A Study on the Characteristics of Data Traffic of Online Social Networks

Shengyong Ding
Guangzhou Research Institute
China Telecom Co. Ltd.
Email: dingsy@gsta.com

Kunfeng Lai
Department of Computing
The Hong Kong Polytechnic University
Email: cskflai@comp.polyu.edu.hk

Dan Wang*
Department of Computing
The Hong Kong Polytechnic University
Email: csdwang@comp.polyu.edu.hk

***Abstract*—In the past few years, we have witnessed a flourish of online social network websites (OSNs). In this kind of websites, the users are not only information consumers, but also actively upload contents of their own to the OSNs. Being sharply different from the conventional sites, OSNs have attracted many studies recently. These studies, however, mainly focus on the social behaviors within OSNs, e.g., the user-user interaction, and the distribution of the time users spend on certain OSNs, etc. There are much fewer studies on the characteristics of the traffic patterns of OSNs upon the Internet infrastructure, even though it is important for the ISPs and the OSNs alike.**

One major difficulty is that it is not easy to obtain data from the private ISP backbone routers. In this paper, we collect data from two backbone routers of China Telecom and present the traffic patterns of several OSNs from the angle of the network layer. We also compare the OSN traffic with three types of traditional websites, namely forums, search engines and news websites. We find that the traffic pattern in network layer show similarities with the user behavior in the application layer.

I. Introduction

The Internet has witnessed a flourish of the online social network sites (OSNs) in the past few years. Among them, the general online communities Facebook and MySpace, video sharing site YouTube, photo sharing site Flickr are some of the most successful examples. The uniqueness of these sites from the traditional websites is that the users of these sites are not only information consumers, but they also actively upload contents of their own. As a new type of websites, there is a great need to understand the user behaviors, so as to improve the organization and regulation of the OSNs.

There are many studies in recent years. These studies usually crawl the data of some representative OSNs for a period of time, and then to study the online social interactions within the OSNs, including user-user interactions (e.g., friendship relationship), user-content interactions (users and their maintained blogs), content-content interactions (content organized/recommended by correlated categories). For example, the characteristics of YouTube are studied in [4], where small world phenomenon between related videos are observed and the video popularity distribution are analyzed. A special function of YouTube, the external link is studied in [8]. Other examples include studies on Facebook [5] [9], Twitter [6], Orkut [3], to name but a few.

These studies provide valuable information of the OSNs from the angle of the users and the contents. There are few understandings on the traffic patterns of these sites on the Internet supporting side, the infrastructures. Clearly, as the contents of the OSNs are not uploaded and regulated by the websites themselves, one may question that the traffic of OSNs can be significantly different from the traditional websites. We believe this question is important to the OSNs and the ISPs alike. Nevertheless, there is a lack of studies towards this direction. One possible reason is that it is not easy to obtain the data traffic of the network layer routers. In sharp contrast to the data of the OSNs, these are private to the ISPs. In this paper, we have collected traffic from two core routers of China-Telecom, the dominant ISP in China. We use NetFlow [17] to collect detailed information from IP layer.

We would like to mention that having the data traffic from the routers does not lead to direct identification of the application layer traffic to the different sites. We use source-destination IP addresses as a criterion to separate the traffic for different websites. We face another difficulty of one IP mapping to multi-domains and/or one domain mapping to multi-IPs. We use DNS query from multiple DNS servers and reverse DNS service to solve these two problems.

In this paper, we first present general statistics of the traffic of the websites. We find that the traffic of the OSNs on backbone routers are generally different from the other websites. We then present some basic features of the traffic of the OSNs. Especially, we compare the differences on the traffic patterns of some representative OSNs and traditional websites. We observed that 1) compared to the search engine sites, the OSNs have a larger percentage of large data upload; 2) compared to news websites, the OSNs have more incoming traffic percentage; and 3) the traffic pattern reflecting user behavior of forum is similar with OSNs; but with the total amount of incoming traffic of forum takes less proportion.

