Analysis of MySpace© User Profiles*

Luisa Massari

Dipartimento di Informatica e Sistemistica
Università di Pavia
I-27100 Pavia, Italy
massari@unipv.it

Abstract. Online social networks have attracted millions of users, who have integrated social network web sites into their daily life. Users participate to the changes and to the evolution of these sites because they are producers and reviewers of contents that help them to maintain the existing social relationships, make new friends, collaborate and enrich experiences. This paper presents a study of the characteristics of the users of MySpace© web site, with the objective of studying relationships and interactions among users and deriving hints about their behavior. The analysis relies on data collected by monitoring the web site for twelve weeks. Typical user behaviors have been derived and classes of users characterized by different levels of participation to the social network have been identified. In particular, the analysis reveals that most of the users actively participate to the social network and specify many personal details. Social networks web sites allow access to such details; the sharing of information about users and their relationships can lead to non-ethic online activities, which threat the privacy and the security of users themselves.

1 Introduction

The Internet has highly increased the ability of individuals to meet, interact, and keep in contact with other individuals having common interests. Moreover, the introduction of Web 2.0 (O'Reilly 2007) has encouraged social interactions by offering integrated services, information and communication tools, such as, blogging, photo and video sharing, organization and report of offline social events, which make it easier for users to generate and share web content and applications, and increase the sense of online social community.
A number of social networking web sites has emerged (e.g., Facebook, Flickr, MySpace©, Orkut, YouTube, LinkedIn), involving both business and private environments, and enabling millions of users to simplify collaboration and data exchange, to increase business and reduce costs, to support education, to make new friends and interact with each other. Hence, people are linked within a social network because they work together, share interests or lifestyles.

* This work has been supported by the Italian Ministry of Education, Universities and Research (MIUR) under the PRIN Project.

Services offered within social network web sites imply the initial creation of a profile. The profile includes users age, location, and interests; moreover, most sites encourage users to enhance their profiles by uploading photos, multimedia contents and applications. Profiles often provide an accurate image of the personality of the users, and hence they can be source of potential risk for privacy. Profiles can be visible to anyone or to a restricted set of friends, according to users choice, and can be downloaded and stored by third parties, creating a database of personal data. This obviously poses many ethical issues, related to the access and use of such data for inappropriate activities, such as, sharing users data with advertisers for business purposes, or spamming.

Recent studies have focused their attention on the analysis of online social networks, on their contents and on users characteristics. In (Kumar et al. 2006) the analysis of the evolution of the structure of social networks shows that users are characterized by different behaviors. There is a large number of passive users who do not contribute to the enrichment and the evolution of a social network, whereas there is a small number of very active users. In (Cha et al. 2008) the mechanism of information exchange and the rules with which the contents are spread over social networks are described by means of epidemiologic models. Some studies (Mislowe et al. 2007; Saha and Getoor 2008) use graphs to describe social networks structural properties and to provide measures on the proximity between groups of users. The impact of social networking services is addressed in (Aguiton and Cardon 2007). Authors show that the majority of collaborations among users results from the opportunities of interactions offered by the services available on the sites. Some papers have addressed the characterization of the technological aspects of the workload of social networking web sites, in particular of sites offering specific services, such as, YouTube for video-sharing (Cha et al. 2007; Gill et al. 2007; Halvey and Keane 2007; Cheng et al. 2008), Wikipedia for the creation of the so called wikis (Urdaneta et al. 2009), and blogs (Cohen and Krishnamurthy 2006). These studies outline the peculiarities of these new types of workloads compared with the characteristics of traditional web workloads.

This paper presents a study of MySpace© web site, with the objective of analyzing the users behavior and participation to social networks and the content they upload to the web. The study is based on the analysis of user profiles, that is, the web pages that specify the identity, tastes, personal details, and cultural interests of each user and than contain links to their friends, to the uploaded content, together with some comments. The study of user profiles is a crucial topic in understanding relationships and connections among users, and in deriving hints about their social interactions. In the first phase of this study, the user profiles have been analyzed with the aim of identifying the various types of information and content. Each profile has been described by means of parameters about user characteristics and social behavior. Statistics and groups of profiles exhibiting similar characteristics have been derived.

The paper is organized as follows. In Section 2, an introduction of MySpace© web site is given and the structure of the user profiles and the type of contents are explained. The preliminary analysis of the profiles

and the analysis of comments associated to each profile are presented in Section 3. In Section 4, clustering techniques are applied in order to identify classes of users with similar behavior. Finally, in Section 5 some conclusions and future work directions are given.

