From the Cover: Semiotic dynamics and collaborative tagging

Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero

PNAS 2007;104;1461-1464; originally published online Jan 23, 2007;
doi:10.1073/pnas.0610487104

Semiotic dynamics and collaborative tagging

Ciro Cattuto*†, Vittorio Loreto†‡, and Luciano Pietronero†

*Museo Storico della Fisica e Centro Studi e Ricerche Enrico Fermi, Compendio Viminale, 00184 Rome, Italy; and †Dipartimento di Fisica, Universita ''La Sapienza,'' Piazzale Aldo Moro 2, 00185 Rome, Italy

Communicated by Nicola Cabibbo, University of Rome, Rome, Italy, November 30, 2006 (received for review June 20, 2006)

Collaborative tagging has been quickly gaining ground because of its ability to recruit the activity of web users into effectively organizing and sharing vast amounts of information. Here we collect data from a popular system and investigate the statistical properties of tag cooccurrence. We introduce a stochastic model of user behavior embodying two main aspects of collaborative tagging: (*i*) a frequency-bias mechanism related to the idea that users are exposed to each other's tagging activity; (*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 modeling is able to account quantitatively for the observed experimental features with a surprisingly high accuracy. This points in the direction of a universal behavior of users who, despite the complexity of their own cognitive processes and the uncoordinated and selfish nature of their tagging activity, appear to follow simple activity patterns.

online social communities | statistical physics | social bookmarking | information dynamics

Recently, a new paradigm has been quickly gaining ground on the World Wide Web: collaborative tagging (1–3). In web applications like *Del.icio.us* (http://del.icio.us), *Flickr* (www.flickr.com), *CiteULike* (www.citeulike.org), and *Connotea* (www.connotea.org), users manage, share, and browse collections of online resources by enriching them with semantically meaningful information in the form of freely chosen text labels (tags). The paradigm of collaborative tagging has been successfully deployed in web applications designed to organize and share diverse online resources such as bookmarks, digital photographs, academic papers, music, and more. Web users interact with a collaborative tagging system by posting content (resources) into the system, and associating text strings (tags) with that content, as shown in Fig. 1. At the global level, the set of tags, although determined with no explicit coordination, evolves in time and leads toward patterns of terminology usage that are shared by the entire user community. Hence, one observes the emergence of a loose categorization system 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 (4–6), a new field that studies how populations of humans or agents can establish and share semiotic systems, typically driven by their use in communication. Indeed, the emergence of a folksonomy exhibits dynamical aspects also observed in human languages (7, 8), such as the crystallization of naming conventions, competition between terms, takeovers by neologisms, and more.

In the following, we adopt the point of view of complex systems science and try to understand how the "microscopic" tagging activity of users causes the emergence of the high-level features we observe for the ensuing folksonomy. We ground our analysis on actual tagging data extracted from *Del.icio.us* and *Connotea* and use standard statistical tools to gain insights into the underlying tagging dynamics. Based on this, we introduce a simple stochastic model for the tagging behavior of an "average" user, and show that such a model, despite its simplicity, is able to reproduce extremely well some of the observed properties. We close giving an interpretation of the model parameters and pointing out directions for future research.

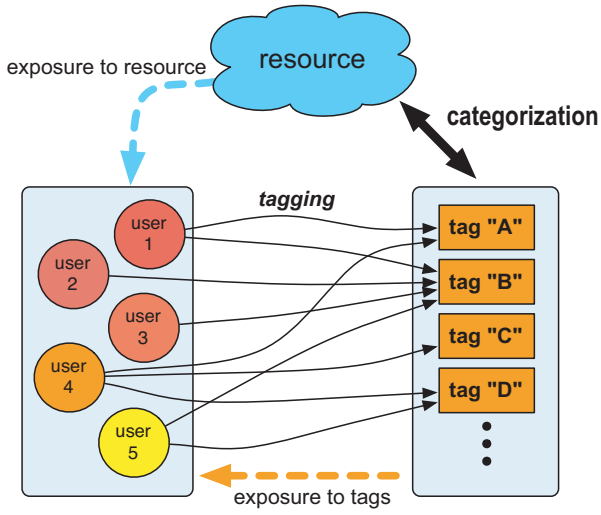

Fig. 1. Collaborative tagging. Schematic depiction of the collaborative tagging process: web users are exposed to a resource and freely associate tags with it. Their interaction with the system also exposes them to tags previously entered by themselves and by other users. The aggregated activity of users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