The remaining part of the paper is organized as follows. In Section II, we discuss some related work. We present the background and measurement methodology in Section III. In Section IV, we give general statistics of the OSN traffic. We provide an in-depth study on the traffic patterns of OSNs in Section V. We conclude our paper in Section VI.

II. Related Work

There are many measurement studies on online social networks recently. In [12], four different OSNs, Flickr, LiveJournal, Orkut and YouTube, are studied and compared. The property of the OSNs is observed that the OSN networks contain a densely connected core of high-degree nodes, and low-degree nodes at the fringes. A general measurement study on YouTube is in [4], and a small world phenomenon has

* Corresponding author. Dan Wang's work is supported by grants Hong Kong PolyU/G-YG78, A-PJ19, 1-ZV5W, and RGC/GRF PolyU 5305/08E.

been observed for the graph constructed by related videos in YouTube. There are also studies for other OSNs, e.g., Slashdot Zoo [7], Orkut [3], Twitter [6], etc.

There are recent studies focusing on some specific features of the OSNs. In [8], the external links of YouTube are studied. It shows that in some cases, as many as 15% of the views can come from the external links, indicating a growing penetration of Youtube to other websites. In [13], tags and user behavior prediction are studied, and also a new tag recommend mechanism is provided. In [14], the negative edges, which indicate the dislike relationship between users, are studied.

Nonetheless, all these studies focus on the impact of the OSNs on the end users and the communities. They have explained from different angles of the reasons for the success and popularity of the OSNs. There are very few studies, however, on the characteristics of the data traffic of OSNs.

Internet infrastructure measurement has attracted studies well before the OSNs came into being. For measurement methodology, a new cross-traffic filtering method called minimum delay difference is recommended in [2]. The method can obtain accurate capacity estimation from the minimal possible delay of packets from different packet pairs. There are tools developed, such as netPolice [15], which can detect the content-based and routing-based differentiations in backbone ISPs. A study on data center traffic measurement is in [1].

To the best of our knowledge, there are no specific studies on the characteristics of the traffic of the OSNs from the angle of Internet infrastructure. In this paper, however, we do not try to understand the total traffic of a individual OSN. This can be done by monitoring the traffic at the Internet end for each individual website. We study the flows of the OSNs from the angle of backbone routers. We hope this can serve as a building block towards a comprehensive understanding of the OSNs.

III. Background and Measurement Methodology

A. The Data Trace

We collect our data from two backbone routers of China-Telecom from 10:00pm June 9th 2010 to 6:00am June 11th 2010. We collect the data traffic for using NetFlow V5 [17] with a linear sampling rate. The total data we collected is more than 300 GBytes. The detail information for one traffic record that we can get is shown in [17]. For commercial reasons, we do not disclose more information of our data collection configurations. We welcome individual request for more details of our data for non-commercial purposes by contacting the corresponding author of this paper.

B. The Websites Studied

From the raw data trace of the backbone routers, we extract data traffic of four types of websites, namely, OSNs, news websites, forums and search engines. The websites we chose are largely from China. This is because some international websites are not popular or accessible. For example, Twitter is generally not accessible in China and Facebook and YouTube are blocked. The websites we chose are representative; as all of which are from the top 30 websites in China according to [16] and all have large data traffic.

For commercial confidential regulations, we cannot reveal the names of the websites. We use labels to denote the websites we sampled. The two sampled OSNs are labeled as OSN_l (with lower traffic) and OSN_h (with higher traffic). Similarly, we label the sampled news websites as News_l and News_h, forums as BBS_l and BBS_h, and the search engines labeled as SE_c (a popular search engine in China) and SE_w (a popular worldwide search engine). In more details, OSN_l is a comprehensive online community similar to facebook. OSN_h is also a online community, but it emphasizes more on photos and has good interfaces to promote photos. News_l and News_h are quite similar in function and are all news webites similar to BBC. BBS_l focuses more on different categories of topics like history, travels and other general topics, while BBS_h focuses more on everyday life.

