Reply networks on a bulletin board system

Kou Zhongbao and Zhang Changshui
State Key Laboratory of Intelligent Technology and Systems, Department of Automation, Tsinghua University, Beijing 100084, People's Republic of China

(Received 11 April 2002; revised manuscript received 16 September 2002; published 24 March 2003)

Using data from a bulletin board system (BBS), we constructed reply networks for various boards, which can be considered as social networks connecting people of the same interests. In these networks, identifications (IDs) are treated as nodes and reply articles set up links. We investigated some statistics on these reply networks, such as clustering coefficients, characteristic path lengths, degree distributions, etc., and showed small-world characteristics and scale-free degree distributions of these reply networks. Then we put forward a model of interest space, which is the basis of the reply networks. We indicated that the hierarchical and clustering structure of the interest space, together with overlapping interests of IDs not only result in small-world characteristics of reply networks on BBS, but also give rise to preferential attachment, which is a popular explanation for scale-free features.

DOI: 10.1103/PhysRevE.67.036117 PACS number(s): 89.75.Hc, 89.65.−s, 89.70.+c

I. INTRODUCTION

The reply networks studied in this paper can be considered as a form of social network belonging to research areas in complex networks [1,2].

A social network is a set of people or groups each of which has connections to some or all of the others [3]. Then each person is treated as a node and each connection as a link in network terminology. In different social networks, nodes and links are assigned different meanings. For example, in film actor collaboration networks [4,5], actors in movies are nodes, and if they perform in the same movie, links are established between them. In scientific collaboration networks [3,6–8], scientists who publish papers are nodes; two scientists are considered connected if they have co-authored one or more papers together. In citation networks [9,10], each paper is treated as a node; the citation references setup links between papers; and in email networks [11], email accounts are nodes and email transmissions construct links.

In this paper, we will construct reply networks according to boards on a bulletin board system (BBS). In these networks, identifications (IDs) registered on BBS are nodes, and a link is set up between two IDs when one of them replies to the articles posted by the other. Two IDs connected by a link have conversed on the same topics, hence they are more likely to enjoy the same interests. So the reply networks can be considered as social networks connecting people of the same interests, in other words, they act as a kind of interest network.

Here, we focus on statistics of different reply networks to acquire some common characteristics, and try to reveal the mechanisms behind these characteristics. It will provide a sample for research into complex networks.

However, research on BBS may not just be limited to statistics of reply networks, other efforts may also be made in future as the following.

(1) Dynamics of reply networks, e.g., forming and evolving of this kind of network.

(2) Research on sociology, e.g., make comparisons among people's interests, or analyze behaviors of individual ID's to tell characteristics of people's interests such as intensity and transition of an interest.

(3) Research on topics, which act as dynamics on BBS, and show propagation of thoughts and opinions, such as how an interesting topic forms and evolves, and how people's discussions diverge from their initial themes.

Besides these, some phenomena on BBS may be similar to those in reality. For example, an analogy can be made between the attraction of boards to users and a retail store's attraction of customers. Data on BBS are much easier to acquire, so research on BBS may provide insights to other social problems where data collection is difficult.

The remainder of this paper is organized as follows. An introduction to BBS and descriptions of research data are given in Sec. II. Statistics and results on reply networks are shown in Sec. III. In Sec. IV, a model of interest space is put forward to explain obtained results. Finally, a conclusion is drawn.

II. REPLY NETWORKS ON BBS

In this section, we introduce BBS and the research data we used, and also define the construction of reply networks.

A. Introduction to BBS

BBS is an electronic message center. Most bulletin boards serve specific interest groups. IDs registered by users are actors on BBS, through which one can review messages left by others and leave one's own messages if one wants.

Bulletin boards are labeled by their names, which are expressive of the content of their articles. Each article on a board is posted by a special ID and involves a special topic. Articles can be classified into two types: initial articles and reply articles. An initial article is the first post of a topic or the origin of a discussion, while a reply article comments on an initial article or another reply article so as to continue the discussion. Only the reply articles can form links between IDs on reply networks.

TABLE I. Statistics on reply networks formed by selected boards.

