Twitter: Network Properties Analysis

Abraham Ronel Martínez Teutle
Universidad de las Américas Puebla, Interactive and Cooperative Technologies Lab
abrahamr.martinezte@udlap.mx

Abstract

Online Social Networks have had a fast growing since the popularization of Web 2.0. This kind of networks provides the basis to find and maintain social relationships with users having diverse interests as well as a current picture of things happening around them. The analysis of the graph structure is necessary to understand the impact of online social networks among people, design better systems, and indirectly measure the Internet expansion. This paper presents a general background of network dynamics and structure, then the social network Twitter [1] is described and classified. Related work is analyzed showing that potential research opportunities might be found in online social network dynamics and suggesting Twitter as a good exercise to perform measurements.

1. Introduction

The development of online social Networks like blogs, wikis, and social networking sites (SNS) has shown the ability to create fast growing online virtual communities, where people communicate, share information, and keep in touch without even know each other. Recently SNS have gained significant popularity on the Web [2]. For example, MySpace (over 190 million users), Orkut (over 62 million), Facebook (over 350 million) [3], Twitter (over 45 million accounts and 18 million active), are popular sites among others [4]. Such trends show that the typical World Wide Web is having a subtle shift to Web 2.0 where people access to information through online communities.

Unlike the Web, which is largely organized around content, the Web 2.0 or social networks are organized around users. Participants join a network, publish their profile and content, and principally they create links to any other users with whom they associate. This network provides a basis for maintaining social relationships, finding users with similar interests, and locating content and knowledge that have been certified by other users. Sometimes, these sites provide a launch platform for free minds and ideas to trigger social movements [5].

In such scenario, SNS in the age of Web 2.0 provide an opportunity to measure how trends, ideas and information travel through social communities. The analysis of the graph structure of SNS is necessary to evaluate current systems, design future online social network based systems, understand the impact of online social networks on the Internet, and indirectly measure the Internet growth. For example, this structure might lead to algorithms that can detect trusted or influential users, similar to the study of the Web graph led to the discovery of authority sites to improve query searches (i.e.: Google pagerank algorithm).

This work provides an initial survey attempting to measure trends along time in Twitter; a rapid growing social network. An empirical analysis of topological structure of the members in this network might provide a first step to understand the various dynamic processes over time. Here, the Twitter network structure is described; related work, network analysis techniques, crawling methods as well as programmatic data structures are considered. Finally graph structure measurements are conducted on Twitter showing some special characteristics.

2. Background and motivation

2.1. Social Networking Sites

Social network sites are usually composed by:

Users: It is the principal actor of these sites. Normally, to be part of the network, a user must register his/her personal information such as *full name, a nickname, a password, e-mail,* etc. Users may give voluntarily more information to enforce their role within the network as a user *profile*.

Links: Users (or nodes) may establish diverse relationships inside the network with other users; these might represent known people, business contacts, similar interests, friendship, likeness, or simply awareness among others.

Groups: Certain social networks enable users to create and join special interest groups. Clients might share resources, post messages, and establish certain rules to be part of that group.

2.2. Network Dynamics

According to [6], there are two types of network dynamics that can be defined: *dynamics **of** the network* and *dynamics **on** the network.*

978-1-4244-5353-5/10/$26.00 ©2010 IEEE

The first type refers to the evolving or changing structure of the networks itself, i.e.: new nodes and the making and breaking of ties. In this case measurements take snapshots of the network over the time showing its evolution.

In the second type, individuals (nodes) are doing something between them. For example, they search for information, learn, share, make decisions, etc. Their acts are influenced by what their neighbors are also doing.

This work will be focused on the significance of the first type of network dynamics.

2.3. Reasons for measuring social networks

Nowadays social networks play an important role in personal and commercial online interaction. According to [7], the growth of social networks is one of the most explosive trends in the Internet traffic; at least seven services present a growth over 100 percent of unique visitors. Moreover, late reports show that from March 2008 to March 2009, Twitter had a growth of 2,565% (13,858,000 unique visitors on march 2009) implying a change in their network dynamics.

