Supporting Information

Cattuto et al. 10.1073/pnas.0901136106

SI Text

Data. The basic unit of information in a collaborative tagging system is a post, i.e., a triple of the form (user, resource, {tags}). In del.icio.us and in BibSonomy (as well as in many other collaborative tagging systems) a post is complemented with a time stamp that records the physical time of the tagging event. Therefore, the temporal ordering of posts can be preserved and the dynamical evolution of the system over time can be reconstructed and investigated. In the following, we describe the 2 datasets that were used for the experimental part of this study.

Data from del.icio.us. The del.icio.us dataset we used consists of $\approx 5 \times 10^6$ posts, comprising $\approx$650,000 users, 1.9×10^6 resources (bookmarks), and 2.5×10^6 distinct tags. It covers almost 3 years of user activity, from early 2004 up to November 2006.

Data were collected in a 2-stage process. The first stage, started in mid-2005, was aimed at having a list of active del.icio.us users as large as possible. The database of users was initially populated with a recursive crawl from the home page of del.icio.us (following resource, tag, and user links). At the same time, the system was continuously monitored to detect new users or the activity of existing users that were not in our database yet. This was achieved by monitoring the home page of del.icio.us for fresh posts and, later, by monitoring the really simple syndication (RSS) feeds provided by del.icio.us. The second stage consisted of a distributed crawl of the system on a per-user basis. During the period 10–24 November 2006, the user pages of all of the users present in our database were crawled and parsed to extract posts. Each downloaded post contains an anonymized user ID, a resource URL (together with a unique resource ID), tags, and a time stamp. Overall, 667,128 user pages of the del.icio.us community were crawled, for a total of 18,782,132 resources, 2,454,546 distinct tags, and 140,333,714 tag assignments (triples).

The data were subsequently postprocessed for the present study. We discarded all posts containing no tags ($\approx$7% of the total). Because del.icio.us is case-preserving but not case sensitive, we ignored capitalization in tag comparison and counted all different capitalizations of a given tag as instances of the same lowercase tag. The time stamp of each post was used to establish post ordering and determine the temporal evolution of the system. Posts with invalid timestamps, i.e., times set in the future or before del.icio.us started operating, were discarded as well ($<$0.5% of the total).

Except for the normalization of character case, no lexical normalization was applied to tags during postprocessing. The notion of identity of tags is identified with the notion of identity of their string representation.

Data from BibSonomy. BibSonomy (3) is a smaller system than del.icio.us, but it was designed keeping data sharing in mind. Because of this, there is no need to crawl BibSonomy by downloading HTML pages and parsing them. Direct access to post data in structured form is available by using the BibSonomy application programming interface (API), www.bibsonomy.org/help/doc/api.html. Moreover, the BibSonomy team periodically releases snapshot datasets of the full system and makes them available to the research community. For the present work, we used the dataset released on January 2008 (www.kde.cs.uni-kassel.de/bibsonomy/dumps/2007–12-31.tgz).

BibSonomy allows 2 different types of resources: bookmarks (i.e., URLs of web pages, similar to del.icio.us) and BibTeX entries. To make contact with the analysis done for del.icio.us, we restricted the dataset to the posts involving bookmark resources only. The resulting dataset we used comprises 1,400 users, 127,115 resources, 37,966 distinct tags, and 503,928 tag assignments (triples). The data from BibSonomy was postprocessed in the same way as the data from del.icio.us.

Although the BibSonomy dataset is much smaller than the del.icio.us dataset, it is a valuable one: direct access to BibSonomy's database guarantees that the BibSonomy dataset is free from biases due to the data collection procedure. This is important because it allows us to show that the investigated features of the data are robust across different systems and not only established in a case where biases due to data collection could be possible.

Network Characterization. In this section, we recall, in more details than in the main text, the definition of various quantities customarily used to characterize complex networks.

Each node i of a network is first characterized by its degree k_i (number of links). In a weighted network, each link i–j carries moreover a weight w_{ij}, and an important measure of a node's importance is given by its strength s_i, defined as

$$s_i = \sum_{j \in \mathcal{V}(i)} w_{ij}, \qquad \textbf{[s1]}$$

where the sum runs over the set $\mathcal{V}(i)$ of i's neighbors. This quantity naturally generalizes the degree by measuring the strength of vertices in terms of the total weight of their connections. A first characterization of a network's properties is obtained by the statistical distributions of the nodes' degree and strength, and the distributions of link weights: $P(k)$, $P(s)$, $P(w)$. Moreover, it is customary to investigate the average strength $s(k)$ of vertices with degree k:

$$s(k) = \frac{1}{N_k} \sum_i \delta_{k,k_i} s_i, \qquad \textbf{[s2]}$$

where N_k is the number of nodes of degree k and $\delta_{k,ki}$ is the Kronecker symbol, taking value 1 if $k_i = k$ and 0 otherwise.

To shed light on a network's topological correlations, 2 main quantities are customarily measured. The clustering coefficient c_i of a node i measures the local cohesiveness around this node (5). It is defined as the ratio of the number of links between the k_i neighbors of i and the maximum number of such links, $k_i(k_i - 1)/2$. The clustering spectrum measures the average clustering coefficient of nodes of degree k, according to

$$C(k) = \frac{1}{N_k} \sum_i \delta_{k,k_i} c_i. \qquad \textbf{[s3]}$$

Moreover, correlations between the degrees of neighboring nodes are conveniently measured by the average nearest neighbors degree of a vertex i, $k_{nn,i} = 1/k_i \Sigma_{j \in \mathcal{V}(i)} k_j$, and the average degree of the nearest neighbors, $k_{nn}(k)$, for vertices of degree k (6)

$$k_{nn}(k) = \frac{1}{N_k} \sum_i \delta_{k,k_i} k_{nn,i}. \qquad \textbf{[s4]}$$

In the absence of correlations between degrees of neighboring vertices, $k_{nn}(k)$ is a constant. An increasing behavior of $k_{nn}(k)$ corresponds to the fact that vertices with high degree have a larger probability to be connected with large-degree vertices (assortative mixing). On the contrary, a decreasing behavior of $k_{nn}(k)$ defines a disassortative mixing, in the sense that high-degree vertices have a majority of neighbors with low degree, whereas the opposite holds for low-degree vertices (7).

These quantities have been generalized to weighted networks (8). The weighted clustering coefficient of a node i,

$$c^w(i) = \frac{1}{s_i(k_i-1)} \sum_{j,h\in\mathcal{V}(i), j\in\mathcal{V}(h)} \frac{(w_{ij}+w_{ih})}{2}, \qquad \textbf{[s5]}$$

considers not only the presence of triangles in the neighborhood of i, but also their total relative edge weights with respect to the vertex's strength. The weighted clustering spectrum $C^w(k)$ is the weighted clustering coefficient averaged over all vertices with degree k. If the weighted clustering is larger than the clustering coefficient, triangles are more likely formed by edges with larger weights and thus carry a strong signification for the network.