2 Structure of MySpace©

MySpace© (MySpace 2009) is currently one of the most popular social networking web site, accounting for about 117 million worldwide unique users per year (comScore 2009). Moreover, MySpace is one of the few social network sites that allow access to user profiles. Users of MySpace© join the site by registering, that is, by filling a profile which specifies their personal identity, and by accepting the "Privacy Policy" (MySpace Privacy 2009), which defines criteria for use and sharing of data specified by users and stored by MySpace©, and for protecting safety and security of users. The identity of a user consists of personally identifiable information, that is, e-mail, full name, postcode, sex and birthday. Privacy settings can be customized to restrict access and contacts from other users and the profile visibility. Users may personalize their profiles by customizing them and changing default layout and settings. Users can specify the style for their pages, insert information as free text, and provide and store non-personally identifiable information, such as, tastes, cultural interests, hobbies, lifestyle choices, groups they belong to. Users upload images and videos, they can make them visible to others, set up contacts and participate in groups, get in touch with other users by sending and accepting friend requests. Indeed, user profiles connect to other user profiles through friend relationships and messaging mechanisms, contain comments and discussions coming from other users, and links to the uploaded content.

3 Preliminary analysis

The analysis of the behavior of the users of MySpace© relies on their profiles. These profiles have been collected by crawling the MySpace© web site for a period of twelve weeks from January to March 2008. About 1.9 millions of user profiles have been captured and analyzed.
Each profile is described by parameters that specify details about the user (e.g., age, location, job, cultural interests, political views), about his popularity within the online social network (e.g., number of friends, number of comments made by other users), and about his activity (e.g., date of the last access to the site, number of uploaded videos, audios, and images, number of subscriptions to groups). In the following analysis, six parameters have been considered, namely: sex, age, number of friends, number of comments, number of images, and number of links. Moreover, another parameter, that is, the amount of details, has been defined to summarize the presence in the profile of other personal information. Indeed, some data, such as, sentimental status, ethnicity, religion, education, children, income, sexual orientation, body type, scope of registering,

occupation, are optional and users do not need to specify all of them. The amount of details expresses, as a percentage, the number of information specified by the users.
Statistics and distributions of the parameters have been derived to characterize the profiles. Table 1 summarizes the basic statistics of the parameters chosen to describe user profiles. Users are in general young: average age specified by users in their profiles is 27.13 years, with the median and the 90th percentile equal to 23 and 37, respectively. The distributions of all the parameters are characterized by a long tail, that is, a small number of profiles is characterized by very large parameter values. By looking at the amount of details provided by the users in their web pages, it comes out that, on average, users tell a lot about themselves, that is, about 52% of possible details. Moreover, the average values of the number of friends and of comments denote users having good social relationships: 50% of the profiles have links to up to 37 friends and contain up to 19 comments. On average, user profiles contain 226 comments. The maximum number of comments is 24,568, and the 90th percentile is equal to 590. On the contrary, the number of images and the number of links in the profiles are low.

	mean	min	max	st.dev	median	90th percentile
Age	27.13	14	108	0.5	23.0	37.0
Number of friends	138.45	2	29,106	530.5	37.0	294.0
Number of comments	226.56	0	24,568	588.5	19.0	590.0
Amount of details	0.52	0	1	0.3	0.6	0.9
Number of images	4.19	0	874	11.7	0.0	11.0
Number of links	6.68	0	984	14.7	3.0	16.0

Table 1. Basic statistics of the parameters chosen to describe user profiles.

Female users represent 47% of the total user population analyzed in our study. Their average age is only slightly lower than males age. Figure 1 shows the distribution of the age specified by users in the profiles, broken down into males and females. Due to the long tail of the distribution, the figure shows ages up to 50, taking into account about 95% of the total user profiles. As can be seen, the shape of the distribution is the same for males and females, independently of the age. The distribution confirms a large presence of young users: users up to 20 years old account for about 25% of the total number of users, and those younger than 18 years old represent about 10%. Figure 2 shows the distribution of the number of friends associated with the various profiles. The figure shows the distribution up to 30 friends, taking into account 47% of the total profiles. Although

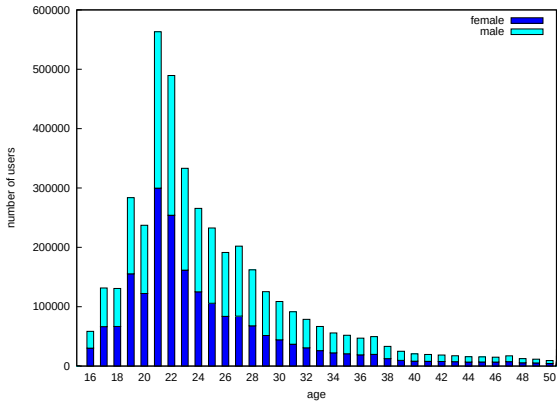

Fig. 1. Distribution of the age of the users.

no correlation between number of friends and number of comments has been discovered, a similar distribution has been derived for the number of comments in a profile (see Fig. 3).

