PAPER

Ontologies and tag-statistics

New Journal of Physics
The open-access journal for physics

Ontologies and tag-statistics

Gergely Tibély[1], Péter Pollner[2], Tamás Vicsek[1,2] and Gergely Palla[2,3]

[1] Department of Biological Physics, Eötvös University, 1117 Budapest, Pázmány P. stny. 1A, Hungary
[2] Statistical and Biological Physics Research Group of HAS, 1117 Budapest, Pázmány P. stny. 1A, Hungary
E-mail: tibelyg@hal.elte.hu, pollner@hal.elte.hu, vicsek@hal.elte.hu and pallag@hal.elte.hu

New Journal of Physics **14** (2012) 053009 (22pp)
Received 3 January 2012
Published 7 May 2012
Online at http://www.njp.org/
doi:10.1088/1367-2630/14/5/053009

Abstract. Due to the increasing popularity of collaborative tagging systems, the research on tagged networks, hypergraphs, ontologies, folksonomies and other related concepts is becoming an important interdisciplinary area with great potential and relevance for practical applications. In most collaborative tagging systems the tagging by the users is completely 'flat', while in some cases they are allowed to define a shallow hierarchy for their own tags. However, usually no overall hierarchical organization of the tags is given, and one of the interesting challenges of this area is to provide an algorithm generating the ontology of the tags from the available data. In contrast, there are also other types of tagged networks available for research, where the tags are already organized into a directed acyclic graph (DAG), encapsulating the 'is a sub-category of' type of hierarchy between each other. In this paper, we study how this DAG affects the statistical distribution of tags on the nodes marked by the tags in various real networks. The motivation for this research was the fact that understanding the tagging based on a known hierarchy can help in revealing the hidden hierarchy of tags in collaborative tagging systems. We analyse the relation between the tag-frequency and the position of the tag in the DAG in two large sub-networks of the English Wikipedia and a protein–protein interaction network. We also study the tag co-occurrence statistics by introducing a two-dimensional (2D) tag-distance distribution preserving both the difference in the levels and the absolute distance in the DAG for the co-occurring pairs of tags. Our most interesting finding is that the local relevance of tags in the DAG (i.e. their rank or significance as

[3] Author to whom any correspondence should be addressed.

characterized by, e.g., the length of the branches starting from them) is much more important than their global distance from the root. Furthermore, we also introduce a simple tagging model based on random walks on the DAG, capable of reproducing the main statistical features of tag co-occurrence. This model has high potential for further practical applications, e.g., it can provide the starting point for a benchmark system in ontology retrieval or it may help pinpoint unusual correlations in the co-occurrence of tags.

Contents

1. Introduction

The network approach has become a ubiquitous tool for analysing complex systems ranging from the interactions within cells through transportation systems, the Internet and other technological networks to economic networks, collaboration networks and society [1, 2]. Over the last decade, it turned out that networks corresponding to realistic systems can be highly non-trivial, characterized by a low average distance combined with a high average clustering coefficient [3], anomalous degree distributions [4, 5] and an intricate modular structure [6–8]. A recently emerged sub-field of growing interest in this area is called *tagged networks, folksonomies* and *hypergraphs*. In general, when studying the topology of the graph corresponding to a real system, the inclusion of *node tags* (also called attributes, annotations, properties, categories and features) leads to a richer structure, opening up the possibility for a more comprehensive analysis. These tags can correspond to any information about the nodes and in most cases a single node can have several tags at the same time. The appearance of tags, e.g. in biological networks, is very common [9–14], where they usually refer to the biological

function of the units represented by the nodes (proteins, genes, etc). Node features are also fundamental ingredients of the so-called *co-evolving* network models, where the evolution of the network topology affects the node properties and vice versa [15–25]. These models are aimed at describing the dynamics of social networks, in which people with a similar opinion are assumed to form ties more easily and the opinion of connected people becomes more similar in time.

The entanglement between tags and the network structure is even deeper in *collaborative tagging systems* or *folksonomies* such as CiteUlike, Delicious or Flickr [26–28], where the network is actually arising in a tagging process. The basic scenario in these systems is that users can tag a certain type of objects (photos, web pages, books, etc) with freely chosen words. Although the limits of the access to objects and tags introduced by others vary from system to system, the arising set of objects with associated free tags is usually referred to as a folksonomy. Since each tagging action is forming a new user–tag–object triple, the natural representation of these systems is given by tri-partite graphs or, in a more general framework, by *hypergraphs* [27, 29, 30], where the hyperedges can connect more than two nodes together. In some cases, the users are also offered the possibility to indicate social contacts (mark each other as a friend), opening up a new dimension for the analysis of the interrelation between tagging and the social ties between users [31, 32].

Folksonomies provide an alternative approach to organize knowledge compared to *ontologies* [33–35]. An ontology usually corresponds to a set of narrower or broader *categories* (capturing the view and concepts of a certain domain, e.g. protein functions), building up a hierarchy composed of 'is a sub-category of' type of relations. The natural representation of this hierarchy is given by a directed acyclic graph (DAG) between the categories. When tagging objects with categories taken from an ontology, we have the benefit that, in principle, all ancestors up to the root in the DAG can be inferred from a single tag on the object. In contrast, the tagging in a folksonomy is either completely 'flat' or at most the users can define a shallow hierarchy for their own tags. Nevertheless, a global hierarchical organization of the tags is not given. One of the very interesting challenges related to folksonomies is to extract an ontology for the tags appearing in the system. Several promising approaches have been proposed: e.g. by aggregating the shallow hierarchies of the individual users [36, 37], using a probabilistic model [38], analysing the node centralities in the co-occurrence network between the tags [39] or integrating information from as many sources as possible [40]. Since a reliable hierarchy between the tags can seriously improve searching, an effective ontology-building algorithm has high potential for practical applications.

Motivated by the ontology extraction problem described above, in this paper we focus on the relation between the structure of the ontology and the distribution of the tags in systems where the DAG describing the hierarchical relations is predefined. The basic idea is that understanding how the ontology effects the tagging can help in improving the methods for reverse engineering the hidden DAG from the tag distribution in folksonomies. Along this line, we examine the statistics of tag occurrence in two large sub-graphs of the English Wikipedia and the protein interaction network of MIPS. We also analyse samples from Flickr, where the user-defined shallow hierarchies are taken into account as individual DAGs. Furthermore, we introduce a simple model for reproducing the observed statistics based on a random walk on the DAG of tags. The paper is organized as follows: in section 2, we define the most important quantities we aim to study, while the details of the investigated networks are given in section 3.

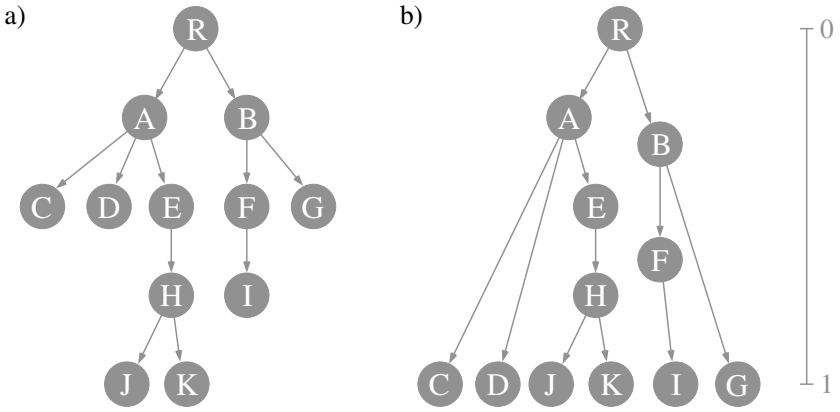

Figure 1. Illustration of the rescaling of the DAG. (a) A small DAG of categories in which leaf nodes appear at various levels. The vertical position of a tag (category) is determined by its distance from the root. (b) After the rescaling the leaves are all at the bottom, and the vertical position of each node is determined by the longest root–leaf path in which it participates.

The obtained statistics are presented in section 4, followed by the description of the random walk model in section 5, with some concluding remarks closing the paper in section 6.

2. Definitions

2.1. Directed acyclic graph (DAG) levels

In the tagged networks we study, the tags are organized into a hierarchy which can be represented by a DAG, where the directed links between two tags correspond to an 'is a sub-category of' type of relation. The tags close to the root in the DAG are usually related to general properties, and as we follow the links towards the leaves, the categories become more and more specific. In some cases, we can find categories in the DAG with more than one in-neighbours, meaning that the given sub-category is a part of several categories that are not in direct ancestor–descendant relation with each other.

Starting from the root, we can define *levels* in the DAG, with the root corresponding to level $l = 0$, the first tags under the root providing level $l = 1$, and so on. To tags that can be reached via multiple paths from the root, we assign the level corresponding to the longest path. (In some cases the level value of a tag is also referred to as the rank of the tag.) One of the simple statistical properties we are interested in is: how does l effect the frequency of the tags, or in other words, are the popular/rare tags close to the root in the DAG or are they more likely to be close to the leaves? At this point, we note that leaves can occur, in principle, at any level in the DAG, since the different branches have usually different maximal depths in a real system. In order to be able to judge the distance of a tag from the leaves as well, we introduce a rescaling of the level values illustrated in figure 1. The rescaled level value, $\tilde{l}$, at the root remains unchanged ($\tilde{l} = 0$), while for any leaf tags we require $\tilde{l} = 1$. For tags in between the two extremes we assign an $\tilde{l} \in [0, 1]$ based on the length of the longest root–leaf branch it takes part in, and $\tilde{l}$ is given by the depth of the tag divided by the maximal depth of the branch.