Understanding the structure of online social networks is critical to be aware of the general impact on the future Internet. It also offers an opportunity for other disciplines such as sociology and marketing to study social networks at a large scale, compare behaviors, relationship patterns, and so on.

As seen on table 1, Twitter is one of the fastest growing SNS (without counting non browser applications). Consequently, it is natural to assume a rapid change on the dynamics of the network, which is suitable for social analysis using graphs.

Table 1. Top 6 fastest growing social networks Mar 08 - Mar 09 unique new visitors.

Name	Mar. 2008	Mar. 2009	Growth
Twitter	520,000	13,858,000	2565%
Ning	1,463,000	5,609,000	283%
Facebook	24,940,000	69,151,000	177%
Bebo	2,483,000	6,149,000	148%
LinkedIn	7,877,000	15,815,000	101%
Multiply	780,000	1,527,000	96%

3. Twitter

Twitter is a free social networking and micro-blogging service where users stay connected through the exchange of quick, short and frequent messages also called "tweets". These messages are text-based posts of up to 140 characters. A common fact in Twitter is "re-tweet", which means that someone forwards a message from another user; thus, making it available for more users.

Tweets are displayed on the user's profile page and delivered to others who signed up to receive them. Such subscription is also known as *following someone*.

A normal web-based profile layout of a Twitter user is composed by the main box where users update their status answering to a simple question: *what's happening?* Messages written here will be shown to all of them *following* this specific user. A brief list of *following or subscribed people* can be found aside. Finally we must assume that anyone can stop following someone.

This social network tool have had great acceptance in the last year due to its real time relevance when searching current facts (i.e. news). According to [8] from December 2008 to April 2009 the Twitter service had its biggest growth. For these statistics we must consider that Twitter service might be accessible using stand-alone programs and mobile devices; therefore, unique visitor numbers might be slightly greater.

Another point is that most of the Twitter profiles are publically available because the micro-blogging service is based on the idea of telling people "*what is currently happening?*" In such scenario is easy to think in a web crawler to collect data for analysis.

Finally the simplicity of Twitter service allows mapping users as nodes and following relations as directed links; which creates a directed graph suitable for further social network analysis (Figure 1). The reader should notice that information flows in the opposite direction of a direct link since it represents the relation of being subscribed to someone.

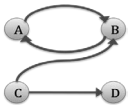


Figure 1: Relationships among Twitter users. User A follows user B, C follows B and so on.

4. Descriptive Measures

Social network analysis is an interdisciplinary methodology developed mainly by sociologist and researchers in social psychology in the 1960s and 1970s. Nowadays in collaboration with mathematics, statistics, graph theory, and computing, this method has led a rapid development of formal analyzing techniques. Social network analysis is based on the importance of relationships or *links* between interactive units or *nodes* [6]. Some descriptive measures to be considered are:

4.1. Strongly connected components

A directed graph is called strongly connected if there is a *path* from each vertex in the graph to every other vertex. Thus, strongly connected components of a directed graph are its maximal strongly connected sub-graphs.

This measure will be helpful when distinguishing the connectivity between groups. Along time, if the number of strongly connected components drop in a certain subgraph it will show that re-tweets might be available for more users. An example of it is shown in figure 2.

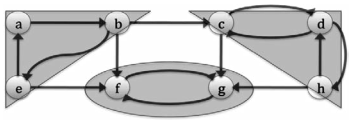


Figure 2: Strongly connected components in a directed graph. This example contains 3

4.2 In-degree and out-degree

The in-degree corresponds to the number of edges coming into a vertex in a directed graph. The in degree shows the growing authority of a node along time. For Twitter, it represents the number of followers.

The out-degree is the number of edges departing from a vertex in a directed graph. The out-degree will be helpful to compute the following ratio.

4.3 Clustering Coefficient

The clustering coefficient of a vertex quantifies how close the vertex and its neighbors are to being a complete graph (a simple graph in which every pair of distinct vertices is connected by an edge).

If clustering coefficient increases, it will show the tendency of a social network or sub-network to become interconnected. This case is different from strongly connected components in that original messages (tweets) will be available for more users.

