

Towards Understanding the Social Characteristic of YouKu: Measurement and Analysis

Yongjun Li ^a, Chun You and Xudong Bao

School of Computer and Engineering, Northwestern Polytechnical University, Xi'an, P.R.China

^alyj@nwpu.edu.cn

Keywords: social network; social characteristic; measurement; YouKu.

Abstract. Online Social networking services are among the most popular sites and become the fast-growing business in the Internet. In-depth understanding the social characteristic of these networks can serve to optimize current systems, to design future social network based systems, and to eventually exploit the user base for commercial purposes. In this paper, we present a large-scale measurement study and analysis on the social structure of YouKu. Our results validate the *power-law*, *small-world* and *clustering coefficient* properties, present the correlation and difference among four *centrality* properties. Finally we discuss the utilization of these structural properties for the commercial purposes.

Introduction

Recently, online social networks have gained significant popularity and are now among the most popular sites on the Web. Unlike the Web, which is largely organized around content, online social networks are organized around users. Participating users join a network, publish their profile and any content, and create links to his friends. The resulting social network provides a basis for maintaining social relationships. Among them, YouKu (www.youku.com) is the most popular video-sharing site in China, which includes a social network.

The research on the graph structure of online social networks can serve to optimize current systems, to design future social network based systems, and to eventually exploit the user base for commercial purposes. For example, understanding the structure of video-sharing networks might lead to algorithms that can detect influential users. If we put advertisement on these users, the advertisement efforts will be promoted significantly. In this paper, we present a large-scale measurement study and analysis on the structure of YouKu. In addition to validating the *power-law*, *small-world* and *clustering coefficient* properties previously observed in other online social networks, we also provide insights into the *centrality* property. We find that some agreement among these centrality properties. For example, as the degree increases, so does the *betweenness* centrality. This suggests that high-degree nodes are critical for the connectivity and the flow of information in these networks. Another finding is those values provided by these centrality algorithms can not reflect the order of nodes consistently. To understand the centrality properties in-depth, we will concentrate on this in the future work.

The rest of this paper is organized as follows: We provide related works in Section 2. We describe our methodology for crawling these networks, and its limitations, in Section 3. We examine structural properties of the networks in Section 4, and discuss the implications in Section 5. Finally, we conclude in Section 6.

Related works

In this section, we describe studies of social network, as well as some work on complex network theory.

As described in [1], sociologists have studied many of properties of social networks. Among them, one of the most famous works is **six degree of separation** done by Milgram [2]. For an overview of social network analysis techniques, we refer the reader to the book by Wasserman and Faust [3].

As online social network gains popularity, many computer scientists concentrate on investigating the characteristic of online social network. Adamic et al. [4] study an early online social network at Stanford University, and find that network exhibits small-world behavior, as well as significant local clustering. In recent works, Alan et al. [1] presents a large-scale measurement study and analysis of the structure of four popular online social networks, and confirms the *power-law*, *small-world*, and *scale-free* properties of these works. Ahn et al. [5] analyze complete data from a South Korean social network (Cyworld), along with data from small sample crawls of MySpace and Orkut. Fu et al. [6] investigates the sample data from RenRen, a popular social network among university student in China. Jiang et al. [7] analyze the latent interactions in RenRen.

There has been much theoretical work on various classes of complex network, such as *Power-law* network, *scale-free* network, and *small-world* network. The online social networks analyzed in existed works have similar properties with these networks. For the detail description, we refer the reader to reference [1].

Crawling the YouKu

We crawled the YouKu site and obtained friendship information through the crawled web pages. The YouKu data we present was crawled on December 20th, 2009 and consists of 0.56 million users and 3.3 million links (friendship).

We consider all the YouKu users to form an undirected graph, where each user is a node and the friendship between these users is link in the graph. Our crawler use breadth-first search (BFS) to obtain the nodes and links included in the graph. Some users were selected randomly as seed set. The crawler read these seed user into a queue at the beginning of the crawl. The crawler selects the first user in queue and obtains its webpage, checks its friend list, and adds any new user to the queue. The crawler repeats the above operations until the queue is exhausted. Given a user, the crawler scrapes the crawled web page and obtains the interest information, including the user's friendship information.

