Clustering of tag-induced sub-graphs in complex networks

Péter Pollner\textsuperscript{a}, Gergely Palla\textsuperscript{a}, Tamás Vicsek\textsuperscript{a,b}

\textsuperscript{a}Statistical and Biological Physics Research Group of HAS, Eötvös University, 1117 Pázmány P. Sány 1/A, Budapest, Hungary
\textsuperscript{b}Department of Biological Physics, Eötvös University, 1117 Pázmány P. Sány 1/A, Budapest, Hungary

Abstract

We study the behavior of the clustering coefficient in tagged networks. The rich variety of tags associated with the nodes in the studied systems provide additional information about the entities represented by the nodes which can be important for practical applications like searching in the networks. Here we examine how the clustering coefficient changes when narrowing the network to a sub-graph marked by a given tag, and how does it correlate with various other properties of the sub-graph. Another interesting question addressed in the paper is how the clustering coefficient of the individual nodes is affected by the tags on the node. We believe these sort of analysis help acquiring a more complete description of the structure of large complex systems.

Keywords: networks, tags, clustering

PACS: 02.70.Rr, 89.20.-a, 89.75.Hc

1. Introduction

A wide range of complex natural, social and technological phenomena can be analyzed in terms of networks capturing the intricate web of connections among the units (building blocks) of the system under study \cite{1,2}. Over the last decade it has 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 \cite{3}, anomalous degree distributions \cite{4,5} and an intricate modular structure \cite{6,7,8}. Although the majority of complex network studies concern simply the topology of the graph corresponding to the investigated system, there is a steadily increasing interest towards tagged networks as well.

The inclusion of node tags (also called as attributes, annotations, properties, categories, features) leads to a richer structure, opening up the possibility for a more comprehensive analysis. 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 appearance of tags e.g., in biological networks is
very common [9, 10, 11, 12, 13, 14], where they usually refer to the biological function of the units represented by the nodes (proteins, genes, etc.). Another field of high interest and special importance from the point of view of practical applications is given by folksonomies and collaborative tagging systems like CiteUlike, Delicious or Flickr [15, 16]. These originate from users associating tags to certain objects (web-pages, photos, etc.), with each tagging action defining a user-tag-object triplet. The natural representation of these systems is given by tri-partite graphs, or in a more general framework by hypergraphs where the hyperedges can connect more than two nodes together. Modeling folksonomies with random hypergraphs is a very interesting new field in complex network theory which is likely to gain serious importance in the close future [17, 18]. Interesting applications of node features can be seen in the studies of co-evolving network models as well, where the evolution of the network topology affects the node properties and vice versa [19, 20, 21, 22, 23, 24, 25, 26, 27, 28]. 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 in time.

Finally, we mention our previous work aimed at the fundamental statistical features of tagged networks where the tags are organized into an ontology [29]. According to our results an interesting self-similarity and scaling can be observed in the link density of the sub-graphs spanning between nodes marked by a given tag or any descendants of this tag in the ontology. Here we continue the study of the relationship between the distribution of tags and the topology by focusing on the clustering coefficient in tagged networks. The clustering coefficient, $C$ is an important measure of the transitivity in a network, measuring the probability of two neighbors of the same node being linked to each other as well [3]. The average $C$ of real networks is usually significantly higher than that of a corresponding Erdős-Rényi (E-R) random graph [30], and for some networks $C$ was claimed to scale as $d^{-1}$ with the node degree $d$ [31].

Related to that, here we shall investigate how the number of tags on the node affect the clustering coefficient. Furthermore, we shall also study the clustering coefficient in the tag-induced sub-graphs. These sub-graphs can be important when e.g., searching in the network. The narrowing or widening of the specificity of the tag according to which we are searching corresponds to switching between sub-graphs embedded in one another, which is presumably accompanied by the change in the cohesiveness of the sub-graph in question. We shall study this change in the cohesiveness via the clustering coefficient. Tagged networks and the change in the cohesiveness of a chosen sub-graph due to the narrowing/widening of the tag specifying the included nodes can be important in the usage of recommendation systems as well. These systems are aimed at offering new services to customers based on previously purchased services. A natural representation of the available services is given by a network with link weights corresponding to the frequency of simultaneous purchasing of the two items, and above a certain number of services the service providers usually organize the services into a hierarchy which can be represented by node tags. The widening or narrowing of the scope of services to take into account during
the recommendation process is similar to switching between more specific or more general tags when searching in a tagged network.