Board	Number of articles	Number of IDs	Average number of links	Clustering coefficient	Characteristic path length	Exponent for degree distribution	Remarks
THUExpress	163011	14838	10.46	0.091	3.501	1.834	Current affairs
AdvancedEdu	109329	6442	12.00	0.129	3.223	1.685	Study abroad
Love	73953	7473	10.11	0.101	3.562	1.807	Love affairs and romance
WorldSoccer	58146	2270	13.97	0.242	3.086	1.585	Football all over the world
TV	50669	4833	7.50	0.112	3.525	1.898	TV
BattleNet	50593	1575	9.11	0.308	3.021	1.743	Network based games

B. Data and preprocessing

The data used in this paper was downloaded from SMTH, which is the biggest bulletin board system of the People's Republic of China and is owned by Tsinghua University. It consists of more than 230 boards. The total number of registered IDs is more than 80 000 and the number of live users is often over 2000, sometimes reaching 3000. The number of articles posted daily is above 50 000. So SMTH provides a very good data resource for our research on reply networks.

In this paper, we select six boards with lots of articles from SMTH to collect statistics for reply networks. The selected boards include THUExpress, AdvanceEdu, Love, WorldSoccer, TV, and BattleNet. The articles posted between Oct. 28, 2001 and Dec. 29, 2001 were used to form our reply networks. Descriptions of these selected boards and statistics on the corresponding reply networks are given in Table I.

Each article on SMTH contains the current ID (the user ID posting the current article), title, board information, date and time, contents; if the post is a reply article, replied ID (the user ID who posted the article that the current article comments on), and replied contents are also included.

Construction of a reply network is straightforward. An article is parsed to extract the different components mentioned above. For an initial article, a node is added to the reply network when the current ID is a new one for the network. For a reply article, when its current ID or replied ID are new ones (the absence of replied IDs in current networks may arise from previous posts not captured in the sample), corresponding nodes are added; at the same time, a link is established between these nodes if they have not been previously connected together. Then, when all the articles on one board are processed, a reply network corresponding to this board is constructed.

Figure 1 shows a typical reply article on SMTH, in which all necessary components are marked. We can get the current ID and the replied ID of this article, which are *bbb* and *aaa*, respectively. Then nodes *bbb* and aaa are added when they are new for the reply network, and a link between them is set up if these two IDs have not been previously connected together.

III. STATISTICS ON REPLY NETWORKS

Statistics such as clustering coefficients, characteristic path lengths, degree distributions, etc. are given in this part to reveal the structure of the reply networks on BBS.

A. Reply networks are small-world networks

A common property of social networks is that clique's form representing circles of friends or acquaintances in which every member knows every other member. This inherent tendency to clustering is quantified by the clustering coefficient. For a selected node i connected with k_i other nodes in a network, if there actually exist e_i links among these k_i neighbors, the value of the clustering coefficient of node i can be calculated by

$$C_i = \frac{e_i}{k_i(k_i-1)/2}, \tag{1}$$

where the denominator $k_i(k_i-1)/2$ is the maximum number of links that may exist among these k_i nodes. The clustering coefficient C of the whole network is the average of all individual C_i. In a real social network of acquaintances, C reflects the average extent to which friends of given individuals' are also friends of each other [1,12].

The path length is the number of links in the shortest path between two nodes. The characteristic path length L is the path length averaged over all pairs of nodes [13].

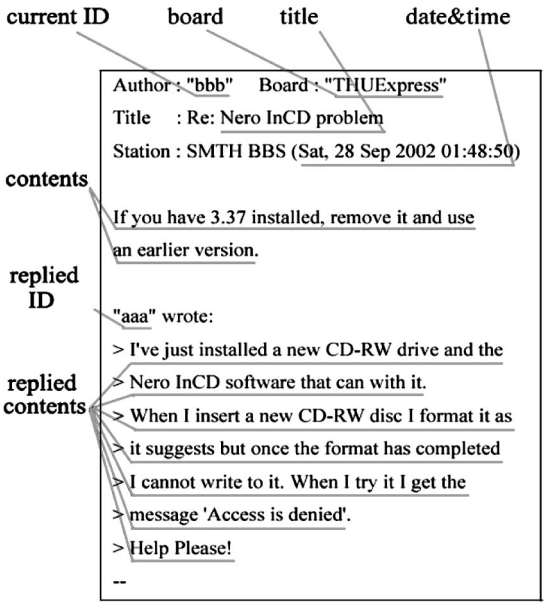

FIG. 1. A typical reply article with necessary components marked.