Similarly, the weighted average nearest-neighbor degree is defined as

$$k^w_{nn,i} = \frac{1}{s_i} \sum_{j\in\mathcal{V}(i)} w_{ij}k_j. \qquad \textbf{[s6]}$$

This quantity performs a local weighted average of the nearest neighbor degrees according to the normalized weight of the connecting edges, w_{ij}/s_i, measuring the effective affinity to connect with high- or low-degree neighbors according to the magnitude of the actual interactions. The average of $k^w_{nn,i}$ over all vertices with degree k, $k^w_{nn}(k)$, marks the weighted assortative or disassortative properties considering the actual interactions among the system's elements.

Robustness of Data Characteristics. To check the robustness of the data analysis, we have considered several tags from the 2 analyzed datasets, del.icio.us and BibSonomy. For each dataset, the tag number corresponds to its popularity ranking, where the popularity is simply given by the number of posts in which it appears. For each tag, we perform the same analysis as in the main text, considering only the posts that contain this tag. We measure the growth of the vocabulary associated with this tag as a function of the number of posts containing it. Typical sublinear power-law growths are obtained, as shown in Fig. S1.

We also build the cooccurrence network of each tag, and characterize these networks using the measures detailed in *Data* above. As Figs. S2 and S3 show, all networks display the same qualitative properties, both at the topological level and for the weighted quantities: broad distributions are observed for the degrees, weights and strengths; the average strength of nodes of degree k is linear at small k and has a superlinear growth at large k; nontrivial correlations are observed, with a disassortative trend, and strong clustering; at large degree, the weighted correlations are stronger than the topological ones, showing that large weights are preferentially connecting nodes with large degree.

Heaps' Law and Zipf's Law. In this short section, we briefly recall how Heap's law can be simply derived from Zipf's law, as presented in ref. 1. Let us consider N_{d} distinct tags corresponding to a total of N annotations. Each tag i has been used n_i times, so that $\sum_{i=1}^{N_{\mathrm{d}}} n_i = N$. Zipf's law states that, if the tags are ranked in decreasing order of frequency, the ith has frequency $n_i = A/i^{\beta}$, where $A = N/\sum_{i=1}^{N_{\mathrm{d}}} i^{-\beta}$ is a normalizing factor. For $\beta > 1$, A is thus of order N for large N and N_{d}. Let us assume that the least-frequent word appears only once (or a number of times of order 1); then $n_{N_{\mathrm{d}}} = \mathcal{O}(1)$, which translates into $\mathcal{O}(1) = A/N_{\mathrm{d}}^{\beta}$ and thus $N = \mathcal{O}(N_{\mathrm{d}}^{\beta})$. Therefore, the number of different tags N_{d} scales as N^{α}, which is the Heap's law, with $\alpha = 1/\beta$.

Random Walks: Number of Distinct Nodes Visited. As explained in the main text, we consider an exploration process on a network which is taken as a sketch of a semantic space, each node corresponding to a tag. Starting from a fixed initial node i_0 (chosen at random), we perform a random walk of a certain length. In this picture, a random walk corresponds to a post (of the same length) in a tagging system: the cooccurrence of 2 tags is represented by the fact that the 2 corresponding nodes are visited in the same random walk. The growth of the number of different tags cooccurring with an initial fixed tag, when the number of posts increases, corresponds then to the growth of the number of distinct visited sites as a function of the number n_{RW} of random walks performed.

Analytics. Let us first consider random walks of fixed length l starting from a given node i_0. We denote by p_i the probability for each of these random walks to visit node i. The probability that i has not been visited after n_{RW} random walks is then simply

$$\mathrm{Prob(i\ not\ visited)} = (1-p_i)^{n_{\mathrm{RW}}}, \qquad \textbf{[s7]}$$

because the random walks are independent stochastic processes, and the probability that i has been visited at least once reads

$$\mathrm{Prob(i\ visited)} = 1 - \mathrm{Prob(i\ not\ visited)} = 1-(1-p_i)^{n_{\mathrm{RW}}}. \qquad \textbf{[s8]}$$

The average number of distinct nodes visited after n_{RW} random walks is then given, without any assumption on the network's structure, by

$$N_{\mathrm{distinct}} = \sum_i \left(1-(1-p_i)^{n_{\mathrm{RW}}}\right), \qquad \textbf{[s9]}$$

where the sum runs over all nodes of the network.

Although this exact expression is not yet really informative, it is possible to go further under some simple assumptions [we also note that analytical results are available in the case of random walks performed either on lattices or on fractal substrates (2)]. Because all of the random walks start from the same origin i_0, it is useful to divide the network into successive "rings" (4), each ring of label l being formed by the nodes at distance l from i_0. The ring $l = 1$ is formed by the neighbors of i_0, the ring $l = 2$ by the neighbors' neighbors that are not part of ring 1, and so forth. We denote by N_l the number of nodes in ring l. We now make the assumption that all N_l nodes at distance l have the same probability to be reached by a random walk starting from i_0 (which is the sole element of ring 0). This is rigorously true for example for a tree with constant coordination number, and more generally will hold approximately in homogeneous networks, whereas stronger deviations are expected in heterogeneous networks. Let us assume moreover that the random walk of length $l_{\max}$ consists, at each step, of moving from one ring l to the next ring $l+1$. This is once again rigorously true for a self-avoiding random-walk on a tree and can be expected to hold approximately if N_l grows fast enough with l: the probability to go from ring l to ring $l+1$ is then larger than to go back to ring $l-1$ or to stay within ring l. For each random walk of length $l_{\max}$, we then have $p_i = 1/N_l$ for each node i in ring $l \le l_{\max}$, and after n_{RW} walks, the average number of distinct visited nodes reads

$$N_{\mathrm{distinct}} = \sum_{l=0}^{l_{\max}} N_l\left(1-(1-1/N_l)^{n_{\mathrm{RW}}}\right). \qquad \textbf{[s10]}$$

The expression (**s10**) lends itself to numerical investigation using various forms for the growth of N_l as a function of l. We obtain that, as n_{RW} increases, N_{distinct} increases, with an approximate power-law form, and saturates as $n_{\mathrm{RW}} \to \infty$ at the total number of reachable nodes $\sum_{l=0}^{l_{max}} N_l$. Moreover, the increase at low n_{RW} is sublinear if N_l grows fast enough with l (at least $\approx l^2$), and is closer to linear if l_{max} increases.

Let us now consider, under the same assumptions, that the successive random walks have randomly distributed lengths according to a certain $P(l)$. Each ring l, on average, is then reached by a random walk $n_{RW} \times \sum_{l' \geq l} P(l') \equiv n_{RW} P_>(l)$ times, so that we have approximately

$$N_{distinct} = \sum_{l=0}^{\infty} N_l (1 - (1 - 1/N_l)^{n_{RW} P_>(l)}), \qquad \textbf{[s11]}$$

where the sum (provided it converges) now runs over all possible lengths.

