic properties including degree distribution, geographical distribution.

From the results of the experiment, the power law exponent of Twitter is approximately -2.4 and this value is similar to that found for Web and blogosphere.

5.2 ASSORTATIVITY

Depending on the type of complex network, nodes may tend to connect to similar or different nodes. Then we may be curious "Does a user of certain popularity follow other users of similar popularity and they reciprocate?" This question is similar to degree correlation. The degree correlation compares a node's degree against those of its neighbors, and tells whether a hub is likely to connect other hubs rather than low-degree nodes in an undirected network. The positive trend in degree correlation is called assortativity and is known as one of the characteristic features of human social networks.

In Kwak et al. [2010], the authors tried to reveal the structure of the users and how the famous people connect with each other. They proposed an experiment to find the most influent people and the communities in Twitter. After the experiment, they summarized that Twitter diverges from well-known traits of social networks: its distribution of followers does not follow a power-law, the degree of separation is shorter than expected, and most links are not reciprocated. But if looking at reciprocated relationships, then they exhibit some level of assortativity.

6. CONCLUDING COMMENTS AND FUTURE WORK

6.1 SUMMARY OF RESEARCH AND TABLE

Crawling complete graphs from OSNs is no longer possible owing to the huge size of the networks. In the recent four years, many frameworks have been built to study the OSNs, and the crawling strategies including API method, HTML method, query method and ID

crawling method are assistant component of these frameworks, they are always combined with sampling methods.

Since Lee et al. [2006] introduced statistics into sampling networks, the research in this field has expanded based on probability distribution model. In this survey, all the sampling methods from surveyed papers are different (Kwak et al. [2006] and Gjoka et al. [2010] both used random walk sampling, but the former one used snowball sampling while the latter one used RW/MHRW sampling method) because of the different purposes. When the accuracy is required in the research, we should avoid the bias of the sampling methods. When the rank of the networks is required, we should make use of the bias so that we can get high ranking users easily.

This survey takes a broad review of the related topic and uses a step-by-step approach to analyze the contribution of online social networks including crawling strategies, sampling methods and topological properties.

Year	Authors	Title	Papers referred to	Major contribution
2006	Kwak et al.	Impact of snowball sampling ratios on network characteristics estimation: A case study of cyworld	Lee et al. 2006	A new approach to evaluate snowball sampling method.
2006	Lee et al.	Statistical properties of sampled networks	none	First paper to combine OSNs' analysis with statistics.
2007	Ahn et al.	Analysis of topological characteristics of huge online social networking services	Lee et al. 2006	First paper to claim that heterogeneous types of users were the force behind the behavior of properties analyzing.
2007	Java et al.	Why we twitter: understanding microblogging usage and	none	First paper to show the intentions of OSNs user.

		communities		
2008	Paolillo et al.	Structure and network in the YouTube core	none	First paper to analyze the structure of YouTube commenting graph using API.
2010	Gjoka et al.	Walking in Facebook: A case study of unbiased sampling of osns	Henzinger et al. 2000 Lee et al. 2006 Ahn et al. 2007	First paper to introduce convergence diagnostics into sampling social networks.
2010	Kwak et al.	What is Twitter, a social network or a news media?	Ahn et al. 2007 Java et al. 2007	First paper to classify the trending topics based on the active period and the tweets.
2010	Ye et al.	Crawling online social graphs	Lee et al. 2006 Ahn et al. 2007	First paper to introduce quantification into evaluating crawling bias.
2011	Catanese et al.	Crawling Facebook for social network analysis purposes	Ahn et al. 2007	First paper to find BFS crawling strategy introduces bias into graph study.
2011	Katzir et al.	Estimating sizes of social networks via biased sampling	Lee et al. 2006	A new algorithm outperforms uniform sampling.

Table 4: Ten surveyed papers

6.2 FUTURE WORK

Online social networks have been carefully studied in recent years. However, the problem of social network data analysis is still in its infancy. There is a tremendous amount of work to be done, particularly in the area of content-based and temporal social networks. Some key research directions for the future are as follows:

Content-based Analysis: Much of the past research in this area has been based on structural analysis of social networks. Such analysis primarily uses linkage structure only in order to infer interesting characteristics of the underlying network. Some recent

research has shown that the inclusion of content information can yield valuable insights about the underlying social network. For example, the content at a given node may provide more information about the expertise and interests of the corresponding actor.

Temporal Analysis: Most of the research in social networks is based on static networks. However, a number of recent studies have shown that the incorporation of temporal evolution into network analysis significantly improves the quality of the results. Therefore, a significant amount of work remains to be done on dynamic analysis of social networks which evolve rapidly over time. Lee et al. [2006] points out that identifying new sampling method to explore complex network is an important research topic in the future.

Adversarial Networks: In adversarial networks, it is desirable to determine the analytical structure of a network in which the actors in the network are adversaries, and the relationships among the different adversaries may not be fully known. For example, terrorist networks would be a typical adversarial network to a law enforcement agency. Such networks are far more challenging because the links may not be known a-priori, but may need to be inferred in many cases. Such inferred links may need to be used for analytical purposes. Kwak et al. [2010] leaves the validation of the reciprocity of Twitter for future study as this property shows the structure of Twitter.

