# Characteristics and evolution of content popularity and user relations in social networks

Claudia Canali, Michele Colajanni, Riccardo Lancellotti
University of Modena and Reggio Emilia
{claudia.canali, michele.colajanni, riccardo.lancellotti}@unimore.it

## Abstract

*Social networks have changed the characteristics of the traditional Web and these changes are still ongoing. Nowadays, it is impossible to design valid strategies for content management, information dissemination and marketing in the context of a social network system without considering the popularity of its content and the characteristics of the relations among its users. By analyzing two popular social networks and comparing current results with studies dating back to 2007, we confirm some previous results and we identify novel trends that can be utilized as a basis for designing appropriate content and system management strategies. Our analyses confirm the growth of the two social networks in terms of quantity of contents and numbers of social links among the users. The social navigation is having an increasing influence on the content popularity because the social links are representing a primary method through which the users search and find contents. An interesting novel trend emerging from our study is that subsets of users have major impact on the content popularity with respect to previous analyses, with evident consequences on the possibility of implementing content dissemination strategies, such as viral marketing*[1].

## 1. Introduction

Social networks represent innovative and constantly evolving applications that are shaping the Web content characteristics, user access patterns and content popularity. According to Nielsen Online's latest research [14], social networks are becoming the most popular applications on the Web, involving two-thirds of the online population. While the Web is largely organized around the consumption of content supplied by third party providers, social networks are centered around the users. The main user activity in social networks includes contacts, upload and exchange of user-generated content and links, novel forms of interactions, such as commenting/tagging contents or establishing friendship relations with other users through so-called *social links*. Differently from the Web, content popularity in a social network is highly affected by *social navigation*, that is, navigation through the social links, which represent a common way for users to search and find contents within the site. The difference of user behavior patterns between the social networks and the Web poses new issues and opens novel possibilities for content and system management strategies, including caching, dissemination, and even marketing. Our study is motivated by the observation that, to the best of our knowledge, no result has been published about the evolutionary trends of different social networks over multiple years. Most papers on social network characterization analyze static properties derived by snapshots of the site structure [13]. Few papers address the issue of network evolution over time and these studies are limited to one social network (e.g., [2, 6, 1]). This paper analyzes the present and evolutionary characteristics of two social networks: Digg, a popular site for bookmark sharing; YouTube, the world's largest video sharing site. The analysis is, as it is typical in related literature, out of the box because we have limited access to social network information. The study has required the implementation of a set of crawlers for each social network, that are able to exploit the public APIs of the social networks and to integrate missing information through the analysis of data available on the users Web pages. The crawling lasted two months for each social network and provided a statistically significant set of data with a dataset referring to about 2 million of users. We compare the data collected in our crawling with previous studies [13, 11], that provide a snapshot of these and other social networks back in 2007. The main findings of this paper can be summarized as follows.

- We confirm the growth in the size of the considered social networks in terms of number of users, shared contents and number of user social links.

1 The authors acknowledge the support of MIUR-PRIN project DOTS-LCCI "Dependable Off-The-Shelf based middleware systems for Large-scale Complex Critical Infrastructures".

- By analyzing the correlation between content popularity and user social links, we evidence the impact of social navigation in the considered social networks, and we show that the content uploaded by users with a larger number of social links tend to have much more visibility.
- We demonstrate that the social networks have changed over the last two years. They have now a more asymmetrical structure that increases the effect of social navigation over the content popularity, with a consequent significant impact on content management and dissemination strategies in the context of social network systems.

Our results provide content and system management algorithms with useful insights to address the issues of social networks, but also to exploit their novel characteristics. For example, an in-depth understanding of the social network structure and user behavior can support the design of novel strategies for content management (e.g., through selective content replication), that can face the challenges of flash crowds and slashdot effects. Furthermore, the knowledge of the user social networks and their evolution is necessary to detect which are the most influential users for content dissemination purposes, for example in the context of viral marketing.

The remainder of this paper is structured as follows. Section 2 describes our methodology for getting information from the considered social networks. Section 3 presents the main results of the analysis of the social networks. Section 4 discusses the related work. We conclude the paper in Section 5 with some final remarks.