For better insight, we analyzed and compared the structure of social network included in YouKu and YouTube, respectively. The YouTube data we present was obtained from <http://an.kaist.ac.kr/traces/IMC2009-kwak.html#dataset>.

Analysis of network structure

In this section, we characterize the structural properties of YouKu, and compare its properties with those previously observed in YouTube. Before we analyze these properties, we feel that need some clarification. As described in section III, YouKu users form an undirected graph. However, all the YouTube users form a directed graph. That is, if a user a is in the friend list of user b , then there is a directed edge from a to b [1]. In the following analysis, we will pay attention to this difference.

Power-law node degrees. We firstly examine the graph structure by considering the node degree distribution. As described in [1], many complex networks, including social network, have been shown to conform to power-law. Thus it is not surprising that YouKu also exhibit power-law degree distribution. However, as the existed analysis shown, there even exists some subtle difference in the degree distribution of social networks, owing to their various application backgrounds.

Figure 1 shows the degree distribution function for YouKu and the indegree/outdegree distribution function for YouTube, respectively. All of distribution functions show behavior consistent with power-law. The majority of nodes have small degree, and a few nodes have significantly higher degree. From their degree complementary cumulative distribution function (CCDF) as shown in figure 2, we can understand this property easily. For an illustrative purpose, we take the YouKu for example. Table 1 shows the number of nodes over their different degree sections. The number of nodes whose degree is less than 10 is 507093. However, there are only five nodes whose degree is more than 10000.

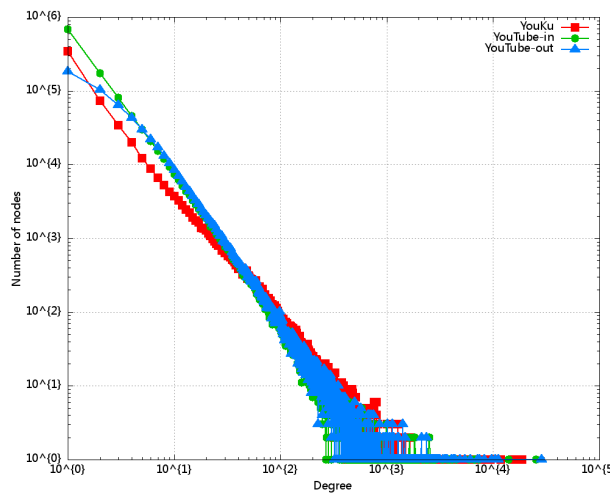


Figure 1. Log-log plot of degree distribution function

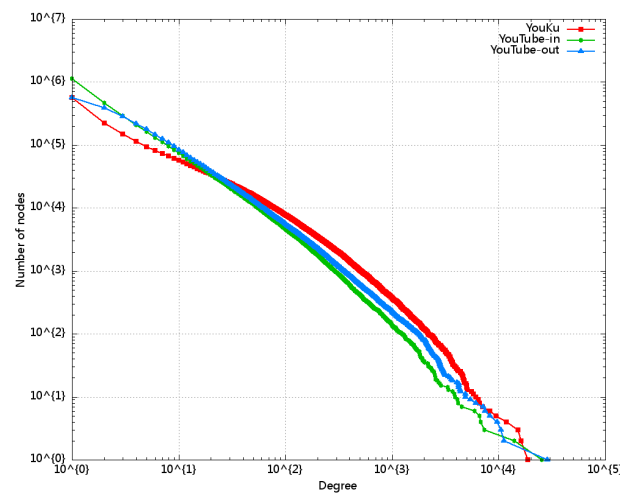


Figure2. Log-log plot of degree complementary cumulative distribution function

Table 1 Num. of nodes over different degree sections

	1-10	10-100	100-1000	1000-10000	>10000
Num. of Nodes	507093	48977	7403	369	5

Compared with YouTube, the slope of degree distribution function for YouKu is smaller. It means that the number of friends of the majority of users in YouKu is larger than the one in YouTube. From figure 2, we can find that the number of nodes of YouKu is larger than the one of YouTube when the node's degree is larger than 20. It is the major factor contributing to this difference that the social network included in YouTube is a directed graph and the one included in YouKu is an undirected graph.

