Folksonomies and clustering in the collaborative system *CiteULike*

2008 J. Phys. A: Math. Theor. 41 224016

(http://iopscience.iop.org/1751-8121/41/22/224016)

IOP PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND THEORETICAL

J. Phys. A: Math. Theor. **41** (2008) 224016 (7pp) doi:10.1088/1751-8113/41/22/224016

Folksonomies and clustering in the collaborative system *CiteULike*

Andrea Capocci[1] **and Guido Caldarelli**[2]

[1] Dip. di Informatica e Sistemistica Università "Sapienza", via Ariosto, 25 00185 Rome, Italy
[2] SMC Centre, CNR-INFM, Dip. di Fisica, Università 'Sapienza', P.le A. Moro 5, 00185-Rome, Italy

Received 1 May 2008
Published 21 May 2008
Online at stacks.iop.org/JPhysA/41/224016

Abstract

We analyze *CiteULike*, an online collaborative tagging system where users bookmark and annotate scientific papers. Such a system can be naturally represented as a tri-partite graph whose nodes represent papers, users and tags connected by individual tag assignments. The semantics of tags is studied here, in order to uncover the hidden relationships between tags. We find that the clustering coefficient can be used to analyze the semantical patterns among tags.

PACS numbers: 89.75.Fb, 89.75.−k

The recent development of the World Wide Web is characterized by a growing number of online social communities. In this case, individuals provide bits of information—about either their tastes, opinions or interests—and software applications gather and organize them into a database. A class of such collaborative systems focuses on collecting users' online bookmarks with either a generalist approach or a more specialized one. The case of study we present here regards the website of the project *CiteULike* whose activity is to store user-generated scientific bibliographies.

In every collaborative system, the elementary contribution, the 'post', is made by three ingredients: a user, an article and an annotation of it by a number of tags chosen by users. In exchange for this voluntary contribution, a user can browse the bibliographies and annotations of other users. Tags are an alternative classification method with respect to traditional taxonomies, where items belong to 'taxa' representing as a tree-like set of categories. Here, each category contains in turn a number of more specialized sub-categories, in a hierarchy of level, until the desired resolution of classification is reached. Instead, in tagging systems items are tagged by users characterized by diverse tagging strategies depending on a number of individual variables. The set of tag-resource relations in such a community is called a 'folksonomy'. Such communities are now extremely popular, storing hundreds of thousands of posts and more. The tagging system we analyze here, *CiteULike* [1], has been built, at the time of our survey, by about 180 000 references annotated by about 48 000 tags supplied by about 6000 users. Our dataset includes about 550 000 'tag assignments': each assignment

is a t-uple (user, resource, tag). The sequence of chronologically ordered tags, in particular, can be interpreted as a stream of words, to which one can apply the traditional statistical text analysis to uncover how human behavior affects it.

The statistical analysis of word occurrences in a written text shows that word frequencies are power-law distributed according to the Zipf's law. This means that a large number of words appear in a text only a few times, while a few words occur orders of magnitude more often [2]. Such a feature has been modeled in many ways all based on the preferential attachment principle. That is, the assumption that authors employ used words with a probability proportional to their frequency. Moreover, it has been observed that the rate of new words decrease with the text length [3], that is, the number of distinct words N_w in a text of length L scales as

$$N_w \propto L^{\beta}, \tag{1}$$

with $\beta < 1$. However, models in the literature assume that new words are introduced at a constant growth rate, so that their total number, i.e. the vocabulary, is a linear constant of the total number of words (both new or repeated ones) used so far [4–6].

Yet, to discover the semantical properties of *CiteULike*, it is more fruitful to represent it by means of a network formalism. This has been proved fruitful in the analysis of many natural and social phenomena involving unsupervised interacting units. In a network perspective, elementary interacting agents or objects are represented by vertices and their interactions by edges connecting them. Interestingly in many real networks common statistical properties are present without external tuning. For example, the degree k, that is, the number of edges pointing to a vertex, follows very often a broad distribution $P(k)$ whose fat tail can be described by a power law of the kind

$$P(k) \propto k^{-\gamma}, \tag{2}$$

with $\gamma < 3$. Edges can also be characterized by a weight w representing their intensity. Accordingly, one can define a weighted degree (*strength*) of a vertex defined as the sum of the weights of edges pointing to it. Also the probability distribution of *weights*, is power law distributed in many real systems. Finally many networks also present a strong transitivity, i.e. with high probability, the neighbors of a node are themselves connected by an edge, with respect to purely random realization of a network with equal number of nodes and links. Even if no formal definition exists, usually networks sharing the above properties are named 'complex networks' [7].