Ahn et al. [2007] points out that the scaling exponents of MySpace and orkut match those from different regions in the Cyworld network is also worthy of note.

To speed-up the data extraction process, Catanese et al. [2011] shows that the development of parallel code is the future work.

6.3 INTERESTING OBSERVATIONS

Almost all the surveyed papers share a common reference which is Lee et al. [2006]. This paper provides mathematical support and build statistical model for sampling the networks. But the problem is that this statistical model only performs well in dense networks, while in sparse social networks, the sample size should be large enough to make the result accurate. Ahn et al. [2007] points out this shortcoming, and they designed another algorithm: the capture/recapture method to overcome this challenge.

7. BIBLIOGRAPHY

AGGARWAL, C. 2011. *Social Network Data Analytics*. Springer-Verlag New York Inc.

AHN, Y.-Y., HAN, S., KWAK, H., MOON, S., AND JEONG, H. 2007. Analysis of topological characteristics of huge online social networking services. *In Proceedings of the 16th International Conference on World Wide Web.* WWW '07. ACM, New York, NY, USA, 835–844.

CATANESE, S., DE MEO, P., FERRARA, E., FIUMARA, G., AND PROVETTI, A. 2011. Crawling facebook for social network analysis purposes. *In Proceedings of the International Conference on Web Intelligence, Mining and Semantics.* 52:1-52:8.

CHAU, D., PANDIT, S., WANG, S., AND FALOUTSOS, C. 2007. Parallel crawling for online social networks. *In Proceedings of the 16th International Conference on World Wide Web*. ACM, 1283–1284.

GJOKA, M., KURANT, M., BUTTS, C., AND MARKOPOULOU, A. 2010. Walking in facebook: A case study of unbiased sampling of osns. *In INFOCOM, 2010 Proceedings IEEE.* Ieee, 1–9.

HANDCOCK, M. S. AND GILE, K. J. 2010. Modeling social networks from sampled data. *The Annals of Applied Statistics 4*, 1, 5-25.

HENZINGER, M., HEYDON, A., MITZENMACHER, M., AND NAJORK, M. 2000. On near-uniform url sampling. *Computer Networks 33*, 1-6, 295-308.

JAIN, A. K., MURTY, M. N., AND FLYNN, P. J. 1999. Data clustering: a review. *ACM Comput. Surv*. 31, 264–323.

JAVA, A., SONG, X., FININ, T., AND TSENG, B. 2007. Why we twitter: understanding microblogging usage and communities. *In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*. ACM, 56–65.

KATZIR, L., LIBERTY, E., AND SOMEKH, O. 2011. Estimating sizes of social networks via biased sampling. *In Proceedings of the 20th International Conference on World Wide Web.* ACM, 597-C606.

KNOKE, D. AND YANG, S. 2008. *Social network analysis*. Number 7. Sage Publications, Inc.

KWAK, H., HAN, S., AHN, Y., MOON, S., AND JEONG, H. 2006. Impact of snowball sampling ratios on network characteristics estimation: A case study of cyworld. *Proc. WWW07*, 835-844.

KWAK, H., LEE, C., PARK, H., AND MOON, S. 2010. What is twitter, a social network or a news media? *In Proceedings of the 19th International Conference on World Wide Web.* ACM, 591–600.

LEE, S., KIM, P., AND JEONG, H. 2006. Statistical properties of sampled networks. *Physical Review E 73*, 1, 016102.

LESKOVEC, J. AND FALOUTSOS, C. 2006. Sampling from large graphs. *In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 631–636.

NAVLAKHA, S., RASTOGI, R., AND SHRIVASTAVA, N. 2008. Graph summarization with bounded error. *In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data.* ACM, 419–432.

PAOLILLO, J. 2008. Structure and network in the youtube core. *In Proceedings of the 41st Annual Hawaii International Conference on System Sciences.* Ieee, 156–156.

RIBEIRO, B. AND TOWSLEY, D. 2010. Estimating and sampling graphs with multidimensional random walks. *In Proceedings of the 10th Annual Conference on Internet Measurement.* ACM, 390–403.

SANTO AND FORTUNATO. 2010. Community detection in graphs. *Physics Reports 486*, 3C5, 75 – 174.

STUMPF, M. AND WIUF, C. 2005. Sampling properties of random graphs: the degree distribution. *Physical Review E 72*, 3, 036118.

WANG, T., CHEN, Y., ZHANG, Z., XU, T., JIN, L., HUI, P., DENG, B., AND LI, X. 2011. Understanding graph sampling algorithms for social network analysis. *In 2011 31st International Conference on Distributed Computing Systems Workshops (ICDCSW).* IEEE, 123–128.

YE, S., LANG, J., AND WU, F. 2010. Crawling online social graphs. *In 12th International Asia-Pacific Web Conference (APWEB), 2010.* IEEE, 236–242.