Path length. Next, we investigate the property of path length distribution between users. We take the number of hops as path length between users. The graph of social network is not strongly connected, making some path length infinity. So it is difficult to calculate the path length exactly. Here we use the Largest Connected Component (LCC) for the measurements. Figure 3 shows the cumulative distribution function (CDF) of path length for YouKu and YouTube, respectively.

As shown in figure 3, two curves are very similar. The majority of path length is 6. The longest path length in YouKu is 8, and the longest one in YouTube is 12. The directed path in YouTube makes its length longer. However, this difference can not hinder two social networks exhibit the small-world property.

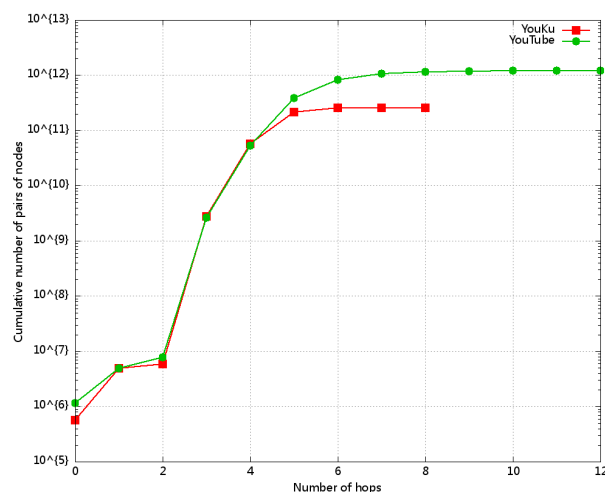


Figure 3. Cumulative distribution function of path length between users

Clustering coefficient. As described in [5], the clustering coefficient of a node is the ratio of the number of existing links over the number of possible links between its neighbors. Given a network $G = (V, E)$, a clustering coefficient, C_i , of node $i \in V$ is:

$$C_i = 2|\{(v,w)|(i,v), (i,w), (v,w) \in E\}|/k_i(k_i - 1) \quad (1)$$

where k_i is the degree of node i . The clustering coefficient of a node represents how well connected its neighbors are. Figure 4 shows how the clustering coefficients of nodes vary with node degree in YouKu and YouTube.

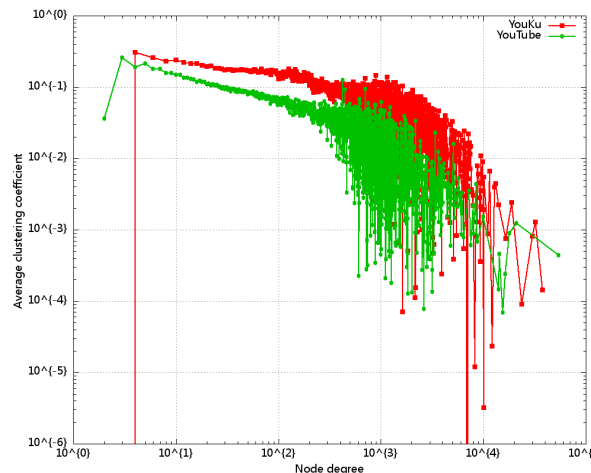


Figure 4. Mean clustering coefficient distribution

From figure 4, it is easy to see that both of the clustering coefficients of YouKu and YouTube are higher for nodes with low degree, suggesting that there is significant clustering among low degree nodes. The clustering coefficient decreased evidently with increase of degree of node. This property has also been found in other social networks, such as Flickr, Orkut. It has a natural explanation in social network: as the number of one's friends increases, so does the probability of his friends do not know each other. On the other hand, the clustering coefficient of social network is several orders of magnitude higher than the one of random graph. This is due to people tending to introduce one of his friends to another, and two of his friends are also friend with high probability.