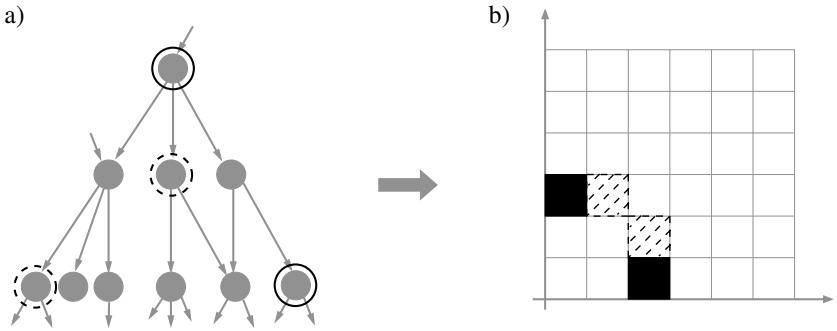

Figure 2. Illustration of the 2D tag-distance distribution for the co-occurring tag pairs. (a) A small part of a DAG with two pairs of tags chosen: the ones marked with continuous circles are in direct ancestor–descendant relation, whereas the tags marked with dashed circles form an 'uncle–nephew' pair. (b) The corresponding cells of the tag-distance distribution are highlighted in solid black colour and with dashed lines, respectively.

2.2. *Tag-frequency*

The *frequency* of the tags in most real systems is heterogeneous, most popular tags occur rather often, whereas others are assigned only to a few objects. A natural choice for the definition of the frequency f_α of a given tag α is simply the number of objects it is assigned to. The probability to find α attached to an object chosen uniformly at random is given by

$$p_\alpha = f_\alpha / N, \tag{1}$$

We note that in systems where a DAG of the hierarchical relations between the tags is given, in principle we could infer all ancestors up to the root from an actually present tag on an object. This enables an alternative definition of the tag-frequency [41, 42], considering the aggregated number of occurrences for all descendants of α and α itself. However, since one of the main motivations of the present work is given by folksonomies (where the DAG is absent), we shall concentrate on the frequency given by simply the number of occurrences.

2.3. *Two-dimensional tag-distance distribution*

Another question of interest is: how does the DAG affect the co-occurrence of different tags on the same object? The simplest idea for measuring the relatedness of a pair of co-occurring tags based on the DAG would be given by their distance. However, for some pairs the connecting path in the DAG is composed of links all going in the same direction, whereas in other cases we might need both upward and downward pointing links to reach one tag from the other. In order to include this aspect in the investigations, we define the *2D tag-distance distribution* for the co-occurring pairs as illustrated in figure 2. The positive quarter plane is divided into unit cells, with the cell at the origin corresponding to distance zero. A given pair of tags contributes to the distribution as follows: starting from one of them we first move upwards in the DAG until the lowest common ancestor is reached. In parallel, we move the same number of cells vertically up in the 2D plane. Next, we move downwards in the DAG to reach the other tag, and

in parallel, we move the same number cells horizontally to the right in the plane, and the number of 'events' in the final cell is increased by one. The contribution from the path going back to the first tag from the second one is taken into account following the same rules: going upwards in the DAG corresponds to moving up in the 2D plane starting from the origin, whereas going down in the DAG corresponds to moving horizontally to the right in the plane. The resulting distribution of the tag-distances is symmetric to the diagonal by construction. The co-occurring pairs of tags which are in direct ancestor–descendant relation contribute to the first column of cells and the bottom row, whereas e.g. the diagonal cells correspond to pairs in which the two tags are equally deep in different branches from their lowest common ancestor (see figure 2 for illustration).

3. The studied systems

We studied the statistical properties of co-occurring tags with predefined DAG in two sub-networks of the English Wikipedia and the protein–protein interaction network of MIPS. Furthermore, we also investigated the tag co-occurrence in the presence of user-defined shallow hierarchies in samples from Flickr.

The protein–protein interaction network of MIPS [43] consisted of $N = 4546$ proteins, connected by $M = 12\,319$ links, and the tags attached to the nodes corresponded to 2067 categories describing the biological processes the proteins take part in. The DAG between these categories was obtained from the Genome Ontology database [44].

In the Wikipedia (http://en.wikipedia.org), the pages are connected by hyperlinks (providing a very interesting network on its own [45–47]), and at the bottom of each page one can find a list of categories, which can be used as tags. We used the same data set as in [41, 42], representing the state of the system in 2008. Since each Wiki category is a page in the Wikipedia as well, these pages were removed from the network to keep a clear distinction between objects and tags. Similarly to the biological processes in the MIPS network, the Wiki categories can have sub-categories and are usually part of a larger Wiki category. (Although the directed graph between the Wiki categories contains a few loops, these can be removed quite easily to obtain a strict DAG [41].) Since the English Wikipedia is quite a large network, we used smaller subsets obtained with a sampling method based on the tag-induced graphs [41]: after choosing a rather general category we keep the pages marked by this tag or any of its descendants. The chosen sub-graphs were induced by the categories 'Japan' (consisting of $N = 61\,581$ nodes, $M = 949\,350$ links and 4939 sub-categories) and 'United Kingdom' (consisting of $N = 318\,183$ nodes, $M = 5432\,914$ links and 30 383 sub-categories).

We also studied a sample from Flickr, corresponding to one of the most popular collaborative tagging systems, designed for tagging photos. We note that due to the lack of a global hierarchy, Flickr seems to be more suitable for, e.g., testing ontology-extracting algorithms than analysing the effect of hierarchy on tag statistics. However, besides tagging, the users can also define a shallow hierarchy for their own tags by grouping their photos into so-called Flickr sets, and putting these Flickr sets into larger Flickr collections up to a limited range of levels. In principle, these user-defined hierarchies allow extraction of similar statistics as in the case of the protein interaction- and Wiki-networks. Our motivation here was to examine whether the statistics of the user-defined small hierarchies show any similarities to the results obtained for the global hierarchies in the case of the other networks. To emphasize

the fundamental difference between Flickr and the other data sets (i.e. the lack of a global hierarchy), the results for Flickr shall be presented in a separate subsection.

The details of the data handling in the case of Flickr are the following. The natural choice for a method for associating a DAG of tags with the user-defined multi-level collections is to link all tags appearing on the photos of a given Flickr set under one ancestor, then link this ancestor tag under an even higher level ancestor corresponding to the Flickr collection the given Flickr set is part of and so on. For ancestor-tags, we used words from titles of the Flicker-sets and Flickr-collections. (For many users these titles consisted actually of only a single word, in a similar fashion to photo tags.)

A single user-defined DAG obtained in the above way is much smaller and is likely to be less well designed compared to, e.g., the overall DAG of the Wiki categories. Thus, the expected outcome for the tag statistics is promiscuous with a lot of noise. In order to increase at least the sample size, instead of using only a single user-defined DAG, we actually prepared the statistics for each user separately and aggregated the results. At this point we note that the contribution of the tags appearing on the photos of a given user to the tag-distance distribution becomes somewhat 'singular' when using the DAG obtained from the Flickr sets and Flickr collections of the same user: all co-occurring tags are siblings. Thus, these 'self-contributions' were left out from the aggregated tag-distance distribution.

4. Applications

4.1. The structure of the DAG

We start our analysis with an interesting effect related to the structure of the DAGs describing the hierarchy of tags in the systems we investigate. In figure 3, we plot the size of the levels (how many tags occur at a given level) as a function of the level depth. For convenience, the vertical axis for each plot is rescaled by the size of the largest level. According to figure 3(a), the size of the levels is small when we are either very close to the root or very far from it, whereas it becomes larger in between. (Since the maximum depth is different in each system, the horizontal axis in this case has been rescaled from l to $l/l_{\max}$, where $l_{\max}$ denotes the length of the longest branch in the DAG.) However, the shape and place of this maximum is unique for each system. (An alternative illustration of this effect is given in the appendix, where in the top panel of figure A.1 the differences between the un-scaled DAGs are more apparent.) In contrast, when we switch to the rescaled level depth $\tilde{l}$, the curves become roughly uniform with a more or less monotonically increasing shape, as shown in figure 3(b). Thus, the rescaling of the level values has an interesting side effect on the shape of the DAG, bringing it closer to a 'triangular' form, similar to the shape of a regular hierarchical graph.