## 2. Methodology of analysis

We now describe the methodology that we used to collect information about Digg and YouTube. We chose these two social networks because of their popularity and because they provide us with APIs that allow to view the social links of any user. Privacy policies of other popular sites, such as Facebook, prevent us to access an amount of user data that may be sufficient for our characterization.

In Digg and YouTube, the social links among users are directional: a user may invite other users to be his/her *friends*; on the other hand, a user may be invited in a friendship relation by other users, who become his/her *fans*. The social network is modeled as a directed graph where each user is a node, connected by social links to other nodes. Each node has outgoing links, that is friends of the user, and incoming links, that is fans of the user. We define the *in-degree* value of a node as the number of incoming links, and the *out-degree* value as the number of outgoing links.

For each user, we also consider additional information, such as content rating activity and uploaded contents. Furthermore, we evaluate the popularity of the contents uploaded by each user. To this aim, we consider for YouTube the number of views received by the user uploaded videos, while for Digg we consider the amount of submitted bookmarks that are marked as popular by the Digg site.

For the social network analysis it was not possible to obtain data directly from the YouTube and Digg site operators. Most sites are hesitant to provide even anonymized data, and signing non-disclosure agreements to obtain data from multiple competing sites is often unfeasible. For this reason, we chose to crawl the social network graph by accessing the public APIs provided by the sites. This approach is commonly used in literature, and gives us access to large data sets from multiple sites. However, crawling a large graph with millions of nodes and links causes issues that may hinder the efficacy of crawling, due to limitations in the social network APIs and to the management of the amount of traffic generated.

We implement two specific crawlers for the YouTube and Digg social networks, that collect data about users and their social links. Each crawler begins with a list of randomly selected users. Then, in each step, our crawlers explore the outgoing links of a not yet visited user to retrieve the list of his/her friends. Finally, we add the retrieved friends to the list of users to visit in the next crawling step. This approach to crawling follows a Breadth-First Search approach that allows us to reduce the bias in the collected data if compared to other methods, like Depth-First Search [13]. Table 1 shows the high level statistics of our crawler activity for YouTube and Digg. For both networks the crawling has been carried out in the second half of 2009. For Youtube we collect data on nearly 2 million users and more than 10 million social links, exploring nearly 1% of the network, while for Digg we crawled more than 10% of the social network users.

| Parameter | YouTube | Digg |
|---|---|---|
| Period of crawling | Jul.-Aug. 2009 | Oct-Nov. 2009 |
| Number of users | 1,708,414 | 349,035 |
| Number of social links | 12,935,561 | 3,212,454 |

**Table 1. High level statistics of crawlers**

It is worth to note that, due to the large number of social links of each user, our crawler had to address a problem related to limitation of the APIs provided by the social network applications: the list of outgoing links returned for each user is truncated to 100 entries for YouTube and Digg. To solve this problem, we exploit the information available on the user home page, that contains a complete list of his/her social links. Our crawlers identify the users where the API returns an incomplete list of outgoing links and integrate it by parsing the HTML code of the Web page.

A further problem that we experienced in our crawling

activity was related to the countermeasures against Denial of Service (DoS) attacks adopted by the social network infrastructures. For example, YouTube servers measure the amount of requests generated from each IP address and, if a security threshold is exceeded, the servers filter out subsequent requests from the same IP for a period of 20 minutes. When the crawler extracts information from user Web pages, the protection against DoS attacks may hinder the ability to crawl the social network. To solve this problem, our crawlers exploit multiple network interfaces and can distribute the crawling traffic among multiple IPs.

## 3. Result analysis

We present the most important results of our analysis concerning the YouTube and Digg social networks. We initially evaluate the social network size, then we study the impact of social navigation on content popularity, and finally we analyze the evolution of the social network structure.

