New Journal of Physics
The open-access journal for physics

Fundamental statistical features and self-similar properties of tagged networks

Gergely Palla[1,3], Illés J Farkas[1], Péter Pollner[1], Imre Derényi[2] and Tamás Vicsek[1,2]

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

New Journal of Physics **10** (2008) 123026 (20pp)
Received 2 October 2008
Published 18 December 2008
Online at http://www.njp.org/
doi:10.1088/1367-2630/10/12/123026

Abstract. We investigate the fundamental statistical features of tagged (or annotated) networks having a rich variety of attributes associated with their nodes. Tags (attributes, annotations, properties, features, etc) provide essential information about the entity represented by a given node, thus, taking them into account represents a significant step towards a more complete description of the structure of large complex systems. Our main goal here is to uncover the relations between the statistical properties of the node tags and those of the graph topology. In order to better characterize the networks with tagged nodes, we introduce a number of new notions, including tag-assortativity (relating link probability to node similarity), and new quantities, such as node uniqueness (measuring how rarely the tags of a node occur in the network) and tag-assortativity exponent. We apply our approach to three large networks representing very different domains of complex systems. A number of the tag related quantities display analogous behaviour (e.g. the networks we studied are tag-assortative, indicating possible universal aspects of tags versus topology), while some other features, such as the distribution of the node uniqueness, show variability from network to network allowing for pin-pointing large scale specific features of real-world complex networks. We also find that for each network the topology and the tag distribution are scale invariant, and this self-similar property of the networks can be well characterized by the tag-assortativity exponent, which is specific to each system.

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

Contents

1. Introduction

Many complex systems in nature and society can be successfully represented in terms of *networks* capturing the intricate web of connections among the units they are made of [1, 2]. In recent years, the research in this field has been focused mainly on the *topology* of the graphs corresponding to these real networks. Since this approach is rooted in, among other things, statistical physics, where often the thermodynamic limit is considered and also the size of the known nets becomes huge, several *large-scale* properties of real-world webs have been uncovered, e.g. a low average distance combined with a high average clustering coefficient [3], the broad (scale-free) distribution of node degree (number of links of a node) [4]–[7] and various signatures of hierarchical/modular organization [8, 9].

On the other hand, there has been a quickly growing interest in the *local structural units* of networks. Small and well defined sub-graphs consisting of a few vertices have been introduced as *motifs* [10, 11], whereas somewhat larger units, associated with more highly interconnected parts [12]–[26] are usually called *communities*, clusters, cohesive groups, or modules. These structural sub-units can correspond to multi-protein functional units in molecular biology [8, 27], a set of tightly coupled stocks or industrial sectors in economy [28, 29], groups of people [19, 30, 31], cooperative players [32]–[34], etc. The location of such building blocks can be crucial to the understanding of the structural and functional properties of the systems under investigation.

The majority of the complex network studies concern 'bare' graphs corresponding to a simple list of connections between the nodes, or at most weighted networks where a connection strength (or intensity) is associated to the links. However, the introduction 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 of the systems under investigation. These tags can correspond to basically any information about the nodes and in most cases a single node can have several tags at the same time. The use of such annotations in biological networks is a common practice [35]–[40], where the tags usually refer to the biological function of the units (proteins, genes, etc). Another interesting application of node features can be seen in the studies of co-evolving network models, where the evolution of the network topology affects the node properties and vice versa [41]–[50]. These models are aimed at describing the dynamics of social networks, in which people with similar opinion are assumed to form ties more easily, and the opinion of connected people becomes more similar over time. Finally, we mention the study of collaborative tagging in [51], where tripartite networks were constructed from data concerning users who associated tags to some kinds of items, (such as music listeners classifying music records). The three types of nodes corresponded to the users, the tags, and the items. The tagging was carried out without any central authority and according to the results, the analysis of the bi- and unipartite projection of the networks can help in structuring the contents (e.g. define a hierarchy between the tags).

In this paper, we study tagged networks from yet another point of view. Our focus is on networks where the links are in principle not related to tagging, however tags can be associated with the nodes quite naturally. The PACS numbers or key-words in the case of co-authorship networks, the scope of business or the industrial sector of companies in the context of financial networks, or the status of employees in the case of a network representing the social ties inside a large firm provide plausible examples for possible tags. The complexity of the networks studied these days is rapidly increasing together with their size. The use of tags associated with the nodes can help in revealing hidden structures or accelerating searching within the networks. Since the usefulness of such attributes has already been proven in biology, the inclusion of tags in the analysis of other networks as well is expected to give a deeper understanding of the interrelations shaping the structure and dynamics of the systems under study.

Along this line, in the present paper we study the fundamental statistics characterizing the distribution of tags in large annotated real networks. By choosing networks representing completely unrelated systems (a co-authorship network, a protein interaction network, and the English Wikipedia), we look for signs of universality in these statistics. Furthermore, we are interested in the relations between the network topology and the distribution of tags. The tags enable the definition of a *similarity function* between the nodes that is *a priori* independent of the topology. We shall refer to this quantity as the *tag-similarity* of the nodes in order to distinguish it from the usual structural similarity of the nodes (based on the similarity between the nearest neighbours). Study of the tag-similarity opens up further directions for exploring the intricate relations between the annotations and the graph structure itself. Interestingly, in all selected systems, the tags form a sort of *taxonomy*: they correspond to features ranging from very specific to rather general ones, which are embedded in a hierarchic structure held together by 'is a sub-category of' type relations. This inter-relatedness of the tags adds an extra twist to the definition of the quantities we study.

The paper is organized as follows. In section 2. we define the most important quantities we aim to study, whereas the construction of the investigated networks (and the hierarchy of the corresponding node labels) is detailed in section 3. The results are presented in section 4 and we close the paper with some concluding remarks in section 5.

2. Definitions

2.1. Basic statistics

2.1.1. Number of tags on a node. In principle, nodes in a network can be tagged with almost anything. Here, we list a few basic types followed by particular examples in parenthesis: real numbers (the accumulated impact factors of authors in a co-authorship network), integers (the number of articles of an author), or character strings (functions of proteins in a protein–protein interaction (PPI) network). However, in most cases, (including the systems we study in the present paper), the node attributes correspond to *character strings*, chosen from a finite set of possible tags. Usually a node can have more than one tag attached to it, e.g. numerous proteins appearing in a PPI network have multiple functions. One of the basic statistics about the annotations is the distribution of the number of tags on the nodes.

2.1.2. Tag frequencies. Similarly to the varying number of tags on the nodes, the *frequency* of the different tags can also be rather heterogeneous. What makes the picture even more complex is that in many cases the tags refer to *categories* of a *taxonomy* or *ontology* (capturing the view of a certain domain, e.g. protein functions). This means that the tags are organized into a structure of relationships which can be represented by a directed acyclic graph (DAG), where the directed links between two categories represent an 'is a sub-category of' relation. The nodes close to the root in the DAG are usually related to general properties, and as we follow the links towards the leafs, 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 part of more than one category (that are not parts of one another). Also note that nodes can be classified not only by the leaf-categories e.g. several proteins in a PPI network can be found with rather general functional descriptions. We illustrate the concept of tagged networks and the corresponding DAG of categories in figure 1 with the help of the English Wikipedia.

Given the DAG between the possible tags, we can define the frequency of a given tag α in two different ways:

$$p_\alpha \equiv N_\alpha / N, \tag{1a}$$

$$\tilde{p}_\alpha \equiv \tilde{N}_\alpha / N, \tag{1b}$$

where N_α denotes the number of nodes tagged with α, $\tilde{N}_\alpha$ stands for the number of nodes tagged with α or any of its descendents, and N is equal to the total number of nodes in the network. From these definitions it follows that when the number of untagged nodes is zero, the root of the annotation DAG will receive $\tilde{p}_\alpha = 1$, whereas for the leaf categories $\tilde{p}_\alpha = p_\alpha$. Furthermore, if category β is a descendent of α, then $\tilde{p}_\alpha \geqslant \tilde{p}_\beta$. Low frequency tags are more specific in an information theoretical sense, whereas high frequency tags carry almost no information (e.g. being tagged by the root in the annotation DAG adds absolutely no information to the description of a node).