4.4. Network Density

Density is defined as the actual number of edges in a network, expressed as a proportion of the maximum possible number of ties. When density is close to 1.0 the network is dense, otherwise it is sparse.

Density is important to measure the "following" rate along time in the Twitter service.

4.5. Betweenness

Represents how much a node is part of all the shortest paths between any given nodes. In other words, vertices that occur on many shortest paths between other two vertices have higher betweenness than those that do not.

4.6. Closeness

It is defined as the mean of all shortest paths between a vertex and all vertices reachable from it. This variable reflects how much a node is immersed within the network. A node with a higher closeness will receive flooding messages more quickly than lower closeness nodes (i.e. breaking news).

4.7. Following ratio

Represents the in-degree (followers) divided by the out-degree (people following). This measure shows how users reveal some profiles within the network:

- Ratio bellow 1: People follows more than being followed, typically seek for knowledge or collect public information. If the ratio is too low, the node might be a web-bot or spider collecting information about trending topics.
- Ratio approximately 1: Related to users with the same in-degree and out-degree that reveal a typical community.
- Ratio above 1: Popular users respected by their community, but still engaged to it. They usually share resources well appreciated by others.
- Ratio above 10 or higher: Represent nodes that have a huge impact around general media, and have no interest on following back, either because they cannot handle following back too many users or do not care about them.

5. Data sets: crawling methods and programmatic data structures

Crawling Twitter data is possible since most of the information is publically available. For this exercise there are two ways to get user's profiles: web and API-based crawling.

Web crawling will process a user's profile web page obtaining data directly from the HTML tags downloaded. However, this procedure is time-consuming since it downloads useless information like user and follower's images.

The API-based crawling uses the public API provided by Twitter to request specific information; consequently, increasing the overall performance when obtaining data. Several libraries are available online; in this case Twitter for Java will be used to retrieve the information.

Twitter for java [9] is a small library for the Java programming language under the GNU Lesser General Public License. It works using an HTTP authentication and retrieves the data wrapped in objects. An example of a authentication and friend list retrieval is shown next.

```
a) Twitter objectTwit =new
Twiter(username,password);
System.out.println("Connection done");
b) List<User>friendsList =
objectTwit.getFriends(userId);
```

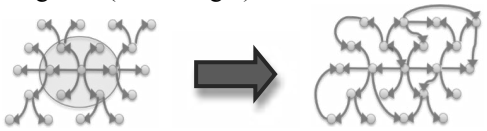
Listing 1: Authentication and friend list retrieval

Similar to [9], JGraphT [10] will be used to process all the crawled data. The graph library uses internally vectors, linked lists, arrays, and the object-oriented paradigm to make easy graphs management.

5.1. Graph snapshoot

To perform the descriptive measurements listed above, an initial graph is captured from the Twitter network. Once authenticated, the snapshot is built as a tree using a Breadth First Search approach starting from a user with a

high out-degree (at least 30 following people). Then, such tree is traversed in order; for each node a link between the current user and any other user within the graph is established when the same relationship is found in the user's "following list" (see listing 1).



a) b)

Figure 6. a) Tree creation and b) final graph snapshot diagram.

5.2. Reproducibility

To guarantee reproducibility and perform measurements over the same users, nodes ID's are saved in a database. Therefore, further crawling will not require taking the graph snapshot.

6. Evaluation

For this study, a graph snapshot of **14,148** nodes was processed. Profile information retrieved corresponds to number of messages posted **(# messages)**, number of following people **(out-degree)**, number of followers **(in-degree)**, and user ID **(nickname)**. Furthermore, following ratio, betweenness, clustering coefficient, and closeness were computed for each node. Finally, strongly connected components and network density were also computed as a general picture of the snapshot. Four different captures were deployed as seen on table 2:

Table 2: Crawling dates

Capture #	Date
1	April 11th, 2009
2	April 22nd, 2009
3	April 23rd, 2009
4	April 29th, 2009

6.1. Messages posted

Figure 3 shows the cumulative distribution function (CDF) of each capture. Around 80% of the users have posted more than 1,500 messages and a few as 5% more than 5,000 messages. The overall distribution maintains its shape along the different captures, which means that most of the active users post regularly. The general average is 9 daily posts per user (see table 3).

