Semiotic dynamics and collaborative tagging

Ciro Cattuto[1,2*], Vittorio Loreto[2†] & Luciano Pietronero[2‡]

[1]*Museo Storico della Fisica e Centro Studi e Ricerche "Enrico Fermi", Compendio Viminale, 00184 Rome, Italy.*

[2]*"La Sapienza" University, Physics Dept., Piazzale A. Moro 2, 00185 Rome Italy*

co-occurring with the selected one, we find a heavy-tailed behavior which is the mark of human activity[8,9] and observe properties that point to an emergent hierarchy of tags. We introduce a stochastic model embodying two main aspects of collaborative tagging: (i) a fundamental multiplicative character closely related to the idea that users are exposed to each other's tagging activity[10,11,12]; (ii) a notion of memory - or aging of resources - in the form of a heavy-tailed access to the past state of the system. Remarkably, our simple modelling is able to account quantitatively for the measured frequency-rank properties of tag association, with a surprisingly high accuracy. This is a clear indication that collaborative tagging is able to recruit the uncoordinated actions of web users to create a predictable and coherent semiotic dynamics at the emergent level.

Eight years after the first vision of the Semantic Web by Tim Berners-Lee[13,14] a set of new semantically-enabled applications is swiftly shaping the next generation of the World-Wide Web (the so-called WEB 2.0). One of the forces driving this change is a distributed classification paradigm known as "*collaborative tagging*", which has been successfully deployed in web application designed to organize and share diverse online

[*] e-mail address: ciro.cattuto@roma1.infn.it (to whom correspondence should be sent)

[†] e-mail address: vittorio.loreto@roma1.infn.it

[‡] e-mail address: luciano.pietronero@roma1.infn.it

resources such as web pages, academic references, photographs and music. Web users interact with a collaborative tagging system by posting content (*resources)* into the system, and freely associating text labels (*tags*) with that content, as shown in Fig. 1. The basic unit of information in a collaborative tagging system is an **entry**, that is a **(user, resource, {tags})** triple, with the largest online communities comprising hundreds of thousands of users and millions of resources. Users are exposed both to the resources and to the tags already existing within the system, and freely associate tags with newly entered resources. At the global level the set of tags, though freely determined, evolves in time and leads towards patterns of terminology usage that are shared by the entire user community. Hence one observes the emergence of a loose categorization system -- commonly referred to as "*folksonomy*" -- that can be effectively used to navigate through a large and heterogeneous body of resources.

Focusing on tags as basic dynamical entities, the process of collaborative tagging falls within the scope of Semiotic Dynamics[5,6], 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 (such as *del.icio.us* or *Flickr*) fall precisely in this perspective and they can be regarded as cases of Semiotic Dynamics at play: the emergence of a folksonomy exhibits dynamical aspects also observed in human languages [15,16], such as the crystallization of naming conventions, competition between terms, takeovers by neologisms, and more.

Fig.1: Schematic depiction of the collaborative tagging process: web users are exposed to a resource and freely associate tags with it. Their interaction with the system also exposes them to tags previously entered by themselves and by other users. The aggregated activity of users leads to an emergent categorization of resources in terms of tags shared by a community.

Data collection and analysis. Here we collect data from *del.icio.us* and *Connotea* (see the Methods section) and investigate the statistical properties of tag association. Specifically, we select a semantic context by extracting all the resources associated with a given tag *X* and study the statistical distribution of tags co-occurring with *X* (see Table 1). Fig 2-a graphically illustrates the associations between tags and entries and Fig. 2-b reports the frequency-rank plots for all the tags co-occurring with a few selected tags (*blog*, *ajax* and *xml for del.icio.us* and *H5N1* for *Connotea*). The high-rank tail of the experimental curves displays a power law behaviour, signature of an emergent hierarchical structure, corresponding to a generalized Zipf's law with a slope larger than one[12]. The low-rank part of the curve, conversely, displays a flattening behaviour typically not observed in systems obeying Zipf's law. Several aspects of the underlying complex dynamics are responsible for this deviation: on the one hand this behaviour points to the existence of semantically equivalent and possibly competing high-frequency tags (e.g. *blog* and *blogs*). More importantly, this flattening behaviour may be ascribed to an underlying hierarchical organization of tags co-occurring with the one we single out. In this scenario, we expect a more shallow behaviour for tags co-occurring with generic tags (e.g. *blog*, see Fig. 2-b) and a steeper behaviour for semantically narrow tags (e.g. *ajax*, denoting a specific technology, also in Fig. 2-b). To better probe the validity of this interpretation, we investigated the co-occurrence relationship that links high-rank tags, lying well within the power-law tail, with low-rank tags, located in the shallow part of the distribution. Our observations (see Supplementary Information) point in the direction of a non-trivial hierarchical organization emerging out of the collective tagging activity, with each low-rank tag leading its own hierarchy of semantically related higher-rank tags, and all such hierarchies merging into the overall power-law tail. We shall elaborate on this point in the theoretical section.