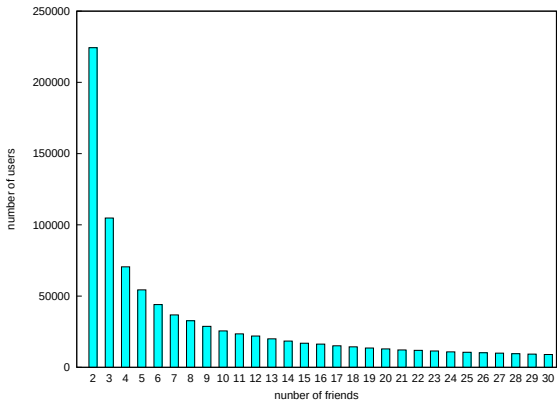

Fig. 2. Distribution of the number of friends associated with the various profiles.

The figure shows the distribution of the number of comments up to 39. The distribution takes into account about 50% of the analyzed profiles. The tail of the distribution is long; 90th percentile is equal to 728, and, as shown in Table 1, maximum value is 24,568. On average, a user profile contains 273.38 comments; moreover, 11% of users have one comment, and 6% only two comments. This denotes tight interactions among users.

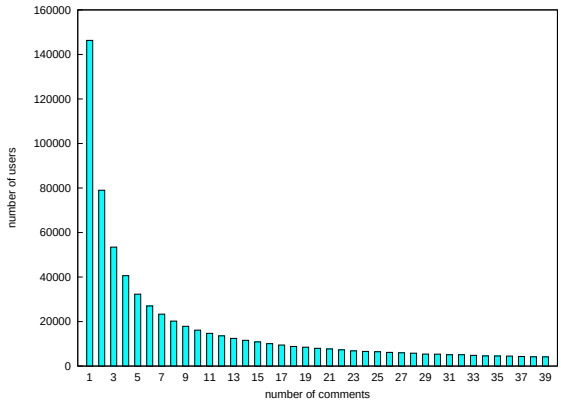

Fig. 3. Distribution of the number of comments added to user profiles.

The number of friends has been further analyzed, in order to investigate whether this parameter is related to the age of the user. Figure 4 plots the number of users having a given number of friends, as a function of their age. The figure shows that young users have more friends. Indeed, about 45% of 20 years old and younger users have more than 100 friends.

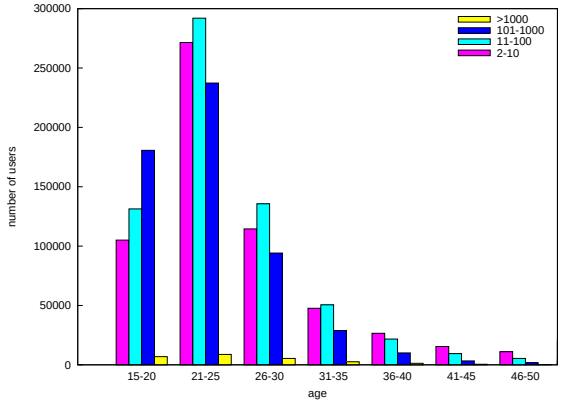

Fig. 4. Number of users as a function of age and number of friends.