Table 3. Messages posted: max, mean and min

	Capture 1	Capture 2	Capture 3	Capture 4
Max:	173,846	176,690	177,004	178,374
Mean:	1,144	1,243	1,254	1,300
Min:	0	0	0	0

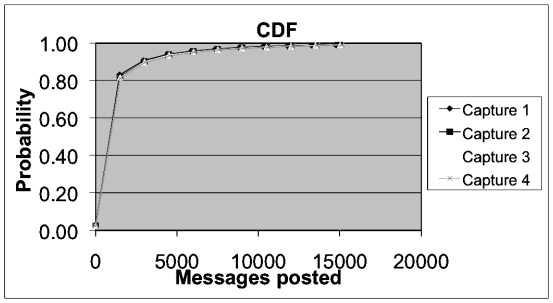


Figure 3. CDF for messages posted per user

6.2. In-degree or followers

Figure 4 represents the CDF for the number of followers for every capture. 25.5 % of the users have at least 50 followers. CDF changes over time suggest that the amount of followers increased quickly within a few days. For example, from capture 1 to capture 2 people with 50 followers decreased from 25.5% to 19.94% of the total, however people with more than 600 followers increased their in-degree rapidly (i.e. the number of nodes with more than 100,000 followers changed from 13 to 35). This suggests that normal and new users are following really famous people within the network. Albanesius [11] mentions that many users created an account to follow someone famous. This scenario is depicted on table 5, where the user *aplusk* got a popularity peak.

With this, it is remarkable that famous people had 140 new followers per day (mean). Table 5 shows one of the peaks observed on the following rate.

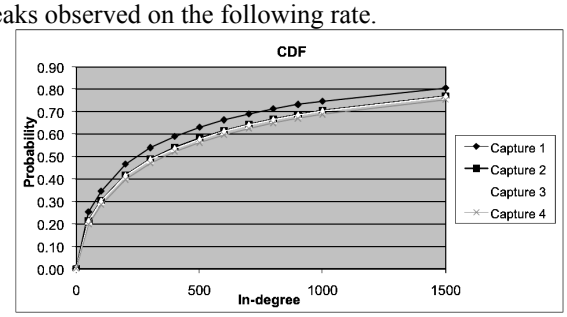


Figure 4. In-degree CDF

Table 4. In-degree: max, mean and min

	Capture 1	Capture 2	Capture 3	Capture 4
Max:	884,631	1,318,608	1,350,777	1,504,914
Mean:	4,925	6,420	6,584	7,293
Min:	0	0	0	0

Table 5.Highest in-degree nodes

Capture#	1st place	2nd place	3rd place
1	cnnbrk 884,643	britneyspears 816,495	aplusk 812,697
2	aplusk 1,318,608	cnnbrk 1,101,475	britneyspears 1,091,719

3	aplusk 1,350,777	britneyspears 1,108,716	cnnbrk 1,101,475
4	aplusk 1,504,914	cnnbrk 1,271,275	britneyspears 1,231,828

6.3. Out-degree

Figure 5 reveals that out-degree rates have not change as much as the in-degree. Change on the different CDF either represent three cases: a) New users follow many people at the beginning that might be promising, then users refine their subscriptions, b) people with higher number of followers are following back to their subscribers, or c) The presence of bots is more common when analyzing Twitter trends, as they collect large quantities of following people. Table 9 summarizes out-degree.

Table 6. Out-degree: max, mean and min.

	Capture 1	Capture 2	Capture 3	Capture 4
Max	628,151	628,151	763,578	765,408
Mean	1,492	1,819	1,846	1,954
Min	0	0	0	0