The paper is organized as follows: in Sect. 2 we specify the two alternative definitions mostly used for the clustering coefficient as well as the various quantities related to the tag-induced sub-graphs. In Sect. 3 we briefly describe the studied networks and show the results concerning the behavior of the clustering coefficient, and finally we conclude in Sect. 4.

2. Definitions

2.1. Clustering coefficient

The clustering coefficient has actually two (slightly different) definitions, one was given by D.J. Watts and S.H. Strogatz based on the local neighborhood around a given node [3]. In this approach, the clustering coefficient, $C_i$ of node $i$ is given by

$$C_i = \frac{2t_i}{d_i(d_i - 1)},$$

where $t_i$ denotes the number of triangles passing through $i$ (equivalent to the number of links between the neighbors of $i$), and $d_i$ is the degree of node $i$. (The clustering coefficient of nodes with less then two links is zero by definition). The clustering coefficient of a sub-graph (or the whole network) is simply $\langle C \rangle$ averaged over its nodes.

To avoid ambiguity, we shall refer to as the transitivity coefficient for the alternative definition of the clustering coefficient, given only for sub-graphs and not for the individual nodes. The transitivity coefficient $T$ of a sub-graph $G$ is given by

$$T = \frac{3t_G}{b_G},$$

where $t_G$ denotes the number of triangles in the sub-graph and $b_G$ stands for the number of connected triples of nodes (equivalent to the number of paths with length two) [32, 33]. The factor of 3 in the numerator accounts for the fact that each triangle contributes to 3 connected triples of nodes, one for each of its 3 nodes. The main difference between the two definitions is that (1) tends to weight the contribution from low degree nodes more heavily, because such nodes have a smaller denominator [34].

2.2. Tag frequencies

As we mentioned in the Introduction, the number of associated tags can vary from one node to the other, and similarly, the frequency of the different tags can also be rather heterogeneous. What can make the picture more complex is that in many systems 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), in which a directed link from a category $\alpha$ pointing to another category $\beta$ represents a “$\beta$ is a sub-category of $\alpha$” relation.
The nodes 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.

Given the DAG between the possible tags, we can define the frequency of a given tag $\alpha$ in two different ways [29]:

$$p_\alpha \equiv N_\alpha/N, \quad (3)$$

$$p_\alpha \equiv \tilde{N}_\alpha/N, \quad (4)$$

where $N_\alpha$ denotes the number of nodes tagged with $\alpha$, $\tilde{N}_\alpha$ stands for the number of nodes tagged with $\alpha$ or any of its descendants, and $N$ is equal to the total number of nodes in the network. 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). The $\tilde{p}_\alpha$ plays an important role in semantic similarity measures [35, 36], e.g., in case of the similarity measure defined by P. Resnik the similarity of two tags is given as $-\log \tilde{p}$ of their lowest common ancestor in the DAG.

2.3. Tag-induced sub-graphs

One of the key objects of the present study is given by the tag-induced sub-graphs, spanning between nodes marked by a given tag $\alpha$ and any of its descendants. (For an illustration see Fig.1). 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 $\tilde{M}_\alpha = \tilde{N}_\alpha(\tilde{N}_\alpha - 1)/2$. According to our previous results, an interesting self-similar property and scaling can be observed when comparing the different tag-induced sub-graphs [29]. The expected number of links in the tag-induced sub-graphs follows a power-law in function of the number of nodes in the sub-graphs, $\tilde{M} \sim \tilde{N}^\mu$, characterized by an exponent $\mu$ related to the tag-assortativity. A network is tag-assortative, if nodes with similar tags are linked with a larger probability than at random.