Results

Data Analysis. The activity of users interacting with a collaborative tagging system consists of either navigating the existing body of resources by using tags, or adding new resources to the system. To add a new resource to 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. Tagging events build a tripartite graph (with partitions corresponding to users, resources and tags, respectively) and such a graph, commonly referred to as folksonomy, can be subsequently used as a navigation aid in browsing tagged information (see Fig. 2). Usually, a post contains also a temporal marker indicating the physical time of the tagging event, so that temporal ordering can be preserved in storing and retrieving posts. Here we analyze data from *Del.icio.us* and *Connotea* and investigate the statistical properties of tag association. Specifically, we select a semantic context by extracting the resources associated with a given tag X and study the statistical distribution of tags cooccurring with X

Author contributions: C.C., V.L., and L.P. designed research; C.C. and V.L. performed research; C.C. and V.L. contributed new reagents/analytic tools; C.C. analyzed data; and C.C., V.L., and L.P. wrote the paper.

The authors declare no conflict of interest.

‡To whom correspondence should be addressed. E-mail: vittorio.loreto@roma1.infn.it.

This article contains supporting information online at www.pnas.org/cgi/content/full/0610487104/DC1.

ajax apple art article blog blogging blogs books browser business code community computer cool css culture database design development diy download firefox flash flickr framework free freeware fun gallery games google graphics hacks howto html humor illustration images imported inspiration internet java javascript library linux live mac media microsoft mp3 music network news online opensource osx photo photography photos photoshop php portfolio productivity programming python rails reference research resources rss ruby search security social software tech technology tips tool tools toread tutorial tutorials ubuntu video web web2.0 webdesign webdev windows xml

Fig. 2. Example of a tag-cloud. A tag-cloud is a common way to visualize tags belonging to a collaborative tagging system. Here, the font size of each tag is proportional to the logarithm of its frequency of appearance within the folksonomy.

(see Table 1). Fig. 3 graphically illustrates the associations between tags and posts, and Fig. 4 reports the frequency-rank distributions for the tags cooccurring with a few selected ones. The high-rank tail of the experimental curves displays a power-law behavior, signature of an emergent hierarchical structure, corresponding to a generalized Zipf's law (9) with an exponent between 1 and 2. Because power laws are the standard signature of self-organization and of human activity (10–12), the presence of a power-law tail is not surprising. The observed value of the exponent, however, deserves further investigation because the mechanisms usually invoked to explain Zipf's law and its generalizations (13) do not look very realistic for the case at hand, and a mechanism grounded on experimental data should be sought.

Moreover, the low-rank part of the frequency-rank curves exhibits a flattening typically not observed in systems strictly obeying Zipf's law. Several aspects of the underlying complex dynamics may be responsible for this feature: on the one hand, this behavior points to the existence of semantically equivalent and possibly competing high-frequency tags (e.g., *blog* and *blogs*). More importantly, this flattening behavior may be ascribed to an underlying hierarchical organization of tags cooccurring with the one we single out: more general tags (semantically speaking) will tend to cooccur with a larger number of other tags. In this scenario, we expect a shallower behavior for tags cooccurring with generic tags (e.g., *blog*) and a steeper behavior for semantically narrow tags (e.g., *ajax*, see also Fig. 3). To better probe the validity of this interpretation, we investigate the cooccurrence 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 supporting information (SI)] point in the direction of a nontrivial 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.