If $P(l)$ is narrowly distributed around an average value, the form (**s11**) will not differ very much from the case of fixed length given by Eq. **s10**. Conversely, for a broad $P(l)$, longer random walks will occur as n_{RW} increases and the tail of $P(l)$ is sampled, allowing visits to nodes situated further from i_0 and avoiding the saturation effect observed for random walks of fixed length.

In some particular cases, a further analytical insight into the form of $N_{distinct}(n_{RW})$ can be obtained:

- assume that $N_l \approx l^a$, and that $P(l)$ is power-law distributed ($P(l) \approx 1/l^b$). Then $N_{distinct}(n_{RW}) \approx \sum_{l=0}^{\infty} l^a (1 - \exp(-n_{RW}/(c\, l^{a+b-1})))$, where c is a constant. The terms in the sum become negligible for l larger than $n_{RW}{}^{1/(a+b-1)}$, whereas they are close to l^a for smaller values of l. The sum therefore behaves as

$$N_{distinct} \sim n_{RW}^{(a+1)/(a+b-1)}, \qquad \textbf{[s12]}$$

  i.e., a power-law. For instance, for $b = 3$ we obtain a sublinear power-law growth with exponent $(a + 1)/(a + 2)$, i.e., 2/3 for $a = 1$, or 3/4 for $a = 2$.
- assume that $N_l \approx z^l$, which corresponds to a tree in which each node has $z + 1$ neighbors, and $P(l) \approx 1/l^b$. Then $N_{distinct}(n_{RW}) \approx \sum_{l=0}^{\infty} z^l (1 - \exp(-n_{RW}/(cz^l l^{b-1})))$. As in the previous case, the terms in the sum become negligible for $l > (\log(n_{RW}) - (b - 1)\log(\log(n_{RW}/\log(z))))/\log(z)$, while they are close to z^l for smaller l. Thus the sum behaves as

$$N_{distinct} \sim n_{RW}/(\log(n_{RW}))^{b-1}, \qquad \textbf{[s13]}$$

  i.e., we obtain a linear behavior with logarithmic corrections, which is known to be very similar to sublinear power-law behaviors.

We have thus shown analytically, under reasonable assumptions, that performing fixed length random walks starting from the same node yields a growth of the number of distinct visited sites (representing the vocabulary size) as a function of the number of random walks (representing posts), which is sublinear with a saturation effect, and that broad distributions of the walks lengths lead to sublinear growths of the vocabulary, and avoid the saturation effect.

Numerics. Model networks considered. We have considered networks with very different topologies to test the robustness of our approach with respect to the structure of the underlying network.

The Watts–Strogatz model has been put forward in ref. 5 as an example of network with large transitivity and at the same time small-world properties, i.e., short distances between nodes. The construction procedure starts with a ring of N vertices in which each vertex is symmetrically connected to its $2m$ nearest neighbors (m vertices clockwise and counterclockwise). Then, for every vertex, each edge connected to a clockwise neighbor is rewired with probability p, and preserved with probability $1 - p$. The rewiring connects the edge's endpoint to a randomly chosen vertex, avoiding self-connections, and thus creating shortcuts between distant parts of the ring. For $1/N \ll p \ll 1$, a network with a large number of triangles (due to the initial ring structure) and small diameter (thanks to the shortcuts) is obtained. The degree distribution is homogeneous, i.e., peaked around its average value.

The random scale-free network obtained from the uncorrelated configuration model (9), in contrast, has a broad degree distribution $P(k) \approx k^{-\gamma}$ (we have used $\gamma = 2.8$ and $\gamma = 2.3$) and a low clustering coefficient.

We have also considered the homogeneous Erdös–Rényi random graph, in which nodes are linked with a uniform probability. In this case, small diameter and small clustering coefficient are obtained, and the degree distribution is homogeneous. Moreover, a model of strongly clustered scale-free network (10) has also been used.

Results. We have performed numerical simulations on the model networks presented above, using random walks of fixed length, or of randomly chosen lengths extracted from a given distribution $P(l)$. Fig. S4 shows the number of distinct visited nodes as a function of the number of random walks performed on the Watts–Strogatz network and on the random scale-free network. Similar results are obtained if the underlying network is an Erdös–Rényi random graph or a strongly clustered scale-free network. For random walks of fixed length, a sublinear behavior is observed, followed by a saturation effect. For broadly distributed lengths, on the other hand, the sublinear power law-like behavior does not show any saturation, in agreement with the analytical insights obtained above.

Artificial Cooccurrence Networks. As explained in the main text, we use the random walks performed on the initial network to construct a synthetic cooccurrence network as follows: Each random walk of length l, which visits nodes $i_0,\ldots,i_l$, is transformed into a clique in the cooccurrence network, connecting the visited nodes. Each link i_a–i_b of the cooccurrence network has a weight corresponding to the number of times i_a and i_b were visited by the same random walk. Because, in our framework, a random walk is associated with a post, this construction mimics exactly the construction of the real tag co-occurrence network.

As a substrate for the random walks (i.e., as artificial semantic space), we have considered networks with markedly different properties, as explained in the previous section: with homogeneous and heterogeneous degree distributions, with small or large clustering. Moreover, we have considered random walks either of fixed length or of broadly distributed lengths. In each case, we have analyzed the artificial cooccurrence network obtained after various numbers of random walks. The characteristics investigated are described in *Network Characterization* above.

Fig. S5 shows the data corresponding to the cooccurrence networks for random walks of lengths distributed according to $P(l) \approx l^{-3}$, performed on a random scale-free network of $N = 5 \times 10^5$ nodes. The top figure corresponds to the cooccurrence network obtained after $n_{RW} = 10^4$, and the bottom figure to $n_{RW} = 5 \times 10^4$. As n_{RW} increases, the size of the cooccurrence network increases (here $\approx 2 \times 10^4$ nodes for $n_{RW} = 10^4$, and $\approx 7 \times 10^4$ nodes for $n_{RW} = 5 \times 10^4$), and the average and the maximal degrees increase as well. The statistical distributions and correlations keep the same qualitative shape, with broadly distributed degrees, strengths and weights, a disassortative trend, and weighted correlations which are systematically larger than the unweighted ones, showing how larger weights are observed between nodes of large degree and on triangles.