From figure 4, we also easily see that the clustering coefficient of YouTube is lower than the one of YouKu. The major factor contributing to this difference is that the social network included in YouTube is a directed graph.

Centrality. In this subsection, we focus on centrality that captures the importance of individuals in a social network. Despite many definitions and implementations of centrality, no clear advantage is given to a particular paradigm for the study of social network characteristics [8]. Here we briefly define the four most used and well known centralities.

Betweenness: Freeman [9] defined betweenness centrality as the ratio of the number of shortest-paths that a node is part of, over all graph shortest-paths.

Eigenvector: Eigenvector [6] is to consider the importance of neighbors of a node; in other words, an important node has important neighbors in the graph topology.

Pagerank: Google's pagerank [11] is another centrality property, and it also being consider to a variant of eigenvector centrality [8].

Degree: it is the simplest form of centrality, and degree centrality assesses the importance of a node according to its degree.

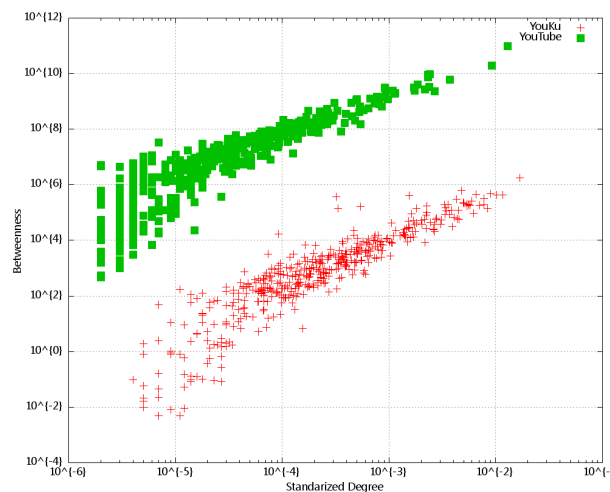


Figure 5. Relationship between Betweenness and Degree

Users expect that exists a centrality algorithm correctly reflect the order of nodes according to their importance. Are there some correlations among these centrality properties, particularly for social networks? To solve this problem, we present the relationship between the first three centrality property and degree centrality in figure 5, figure 6 and figure 7. From these plots, we can easily find some agreement among these centrality properties. That is, as the degree increases, so do the other three centrality properties. However, the values provided by these centrality algorithms can not reflect the order of nodes consistently. More specifically, the agreement between Pagerank and degree is more consistent. The detail research is our future work.