In the absence of correlations between the tags and the graph topology the mentioned tag-assortativity exponent equals $\mu = 2$. (In this case a tag-induced sub-graph corresponds to just a random sample from the network with $\tilde{N}(\tilde{N} - 1)/2$ possible places for links which are filled with a uniform probability independent of the sub-graph size $\tilde{N}$). In contrast, for the studied systems $\mu$ was found to be between 1 and 1.5, which is a signature of tag-assortativity as we shall see shortly. The probability to find a link between a randomly chosen pair of nodes in the tag-induced sub-graph denoted by $\rho$ scales as $\rho \sim \tilde{M}/\tilde{N}^2 \sim \tilde{N}^{\mu-2}$. When $\mu < 2$, this linking probability $\rho$ becomes larger for the smaller sub-graphs, corresponding to more specific tags, thus, the network is tag-assortative.

Based on the above behavior one expects that the clustering coefficient (transitivity coefficient) in the tag-induced sub-graphs of more specific tags should be higher on average as well. In an Erdős-Rényi (E-R) graph [30] with the same
number of nodes and links as a chosen tag-induced sub-graph the clustering coefficient would be equal to the linking probability $\rho$. Thus, we expect $\langle C \rangle$ (and $T$) to grow at least as $N^{\mu-2}$ on average when moving from the tag-induced sub-graph of a general tag to the tag-induced sub-graph of a more specific one.

3. Applications

We studied the behavior of the clustering coefficient in the same three networks of high interest as in [29], capturing the relations between interacting proteins, collaborating scientists, and pages of an on-line encyclopedia. The protein-protein interaction network of MIPS [37] consisted of $N = 4546$ proteins, connected by $M = 12319$ 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 [38].
The co-authorship network originated from MathSciNet (Mathematical review collection of the American Mathematical Society) [39], with \( N = 391529 \) scientists connected by \( M = 873775 \) links of collaboration. 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 third network was given by a subset of pages from the English Wikipedia [40, 41, 42, 43], connected by 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. Thus, we removed a few relations from this graph until it turned into a DAG, following a method detailed in the Appendix of [29]. The chosen subset of pages corresponded to the tag-induced sub-graph of “Japan”, consisting of \( N = 43307 \) nodes, \( M = 102753 \) links with 3197 sub-categories appearing as tags on the nodes.

![Figure 2: The average clustering coefficient of nodes in function of their number of tags for the Wiki-Japan (crosses), the MathSciNet (boxes) and the MIPS network (circles).](image)

In Fig. 2, we show the average clustering coefficient of the nodes, \( \langle C \rangle \) in function of their number of tags. As we already pointed out, the studied systems are all tag-assortative, thus, neighboring nodes are likely to share common tags. From this it follows that nodes with a lot of tags are likely to have different neighbors which are not linked to each other, thus, we expect the clustering coefficient of such nodes to be lower than those with few tags. We see this expected decreasing tendency with some fluctuations towards the large number
of tags for the Wiki-Japan network. In case of the MathSciNet the overall tendency is decreasing as well, however, the $\langle C \rangle$ of nodes with one tag is larger than that of those with none, producing a maximum at the beginning of the curve. Most peculiar is the curve of the MIPS (circles), showing an increasing tendency for low number of tags and a fluctuating plateau for larger values. This network has shown interesting differences from the other two networks in our previous study as well [29]. E.g., hubs with a rather special function (described by a single- or only few tags) could be found, contradicting the simple argument of large node degree correlating with large number of tags on the node. (The proteins helping other proteins to fold are good examples for this). Naturally, the clustering coefficient is expected to be low for these nodes due to the large degree.

