Characterizing a social bookmarking and tagging network

Ralitsa Angelova[1], Marek Lipczak[2], Evangelos Milios[2], Paweł Prałat[3] [4]

Abstract.

Social networks and collaborative tagging systems are rapidly gaining popularity as a primary means for storing and sharing data among friends, family, colleagues, or perfect strangers as long as they have common interests. `del.icio.us`[5] is a social network where people store and share their personal bookmarks. Most importantly, users tag their bookmarks for ease of information dissemination and later look up. However, it is the friendship links, that make delicious a social network. They exist independently of the set of bookmarks that belong to the users and have no relation to the tags typically assigned to the bookmarks. To study the interaction among users, the strength of the existing links and their hidden meaning, we introduce implicit links in the network. These links connect only highly "similar" users. Here, similarity can reflect different aspects of the user's profile that makes her similar to any other user, such as number of shared bookmarks, or similarity of their tags clouds. We investigate the question whether friends have common interests, we gain additional insights on the strategies that users use to assign tags to their bookmarks, and we demonstrate that the graphs formed by implicit links have unique properties differing from binomial random graphs or random graphs with an expected power-law degree distribution.

1 INTRODUCTION

Social bookmarking and collaborative tagging services lead to the formation of a new type of organically grown network structure. In such networks, users are linked to other users through social connections (e.g. directed friendship links) and to network specific online resources (e.g. bookmarks, photos, books, etc.) by either explicitly linking to them, tagging them with appropriate terms, or commenting on them. Clustering users in this context is a challenging problem, as it involves accounting for multiple types of social linkage among users and diversity of content ranging from personal photos (`flickr.com`) or bookmarks (`del.icio.us`) to whole libraries of read books throughout the user's lifetime (The Personal Library, `librarything.com`). The complexity of the clustering problem raises dramatically if we look at the overall electronic fingerprint of these users after connecting all their profiles from the various social networks they actively contribute to [12]. Not only is clustering itself challenging but evaluation of the clustering solution is also very hard as reference class assignments are typically missing or very expensive to manually gather. These class assignments (also known as ground truth) are ignored in the clustering process. They are used exclusively in the evaluation phase to compare the groups produced by the clustering technique to the known classes it comprises. Modelling social bookmarking and tagging services is a way to generate synthetic data sets that mimic the behaviour of such social networks. Moreover, the synthetic data generative model also provides the corresponding ground truth for performance evaluation and comparison. A requirement for the design of useful models is an in-depth understanding of the properties of real-life data sets obtained from on-line social networks such as `del.icio.us`.

A collaborative tagging system like `del.icio.us` can be visualized as a tripartite structure [7], where links (edges) are established between users, tags and bookmarks. Additionally, the social dimension introduces "friendship" links between users. Several research questions about the structure of the social network and its implications arise:

- What is the role of friendship in relation to interest sharing as reflected in the bookmarks and tags of users. Do friends appear to have more common interests than non-friends?
- Do "highly social" users share more topics of interest with others than the "less social" users?
- What are the structural properties of the friendship graph and the graphs induced by the implicit similarity-based links among users? Is their degree distribution indicative of power-law graphs? What are their connectivity and local density properties, measured by the clustering coefficient, as a function of k in their k-core analysis [8]? How do they compare with binomial random graphs [10] and random graphs with an expected power-law degree distribution [4]?
- What are the common properties of the friendship, bookmark-based and tag-based links? In particular, how do the three types of links correlate for individual users?

We are not the first to analyze social collaboration on the Web. Evolution models of two online social networks - Flickr and Yahoo!360 are examined in [11]. In the experiments performed in [13] on the photo sharing network Flickr, after taking a random subset of photos and their owners or users, it is demonstrated that Flickr exhibits the characteristics of small-world and scale-free networks described earlier by [2, 5]. Search and ranking techniques applied to social networks are discussed in [9].