This network approach has also been recently adopted to analyze the semantical structure of tagging systems [11–13]. Tags can be represented by networks in different ways, in order to study how the behavior of users maps into the dynamical or topological features. For example, tags can be implicitly linked by hierarchical and logical associations emerging despite the diversity of users when their number is large enough. The underlying semantical organization of tags reveals the dominant trends within a tagging community and allows us to improve its navigability. Recently, algorithms have been introduced in order to infer a taxonomy of tags from a folksonomy [8, 9]. The statistical properties we observed in the *CiteULike* data are consistent with the findings obtained in similar surveys, confirming that tags in collaborative systems tend to form complex networks. Moreover we have investigated how the underlying semantics of tags reflects on the topology of the network.

As a matter of fact, tags provided by users come with no explicit hierarchy beside the chronological ordering, leading authors to analyze the stream of tags as a text-like sequence of words. Interestingly, the time-ordered sequence of tags displays statistical properties already observed in written texts, such as the fat tails in the word frequency distributions or the sublinear vocabulary growth. Our analysis confirms the sublinear vocabulary growth observed

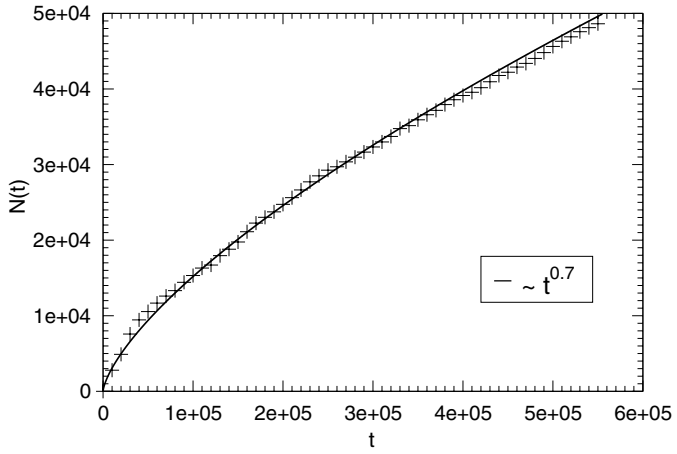

Figure 1. The number of tags $N(t)$ as a function of time t, where time is measure in chronologically ordered tag assignments (plus symbols). Solid line represents $t^{0.7}$ for a comparison.

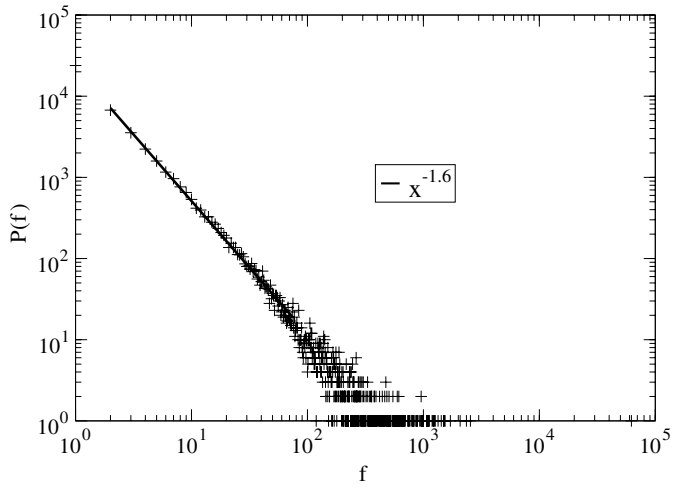

Figure 2. The statistical distribution $P(f)$ of tag frequencies f (plus symbols). Solid line represents $f^{-1.6}$ for a comparison.

in written texts. The number of distinct tags $N(t)$ introduced by users after t assignments grows approximately as $N(t) \propto t^{0.7}$, as shown in figure 1 although the pace is slightly smaller than in other collaborative tagging systems already surveyed [10]. The frequency of tags, too, reported in figure 2, reminds that of words observed in written texts, algebraically decaying according to the Zipf law [2]. As in the case of many models, new 'words' can be introduced at any time in the tag vocabulary.

Anyway, the frequency of tags as a function of time does not convey information about the semantics. On a large scale it only reflects the different centrality of associated concepts in the underlying knowledge organization. To investigate tag pair relations, one has to represent the unit elements of a tagging system as nodes of a network.