4.2. Tag-frequencies and level values

As our interest is mainly in the interplay between the tag-hierarchy and the statistical properties of tag-occurrences, in figure 4 we show the average tag-frequency as a function of the level depth. When no rescaling is applied (apart from dividing l by the maximal level depth $l_{\max}$), the tag-frequency is almost completely independent of the level depth in a wide range of l (figure 4(a)). In contrast, when switching to the rescaled $\tilde{l}$, a clear decreasing tendency can be

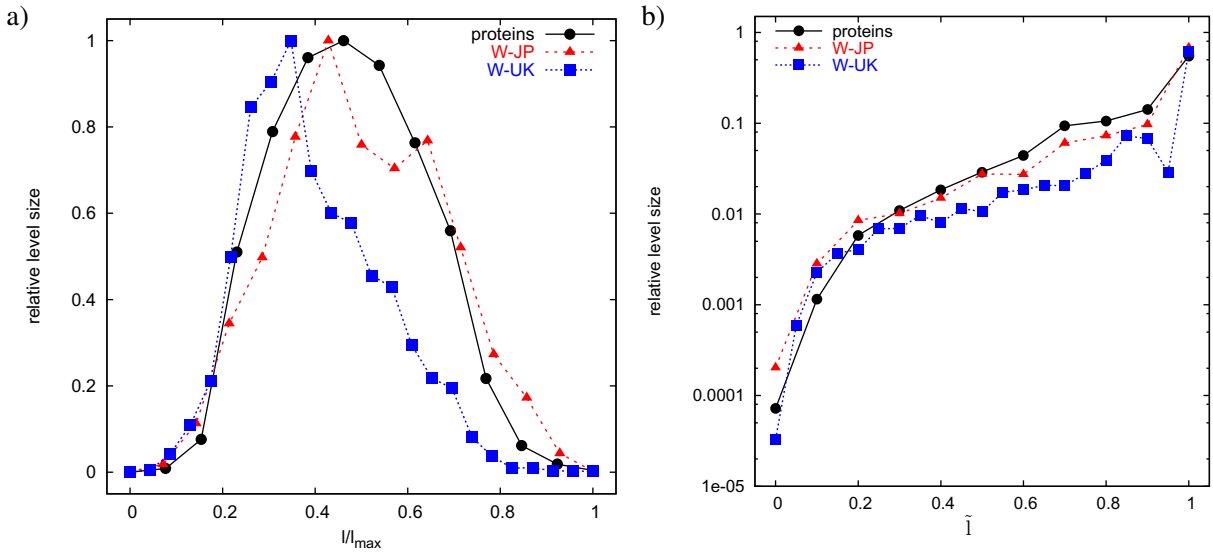

Figure 3. (a) The relative size of the levels in the DAG (scaled with the largest level) as a function of $l/l_{\max}$, where $l_{\max}$ denotes the length of the longest branch in the DAG. (b) The relative size of the levels in the DAG as a function of the rescaled level depth $\tilde{l}$. Note that the vertical axis is logarithmic.

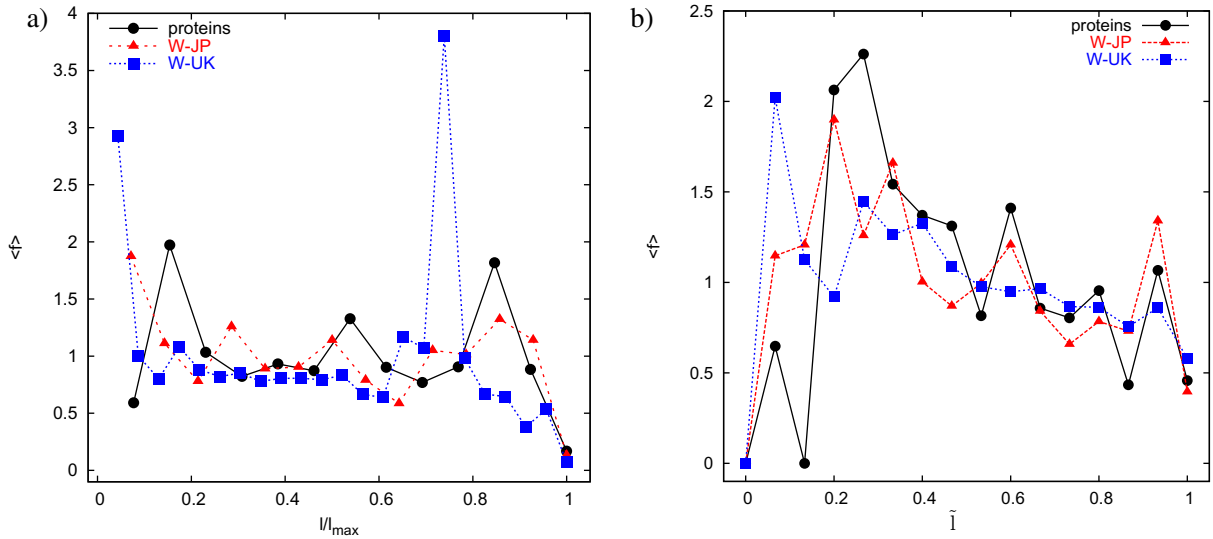

Figure 4. (a) The average frequency $\langle f \rangle$ of the tags on a given level as a function of $l/l_{\max}$. (b) The average frequency $\langle f \rangle$ of the tags on a given level as a function of the rescaled level depth $\tilde{l}$.

observed, apart from the very low $\tilde{l}$ region (corresponding to levels close to the root). This non-trivial result indicates that the frequency of a tag is more sensitive to the depth of the branches starting from it compared to its distance from the root. A plausible explanation of this effect is the following: we have already pointed out that leaves can occur basically at any level in a large enough real world DAG. As we move upwards from the leaves, presumably the importance

(relevance, rank, significance, standing, etc) of the tags is increasing at least in the first few steps. However, since the leaves we started from were located at various levels, we arrive at the conclusion that tags with higher relevance can also occur at a wide range of levels in the DAG. Thus, the level value l of a tag, measuring its global distance from the root, is not very informative in this respect, and accordingly, it has no significant effect on the average frequency of the tags. In contrast, by switching to the rescaled level value $\tilde{l}$, we also take into account the depth of the local branches starting from the given tag, which seem to be more relevant for evaluating the standing of a tag in the hierarchy, as the frequency of tags is decreasing with $\tilde{l}$. (The more important tags have longer sub-branches starting from them and thus, on average, have lower $\tilde{l}$ values.)

4.3. Tag-distance and co-occurrence

Next we move onto the examination of the 2D tag-distance distributions defined in section 2.3. For illustration, in figure 5(a) we show the contribution from the tag 'British kings involved in Caesar's invasion of Britain' in the case of the Wiki-UK network, where the number of occurrences together with other tags at a given distance l_{diff} are indicated by the colour of the corresponding cell. The sub-graph between the tags and the lowest common ancestors in the DAG is given in the inset. For comparison, in figure 5(b) we show the contribution from 'World War I poems' in a similar fashion. Here the routes through the lowest common ancestors between the co-occurring tag pairs are much longer; thus, they are displayed separately in figure 5(c). The tags co-occurring with 'British kings...' are close in the DAG, and accordingly, their contribution in the 2D distribution is close to the origin. In contrast, the distances to the tags co-occurring with 'World War I poems' are long; thus, their contribution falls in cells far from the origin.

In figure 6, we plot the complete 2D tag-distance distributions for the networks we investigated. According to figure 6(a), the maximum of the plots is a few steps away from the origin, which might seem a bit surprising at first sight. In order to reveal the background of this effect we also measured the average tag-distance distribution for a random tag assignment analogous to the configuration model in the networks literature. Here the DAG is taken from the system under study, and we consider the ensemble of all possible associations of tags to the objects consistent with the observed number of occurrences for the tags and the observed number of tags on the individual objects. To simulate draws from this ensemble one can apply a randomization procedure, in which a pair of tags is swapped between two randomly chosen objects in each step. This way both the number of tags on the objects and the tag-frequencies are preserved. The average tag-distance distributions for this random tag assignment are shown in figure 6(b), with the maximums even further away from the origin compared to the original data. An alternative possibility for randomization is to replace the DAG of the original system by a random DAG of the same size. For this we used the random DAG model introduced in [48] with a fixed number of nodes and links. The results for random DAGs and the original distribution of the tags on the objects are displayed in figure 6(c), showing a picture somewhat similar to figure 6(b). Finally, to highlight the part of the original tag-distance distribution that cannot be accounted for by random effects, in figure 6(d) we show the z-score of the individual cells, defined as the difference between the original distribution (figure 6(a)) and the average for the random tag assignment (figure 6(b)) scaled by the standard deviation of the random tag assignment. The maximums in these plots have clearly moved close to the origin, showing that