In the following, we shall refer to the sub-graph induced by the nodes (i.e. constituted by these nodes and all links between them) marked by the tag α and any of its descendents as the *tag-induced sub-graph* of α. The number of nodes in this sub-graph is given by $\tilde{N}_\alpha$, whereas the number of links can vary between $\tilde{M}_\alpha = 0$ and $\tilde{M}_\alpha = \tilde{N}_\alpha(\tilde{N}_\alpha - 1)/2$. It is interesting to compare $\tilde{M}_\alpha$ to the number of links $\tilde{M}_{\text{rand}}$ one would expect in a random sub-graph of the same size: if $\tilde{M}_\alpha$ is significantly larger/smaller than $\tilde{M}_{\text{rand}}$, then nodes sharing the tag α attract/repel each other in the sense that they are linked with higher/smaller probability than at random.

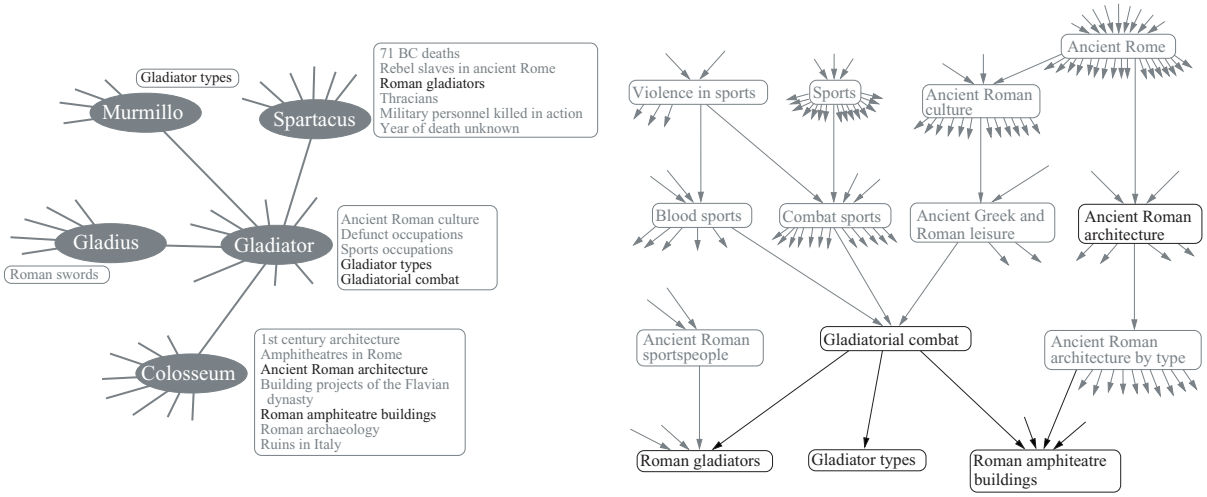

Figure 1. A small labelled sub-graph and the corresponding DAG of categories in the English Wikipedia. In the left panel we show a few neighbours of the page 'Gladiator', where the connections correspond to mutual hyperlinks between the pages embedded in the text of the page. At the bottom of each page we can find a list of categories, which we use as tags. These are listed in the frames appearing near the nodes. These categories are organized into a DAG, as demonstrated in the right panel, where e.g. 'Gladiator types' is a sub-category of 'Gladiatorial combat'. The categories appearing in both panels are emphasized in black.

2.2. Tag-similarity

Our aim in this section is to define a similarity function between the nodes which is based solely on the tags, therefore, it can be evaluated without any knowledge about the graph structure. Although we refer to this quantity as the tag-similarity of the nodes in general, we shall use the term similarity in the same sense for short.

2.2.1. Simple similarity measures. To what extent two nodes i and j having a set of tags Ω_i and Ω_j are similar is a far from trivial question, as the number of possible similarity measures is vast. A simple approach is to use the Jaccard-similarity [52] defined as

$$s_{ij}^{(\mathrm{J})} \equiv \frac{|\Omega_i \cap \Omega_j|}{|\Omega_i \cup \Omega_j|}, \tag{2}$$

where $|\Omega_i \cap \Omega_j|$ is equal to the number of common tags and $|\Omega_i \cup \Omega_j|$ is equal to the total number of different tags in Ω_i and Ω_j. Another possibility is to represent the annotations as vectors $\mathbf{v}_i$ and $\mathbf{v}_j$, where the number of entries in the vectors is equal to the number of different tags in the network, and the nonzero elements indicate the presence (or possibly the weight) of the actual tags on the given node. In this approach the cosine similarity

$$s_{ij}^{(\mathrm{c})} \equiv \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{|\mathbf{v}_i||\mathbf{v}_j|} \tag{3}$$

yields a simple similarity value for a pair of nodes i and j.

The advantage of the above methods is that they do not depend on the DAG between the tags, therefore, they can be applied even when the tags are not part of a structured taxonomy. However, when a tag refers to a sub-category of another tag, the similarity measure should be refined. As an example, let us assume that node i is tagged exclusively with category α, and node j has a single tag β, that is a direct descendent of α (e.g. $\alpha \equiv$'knife' and $\beta \equiv$'kitchen knife'). In this case both (2) and (3) yield $s_{ij}^{(\mathrm{J})} = s_{ij}^{(\mathrm{c})} = 0$, which is not what we would expect.

2.2.2. Semantic similarities. To overcome the problem raised above, we should use a similarity measure which takes into account the structure of the annotation DAG. At this point we divide the evaluation of similarity into two parts: first we deal with the similarity $s_{\alpha\beta}$ between a pair of tags, then elaborate on how to combine the pairwise similarities $s_{\alpha\beta}$, $\alpha \in \Omega_i$, $\beta \in \Omega_j$ between the sets of tags Ω_i, Ω_j associated with a pair of nodes i and j to obtain s_{ij}.

A simple choice for determining the similarity between two tags is the length of the longest shared path towards the root of the annotation DAG. A somewhat more sophisticated approach is to use *semantic similarities*. The basic idea behind these methods is to take into account the frequency of the tags: sharing a rare tag by two nodes should indicate high similarity, whereas sharing a frequent tag should not. The semantic similarity between tags α and β is derived by Resnik [53] as

$$s_{\alpha\beta}^{(\mathrm{R})} \equiv \max_{\gamma \in \Gamma(\alpha,\beta)} \left[-\log \tilde{p}_\gamma \right], \tag{4}$$

where $\Gamma(\alpha, \beta)$ denotes the set of common ancestors of α, β, and $-\log\tilde{p}_\gamma$ corresponds to the information content of category γ. From this definition it follows that if β is a descendent of α, then $s_{\alpha\beta}^{(\mathrm{R})} = -\log\tilde{p}_\alpha$, and when the two compared tags are not connected by a directed path, then $s_{\alpha\beta}^{(\mathrm{R})}$ is equal to the information content of one of their nearest common ancestors. A closely related similarity measure was proposed by Lin [54] as

$$s_{\alpha\beta}^{(\mathrm{L})} \equiv \frac{2 \max_{\gamma \in \Gamma(\alpha\beta)} [-\log \tilde{p}_\gamma]}{|\log \tilde{p}_\alpha + \log \tilde{p}_\beta|}. \tag{5}$$

In practice (5) was reported to slightly underperform (4) [55], however the big advantage of (5) is that $s_{\alpha\beta}^{(\mathrm{L})}$ becomes bounded in [0, 1]. The maximal possible $s_{\alpha\beta}^{(\mathrm{R})}$ obtained from (4) depends on the frequency of the rarest tag, which in our case is strongly varying from system to system. For this reason, we shall use (5) for calculating the similarity between categories.