The dataset analyzed can be naturally represented as a tri-partite network, where each node corresponds either to a user u, to a resource r or to a tag t. If a tag assignment (u, r, t) exists,

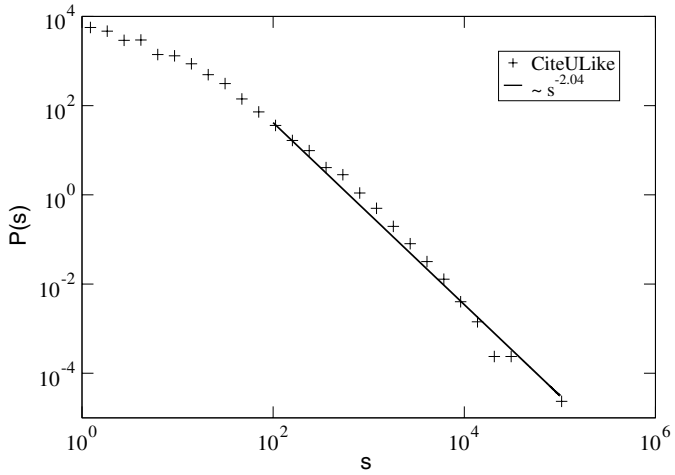

Figure 3. The distribution $P(s)$ of the node strengths s in the tag co-occurrence network. The best fitting power-law exponent, represented by the solid curve, yields $\gamma = 2.04$.

an edge is drawn from u to r, and from r to t. Since in a single user's post a resource can be tagged more than once, one post can correspond to multiple tag assignments [12]. Although an efficient algorithm has been developed to analyze such a tri-partite network [14], the heterogeneity of nodes discourages in general the application of traditional network methods, mainly conceived to deal with network connections representing peer-to-peer relationships. Thus, to study how tags are organized we chose to project the tri-partite networks on the tag space. As a result, the tag co-occurrence network we study is composed by nodes representing tags only, between which an undirected edge of weight w is drawn if w distinct resources are labeled by both tags.

The resulting network displays some of the typical features of weighted scale-free networks. We have measured the distribution of the sum s of the weights of edges pointing to a given node, or the *strength* of the node: such a distribution $P(s)$, plotted in figure 3, exhibits a clear power-law decay $P(s) \propto s^{-\gamma}$, with $\gamma = 2.04 \pm 0.02$ for large values of s. Interestingly, the heterogeneity of the observed node weights does not necessarily reflects the centrality of corresponding concepts in the underlying hierarchy of tags. As it has been already shown [15], reshuffling the tag assignment in order to destroy the logical association among words does not change dramatically the shape of $P(s)$, which proves that the *weight* heterogeneity is more a consequence of frequency distribution broadness than of the varying roles of concepts in the semantical organization of the whole vocabulary.

Nevertheless, the tag co-occurrence network unveils some semantical feature of the underlying structure. This can be obtained by focusing on the inspection of quantities involving its environment. An example of such is represented by the analysis of the neighbor average degree $K_{nn}(k)$ of nodes with degree k, where the degree is the number of incoming edges of a node or the study of the clustering coefficient. Here we focus on this latter property by considering the clustering properties of the tag co-occurrence network. The coefficient $C(k)$ counts the average density of triangles involving nodes with degree k or, in other words, the probability that the nearest neighbors of a node with degree k are in turn connected to each other. This reads

$$C(k) = \frac{2 \sum_{i,k(i)=k} \sum_{j>h}^{1,k} a_{ji} a_{hi}}{N_k k(k-1)}, \tag{3}$$

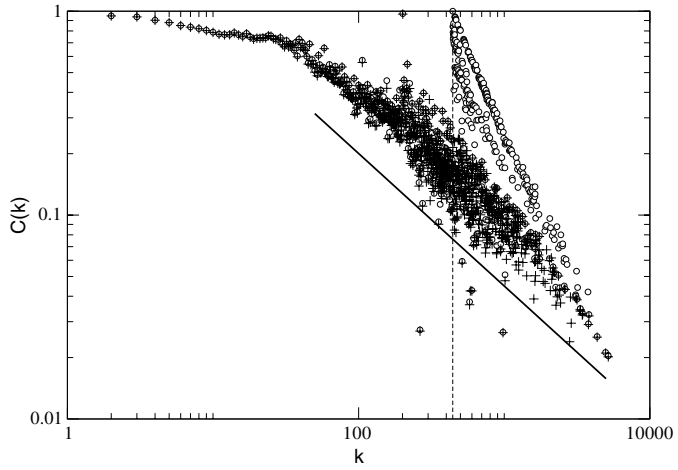

Figure 4. The clustering coefficient $C(k)$ of the tag co-occurrence network as a function of the nodes' degree k before (circles) and after (plus symbols) the removal of a spam post from the dataset. The solid line represents a decay $k^{-0.64}$ and the dashed vertical ruler is set at $k = 443$.