Table 1. Statistics of the data sets used for the cooccurrence analysis

Tag	No. of posts	No. of tags	No. of distinct tags	No. of resources
Blog	37,974	24,171	10,617	16,990
Ajax	33,140	108,181	4,141	2,995
Xml	24,249	108,013	6,035	7,364
H5N1	981	5,185	241	969

For each tag, we report the number of posts marked with that tag, the number of total and distinct tags cooccurring with it, and the corresponding number of resources. The data were retrieved during May of 2005.

A Yule–Simon Model with Long-Term Memory. We now aim at gaining a deeper insight into the phenomenology reported above. 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 (14, 15) by enhancing it with a fat-tailed memory kernel. The original model can be described as the construction of a text from scratch. At each discrete time step one word is appended to the text: with probability p the appended word is a new word, never occurred before, whereas with probability $1 - p$ one word is copied from the existing text, choosing it with a probability proportional to its current frequency of occurrence. This simple process yields frequency-rank distributions that display a power-law tail with exponent $\alpha = 1 - p$, lower than the exponents we observe in actual data. This happens because the Yule–Simon process has no notion of "aging," i.e., all positions within the text are regarded as identical.

In our construction, we moved from the observation that actual users are exposed in principle to all of 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 resource, is far from being uniform in time (see also refs. 16 and 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

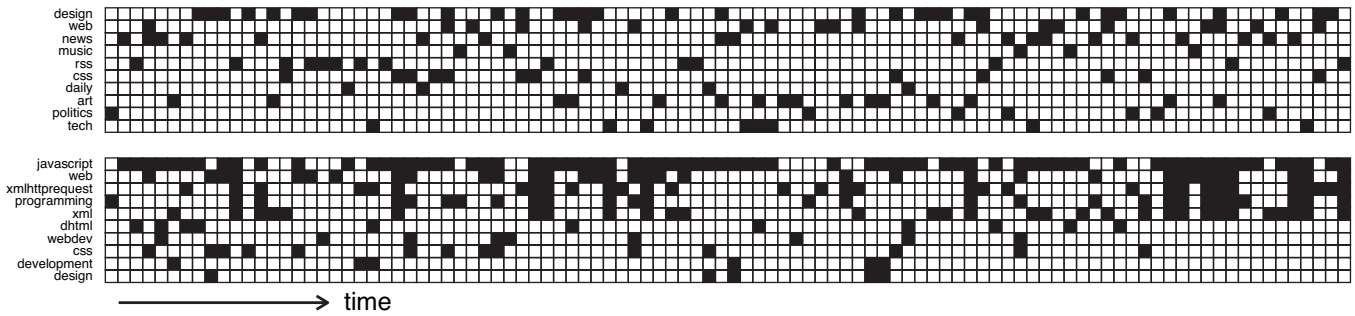

Fig. 3. Tagging activity: a time-ordered sequence of tagging events is graphically rendered by marking the tags cooccurring with *blog* (*Upper*) or *ajax* (*Lower*) in an experimental sequence of posts on *del.icio.us*. In each panel, columns represent single tagging events (posts) and rows correspond to the 10 most frequent tags cooccurring with either *blog* (*Upper*) or *ajax* (*Lower*). One hundred tagging events are shown in each panel, temporally ordered from left to right. Only posts 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, whereas an empty cell indicates its absence. A qualitative difference between *blog* (*Upper*) and *ajax* (*Lower*) is clearly visible, where a higher density at low-rank tags characterizes the semantically narrower *ajax* term. This corresponds to the steeper low-rank behavior observed in the frequency-rank plot for *ajax* (Fig. 4).