A detailed analysis of three other online social networks is presented in [1]. Tagging distributions in `del.icio.us` are shown to stabilize into power law distribution with a limited number of stable tags and a much larger "long-tail" of more idiosyncratic tags. Similar results are noted in [6]. Both results give strong evidence that collaborative tagging systems (like `del.icio.us`) can be exploited for reliable automatic creation of taxonomies. More recently, a study of a tag co-occurrence network was carried out [14]. The nodes of this network are tags, and tags are linked when the two tags occur together in the set of tags assigned to a specific bookmark by a user

[1] Max Planck Institut für Informatik, Saarbrücken, Germany
[2] Faculty of Computer Science, Dalhousie University, Halifax, Canada, eem@cs.dal.ca
[3] Department of Mathematics and Statistics, Dalhousie University, Halifax, Canada
[4] Authors listed in alphabetical order.
[5] http://del.icio.us

(the post). A weight is given to a link that depends on the number of posts in which the two tags co-occur. The tag co-occurrence network was shown to reveal spamming behaviour.

2 Definitions of graphs and associated metrics

To answer the questions above, we study various graphs on the data provided by the social bookmarking network `del.icio.us`. For all graphs, vertices correspond to users. The edges are defined differently for each graph as follows.

The friendship graph. Edges correspond to *directed* friendship links between users. In this paper, as discussed in Sec. 3, we ignore the direction of friendship links and obtain an undirected friendship graph. Bidirectional edges (i.e. representing mutual friendship) are included only once.

Common entity graphs. In this type of implicit graph two vertices are connected by an *undirected* weighted edge.

The edge weight reflects the number of entities that the connected users have in common. A drawback of the common entity method is that it does not take into consideration the number of terms in both sets - having three common terms has the same meaning, no matter if the total number of user's terms is ten, or fifty. This similarity metric is symmetric, and its range is from 0 to the maximum number of entities. The two subtypes of graph under the common entity type are the following:

- **The common bookmark graph**, where the set of entities of a user is the set of bookmarks of that user.
- **The common tag graph**, where the set of entities of a user is the set of tags the user assigned to all her bookmarks.

Similarity graphs. In this type of implicit graph, two vertices are connected by an *undirected* weighted edge.

The edge weight reflects the cosine similarity between the entity vectors of the connected users. In these similarity graphs, entities can be either bookmarks or tags. The user entity vector belongs to a vector space defined on the entire data set, but captures only the individual entities belonging to the specific user. The weight of a vector coordinate is the binary or tf-idf score for the associated entity of the user. More specifically, the cosine similarity between vectors, Eq. 1 is used to define the distance between users, where U_i is the vector of user i, and $w_{i,j}$ is its j-th coordinate. The cosine distance has the advantage that it is normalized with respect to the length of vectors. Cosine similarity is symmetric, and its range is from 0 to 1.

$$cos\_sim(U_1,\ U_2) = \frac{\sum_{j=1}^{n} w_{1,j} * w_{2,j}}{\sqrt{\left(\sum_{j=1}^{n} w_{1,j}^2\right) * \left(\sum_{j=1}^{n} w_{2,j}^2\right)}} \quad (1)$$

Neither bookmarks nor tags have the equivalent of stop words, as discussed in Section 3, so we include their entire sets in computing the above user similarity metrics. The two subtypes of similarity graph that we study are derived as follows:

- **The bookmark similarity graph** is derived by considering bookmarks as entities. The weight of a bookmark in any user vector is binary.
- **The tag similarity graph** The tag similarity graph is derived by considering tags as entities. The weight of a tag in any user vector is the tf-idf score for this tag. The term frequency factor (tf) is the number of times a user used a given tag. The inverted document frequency factor (idf) is the inverse of the natural logarithm of the number of users that used a given tag plus one.

Both versions of graphs were generated after low frequency entities were removed. We ignored tags and bookmarks that were used by less than five users.

Unweighted graph versions. Many interesting graph properties (e.g., clustering coefficient, diameter) are defined for unweighted graphs. To measure these properties, we transform the weighted common entity and similarity graphs to unweighted versions. A single threshold is selected for each graph such that after removing connections with weight lower than the threshold one million edges are left.