Fig. S6 displays data for a synthetic co-occurrence network built from random walks performed on the same Watts–Strogatz network as in the main text, but with fixed length $l = 5$. Although the main qualitative features of the cooccurrence network appear to be robust and can be observed even at fixed l, some differences are detectable. These differences are also clear in Fig. S7, which allows the direct comparison of the characteristics of cooccurrence networks built from random walks on a random graph (Erdös–Rényi network) for a fixed random walk length and for 2 different length distributions, namely $P(l) \approx l^{-4}$ and $P(l) \approx l^{-3}$. Clearly, all features are qualitatively robust, with a better quantitative agreement when $P(l)$ is closer to the experimental distribution of post lengths. For fixed l, in particular, we observe that the separation of the curves representing weighted and unweighted correlation properties occur at very small k, whereas the weighted and unweighted curves are equal over a certain range of degrees for broadly distributed random walk lengths. This is consistent with the results shown in Fig. 4 of the main text concerning the correlations between the weight of a link and the degrees of its extremities: Equality between weighted and unweighted quantities is indeed observed here for nodes with $s = k$, i.e., which are typically visited only once (by the rare walks with large l); when l is fixed instead, such rare events do not occur, and each node is visited several times, so that typically $s > k$.

An Even Finer Characterization: Structural Similarity of Nodes. We present here additional data and discussion on the distribution of tag similarities in real and synthetic cooccurrence networks. We recall that the similarity of 2 nodes i_1 and i_2 can be defined as

$$\mathrm{sim}(i_1, i_2) \equiv \sum_j \frac{w_{i_1 j} w_{i_2 j}}{\sqrt{\sum_\ell w_{i_1 \ell}^2 \sum_\ell w_{i_2 \ell}^2}}, \qquad \text{[s14]}$$

which is simply the scalar product of the vectors of normalized weights of nodes i_1 and i_2. This quantity measures the similarities between neighborhoods of nodes and is therefore a correlation of a higher order than the ones previously presented (clustering or assortativity properties).

In Fig. S8, we report the histograms of pairwise similarities between nodes in various cooccurrence networks. We first present the data for the tag cooccurrence network of various tags of del.icio.us and BibSonomy: A clearly skewed character of the distributions is observed, with a peak for low values of the similarities. The data are quite similar for the various tags investigated in BibSonomy, whereas the peak can be more or less broad for del.icio.us. Fig. S8 also displays the similarity histograms observed for the networks constructed according to the mechanism proposed in the main text, i.e., random walks of broadly distributed lengths, performed on underlying networks with various properties. A very similar behavior is observed, with a (more or less pronounced) peak at low values of the similarity. We emphasize that the shape observed for these histograms is quite specific of the cooccurrence networks studied here: we have indeed computed such histograms for a set of well known network models and real-world networks, obtaining typically a strong peak at zero similarity and qualitatively different tails.

For comparison, we also display the histogram of similarities for networks constructed from artificial tags with the following null model: (*i*) we start with a list of tags whose a priori frequencies follow a Zipf's law of exponent α; (*ii*) we construct artificial posts of length l distributed as l^{-3} by choosing at random l tags with probability proportional to their a priori frequency (the first tag of each post is always the same, because for the real data we are considering the posts containing all a given tag); (*iii*) we build the corresponding cooccurrence network. Clearly, this null model (which contains the Zipf's law as an ingredient, in contrast to our mechanism) does not contain any semantic correlations as the tags are used without any correlations. As shown in Fig. S8, the distribution of similarities is then indeed very different, and in particular it is much less skewed. Interestingly, the similarity distributions of some del.icio.us tags are in fact closer to the ones obtained for the null model than to the ones obtained from the random walk mechanisms. First of all, different tags have cooccurrence networks with distinct properties, corresponding to distinct semantic neighborhoods. It is therefore not surprising that the distributions of cosine similarities differ from one tag to another. In the framework of our model, this means that different models of semantic networks on which the random walks are performed are needed to describe the different tags' cooccurrence networks. In particular, we hypothesize that the tags whose cosine similarity distribution is less skewed correspond to tags whose "semantic neighborhood" has weaker correlations: The null model (in which no correlations are present) can then reproduce reasonably well their cooccurrence network's properties.

We emphasize however, on the one hand, that this null model contains as an ingredient the Zipf's law, in contrast with our mechanism and, on the other hand, that a precise fitting of the cooccurrence networks properties, although in principle doable by specifically fine-tuning the properties of the network on which random walks are performed, is beyond the purpose of our work.

1. Baeza-Yates R, Navarro G, (2000) Block addressing indices for approximate text retrieval. *J Am Soc Info Syst* 51:69–82.
2. Acedo L, Yuste SB (2002) Multiparticle random walks. *Recent Res Dev Stat Phys* 2:83.
3. Hotho A, Jäschke R, Schmitz C, Stumme G (2006) BibSonomy: A social bookmark and publication sharing system. *Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures, Aalborg, Denmark, July 2006*, de Moor A, Polovina S, Delugach H, eds (Aalborg Univ Press, Aalbord, Denmark).
4. Baronchelli A, Loreto V (2006) Ring structures and mean first passage time in networks. *Phys Rev E* 73:026103.
5. Watts DJ, Strogatz SH (1998) Collective dynamics of "small-world" networks. *Nature* 393:440–442.
6. Pastor-Satorras R, Vázquez A, Vespignani A (2001) Dynamical and correlation properties of the Internet. *Phys Rev Lett* 87:258701.
7. Newman MEJ (2002) Assortative mixing in networks. *Phys Rev Lett* 89:208701.
8. Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. *Proc Natl Acad Sci (USA)* 101:3747–3752.
9. Catanzaro M, Boguná M, Pastor-Satorras R (2005) Generation of uncorrelated random scale-free networks. *Phys Rev E* 71:027103.
10. Dorogovtsev SN, Mendes JFF, Samukhin AN (2001) Size-dependent degree distribution of a scale-free growing network. *Phys Rev E* 63:062101.