In Fig.3 we display the transitivity coefficient of the tag-induced sub-graphs in function of $\bar{p}$ corresponding to their relative size. For each network we also plotted the values one would obtain in E-R graphs with the same number of

![Graph showing transitivity coefficient in function of relative size.](image)

Figure 3: The transitivity coefficient of tag-induced sub-graphs in function of their relative size, $\bar{p}$. In each panel, the dark symbols show the measured value, whereas the gray symbols correspond to the estimated value in an E-R graph with the same number of nodes and links as the tag-induced sub-graph. The continuous- and dashed black curves show the average values. For each network there is a sub-range of $\bar{p}$ values in which the average of the measured $T$ decays more or less as a power-law, this is shown by the dashed gray lines.
nodes and links as the tag-induced sub-graphs, showing the $\tilde{N}^{p-2}$ scaling as discussed in Sect. 2.3. For all three networks the average $T$ is clearly higher than what we would get in the E-R counterpart, which is a signature of correlations making these sub-graphs more cohesive. This difference becomes really significant for the larger sub-graphs (more general tags). Although the average of the actual $T$ of the tag-induced sub-graphs can be fitted with a power-law only in a limited range, an apparently decreasing $\langle T \rangle$ curve can be observed in function of $\tilde{p}$ for the MIPS and the MathSciNet. In case of the Wiki-Japan the overall nature of the same curve is decreasing as well with an increasing tail at large $\tilde{p}$ values. This may be an effect of the poorer statistics (smaller number of sub-graphs) in this region. The exponents of the fitted power-laws are between -0.1 and -0.3, thus, the decrease in the transitivity with increasing $\tilde{p}$ is quite slow and the larger tag-induced sub-graphs remain rather clustered on average.

![Figure 4](image-url)

Figure 4: Scatter plot of the linking probability $\rho \sim M/N^2$ of the tag-induced sub-graph in function of the $\rho \sim M/N^2$ in the tag-induced sub-graph of the direct ancestor (“parent” in the DAG) of the tag for the MIPS (a), the Wiki-Japan (b) and the MathSciNet networks (c). (Each point corresponds to a direct ancestor-descendant pair.) Panel (d) depicts a contour plot of the point densities obtained from the scatter plots.

Next, we aim at investigating the direct ancestor-descendant relation of the tags by comparing various statistics of the corresponding tag-induced sub-graphs. In other words, we study how the sub-graphs change if we narrow our field of interest by moving in the DAG of tags along a directed link to a more specific tag. In Fig. 4 we simply plot the link probability $\rho \sim M/N^2$ of the descendant ("child") in function of the $\rho$ of its direct ancestor ("parent"). As
expected, the vast majority of the points falls above the $y = x$ curve, corresponding to a larger linking probability in the descendants sub-graph. Some differences can be observed between the three systems which are emphasized in Fig. 4 showing a contour plot of the point densities obtained from the scatter plots Fig. 4a-c: For the Wiki-Japan the points spread all over the range above the $y = x$ line, thus, an ancestor having a low $\rho$ value can still have a direct descendant with a high $\rho$. In case of the other two networks the points remain closer to the diagonal, corresponding to more correlated “child-parent” $\rho$ values.

Figure 5: The transitivity coefficient of the tag-induced sub-graph in function of the transitivity coefficient of the tag-induced sub-graph of its direct ancestor for the MIPS (a), the Wiki-Japan (b) and the MathSciNet networks (c). (Each point corresponds to a direct ancestor-descendant pair.) For a comparison, in panel (d) we show a contour plot of the point densities obtained from the scatter plots.

In Fig. 5 we show the transitivity coefficient $T$ in the induced sub-graph of the descendants in function of $T$ of their direct ancestors induced sub-graph. The majority of the points falls above the $y = x$ line, thus, the increase of $T$ when moving from the ancestor to the descendant is more common than the decrease. This tendency is most pronounced in case of the MathSciNet. In case of the MIPS and especially the Wiki-Japan we can also find numerous ancestor-descendant pairs where $T$ is actually smaller for the descendant. By examining such pairs in more details it turned out that the usual cause for this effect is another descendant of the ancestor having a high transitivity: in such settings the transitivity of the ancestor becomes roughly the average of the transitivities of its descendants. According to the contour plot (Fig. 5d) obtained from the indi-
individual scatter plots the transitivity coefficients of the direct ancestor-descendant pairs are more correlated than e.g., the corresponding linking probabilities, as the majority of the points in the scatter plots gather around the diagonal.