where a_{ij} is 1 if a link exists between i and j and 0 otherwise, and N_k is the frequency of nodes with degree k. This quantity has been found to characterize the most complex networks found in nature and society, where it takes substantially larger values with respect to a purely random network [16]. The properties of the clustering coefficient are often associated with the hierarchical organization of nodes [17].

Indeed, the clustering coefficient appears to encode a signature of semantical relations between words. As represented in figure 4, the clustering coefficient $C(k)$ in *CiteULike* decays algebraically for large values of the degree k, according to $C(k) \propto k^{-0.64}$. However, the clustering value displays an apparent fluctuation at $k = 443$. By inspecting the nodes corresponding to such value, one discovers that the sharp rise taking place at $k = 443$ corresponds to a non-existing resource labeled by 444 distinct uncorrelated randomly chosen tags. This feature closely mimics a spam contribution to the collaborative systems. One can think that the overall semantical organization of concept represented by tags is encoded in a characteristic behavior of the clustering coefficient $C(k)$. If this is true, tags assigned in a semantically inconsistent way fall far away from this behavior. To verify such conjecture, we have performed the same statistical analysis after removing from the data set the tag assignments related to the spam-like page. As shown in figure 4, after the removal the clustering coefficient follows a more regular behavior, confirming that the strong fluctuation observed above was indeed due to the presence of a single meaningless set of assignments involving a single resource. Tags assigned only to the spam resource form a complete co-occurrence network, so that their clustering coefficient is equal to 1. Thus, the behavior of the clustering coefficient of the tag co-occurrence networks can be used as a test for models representing the tag semantical organization or, equivalently, how users choose tags when annotating a resource. As noted in the literature [8], users typically use tags hierarchically, labeling a resource by tags related to the same topics but with different generality, adding more specialized tags as the number of collected resources grows.

On a very basic level, we have tested how such hierarchical tagging, affects the topology of the tag co-occurrence network by a simple toy model defined in the following. Let us assume that tags are organized on a taxonomy, that is, a tree-like structure stemming from a

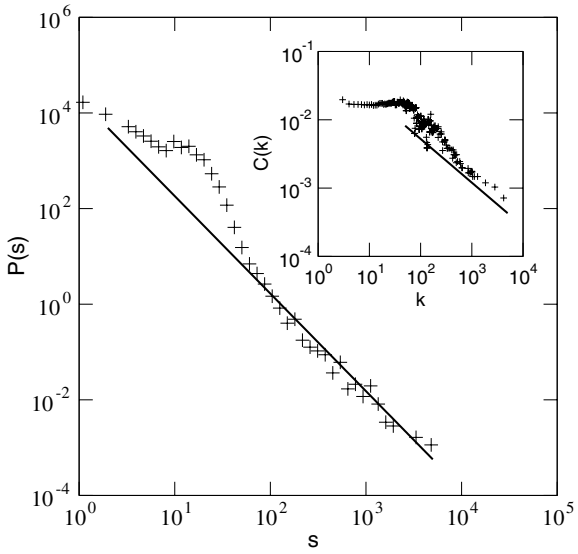

Figure 5. Main plot: the strength distribution in the co-occurrence network derived from the model with $p_b = 0.25$ (plus symbols); the solid line represents the decay $s^{-2.04}$ for a comparison with real data. Inset: the clustering coefficient in the co-occurrence network derived from the model with $p_b = 0.25$ (plus symbols). The solid line represents the decay $k^{-0.64}$ for a comparison with real data.

seed node, where each node corresponds to a tag and is an offspring of another tag belonging to the same branch of knowledge with higher generality. At discrete time steps, a new post is added to the system, with a new resource and two tags. The first tag can be a new one, with probability p_g: in such a case, the new tag is an offspring of a tag randomly chosen among the already employed ones. Otherwise, the first tag is chosen at random among the already employed ones. The second tag is either chosen at random from within the whole set of used tags or, with probability p_b, it is chosen according to hierarchy. In this latter case, the second tag is drawn randomly among the nodes that lie on the shortest path length from the first tag to the seed node on the tree-like taxonomy.