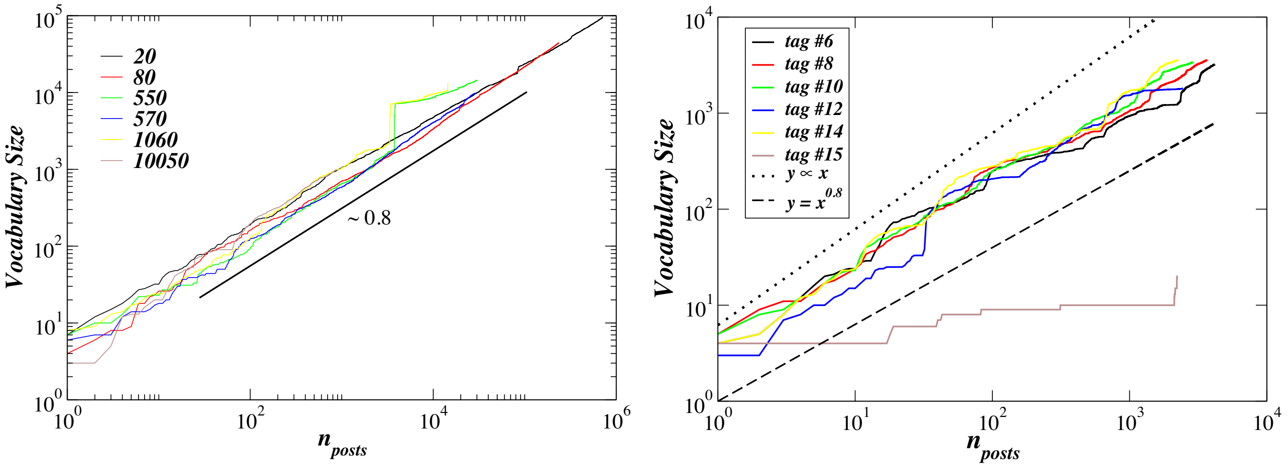


Fig. S1. Experimental data for the growth of the cooccurring tag vocabulary in del.icio.us (*Left*) and BibSonomy (*Right*), for a few tags of different popularities (the tag number corresponds to its popularity rank). For each tag, the time-ordered list of posts containing it was extracted, and the number of distinct tags cooccurring with the chosen tag is plotted here as a function of the number of posts (i.e., time increases along the *x* axis). The number associated with each tag corresponds to its global rank in terms of usage frequency. The dotted line corresponds to a linear growth, whereas the continuous line is a power-law growth with exponent 0.8. A sublinear power-law behavior across several orders of magnitude is visible for all tags and for both systems. The steps in the del.icio.us data (tags 550 and 1,060) are caused by spam posts containing a very large number of spurious tags. The curve for tag 15 in BibSonomy exhibits a different behavior: Direct inspection shows that the corresponding posts were all injected into BibSonomy at the same time by a single user, probably using an automated procedure. Thus, the evolution of the vocabulary cooccurring with this tag is neither social nor driven by human dynamics.

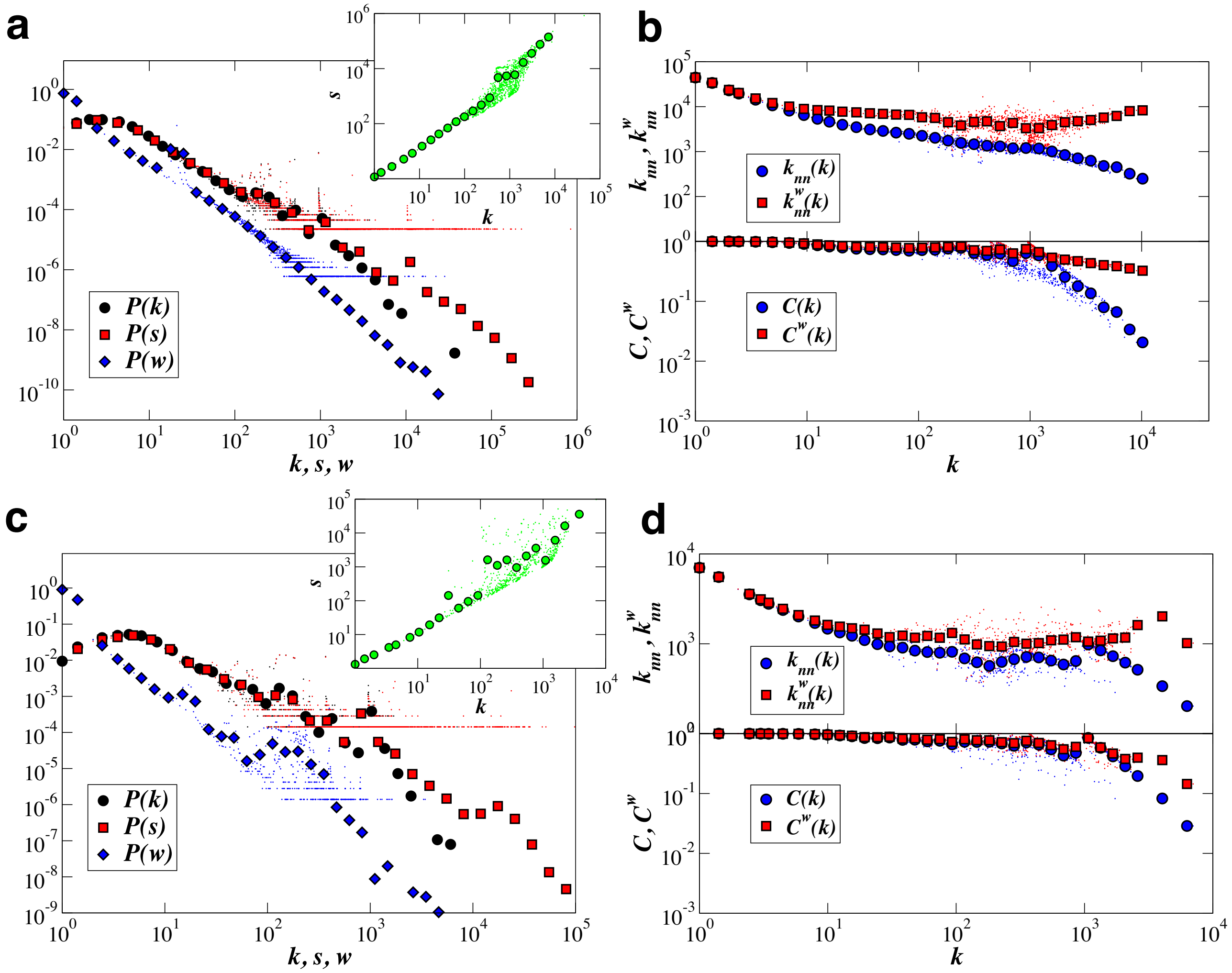


Fig. S2. Properties of the cooccurrence networks of the tags 80 (*a* and *b*) and 1,060 (*c* and *d*) in del.icio.us. The tag number corresponds to its popularity rank in the database analyzed. *a* and *c* show the broad distributions of degrees *k*, strengths *s*, and weights *w*. The *Insets* show the average strength of nodes of degree *k*, with a superlinear growth at large *k*. *b* and *d* display the weighted (k_{nn}^w) and unweighted (k_{nn}) average degree of nearest neighbors, and weighted (C^w) and unweighted (*C*) average clustering coefficients of nodes of degree *k*. In all figures, both raw data and logarithmically binned data are shown.