When moving from the similarity of tags to the similarity of nodes, again we have a number of possibilities to choose from. A simple approach is to use the average of the pairwise similarities as

$$s_{ij} \equiv \frac{1}{n_i n_j} \sum_{\alpha \in \Omega_i, \beta \in \Omega_j} s_{\alpha\beta}^{(\mathrm{L})}, \tag{6}$$

where n_i and n_j denote the number of tags on nodes i and j, respectively. The problem with the expression above is that if the labels associated with a given node are very different from each other, then by comparing this node even to itself, the 'cross-terms' reduce the similarity value. A simple solution is to replace the average in (6) by the maximal pairwise similarity amongst the tags:

$$s_{ij} \equiv \max_{\alpha \in \Omega_i, \beta \in \Omega_j} s_{\alpha\beta}^{(\mathrm{L})}. \tag{7}$$

Another possibility along this line is to organize the pairwise similarities between the tags into an n_i by n_j matrix, and define the quantities $rowScore$ and $columnScore$ as the average of the maximal values in the rows and columns of this matrix, respectively. The similarity between the two annotation vectors can then be given as either the average or the maximum of $rowScore$ and $columnScore$ [56]. In our studies we shall use (7) due to its computational simplicity and the fact that it is analogous to the concept of minimum linkage clustering, where the distance between two sets of elements (the tags) is defined as the minimum pairwise distance between the elements.

2.3. Tag-assortativity

A plausible hypothesis about tagged real networks is that links are likely to form ties between similar nodes and vice versa, we expect connected nodes to share common tags with enhanced probability. However, this property is not evident in all cases. For example, if we colour the nodes in a network according to the famous vertex colouring problem [57], (namely we look for the minimal number of colours which can be distributed in such a way that no neighbours have the same colour), and identify the node colours as the tags, then similar nodes are actually never connected.

In general, the property that nodes are more frequently connected to others that are similar/different in some quality is referred to as assortativity/disassortativity. The most typically considered quality—which is based on the network's topology—is the degree of the nodes. In tagged networks, however, another natural way of comparing nodes can be based on the above defined tag-similarity. We can thus introduce the notion of *tag-assortativity* (to distinguish this property from the degree-assortativity), and call a network *tag-assortative*/*tag-disassortative* if nodes having similar tags are linked with higher/lower probability than at random.

2.4. Uniqueness

Interestingly it is not uncommon to find tags associated with the same node which are rather different from each other, e.g. in the PPI network studied in this paper more than 10% of the nodes have at least one pair of tags for which the nearest common ancestor is actually the root of the annotation DAG. This means that the given protein can take part in very different biological processes. On the other hand, many nodes have more or less similar categories in their annotation, so they take part in more or less similar processes.

To quantify the above aspect, we introduce the *node uniqueness*, defined as

$$u_i \equiv \min_{\alpha,\beta\in\Omega_i} s^{(\mathrm{R})}_{\alpha\beta}. \tag{8}$$

In principle, we could have chosen $s^{(\mathrm{L})}_{\alpha\beta}$ rather than $s^{(\mathrm{R})}_{\alpha\beta}$ in the definition above. However, since $s^{(\mathrm{L})}_{\alpha,\alpha} = 1$ for every α, if node i has only a single tag, then u_i would be unity independent of whether this tag is frequent or not. By using the Resnik-similarity (4) for which $s^{(\mathrm{R})}_{\alpha,\alpha} = -\log\tilde{p}_\alpha$, we can differentiate between nodes with single tags as well, based on the tag frequencies. The lowest possible value for u occurs in the case where a node belongs to more than one category, out of which at least two have the root of the DAG as their nearest common ancestor. The highest

possible value for u occurs if a node belongs to a single category, and this category happens to be the rarest among all. We note that in [51] a closely related quantity called node diversity was defined for the case where the tags are not part of a hierarchical taxonomy.

3. Applications

We studied the node annotations in three networks of high importance from the aspect of practical applications, capturing the relations between interacting proteins, collaborating scientists, and pages of an on-line encyclopedia. The PPI network of MIPS [58] contained $N = 4546$ 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 [59].

The investigated co-authorship network is known as the MathSciNet (Mathematical review collection of the American Mathematical Society) [60], and represents the $M = 873\,775$ links of collaboration between $N = 391\,529$ mathematicians. The node tags were obtained from the 6499 different subject classes of the articles, which were organized into a DAG. Thus, the set of tags attached to each author was the union of all subject-classes that appeared on her/his papers.

Finally, the nodes in the third studied network corresponded to the $N = 1473\,894$ pages of the English Wikipedia [61]–[64], connected by the $M = 3755\,485$ hyperlinks embedded in the text of the pages. At the bottom of each page, one can find a list of categories, which were used as node tags. Since each wiki-category is a page in the Wikipedia as well, we removed these pages from the network to keep a clear distinction between nodes and attributes. Furthermore, we kept only the mutual links between the remaining pages. Similarly to the biological processes in the MIPS network or the subject classes in the MathSciNet, the wiki-categories can have sub-categories and are usually part of a larger wiki-category. However, when representing these relations as a directed graph, some directed loops appear, therefore, they do not form a strict DAG as required for e.g., the semantic similarity measures (4) and (5). In order to be able to use these similarity functions, we removed a few relations from this graph until it turned into a DAG, following a method detailed in the appendix.

Due to the very large size of this network, some of the analysis we carried out turned out to be very time consuming, therefore, in certain cases we used only smaller sub-graphs of Wikipedia, induced by rather general categories e.g. ‘Soccer’, ‘Japan’, etc. (The tags which were not descendents of the chosen category were naturally dropped from the nodes in the tag-induced sub-graph.) The advantage of this method is that the categories appearing as node tags in the resulting sub-graph also form a DAG which is equivalent to the DAG of the descendents of α (in which the root is α). In this paper, we show the results for the case where $\alpha \equiv$‘Japan’, (altogether $N = 43\,307$ nodes, $M = 102\,753$ links and 3197 sub-categories), however other choices resulted in very similar results as well.

We also checked whether this sort of sampling from the networks distorts the studied statistics by examining tag-induced sub-graphs in the other two networks (and smaller tag-induced sub-graphs in the Wikipedia/Japan network) as well. We found that for all statistics studied in this paper the results in a large enough tag-induced sub-graph are very similar to those for the whole network, and the differences can be mostly attributed to the different system sizes.

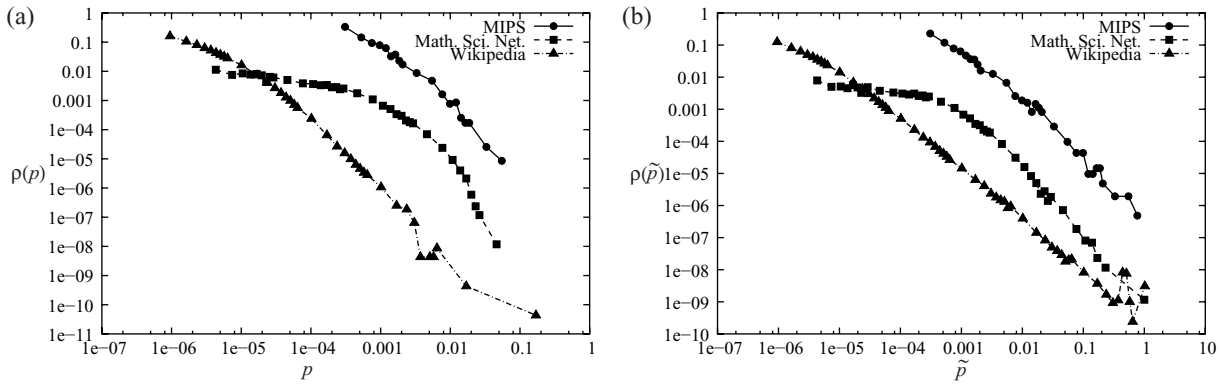

Figure 2. The density distributions of the tag frequencies (a) p_α and (b) $\tilde{p}_\alpha$ on logarithmic scale.