Figure 6: The ratio of the direct ancestor-descendant link probabilities in function of the ratio of the transitivity coefficient for the same pairs in the MIPS (a), the Wiki-Japan (b) and the MathScinet networks (c). Similarly to the previous Figs., in panel (d) we show a contour plot of the point densities obtained from the scatter plots.

Finally, in Fig.6 we plot the ratio of the transivities obtained in the tag-induced sub-graphs corresponding to a direct ancestor-descendant pair in function of the ratio of the linking probabilities in the same pair of sub-graphs. In case of the MIPS (Fig.6a) and the Wiki-Japan networks (Fig.6b) the plots are rather scattered, with a weak increasing tendency. This means that when moving from the induced sub-graph of a more generic tag to the induced sub-graph of its direct descendant, an increase in the linking probability will more likely induce an increase in the transitivity as well, but not necessarily. The points in case of the MathSciNet (Fig.6c) form a much more concentrated cloud than in the previous examples, corresponding to a more pronounced positive correlation between the change in $T$ and the change in $\rho$. This behavior can be seen in Fig.6d comparing the contour plots corresponding to the individual scatter plots.
4. Summary and conclusions

We studied the behavior of the clustering coefficient in tagged networks where the tags are organized into an ontology. The investigated systems showed universal features in some aspects with interesting differences from other perspectives. The average $C$ showed a decaying tendency in function of the number of tags on the nodes in case of the MathSciNet and Wiki-Japan networks. This sort of correlation was absent in case of the MIPS. The tag-induced sub-graphs showed an increasing transitivity coefficient on average with decreasing size in all networks. Furthermore, for the vast majority of the tags the transitivity in the induced sub-graph was much higher than in an Erdős-Rényi random graph with the same number of nodes and links. This difference became really significant towards the larger sub-graphs, corresponding to more general tags. In other words, when widening the specificity of tags, the transitivity in the corresponding sub-graphs decreases at a much slower rate on average than expected based on the decrease of the linking probability in the sub-graphs.

We also compared various properties of the tag-induced sub-graphs to the same properties in the induced sub-graph of the tags direct ancestor in the DAG. According to our results for the majority of the ancestor-descendant pairs the transitivity and the linking probability is larger in the sub-graph of the descendant. The correlation between the transitivity values is also stronger than the correlation between the linking probabilities.

References

[1] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. Rev. Mod. Phys., 74:47–97, 2002.
[2] J. F. F. Mendes and S. N. Dorogovtsev. Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, Oxford, 2003.
[3] D. J. Watts and S. H. Strogatz. Collective dynamics of ’small-world’ networks. Nature, 393:440–442, 1998.
[4] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. Comput. Commun. Rev., 29:251–262, 1999.
[5] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. Science, 286:509–512, 1999.
[6] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA, 99:7821–7826, 2002.
[7] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435:814–818, 2005.
[8] S. Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.

[9] O. Mason and M. Verwoerd. Graph theory and networks in Biology. *IET Systems Biology*, 1:89–119, 2007.

[10] X. Zhu, M. Gerstein, and M. Snyder. Getting connected: analysis and principles of biological networks. *Genes & Development*, 21:1010–1024, 2007.

[11] T. Aittokallio and B. Schwikowski. Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics*, 7:243–255, 2006.

[12] G. Finocchiaro, F. M. Mancuso, D. Cittaro, and H. Muller. Graph-based identification of cancer signaling pathways from published gene expression signatures using PubLiME. *Nucl. Ac. Res.*, 35:2343–2355, 2007.

[13] P. F. Jonsson and P. A. Bates. Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 22:2291–2297, 2006.

[14] P. F. Jonsson, T. Cavanna, D. Zicha, and P. A. Bates. Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC Bioinformatics*, 7:2, 2006.

[15] C. Cattuto, V. Loreto, and L. Pietronero. Semiotic dynamics and collaborative tagging. *Proc. Natl. Acad. Sci. USA*, 104:1461–1464, 2007.