### 3.1. Growth of network content and structure

The evolution of the YouTube and Digg networks over the last two years is studied in terms of number of shared contents and user social links. Table 2 shows the absolute number of users registered to each social network, the average number of daily uploads of resources, the average number of monthly unique users visiting the site and the median number of social links (incoming and outcoming) per user. As a measure of the social links, we consider the median instead of the average value, because the former is statistically more significant in contexts characterized by heavy-tailed distributions [13].

Table 2 shows a clear growth for all these parameters over the last two years. In particular, the number of users increased by one and two orders of magnitude for Digg and YouTube, respectively. For both social networks the amount of shared content is more than doubled, and the number of monthly unique visitors increased by a factor of five. Furthermore, users are better connected with each others, with a growth in the social links of nearly two orders of magnitude.

This growth demonstrates the increasing importance of social networks for user online activity. The increase in the number of users and social links augments the impact of social navigation on content popularity. Hence, it is important to understand *how* user behaviors and network structure have evolved in order to achieve useful insights about how to manage information and systems supporting social networks and possible related applications, such as publicity and marketing.

### 3.2. Impact of social navigation on content popularity

To understand the impact of social navigation on content popularity, we should evaluate to which extent users follow social links to look for contents within the site. To this aim, we evaluate the popularity of the contents provided by each user with respect to the number of user social links. As a measure of the content popularity, we consider for YouTube the number of views to the content uploaded by the user, while for Digg we consider the number of bookmarks submitted by the user that are marked as *popular* by the Digg site.

Figures 1(a) and 1(b) report scatter plots which show the popularity of contents versus the in-degree of the supplier user for YouTube and Digg, respectively. For the most part, a higher user in-degree seems to correspond to a higher popularity of the supplied contents for both the social networks. This impression is confirmed by the Pearson correlation coefficient between user in-degree and content popularity reported in Table 3: a value of the Pearson coefficient $> 0.75$ indicates a high correlation degree, thus meaning that contents uploaded by a user with high in-degree are likely to receive a large number of accesses. Similar effects were identified in the case of Flickr [6], and Digg [11] in 2007, while it was never demonstrated for YouTube. Our results confirm social navigation to be a common trait of social networks. Furthermore, for the Digg case, we found a higher value of the correlation coefficient between user in-degree and content popularity with respect to 2007. This suggests that users exploiting social links to find interesting contents are continuously augmenting.

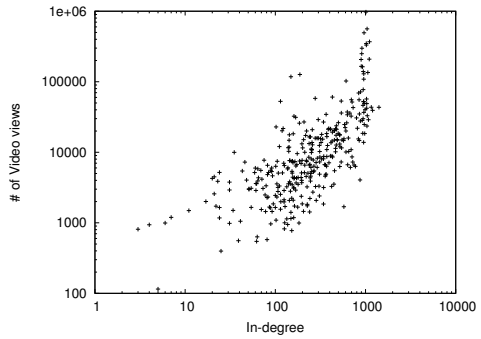
It is important to note that social navigation has a significant impact on the strategies for content management and on the opportunity for deploying novel applications exploiting social networks for content dissemination and marketing.

With respect to content management, the strong correlation between social links and content popularity suggests that social navigation is and will likely be in the future a key factor to access contents in social networks. Knowledge of the user social links allows a system manager to predict whether a newly uploaded content is likely to become popular, without the need to rely on past access history for that content [4]. This knowledge may assume a critical relevance for content management in social networks, where slashdot effects and flash crowd can shape in very short time the access patterns to the server infrastructure and can represent a challenge even for existing distributed server infrastructures, such as CDNs [3].

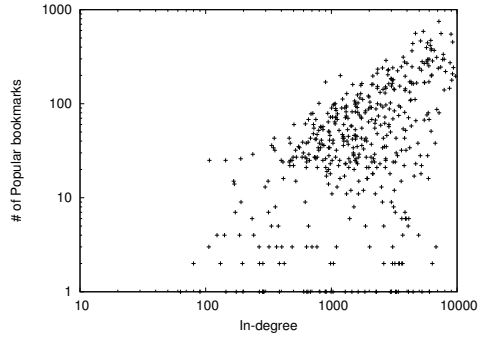
If we consider the potential impact of social navigation on novel applications, such as viral marketing, the high correlation between user in-degree and content popularity suggests that some users, thanks to their social network, can