The tag co-occurrence network resulting from the above algorithm share some features of the *CiteULike* one, if we assume a time-dependent p_g which reproduces the sublinear vocabulary growth observed in reality, and by a suitable choice of the parameter p_b, which mimics the relevance of hierarchy in tagging activity. The simplest functional form that reproduce the observed behavior is

$$p_g(t) = At^{-B} \tag{4}$$

then we have that

$$\int_1^{N_{\text{res}}} p_g(t)\,\mathrm{d}t = N_{\text{tag}} \tag{5}$$

and

$$N_{\text{tag}} = N_{ass}^{\beta}, \tag{6}$$

where the number of resources N_{res}, the number of tag assignments N_{ass}, the number of tags N_{tag} and β are set to the same values they have in the real case of *CiteULike*. As a result, this yields $A = 5$ and $B = 0.3$.

As shown in figure 3, the strength distribution $P(s)$ of tags is a scale free one with a good agreement with reality in the decaying exponent for large values of s if one sets $p_b = 0.25$. For such choice of the parameter, the clustering coefficient reproduces qualitatively the algebraic decay observed in *CiteULike*, as shown in figure 5, although the absolute value differs of orders of magnitude.

In conclusion we present here a simple model that captures the complex features of a tag co-occurrence network issued from the dataset describing an online collaborative tagging system, *CiteULike*. In particular, (by assuming that users label resources by hierarchically associated tags) the probability distribution of strength is reproduced for a suitable choice of the parameters. Moreover, the model reproduces qualitatively the decaying asymptotic behavior of the clustering coefficient $C(k)$. Such quantity encodes a signature of the semantical organization of concepts represented by tags, so that malicious or meaningless tag assignments can be detected by inspecting the perturbation to the clustering coefficient they generate. Establishing a relationship between clustering and semantics may suggest tools and algorithms for technological tasks such as automatic categorization of resources, recommendation and spam detection techniques.

Acknowledgments

The authors acknowledge useful discussions with Francesca Colaiori, Stefano Leonardi, Ciro Cattuto, Vito DP Servedio and Andrea Baldassarri. The authors acknowledge the European project DELIS for support.

References

[1] *See the online URL http://www.citeulike.org*
[2] Zipf G K 1949 *Human Behavior and the Principle of Least Effort* (Reading, MA: Addison-Wesley)
[3] Zanette D H and Montemurro M A 2005 *J. Quant. Linguist.* **12** 29
[4] Simon H A 1955 *Biometrika* **42** 425
[5] Yule G U 1925 *Phil. Trans. R. Soc.* B **213** 21
[6] Ferrer i Cancho R and Servedio V D P 2005 *Glottometrics* **11** 1
[7] Caldarelli G 2007 *Scale-Free Networks* (Oxford: Oxford University Press)
[8] Heyman P and Garcia-Molina H 2006 Collaborative creation of communal hierarchical taxonomies in social tagging systems *Technical Report InfoLab 2006-10* (Stanford, CA: Department of Computer Science, Stanford University)
[9] Berendt B, Hotho A and Stumme G 2002 *Lecture Notes Comput. Sci.* **2342** 264
[10] Cattuto C and Loreto V 2006 Vocabulary growth in collaborative tagging systems *Preprint* 0704.3316
[11] Huberman B A and Golder S 2006 *J. Inf. Sci.* **32** 198
[12] Lambiotte R and Ausloos M 2006 *Lecture Notes Comput. Sci.* **3993** 1114
[13] Santos-Neto E, Ripeanu M and Iamnitchi A 2007 *Proceedings of International ACM/IEEE Workshop on Contextualized Attention Metadata: personalized access to digital resources*
[14] Hotho A, Jaeschke R, Schmitz C and Stumme G 2006 *Lecture Notes Artif. Intell.* **4011** 411
[15] Schmitz C, Grahl M, Hotho A, Stumme G, Cattuto C, Baldassarri A, Loreto V and Servedio V D P 2007 Network properties of folksonomies *Proc. 16th Int. World Wide Web Conf. (WWW2007)*
[16] Albert R and Barabási A-L 2002 *Rev. Mod. Phys.* **74** 47
[17] Ravasz E and Barabási A-L 2003 *Phys. Rev.* E **67** 026112