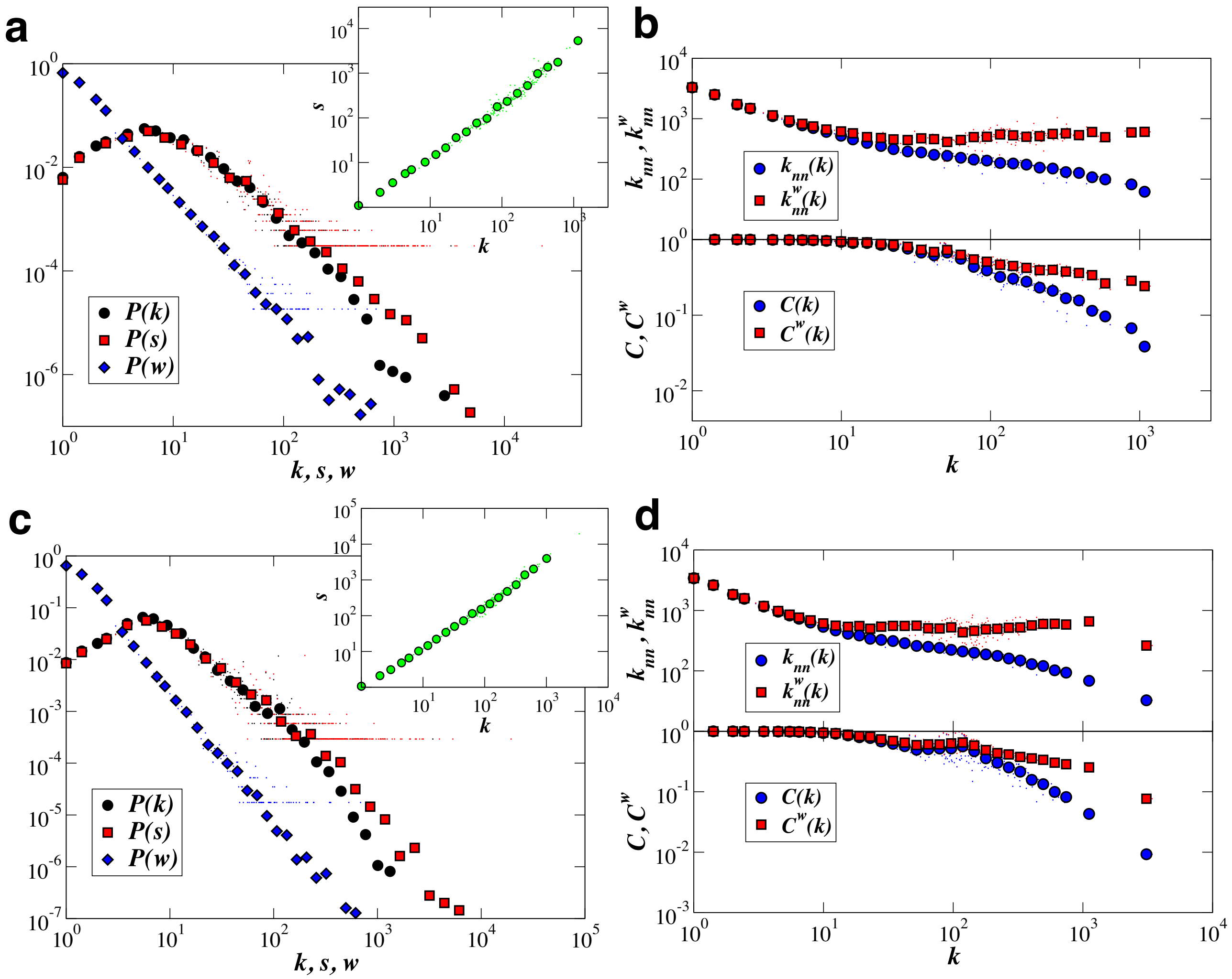
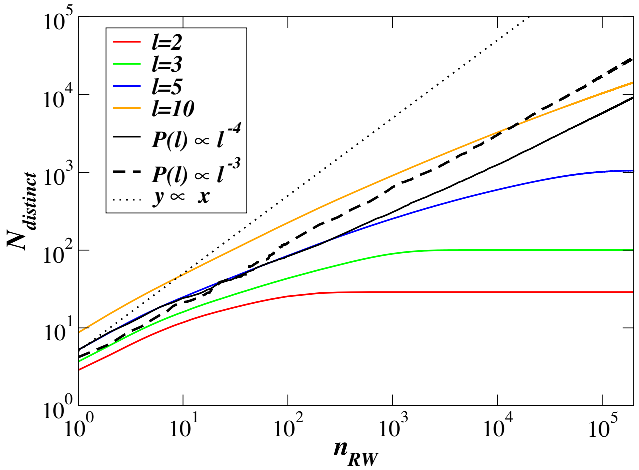


Fig. S3. Same as Fig. S2 for the cooccurrence networks of the tags 6 (*a* and *b*) and 8 (*c* and *d*) in BibSonomy.



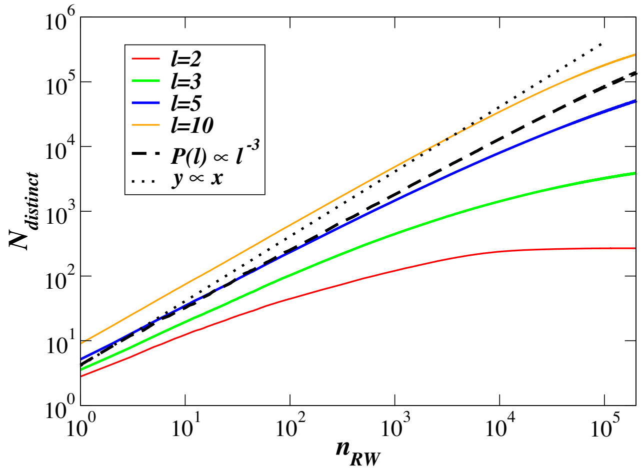


Fig. S4. Growth of the number of visited sites as a function of the number of random walks, for fixed lengths and for randomly distributed lengths with a broad distribution. (*Left*) Walks performed on a Watts–Strogatz network of 5×10^5 nodes, $m = 4$, $p = 0.1$. (*Right*) Walks performed on a random scale-free network with $\gamma = 2.8$, 5×10^5 nodes. The dotted line corresponds to a linear growth, for reference.