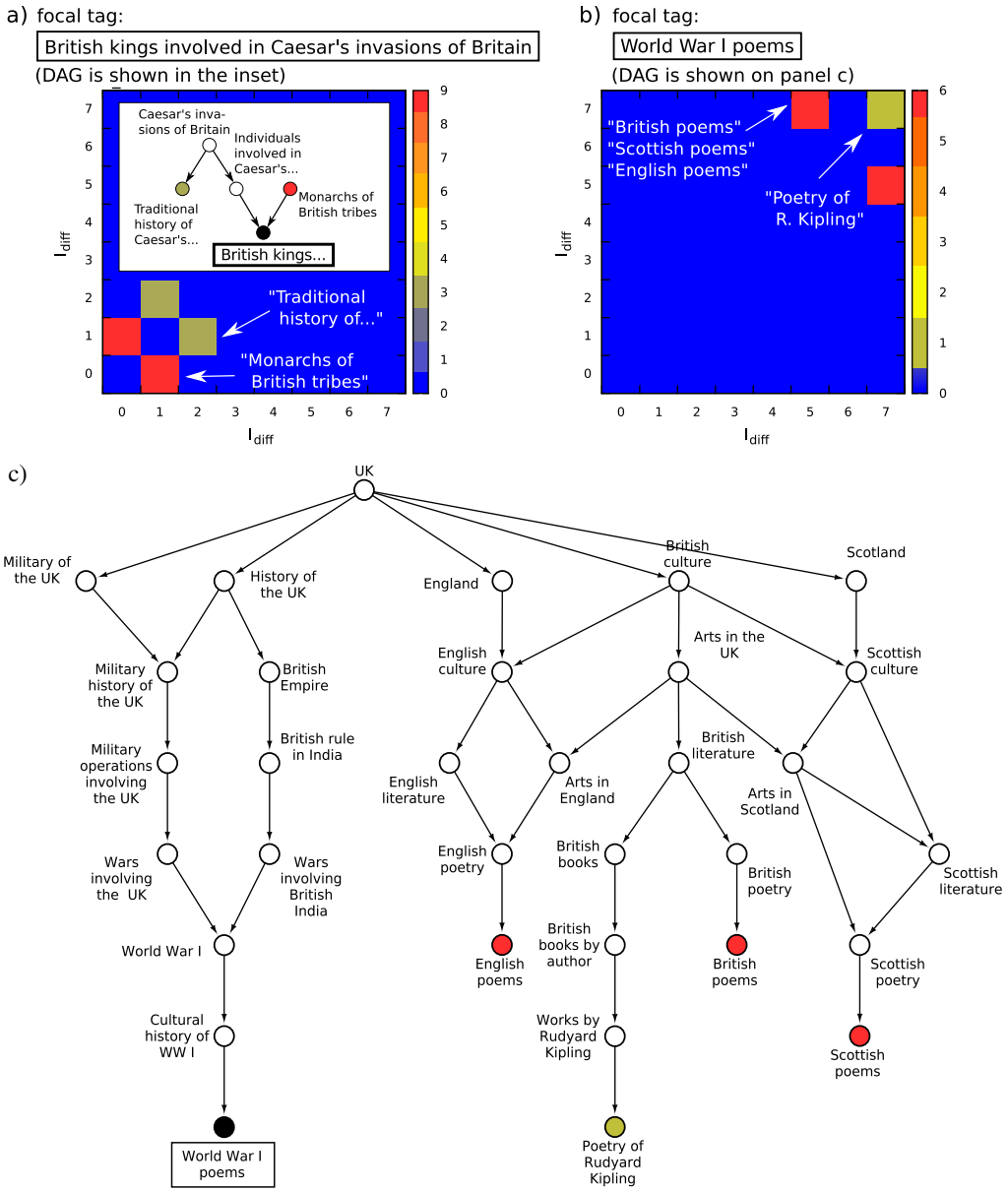

Figure 5. Illustration of the calculation of the tag-distance distribution in case of the Wiki-UK network. (a) The contribution from the tag 'British kings involved in Caesar's invasion of Britain'. The colours indicate the number of co-occurrences of other tags with the focal tag; the inset shows the corresponding sub-graph in the DAG. (b) The contribution from the tag 'World War I poems'. (c) The shortest paths to the lowest common ancestors in the DAG for tag pairs considered in panel (b).

the co-occurrence of tags only a few steps away in the DAG is far more probable than at random in the systems we investigated.

The three z-score plots in figure 6(d) also reveal an interesting difference between the systems: in the case of the protein interaction network the maximum is in the diagonal, while

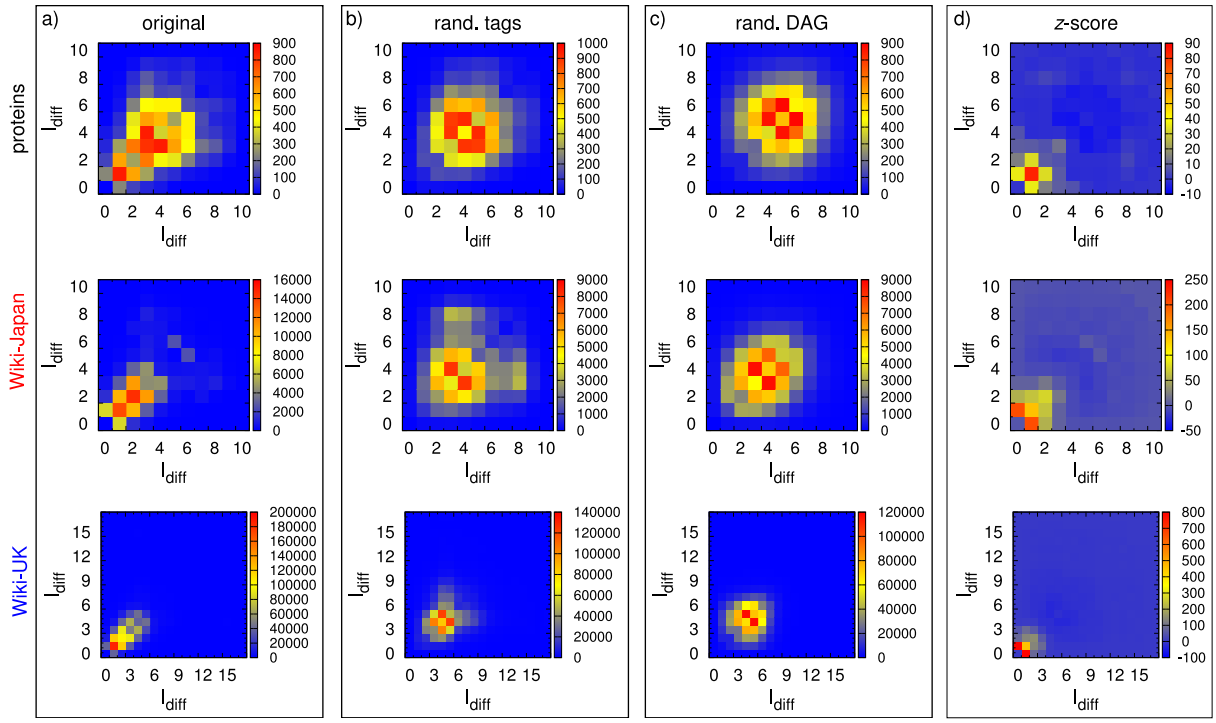

Figure 6. (a) The 2D tag-distance distribution for co-occurring tag pairs in the studied systems (colour-coded). (b) The average tag-distance distribution when the tags are randomized keeping the tag frequencies and the number of tags on the objects fixed. (c) The average tag-distance distribution for random DAGs. (d) The z-score corresponding to the difference between the original data (panel (a)) compared to the random tag assignment (panel (b)) in units of the standard deviation of the random tag assignment.

for the two Wiki networks it is in the first row (or the first column). This means that for the protein interaction network the most enhanced co-occurring tags are like 'brothers', i.e. they are at the same depth from their lowest common ancestor on different branches. In contrast, the maximum places for the Wiki networks correspond to tag pairs in direct ancestor–descendant relation with each other.

The results discussed in sections 4.1–4.2 have shown that switching to the rescaled DAG levels $\tilde{l}$ can reveal interesting effects otherwise hidden when using the original l. Therefore, in figure 7 we replot the tag-distance distributions shown in figure 6 when the distance between the tags is measured according to $\tilde{l}$. Since $\tilde{l}$ can take up real values in [0,1], we introduced bins of size 0.05. Similarly to figure 6, the maximums are far from the origin for both the original data sets (panel (a)) and their random counterparts (panels (b)–(c)), while they are shifted close to the origin for the z-score (panel (d)). The difference between the z-score of the protein interaction network and that of the Wiki networks is somewhat more apparent here compared to figure 6: the maximum is along the diagonal in the first case, while it is concentrated in the first row (or column) in the latter cases. A further nice feature of using the rescaled levels is that the

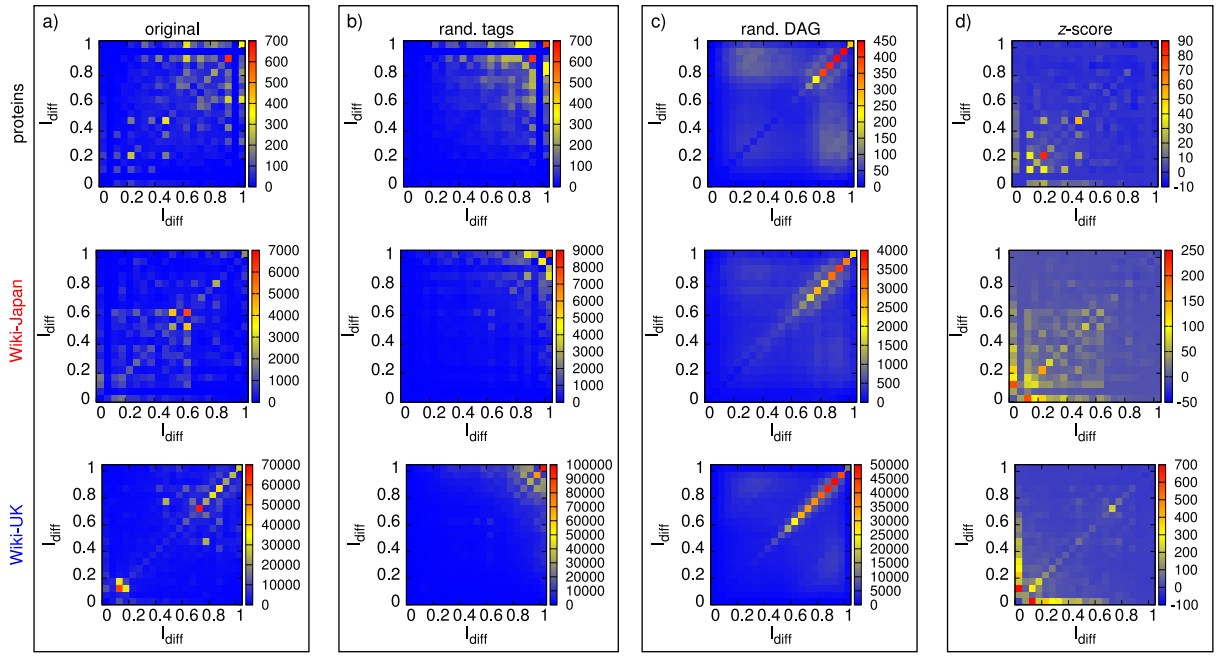

Figure 7. The tag-distance distributions shown in figure 6 when the distances are measured according to the rescaled level value $\tilde{l}$. Since $\tilde{l}$ can take up real values (not only integers as l), we introduced bins of size 0.05. Similarly to figure 6, besides the original data (panel (a)) we also show the results for random tag assignment (panel (b)), random DAGs (panel (c)) and the z-scores (panel (d)).

