

# Preferential attachment of communities: the same principle, but a higher level

P. POLLNER<sup>1</sup>, G. PALLA<sup>1,2</sup> and T. VICSEK<sup>1,2</sup>

<sup>1</sup> Department of Biological Physics, Eötvös University, Pázmány P. sny. 1A, H-1117 Budapest, Hungary

<sup>2</sup> Biological Physics Research Group of HAS, Eötvös University, Pázmány P. sny. 1A, H-1117 Budapest, Hungary

PACS. 89.75.Hc – Networks and genealogical trees.

PACS. 89.75.Fb – Structures and organization in complex systems.

PACS. 89.75.Da – Systems obeying scaling laws.

## Abstract. –

The graph of communities is a network emerging above the level of individual nodes in the hierarchical organisation of a complex system. In this graph the nodes correspond to communities (highly interconnected subgraphs, also called modules or clusters), and the links refer to members shared by two communities. Our analysis indicates that the development of this modular structure is driven by preferential attachment, in complete analogy with the growth of the underlying network of nodes. We study how the links between communities are born in a growing co-authorship network, and introduce a simple model for the dynamics of overlapping communities.

*Introduction.* – A wide class of complex systems occurring from the level of cells to society can be described in terms of networks capturing the intricate web of connections among the units they are made of. Graphs corresponding to these real networks exhibit unexpected non-trivial properties, *e.g.*, new kinds of degree distributions, anomalous diameter, spreading phenomena, clustering coefficient, and correlations [1–5]. In recent years, there has been a quickly growing interest in the structural sub-units of complex networks, associated with more highly interconnected parts [6–18]. These sets of nodes are usually called clusters, communities, cohesive groups, or modules, with no widely accepted, unique definition. Such building blocks (functionally related proteins [19, 20], industrial sectors [21], groups of people [14, 22], cooperative players [23, 24], *etc.*) can play a crucial role in forming the structural and functional properties of the involved networks. On the other hand, the presence of communities in networks is a relevant and informative signature of the *hierarchical nature* of complex systems [19, 25, 26].

Typically, the communities in a complex system are not isolated from each other, instead, they have overlaps, *e.g.*, a protein can be part of more than one functional unit [27], and people can be members in different social groups at the same time [28]. This observation naturally leads to the definition of the *community graph*: a network representing the connections between

the communities, with the nodes referring to communities and links corresponding to shared members between the communities. Accordingly, the community degree  $d^{\text{com}}$  of a community is given by the number of other communities it overlaps with, and is equal to the degree of the corresponding node in the community graph. The studies of the relevant statistics describing the community graph (*i.e.*, the degree distribution, clustering *etc.*) of real networks have just begun [29]. So far, in the networks investigated, the community degree distribution was shown to decay exponentially for low and as a power law for higher community degree values. This means that fat tailed degree distributions appear at two levels in the hierarchy of these systems: both at the level of nodes (the underlying networks are scale free), and at the level of the communities as well.

*Preferential attachment* is a key concept in the field of scale-free networks. In a wide range of graph models the basic mechanism behind the emerging power law degree distribution is that the new nodes attach to the old ones with probability proportional to their degree [2–4]. Furthermore, in earlier works the occurrence of preferential attachment was directly demonstrated in several real world networks with scale free degree distribution [30–32]. The observed fat tails in the degree distribution of the community graphs indicate that the mechanism of preferential attachment could be present *at the level of communities* as well. Our aim in the present manuscript is to examine the attachment statistics of communities in order to clarify this question. Our investigations focus on the development of the communities in the growing co-authorship network of the Los Alamos cond-mat e-print archive [33], in which the nodes correspond to authors, and two authors become linked if they publish an article together. By studying the time evolution of this system, we investigate the dynamics of the new community links. For example, when a previously unlinked community is attached to another one, what are the size and community degree statistics of that other community? Another, closely related issue addressed in this paper is the appearance of new members in the communities. The size distribution of the communities was found to be a power-law in the system to be investigated [29]. Thus it is natural to address questions such as: What happens when a node belonging to none of the communities suddenly joins a community? What are the size and community degree statistics of the community chosen?