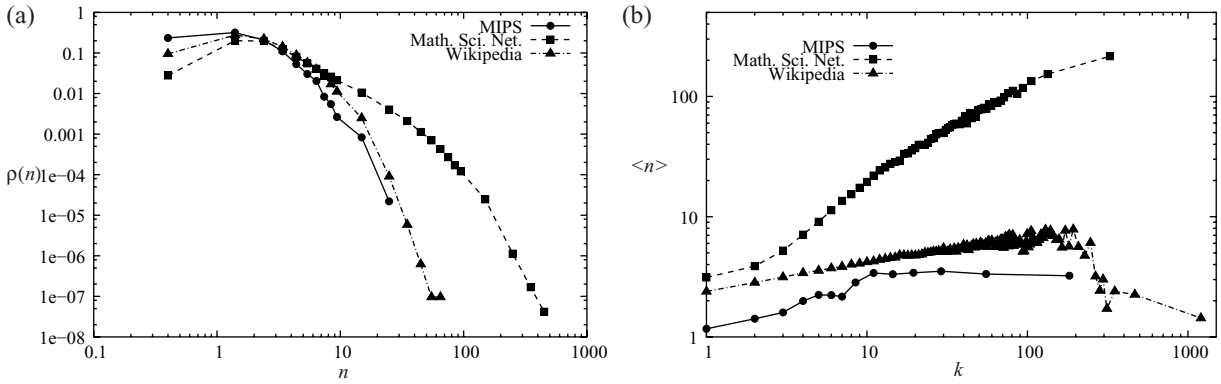

Figure 3. (a) The density distributions of the number of tags n per node. (b) The average number of tags $\langle n \rangle$ as a function of the node degree k.

4. Results

4.1. Basic statistics

We begin our investigations in figure 2 with the distribution of the tag frequencies in the three networks. According to figure 2(a), the distribution of p_α resembles a power-law for the MIPS network and Wikipedia, whereas it resembles an exponential distribution for the MathSciNet. When moving from p_α to $\tilde{p}_\alpha$ (by including the nodes tagged with any descendents of α as well), the tail of the distribution becomes power-law like for each network, as shown in figure 2(b). This is consistent with the hierarchical nature of the annotation DAG: categories high up in the DAG correspond to general concepts, therefore apply to a vast number of nodes, whereas leaf categories (without any descendents) refer to something specific, therefore occur rarely [65]–[67].

Our main goal in this paper is to study the relations between the distribution of node tags and the network topology. One of the most basic statistical quantity which can be studied in this respect is the number of n_i tags for each node i. In figure 3(a), we display the density distribution of n_i in the studied systems, whereas figure 3(b) shows the average number of tags, $\langle n \rangle$ as a function of the node degree. Since the range of the possible n values is rather wide (especially

in case of the MathSciNet), we used exponentially increasing bin sizes in figure 3(a). The decay of the distributions towards large n values seems exponential. Concerning the curves shown in figure 3(b), a plausible hypothesis about tagged real networks is that they show tag-assortativity, namely links form ties between similar nodes more frequently than at random. Therefore, we expect connected nodes to share common tags with enhanced probability. Consequently hubs are expected to have a larger number of tags than nodes with small degrees, since they have to share common attributes with a large number of other nodes. Interestingly, in figure 3(b) the MathSciNet behaves as expected from this point of view (with a monotonically increasing $\langle n \rangle(k)$ curve), whereas the MIPS network and Wikipedia do not. For both networks, $\langle n \rangle(k)$ increases at small degrees, then in case of the MIPS network it saturates, whereas for Wikipedia it even drops down at high degrees. This implies that the simple picture shown above, in which the hubs correspond to versatile nodes with a large number of different tags, does not hold in these systems.

4.2. Tag-induced sub-graphs

Due to the size of the entire Wikipedia, we have been able to analyse only some of its tag-induced sub-graphs, as described in section 3. To get a better understanding of the relationship between tag distribution and network topology, it is very insightful to go further down this line, and compare some of the basic properties of the tag-induced sub-graphs for every category (all the way from the root to the leaves) in all of our three networks. The scatter plots in figure 4 with grey symbols depict the link number ($\tilde{M}$) versus node number ($\tilde{N}$) relation for each category. $\tilde{M}$ has a maximum of $\tilde{M}_{\max}(\tilde{N}) = \tilde{N}(\tilde{N}-1)/2$, when the sub-graph forms a clique, i.e. each node is linked to all the others. This upper bound is shown with a dashed–dotted line. The estimate of the number of links $\tilde{M}_{\text{rand}}(\tilde{N}) = p\tilde{N}(\tilde{N}-1)/2 = p\tilde{M}_{\max}(\tilde{N})$ between randomly selected $\tilde{N}$ nodes, is also plotted with a dashed line, where the linkage probability is defined as $p = M/[N(N-1)/2]$. According to the scatter plots, in all the three systems the number of links $\tilde{M}$ in every tag-induced sub-graph (with some exception at $\tilde{M} = 0$) exceeds the number of links $\tilde{M}_{\text{rand}}$ expected for a link distribution that is uncorrelated to the tag distribution. This strongly indicates that the networks under study are *tag-assortative*.

An even more intriguing property of the scatter plots is that if the average number of links $\langle \tilde{M} \rangle$ are plotted (with black symbols) as a function of the number of nodes $\tilde{N}$ (using logarithmic binning), then they strictly follow a power-law $\langle \tilde{M} \rangle \sim \tilde{N}^{\mu}$ (solid lines) for several orders of magnitude (with a deviation only at the smallest sub-graphs). The *tag-assortativity exponent* μ, defined by this power law, takes the values of 1.30 ± 0.02, 1.16 ± 0.02, and 1.18 ± 0.01 for the MIPS, the Wiki-Japan, and the MathSciNet networks, respectively. The physical meaning of this exponent can be demonstrated by considering the relation between the tag-induced sub-graph of some category and those of its sub-categories. If the tag-induced (not necessarily disjoint) sub-graphs of the sub-categories inherit all the links of the parent category 'homogeneously' and without having inter-sub-graph links (i.e. having no links between any pair of sub-graphs other than those originating in the intersection), then the number of links corresponding to a sub-category is expected to scale linearly with the number of its nodes, implying $\mu = 1$. If, however, inter-sub-graph links also appear (cf figure 5), then the number of links corresponding to a sub-category is expected to drop faster than linearly, leading to $\mu > 1$. Although $\mu < 1$ cannot be ruled out (at least locally, between a particular category and its sub-categories), it requires very peculiar topologies (e.g. large link density in the intersection between the