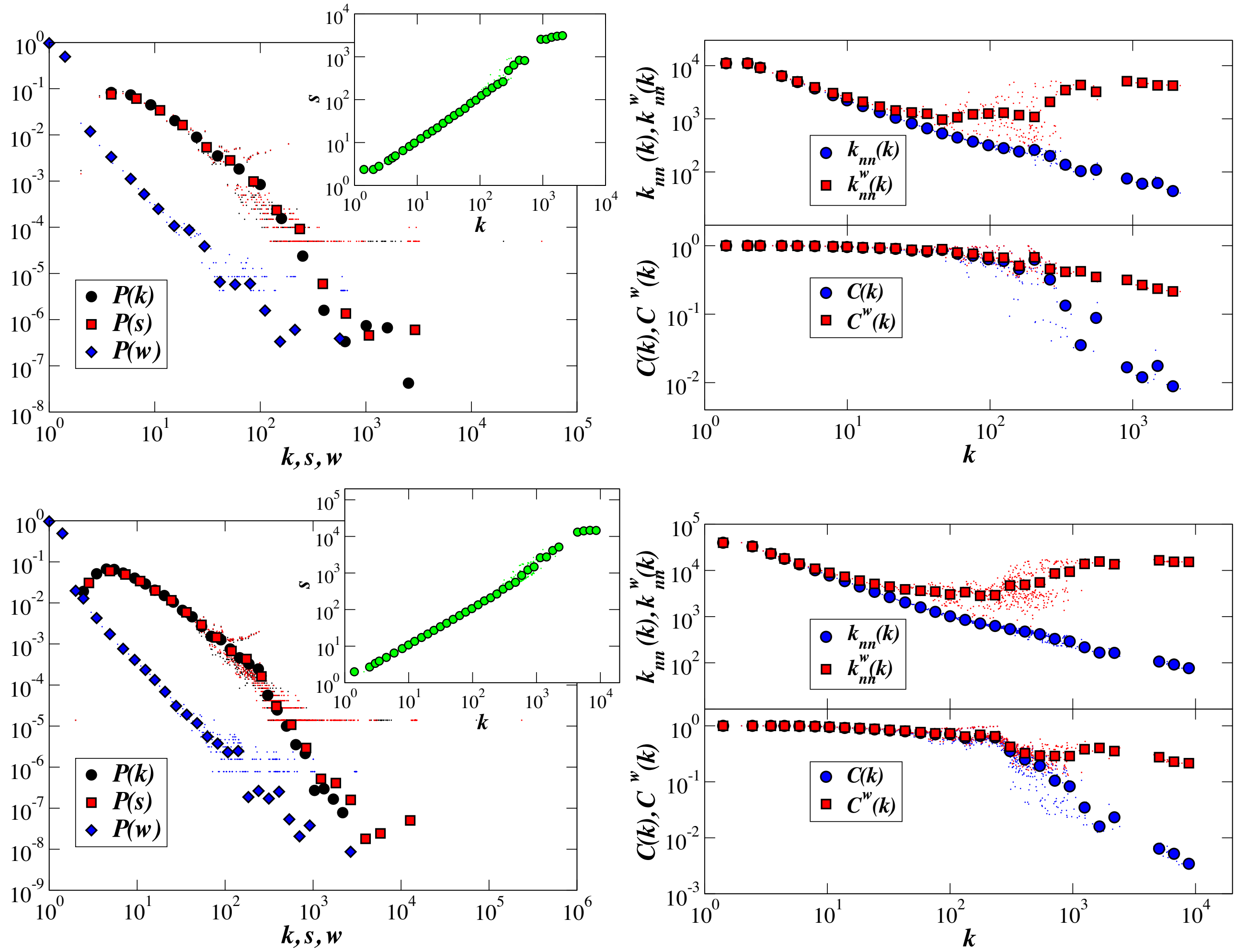


Fig. S5. Properties of the synthetic cooccurrence networks obtained from random walks on an uncorrelated scale-free graph [obtained from the uncorrelated configuration model (9)], with degree distribution $P(k) \propto k^{-\gamma}$, $\gamma = 2.3$; the cooccurrence network is obtained from $n_{RW} = 10^4$ (*Upper*) and 5×10^4 (*Lower*) random walks of random lengths distributed according to a broad $P(l)$.

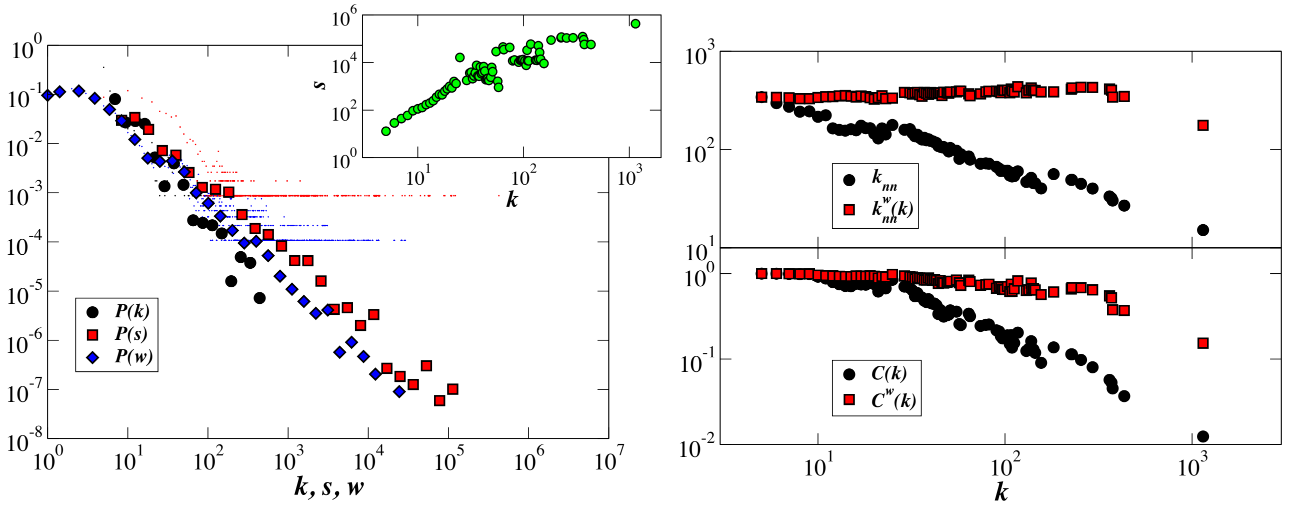


Fig. S6. Properties of the synthetic cooccurrence networks obtained from $n_{RW} = 10^5$ random walks of fixed length $l = 5$ performed on a Watts–Strogatz network of size 5×10^4 nodes and average degree 8, rewiring probability $P = 0.1$.

