Eur. Phys. J. C 46, s02, 33–37 (2006)
Digital Object Identifier (DOI) 10.1140/epjcd/s2006-03-004-4

EPJ C direct
electronic only

Young Scientist

Semiotic dynamics in online social communities

Ciro Cattuto[1,2,a]

[1] Museo Storico della Fisica e Centro Studi e Ricerche "Enrico Fermi" Compendio Viminale, 00184 Roma, Italy
[2] Dipartimento di Fisica, Università di Roma "La Sapienza" P.le A. Moro, 2, 00185 Roma, Italy

Received: 30 June 2006 /
Published online: 8 August 2006 – © Springer-Verlag / Società Italiana di Fisica 2006

Abstract. A distributed classification paradigm known as *collaborative tagging* has been successfully deployed in large-scale web applications designed to manage and share diverse online resources. Users of these applications organize resources by associating with them freely chosen text labels, or *tags*. Here we regard tags as basic dynamical entities and study the *semiotic dynamics* underlying collaborative tagging. We collect data from a popular system and focus on tags associated with a given resource. We find that the frequencies of tags obey to a generalized Zipf's law and show that a Yule–Simon process with memory can be used to explain the observed frequency distributions in terms of a simple model of user behavior

1 Collaborative Tagging

A new paradigm has been quickly gaining ground in information systems on the World Wide Web: collaborative tagging [1, 2]. In web-based applications like *Del.icio.us*[1], *Flickr*[2], *CiteULike*[3], users enrich diverse resources – ranging from photographs to scientific references and web pages – with semantically meaningful information in the form of text labels, or "tags". Tags are freely chosen and users associate resources with them in a totally uncoordinated fashion, for their own use.

The tagging activity of each user is globally visible to the user community (see Fig. 1) and the tagging process develops genuine social aspects and complex interactions. Remarkably, despite the selfish nature of users' behavior, tagging systems exhibit cooperative dynamics that eventually lead to a bottom-up categorization of resources, shared throughout the user community. The open-ended set of tags used within the system – commonly referred to as "folksonomy" – can be used as a sort of semantic map to navigate the contents of the system itself. It has been argued that the surging popularity of collaborative tagging is due to its comparatively small cognitive overhead with respect to taxonomic categorization [1, 3], so that tagging is a very natural activity for web users. Collaborative tagging systems leverage this aspect, recruiting simple and robust behaviors of individual users in order to create cooperation and foster the emergence of shared conventions at the system level.

Focusing on tags as basic dynamical entities, collaborative tagging falls within the scope of semiotic dynamics [4, 5], a new field that studies how populations of humans or agents can establish and share semiotic systems, typically driven by their use in communication. New web applications hinged on collaborative tagging fall precisely in this perspective and can be regarded as cases of semiotic dynamics at play: folksonomies, in fact, do exhibit dynamical aspects also observed in human languages [6, 7], such as the emergence of naming conventions, competition between terms, takeovers by neologisms, and more.

In the following we briefly describe the structure of collaborative tagging systems and discuss the experimental procedures that we employ to collect data from *Del.icio.us*, one of the most popular and paradigmatic social bookmarking systems. We adopt a resource-centric view of the system and focus on the tags that have been associated with a given resource. We report on the distribution of tag frequencies and show that a stochastic Yule–Simon process with memory can be used to successfully describe the observed frequency distributions. We close by casting our work in a more general research perspective.

2 Experimental data

The activity of users interacting with a collaborative tagging system consists of either navigating the existing body