Table 1: Dataset statistics

Tag	No. entries	No. tags	No. different tags	No. resources
Blog	37974	124171	10617	16990
Ajax	33140	108181	4141	2995
Xml	24249	108013	6035	7364
H5N1	981	5185	241	969

Table 1: Statistics of the datasets used for the co-occurrence analysis. For each tag in the first column we report the number of entries marked with that tag, the number of total and distinct tags co-occurring with it and the corresponding number of resources.

Fig. 2: **a)** Tagging activity: a time-ordered sequence of tagging events is graphically rendered by marking the tags co-occurring with *blog* (top panel) or *ajax* (bottom panel) in an experimental sequence of resources entered into *del.icio.us*. In each panel, columns represent single tagging events and rows correspond to the 10 most frequent tags co-occurring with either *blog* (top panel) or *ajax* (bottom panel). 150 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. A qualitative difference between *blog* (top panel) and *ajax* (bottom panel) is clearly visible, where a higher density at low-rank tags characterize the semantically narrower *ajax* term. This corresponds to the steeper low-rank behavior observed in the frequency-rank plot for *ajax* (Fig. 2-b). **b)** Frequency-rank plots for tags co-occurring with a selected tag: experimental data (black symbols) are shown for *del.icio.us* (circles for tags co-occurring with the popular tag *blog*, squares for *ajax* and triangles for *xml*) and *Connotea* (inset, black circles for the *H5N1* tag). For the sake of clarity, the curves for *ajax* and *xml* are shifted down by one and two decades, respectively. Details about the

experimental datasets are reported in Table 1. All curves exhibit a power-law decay for high ranks (a dashed line corresponding to the power law $R^{-5/4}$ is provided as an aid for eye) and a more shallow behaviour for low ranks. To make contact with Fig. 2-a, some of the top-ranked tags co-occurring with *blog* and *ajax* are explicitly indicated with arrows. Red symbols are theoretical data obtained by computer simulation of the stochastic model described in the text (Fig. 3). The parameters of the model, i.e. the probability p, the short-term memory parameter t and the initial number of words n_0 were adjusted to match the experimental data, giving approximately $p = 0.06$, $t = 100$ and $n_0 = 100$ for *blog*, $p = 0.03$, $t = 20$ and $n_0 = 50$ for *ajax*, and $p = 0.034$, $t = 40$ and $n_0 = 110$ for *xml*. *Connotea* is a much younger system than *del.icio.us* and the corresponding data set is smaller and noisier. Nevertheless, a good match with experimental data can be obtained for $p = 0.05$, $t = 120$ and $n_0 = 7$ (inset, red circles), demonstrating that our model can be also applied to the early stages of development of a folksonomy. **c)** Tag-tag correlation functions and non-stationarity. The tag-tag correlation function $C(\Delta t, t_w)$ is computed over three consecutive and equally long ($T = 30000$ tags each) subsets of the *blog* dataset, starting respectively at positions $t_w^1 = 10000$, $t_w^2 = 40000$ and $t_w^3 = 70000$ within the collected tag stream. Short-range correlations are clearly visible, slowly decaying towards a long-range plateau value. The non-stationary character of correlations is visible both at short range, where the value of the correlation function decays with t_w, and at long range, where the asymptotic correlation increases with t_w. The long-range correlations (dashed lines) can be estimated as the natural correlation present in a random stream containing a finite number of tags: on using the appropriate ranked distribution of tag probabilities within each window (see text) the values $c(t_w^1)$, $c(t_w^2)$ and $c(t_w^3)$ can be computed, matching the meas ured plateau of the correlation functions. The thick solid line is a fit with a long-range memory kernel (see text).