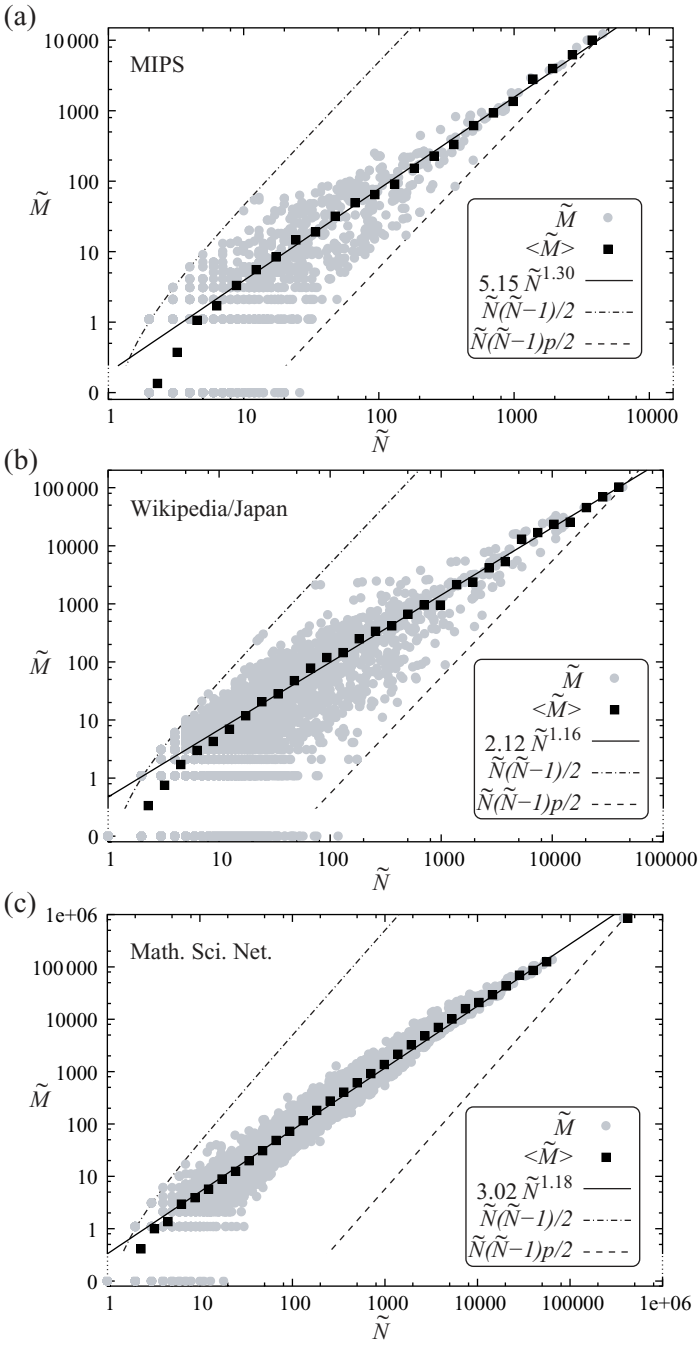

Figure 4. Scatter-plots of the number of links, $\tilde{M}$ versus the number of nodes, $\tilde{N}$ in the tag-induced sub-graphs of the different categories (grey symbols) for the MIPS network (a), the Wiki-Japan network (b) and the MathSciNet. The black symbols show $\langle\tilde{M}\rangle$, whereas the solid lines correspond to the best power-law fit to $\langle\tilde{M}\rangle(\tilde{N})$. The dot-dashed lines and the dashed lines in each plot show the upper bound in $\tilde{M}$ and the expected number of links for a randomly chosen nodes, respectively.

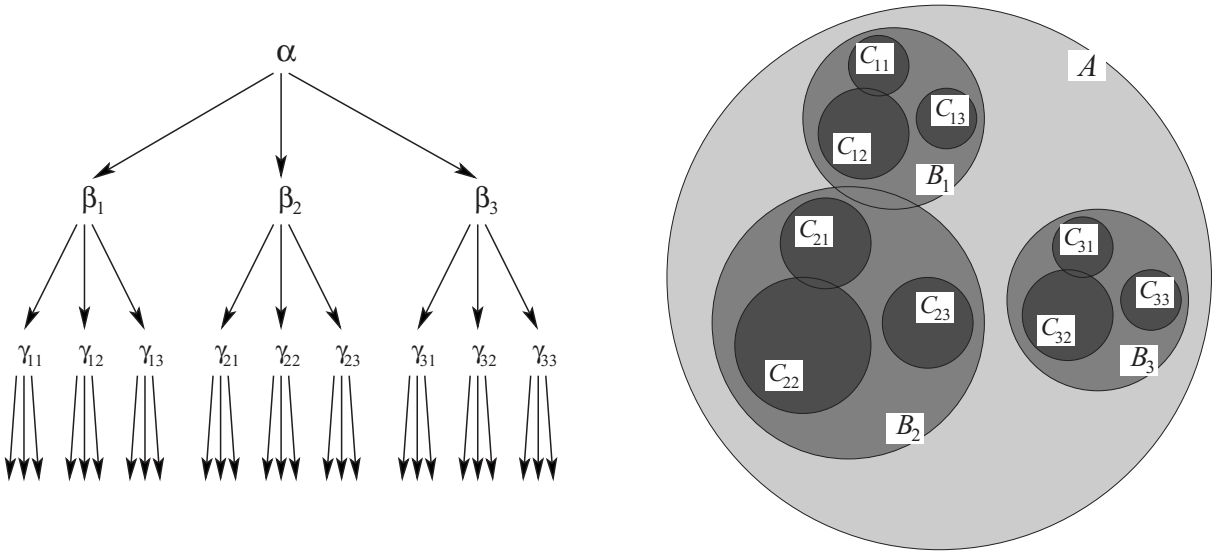

Figure 5. Demonstration of the *self-similar* nature of recursively embedded tag-induced sub-graphs $C_{ij} \subset B_i \subset A$ generated by the DAG of hierarchically organized categories $\gamma_{ij} \subset \beta_i \subset \alpha$. The grey level is indicative of the link density.

tag-induced sub-graphs of two sub-categories) and, thus, we do not anticipate obtaining such values for real systems.

In brief, a value of $\mu > 2$ indicates tag-disassortativity; $\mu = 2$ characterizes no correlation between tag-similarity and link distribution (cf $\tilde{M}_{\text{rand}}$); whereas $0 < \mu < 2$ is the regime of tag-assortativity with the amendment that $0 < \mu < 1$ would represent extreme tag-assortativity. This classification scheme affirms that the tag-assortativity exponent μ defined above is indeed an appropriate quantity for characterizing the extent of tag-assortativity. Our finding that its value for the three networks we have studied is closer to 1 than to 2 suggests that these networks exhibit a significant tag-assortativity, MIPS being somewhat less tag-assortative than the other two.

Both the fact that the statistical properties of tag-induced sub-graphs are similar to those of the entire graph and also the fact that a single well defined exponent characterizes tag-assortativity over several orders of magnitudes of the sub-graph size imply prominent *self-similarity* in the structure of tagged networks. Briefly speaking, the tag-induced sub-graph A of some category α is related to the tag-induced sub-graphs $B_i \subset A$ of its sub-categories $\beta_i \subset \alpha$ statistically the same way as the sub-graphs B_i of categories β_i to the tag-induced sub-graphs $C_{ij} \subset B_i$ of their sub-categories $\gamma_{ij} \subset \beta_i$, as demonstrated in figure 5, i.e. both the network topology and the tag distribution appear to be *scale invariant*.

4.3. Similarity

The introduction of a similarity measure based on the node tags enables us to study other type of relations between the topology and the annotations as well. In figure 6, we follow the change of the similarity between the nodes with the distance in the three networks. The right column of the figure shows the density distribution $\rho(s_{ij})$ for s_{ij} obtained from (7), whereas the left column displays the corresponding average similarity, $\langle s_{ij} \rangle$ as a function of the node distance d.