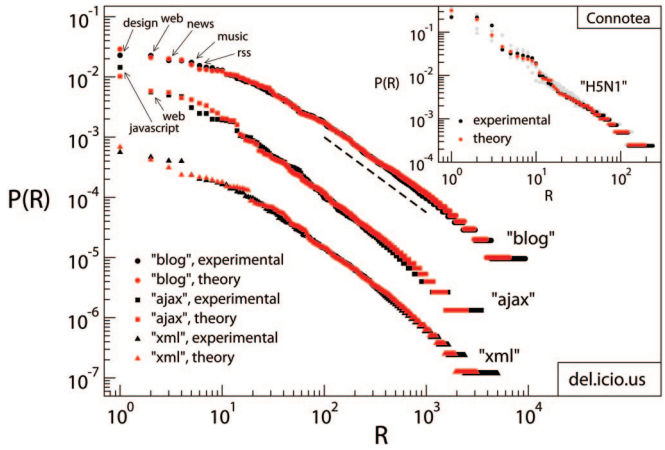

Fig. 4. Frequency-rank plots for tags cooccurring with a selected tag: experimental data (black symbols) are shown for *Del.icio.us* (circles for tags cooccurring with the popular tag *blog*, squares for *ajax*, and triangles for *xml*) and *Connotea* (*Inset*, black circles for the *H5N1* tag). For the sake of clarity, the curves for *ajax* and *xml* are shifted down by one and two decades, respectively. Details about the experimental data sets 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 shallower behavior for low ranks. To make contact with Fig. 3, some of the highest-frequency tags cooccurring with *blog* and *ajax* are explicitly indicated with arrows. Red symbols are theoretical data obtained by computer simulation of the stochastic process 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, giving approximately $p = 0.06$, $\tau = 100$, and $n_0 = 100$ for *blog*, $p = 0.03$, $\tau = 20$, and $n_0 = 50$ for *ajax*, and $p = 0.034$, $\tau = 40$, and $n_0 = 110$ for *xml*. (*Inset*) *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 $p = 0.05$, $\tau = 120$, and $n_0 = 7$ (red circles), demonstrating that our model also applies to the early stages of development of a folksonomy. Gray circles correspond to different realizations of the simulated dynamics.

findings about human activities (12) support the idea that the access pattern to the past of the system should be fat-tailed, suggesting a power-law memory kernel.

We tested this hypothesis with real data extracted from *Del.icio.us*: Fig. 5 shows the temporal autocorrelation function for the sequence of tags cooccurring with *blog*. Such a sequence is constructed by consecutively appending the tags associated with each post, respecting the temporal order of posts. Correlations are computed inside three consecutive windows of length T, starting at different times t_w,

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

where $\delta(\text{tag}(t + \Delta t), \text{tag}(t))$ is the usual Kronecker delta function, taking the value 1 when the same tag occurs at times t and $t + \Delta t$. From Fig. 5, it is apparent that the correlation function is nonstationary over time. Moreover, for each value of the initial time t_w a power-law behavior is observed: $C(\Delta t, t_w) = a(t_w)/(\Delta t + \delta(t_w)) + c(t_w)$, where $a(t_w)$ is a time-dependent normalization factor and $\delta(t_w)$ is a phenomenological time scale, slowly increasing with the "age" t_w of the system. $c(t_w)$ is the correlation that one would expect in a random sequence of tags distributed according to the frequency-rank distribution $P_{T,tw}(R)$ pertaining to the relevant data window. Denoting by $R = R_{max}(T, t_w)$ the number of distinct tags occurring in the window $[t_w, t_w + T]$, we have $c(t_w) = \Sigma_{R=1}^{R=R_{max}}(T,t_w)\, P^2_{T,t_w}(R)$.