As already pointed out, MySpace© users can add comments in users profiles, that is, annotate them and insert links to various types of objects. Comments have been considered an important characteristic of the profiles, because they allow to better understand relationships and interactions among users, and to quantify the popularity

of users in terms of their ability in involving other users in discussions and obtaining comments from them. The analysis did not consider the profiles without comments, hence focusing on about 1.4 millions users.
All profiles contain globally more than 365 millions comments. This analysis has focused on users who leave their comments on a profile, by studying the relationships between their number and the number of comments and by investigating their behavior in terms of number of added comments and number of different user profiles involved by these comments. The comments on the analyzed user profiles are added by more than 21 millions different users. On average, the comments of each profile are added by 58 users, and the comments on half of the analyzed profiles are contributed by less than 15 users. The correlation computed between number of comments and number of commenting users is equal to 0.738, which does not indicate any significant correlation between these two parameters. Hence, this relationship has been further investigated.
The number of comments added, on average, by a user to a profile has been computed as the number of comments over the number of users who leaved comments. A value close to one means that comments to the profile are added by different users, and gives an idea of the popularity of profiles in terms of the number of users involved in comments. Moreover, a small value also indicates that a user does not add many comments in the same profile, hence representing an index of the "loyalty" of the user. Figure 5 shows the distribution of the number of comments added on average by a user. The figure shows the distribution up to 20 comments. The distribution takes into account 99% of the profiles. As can be seen, users add few comments to the majority of profiles: 50% of profiles have on average just one or two comments from a user.

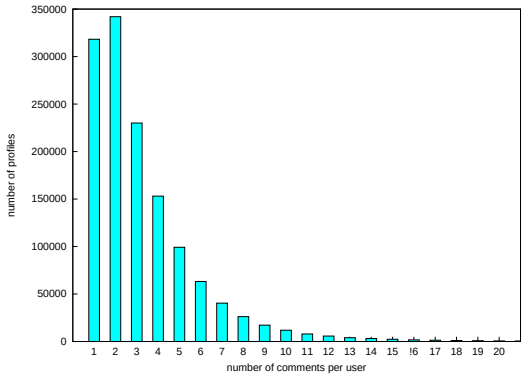

Fig. 5. Distribution of the number of comments per user.

The number of comments added by a user has also been analyzed as a function of the number of comments if the profile. Figure 6 shows this analysis up to 350 comments. As can be seen, the number of comments per user increases rapidly when the profile contains few comments. Then, as the number of comments increases, the behavior is almost stable. More in detail, it has been noticed that, independently on the number of comments in the profile, the minimum of the number of comments per user is close to one, that is, comments are added by different users. Moreover, the maximum is equal to the number of comments, that is, just one user adds all comments, only when few comments are in the profile. This means that popular user profiles capture many comments from different "non-loyal" users.

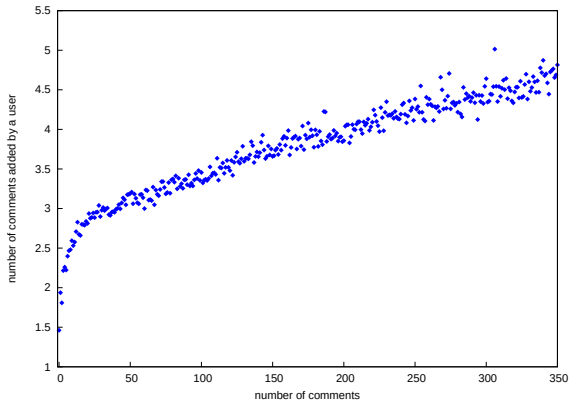

Fig. 6. Number of comments added by a user to a profile as a function of the number of comments in the profile.

The analysis has also shown that a small number of users (57,821) adds comments to their own profile. More than 83% of these users add less than three comments and in about 10% of these profiles, self comments are the only comments. It is interesting to point out that the average number of comments in these profiles is 886, that is, much higher than the average computed over all profiles. This means that self–commenting users tend to attract comments from other users.

Figure 7 shows the number of users who add comments as a function of the number of added comments. Up to 37 comments are shown, taking into account 90% of the users. On average, a user adds 17.3 comments in one or more profiles and 25% of users add one comment, whereas 55% of users add up to four comments. Few users, namely 0.04%, have been identified as having an intense ac-

tivity, adding from 1,000 up to 33,572 comments. Moreover, it has been noticed that about 70% of users add comments only to one or two profiles. Looking at the activity of these users, we discover that the number of comments they add is only 19% of the total number of added comments. Hence, a large number of users exists which add few comments on one or two profiles.

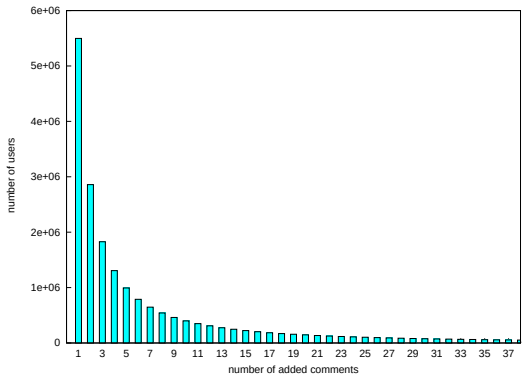

Fig. 7. Distribution of the number of comments added by each user.