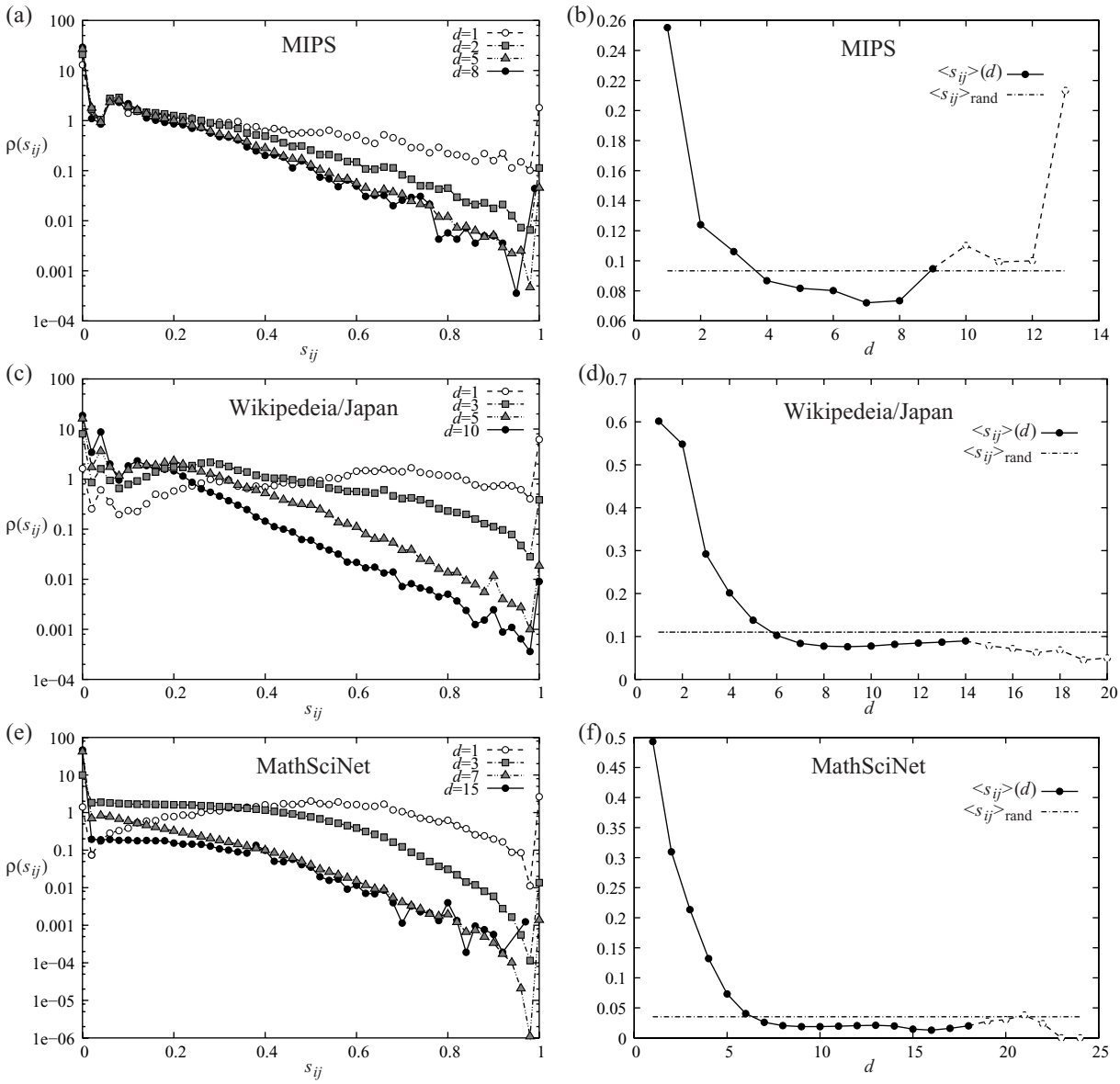

Figure 6. The similarity s_{ij} as a function of the distance between the nodes. The density distribution of s_{ij} at various distances is plotted on semi-logarithmic scale for the MIPS network, the Wiki-Japan network and the MathSciNet in panels (a), (c) and (e), respectively. The corresponding average similarity, $\langle s_{ij} \rangle$ as a function of the node distance d is shown in panels (b), (d) and (f). The number of pairs at large d becomes small, therefore, the results for $\langle s_{ij} \rangle$ in this regime cannot be trusted. The empty symbols and dashed lines indicate that the number of pairs has decreased below the total number of links in the network.

The $\rho(s_{ij})$ distributions are shifted towards lower s_{ij} values with increasing distance d between the nodes and accordingly a rapid decreasing tendency can be observed in the $\langle s_{ij} \rangle(d)$ function at small distances. At medium node distances $\langle s_{ij} \rangle$ becomes more or less constant, suggesting that the nodes become independent of each other. In consistency with the results of section 4.2,

this is another indication of *tag-assortativity*: if links were drawn between the nodes at random, the $\langle s_{ij} \rangle$ would be independent of the distance between the nodes (the $\langle s_{ij} \rangle(d)$ would resemble a flat line). The prominent peak at distance $d = 1$ signals that neighbouring nodes are much more similar to each other than at larger distances and much more similar to each other than at random as well.

At large node distances the number of pairs is rapidly decreasing (i.e. at the possible maximum distance only a few pairs of nodes can contribute to $\langle s_{ij} \rangle$). To indicate that the number of samples in this regime is not enough for a significant statistical analysis, we changed the filled symbols (and solid lines) to empty symbols (and dashed lines) in figures 6(b), (d) and (e). Interestingly, for the MIPS network $\langle s_{ij} \rangle(d)$ increases in this region, reaching a value at the maximal distance $d_{\max}$ almost as high as at $d = 1$. However, this is due the fact that the five nodes making up the pairs at $d_{\max}$ happen to be more similar to a randomly chosen node than average. (Nodes having a couple of non-specific tags can be indeed quite similar to the majority of the nodes). Since the number of pairs at large d is small, the contribution from these nodes is significant, and $\langle s_{ij} \rangle$ becomes larger than at medium d, where the vast number of other nodes counter balance this distortion.

4.4. Node uniqueness

We now move on to the investigation of the node uniqueness, defined in (8). Our main interest concerns the dependence of u on the node degree. We divide the nodes into three classes of equal size depending on their u value: *specific* nodes have relatively high u (marked by either a rare label or a few closely related rare labels), *medium* nodes have a u value around the average, whereas *diverse* nodes have a relatively low u value (marked by frequent or un-related labels). In figure 7 we show the participation ratio of the nodes in the three classes as a function of the node degree k. Again, the three systems show different behaviour. In case of the Wikipedia and the MathSciNet, the ratio of diverse nodes is increasing monotonically with the node degree. This tendency is very pronounced in the latter network (figure 7(c)), where in fact all hubs are classified as diverse above a certain degree. This is consistent with the steady increase in the average number of tags as a function of the node degree in figure 3(b) (square symbols): for nodes with $n \sim 10^2$ tags we expect to find at least a few pairs of rather un-related categories resulting in a low u value. In contrast, for the MIPS network, the monotonic increase in the ratio of diverse nodes with the node degree is followed by a sudden drop at the largest degrees. This means that a significant portion of the hubs in this network have rather specific functions.

5. Summary and conclusion

We studied the basic statistical properties of tags in real networks, with an interest in the relation between the topology and the tag distribution. We found that the investigated systems show universal features in some aspects with interesting differences from other perspectives. At small and intermediate degrees the average number of tags per node increases with the degree, and accordingly the node uniqueness decreases. For the MathSciNet this tendency is prolonged in the high degree regime as well. In contrast, the number of tags on the hubs in the MIPS network drops down and simultaneously the ratio of nodes with large uniqueness increases. The behaviour of the English Wikipedia is somewhere in between: the number of tags saturates for the hubs and the further increase in the ratio of nodes with low uniqueness is marginal.

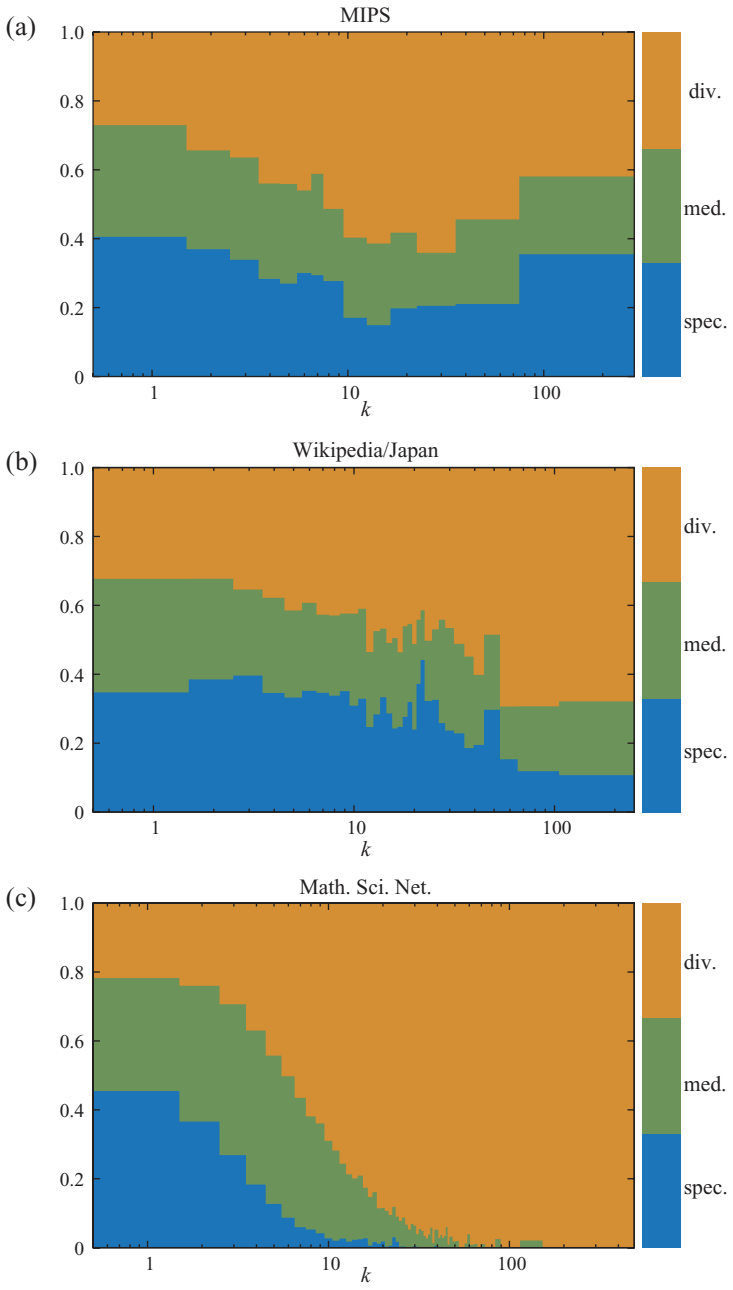

Figure 7. The participation ratio of the nodes in the three node uniqueness classes as a function of the node degree k for the MIPS network (a), the Wikipedia/Japan network (b) and the MathSciNet (c).