A Yule-Simon's model with long-term memory. We now aim at gaining a deeper insight in the phenomenology reported above. In order to model the observed frequency-rank behavior for the full range of ranking values, we introduce a new version of the "rich-get-richer" Yule-Simon's stochastic model[10,11] by enhancing it with a long-term memory kernel. The original model can be described as the construction of a text from scratch. At each discrete time step one appends a word to the text: with probability p the appended word is a new word, that has never occurred before, while with probability $1-p$ the word is copied from the existing text, choosing it with a probability proportional to its current frequency of occurrence. This simple process produces frequency-rank distributions with a power-law tail whose exponent is given by $a = 1 - p$. In our construction we moved from the observation that actual users are exposed in principle to all the tags stored in the system (like in the original Yule-S imon model) but the way in which they choose among them, when they tag a new entry, is far from being a simple uniform distribution (see also [17]). It seems more realistic to assume that users tend to apply recently added tags more frequently than old ones, according to a memory kernel which might be highly skewed. Indeed, recent findings about human activities[9] support the idea that the access pattern to the past of the system should be fat-failed, suggesting a power-law memory kernel. Let us test this hypothesis with real data extracted from *del.icio.us*: in Fig. 2-c we show the temporal auto-correlation function of the sequence of tags co-occurring with *blog*. Correlations are computed inside three consecutive windows of length T, starting at different times t_w,

$$C(\Delta t, t_w) = \frac{1}{T-\Delta t} \sum_{t=t_w}^{t=t_w+T-\Delta t} d\,(tag(t+\Delta t), tag(t)),$$

where $d(tag(t+\Delta t), tag(t))$ is the usual Kronecker delta function, taking the value 1 when the same tag occurs at times t and $t+\Delta t$. From Fig.2-c it is apparent how the correlation function is non-stationary over time. Moreover, for each value of the initial time t_w it displays a power-law behaviour $a(t_w)/(\Delta t + d(t_w)) + c(t_w)$, where $a(t_w)$ is a time-dependent normalization factor and $d(t_w)$ is a phenomenological time-scale, slowly increasing with the "age" of the system t_w, whose interpretation is related to the number of independent hierarchies nested into the emergent categorization. $c(t_w)$ is the correlation that one would expect in a random sequence of tags distributed according to the frequency-rank distribution $P_{T,t_w}(R)$ pertaining to the data windows we selected. Denoting by $R_{\max}(T,t_w)$ the number of distinct tags occurring in the window $[t_w, t_w + T]$, we have $c(t_w) = \sum_{R=1}^{R=R_{\max}(T,t_w)} P^2_{T,t_w}(R)$.

Fig. 3: A Yule-Simon's process with long-term memory. A synthetic 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, its probability being weighted by the long-range memory kernel $Q_t(x)$, which provides a fat-tailed access to the past of the stream.

Our modification of the Yule-Simon's model thus consists in weighting the probability of choosing an existing word (tag) according to a power-law. This hypothesis about the functional form of the memory kernel is also supported by findings in Cognitive Psychology[18] (where power laws of latency and frequency have been shown to model human memory) as well as recent analysis on patterns of human activity[9], as mentioned above. Thus, our model 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 behaviour of an effective average user in the context identified by a specific tag. At a generic (discrete) time step t, a brand 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 x steps with a probability that decays as a power law, $Q_t(x) = a(t)/(x+t)$. $a(t)$ is a normalization factor and t is the characteristic time-scale over which the recently added words have comparable probabilities. Its interpretation is similar to that of the d parameter we used to model the tag-tag correlation functions (Fig. 2-c). In our simple model the average user is exposed to a few equivalent top-ranked tags, and this is translated mathematically in a low-rank cut-off of the power law. Fig.2-b shows the excellent agreement between the experimental data and the numerical predictions of our Yule-Simon's model with long-term memory. The parameter t can be interpreted as the signature of a semiotic dynamics underlying the categorization process, since it affects the number of different top-ranked tags which are perceived by users as semantically independent, loosely corresponding to co-existing and non-overlapping categories. This picture is confirmed by that fact that the value of t needed to match the experimental data for *blog* (a rather generic tag) is larger than the one needed for *ajax* (a pretty specific tag). This suggests that users of collaborative tagging systems share a universal behaviour which, despite the intricacies of personal categorization, tagging procedures and user interactions, appears to obey simple activity patterns.