[a] e-mail: ciro.cattuto@roma1.infn.it

[1] http://del.icio.us/

[2] http://flickr.com/

[3] http://citeulike.org/

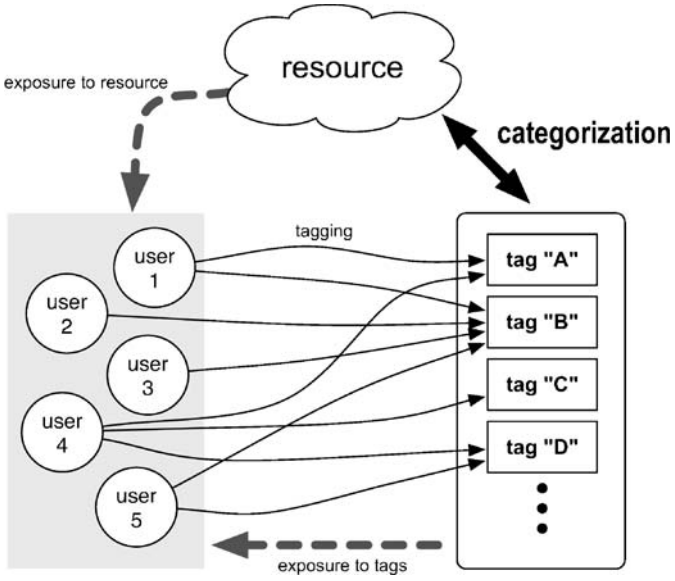

Fig. 1. Schematic depiction of the collaborative tagging process: web users (*circles on the left*) are exposed to a resource and freely associate tags with it (*rectangles on the right*). In their interaction with the system users are also exposed to tags previously entered by themselves and by other users. The collective tagging activity creates a dynamical correspondence between a resource and a set of tags, i.e. an emergent categorization in terms of tags shared by a community

of resources by using tags, or of adding new resources. In order to add a new resource into the system the user is prompted for a reference to the resource and a set of tags to associate with it. Thus, the basic unit of information in a collaborative tagging system is a (user, resource, {tags}) triple, here referred to as "post" (see Fig. 2). Tagging events build a tripartite graph with partitions corresponding to users, resources and tags, respectively. A post typically contains a temporal marker recording the (physical) time of the tagging event, so that temporal ordering can be preserved in storing and retrieving posts.

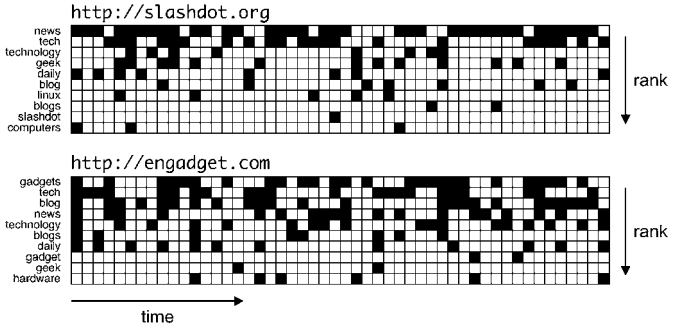

Fig. 2. Tagging activity: an experimental sequence of tagging events on *Del.icio.us* is graphically rendered by displaying the tags that were used in posts associated with two popular resources (web pages). In each panel, *columns* represent single tagging events (posts) and *rows* correspond to the 10 tags most frequently associated with the given resource. 50 tagging events are shown in each panel, temporally ordered from left to right. Only events involving at least one of the 10 top-ranked tags are shown. For each tagging event (*column*), a *filled cell* marks the presence of the tag in the corresponding row, while an *empty cell* indicates its absence

Our analysis will focus on *Del.icio.us* for several reasons: i) it was the first system to deploy the ideas and technologies of collaborative tagging, so it has acquired a paradigmatic character and it is the natural starting point for any quantitative study. ii) because of its popularity, it has a large community of active users and comprises a precious body of raw data on the static and dynamical properties of a folksonomy. iii) it is a *broad folksonomy* [8], i.e. single tagging events (posts) retain their identity and can be individually retrieved. This affords unimpeded access to the "microscopic" dynamics of collaborative tagging, providing the opportunity to make contact between emergent behaviors and low-level dynamics. It also allows us to define and measure the multiplicity (or frequency) of tags in the context of a single resource. Contrary to this, popular sites falling in the *narrow folksonomy* class (*Flickr*, for example) foster a different model of user interaction, where tags are mostly applied by the content creator, no notion of tag multiplicity is possible in the context of a resource, and no access is given to the raw sequence of tagging events. On studying *Del.icio.us* we adopt a resource-centric view of the system, that is we investigate the dynamical correspondence between a given resource and the tags that users associate with it. In line with our focus on semiotic dynamics, we factor out the detailed identity of the users involved in the process and only deal with streams of tagging events and their statistical properties.