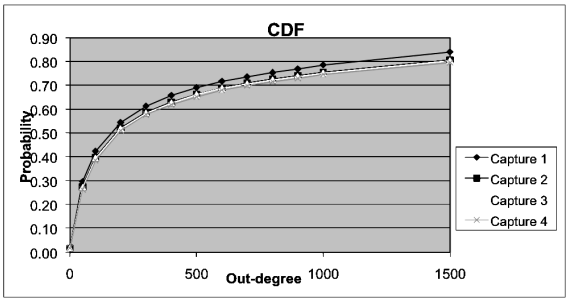


Figure 5. Out-degree CDF

6.4. Following ratio (Followers / Following)

The following ratio shows that around 50% of the users behave in a 1:1 relationship. In other words, these users keep the same amount of in-degree and out-degree. Figure 6 shows that following ratio maintains its proportion along time, suggesting that major changes on in-degree are due to many users following someone specific, than one node following suddenly many other people. Slight variations were registered for those users with high following ratio, confirming that those users are indeed getting more followers (i.e. from capture 1 to capture 4 there are 3.5% new high following ratio users). For displaying purposes those users are not shown on the graph.

Table 7. Following ratio: max, mean and min.

	Capture 1	Capture 2	Capture 3	Capture 4
Max:	884,631	183,579	183,579	211,879
Mean:	194	194	197	215
Min:	0	0	0	0

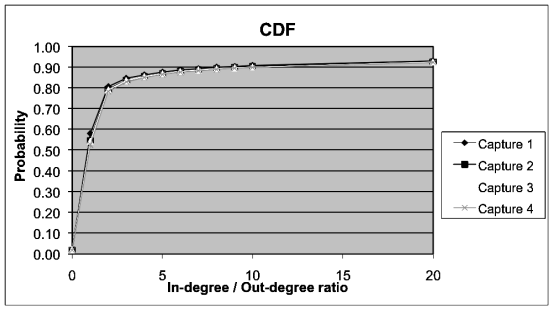


Figure 6. Following ratio

6.5. Clustering coefficient

This descriptive measurement was performed using capture 2 and 3 (one day of difference), thus the CDF does not vary at all. However Figure 7 shows that potentially 45% of the users are leafs within the network, which means that they have a very low or inexistent out-degree. This percentage might indicate that a considerable number of nodes are users that signed in for a Twitter account but do not use it anymore. These profiles are called inactive users or inactive community users.

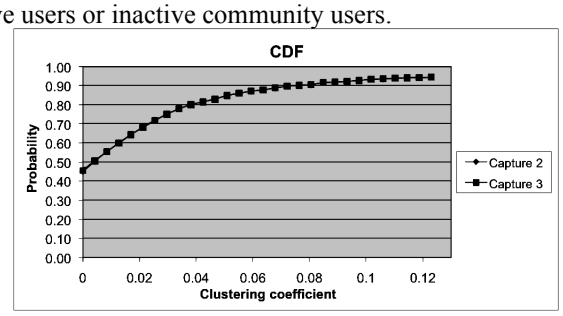


Figure 7.Clustering coefficient CDF

On the other hand, users with more clustering coefficient mean that they deploy stronger communities that exchange messages among them. These people are able to know about events happening in the network quicker than those who do not have higher clustering. For example, at least 30% of the users are able to receive direct or forwarded (re-tweets) messages quickly. This phenomenon is important since Twitter now plays a role where people say "*what is happening out there*" rather than "*what are you doing*" as it was focused at the begining.

6.6. Betweenness

This descriptive measurement shows that 20% users have very low or inexistent betweenness coefficient, meaning that those users do not have a relevant role within the network. In other words, they have a very low in-degree and out-degree that might show the real amount of inactive users. Figure 8 shows Betweenness CDF.

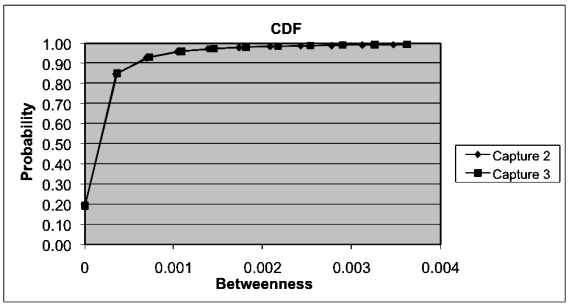


Figure 8. Betweenness CDF

On the other hand, high betweenness means that the following ratio tends to be 1; users follow and are being followed. It should be noticed that betweenness differs from clustering coefficient; the first one shows 50% of the users "playing a role" within the network, while the second one shows that 30% of the total nodes is clusterable or tend to be in communities.