Our modification of the Yule–Simon model thus consists in weighting the probability of choosing an existing word (tag) ac-

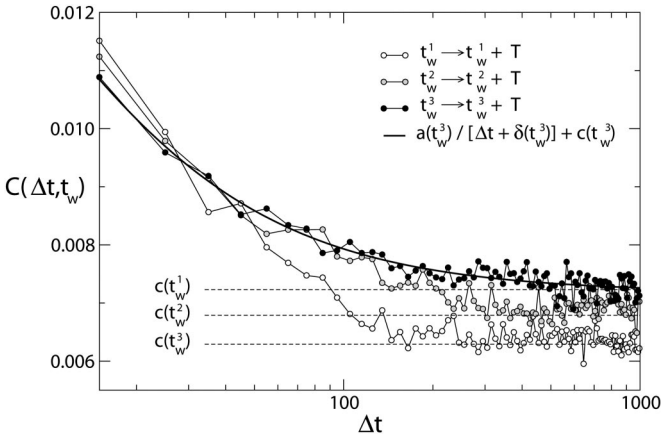

Fig. 5. Tag–tag correlation functions and nonstationarity. The tag-tag correlation function $C(\Delta t, t_w)$ is computed over three consecutive and equally long ($T = 30{,}000$ tags each) subsets of the *blog* data set, starting respectively at positions $t_w^1 = 10{,}000$, $t_w^2 = 40{,}000$, and $t_w^3 = 70{,}000$ within the collected sequence. Short-range correlations are clearly visible, slowly decaying toward a long-range plateau value. The nonstationary 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 sequence containing a finite number of tags: on using the appropriate ranked distribution of tag frequencies 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 measured plateau of the correlation functions. The thick line is a fit to the fat-tailed memory kernel described in the text.

cording to a power-law kernel. 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.

In summary, our model of users' behavior 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 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, whereas with probability $1 - p$, one word is copied from the existing text, going back in

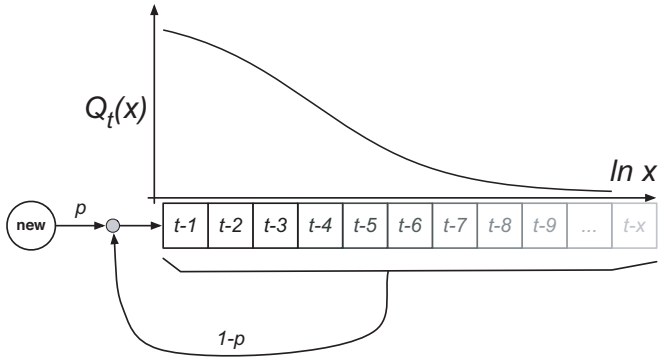

Fig. 6. A Yule–Simon's process with long-term memory. A synthetic stream of tags is generated by iterating the following step: with probability p a new tag is created and appended to the stream, whereas with probability $1 - p$, a tag is copied from the past of the stream and appended to it. The probability of selecting a tag located x steps into the past is given by the long-range memory kernel $Q_t(x)$, which provides a fat-tailed access to the past of the stream.

time by x steps with a probability $Q_t(x)$ that decays as a power law, $Q_t(x) = a(t)/(x + \tau)$ (see Fig. 6). $a(t)$ is a normalization factor, and τ is a characteristic time scale over which recently added words have comparable probabilities. Fig. 4 shows the excellent agreement between the experimental data and the numerical predictions of our Yule–Simon model with long-term memory. Our model, unsurprisingly, also reproduces the temporal correlation behavior observed in real data.

The interpretation of τ (similar to that of the δ parameter introduced above for tag-tag correlations) is related to the number of equivalent top-ranked tags perceived by users as semantically independent (see SI). In our model, in fact, the average user is exposed to a few roughly equivalent top-ranked tags and this is translated mathematically into a low-rank cutoff of the power law, i.e., the observed low-rank flattening.