Clustering coefficient. A quantity of interest in measuring the local density properties of the various graphs is the *clustering coefficient* [1]. Given an undirected graph $G = (V, E)$, the clustering coefficient of vertex $i \in V$ is defined as:

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

where k_i is the degree of vertex i. The clustering coefficient is not defined for vertices of degree at most 1. The clustering coefficient of a graph G is the average over all vertices (of degree at least 2) in the graph, that is, $C(G) = \sum_{i \in V, \deg(v) \geq 2} C_i / |\{v \in V : \deg(v) \geq 2\}|$. Its values always lie between 0 and 1. The clustering coefficient is asymptotic to $|E|/\binom{|V|}{2}$ for a binomial random graph. For real-world networks it is usually much larger.

K-cores. A k-core of a graph is a maximal induced subgraph of minimum degree at least k. If no subgraph has this property, then we say that the k-core is empty. It is possible to show [3] that the degree core is unique for a given graph and a given k, and can be obtained by recursively removing all vertices with degree less than k. The k-cores of a graph can consist of multiple components. The difference between the k-core and simply filtering out all vertices with degree less than k is best illustrated by comparing their effects on a simple tree. In the case of a tree, the filtering of all degree-one vertices results in the pruning of all of a tree's leaves, whereas the degree core with $k = 2$ would prune back the leaves of a tree at each recursion, thus destroying the tree completely. Cores were first introduced in studying social networks by Seidman [15] and popularized by Wasserman and Faust[16]. Batagelj and Zaversnik [3] generalize Seidman's work beyond simple degree to include any monotone function p.

In this paper, we perform degree-based k-core analysis of the friendship, common entity and similarity graphs, by repeatedly increasing k by one until the k-core is reduced to empty. We then plot various properties of the k-core sequence of graphs, including the diameter, size, average distance and average clustering coefficient of the vertices between vertices of the largest component, and the number of components.

Note that the k-core analysis becomes prohibitively expensive as the input graphs get denser. This made the use of a computing cluster necessary for most of the k-core analysis presented in this paper.

3 Exploring the data set

We perform our experiments on a data set obtained by an automated Web crawl on the social community platform `del.icio.us`. The subset of `del.icio.us` we have at hand consists of $13,514$ users, $4,574,587$ bookmarks, and $47,807$ friendship connections, $6,876$ of which are mutual (bidirectional) connections. Most users

(13, 238) have at least one bookmark. Most users (13, 439) also point to at least one friend. The total number of tags used in our subset is 643, 889.

The crawl on del.icio.us is a Breadth First Search (BFS) of the friendship graph, starting with the user with the highest number of friends as a seed node. This resembles the so called *snowball sampling* technique which is argued to be the only feasible sampling method for crawling such networks [1].

The friendship graph It is a straightforward representation of the directional connections between users. We observe that this graph is quite sparse, with 40,931 uni-directional edges, and 6,876 bi-directional edges corresponding to mutual friendships. Only 14% of friendships are mutual, due to the fact that by design friendships do not have to be confirmed in in the online system supporting del.icio.us.

Bookmark distributions and properties Intuitively two users are connected when they point to similar sets of bookmarks. We investigate the hypothesis that bookmarks can be treated as terms in the standard Information Retrieval sense, where each user is viewed as a document, and the user's bookmarks are viewed as terms. The proportion of bookmark urls used only once over the total number of unique bookmark urls is 78%, where the proportion of unique words in Wikipedia used only once over the total number of unique words is 52% (second and fifth column of Table 1).