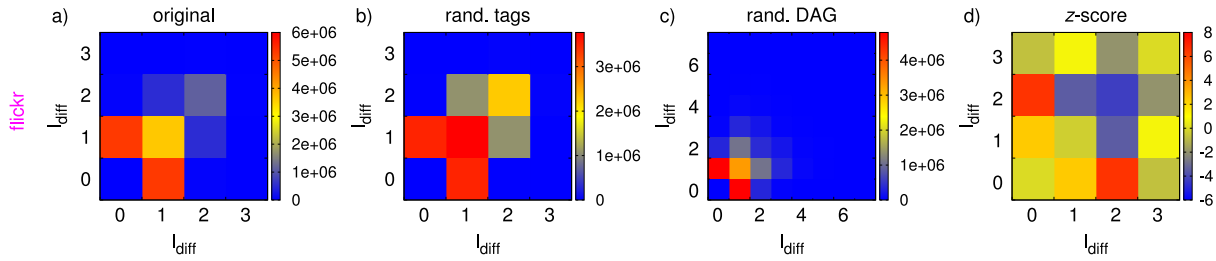

Figure 8. (a) The tag-distance distribution obtained for a sample from Flickr. We measured the distances between the co-occurring tags on a given photo belonging to a given user using the shallow hierarchies of the other users in the sample, and aggregated the results for all photos and all DAGs. (b) The average tag-distance distribution for random tag assignment. (c) The average tag-distance distribution for random DAGs (d) The z-score obtained by comparing the results shown in panel (a) and the null-model displayed in panel (b).

tag-distance distribution of the random tag assignment has high values around (1,1), in contrast to the traditional level-based distribution, which has high values in a non-trivial, case-specific region (see figure 6(b)). In summary, the overall behaviour of the tag-distance distribution is similar when using $\tilde{l}$ instead of l, with some slight changes in the details, which can help pinpoint specific features varying from system to system.

4.4. Results for Flickr

We also prepared the 2D tag-distance distribution using the user-defined shallow hierarchies for a sample from Flickr, as described in section 3. We emphasize that in this case the main question is whether these small user-defined DAGs show a similar behaviour, on average, to the overall global DAGs seen in the protein- and Wiki-networks. Since the maximal level depth in this case is only $l = 3$, the distinction between close and far away tags becomes a bit artificial, (e.g. for direct descendants the largest possible distance is 3). According to the results shown in figure 8, the maximum of the 2D tag-distance distribution is close to the origin for both the original data (figure 8(a)) and its randomized counterparts (figures 8(b) and (c)). From the z-score (figure 8(d)) we can see that the cells having the most significant enhancement in the number of tag-pairs compared to the random tag assignment correspond to direct descendants within distances 1 and 2. Our conclusion is that the behaviour of the small user-defined DAGs is consistent with the results shown previously for the three tagged networks with predefined global hierarchies; however, the enhancement in the number of close tag-pairs is far less striking. An interesting question (which is outside the scope of this work) related to the above is the following: how would the tag-distance distribution behave in the Flickr data set if the set of user-defined shallow hierarchies is replaced by a unique overall DAG obtained from an ontology-extraction algorithm?

4.5. Local standing versus global rank in the tag-distance distribution

In section 4.2, it has been seen that the length of the local branches starting from a given tag in the DAG has a much larger effect on its frequency compared to its global distance from the root. An interesting question related to this is the following: can we observe a similar effect in the behaviour of the 2D tag-distance distribution as well? Since the distribution depends on the relative distance between the co-occurring tags, the answer to this question is not straightforward. However, as we shall see, the partial randomization of the DAG can reveal a difference between the influences of the 'local' and the 'global' configuration of the DAG on the tag-distance distribution, and this effect is consistent with the findings detailed in section 4.2.

The basic idea is the following: according to the previous results the shape of the 2D tag-distance distribution is modified when randomizing the tag assignment of the objects, and these changes can be tracked by the z-score. Here we are going to use the z-score to monitor the effect of changes in the DAG. First let us assume a completely random DAG: in this case the tag-distance distribution is already random even for the original tag assignment of the objects, and switching to random tag assignment does not make much difference, and accordingly, the z-score becomes flat. Now let us assume randomizing only a smaller sub-graph in the DAG. If this chosen smaller sub-graph has only a slight influence on the tag co-occurrence, then its randomization should not make a significant difference compared to the original DAG, and the z-score is expected to be very similar to the case when the DAG is not modified at all. In contrast, if the chosen smaller sub-graph is crucial, then its randomization has a large effect, and the z-score becomes more similar to that of the completely random DAG.

In order to apply the above framework for comparing the importance of the global and the local structure of the DAG, we need to find sub-graphs in the DAG which are more related to the global structure, and other sub-graphs entangled with the local structure, respectively. As a starting point, let us first consider the swapping of tags just below the root of the DAG with each other (in other words, the randomization of level $l = 1$): for the vast majority of the tags the

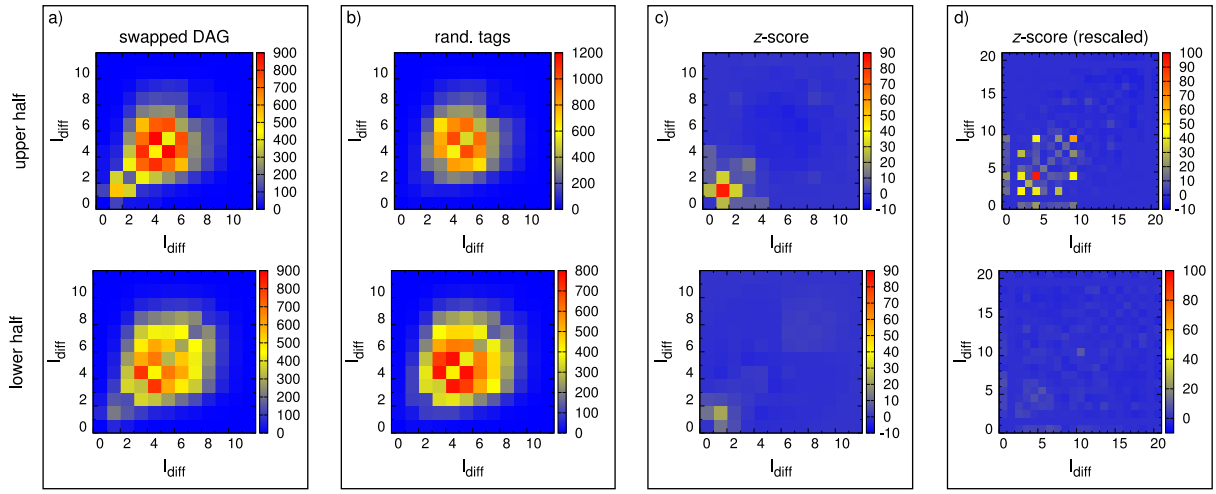

Figure 9. Comparison between the tag-distance distributions obtained after randomization of the 'upper half' of the DAG (top row) and the 'lower half' of the DAG (bottom row) in the case of the protein interaction network. (a) The obtained 2D tag-distance distributions for the partially randomized DAGs and the original tag assignment of the objects. (b) The average 2D tag-distance distributions for the partially randomized DAGs and random tag assignment of the objects. (c) The z-score corresponding to the difference between panel (a) and panel (b) in units of the standard deviation of panel (b). (d) The z-score when the distance between the tags is measured according to the rescaled level value $\tilde{l}$.

local neighbourhood in the DAG is left unchanged, whereas the global path leading to the root has been modified. In contrast, if we start swapping leaf nodes, the global structure of the DAG is not altered, while the local neighbourhood for some of the tags is changed. Along this line we divided the DAG of the systems we investigated into an 'upper half', corresponding to levels close to the root, which are more related to the global structure of the DAG, and a 'lower half', composed of bottom levels far from the root, which have more to do with the local structure of the DAG. (The two parts contained the same number of tags.)

In figure 9, we show the results for randomizing the 'upper half' (top row) and the 'lower half' (bottom row) of the DAG separately in the protein interaction network. During the randomization process at each step a pair of tags from the given half was randomly swapped in the DAG. In figure 9(a), we show the 2D tag-distance distribution for the partially randomized DAGs with the original tag assignment of the objects. Similarly to figures 6 and 7, in figure 9(b) we display the results for the same partially randomized DAGs when the tag assignment of the objects is also randomized. The corresponding z-scores are given in figures 9(c) and (d) for both for the original level values l and the rescaled level values $\tilde{l}$, showing striking differences between randomizing the 'upper half' (top row) and the 'lower half' (bottom row). In the case of randomizing the 'upper half', the z-scores are quite similar to the z-scores shown in figures 6 and 7 (although some small details look slightly different). In contrast, the z-scores for randomizing the 'lower half' show drastic deviations from the original z-scores: we can observe only weak reminiscences of the maximum close to the origin, and the landscape becomes almost

completely flat. This enhanced sensitivity of the 2D tag-distance distribution to the changes in the 'lower half' of the DAG compared to changes in the 'upper half' is in agreement with the enhanced sensitivity of the tag-frequencies to the local position of the tags in the hierarchy compared to the global distance from the root: when randomizing the 'upper half', the local position for at least the tags in the 'lower half' is preserved, whereas the global routes to any tag are messed up. In contrast, when randomizing the 'lower half', while preserving the global structure, we mess up the local position for the majority of the tags (as 'upper half' tags are also likely to have branches reaching into the 'lower half'). We observed a similar behaviour in the case of randomizing partly the DAG of either the Wiki-Japan or the Wiki-UK network as well.