*The communities.* – In the present work we study the dynamics of the communities in the Los Alamos cond-mat e-print archive [33], in which an article with  $n$  authors contributes with  $(n - 1)^{-1}$  to the weight of the links between every pairs of its authors. (The dataset contains altogether 30739 nodes and 136065 links). The communities are extracted with the Clique Percolation Method (CPM) [29, 34] at each time step, using the CFinder package freely downloadable from [35]. (Each time step corresponds to one month, and the data set contained 143 time steps from February 1992 to April 2004). The communities obtained by the CPM correspond to  $k$ -clique percolation clusters in the network. The  $k$ -cliques are complete subgraphs of size  $k$  (in which each node is connected to every other nodes). A  $k$ -clique percolation cluster is a subgraph containing  $k$ -cliques that can all reach each other through chains of  $k$ -clique adjacency, where two  $k$ -cliques are said to be adjacent if they share  $k - 1$  nodes. The  $k$ -clique percolation clusters can be best visualised with the help of  $k$ -clique templates, that are objects isomorphic to a complete graph of  $k$  vertices. Such a template can be placed onto any  $k$ -clique in the graph, and rolled to an adjacent  $k$ -clique by relocating one of its vertices and keeping its other  $k - 1$  vertices fixed. Thus, the  $k$ -clique percolation clusters ( $k$ -clique communities) of a graph are all those subgraphs that can be fully explored by rolling a  $k$ -clique template in them but cannot be left by this template.

The main advantages of this community definition are that it is not too restrictive, it is local, it is based on the density of the links and it allows overlaps between the communities:

a node can be part of several  $k$ -clique percolation clusters at the same time. The number of communities a given node  $i$  belongs to shall be referred to as the membership number  $m_i$  of the node from now on.

When applied to weighted networks (such as the present co-authorship network), the CPM method has two parameters: the  $k$ -clique size  $k$ , and a weight threshold  $\omega^*$  (links weaker than  $\omega^*$  are ignored). The criterion used to fix these parameters is based on finding a community structure as highly structured as possible. In the present paper we stick to the optimal parameter values found in earlier studies of the same co-authorship network [29], given by  $k = 6$  and  $\omega^* = 0.1$ .

*Determining attachment probabilities.* – The method presented below can be applied in general to any empirical study of an attachment process where the main goal is to decide whether the attachment is uniform or preferential with respect to a certain property (*e.g.*, degree, size, *etc.*) of the attached objects (*e.g.*, nodes, communities *etc.*). If the process is uniform with respect to a property  $\rho$ , then objects with a given  $\rho$  are chosen at a rate given by the distribution of  $\rho$  amongst the available objects. However, if the attachment mechanism prefers high (or low)  $\rho$  values, then objects with high (or low)  $\rho$  are chosen with a higher rate compared to the  $\rho$  distribution of the available objects. To monitor this enhancement, one can construct the cumulative  $\rho$  distribution  $P_t(\rho)$  of the available objects at each time step  $t$ , together with the un-normalised cumulative  $\rho$  distribution of the objects chosen by the process between  $t$  and  $t + 1$ , denoted by  $w_{t \rightarrow t+1}(\rho)$ . The value of  $w_{t \rightarrow t+1}(\rho^*)$  at a given  $\rho^*$  equals to the number of objects chosen in the process between  $t$  and  $t + 1$ , that had a  $\rho$  value larger than  $\rho^*$  at  $t$ . To detect deviations from uniform attachment, it is best to accumulate the ratio of  $w_{t \rightarrow t+1}(\rho)$  and  $P_t(\rho)$  during the time evolution to obtain

$$W(\rho) = \sum_{t=0}^{t_{\max}-1} \frac{w_{t \rightarrow t+1}(\rho)}{P_t(\rho)}. \quad (1)$$

If the attachment is uniform with respect to  $\rho$ , then  $W(\rho)$  becomes a flat function. However, if  $W(\rho)$  is an increasing function, then the objects with large  $\rho$  are favoured, if it is a decreasing function, the objects with small  $\rho$  are favoured in the attachment process. The advantage of this approach is that the rate-like variable  $w_{t \rightarrow t+1}(\rho)$  associated to the time step between  $t$  and  $t + 1$  is always compared to the  $P_t(\rho)$  distribution at  $t$ . Therefore  $W(\rho)$  is able to indicate preference (or the absence of preference) even when  $P_t(\rho)$  is slowly changing in time (as in the case of the community degree in the co-authorship network under investigation).