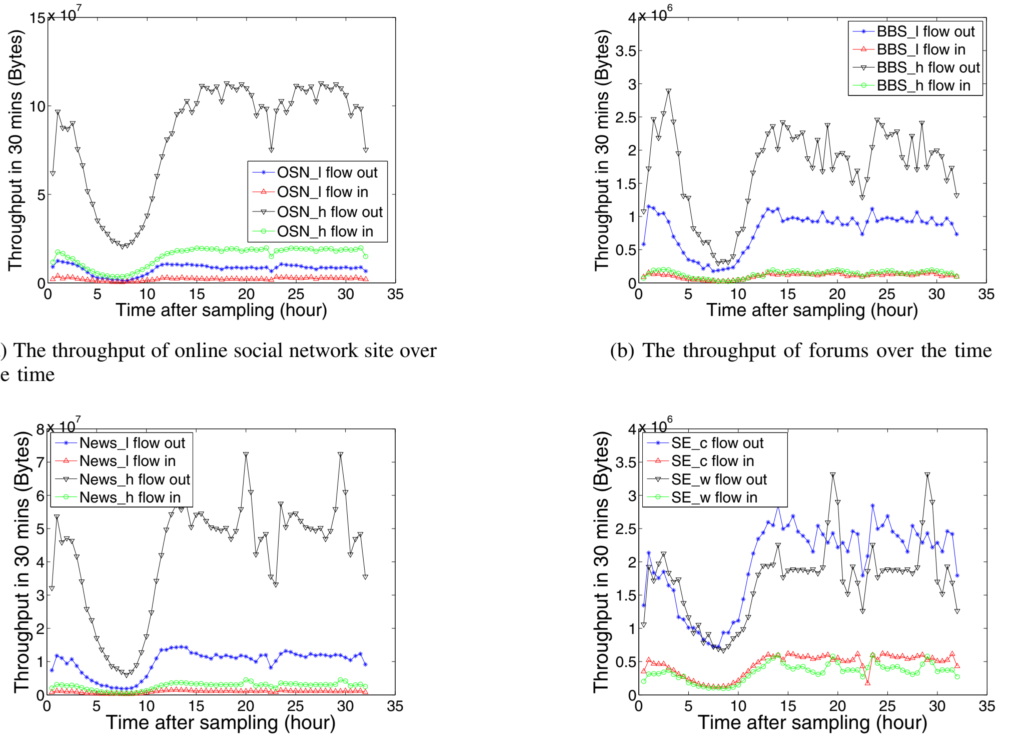
C. Traffic Extraction

The data packets we obtained are from the network layer and only provide limited information for the application layer. To identify the traffic of the different websites, we use their source and destination IP addresses. We face two difficulties: 1) the websites usually have widely distributed CDN servers and multiple IP addresses, and 2) some IP addresses may concurrently host more than one domain names. To handle these obstacles, we sample the IP addresses of the websites with the following steps: (1) We first perform some regular operations towards a website and use Wireshark [19] to capture every interacting packet. We thus collected, as many as possible, the related domain names of the website from these captured packets; (2) We look up the IP addresses for these domains with ‘nslookup’ initiated from different DNS servers in five geographically diverse cities, namely, Hong Kong, Shenzhen, Chengdu, Taiyuan, and Xiamen. Although we cannot collect all the IP addresses of each website, we maximally sample the traffic related to the behaviors of clients using these DNS servers; (3) We verify that these IP addresses to see whether they host any other domain names. We manually query reverse DNS website in [20]. The result shows that all the websites we have sampled have individual IP addresses. This is not surprising as all the websites we studied are flagship websites. For the two news websites, News_l and News_h, their news publication domain names also do not share IP address with their blogs and forums. In this case, we believe the packets we have sampled are all correlated with the websites that we are looking for.

IV. A General View of the Data Traffic

The first thing that we are interested in is the sheer amount of the traffic of different websites: in particular, the incoming flow (i.e., the data packets with the website as their destinations) and the outgoing flow (i.e., the data packets with the website as their source). Fig. 1 shows the result. We record the traffic volume every 30 minutes, and the data scales for 32 hours from 10:00pm June 9th to 6:00am June 11th, 2010. We observe that for all websites, the incoming and the outgoing flow follow a similar trend. That is, the traffic drops from 1:00am (the 3rd hour in the time line), and reaches the bottom at 5:00am (the 7th hour), while the flow recovers to normal level at 10:00am (the 12th hour) in the next day. This is not surprising since most people use the Internet during daytime.