In summary, the partial randomization of the DAG has revealed an interesting difference between the sensitivity of the 2D tag-distance distribution to the global and the local structure of the DAG. This effect is in complete agreement with the previously observed enhancement of the importance of the local position compared to the global position in the DAG from the point of view of tag frequencies.

5. Random walk model

According to the results of section 4, the DAG between the tags has indeed an effect on the co-occurrence of tags. In this section, we demonstrate that a rather simple model can reproduce the main statistical features observed for the real systems. Since the co-occurring tags were closer to each other in the DAG than at random, the model has to provide a mechanism for choosing pairs from the DAG with an enhanced probability for close tags. A natural idea is to pick the first tag at random, then start an undirected random walk on the DAG from the chosen tag, and after a few steps pick the target reached. (In some respects this approach is a sort of 'dual model' of the random walk model introduced in [49] on the network of word associations, which was used for inferring similarity relations between words.)

If we take the DAG as predefined (e.g. the DAG of the system we would like to model), then the two 'parameters' of the model are given by the frequency distribution of the tags and the length distribution of the random walks. For the tag-frequencies the first natural idea is to use the frequency distribution measured in the real data. However, we also worked with uniform tag-frequencies set to the average value measured in the real data. In the case of the random walk length distribution we tried out the gamma-distribution, the uniform distribution, the lognormal distribution, and the Poisson distribution. For all choices the average length of the walks was set to a value ranging between 3 and 10. According to the results, the tag-distance distribution is very robust against changes in any parameters.

In figure 10(a), we show typical tag-distance distribution results for the random walk model. The DAGs used in these simulations (indicated on the left of each row) were taken from the tagged networks we studied in section 4.3, and the frequency of the tags as well as the number of tags on the objects were set to the average values measured in the corresponding real system. The length distribution for the random walks was a uniform distribution in the [3,10] interval. (In the appendix we show very similar results for different random walk length distributions.) Similarly to the case of the real systems we studied, in figure 10(b) we also show the results for a suitably chosen null-model, which in this case corresponds to choosing the tag-pairs at random, irrespective of the DAG. In figure 10(c), we show the results for random walks

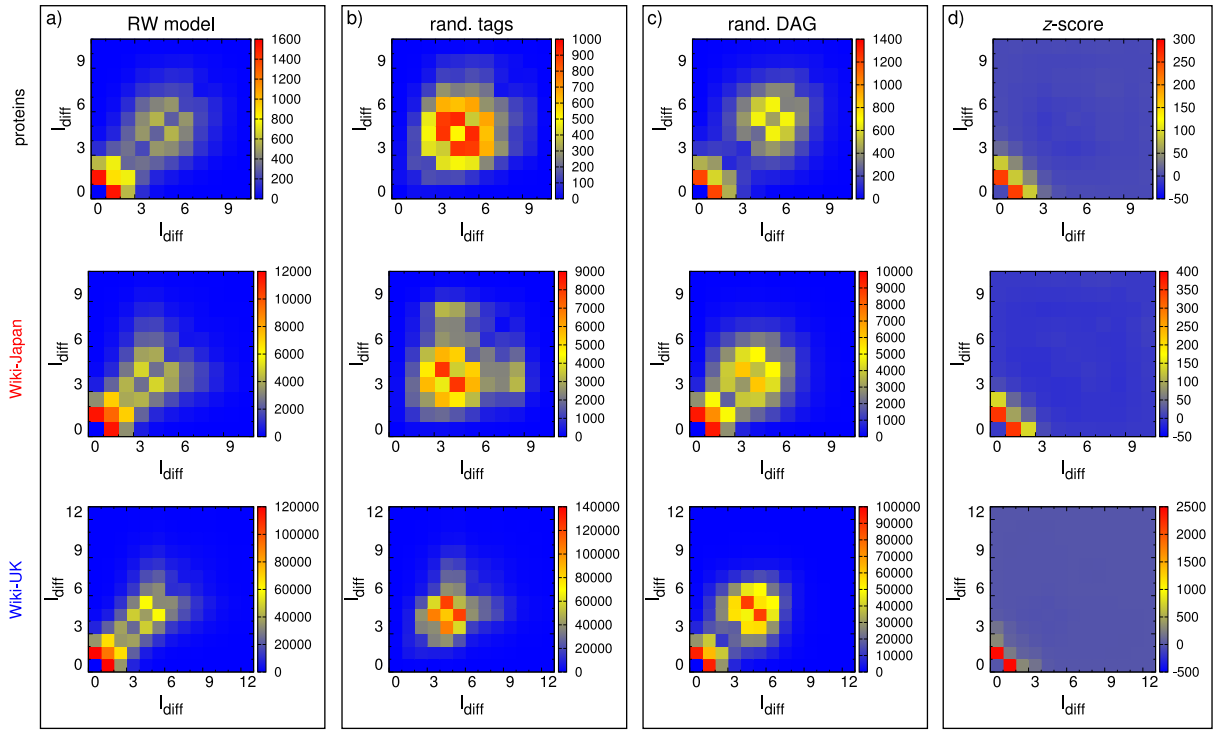

Figure 10. The tag-distance distributions for co-occurring pairs of tags in the random walk model (colour coded), where the DAG was taken from the protein interaction network (first row), the Wiki-Japan network (second row) and the Wiki-UK network (third row). Similarly to figure 6, besides the actually measured values (panel (a)), for comparison the results for un-correlated tag assignment (panel (b)), the results for random DAGs (panel (c)) and the z-scores (panel (d)) are also shown.

on random DAGs generated using the model introduced in [48]. For highlighting the part which is only present due to the correlations induced by the random walk, in figure 10(d) we also display the z-scores (corresponding to the difference between figures 10(a) and (b) in units of the standard deviation of figure 10(b)). Similarly to the behaviour observed in the real systems, the maximum in the z-score is shifted close to the origin in all cases.

In summary, according to the simulations, our random walk model qualitatively reproduces the main properties of the tag-distance distribution for the co-occurring tags observed in real systems. Although the model is rather simple, it has high potential for further applications. On the one hand, with the further development of ontology extraction algorithms sooner or later the need for a controllable benchmark system will arise. The rough outline of this benchmark is that stochastic collections of tagged 'objects' (sets of co-occurring tags) are generated based on given input DAG, and using these collections as input we can test how well does a given ontology extraction algorithm recover the DAG. Our random walk model can provide a starting point for generating random collections of co-occurring tags with some sort of 'memory' of the underlying DAG of hierarchy between the tags.

On the other hand, the tag-distance distribution generated by the random walk model can also help point out non-trivial effects in the tag-distance distribution of the original data where the DAG was taken from. For example, when using the DAG taken from the Wiki networks, the behaviour of the tag-distance distribution in the random walk model is very similar compared to the original one. However, when repeating this experiment with the protein interaction network, the two tag-distance distributions are only roughly similar: although the maximum in the z-scores is shifted close to the origin in both cases, for the original data it remains in a diagonal position (see figures 6 and 7), while it becomes off-diagonal for the random walk model (see figure 10). Thus, the co-occurrence patterns of the protein interaction network have features which cannot be explained by a simple random walk on the DAG.

6. Conclusions

Motivated by the ontology extraction problem in collaborative tagging systems and folksonomies, we studied the statistical properties of tag occurrence in tagged networks where the DAG of hierarchy between the tags is predefined. In order to be able to give support to the further development of ontology-extraction algorithms, this research was focused on the interaction between the DAG and the tag-statistics. Our most interesting result is that the local standing (rank, significance, etc) of the tags in the DAG has a much more relevant effect on the tag-statistics compared to the global distance from the root. This is supported, on the one hand, by the change in the behaviour of the tag-frequency as a function of the level value when switching to the rescaled levels $\tilde{l}$, and, on the other hand, by the different sensitivity of the 2D tag-distance distribution for co-occurring tags to randomizing the 'upper half' or the 'lower half' of the DAG.

According to our studies on a protein interaction network and two sub-networks from the English Wikipedia, the average frequency of the tags is more or less independent of the level value (distance from the root) in the hierarchy of the tags. In contrast, if we switch to a rescaled level value $\tilde{l}$ taking into account also the length of the sub-branches starting from the given tag in the DAG, we see a decreasing tendency in the tag-frequency with growing $\tilde{l}$ in a wide range of $\tilde{l}$. A plausible explanation for this interesting effect is that the distance from the root is not a good indicator of the importance (significance, rank, etc), e.g. in the DAGs we studied leaves (corresponding probably to the most specific tags) occurring at a wide range of levels. However, the lengths of the branches starting from a given tag provide an alternative candidate for evaluating its importance, and in contrast to the distance from the root, this measure is of local nature. The above result suggests that also taking into account this local information when evaluating the rank of a tag yields a quantity that is much more entangled with the tag-frequency compared to the traditional level value.