We have tested the above method on simulated graphs grown with known attachment mechanisms, *i*) uniform attachment (new nodes are attached to a randomly selected old node), *ii*) linear preferential attachment (new nodes are attached to old ones with a probability proportional to the degree), *iii*) and anti-preferential attachment (new nodes are attached to the old ones with a probability proportional to  $\exp(-d)$ , where  $d$  is the degree). In these cases the degree  $d$  of the individual nodes plays the role of the parameter  $\rho$ . For each time step, we recorded the cumulative degree distribution of the nodes  $P_t(d)$ , together with the number of nodes gaining new links with a degree higher than a given  $d$ , labelled by  $w_{t \rightarrow t+1}(d)$ . By summing the ratio of these two functions along the time evolution of the system one gets  $W(d) = \sum_{t=0}^{t_{\max}-1} w_{t \rightarrow t+1}(d)/P_t(d)$ . In fig.1a. we show the empirical results for  $W(d)$  obtained for the simulated networks grown with the three different attachment rules. The curves reflect the difference between the three cases very well: for the uniform attachment probability  $W(d)$  is flat, for the preferential attachment  $W(d)$  is clearly increasing, and for the anti-preferential attachment  $W(d)$  is decreasing. We have also calculated the attachment statistics of the nodes

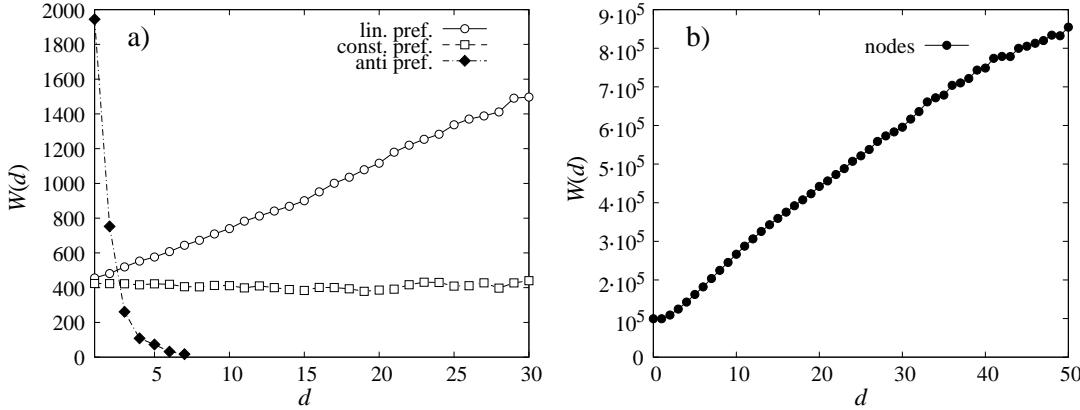


Fig. 1 – a) The  $W(d)$  function for networks grown with known attachment rules: uniform probability (squares), linear preferential attachment (open circles), and anti preferential attachment (diamonds). b) The  $W(d)$  function in the co-authorship network of the Los Alamos cond-mat archive.

in the studied co-authorship network. As it can be seen in fig.1b., the corresponding  $W(d)$  curve is increasing, therefore preferential attachment is present at the level of nodes in the system.

*Results.* – In case of the communities of the investigated co-authorship network, the two properties to be substituted in place of  $\rho$  are the community degree  $d^{\text{com}}$  and the community size  $s$ , therefore, the cumulative community size distribution  $P_t(s)$  and the cumulative community degree distribution  $P_t(d^{\text{com}})$  were recorded at each time step  $t$ . To study the *establishment of the new community links*, we constructed the un-normalised cumulative size distribution  $w_{t \rightarrow t+1}(s)$  and the un-normalised cumulative degree distribution  $w_{t \rightarrow t+1}(d^{\text{com}})$  of the communities gaining new community links to previously unlinked communities. The value of these distributions at a given  $s$  (or given  $d^{\text{com}}$ ) is equal to the number of unlinked communities at  $t$  that establish a community link between  $t$  and  $t+1$  with a community larger than  $s$  (or having larger degree than  $d^{\text{com}}$ ) at  $t$ . By accumulating the ratio of the rate-like variables and the corresponding distributions we obtain

$$W(s) = \sum_{t=0}^{t_{\max}-1} \frac{w_{t \rightarrow t+1}(s)}{P_t(s)}, \quad W(d^{\text{com}}) = \sum_{t=0}^{t_{\max}-1} \frac{w_{t \rightarrow t+1}(d^{\text{com}})}{P_t(d^{\text{com}})}. \quad (2)$$