| Network | Year | # of Users | Daily Uploads | Monthly Unique Visitors | Social Links per User |
|---|---|---|---|---|---|
| Youtube | 2009 | 258M [10] | 200,000 [18] | 92M [8] | 466 |
| Youtube | 2007 | 2.86M [7] | 65,000 [9] | 20M [9] | 3 [13] |
| Digg | 2009 | 19M | 18,000 | 43M [8] | 1170 |
| Digg | 2007 | 2.7M [16] | 8,000 | 7M [16] | 50 [11] |

**Table 2. Evolution of Social network size**




(a) YouTube




(b) Digg

**Figure 1. Content popularity versus user in-degree**

be very effective in disseminating information. For example, in the case of an advertisement campaign, identifying a limited set of users that can provide high visibility to the advertisement message is a critical task for the campaign success. From a user point of view, this also means that having a high in-degree may result in an economic benefit, if the user hosts or supports the advertisement in his homepage. We can conclude that social networks may open a novel market where the social links of each user are a valuable asset for content dissemination purposes.

As a further point, while the correlation between user in-degree and content popularity is high, there is no significant correlation between the content popularity and other parameters characterizing the user in the network, such as user out-degree and content rating activities. The lack of

| Network | Parameter | Correlation |
|---|---|---|
| Youtube | in-degree | 0.87 |
| Youtube | out-degree | 0.10 |
| Youtube | favorites | 0.04 |
| Digg | in-degree | 0.83 |
| Digg | out-degree | 0.19 |
| Digg | diggs | 0.44 |

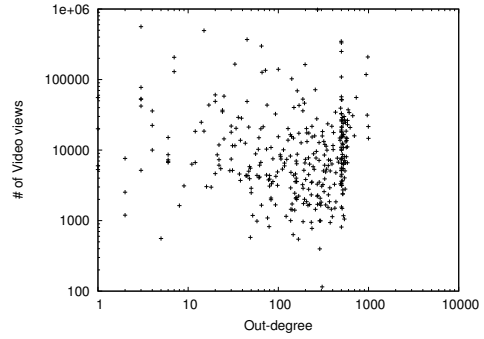
**Table 3. Correlation with content popularity**

correlation is demonstrated by the low values of the Pearson coefficient in Table 3 and by the scatter plot of the user out-degree vs. content popularity in Figure 2. In order to exploit social network structure for content dissemination, this result is complementary to the importance of user in-degree for content popularity. Indeed, actions directly carried out by the users (such as creating new outgoing links, voting or marking contents as favorites) have little effect on the popularity of the uploaded content. Hence, each application aiming to exploit the user social links for dissemination purposes should rely on existing social connections, because it is not simple to create *ex-novo* a user profile and make that user popular through some direct actions.

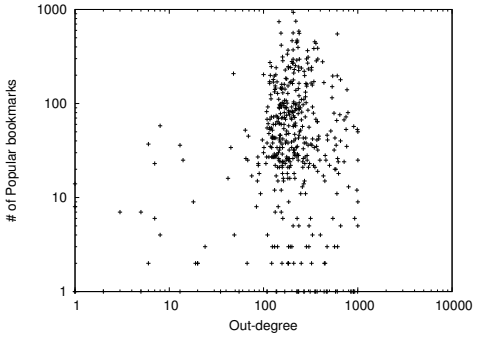
### 3.3. Network structure evolution

We now analyze how the structural properties of the social networks have evolved over time by comparing present results on YouTube and Digg with previous measurements on YouTube, Flickr and LiveJournal dating back to 2007 [13].