We studied the statistical properties of co-occurring tag pairs by introducing a 2D tag-distance distribution for the relative positions in the DAG. We compared this distribution for the three investigated systems with the distribution obtained for a random tag assignment analogous to the configuration model in the complex network literature. According to the z-scores, close pairs of tags co-occur in these systems far more often than expected at random. Furthermore, these 2D plots also reveal an interesting difference between the protein interaction network and the Wiki networks: in the first system the co-occurring tag pairs are much less likely to be

direct descendants of each other compared to the other two networks; instead they are often like 'cousins', 'brothers' or 'nephews'. We also analysed the 2D tag-distance distribution obtained for a sample from Flickr using the shallow hierarchies defined by the users. The results were consistent with the behaviour seen for the tagged networks with predefined DAG; however, the increase in the number of close tag pairs compared to the random null model was far less striking.

In order to examine the difference between the importance of the local and the global position of the tags in the hierarchy from another perspective, we applied restrictive randomization to the DAG by dividing it into an 'upper-half' and a 'lower-half' of equal size. The induced changes in the 2D tag-distance distribution showed significant differences: the effect of randomizing the 'upper-half' is marginal, whereas the structure of the z-score undergoes a drastic transformation when randomizing the 'lower-half'. Since randomizing the 'upper-half' modifies mainly the global structure, while randomizing the 'lower-half' reshuffles mainly the local structure, this effect is in complete agreement with the previously observed imbalance between the importance of the local and the global standing of tags (in favour of the local one) from the point of view of tag-frequencies.

Finally, we introduced a simple model based on random walks on the DAG for describing the enhancement of close tag-pairs in the tag-distance distribution. According to our simulations, this approach can reproduce the shift of the maximum towards the origin in the z-score in a robust way. Although simple in nature, this model has relevant potential for further applications, e.g. it can provide a starting point in constructing benchmark systems for ontology-extraction algorithms, and can also help in pinpointing non-trivial effects in the tag-distance distribution of real systems.

Acknowledgments

This work was supported by the European Union and co-financed by the European Social Fund (grant agreement no. TAMOP 4.2.1/B-09/1/KMR-2010-0003).

Appendix

A.1. The structure of the DAGs

The DAG capturing the hierarchical relations between the tags plays a crucial role in our analysis, and in the systems that we investigate, the structure of the DAG is not trivial, i.e. its shape is far from, e.g., the shape of a regular hierarchical graph in which the level sizes are increasing as a power-law with the level depth. In figure A.1(a), we show a schematic illustration of the level sizes for the networks under study, where the width of the bars indicates the number of tags on a given level, while the vertical position of the bar corresponds to l. This representation shows significant differences between the three DAGs. However, when switching to the rescaled level value $\tilde{l}$, according to figure A.1(b), the shape of the DAGs becomes more or less uniformly 'triangular'. (Since $\tilde{l}$ can take up real values instead of integers, we used binning similarly to the case of figure 7 in the main text.)

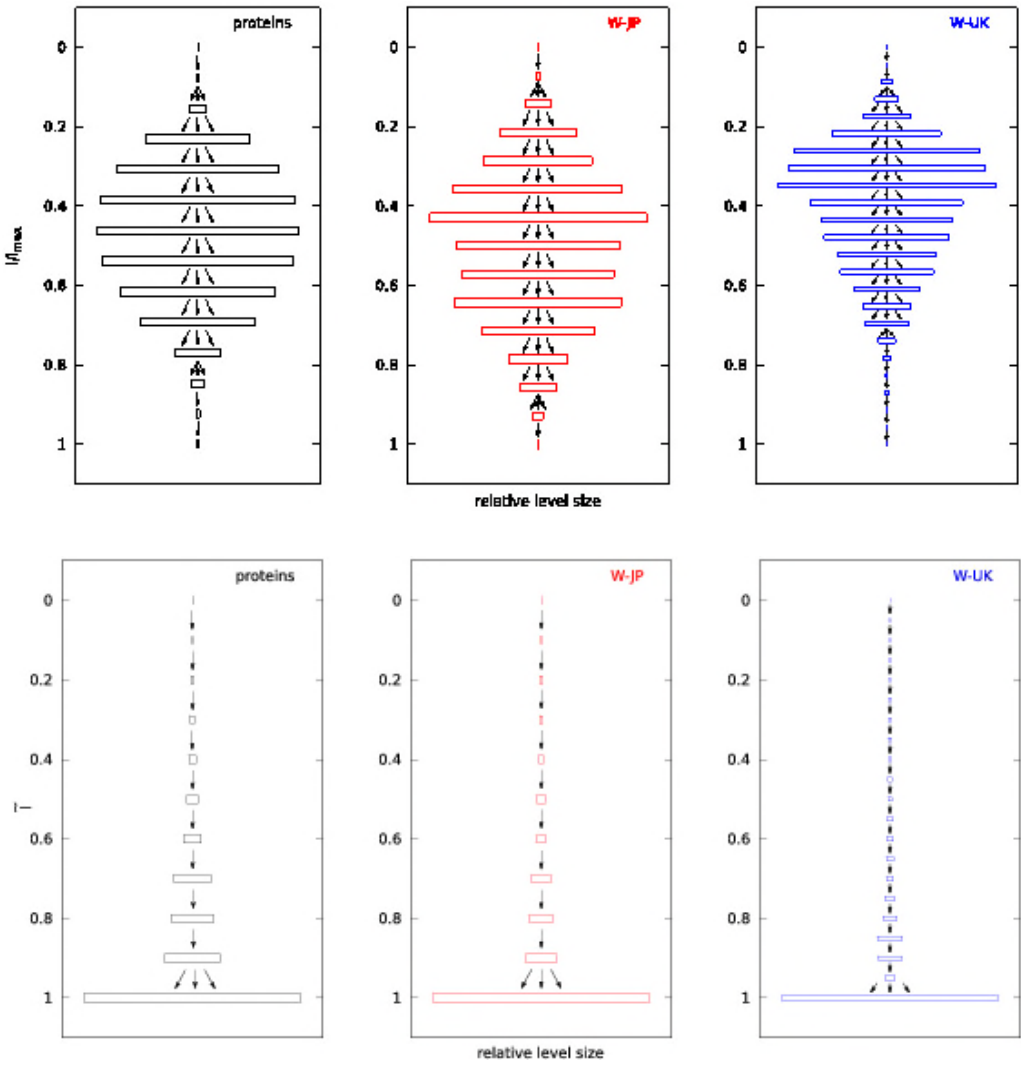

Figure A.1. (a) Schematic illustration of the DAGs in the tagged networks we investigate. The width of the bars corresponds to the number of tags at the given level. (b) After switching to the rescaled level value $\tilde{l}$ the shape of the DAGs becomes rather uniform.

A.2. Robustness of the random walk model

As mentioned in the main text, the random walk model turned out to be quite robust against changes in the details such as the frequency distribution of the tags, the distribution of the number of tags on the objects or the length distribution of the random walk on the DAG. For illustration, in figure A.2, we show the results for replacing the uniform distribution of the random walk lengths in figure 10 in the main text by γ-distribution (first row), uniform distribution with different ranges (second row), lognormal distribution (third row) and Poisson distribution (fourth row). Apparently, the qualitative behaviour of the 2D tag-distance distribution is the same as before: the maximum is shifted close to the origin in the z-score.