The four types of websites all have a larger outgoing flow than the incoming flow. For example, in Fig. 1 (a) the outgoing flow of OSN_l is approximately 5.5 times of its incoming flow,



(a) The throughput of online social network site over the time

(b) The throughput of forums over the time

(c) The throughput of news websites over the time

(d) The throughput of search engine sites over the time

Fig. 1: The traffic of websites

and for OSN_h, the ratio is is 3.5:1. In Fig. 1 (b), we see that the ratios for two forums are 12:1 and 8.95:1 for BBS_l and BBS_h respectively. The news websites have ratios of 9.8:1 in News_l and 12.9:1 in News_h (see Fig. 1 (c)). In Fig. 1 (d), the search engines have ratios of 4.2:1 and 4.3:1 for SE_c and SE_w respectively. The large gap between the outgoing flow and incoming flow suggests most users in these websites serve as information consumers. The gap for OSNs and search engines are relatively smaller as they have larger interactions. OSN_h has the smallest ratio as it emphasizes on photos, resulting a larger upload traffic. The ratios of two search engines are very similar, showing that there is little variation on the ratio of user keyword upload packets and the return content packets.

As a summary, our observation is not very surprising. We quantify the magnitude, however, for different websites. The OSNs has a larger outgoing flow, (even higher than search engines for OSN_l), showing that users of OSNs are still mainly content consumers. The biggest difference between outgoing and incoming flow comes from the news sites, which further validates the user behavior in the Internet.

V. The Traffic Patterns of the Websites

We inspect the traffic flow of the websites related to different type of contents, such as texts and photos. We distinguish different traffic flows from different content by host names. The reasoning is that large scale websites usually have multiple web servers with different host names to improve the web viewing experiences, and different types of servers are used to provide different types of contents. In this regard, we differentiate four types of flows by host names: the text content download flow, the text message upload flow, the photo content download flow and the photo upload flow. Besides, we study the comment download and upload flow for news websites.

We are also specifically interested in the upload packets. The length of the upload packets provide additional information to scrutinize the incoming traffic differences from network layer. According to the traces from Wireshark [19], we divide the incoming packets of websites into three types: 1) Type I packets with packet length from 40 bytes to 60 bytes. These packets are the TCP control packets like ACK or FIN; 2) Type II packets with packet length from 61 bytes to 1399 bytes. These packets are largely the HTTP request packets such as HTTP Get and HTTP Post; 3) Type III packets with packet length greater than 1400 bytes and these packets are generally the packet segmentations involving in large data transmission. As Type III packets are caused by user data upload, they are our focus of study in this section.

A. OSNs and Search Engine Websites

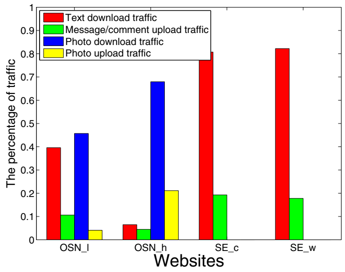
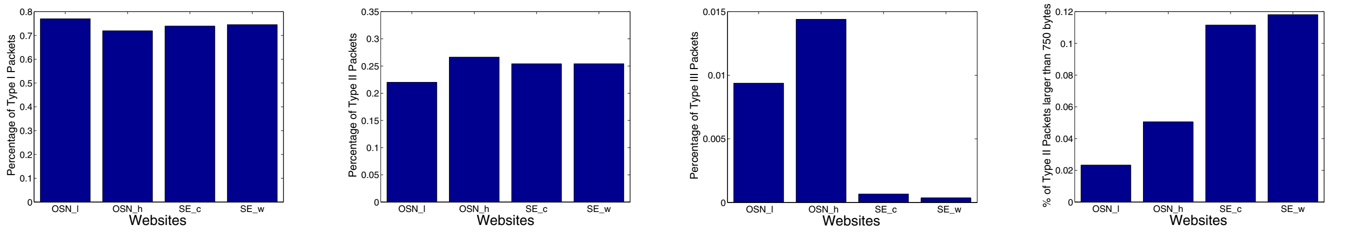