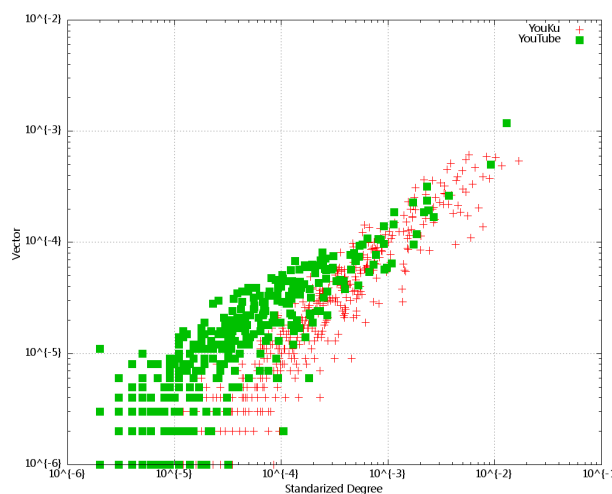


Figure 6. Relationship between Vector and Degree

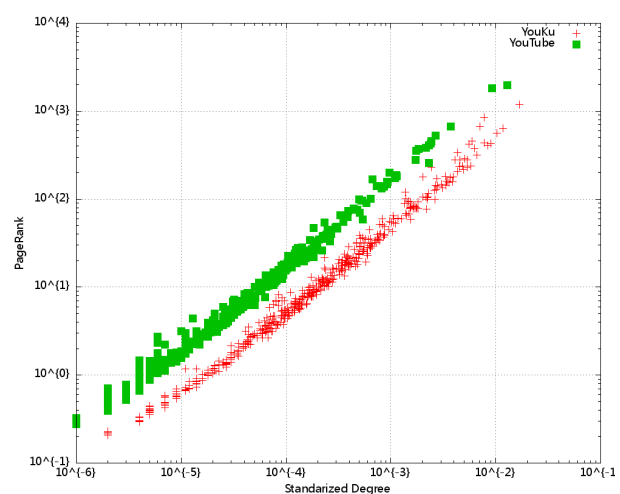


Figure 7. Relationship between Pagerank and Degree

Discussion

In this section, we discuss some implications of our findings. Our measurement results indicate that the degree distribution of YouKu fits the propriety of *power-law*, *small-world* which is the same as other online social networks. We also find that some agreement among these *centrality* properties. That is, the node with high degree plays a more important role than other nodes in online social network. These findings are likely applicable to many different online social network based applications. Due to limited space, here we concentrate on their effect on information dissemination.

Recently, many online events prove that the online social networks have been used as a means for rapidly and easily disseminating information. The *small-world* property implies that information seeded via these high-degree nodes will rapidly spread through the entire network within a few hops. This can be used in advertising. However, it is also well worth paying attention that these properties not only help positive information but also help negative information such as the spam or viruses disseminate quickly and widely.

Conclusions

We have presented an analysis of the structural properties of YouKu and Youtube. Our results validate the *power-law*, *small-world* and *clustering coefficient* properties previously observed in other online social networks, present that some agreement among four *centrality* properties. Another finding is those values provided by these centrality algorithms can not reflect the order of nodes consistently. We have outlined how these properties promote advertisement effects. To utilize *centrality* property effectively, we will concentrate on in-depth understanding *centrality* property in future work.

Acknowledgment

This work was supported partly by Open Research Fund from Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China (K93-9-2010-09).

References

- [1] Alan M, Massimiliano M, Krishna P G. Measurement and analysis of Online Social Networks. Proc. Internet Measurement Conference 2007 (IMC 07), ACM Press, Oct. 2007, pp.29-42.
- [2] Milgram S. The small world problem. PsychologyToday, 2(60), 1967.
- [3] S. Wasserman and K. Faust. Social Networks Analysis: Methods and Applications. Cambridge University Press, Cambridge, UK, 1994.
- [4] Adamic L. A., Buyukkokten O., Adar E.. A socialnetwork caught in the Web. First Monday, 8(6), 2003.
- [5] Ahn Y.Y., Han S., Kwak H., Moon S., and Jeong H.. Analysis of Topological Characteristics of Huge Online Social Networking Services. Proc. the 16th international conference on World Wide Web (WWW'07), Banff, Canada, May 2007.
- [6] Fu F, Chen X, Liu L et al. Social dilemmas in an online social network: the structure and evolution of cooperation. Physics Letters A, 2007, 371(1-2):58-64.
- [7] Jiang J, Wilson Ch, Wang X, Huang P, et al. Understanding Latent Interactions in Online Social Networks. Proc. Internet Measurement Conference 2010 (IMC 10), ACM Press, Nov. 2010.
- [8] Erwan L M, Gilles T. Centralities: Capturing the Fuzzy Notion of Importance in Social Graphs. Proc. 2nd ACM Workshop on Social Network Systems (SNS'09) , ACM Press, 2009
- [9] Linton C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):35–41, March 1977.
- [10] Bonacich P. Factoring and weighting approaches to status scores and clique identification. J. Math. Sociol. 2, pages 113–120, 1972.
- [11] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.