6.7. Closeness

Finally this measurement shows that around 70% of the nodes have closeness between 0.22 and 0.3 units, suggesting that many of the users are highly close each other. In other words, a forwarded message or a topic trend is more likely to be spread quickly among those users with high closeness. Also, around 8% of the users have no closeness, meaning that those nodes are new to the network at the moment of the snapshot (Figure 9).

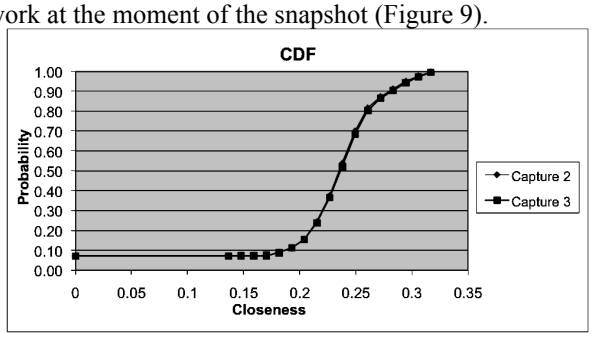


Figure 9. Closeness CDF.

6.8. Network density and strongly connected components

Results of the overall network show that not all the possible links within the Twitter network are exhausted (see Table 8). In a similar way, the strongly connected communities varied from one day to another a very low percentage suggesting that building communities takes time within network, rather than those explosive following peaks.

Given the growth of twitter we appreciate this low increase of the strongly connected components, since most of the new and inactive users are represented with a single strongly component.

Finally, the network density coefficient is not expected to grow because people are getting a level where no more information or subscriptions can be handled. On the other hand, network density is expected to diminish while the exponential growth of Twitter keeps going.

Table 8. Network density and strongly connected components measurement.

	Capture 2	Capture 3
Network density	0.0009843931	0.0009687799
SCC	2606	2651

7. Open issues and future work

For this project a low quality GUI was developed to show a map of the snapshot taken (Figure 10). However this output does not provide feasible human-computer interaction to analyze data. A future release will improve such output.

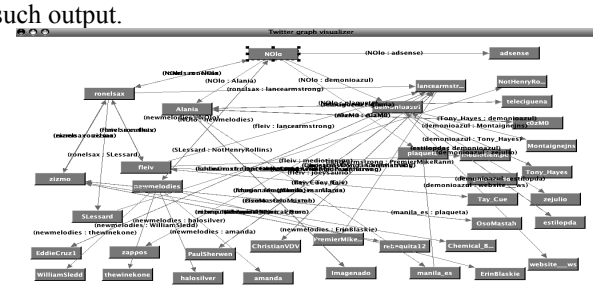


Figure 10. Output of the graph snapshot for the Twitter Crawler tool (36 nodes).

This work provides initial profiling of the Twitter graph structure, however measurements with bigger snapshots and regular time intervals might provide a better picture of the Twitter network dynamics. In a similar way, correlation between different descriptive measurements will provide a better understanding and profiling of complex relations within the network. For example, given the clustering coefficient, what is the probability of a higher betweenness?

Finally a better and formal test bed must be identified using real information from Twitter, however due to disclosing agreements, that information is not yet available.

8. Related work

In [2] is presented an empirical study for online social networks focused on blogging networks (blogosphere), with initial data collected authors demonstrate that such networks possess small-world and scale-free features. In addition, the distribution of topological distance, in degree, out degree, and clustering coefficient degree are analyzed too. In a similar research approach [4], [12] and [13] present a large-scale study and analysis of the structure and geometry of multiple online social networks (Flickr, YouTube, LiveJournal, and Orkut, My Circle among others). [14] also lies in this category but analyses

the relations between social network characteristics in an online class and learning outcomes. Nonetheless, measurements along time are not performed.

The concept of collaborative work and privacy is discussed in [15] and [16].

In [17] MySpace is studied as one of the most visited social networks. This work explores the demographics of its users focusing in how specific age ranges create their own social norms within social networking. However, this work does not study the dynamic of such networks.