Fig. 2: The traffic decomposition of OSNs and search engines



(a) The percentage of the number of Type I upload packets

(b) The percentage of the number of Type II upload packets

(c) The percentage of the number of Type III upload packets

(d) The percentage of the number of Type II upload packets (> 750 bytes)

Fig. 3: Online social network and search engine

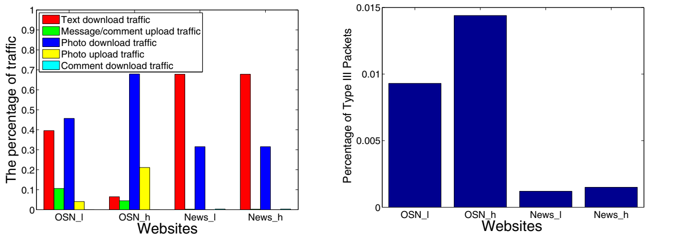
Fig. 2 plots the traffic decomposition of OSN_l, OSN_h, SE_c and SE_w. we have two observations. First, when we compare the traffic decomposition between the two OSNs, we see the OSN_h has a larger proportion of traffic on photos: the combined traffic of photo is greater than 90% with upload and download traffic 70% and 22% respectively. The photo upload and download traffic in OSN_l is only 50% of the OSN_l total traffic. This is because OSN_h emphasizes more on photo sharing, and it has better user interfaces for photo viewing. Second, compared with OSNs, the overall outgoing and incoming flow comparison ratios in search engines are more stable. As studied in the previous section, the outgoing and incoming flow ratios in search engine websites are around 4:1 and those in OSNs are 3.5:1 and 5.5:1. This indicates that search engines are stable in user behavior but in OSNs users act differently.

Fig. 3 (a) and Fig. 3 (b) plot the percentages of the number of Type I upload packets and Type II upload packets. We can see that the percentages of the two type packets do not vary between OSNs and search engines. Type I packet is about 75% in OSNs and search engines, and Type II packet is about 22.5% to about 26%. Note that we also see that the Type I and Type II packet percentages in OSNs are also close to that in the news websites as well as in forums. Due to the limitation of the space, we will not present the percentages of Type I packets and Type II packets in the rest of the paper.

Fig. 3 (c) compares the percentages of the number of Type III upload packets between the OSNs and the search engines. Clearly, the Type III upload packets in OSNs take about ten times larger than that in search engines. They are 0.9% (out of 467K packets) and 1.4% (out of 2.7 million packets) in OSNs, and they are less than 0.1% (out of 250K packets from both sites) in the two search engine websites. This shows that OSN websites have larger data uploads than search engine websites. It is not surprising as OSNs have longer articles and larger photo uploads, while the search engines only have short keyword inputs. To further examine the data uploads, we study the Type II packets with a size greater than 750 bytes. These packets also contain upload messages. The result is shown in Fig. 3 (d). The percentages in OSNs are 2.3% in OSN_l and 4.9% in OSN_h; and the percentages in search engine websites are 11.2% in SE_c and 11.9% in SE_w. The higher percentage of the number of Type II upload packets with a length over than 750 bytes in search engines can be explained by that the HTTP request packets in these sites are always enough to serve the keyword search.

B. OSNs and the News Websites

Fig. 4 (a) compares the traffic decomposition of the OSN websites and the news websites. We have two observations. First, the upload flow sizes are close to zero in News_l and News_h. This shows that for news websites, people always act as information consumers. It is not surprising if we take a look at the outgoing traffic of the comment publish servers from the two news websites: traffic from the comment publish servers from News_l is nearly zero and that from News_h is less than 10%. Such comment upload can be attributed to the replies of comment viewing. We show the Type III upload packets in Fig. 4 (b). We can also see that the percentages of Type III packets in news websites are small (0.12% and 0.14%).