4 Cluster analysis

Clustering techniques have been applied to discover groups of profiles having similar characteristics. From the point of view of this multidimensional statistical analysis technique, a user profile is represented as a point in a multidimensional space, the number of dimensions being the number of parameters used to characterize each profile. Hence, the problem is the identification of groups of profiles with similar characteristics in a multidimensional space. Clustering algorithms (Hartigan 1975) partition a set of points into groups, or clusters, such that points belonging to the same cluster exhibit similar behavior, that is, distance among points within a cluster is smaller than the distance among points of different groups. For the analysis of the user profiles we used the k-means clustering algorithm, which uses the Euclidean distance as a similarity criterion. A cluster is represented by its centroid, that is, the geometric center of the group. Since the objective of the study was to evaluate user popularity, the number of friends, the number of comments, and the amount of details specified in each profile have been used as characterizing parameters.
Table 2 shows the subdivision of the profiles in four groups, and in

particular the number of profiles per cluster and the corresponding centroids. The first and second cluster group about 95% of the total number of the profiles, whereas the third and fourth clusters group the remaining 5%.

	cluster 1 783,964 profiles	cluster 2 1,034,879 profiles	cluster 3 81,716 profiles	cluster 4 3,102 profiles
Number of friends	52.57	124.46	766.33	10,288.00
Number of comments	66.81	184.13	2,183.81	3,372.88
Amount of details	0.15	0.79	0.59	0.58

Table 2. Centroids of the four clusters.

Figure 8 shows the Box-and-Whisker plot for the parameter that specifies the amount of details. The plot helps in analyzing the parameter distributions, by representing the median, first and third quartile, and minimum and maximum values. The figure shows that profiles in the first cluster are characterized by small values and small variability. Moreover, the minimum is equal to zero, and it overlaps with first quartile, suggesting a distribution skewed towards zero. Profiles characterized by high values for the amount of details belong to the second cluster. Cluster 3 and 4 are both characterized by parameter values ranging from zero to one. The amount of details has a high variability, with high values, as suggested by the corresponding median that is equal to 0.7.

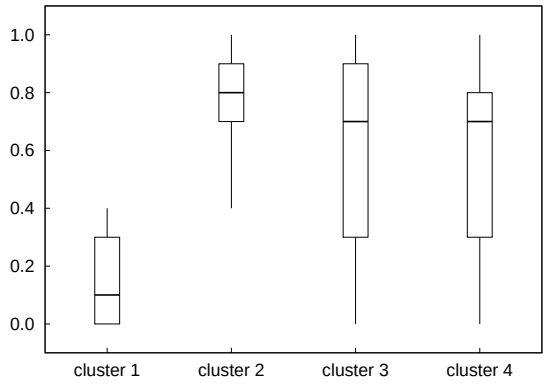

Fig. 8. Box-and-Whisker plot of the parameter describing the amount of details on each profile.

By looking in detail at values in Table 2 it comes out that profiles in the first cluster contain on average just 15% of personal data, and a number of friends and a number of comment equal to 55 and 67, respectively, which are far below the mean of the analyzed profiles (equal to 138 and 226, respectively). The second cluster is the most numerous and groups profiles containing many details about users, as shown by Fig. 8. Moreover, the number of friends is equal to 125 and the number of comments is equal to 184. The third and fourth clusters group profiles with a number of friends and a number of comments belonging to the tail of the corresponding distributions. In particular, the geometric centers of the third cluster are of 766 friends and 2,183 comments. The fourth cluster groups profiles of very popular users, and by a very high number of friends and of comments. Even though these parameters are not correlated, the number of friends influences the number of comments. The Box-and-Whisker plot of Fig. 9 shows the distributions of the parameters in the fourth cluster. Note that values have been scaled between zero and one to better highlight parameter distributions. As can be seen, the range of variability is high for all parameters. In particular, the distribution of the number of friends and of the number of comments are both characterized by a long tail; third quartile, median and first quartile are close to the minimum and very far from the maximum.