[16] R. Lambiotte and M. Ausloos. Collaborative tagging as a tripartite network. *Lect. Notes in Computer Sci.*, 3993:1114–1117, 2006.

[17] G. Ghosal, V. Zlatić, G. Caldarelli, and M. E. J. Newman. Random hypergraphs and their applications. *Phys. Rev. E*, 79:066118, 2009.

[18] V. Zlatić, G. Ghosal, and G. Caldarelli. Hypergraph topological quantities for tagged social networks. *Phys. Rev. E*, 80:036118, 2009.

[19] M. G. Zimmermann, V. M. Eguíluz, and M. S. Miguel. Coevolution of dynamical stats and interactions in dynamic networks. *Phys. Rev. E*, 69:065102(R), 2004.

[20] V. M. Eguíluz, M. G. Zimmermann, and C. J. Cela-Conde. Cooperation and the emergence of role differentiation in the dynamics of social networks. *Am. J. Sociol.*, 110:977–1008, 2005.

[21] G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311:88–90, 2006.

[22] G. C. M. A. Ehrhardt and M. Marsili. Phenomenological models of socioeconomic network dynamics. *Phys. Rev. E*, 74:036106, 2006.
[23] P. Holme and M. E. J. Newman. Nonequilibrium phase transition in the coevolution of networks and opinions. *Phys. Rev. E*, 74:056108, 2006.

[24] S. Gil and D. H. Zanette. Coevolution of agents and networks: Opinion spreading and community disconnection. *Phys. Lett. A*, 356:89–94, 2006.

[25] F. Vazquez, J. C. González-Avella, V. M. Eguíluz, and M. S. Miguel. Time-scale competition leading to fragmentation and recombination transitions in the coevolution of network and states. *Phys. Rev. E*, 76:046120, 2007.

[26] F. Vazquez, V. M. Eguíluz, and M. S. Miguel. Generic absorbing transition in coevolution dynamics. *Phys. Rev. Lett.*, 100:108702, 2008.

[27] B. Kozma and A. Barrat. Consensus formation on adaptive networks. *Phys. Rev. E*, 77:016102, 2008.

[28] I. J. Benczik, S. Z. Benczik, B. Schmittmann, and R. K. P. Zia. Lack of consensus in social systems. *Europhys. Lett.*, 82:48006, 2008.

[29] G. Palla, I. J. Farkas, P. Pollner, I. Dernyi, and T. Vicsek. Fundamental statistical features and self-similar properties of tagged networks. *New Journal of Physics*, 10:123026, 2008.

[30] P. Erdős and A. Rényi. On the evolution of random graphs, 1960.

[31] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551 – 1555, 2002.

[32] A. Barrat and M. Weigt. On the properties of small-world network models. *Eur. Phys. J. B*, 13:547–560, 2000.

[33] M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proc. Natl. Acad. Sci. USA*, 99:2566–2572, 2002.

[34] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[35] P. Resnik. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intel. Res.*, 11:95–130, 1999.

[36] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, San Francisco CA, 1998.

[37] H. W. Mewes, S. Dietmann, D. Frishman, R. Gregory, G. Mannhaupt, K. Mayer, M. Muensterkötter, A. Ruepp, M. Spannagl, V. Stuempflen, and T. Rattei. Mips: Analysis and annotation of genome information in 2007. *Nucl. Acids Res.*, 36:D196–D201, 2008.
[38] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[39] [http://www.ams.org/mathscinet](http://www.ams.org/mathscinet).

[40] [http://en.wikipedia.org](http://en.wikipedia.org).

[41] V. Zlatić, M. Božičević, H. Stefančić, and M. Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Phys. Rev. E*, 74:016115, 2006.

[42] A. Capocci, V. D. P. Servedio, F. Colaioni, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. *Phys. Rev. E*, 74:036116, 2006.

[43] A. Capocci, F. Rao, and G. Caldarelli. Taxonomy and clustering in collaborative systems: The case of the on-line encyclopedia wikipedia. *Europhys. Lett.*, 81:28006, 2008.