Finally other works like [18] and [12] are focused on social networking growth; nevertheless the focus is for marketing purposes, showing which service provides a bigger audience.

Further work: correlate information i.e. people with following ratio 1 and low closeness means that they form small and "internal" communities, which are not highly interested on people outside of their network (i.e. job or corporate communities). On the other hand, users with following ratio 1 and high closeness will be related to large communities interested on "*what is happening now*" through people they might not even know.

9. Conclusions

This paper presented a general background of network dynamics as well as motivations related to the growth, structure and dynamic **of** the network.

Twitter was described and classified as one of the most rapidly growing social networks; its simplicity allows applying Social Network methodologies along time, such as strongly connected components, in degree, clustering coefficient, following ratio, and network density. The technical mechanisms were also mentioned.

A series of measurements were performed over the social network, providing a better introspective of its growth.

Finally related work was surveyed showing that potential research opportunities might be found in online social network dynamics, suggesting Twitter and its rapid growth as a good exercise to perform measurements.

10. References

[1]Twitter Inc, *Twitter: What's happening?*, retrieved online on January 10th 2010 from: http://www.twitter.com, 2010.

[2] Fu, Feng, et al., "Empirical analysis of online social networks in the age of Web 2.0", *ScienceDirect*, retrieved online on February 20th 2009 from: http://bit.ly/4F0xwM, 2007.

[3] Facebook, "Press Room", *Statistics*, retrieved online on January 12th 2010 from: http://bit.ly/12oAN, 2010.

[4] Mislove, Alan, et al., "Measurement and Analysis of Online Social Networks", *In procedings of the 7th SIGCOM conference of Internet measurement*, ACM, 2007.

[5] Twitter.com. Twitter blog. "Tracking Candidates on Twitter", retrieved online on February 12th 2009 from: http://bit.ly/pHseH, 2008.

[6] Coulon, Fabrice, "The use of Social Network Analysis in Innovation Research: A literature review", retrieved online on March 2nd 2009 from: http://bit.ly/2Im7IM, 2005.

[7] Nielsen Analytics, "Social networking stats", Press release, retrieved online on January 11nd 2010 from: http://bit.ly/2R1Ro, 2009.

[8]. Syte Analytics, "Twitter stats", retrieved online on January 11th from: http://bit.ly/mivD, 2010.

[9] Twitter for java. "Twitter 4Java: the Java library for the Twitter API", retrieved online on March 4th 2009 from: http://bit.ly/7LiSC, 2008.

[10] JGraphT, "The JGraphT library", retrieved online on March 4th 2009 from: http://bit.ly/cxk8k, 2005

[11] Albanesius, Chloe, "Many new Twitter users not coming back for more", *PC Magazine news and analysis*, retrieved online on April 16th 2009 from: http://bit.ly/KrilM, 2009.

[12] Ahn, Yong-Yeol, et al, "Analysis of Topological Characteristics of Huge Online Social Networking Services", *In Proceedings of the 16th international conference on WWW*, 2007.

[13] Zinoviev, Dmitry, "Topology and Geometry of Online Social Networks", *Proc. 12th World Multi-Conference on Systemics, Cybernetics and Informatics VI*, pp 138-143, 2008.

[14] Russo, C. Tracy and Koesten, Joy, "Prestige, Centrality, and Learning: A social Network Analysis of an Online Class", *Customer Services for Taylor & Francis Group Journals*, 2005.

[15] Wellman, Barry, et al, "Computer Networks as Social Networks: Collaborative Work, Telework, and Virtual Community", *Centre for Urban and Community Studies,* University of Tornoto, 1996.

[16] Gross, Ralph, et al, "Information Revelation and Privacy in Online Social Networks", *In ACM Workshop on Privacy in the Electronic Society*, ACM, 2005.

[17] O'Malley, James, et al, "The analysis of social networks", *Department of Sociology,* University of Arizona, 2005.

[18] Krzykowski, Matthaus, "Analysis: MySpace and Facebook challenge mobile-only social networks", retrieved online on February 25th 2009 from: http://bit.ly/8tMic, 2008.