To perform automated data collection of raw data we use a custom web (HTTP) client that connects to *Del.icio.us* and navigates the system's interface as an ordinary user would do, extracting the relevant metadata and storing it for further post-processing. *Del.icio.us* allows the user to browse its content by resource: our client requests the web page associated with the resource under study and uses an HTML parser to extract the post information (user, tags and time stamp). Figure 2 graphically depicts the raw data we gather for the case of two popular resources on *Del.icio.us*. The data used for the present analysis were retrieved in November 2005.

3 Temporal evolution

Figure 3 displays the amount of tagging data associated with a single popular resource (the same shown in the top panel of Fig. 2) as a function of time. The time stamp attached to each post is used to build a time series of tagging events and cumulated values are shown for the number of entries associated with the resource and the total number of tags used. The data shown in Fig. 3 span a time interval of about 2 years, during which the popularity of *Del.icio.us* surged and its user base increased enormously. Correspondingly, the amount of tagging data associated with the selected resource increased by several orders of magnitude. In striking contrast with this, the relative proportions of

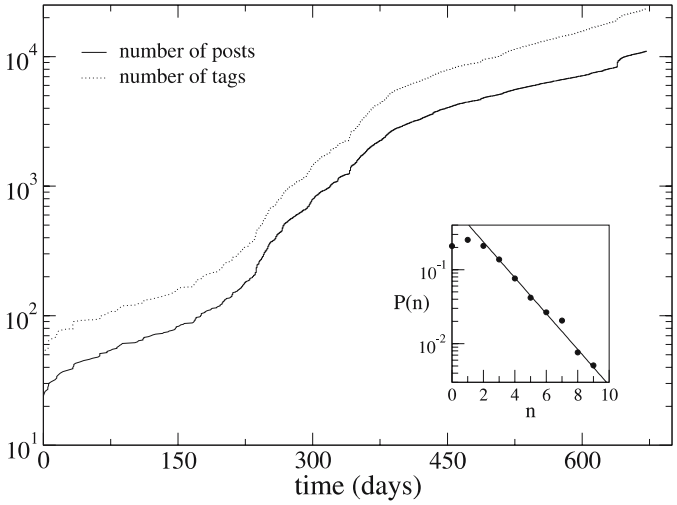

Fig. 3. Amount of metadata associated with a popular resource in *Del.icio.us*, as a function of time. Data are shown for the same resource of Fig. 2, top panel. Over a time interval of about 2 years, the number of posts (*solid lines*) and the number of total tags (*dotted line*) increased by several orders of magnitude. *Inset*: the fraction $P(n)$ of posts containing n tags displays an exponential tail for high values of n. The average value $\bar{n}$ corresponds to the vertical offset between the *solid line* and the *dotted line*

tags associated with a given resource quickly approach a quasi-stationary condition: in Fig. 4 we plot the fraction of occurrence of the 10 top-ranked tags of Fig. 2 (top panel) as a function of the number of posts associated with