TABLE II. C and L for reply networks of the selected boards.

Board	C	C_{rand}	L	L_{rand}
THUExpress	0.091	0.000705	3.501	3.614
AdvancedEdu	0.129	0.001863	3.233	3.530
Love	0.101	0.001353	3.562	3.855
WorldSoccer	0.242	0.006157	3.086	2.930
TV	0.112	0.001552	3.525	4.210
BattleNet	0.308	0.005788	3.521	3.332

In reply networks of different boards on BBS, C and L can be acquired in the same manner.

Watts and Strogatz first pointed out that in most, if not all, real networks in nature, in man-made systems, and in society, there are comparable characteristic path lengths and much larger clustering coefficients than those in random networks with equal number of nodes and links. These kinds of networks are so-called small-world networks [5].

In Table II, C and L of the selected boards are listed. For comparison, C_{rand} and L_{rand} of corresponding random networks are listed together, which are calculated by formulas in Ref. [14]. It seems that reply networks on BBS also exhibit small-world topologies.

Large C values indicate that discussions can be put forward among bunches of IDs easily. Small L values indicate that ideas and opinions can propagate rapidly from one person to another. So the small-world topologies of reply networks on BBS ensure the propagation of discussions among IDs.

B. Degree distributions are scale-free

Not all nodes in a network have the same number of links. The spread in the number of links a node has, i.e., degree of a node, is characterized by degree distribution marked as $P(k)$. $P(k)$ gives the number of IDs which have exactly k links on reply networks. Here we only use the word ''distribution'' but no normalization is done.

The degree distributions for reply networks of selected boards on SMTH are all roughly scale-free, which accords with a large number of real networks [1,2]. That is, the distributions follow the formula

$$P(k) \sim k^{-\gamma_k}. \tag{2}$$

A scale-free distribution shows a straight line on a logarithmic scale histogram [15], and in which, the value of exponent γ_k corresponds to the slope of the straight line.

Figure 2 shows degree distributions of the boards THUExpress and WorldSoccer on logarithmic scale histograms, which approximate straight lines with exponents 1.834 and 1.585, respectively. As a result, they show scale-free distributions. γ_k values of the selected boards vary from 1.585 for board WorldSoccer to 1.898 for board TV. In Fig. 2, the data points of board THUExpress are higher than those of board WorldSoccer, this is because THUExpress has more visitors.

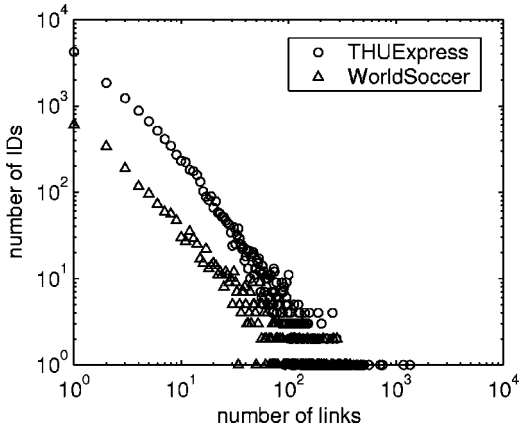

FIG. 2. Degree distributions for reply networks of boards THUExpress and WorldSoccer. Both of them fit scale-free forms with different exponents.

IV. DISCUSSION

A. Previous explanations

Small-world properties and scale-free degree distributions are shared by many complex networks [5,1,2] including many social networks [1,2]. Some models have been proposed to explain their formation.

Quantitative studies of the ''small-world phenomenon'' were first performed by Milgram in 1967 [16], which was popularly known as ''six degrees of separation'' between any two people in the United States. Watts and Strogatz [5] first proposed a model to generate graphs with high clustering coefficients and small L. They also concluded that small-world phenomenon arises by a few ''long-range'' connections in the otherwise short-range structure of a social network. Alternatively, Kasturirangan [17] put forward that a few nodes that are linked to a widely distributed set of neighbors cause the small L of the network. However, Newman [18] argued that real networks are perhaps roughly regular lattices of very high dimension, which may cause small-world characteristics of networks. Later, Davidsen and coworkers [14] modeled acquaintance networks and pointed out that introduction by common acquaintance gives rise to small L, together with large C in these kind of networks. Mathias and Gopal [19] studied that the small-world topology arises as a consequence of a trade-off between maximal connectivity and minimal wiring.