(a) The traffic decomposition

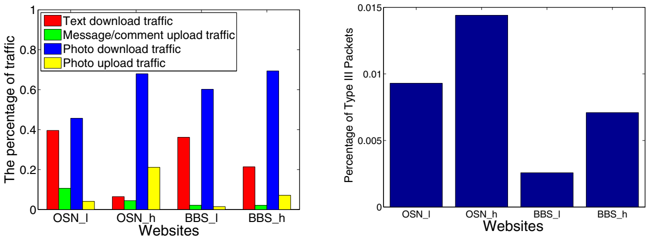
(b) The percentage of Type III packets

Fig. 4: Online social network and news websites

Second, we can see that the photo download traffic is larger than the text download traffic in the OSNs, but it is opposite in the news websites. Photo download traffic is 5% greater than text download traffic in OSN_l and it is seven times greater than text download traffic in OSN_h. The photo download traffic sizes in both news websites are only about half of the text download traffic sizes. This can be explained by the fixed news page format in the news websites. In addition, compared with large sizes of news pages with news contents, advertisements and JavaScript codes, the photos are compact and their sizes are fixed in these new websites. OSNs are quite different. The photos in the OSNs are often large in size and there are many photos embedded in one article. This difference in article formats between OSNs and the news websites causes the different text-photo download ratios between OSNs and the news websites.

C. OSNs and Forums

Fig. 5 (a) shows the traffic decomposition of the two OSNs and the two forums. The traffic patterns are similar in both types of sites. Especially the photo content also has



(a) The traffic decomposition (b) The percentage of Type III packets

Fig. 5: Online social network and forums

greater traffic than the text content in the forums and there is considerable uploading traffic. This may be explained as the users in forums have similar activities with the users in OSNs, such as viewing the articles and photos, posting comments and uploading photos. Nevertheless, the outgoing and incoming traffic ratios in forums are large (12.0:1 and 8.95:1). This shows that the user behavior of forums is similar to OSNs but users are slightly less active in information generation.

We present the percentage of Type III upload packets in Fig. 5 (b). The Type III upload packets in the forums has a greater proportions than that of search engines and news websites (2.6% and 7.2%, respectively) and is closer to those in the OSNs. This shows that the users in forums are considerably active. However, only the OSNs fully unleash the interaction between people. We can think the foundation of forums is posts (i.e., articles); this is similar to news websites. On the contrary, the foundation of OSNs is people.

D. Summary

In this section, we study the traffic features of the OSNs, and we compare these features to the traditional websites: search engines, news websites and forums. We see that the traffic in the network layer generally reflects the user behaviors in the application layer: 1) The OSNs and the search engines have the largest proportion of uploading traffic of the four types of websites; 2) compared with the search engines, the OSNs have a larger traffic percentage of large data uploads; 3) The news websites have the least percentage of uploading traffic, and most of the data transmission are caused by the text content transmission; 4) Forums websites have the similar traffic pattern with OSNs. But compared with OSNs, the upload traffic of forums are of moderately smaller percentage.

These phenomenon indicates the different user behaviors in these websites. We can see that the users are active as information generators in OSNs and fourms. Comparatively, users are much less active in news websites and search engines. Especially in news websites, it is interesting that the user comments take a minor part in the total traffic. Maybe adopting the idea of OSNs into the user comment module in the news websites can encourage the user activities.

In the view of the Internet infrastructure, the ONSs and the forums bring a close packet forwarding burden both sides, while for other two types of websites, the major burden comes from the outgoing traffic. This may provide us another perspective to improve the efficiency of the website gateways.

VI. Conclusion

In this paper, we studied the characteristics of OSNs from the angle of the traffic pattern upon Internet infrastructure. We collected the data from two backbone router in China Telecom. We observed that the OSNs and search engines have the smallest differences in the amount of incoming flow and the outgoing flow, as compared to the other types of traditional websites. Besides that, we found that the OSNs have a higher percentage of packets involving in large object uploading than search engines. We also compared the OSN traffic flow with the traditional news websites, and we found that the news websites have least percentage of incoming flow and most of the traffic are generated by text content; at last, we also found that the OSNs have a similar traffic style with forums but with a larger percentage of incoming flow.