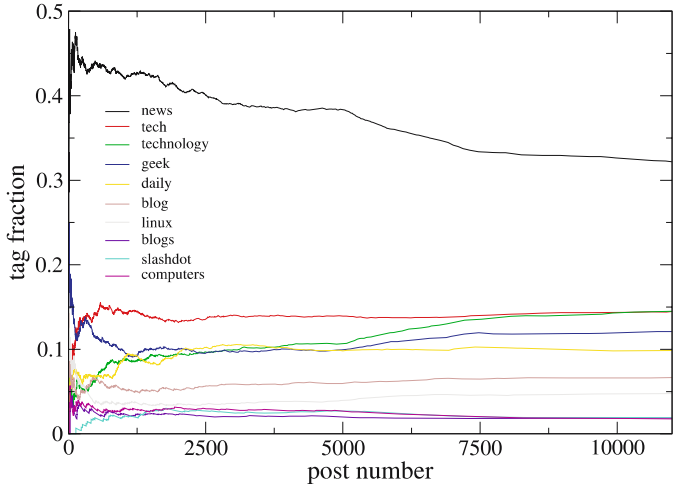

Fig. 4. For a given resource (Fig. 2, *top panel*) the relative fraction of cumulated tag occurrences is shown as a function of time, measured by the number of posts associated with the resource. Tag fractions are plotted for the 10 most frequent tags associated with the resource (legend, same tags as Fig. 2, top panel). After an initial transient, the relative proportions of tag occurrence freeze towards (approximately) steady values, even though the amount of accumulated metadata never levels off (Fig. 3)

the resource. After an initial transient, during which the fractions fluctuate significantly, the relative proportion of each tag settles towards an approximately constant value. This is reminiscent of the behavior observed in Polya's urn problems [9] and suggests the existence of a multiplicative process underlying the tagging activity of users [10], where the usage pattern of tags is subjected to some kind of frequency-bias. Once the number of posts associated with a resource is sufficiently large, single tagging events have a negligible effect on the global distribution of tags and the existing distribution is reinforced, generally becoming more and more stable. This kind of robustness is a very important property of collaborative tagging: on the one hand, the fact that tag fractions stabilize quickly allows the emergence of a clearly defined categorization of the resource in terms of tags, with a few top-ranked tags defining a semantic "fingerprint" of the resource they refer to (see also Fig. 2). On the other hand, the long-term stability of tag proportions makes the emergent categorization robust against noise. Both aspects contribute greatly to the actual usability of collaborative tagging systems. Occasionally, interesting non-stationary behaviors and transitions can be observed, where new tags are invented and become socially adopted by the user community [10].

4 Tag frequencies

To better probe the emergent categorization of a resource in terms of tags, we compute the standard frequency-rank distributions for the tags associated with the resources of Fig. 2, i.e. we compute the number of occurrences of tags and rank them: our results are shown in Fig. 5 (black dots). The high-rank tails of the experimental curves display a power-law behavior reminiscent of Zipf's law [11] which is characteristic of self-organized communication systems and is commonly observed in natural languages and written text [12]. The observed exponent of the power law is greater than 1 (the lines in Fig. 5 have slope 4/3) and its explanation requires more complex microscopic mechanisms than those usually invoked to explain Zipf's law [13]. Moreover, the low-rank part of the distributions of Fig. 5 displays a flattening behavior typically not observed in systems strictly obeying Zipf's law, related to the co-existence (and possibly the competition) of the few low-rank tags which characterize the resource in the statistically strongest way.

In order to model the observed frequency-rank behavior, we use a Yule–Simon stochastic process [14–16] with long-term memory, introduced in [17] (Fig. 6). We move from the observation that actual users are in principle exposed to all the tags stored in the system (like in the original Yule–Simon model), but the way in which they choose among them, when tagging a new post, is far from being a simple uniform distribution (see also [18]). It seems more realistic to assume that users tend to use recently added tags more frequently than old ones, according to a skewed memory kernel. Our modification of the Yule–Simon's model consists in weighting the probability