For the investigation of the *appearance of new members* in the communities, we recorded the un-normalised community size distribution  $\hat{w}_{t \rightarrow t+1}(s)$  and the un-normalised community degree distribution  $\hat{w}_{t \rightarrow t+1}(d^{\text{com}})$  of the communities gaining new members (belonging previously to none of the communities) between  $t$  and  $t+1$ . The corresponding distributions that can be used to detect deviations from the uniform attachment are

$$\widehat{W}(s) = \sum_{t=0}^{t_{\max}-1} \frac{\hat{w}_{t \rightarrow t+1}(s)}{P_t(s)}, \quad \widehat{W}(d^{\text{com}}) = \sum_{t=0}^{t_{\max}-1} \frac{\hat{w}_{t \rightarrow t+1}(d^{\text{com}})}{P_t(d^{\text{com}})}. \quad (3)$$

In fig.2a. we show the empirical  $W(s)$  and  $\widehat{W}(s)$  functions, whereas in fig.2b. the empirical  $W(d^{\text{com}})$  and  $\widehat{W}(d^{\text{com}})$  are displayed. All four functions are clearly increasing, therefore we can draw the following important conclusions:

- When a previously unlinked community establishes a new community link, communities with large size and large degree are selected with enhanced probability from the available other communities.
- When a node previously belonging to none of the communities joins a community, communities with large size and large degree are selected with enhanced probability from the available communities.

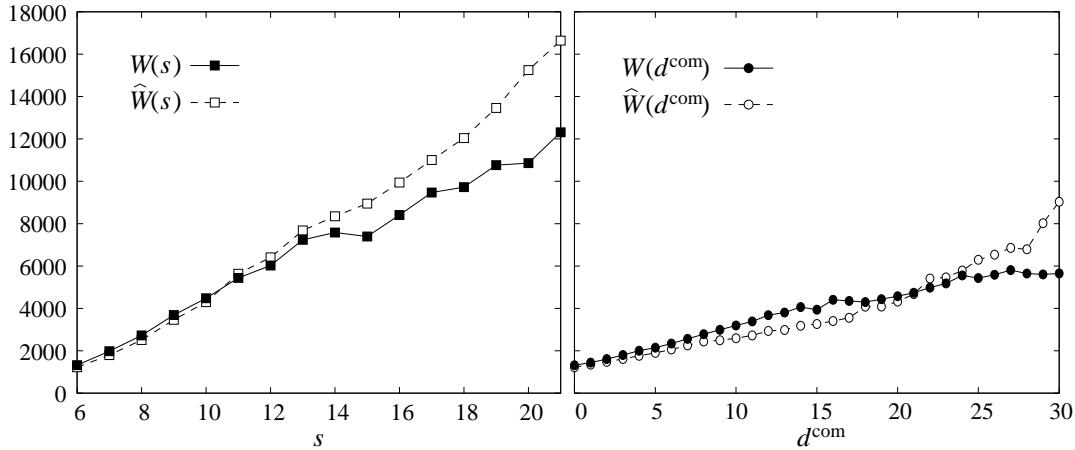


Fig. 2 – a) The  $W(s)$  and  $\hat{W}(s)$  functions for the communities of the co-authorship network of the Los Alamos cond-mat e-print archive. b) The  $W(d^{\text{com}})$  and  $\hat{W}(d^{\text{com}})$  functions of the same network. The increasing nature of these functions indicates preferential attachment at the level of communities in the system.

We note that the community size and the community degree are strongly for higher sizes and degrees: large communities have large community degree and vice versa. Therefore, if one observes an attachment mechanism that is preferential with respect to either the community size, or the community degree, than it must be preferential for both of them.

*A toy model.* – In this section we outline a simple model for the growth of overlapping communities, in which the preferential attachment of the node to communities results in the emergence of a community system with scaling community size and community degree distribution. We note that when using the well known models based on preferential attachment solely between the nodes [2–4], the resulting graph contains no communities at all at  $k = 6$ .

In our model the underlying network between the nodes is left unspecified, the focus is on the content of the communities. During the time evolution, similarly to the models published in [36–38], new members may join the already existing communities, and new communities may emerge as well. The new nodes introduced to the system choose their community preferentially with the community size, therefore the size distribution of the communities is expected to develop into a power-law. The appearance of the new community links originates in new nodes joining several communities at the same time. The detailed rules of the model are the following:

- The initial state of the model is a small set of communities with random size.
- The new nodes are added to the system separately. For each new node  $i$ , a membership  $m_i$  is drawn from a Poissonean distribution with an expectation value of  $\lambda$ .