We further examine whether the distribution of bookmark urls follows Zipf's law. The log-log plot of the url frequency against rank of urls sorted by decreasing order of frequency, has the following characteristics (a) a slowly declining part for the top 1000 urls, (b) a fairly straight part up to rank of about 10^6, and (c) a horizontal part at frequency equal to 1 up to rank of about 0.25×10^7 corresponding to a large number of bookmarks appearing only once (Fig. 1). Surprisingly, bookmarks to general use sites (e.g., google.com) are not among the ten most frequent bookmarks. We would expect such sites to be analogous to stop words in standard text repositories. One possible explanation is that users avoid adding these stop-word-like urls to their bookmarks to keep the bookmark list smaller. The addresses of general purpose websites are easy to remember and do not have to be stored in del.icio.us. If we prune urls keeping only their domain, the frequency against rank plot shows that the domains of bookmarks do match Zipf's law. The proportion of bookmark url domains used only once over the total number of unique bookmark url domains is 56%, a very close match to the proportion of unique words in Wikipedia used only once over the total number of unique words, which is 52% (third column of Table 1). Furthermore, the list of ten most frequently bookmarked domain addresses contains indeed the most popular websites. Despite a better fit to Zipf's law, we do not think that the list of pruned bookmarks is useful for finding real connections between users, as domains are usually too general to capture users' interests.

Tag distributions and properties Shared tags between two users may be interpreted as showing their overlapping interests. Although tags are mostly nouns and adjectives, and they do not form grammatically correct sentences, they potentially match many of characteristics of text repositories. The proportion of tags used only once over the total number of unique tags is 53%, matching very closely the proportion of unique words in Wikipedia used only once over the total number of unique words of 52% (fourth column of Table 1). Similarly to bookmarks, tag frequency is not distributed according to the Zipf's law for the highest ranks, as there are no stop-word-like tags (Fig. 1). The most frequently used tags correspond to general categories, therefore they still convey content information.

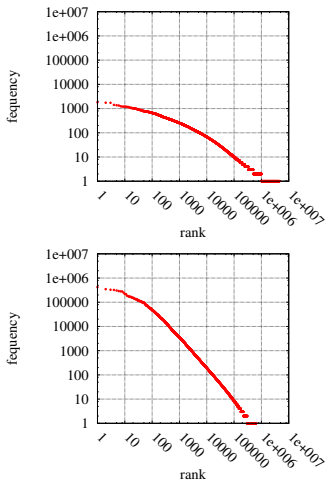

Figure 1. Frequency distribution of bookmark urls (top), and tags (bottom).

Counts (in K)	urls	url domains	Tags	Wikipedia words
Total number of unique terms	4,575	1,106	644	4,098
Terms used more than once	1,017	483	303	1,978
Terms used only once	3,558	623	341	2,120

Table 1. Counts (in thousands) of bookmark urls and tags, treating bookmarks and tags as "terms", the user as the "document", and the set of all users as the corpus. The same counts are shown for the Wikipedia text corpus. We see that the proportion of terms used only once is much higher in the bookmark url corpus than either the tag or the Wikipedia corpus.

4 Relating friendship with bookmark and tag similarity

We now study the question whether friends have more similar bookmarks or tags than non-friends. We calculate the average connection strengths (according to the previously defined user similarity metrics) over pairs of friends and pairs of non-friends, shown in Table 2. We observe that friends have significantly stronger connections than non-friends based on the bookmark similarity metrics, whereas this is not true for the tag based similarity metrics. We conjecture that the reason for this is that the majority of the tags are individualized, dependent on the individual user's way of organizing their bookmarks and their background in the corresponding areas. The number of tags that capture the generic meaning of a bookmarked page is a small fraction of the total number of tags associated with that page. In other words, friends who share bookmarks choose mostly different tags for them. To further explore this conjecture, we inspect the Zipf distribution of tags for individual users, and bookmarked web sites. In the plots for individual users, we notice the most frequently used tags, which are likely to be the ones recommended by del.icio.us, followed by the tail of infrequently used tags, which are likely to be individualized tags, specific to the user's organization of her bookmarks. A significant fraction of the tags of a user is used only once. In the plots for individual bookmarked web sites, we also notice a small number of most frequently used tags, followed by a large number of user-specific tags, used only once.