We examine the network structure by initially considering the user link distribution. Figure 3 shows the complementary cumulative distribution function (CCDF) on a doubly logarithmic axes for the in-degree and out-degree of social networks in 2007 [13] and 2009. This log-log plot is commonly used to describe statistical distribution properties and verify whether a distribution follows a power-law behavior [5, 13]. All the curves in Figure 3 show a behavior consistent with a power-law network for both incoming and outgoing links: the majority of nodes have small degree, and a few nodes have significantly higher degree. If we pass to observe the curves related to the 2009 dataset, we note two main differences with respect to previous results. The 2009 curves are shifted to the right with respect

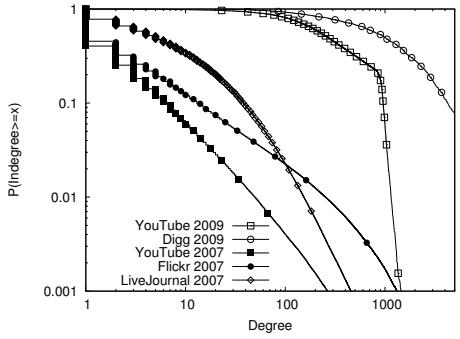



(a) YouTube

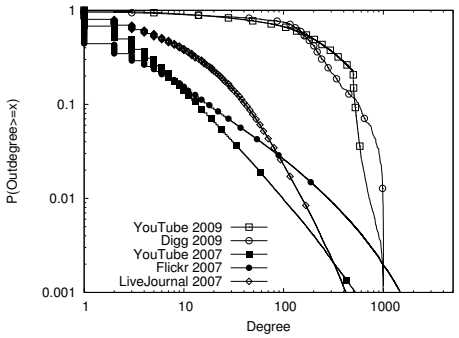



(b) Digg

**Figure 2. Content popularity versus user out-degree**




(a) In degree




(b) Out degree

**Figure 3. Out- and In-degree**

to the curves related to the 2007 datasets. This is due to a significant increase in the networks size in terms of number of social links for each user, as anticipated by the results in Table 2. Moreover, the curves present a truncated behavior, that is much more evident now with respect to the 2007 dataset, with the leftmost part of the curve that is nearly constant and a sudden drop at the right side of the curve. The truncated tail in this kind of networks has been already observed in other studies on social networks [9]. A pure power-law behavior would imply the existence of a very few users with a so much high degree that they are connected almost to the entire social network. On the other hand, the effect of a truncated tail poses a limit to the maximum achievable user degree, implying a more irregular social network structure where a limited set of users show a similar high degree level. Such irregular structure may arise issues in the identification of the best-connected users, that we expect to play a key role for content management, dissemination and marketing purposes (see Section 3.2).

Another important characteristics of the social network structure is the possible (a)symmetry between user in-degree and out-degree. The analysis in Section 3.2 shows that the user in-degree is more relevant than his out-degree in determining the content popularity. It is then important to identify the population of users that may be critical sources of popular information (users with high in-degree) and understand whether they are the same as the population of the most active users (users with high out-degree). To this purpose, we evaluate whether users with high in-degree also have high out-degree. Figure 4 shows the overlap between the top x% of users ranked by in-degree and out-degree. In the 2007 datasets the top 20% of the users ranked by in-degree had more than 80% overlap with the top 20% users ranked by their out-degree. This high overlap in the set of most well connected users according to their in- or out-degree suggests a symmetric social network structure. Hence, we can conclude that in the social network snapshot dating back to 2007, the most active users tended to be critical sources of popular information.

On the other hand, in the 2009 dataset, we observe a significantly lower overlap between users ranked by in- and out-degree. For both considered social networks, the top 20% of the users have an overlap below 45%. This result suggests an interesting novel trend of the social network structure that is becoming more asymmetric.

To confirm the asymmetry in link distribution of current social networks, we examine the in-degree and out-degree of individual users. Figure 5 shows the cumulative distri-