- If  $m_i \geq 1$ , communities are successively chosen with probabilities proportional to their sizes, until  $m_i$  is reached, and the node  $i$  joins the chosen communities simultaneously.
- If  $m_i = 0$ , the node  $i$  joins the group of unclassified vertices.
- When the ratio  $r$  of the group of unclassified nodes compared to the total number of nodes  $N$  exceeds a certain limit  $r^*$ , a number of  $q$  vertices from the group establish a new community. (Obviously,  $q$  must be smaller than  $Nr$  even in the initial state).

To be able to compare the results of the model with the community structure of the co-authorship network, the runs were stopped when the number of nodes in the model reached the size of the co-authorship network.

Our experience showed that the model is quite insensitive to changes in  $r$  or  $q$ , and  $\lambda$  is the only important parameter. For small values ( $\lambda < 0.3$ ) the resulting community degree distribution is truncated, whereas when  $\lambda$  is too large ( $\lambda > 1$ ), a giant community with abnormally large community degree appears. For intermediate  $\lambda$  values ( $0.3 < \lambda < 1$ ), the community size- and community degree distributions become fat tailed, similarly to the co-authorship network. In fig.3. we show the cumulative community size distribution  $P(s)$  and

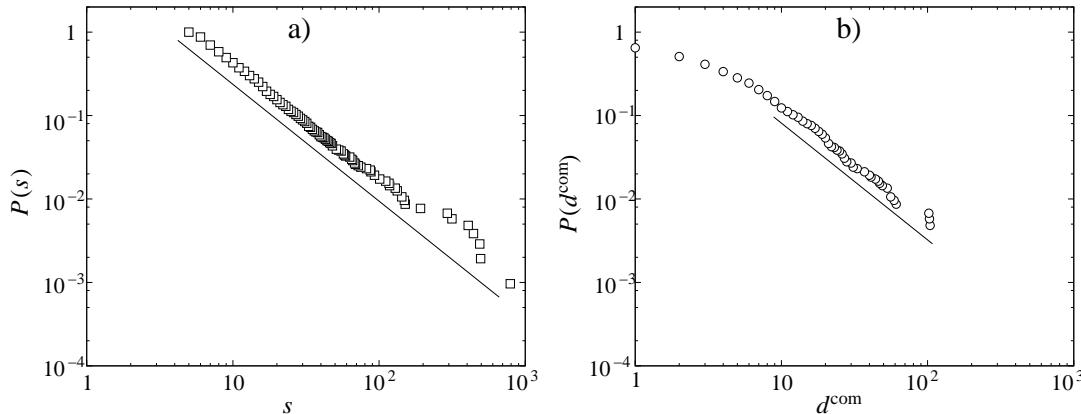


Fig. 3 – a) the cumulative community size distribution  $P(s)$  (open circles) in our model at  $\lambda = 0.6$  follows a power-law with an exponent of  $-1.4$  (straight line) ( b ) the cumulative community degree distribution  $P(d^{\text{com}})$  (filled circles) in our model at the same  $\lambda$ . The tail of this distribution follows the same power-law as the community size distribution (straight line), similarly to the communities found in the co-authorship network [29].

the cumulative community degree distribution  $P(d^{\text{com}})$  of the communities obtained in our model at  $\lambda = 0.6$ . (Changes in the parameters  $r$  and  $q$  only shifts these curves, their shape remains unchanged). Our model grasps the relevant statistical properties of the community structure in the co-authorship network [29] quite well: the community size distribution and the tail of the community degree distribution follow a power-law with the same exponent.

*Conclusions.* – We studied the evolution of the community graph in a growing co-authorship network. We found that similar processes control the growth of the system at different levels in the hierarchy, as the growth of the communities, the development of the community graph and the growth of the underlying network are all driven by preferential attachment. Inspired by these results, we introduced a simple model for the dynamics of overlapping communities leading to scaling size- and community degree distribution.

\* \* \*

This work has been supported in part by the Hungarian Science Foundation (OTKA), grant Nos. F047203 and T034995. We thank A.-L. Barabási for useful discussions.