Conclusions. Uncovering the mechanisms governing the emergence of shared categorizations or vocabularies in absence of global coordination is a key problem with significant scientific and technological potential. The theoretical challenges are relevant for diverse areas where researchers are faced with the problem of taming the information overload and unleashing the full potential of computer-mediated social interaction. Collaborative tagging can provide precious hints and tools to both analyze and design large (human or artificial) communicating systems. Here we report for the first time about the emergence of a self-organized hierarchical structure in the co-occurrence of tags. Furthermore we show how a simple modelling scheme for collaborative tagging, which only takes into account two basic elements underlying the tagging process at the user level, is able to reproduce to a surprisingly level of accuracy the phenomenology of online social systems at the emergent level. In addition to the findings reported and discussed in this Letter, our approach constitutes a starting point upon which studies of greater complexity can be based, with the final goal of understanding, predicting and controlling the Semiotic Dynamics of on-line social systems.

Methods. Collaborative tagging systems are designed from the ground up to be information sharing systems for web-based interactive use, so that most of the information they contain is accessible by using a web browser. To perform automated data collection, a custom web (HTTP) client was written that connects to *del.icio.us* (or *Connotea*) and navigates the system's interface as an ordinary user would do, logging the relevant information for further post-processing. Both *del.icio.us* and *Connotea* allow the user to browse their content by tag: our client requests all the entries associated to the tag under study and uses an HTML parser to extract the post information (resource, user, tags) from the returned HTML code. Usually, only the most recent posts are returned, together with links to the previous history of the system: our client follows those links and repeats the parsing stage over and over to trace back the history of all posts associated to a given tag.

[1] Mates A, Folksonomies - Cooperative Classification and Communication Through Shared Metadata, *Computer Mediated Communication*, LIS590CMC.

[2] Hammond T., Hannay T., Lund B. & Scott J., Social Bookmarking Tools (I): A General Review, *D-Lib Magazine* **11**(4), 2005.

[3] Lund B., Hammond T., Flack M. & Hannay T., Social Bookmarking Tools (II): A Case Study - Connotea, *D-Lib Magazine* **11**(4), 2005.

[4] Golder S. & Huberman B.A., The Structure of Collaborative Tagging Systems, cs.DL/0508082 (2005).

[5] Steels L. & Kaplan F., Collective learning and semiotic dynamics. In Floreano, D. and Nicoud, J-D and Mondada, F., editor, Advances in Artificial Life: 5th European Conference (ECAL 99), *Lecture Notes in Artificial Intelligence* **1674**, 679-688, Berlin, (1999).

[6] Ke J., Minett J.W., Ching-Pong A., Wang W.S.-Y., Self-organization and selection in the emergence of vocabulary, *Complexity* **7**, 41-54 (2002).

[7] Zipf G.K., *Human Behavior and the Principle of Least Effort*, Addison-Wesley, Cambridge, MA, (1949).

[8] Newman M.E.J., Power laws, Pareto distributions and Zipf's law, *Contemporary Physics* **46**(5), 323-351 (2005).

[9] Barabasi A.-L., The origin of bursts and heavy tails in human dynamics, *Nature* **435**, 207 (2005).

[10] Yule G.U., A Mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, *Philos. Trans. R. Soc. London B* **213**, 21-87 (1925).

[11] Simon H.A., On a class of skew distribution functions, *Biometrika* **42**, 425 (1955).

[12] Ferrer Cancho R. & Servedio V.D.P., Can simple models explain the Zipf's law for all exponents? *Glottometrics* **11**, 1 (2005).

[13] Berners-Lee, T.: Realising the full potential of the Web, 1997. http://www.w3.org/1998/02/ Potential.html/

[14] Berners-Lee, T.: A roadmap to the Semantic Web, 1998. http://www.w3.org/DesignIssues/Semantic.html

[15] Nowak M.A., Komarova N.L. and Niyogy P., Computational and evolutionary aspects of language, *Nature* **417**, 611-617 (2002).

[16] Kirby S., Natural language and artificial life, *Artificial Life* **8**, 182-215 (2002).

[17] Zanette D.H. & Montemurro M.A., Dynamics of text generation with realistic Zipf's distribution, *Journal of Quantitative Linguistics* **12**, 29-40 (2005).

[18] Anderson J.R., Cognitive Psychology and its implications, fifth edition, Worth Publisher, New York, (2000).

Acknowledgements The authors wish to thank A. Baronchelli and A. Baldassarri for many interesting discussions 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.

Competing Interests statement: The authors declare that they have no competing financial interests.