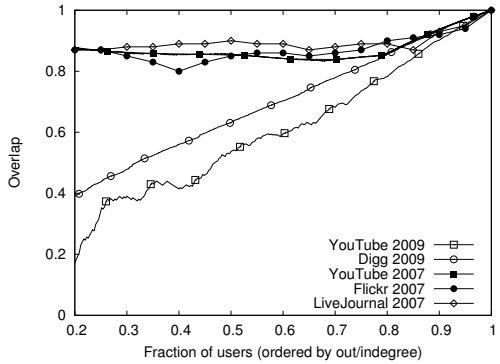



**Figure 4. Overlap of top x% of users ranked by out- and in-degree**

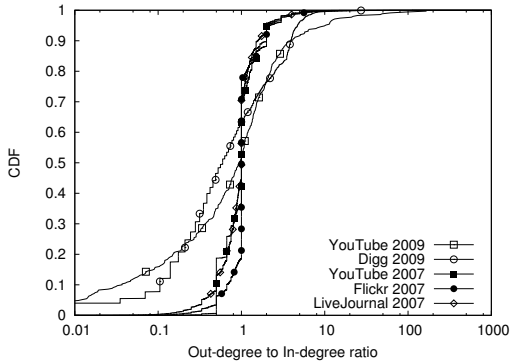



**Figure 5. Out-degree to In-degree ratio**

bution of the ratio between user out-degree and in-degree. If we consider the 2007 datasets, we observe that a large fraction of users has an out-degree to in-degree ratio close to 1 for both Digg and YouTube. On the other hand, we observe that the 2009 dataset is characterized by a much higher variability in the out-degree to in-degree ratio, thus confirming the evolution of current social networks towards a more asymmetrical structure. This result determines new challenges for content management algorithms exploiting social information and even for effective content dissemination strategies based on the identification of key users. Unlike the traditional Web and initial social network sites, there is the need of proposing novel methodologies and algorithms that can cope with an irregular social link structure spanning across many users.

From Figure 5 we also note that in the 2009 dataset a significant fraction of users has an in-degree much higher than its out-degree. This means that there are users that act as *hub* for incoming links. Due to the already demonstrated correlation between content popularity and social network structure (see Section 3.2), we can conclude that the contents uploaded by these hub users have more probabilities to become popular. From a content management point of view, this information is very important, because it may help the identification of a set of contents that should be replicated faster than others, as they are more likely to generate flash crowds and slashdot effects. Furthermore, hub users may also play a key role in content dissemination strategies, because they are the first users that should be exploited to guarantee a high visibility to newly added information.

## 4. Related Work

The results on the characterization of social networks consider three main areas: social network workload [9, 7, 5], social network structure [2, 1, 13], and effects of social navigation [12, 17, 11].

The most relevant studies on workload characterization refer to YouTube, where serving video content represents a challenge due to resource size and image quality requirements. Gill *et al.* [9] track YouTube transactions in a campus network and consider content access patterns and video characterization in terms of size and category. The paper in [7] extends [9] by measuring global properties through a larger set of considered videos. Cha *et al.* [5] present an extensive analysis of the video distribution, age, and content duplication in YouTube. We exploit some of these results to evaluate how the workload characteristics of the YouTube workload have evolved over the last two years. We then expand the analysis to evaluate the correlation between content popularity and network structure.

A characterization of the social network structure was initially proposed in [2, 1], that analyze the evolution of one site over time, and in [13], that compares the properties of static snapshots of multiple social networks. In [2] it is presented a study about the user group formation and evolution in LiveJournal and models for group evolution are proposed. The authors in [1] analyze Cyworld, that is the largest Korean social networking site. To the best of our knowledge, this is the only study on social networks that is based on internal data obtained from the CyWorld operator. Thanks to this amount of information, the authors were able carry out an in-depth study of the site and of its evolution over time.

Our paper extends the analysis of the evolutionary trends beyond the limitation of one social network, and aims to identify common trends in the evolution of two social networks over a period of two years. Mislove *et al.* [13] present some structural properties in a snapshot of multiple social networks. The results confirm the power-law, small-world and scale-free properties of the considered online social networks, and also reveal a high level of link symmetry among users. Our analyses evidence that significant changes occurred in the social network structure. In particular, we show a novel trend that is leading the distribution of user links towards a larger and more asymmetrical topol-

ogy. These characteristics have possible relevant effects on the distribution function of the content popularity.