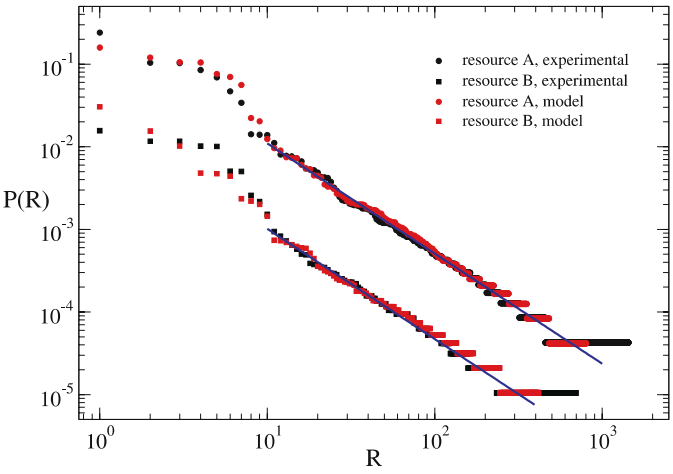

Fig. 5. Frequency-rank plots for the tags associated with a given resource. Experimental data (black symbols) are shown for the resources of Fig. 2 (top panel and bottom panel correspond to resource A and B, respectively). For the sake of clarity the data for resource B were shifted down by one decade. The high-rank tail of the curves displays a power-law behavior corresponding to a generalized Zipf's law. For reference, the *blue lines* correspond to the power law $R^{-4/3}$. *Red symbols* are theoretical data obtained by computer simulation of the stochastic model described in the text (Fig. 6). The parameters of the model, i.e. the probability p, the memory parameter τ and the initial number of words n_0 were adjusted to match the experimental data, yielding approximately $p = 0.035$, $\tau = 175$ and $n_0 = 10$ for resource A, and $p = 0.04$, $\tau = 170$ and $n_0 = 10$ for resource B

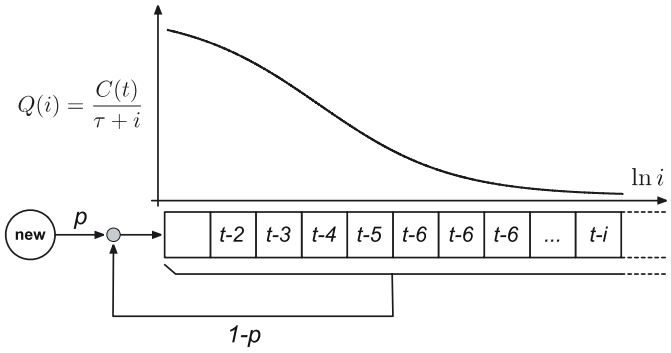

Fig. 6. A Yule–Simon's process with memory. A stream of tags is generated by iterating the following step: with probability p a new tag is created and appended to the tag stream, while with probability $1-p$ a tag is extracted from the past history of the system and appended to the text, going back in time by i steps with a probability $Q(i) \sim 1/(\tau + i)$

of choosing an existing tag according to a power-law memory kernel which controls the visibility of the past history of the resource. This hypothesis about the functional form of the memory kernel is supported by findings in cognitive psychology [19] (where power laws of latency and frequency have been shown to model human memory) as well as by recent analysis on patterns of human activity [20]. Specifically, previous work on tag co-occurrence [17] shows that a hyperbolic memory kernel is needed to match the experimental data. Thus, our model (Fig. 6) can be stated as follows: the process by which users of a collaborative tagging system associate tags to resources can be regarded as the construction of a "text", built one step at a time by adding "words" (i.e. tags) to a text initially comprised of n_0 words. This process is meant to model the behavior of an effective average user in the context of a single resource. At a generic time step t, a new word may be invented with probability p and appended to the text, while with probability $1-p$ one word is copied from the existing text, going back in time by i steps with a probability that decays with the offset i as a power law, $Q(i) = C(t)/(\tau + i)$. $C(t)$ is a time-dependent normalization factor and τ is a characteristic time-scale over which recently added words have comparable probabilities. By simulating the stochastic process described above and adjusting the values of the model parameters in order to match the experimental data, we obtain a very good agreement between experimental data (black dots) and model (red dots), as shows in Fig. 5.

This proves that the simple Yule–Simon process with fat-tailed memory can be successfully applied to model the frequency-rank distribution of tags associated with a given resource, providing insights into the average behavior of users in the context of that resource.