Fitting the parameters of the model to match its predictions (obtained by computer simulation) against the experimental data, we obtain an excellent agreement for all of the frequency-rank curves we measured, as shown in Fig. 4. This is a clear indication that the behavior encoded in our simple model is able to capture some key features of the tagging activity. The parameter τ controls the number of top-ranked tags that are allowed to cooccur with comparable frequencies, so that it can be interpreted as a measure of the "semantic breadth" of a tag. This picture is consistent with the fact that the fitted value of τ obtained for *blog* (a rather generic tag) is larger than the one needed for *ajax* (a pretty specific tag). It is worth remarking that, despite the agreement between the experimental data and our model predictions, our simple modeling is an attempt toward the modeling of user behaviors, not meant to be neither unique or exclusive of other generative models (10, 11).

Discussion and 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. Collaborative tagging provides a precious opportunity to both analyze the emergence of shared conventions and inspire the design of large (human or artificial) agent systems. Here we report a statistical analysis of tagging activity in a popular social bookmarking system, and introduce a simple stochastic model of user behavior which is able to reproduce the measured cooccurrence properties to a surprisingly level of accuracy. Our results suggest that users of collaborative tagging systems share universal behaviors that, despite the intricacies of personal categorization, tagging procedures, and user interactions, appear to follow simple activity patterns. In addition to the findings reported and discussed in this paper, our approach constitutes a starting point upon which more cognitively informed studies can be based, with the final goal of understanding and engineering the semiotic dynamics of online social systems.

Experimental Data

Our analysis focuses on *Del.icio.us*, for several reasons: (*i*) it was the first system to deploy the ideas and technologies of collaborative tagging, and the paradigmatic character it acquired makes it a 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 (as Thomas Vander Wal argued, www.personalinfocloud.com/2005/02/explaining_and_.html), and 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 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 available 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 tag-centric view of the system; i.e., we investigate the evolving relationship between a given tag and the set of tags that cooccur 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 postprocessing. *Del.icio.us* allows the user to browse its content by tag: our client requests the web page associated with the tag under study and uses an HTML parser to extract the post information (user, resource, tags, time stamp) from the returned HTML code. Fig. 3 graphically depicts the raw data we gather, for the case of two popular tags on *Del.icio.us*. Table 1 describes the data sets we used for the present analysis.

We thank A. Baronchelli, A. Baldassarri, and V. Servedio for many interesting discussions and suggestions. This research has been partly supported by the TAGora project funded by the Future and Emerging Technologies program (IST-FET) of the European Commission under the European Union RD contract IST-034721.

1. Mathes A (2004) *Computer Mediated Commun*, LIS590CMC.
2. Hammond T, Hannay T, Lund B, Scott J (2005) *D-Lib Magazine* 11:www.dlib.org/dlib/april05/hammond/04hammond.html.
3. Golder S, Huberman BA (2006) *J Information Sci* 32:198–208.
4. Steels L, Kaplan F (1999) *Lect Notes Artificial Intell* 1674:679–688.
5. Steels L (2006) *IEEE Intelligent Syst* 21:32–38.
6. Ke J, Minett JW, Ching-Pong A, Wang WS-Y (2002) *Complexity* 7:41–54.
7. Nowak MA, Komarova NL, Niyogy P (2002) *Nature* 417:611–617.
8. Kirby S (2002) *Artificial Life* 8:182–215.
9. Zipf GK (1949) *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Cambridge, MA).
10. Mitzenmacher M (2004) *Internet Math* 1:226–251.
11. Newman MEJ (2005) *Contemporary Phys* 46:323–351.
12. Barabasi A-L (2005) *Nature* 435:207–211.
13. Ferrer Cancho R, Servedio VDP (2005) *Glottometrics* 11:1.
14. Yule GU (1925) *Philos Trans R Soc London B* 213:21–87.
15. Simon HA (1955) *Biometrika* 42:425–440.
16. Zanette DH, Montemurro MA (2005) *J Quant Ling* 12:29–40.
17. Dorogovtsev SN, Mendes JFF (2000) *Phys Rev E* 62:1842–1845.
18. Anderson JR (2000) *Cognitive Psychology and its Implications* (Worth, New York), 5th Ed.