More recently, several studies have been devoted to the impact of social navigation on content popularity in social networks. For example, [15] shows that, in e-commerce systems, the knowledge of the other customers choice affects the collective decision of large groups of users. The same behavior has been verified within some social networks, such as Flickr and Digg [12, 17, 11]. Lerman and Jones [12] consider the Flickr site and show the correlation between the number of incoming links of a user and the popularity of the uploaded images in terms of number of views. Also the study in [17] concerns the Flickr network and analyzes the users behavior in relation to the age of the uploaded contents. In [11] social navigation has been shown within the Digg network. Our study evaluates whether the impact of social navigation on content popularity has increased in the Digg social network over the last two years, and verifies for the first time the evidence of this behavior within the YouTube social network. Furthermore, by considering the combined effects of social navigation and of the evolution of the network structure, we derive useful insights for the design of content management, content dissemination and marketing strategies for systems supporting future social networks.

## 5. Conclusions

This study characterizes the structure of social networks and their evolution over a two year period. We consider two popular social networks, Digg and YouTube, and compare 2009 results against those referring to 2007. Our analyses confirm that social networks are experiencing a fast growth in terms of shared content and number of social links among users. We also show that social navigation has a primary role in determining the user access patterns and content popularity. We observe that the social network structure is becoming highly asymmetric, with a set of *hub* users that are characterized by an in-degree far higher than their out-degree. We expect these hub users may play a critical role in current and future social networks. From the point of view of content dissemination, these users can be extremely useful because their content are likely to gain high popularity thanks to social navigation. If we consider applications of social networks, such as viral marketing, hub users are expected to play a key role to disseminate information through their social links. Hub users are important also for the design of content management strategies, because their uploads can generate slashdot effects or flash crowds.

## References


[1] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *Proc. of the 16th International Conference on World Wide Web (WWW'07)*, May 2007.

[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD'06)*, Aug. 2006.

[3] R. Buyya, M. Pathan, and A. Vakali, editors. *Content Delivery Networks*. Springer-Verlag, 2008.

[4] C. Canali, M. Colajanni, and R. Lancellotti. Hot set identification for social network applications. In *Proc. of the 33rd Annual IEEE International Computer Software and Applications Conference (COMPSAC'09)*, Jul. 2009.

[5] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proc. of the 7th ACM SIGCOMM conference on Internet measurement (IMC'07)*, San Diego, CA, Oct 2007.

[6] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi. Characterizing social cascades in Flickr. In *Proc. of the 1st Workshop on Online Social Networks (WOSP'08)*, Aug. 2008.

[7] X. Cheng, C. Dale, and J. Liu. Statistics and Social Network of YouTube Videos. In *Proc. of 16th International Workshop on Quality of Service (IWQoS'08)*, Enschede, The Netherlands, Jun. 2008.

[8] Compete. Search Analytics Research Report, 2009. – http://siteanalytics.compete.com/.

[9] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. YouTube traffic characterization: A view from edge. In *Proc. of Internet Measurement Conference (IMC'07)*, Oct. 2007.

[10] L. Lake. YouTube: Social Media Marketing via Video. Research Report About.com Marketing, 2009.

[11] K. Lerman. Social Information Processing in News Aggregation. *IEEE Internet Computing*, 11(6):16–28, 2007.

[12] K. Lerman and L. Jones. Social Browsing on Flickr. In *Proc. of ICWSM Conference*, Mar. 2007.

[13] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proc. of the 7th ACM SIGCOMM conference on Internet measurement (IMC'07*, Oct. 2007.

[14] Nielsen Online Report. Social networks and blogs now 4th most popular online activity. 2009.

[15] M. J. Salganik, P. S. Dodds, and D. J. Watts. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science*, 311(5762):854 – 856, Feb. 2006.

[16] E. Schonfeld. Digg Nearly Triples Registered Users In a Year, Says Sleuth Programmer. Research Report TechCrunch, 2008.

[17] M. Valafar, R. Rejaie, and W. Willinger. Beyond friendship graphs: a study of user interactions in Flickr. In *Proc. of the 2nd ACM Workshop on Online Social Networks (WOSN'09)*, Aug. 2009.

[18] M. Wesch. YouTube Statistics. Digital Ethnography, Kansas State University, 2009.