While the current trend of researches on OSNs is still on the user behavior, we have shown in this paper that the traffic pattern in network layer is worth consideration as well. In the future, we are going to give further measurement on the traffic patterns of OSNs and more in-depth explanation. We hope our work can foster an interest in understanding the OSN flow features from the perspective other than the application layer.

References

[1] T. Benson, A. Anand, A. Akella, and M. Zhang, "Understanding Data Center Traffic Characteristics", In *Proc. WREN'09*, Barcelona, Spain, August 21, 2009.

[2] E. Chan, X. Luo, and R. Chang, "A Minimum-Delay-Difference Method for Mitigating Cross-Traffic Impact on Capacity Measurement", In *Proc. ACM CoNext'09*, Rome, Italy, December 1 - 4, 2009.

[3] W. Chen, J. Chu, and J. Luan, "Collaborative Filtering for Orkut Communities: Discovery of User Latent Behavior", In *Proc. ACM WWW'09*, Madrid, Spain, April 20 - 24, 2009.

[4] X. Cheng, C. Dale, and J. Liu, "Statistics and Social Network of YouTube Videos", In *Proc. IEEE IWQoS'08*, Enschede, The Netherlands, May 28 - 30, 2008.

[5] A. Joinson, " 'Looking at', 'Looking up' or 'Keeping up with' People? Motives and Uses of Facebook", In *Proc. ACM CHI'08*, Florence, Italy, April 5 - 10, 2008.

[6] B. Krishnamurthy, P. Gill, and M. Arlitt, "A Few Chirps about Twitter", In *Proc. ACM WOSN'08*, Seattle, USA, August 18, 2008.

[7] J. Kunegis, A. Lommatzsch, and C. Bauckhage, "The Slashdot Zoo: Mining a Social Network with Negative Edges", In *Proc. ACM WWW'09*, Madrid, Spain, April 20 - 24, 2009.

[8] K. Lai and D. Wang, "A Measurement Study of External Links of YouTube", In *Proc. IEEE Globecom'09*, Hawaii, USA, November 30 - December 04, 2009.

[9] C. Lampe, N. Ellison, and C. Steinfield, "A Face(book) in the Crowd: Social Searching vs. Social Browsing", In *Proc. ACM CSCW'06*, Banff, Canada, November 4 - 8, 2006.

[10] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney, "Statistical Properties of Community Structure in Large Social and Information Networks", In *Proc. ACM WWW'08*, Beijing, China, April 21 - 25, 2008.

[11] Y. Matsuo and H. Yamamoto, "Community Gravity: Measuring Bidirectional Effects by Trust and Rating on Online", In *Proc. ACM WWW'09*, Madrid, Spain, April 20 - 24, 2009.

[12] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks", In *Proc. IMC'07*, San Diego, California, USA, October 24 - 26, 2007.

[13] S. Sen, J. Vig, and J. Riedl, "Tagommenders: Connecting Users to Items through Tags", In *Proc. ACM WWW'09*, Madrid, Spain, April 20 - 24, 2009.

[14] J. Tang, M. Musolesi, and C. Mascolo, "Temporal Distance Metrics for Social Network Analysis", In *Proc. ACM SIGCOMM WOSN'09*, Barcelona, Spain, August 17, 2009.

[15] Y. Zhang, Z. Mao, and M. Zhang, "Detecting Traffic Differentiation in Backbone ISPs with NetPolice", In *Proc. ACM IMC'09,* Chicago, USA, November 4 - 6, 2009.

[16] http://www.Alexa.com, "Alexa".

[17] http://www.cisco.com/warp/public/732/tech/netflow, "Cisco NetFlow".

[18] http://tools.whois.net/whoisbyip/, "Whois By IP Address".

[19] http://www.wireshark.org/, "Wireshark".

[20] http://www.yougetsignal.com/tools/web-sites-on-web-server/, "You Get Signal- Reverse IP Domain Check".