The origin of the scale-free degree distribution was first addressed by Barabási and Albert [20]. Later, others improved and expanded it [15,21,22]. In these models, the two basic requirements which result in scale-free degree distributions lie in the fact that networks expand continuously by the addition of new vertices, and new vertices attach preferentially to the already well connected ones. Instead of introducing preferential attachment explicitly, mechanisms of placing nodes and edges in some models are designed to introduce it implicitly [23,24]. There are also mechanisms of growth and preferential attachments in many social networks [8,25].

Evolving models of networks mentioned above are usually good for explaining scaling, but not for explaining high clustering or short path length. Therefore, models simulating

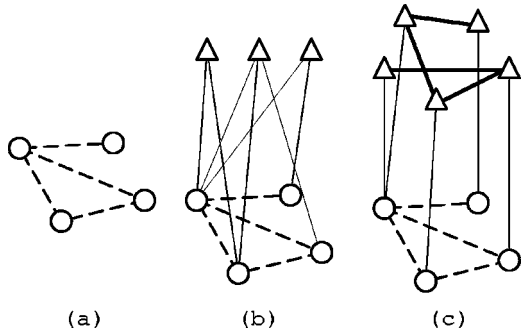

FIG. 3. Types of social networks. In all cases, circles represent nodes of the formed networks (i.e., actors), triangles denote mediums, dash lines are links between nodes, solid lines tie actors and mediums together, and bold lines connect mediums. (a) Nonbipartite networks; (b) separate bipartite networks, where nodes are connected by their common mediums; (c) connected bipartite networks, where mediums themselves form a network, and a network of actors is formed indirectly.

different mechanisms in real networks are also proposed to acquire networks concerning both aspects [26–29].

B. Two-level structure of reply networks

Many social networks can be represented by the bipartite graph [2] containing two distinct types of vertices, which are called actors and mediums. In this kind of network, two actors are often tied together by the mediums that connect them. For example, in collaboration networks of film actors or scientists, performing artists are tied by movies, and scientists are tied by their common papers; in email networks, accounts are tied by emails, etc. There are also some social networks whose links are set up directly with no medium involved, such as acquaintance networks, WWW, etc.

Then, according to the manner of links being established, social networks can be classified into following three types.

(1) Nonbipartite networks, in which nodes are linked directly, shown in Fig. 3(a). Examples include acquaintance networks, WWW, etc.

(2) Separate bipartite networks, in which mediums are not connected directly, shown in Fig. 3(b). The cases stated above reside in this type, such as collaboration networks, email networks, etc.

(3) Connected bipartite networks, in which mediums themselves form a network, shown in Fig. 3(c).

Reply networks on BBS belong to connected bipartite networks. In reply networks, IDs are actors and articles they post are mediums. Each article links one ID, i.e., its author, and articles are connected by replying behaviors. There are actually two levels of networks: The network of articles and the network of IDs. The latter are set up according to the former. In fact, in citation networks, if the objects to be considered are not papers but scientists, the citation networks of authors discussed in Ref. [30] are acquired, which are also connected bipartite networks.

The process of articles being posted on one board is just the process of reply networks forming on BBS, so reply networks are surely evolved gradually. It seems reasonable to propose an evolving model; but complexity of the structure makes them difficult to analyze. Alternatively, we put forward a model of interest space to explain both small-world and scale-free features of reply networks.

C. Explanations by the model of interest space

One article on BBS may involve different aspects of interests, and different articles that reply to it may focus on different aspects. We call them different interest points. For example, for an article concerning ''A privately held company said on Friday it had produced the first clone of a human being, without offering any proof,'' different persons may care for different things, someone may be excited, some may doubt it, and others may make criticisms from the viewpoint of humanitarianism. Then the articles they post to discuss the former one may focus on these different aspects, which do just act as interest points.