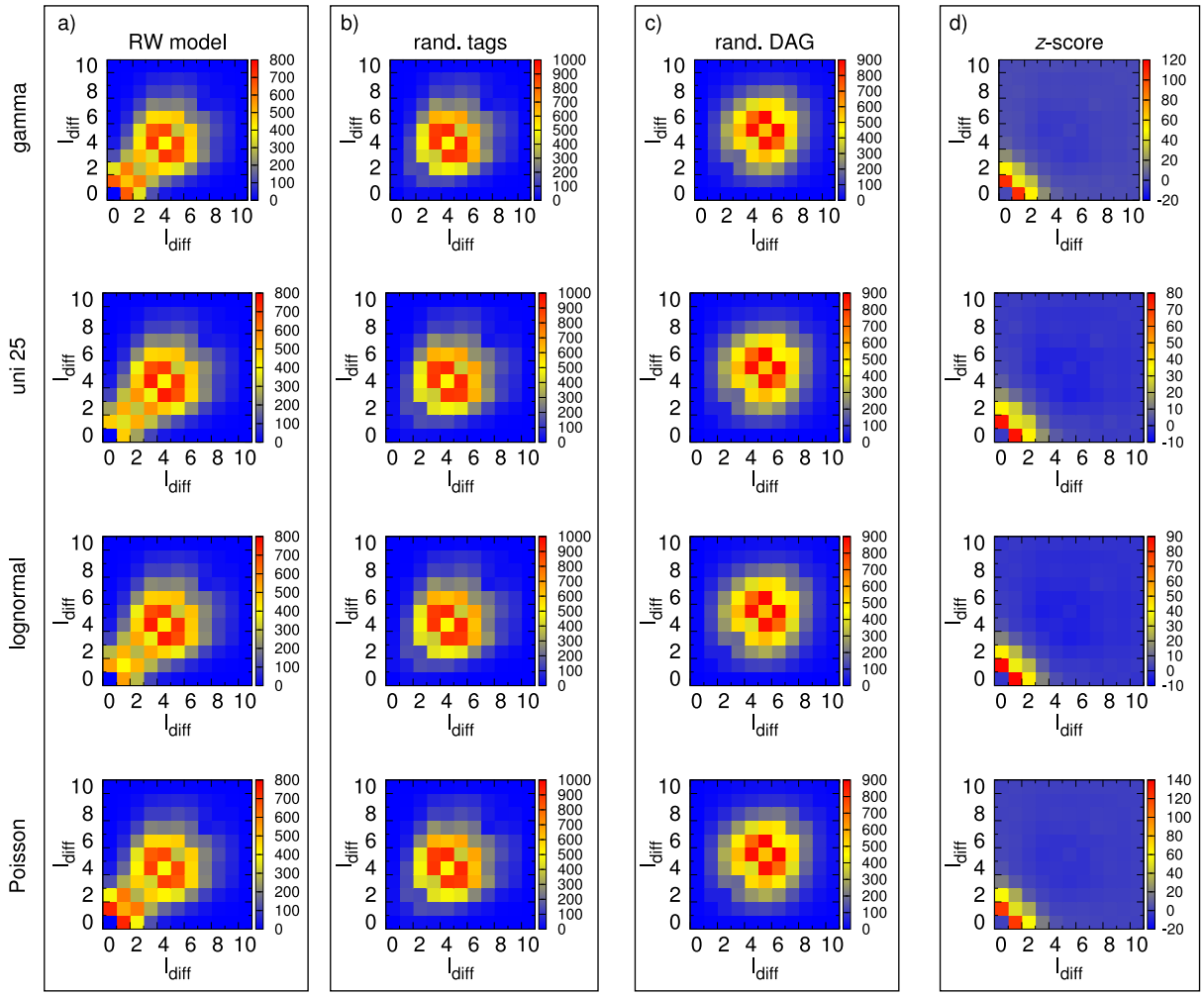

Figure A.2. The 2D tag-distance distribution in the random walk model when changing the walk length distribution to γ-distribution (first row), uniform distribution (second row), lognormal-distribution (third row) and Poisson-distribution (fourth row) for the DAGs taken from the protein interaction network. Similarly to figure 10, besides the actually measured values (panel (a)), for comparison the results for un-correlated tag assignment (panel (b)), the results for random DAGs (panel (c)) and the z-scores (panel (d)) are also shown.

References

[1] Albert R and Barabási A-L 2002 Statistical mechanics of complex networks *Rev. Mod. Phys.* **74** 47–97

[2] Mendes J F F and Dorogovtsev S N 2003 *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford: Oxford University Press)

[3] Watts D J and Strogatz S H 1998 Collective dynamics of 'small-world' networks *Nature* **393** 440–2

[4] Faloutsos M, Faloutsos P and Faloutsos C 1999 On power-law relationships of the internet topology *Comput. Commun. Rev.* **29** 251–62

[5] Barabási A-L and Albert R 1999 Emergence of scaling in random networks *Science* **286** 509–12

[6] Girvan M and Newman M E J 2002 Community structure in social and biological networks *Proc. Natl Acad. Sci. USA* **99** 7821–6
[7] Palla G, Derényi I, Farkas I and Vicsek T 2005 Uncovering the overlapping community structure of complex networks in nature and society *Nature* **435** 814–8
[8] Fortunato S 2010 Community detection in graphs *Phys. Rep.* **486** 75–174
[9] Mason O and Verwoerd M 2007 Graph theory and networks in biology *IET Syst. Biol.* **1** 89–119
[10] Zhu X, Gerstein M and Snyder M 2007 Getting connected: analysis and principles of biological networks *Genes Dev.* **21** 1010–24
[11] Aittokallio T and Schwikowski B 2006 Graph-based methods for analysing networks in cell biology *Briefings Bioinformatics* **7** 243–55
[12] Finocchiaro G, Mancuso F M, Cittaro D and Muller H 2007 Graph-based identification of cancer signaling pathways from published gene expression signatures using PubLiME *Nucl. Acids. Res.* **35** 2343–55
[13] Jonsson P F and Bates P A 2006 Global topological features of cancer proteins in the human interactome *Bioinformatics* **22** 2291–7
[14] Jonsson P F, Cavanna T, Zicha D and Bates P A 2006 Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis *BMC Bioinformatics* **7** 2
[15] Zimmermann M G, Eguíluz V M and Miguel M S 2004 Coevolution of dynamical stats and interactions in dynamic networks *Phys. Rev.* E **69** 065102
[16] Eguíluz V M, Zimmermann M G and Cela-Conde C J 2005 Cooperation and the emergence of role differentiation in the dynamics of social networks *Am. J. Sociol.* **110** 977–1008
[17] Kossinets G and Watts D J 2006 Empirical analysis of an evolving social network *Science* **311** 88–90
[18] Ehrhardt G C M A and Marsili M 2006 Phenomenological models of socioeconomic network dynamics *Phys. Rev.* E **74** 036106
[19] Holme P and Newman M E J 2006 Nonequilibrium phase transition in the coevolution of networks and opinions *Phys. Rev.* E **74** 056108
[20] Gil S and Zanette D H 2006 Coevolution of agents and networks: opinion sperading and community disconnection *Phys. Lett.* A **356** 89–94
[21] Vazquez F, González-Avella J C, Eguíluz V M and Miguel M S 2007 Time-scale competition leading to fragmentation and recombination transitions in the coevolution of network and states *Phys. Rev.* E **76** 046120
[22] Vazquez F, Eguíluz V M and Miguel M S 2008 Generic absorbing transition in coevolution dynamics *Phys. Rev. Lett.* **100** 108702
[23] Kozma B and Barrat A 2008 Consensus formation on adaptive networks *Phys. Rev.* E **77** 016102
[24] Benczik I J, Benczik S Z, Schmittmann B and Zia R K P 2008 Lack of consensus in social systems *Europhys. Lett.* **82** 48006
[25] Castellano C, Fortunato S and Loreto V 2009 Statistical physics of social dynamics *Rev. Mod. Phys.* **81** 591–646
[26] Cattuto C, Loreto V and Pietronero L 2007 Semiotic dynamics and collaborative tagging *Proc. Natl Acad. Sci. USA* **104** 1461–4
[27] Lambiotte R and Ausloos M 2006 Collaborative tagging as a tripartite network *Lect. Notes Comput. Sci.* **3993** 1114–7
[28] Cattuto C, Barrat A, Baldassarri A, Schehr G and Loreto V 2009 Collective dynamics of social annotation *Proc. Natl Acad. Sci. USA* **106** 10511–5
[29] Ghosal G, Zlatić V, Caldarelli G and Newman M E J 2009 Random hypergraphs and their applications *Phys. Rev.* E **79** 066118
[30] Zlatić V, Ghosal G and Caldarelli G 2009 Hypergraph topological quantities for tagged social networks *Phys. Rev.* E **80** 036118

[31] Schifanella R, Barrat A, Cattuto C, Markines B and Menczer F 2010 Folks in folksonomies: social link prediction from shared metadata *WSDM'10, Proc. 3rd ACM Int. Conf. on Web Search and Data Mining* pp 271–80
[32] Aiello L M, Barrat A, Cattuto C, Ruffo G and Schifanella R 2010 Link creation and profile alignment in the aNobii social network *Proc. 2nd IEEE Int. Conf. on Social Computing, SocialCom 2010* pp 249–56
[33] Mika P 2005 Ontologies are us: a unified model of social networks and semantics *Int. Semantic Web Conf.* vol 3729 pp 522–36
[34] Spyns P, Moor A D, Vandenbussche J and Meersman R 2006 From folksologies to ontologies: how the twain meet *Proc. OTM Conf. (1)* pp 738–55
[35] Voss J 2007 Tagging, folksonomy and co-renaissance of manual indexing? arXiv:cs/0701072v2
[36] Plangprasopchok A and Lerman K 2009 Constructing folksonomies from user-specified relations on Flickr *Proc. World Wide Web Conf.* pp 781–90
[37] Plangprasopchok A, Lerman K and Getoor L 2011 A probabilistic approach for learning folksonomies from structured data *Fourth ACM Int. Conf. on Web Search and Data Mining (WSDM)* pp 555–64
[38] Schmitz P 2006 Inducing ontology from Flickr tags *Collaborative Web Tagging: Proc. 15th Int. Conf. on World Wide Web (WWW)*
[39] Heymann P and Garcia-Molina H 2006 Collaborative creation of communal hierarchical taxonomies in social tagging systems *Technical Report* Stanford InfoLab
[40] Van Damme C, Hepp M and Siorpaes K 2007 Folksontology: an integrated approach for turning folksonomies into ontologies *Soc. Netw.* **2** 57–70
[41] Palla G, Farkas I J, Pollner P, Derényi I and Vicsek T 2008 Fundamental statistical features and self-similar properties of tagged networks *New J. Phys.* **10** 123026
[42] Pollner P, Palla G and Vicsek T 2010 Clustering of tag-induced subgraphs in complex networks *Physica* A **389** 5887–94
[43] Mewes H W *et al et al* 2008 MIPS: analysis and annotation of genome information in 2007 *Nucl. Acids Res.* **36** D196–201
[44] The Gene Ontology Consortium 2000 Gene ontology: tool for the unification of biology *Nature Genet.* **25** 25–9
[45] Zlatić V, Božičević M, Štefančić H and Domazet M 2006 Wikipedias: collaborative web-based encyclopedias as complex networks *Phys. Rev.* E **74** 016115
[46] Capocci A, Servedio V D P, Colaiori F, Buriol L S, Donato D, Leonardi S and Caldarelli G 2006 Preferential attachment in the growth of social networks: the internet encyclopedia wikipedia *Phys. Rev.* E **74** 036116
[47] Capocci A, Rao F and Caldarelli G 2008 Taxonomy and clustering in collaborative systems: the case of the on-line encyclopedia wikipedia *Europhys. Lett.* **81** 28006
[48] Karrer B and Newman M E J 2009 Random acyclic networks *Phys. Rev. Lett.* **102** 128701
[49] Borge-Holthoefer J and Arenas A 2010 Categorizing words through semantic memory navigation *Eur. Phys. J.* B **74** 265