	Friend pair average	Non-friend pair average
k-common bookmarks	1.931	0.372
bookmark cosine sim	0.011	0.004
k-common tags	54.157	41.816
tag cosine sim	0.081	0.085

Table 2. Average strength of the connection between friends (column 2) and non-friends (column 3). We observe that friends have significantly stronger connections than non-friends based on the bookmark similarity metrics, whereas this is not true for the tag-based similarity metrics.

5 Density properties of the friendship graph and the content-based social graphs

We now apply the concept of k-core introduced in Sec. 2 to discover the density properties of the friendship graph and our content-based social graphs (common bookmark/tag, and bookmark/tag similarity graphs), and compare them with the properties of binomial random graphs [10] and random graphs with power-law degree distribution [4]. We focus on plots of properties of the graphs as a function of increasing k in the k-core computation. For fairly small k the smaller components disappear and we are left with a single component, which becomes progressively denser with increasing k, and eventually disappears. The properties of interest are: the number of components, and, for the largest component, its size (number of vertices), its average distance between vertex pairs, and its diameter. We furthermore generate scatter plots of the clustering coefficients versus degree for the graphs' vertices, and the average clustering coefficient of the vertices of the largest component.

Basic k-core properties of the content-based social graphs The average distance between pairs of points of the largest component in the k-core analysis of common bookmark graphs is shown in Fig. 2. The linear nature of the curves is also worth noting.

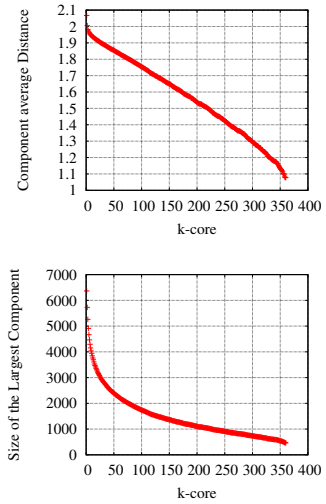

Figure 2. This plot focusses on the largest component of the common bookmark graphs, and it depicts average distance between pairs of vertices (top) and size (bottom) as a function of k in the k-core analysis.

The diameter of the largest component in the k-core analysis of common bookmark graphs is lower than 7.

The number of components in the k-core analysis of common bookmark graphs drops very quickly to one, as k increases in the k-core analysis. Having a single component is similar to binomial random graph behaviour.

The size of the largest component (number of vertices) in the k-core analysis of common bookmark graphs is shown in Fig. 2. The size of the last non-trivial k-core is very close to k, so these vertices form an induced subgraph that is close to a clique. This is very different behaviour compared to binomial random graphs, where the last non-empty k-core still contains a positive fraction of vertices with high probability.

Clustering coefficient of k-cores In this subsection we examine the clustering coefficient and its relation to the degree of the graph vertices. A plot of the average clustering coefficient for vertices of a given degree is shown as a function of degree in the left column of Fig. 3. For a binomial random graph and random graph with expected power law degree distribution, the corresponding theoretical curve would be horizontal, since the fact that two vertices are friends of a third vertex does not affect the probability of them being linked.

The average clustering coefficient over all vertices as a function of k in the k-core analysis is shown in the right column of Fig. 3. These plots are consistent with the plots in the left column, considering that the low-degree vertices are dropped first as k increases in the k-core analysis. For low values of k, low degree vertices are lost. When such vertices have high clustering coefficient, we see a drop in the average clustering coefficient, down to a minimum for a value of k. As k increases past that minimum, the densification process prevails and we see the expected monotonic increase in the average clustering coefficient. In the friendship graph, we observe the same trend of low degree vertices having high clustering coefficients, with the average monotonically decreasing with increasing degree. This implies that friends of users (represented by vertices) with low degrees, or equivalently, few friends, tend to be friends themselves, while friends of users with large degrees are not necessarily connected.