5 Conclusions and perspective

Information systems on the World Wide Web have been increasing in size and complexity to the point that they presently exhibit features typically attributed to *bona fide* complex systems. They display rich high-level behaviors that are causally connected in non-trivial ways to the dynamics of their interacting parts. Because of this, concepts and formal tools from the science of complex systems can play a potentially important role in understanding and designing the behavior of such systems. Here we made a first step in this direction and focused on the new paradigm of collaborative tagging. Collaborative tagging systems involve the computer-mediated interaction of a large number of human agents. They have an emergent semiotic system (the folksonomy) at their core and can be regarded as a sort of "laboratory" of semiotic dynamics. In addition to this, their evolution can be easily monitored both at the global level (folksonomy) and at the agent-level (single tagging events of users), providing an interesting opportunity to connect the two levels by means of a suitable theoretical framework. We grounded our work on actual tagging data and made use of statistical tools and stochastic models to gain insights into the dynamical properties of collaborative tagging. We studied the frequency distributions of tags and found that a simple stochastic process, a long-term memory version of the classic Yule–Simon process, is able to describe very accurately the experimental data, allowing us to develop a model of user behavior and link it to the statistical properties of the folksonomy. Overall, our findings suggest that users of collaborative tagging systems share universal behaviors which, despite the intricacies of

personal categorization, tagging procedures and user interactions, appear to obey to simple activity patterns.

Acknowledgements. The author wishes to thank V. Loreto, L. Pietronero, A. Baldassarri, A. Baronchelli, V. Servedio and L. Steels for stimulating discussion and suggestions. This research has been partly supported by the ECAgents project funded by the Future and Emerging Technologies program (IST-FET) of the European Commission under the EU RD contract IST-1940. The information provided is the sole responsibility of the authors and does not reflect the Commission's opinion. The Commission is not responsible for any use that may be made of data appearing in this publication.

References

1. A. Mates, Folksonomies – Cooperative Classification and Communication Through Shared Metadata, Computer Mediated Communication, LIS590CMC, Graduate School of Library and Information Science, University of Illinois Urbana-Champaign (2004)
2. T. Hammond, T. Hannay, B. Lund, J. Scott, Social Bookmarking Tools (I): A General Review, D-Lib Magazine **11**, (2005)
3. R. Sinha, http://rashmisinha.com/archives/05_09/tagging-cognitive.html (2005)
4. L. Steels, F. Kaplan, Collective learning and semiotic dynamics. In D. Floreano J.-D. Nicoud, F. Mondada, (Eds.), Advances in Artificial Life: 5th European Conference (ECAL 99), Lecture Notes in Artificial Intelligence **1674**, 679-688, Berlin, (1999)
5. J. Ke, J.W. Minett, A. Ching-Pong, W.S-Y. Wang, Self-organization and selection in the emergence of vocabulary, Complexity **7**, 41–54 (2002)
6. M.A. Nowak, N.L. Komarova, P. Niyogy, Computational and evolutionary aspects of language, Nature **417**, 611–617 (2002)
7. S. Kirby, Natural language and artificial life, Artificial Life **8**, 182–215 (2002)
8. T. Vander Wal, Explaning and Showing Broad and Narrow Folksonomies, http://www.personalinfocloud.com/2005/02/explaining_and_.html, (2005)
9. F. Eggenberger, G. Polya, Über die Statistik verketter vorgage, Zeit. Angew. Math. Mech. **1**, 279 (1923)
10. S. Golder B.A. Huberman, J. Inf. Sci. **32**, 198 (2006); e-print cs/0508082
11. G.K. Zipf, Human Behavior and the Principle of Least Effort (Addison-Wesley, Cambridge, MA, 1949)
12. http://linkage.rockefeller.edu/wli/zipf/index_ru.html
13. R. Ferrer Cancho, V.D.P. Servedio, Glottometrics **11**, 1 (2005)
14. M.E.J. Newman, Contemp. Phys. **46**, 323 (2005)
15. G.U. Yule, Philos. Trans. R. Soc. London B **213**, 21 (1925)
16. H.A. Simon, Biometrika **42**, 425 (1955)
17. C. Cattuto, V. Loreto, L. Pietronero, Semiotic Dynamics and Collaborative Tagging, submitted (2006); e-print cs/0605015
18. D.H. Zanette, M.A. Montemurro, J. Quant. Linguist. **12**, 29 (2005)
19. J.R. Anderson, Cognitive Psychology and its Implications, fifth edition, (Worth Publisher, New York, 2000)
20. A.-L. Barabasi, Nature **435**, 207 (2005)