This comparison reflects the difference in the behaviour of hubs in these networks: the hubs of the MathSciNet are very versatile with huge amounts of different tags and low values of uniqueness, whereas in the MIPS network a significant portion of the hubs correspond to proteins with rather specific functions.

We introduced the tag-similarity of nodes, which (in contrast to the usual structural similarity) can be calculated independently of the graph topology, and is based on the set of

tags associated with the nodes. According to our results, the studied real networks show tag-assortativity: the similarity is decreasing with the node distance at small range and reaches a minimum at medium distances. In other words, tag-similar nodes are linked with each other at higher probability than at random. The tag-assortativity is supported by the investigation of the tag-induced sub-graphs as well, since the number of links between the nodes sharing a given tag is always larger than (or at least equal to) the number of links expected at random.

An even more interesting property of the tag-induced sub-graphs is that the average number of their links follow a power-law as a function of the number of their nodes for several orders of magnitude. The magnitude of the *tag-assortativity exponent* μ, defined by this power-law is in close relation with the tag-assortativity property of the network: a value of $\mu > 2$ indicates tag-disassortativity; $\mu = 2$ characterizes no correlation between tag-similarity and link distribution; whereas $0 < \mu < 2$ is the regime of tag-assortativity (with $0 < \mu < 1$ representing extreme tag-assortativity). The tag-assortativity exponent was slightly above 1 for all studied networks in our case.

The above scaling also reveals that the structure of the studied tagged networks is *self-similar*. This is supported by the fact that the statistical properties of tag-induced sub-graphs are similar to those of the entire graph. This means that in the statistical sense, the network is related to a sub-graph induced by a given category α in the same way as this sub-graph is related to the tag-induced sub-graph of a descendent of α, i.e. both the network topology and the tag distribution are *scale invariant*.

Acknowledgments

We thank E Gabrilovich for public domain pre-processing software of Wikipedia. This work was supported by the Hungarian National Science Fund (OTKA K68669, K75334 and T049674), the National Research and Technological Office (NKTH, CellCom RET, Textrend) and the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

Appendix

A.1. Preparing a DAG from the category hierarchy of the English Wikipedia

In Wikipedia the classification terms of each page (appearing at the bottom of the page) are called categories and are arranged into a hierarchy, i.e. a directed network where a more general term is connected to each of its child terms via a directed link. It is important to note that this directed graph contains cycles (loops): a closed path of nodes where each node (a category) is a sub-category of the previous one and the first is a sub-category of the last. Many of these loops are short and are made up of a small group of synonymous terms, e.g. the categories Hindustani and Urdu are very closely related and are both sub-categories of the other. An example for a longer loop is Education: Social sciences: Academic disciplines: Academia: Education, and a loop of length 22 has been found, too, in the English Wikipedia [68].

Loops in the category hierarchy can confuse both readers and search engines, and prohibit a tree-based semantic analysis of annotations. For example, with loops it would be impossible to identify the closest common ancestor(s) of two arbitrary terms and decide their level of relatedness. To delete all loops from the hierarchy of Wikipedia categories, first we devised an

algorithm eliminating all loops from a generic directed network by sequentially removing single directed links and modifying the directed network by the smallest possible amount. Then, we applied the algorithm to the directed network defined by the category hierarchy of the English Wikipedia.

The algorithm can be applied to an arbitrary directed network (nodes connected with directed links) and it has two parts. First, it identifies the 'loop sub-graph' of the full directed graph, the set containing precisely the directed links of all loops. This is achieved by an iteration where in each step all directed links are removed that have either a start node that is a source (no incoming link) or an end node that is a drain (no outgoing links). Neither of these two node types (source and drain) can be in a loop. Repeating this removal step until at least one node is removed lead to a sub-graph containing precisely the loops of the full graph. Note that the loop sub-graph may have more than one graph component.

The second step of the algorithm identifies a set of directed links (L) whose removal from the loop sub-graph eliminates all of its loops. As the loop sub-graph is by definition the set of loops of the original graph, removing the same directed links from the full graph will eliminate its loops. We selected the set of removed links, L, with the goal to modify the full graph by the smallest possible amount. This concerns not only the size of L (the number of links removed), but also selecting links with the smallest significance as viewed from the full graph. Turning back to one of the above examples, one has to decide which of the two directed links 'Urdu is a sub-category of Hindustani' or 'Hindustani is a sub-category of Urdu' is less relevant from the point of view of the entire directed network. More generally, suppose that in a (directed) network the directed links A $\rightarrow$ B and B $\rightarrow$ A are both present. To eliminate the loop A $\rightarrow$ B $\rightarrow$ A, one of the two links has to be removed.

To decide which of the two links is less significant, consider another example. In a directed network with the four links M $\rightarrow$ A, A $\rightarrow$ B, B $\rightarrow$ A and B $\rightarrow$ N, the link A $\rightarrow$ B is more important than B $\rightarrow$ A, because it is contained by a long continuous path, M $\rightarrow$ A $\rightarrow$ B $\rightarrow$ N. On the other hand, B $\rightarrow$ A points in the opposite direction, thus, it is likely to be a 'side effect'. The difference between these two links can be measured. The number of point-to-point shortest directed paths passing through A $\rightarrow$ B is larger (3: M $\rightarrow$ N, A $\rightarrow$ N and A $\rightarrow$ N) than the number of those containing B $\rightarrow$ A (only 1: B $\rightarrow$ A). In a directed network the number of shortest paths passing through a given (directed) link is called the directed betweenness centrality of that link. Multiple shortest paths between two nodes are accounted for by weighting, see e.g. [2] for the undirected case. Based on the above observation, we quantified the significance of each directed link by its directed betweenness centrality, $\mathcal{B}$, as measured in the full network.

Now let us return to the second part of the algorithm starting from the loop sub-graph. Knowing $\mathcal{B}$ of each link in this sub-net, we can select and remove the least important link, i.e. the one with the lowest $\mathcal{B}$ value. This link removal may produce source nodes (only outgoing links) and drain nodes (only incoming links). Again we iteratively remove links not contained by loops until the remaining network 'melts down' to the set of remaining loops. We repeat this step—deleting the link with smallest $\mathcal{B}$ and then iteratively removing all non-loop links—until no more links remain. We save the set of removed links, L, and remove the same set of links from the full graph to eliminate all of its loops by modifying it by the smallest possible amount.

The full category hierarchy of the English Wikipedia (17 October 2007 version) contains 265 432 nodes (categories) and 543 722 directed links (category–sub-category connections). The loop sub-graph has 4980 nodes and 13 164 (directed) links. The total number of removed

links was $|L| = 3977$. Data together with processing programs can be downloaded from the website http://CFinder.org.