6 Discussion

In this paper, we closely examined the properties of a typical social bookmarking and tagging data set, to obtain insights for facilitating creation of models for such data. We summarize our observations as follows:

- Friendship correlates well with common bookmarks or similar bookmark vectors of users, but not well with common tags or similar tag vectors. This implies that the majority of tags that users assign to bookmarks are user-specific tags, and only a small fraction of the tags capture the generic meaning of the bookmarked web page.
- Tags behave more like words in text, while bookmarks less so, in the sense that a much higher proportion of bookmarks are used only once compared to the proportion of tags and words in the Wikipedia corpus that are used only once.
- There are no tags or bookmarks behaving like stop words in text. Even the highest frequency tags or bookmarks do not appear that frequently to deserve the characterization of stop words.
- The clustering coefficient as function of k in the k-core analysis displays a U-shaped curve. This is not consistent with binomial random graphs or random graphs with a power law degree distribution.

Acknowledgements We are grateful for the financial support of the Natural Sciences and Engineering Research Council of Canada, the MITACS Network of Centres of Excellence, and Genieknows.com. Tom Crecelius, and Mouna Kacimi of the Databases and Information Systems Department, Max

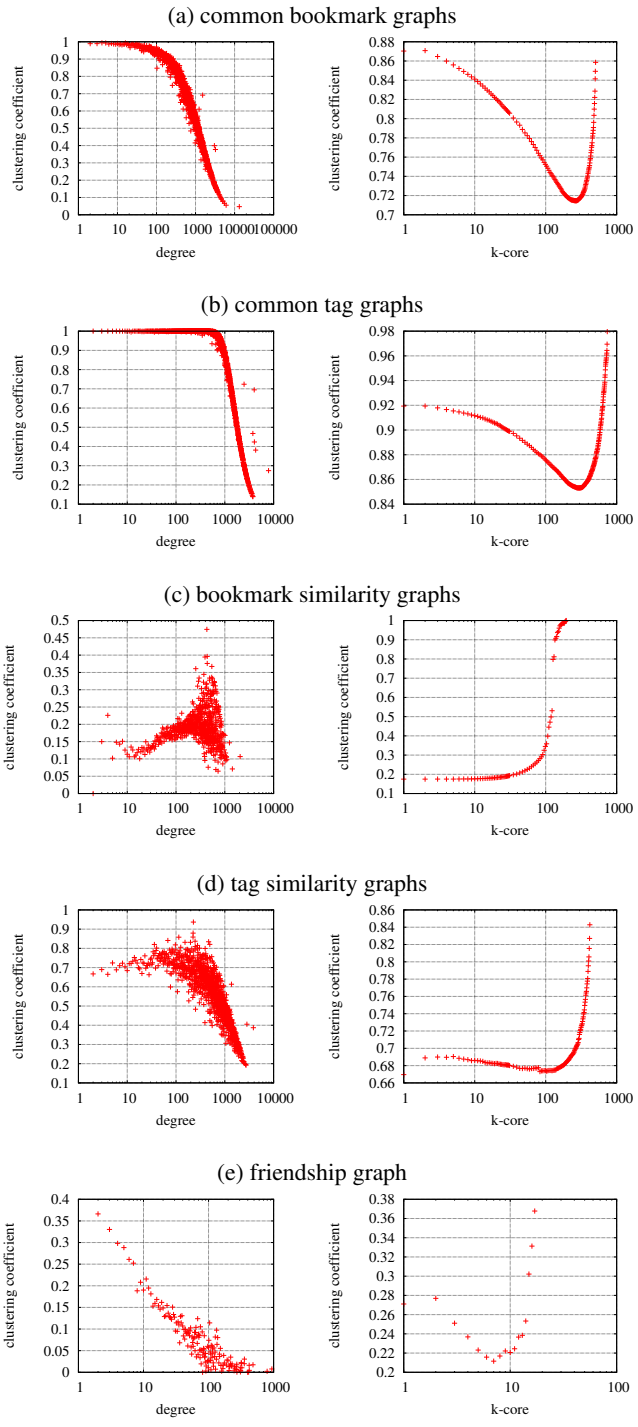

Figure 3. The average clustering coefficient over the vertices of a given degree as a function of degree for the common bookmark graph (left). The average clustering coefficient over all vertices of the largest component in the k-core analysis of the common bookmark graph, as a function of k (right). We observe that less "social" individuals tend to have a closely knit set of neighbours, who are likely to be connected to each other, because of their very specialized interests.