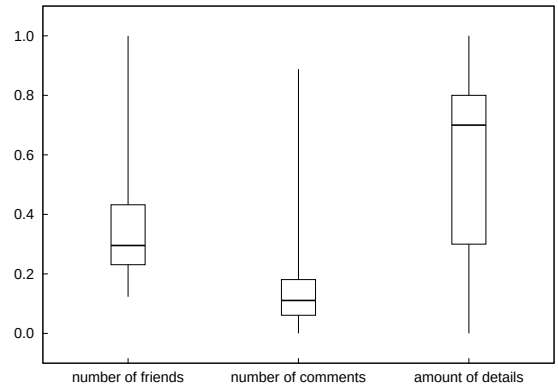

Fig. 9. Box-and-Whisker plot of the parameters of the profiles belonging to the fourth cluster.

5 Conclusions

This study provides the characterization of the behavior of users of MySpace© web site. The analysis relies on user profiles collected by crawling the web site for a period of twelve weeks. Profiles have been analyzed by means of statistical techniques, and typical user characteristics have been derived in terms of their age, number of friends and amount of personal details specified in the profiles. Moreover, the analysis of the comments contained in user profiles has highlighted the behavior of the users who leave a comment. A large number of users exists which add few comments on one or two profiles, while a small percentage of users having an intense activity have been identified.

Clustering techniques have been applied to identify classes of users with similar behavior. The groups of users identified by the clustering technique are, characterized by different behavior in terms of number of friends, number of comments and amount of details contained in the profiles. In particular, it has been discovered that the majority of users specify in their profiles many personal details, and participate actively to the social network, as derived from the high number of friends and of comments. Moreover, a small group of very popular users has been identified, whose profiles contain a very high number of friends.

Future work will be dedicated to a deeper investigation of the social relationships among users and of the evolution of their profiles, that is, changes due to user updates. Moreover, the analysis of the actual content of the comments will provide better understanding of the types of objects uploaded by the users and on the load induced on the servers that have to manage the web site.

References

Aguiton, C., & Cardon, D. (2007). The Strength of Weak Cooperation: An Attempt to Understand the Meaning of Web 2.0. *Communications & Strategies*, 65, 51–65. Available at http://ssrn.com/paper=1009070.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y. & Moon, S. (2007). I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *ACM Internet Measurement Conference (IMC'07)*, pp. 1–14.

Cha, M., Mislove, A., Adams, B., & Gummadi, K. (2008). Characterizing Social Cascades in Flickr. In *Proc. of the First Workshop on Online Social Networks*, pp. 13–18. ACM Press.

Cheng, X., Dale, C., & Liu, J. (2008). Characteristics and Potential of YouTube: a Measurement Study. In E.M. Noam and

L.M. Pupillo (Eds.), *Peer-to-Peer Video - The Economics, Policy and Culture of Today's New Mass Medium*, pp. 205–217. Springer.

Cohen, E., & Krishnamurthy, B. (2006). A short walk in the Blogistan. *Computer Networks*, 50(6), 615–630.

comScore web site. http://www.comscore.com/press/release.asp?press=2396. Accessed 12 May 2009.

Gill, P., Arlitt, M., Li, Z., & Mahanti, A. (2007). YouTube Traffic Characterization: A View From the Edge. In *ACM Internet Measurement Conference (IMC'07)*, pp. 15–28.

Halvey, M.J., & Keane, M.T. (2007). Analysis of online video search and sharing. In *Proc. of the 18th Conference on Hypertext and Hypermedia (HT07)*, pp. 217–226.

Hartigan, J.A. (1975). *Clustering Algorithms*. John Wiley & Sons.

Kumar, R., Novak, J., & Tomkins, A. (2006). Structure and Evolution of Online Social Networks. In *Proc. of the 12th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 611–617.

Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., & Bhattacharjee, B. (2007). Measurement and Analysis of Online Social Networks. In *ACM Internet Measurement Conference (IMC'07)*, pp. 29–42.

MySpace web site. http://www.myspace.com. Accessed 12 May 2009.

MySpace privacy policy. http://www.myspace.com/index.cfm?fuseaction=misc.privacy. Accessed 12 May 2009.

O'Reilly, T. (2007). What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications & Strategies*, 65, 17–37. Available at http://ssrn.com/paper=1008839.

Saha, B., & Getoor, L. (2008). Group Proximity Measure for Recommending Groups in Online Social Networks. In *Proc. of the Second SNA-KDD Workshop on Social Network Mining and Analysis*. ACM Press.

Urdaneta, G., Pierre, G., & van Steen, M. (2009). Wikipedia Workload Analysis for Decentralized Hosting. *Computer Networks*. doi:10.1016/j.comnet.2009.02.019.