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] Boccaletti S, Latora V, Moreno Y, Chavez M and Hwang D-U 2006 Complex networks: structure and dynamics *Phys. Rep.* **424** 175–308
[7] Jeong H, Tombor B, Albert R, Oltvai Z N and Barabási A-L 2000 The large-scale organization of metabolic networks *Nature* **407** 651–4
[8] Ravasz E, Somera A L, Mongru D A, Oltvai Z N and Barabási A-L 2002 Hierarchical organization of modularity in metabolic networks *Science* **297** 1551–5
[9] Han J-D J, Bertin N, Hao T, Goldberg D S, Berriz G F, Zhang L V, Dupuy D, Walhout A J M, Cusick M E, Roth F P and Vidal M 2004 Evidence for dynamically organized modularity in the yeast protein–protein interaction network *Nature* **430** 88–93
[10] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D and Alon U 2002 Network motifs: simple building blocks of complex networks *Science* **298** 824–7
[11] Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, Sheffer M and Alon U 2004 Superfamilies of evolved and designed networks *Science* **303** 1538–42
[12] Blatt M, Wiseman S and Domany E 1996 Super-paramagnetic clustering of data *Phys. Rev. Lett.* **76** 3251–4
[13] Girvan M and Newman M E J 2002 Community structure in social and biological networks *Proc. Natl Acad. Sci. USA* **99** 7821–6
[14] Zhou H 2003 Distance, dissimilarity index, and network community structure *Phys. Rev.* E **67** 061901
[15] Newman M E J 2004 Fast algorithm for detecting community structure in networks *Phys. Rev.* E **69** 066133
[16] Radicchi F, Castellano C, Cecconi F, Loreto V and Parisi D 2004 Defining and identifying communities in networks *Proc. Natl Acad. Sci. USA* **101** 2658–63
[17] Wilkinson D M and Huberman B A 2004 A method for finding communities of related genes *Proc. Natl Acad. Sci. USA* **101** 5241–8
[18] Reichardt J and Bornholdt S 2004 Detecting fuzzy community structures in complex networks with a Potts model *Phys. Rev. Lett.* **93** 218701
[19] Scott J 2000 *Social Network Analysis: A Handbook* 2nd edn (London: Sage Publications)
[20] Shiffrin R M and Börner K 2004 Mapping knowledge domains *Proc. Natl Acad. Sci. USA* **101** (Suppl 1) 5183–5
[21] Everitt B S 1993 *Cluster Analysis* 3rd edn (London: Edward Arnold)
[22] Knudsen S 2004 *A Guide to Analysis of DNA Microarray Data* 2nd edn (Wilmington, DE: Wiley-Liss)
[23] Newman M E J 2004 Detecting community structure in networks *Eur. Phys. J.* B **38** 321–30
[24] 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
[25] Fortunato S and Castellano C 2009 Community structure in graphs *Encyclopedia of Complexity and System Science* (Berlin: Springer) (arXiv:0712.2716)
[26] Lancichinetti A, Fortunato S and Kertész J 2008 Detecting the overlapping and hierarchical community structure of complex networks arXiv:0802.1218

[27] Spirin V and Mirny K A 2003 Protein complexes and functional modules in molecular networks *Proc. Natl Acad. Sci. USA* **100** 12123–8
[28] Onnela J-P, Chakraborti A, Kaski K, Kertész J and Kanto A 2003 Dynamics of market correlations: taxonomy and portfolio analysis *Phys. Rev.* E **68** 056110
[29] Heimo T, Saramäki J, Onnela J-P and Kaski K 2007 Spectral and network methods in the analysis of correlation matrices of stock returns *Physica* A **383** 147–51
[30] Watts D J, Dodds P S and Newman M E J 2002 Identity and search in social networks *Science* **296** 1302–5
[31] Palla G, Barabási A-L and Vicsek T 2007 Quantifying social group evolution *Nature* **446** 664–7
[32] Szabó G, Vukov J and Szolnoki A 2005 Phase diagrams for an evolutionary prisoner's dilemma game on two-dimensional lattices *Phys. Rev.* E **72** 047107
[33] Vukov J, Szabó G and Szolnoki A 2006 Cooperation in the noisy case: prisoner's dilemma game on two types of regular random graphs *Phys. Rev.* E **73** 067103
[34] Szabó G and Fáth G 2007 Evolutionary games on graphs *Phys. Rep.* **446** 97–216
[35] Mason O and Verwoerd M 2007 Graph theory and networks in biology *IET Syst. Biol.* **1** 89–119
[36] Zhu X, Gerstein M and Snyder M 2007 Getting connected: analysis and principles of biological networks *Genes Dev.* **21** 1010–24
[37] Aittokallio T and Schwikowski B 2006 Graph-based methods for analysing networks in cell biology *Brief. Bioinform.* **7** 243–55
[38] 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
[39] Jonsson P F and Bates P A 2006 Global topological features of cancer proteins in the human interactome *Bioinformatics* **22** 2291–7
[40] 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 Bioinform.* **7** 2
[41] Zimmermann M G, Eguíluz V M and Miguel M S 2004 Coevolution of dynamical states and interactions in dynamic networks *Phys. Rev.* E **69** 065102
[42] 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
[43] Kossinets G and Watts D J 2006 Empirical analysis of an evolving social network *Science* **311** 88–90
[44] Ehrhardt G C M A and Marsili M 2006 Phenomenological models of socioeconomic network dynamics *Phys. Rev.* E **74** 036106
[45] Holme P and Newman M E J 2006 Nonequilibrium phase transition in the coevolution of networks and opinions *Phys. Rev.* E **74** 056108
[46] Gil S and Zanette D H 2006 Coevolution of agents and networks: opinion sperading and community disconnection *Phys. Lett.* A **356** 89–94
[47] 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
[48] Vazquez F, Eguíluz V M and Miguel M S 2008 Generic absorbing transition in coevolution dynamics *Phys. Rev. Lett.* **100** 108702
[49] Kozma B and Barrat A 2008 Consensus formation on adaptive networks *Phys. Rev.* E **77** 016102
[50] Benczik I J, Benczik S Z, Schmittmann B and Zia R K P 2008 Lack of consensus in social systems *Europhys. Lett.* **82** 48006
[51] Lambiotte R and Ausloos M 2006 Collaborative tagging as a tripartite network *Lect. Notes Comput. Sci.* **3993** 1114–7
[52] Jaccard P 1908 Nouvelles recherches sur la distribution florale *Bull. Soc. Vandoise Sci. Nat.* **44** 223–70
[53] Resnik P 1999 Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language *J. Artif. Intell. Res.* **11** 95–130

[54] Lin D 1998 An information-theoric definition of similarity *Proc. 15th Int. Conf. on Machine Learning San Francisco, CA* pp 296–304
[55] Guo X, Liu R, Shriver C D, Hu H and Liebman M N 2006 Assessing semantic similarity measures for the characterization of human regulatory pathways *Bioinformatics* **22** 967–73
[56] Schlicker A, Domingues F S, Rahnenführer J and Lengauer T 2006 A new measure for functional similarity of gene products based on gene ontology *BMC Bioinform.* **7** 302–17
[57] Jensen T R and Toft B 1995 *Graph Coloring Problems* (New York: Wiley-Interscience)
[58] Mewes H W *et al* 2008 MIPS: analysis and annotation of genome information in 2007 *Nucl. Acids Res.* **36** D196–201
[59] The Gene Ontology Consortium 2000 Gene ontology: tool for the unification of biology *Nat. Genet.* **25** 25–9
[60] http://www.ams.org/mathscinet
[61] http://en.wikipedia.org
[62] Zlatić V, Božičević M, Štefančić H and Domazet M 2006 Wikipedias: collaborative web-based encyclopedias as complex networks *Phys. Rev.* E **74** 016115
[63] 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
[64] 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
[65] De Los Rios P 2001 Power law size distribution of supercritical random trees *Europhys. Lett.* **56** 898–903
[66] Caldarelli G, Caretta Cartozo C, De Los Rios P and Servedio V D P 2004 Widespread occurrence of the inverse square distribution in social sciences and taxonomy *Phys. Rev.* E **69** 035101
[67] Caretta Cartozo C, Garlaschelli D, Ricotta C, Barthélemy M and Caldarelli G 2008 Quantifying the taxonomic diversity in real species communities *J. Phys. A: Math. Theor.* **41** 224012
[68] http://en.wikipedia.org/wiki/Wikipedia:Categorization.