Planck Institut für Informatik directed by Prof. Gerhard Weikum, kindly provided the `del.icio.us` data set for this project. Jacek Wołkowicz gave us the Wikipedia statistics. The computations required for the k-core results would not be possible without access to the Atlantic Computational Excellence Network ACEnet and the Shared Hierarchical Academic Research Computing Network SHARCNET computing clusters.

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 *World Wide Web (WWW) Conference*, pp. 835–844, Banff, Alberta, (May 8-12 2007).

[2] A-L. Barabasi, *Linked: The New Science of Networks*, Perseus, Cambridge, MA, 2002.

[3] V. Batagelj and M. Zaveršnik, 'Generalized Cores', *ArXiv Computer Science e-prints*, (2002).

[4] F.R.K. Chung and L. Lu, *Complex graphs and networks*, American Mathematical Society, U.S.A., 2004.

[5] R. Cohen, D. Avraham, and S. Havlin, 'Structural properties of scale free networks', in *Handbook of graphs and networks*, eds., S. Bornholdt and H. G. Schuster. Wiley-VCH, (2003).

[6] S. Golder and B. Huberman, 'The structure of collaborative tagging systems', Technical report, Information Dynamics Lab, HP Labs, (August 2005).

[7] H. Halpin, V. Robu, and H. Shepherd, 'The complex dynamics of collaborative tagging', in *World Wide Web (WWW) Conference*, pp. 211–220, Banff, Alberta, (May 8-12 2007).

[8] John Healy, Jeannette Janssen, Evangelos Milios, and William Aiello, 'Characterization of graphs using degree cores', in *Algorithms and Models for the Web-Graph: Fourth International Workshop, WAW 2006*, volume LNCS-4936 of *Lecture Notes in Computer Science*, Banff, Canada, (Nov. 30 - Dec. 1, 2006 2008). Springer Verlag.

[9] Andreas Hotho, Robert Jaeschke, Christoph Schmitz, and Gerd Stumme, 'Information retrieval in folksonomies: Search and ranking', in *Proceedings of the 3rd European Semantic Web Conference*, Lecture Notes in Computer Science. Springer, (2006).

[10] S. Janson, T. Łuczak, and A. Ruciński, *Random Graphs*, Wiley, New York, 2000.

[11] R. Kumar, J. Novak, and A. Tomkins, 'Structure and evolution of online social networks', in *Knowledge Discovery in Databases (KDD) Conference*, pp. 611–617, Philadelphia, PA, (August 20-23 2006). ACM.

[12] Flavia Moser, Rong Ge, and Martin Ester, 'Joint cluster analysis of attribute and relationship data without a-priori specification of the number of clusters', in *Knowledge Discovery in Databases (KDD) Conference*, pp. 510–519, San Jose, CA, (August 12-15 2007). ACM.

[13] Radu Negoescu, 'An analysis of the social network of flickr', Technical report, Laboratory of nonlinear systems - LANOS, School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne, Switzerland, (July 2007).

[14] Christoph Schmitz, Miranda Grahl, Andreas Hotho, Gerd Stumme, Ciro Cattuto, Andrea Baldassarri, Vittorio Loreto, and Vito D.P. Servedio, 'Network properties of folksonomies', in *World Wide Web Conference (WWW)*, Banff, Canada, (May 8-12 2007).

[15] S. B. Seidman, 'Network structure and minimum degree', *Social Networks*, 269–287, (1983).

[16] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*, Cambridge University Press, Cambridge, 1994.