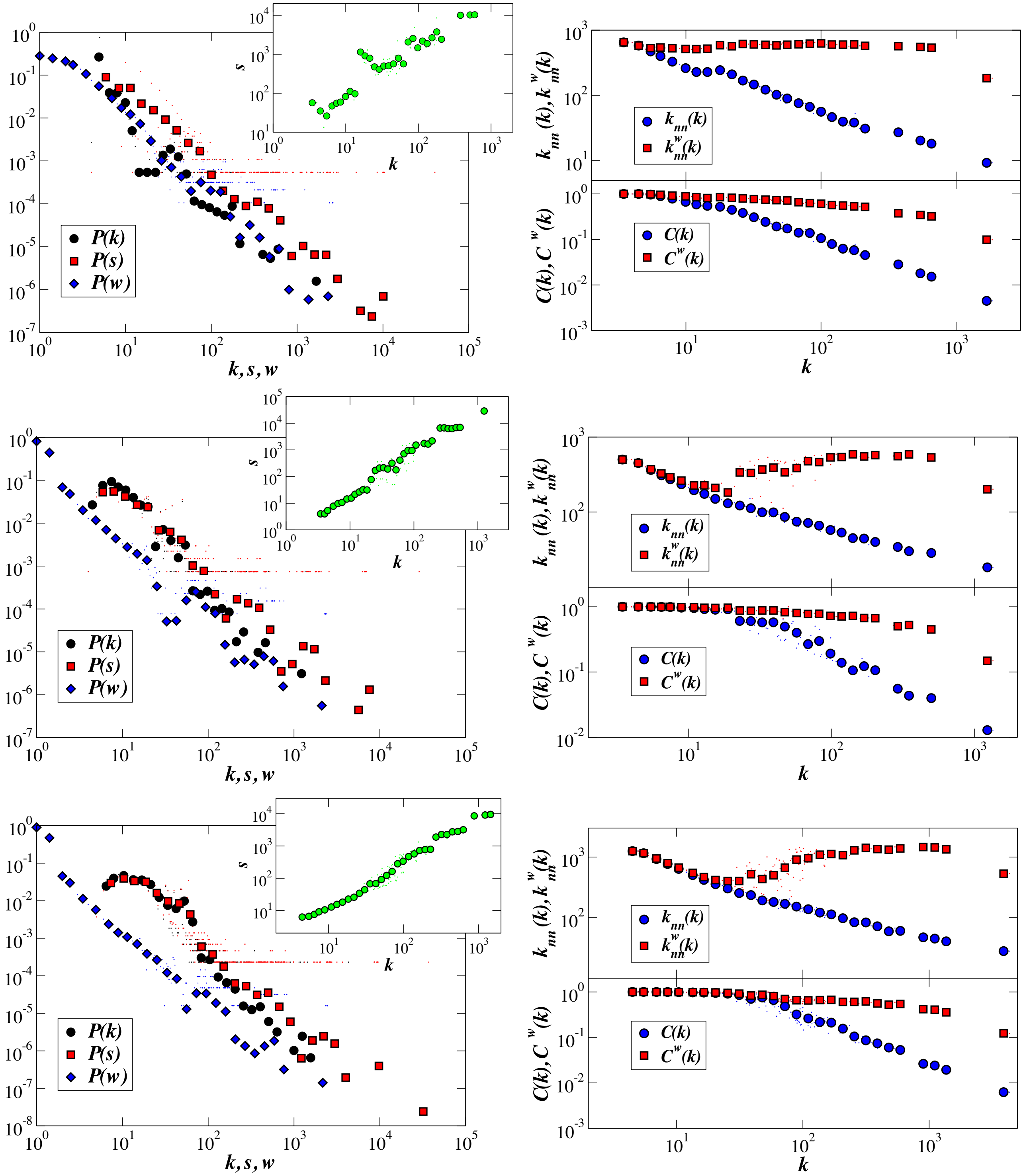


Fig. S7. Properties of the synthetic cooccurrence networks obtained from $n_{RW} = 10^4$ random walks performed on an Erdös–Renyi network of $N = 10^5$ nodes with average degree $\langle k \rangle = 5$. The random walks have fixed length $l = 5$ (*Top*), random lengths according to $P(l) \approx l^{-4}$ (*Middle*), and broad $P(l) \approx l^{-3}$ (*Bottom*).

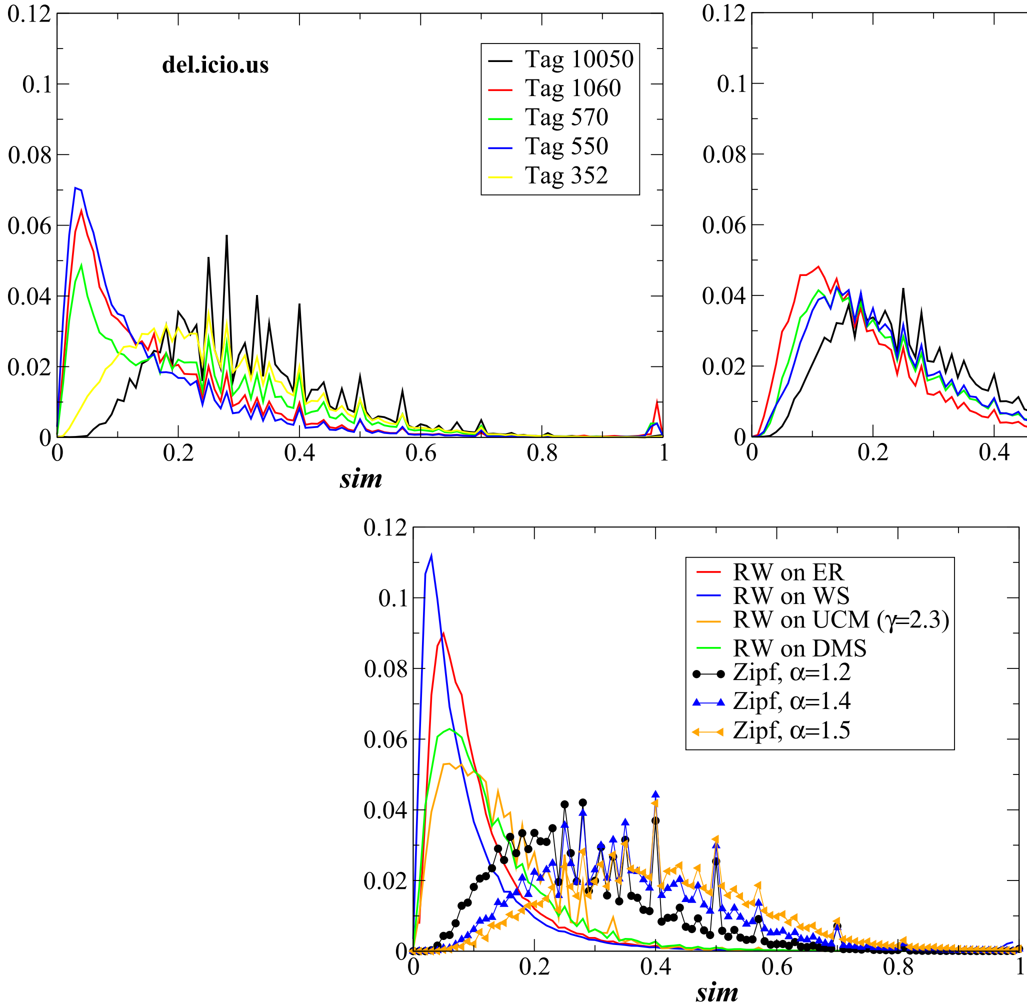


Fig. S8. Distributions of cosine similarities for real and synthetic cooccurrence networks. For del.icio.us and BibSonomy, the tag number represents its popularity rank in the database. Two types of processes are considered for building synthetic cooccurrence networks: random walks (RW) on various types of networks [ER, Erdös–Rényi random graph; WS, Watts–Strogatz network; UCM, uncorrelated configuration model (9) with broad degree distribution $P(k) \approx k^{-\gamma}$; DMS, highly clustered scale-free network with degree distribution $P(k) \approx k^{-3}$ (10)] and artificial posts built from a list of tags whose a priori frequencies follow a Zipf's law of exponent α (symbols).