## REFERENCES

- [1] D. J. WATTS and S. H. STROGATZ, *Nature*, **393** (1998) 440.
- [2] A.-L. BARABÁSI and R. ALBERT, *Science*, **286** (1999) 509.
- [3] R. ALBERT and A.-L. BARABÁSI, *Rev. Mod. Phys.*, **74** (2002) 47.
- [4] J. F. F. MENDES and S. N. DOROGOVTCSEV, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford) 2003.
- [5] A. BARRAT and M. BARTHELEMY and A. VESPIGNANI, *Phys. Rev. Lett.*, **92** (2004) 228701.
- [6] M. BLATT and S. WISEMAN and E. DOMANY, *Phys. Rev. Lett.*, **76** (1996) 3251.
- [7] M. GIRVAN and M. E. J. NEWMAN, *Proc. Natl. Acad. Sci. USA*, **99** (2002) 7821.
- [8] H. ZHOU, *Phys. Rev. E*, **67** (2003) 061901.
- [9] M. E. J. NEWMAN, *Phys. Rev. E*, **69** (2004) 066133.
- [10] F. RADICCHI and C. CASTELLANO and F. CECCONI and V. LORETO and D. PARISI, *Proc. Natl. Acad. Sci. USA*, **101** (2004) 2658.
- [11] L. DONETTI and M. A. MUÑOZ, *cond-mat/0404652*, (2004) .
- [12] D. M. WILKINSON and B. A. HUBERMAN, *Proc. Natl. Acad. Sci. USA*, **101** (2004) 5241.
- [13] J. REICHARDT and S. BORNHOLDT, *Phys. Rev. Lett.*, **93** (2004) 218701.
- [14] J. SCOTT, *Social Network Analysis: A Handbook*, 2nd ed. (Sage Publications, London) 2000.
- [15] R. M. SHIFFRIN and K. BÖRNER, *Proc. Natl. Acad. Sci. USA*, **101** (2004) 5183.
- [16] B. S. EVERITT, *Cluster Analysis*, 3th ed. (Edward Arnold, London) 1993.
- [17] S. KNUDSEN, *A Guide to Analysis of DNA Microarray Data*, 2nd ed. (Wiley-Liss) 2004.
- [18] M. E. J. NEWMAN, *Eur. Phys. J. B*, **38** (2004) 321
- [19] E. RAVASZ, and A. L. SOMERA and D. A. MONGRU and Z. OLTVAI and A.-L. BARABÁSI, *Science*, **297** (2002) 1551.
- [20] V. SPIRIN and L. A. MIRNY, *Proc. Natl. Acad. Sci. USA*, **100** (2003) 12123.
- [21] J. P. ONNELA and A. CHAKRABORTI and K. KASKI and J. KERTÉSZ and A. KANTO, *Phys. Rev. E*, **68** (2003) 056110.
- [22] D. J. WATTS and P. S. DODDS and M. E. J. NEWMAN, *Science*, **296** (2002) 1302.
- [23] J. VUKOV and Gy. SZABÓ, *Phys. Rev. E*, **71** (2005) 036133.
- [24] Gy. SZABÓ and J. VUKOV and A. SZOLNOKI, *cond-mat/0506433*, (2005) .
- [25] T. VICSEK, *Nature*, **418** (2002) 131.
- [26] R. GUIMERÀ and L. DANON and A. DÍAZ-GUILERA and F. GIRALT and A. ARENAS, *Phys. Rev. E*, **68** (2003) 065103.
- [27] A. C. GAVIN *et al.*, *Nature*, **415** (2002) 141.
- [28] P. EDWARDS CARRINGTON and J. SCOTT and S. WASSERMAN, *Models and Methods in Social Network Analysis*, edited by K. FAUST, Ch. 7 (Cambridge University Press, New York) 2005.
- [29] G. PALLA and I. DERÉNYI and I. FARKAS and T. VICSEK, *Nature*, **435** (2005) 814.
- [30] A. L. BARABÁSI and H. JEONG and Z. NÉDA and E. RAVASZ and A. SCHUBERT and T. VICSEK, *Physica A*, **311** (2002) 590.
- [31] H. JEONG and Z. NÉDA and A.-L. BARABÁSI, *Europhysics Letters*, **61** (2003) 567.
- [32] M. E. J. NEWMAN, *Phys. Rev. E*, **64** (2001) 025102.
- [33] S. WARNER, *Library Hi Tech*, **21** (2003) 151.
- [34] I. DERÉNYI and G. PALLA and T. VICSEK, *Phys. Rev. Lett.*, **94** (2005) 160202.
- [35] <http://angel.elte.hu/clustering>.
- [36] J. D. NOH and H.-C. JEONG and Y.-Y. AHN and H. JEONG, *Phys. Rev. E*, **71** (2005) 036131.
- [37] R. GUIMERÀ and B. UZZI and J. SPIRO and L. A. N. AMARAL, *Science*, **308** (2005) 697
- [38] J. J. RAMASCO and S. A. MORRIS, *physics/0509247*, (2005) .