Suppose there is an interest space, i.e., a space composed of interest points. In this space, points which are near to each other in distance represent ''similar'' interests. The definition of similarity here may lie in semantics, or may be acquired by statistics of people's behaviors. For example, the number of people who like both TV and movies are more than those who care for religion and pop music together, so the interest points concerning TV and movie should be nearer to each other than the latter ones.

There may be different aspects for a special kind of interest dependency, and each aspect may also include more subordinate ones, etc. That is to say, interest space is hierarchical. For example, all interest points concerning sports may involve different kinds of sports, such as football, basketball, tennis, etc. Furthermore, interest points regarding a special kind of sport, such as football, often focus on various things about it, such as players, teams, and matches, etc. Besides this, the interest points concerning a narrower scope are often about the more correlated interests, and they are usually closer to each other in interest space. In the example mentioned above, interest points on football are more collective than those on sports. So the interest space is locally clustering.

Ravasz [31] showed that clustering and scale-free of a network are the consequences of its hierarchical organization. However, the hierarchy of the interest space is imaginary, different from the hierarchical structure of a real network. Then do the network based on the hierarchical interest space still own the small-world and scale-free characteristics?

Based on hierarchical interest space, article level networks, together with ID level networks are set up, as shown in Fig. 4. For each reply network corresponding to a board on BBS, a pair of article level and ID level networks are based on a subspace of the interest space.

Each article may cover several interest points in interest space and each ID may post several articles, so a reply network of IDs may be caused by overlap of interest points that each ID covers. That is to say, each ID may cover a few interest points, and IDs who own common interest points are easier to be connected together in a reply network.

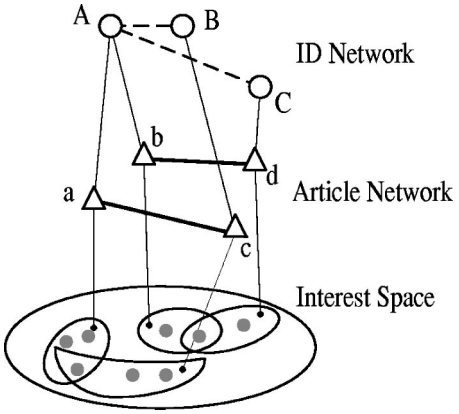

FIG. 4. Construction of a reply network based on interest space. An interest subspace that consists of eight interest points (solid circle) is shown. Each article (triangles) covers several interest points, where *a* and *c* cover a common point and are connected by replying; so do *b* and *d*. IDs (circles) *A*, *B*, and *C* post the four articles, and they are tied when their articles have been connected. So networks of the three IDs form.

The structure of interest space and overlapping interests of IDs may influence the structure of reply networks in two aspects.

First, when the interest subspace used to construct the reply networks keeps a narrow range, there may be some IDs who cover almost all its interest points, which are often highly collective. These IDs may connect to many other IDs and own a high degree in corresponding reply networks. As in Ref. [17], the existence of these IDs makes a small characteristic path length in reply networks of IDs.

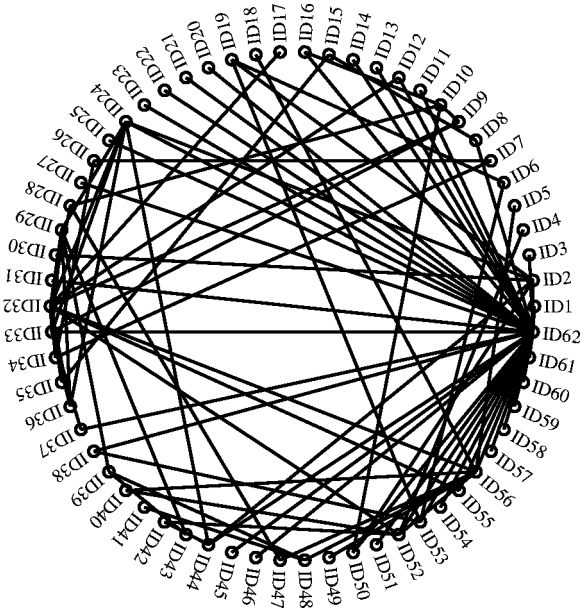

FIG. 5. A small network constructed from an interesting topic. Each point on the circle denotes an ID who engages in the topic, and lines between points denote links set up by reply articles. IDs are all marked beside the corresponding points. ID62 initiated the discussion.

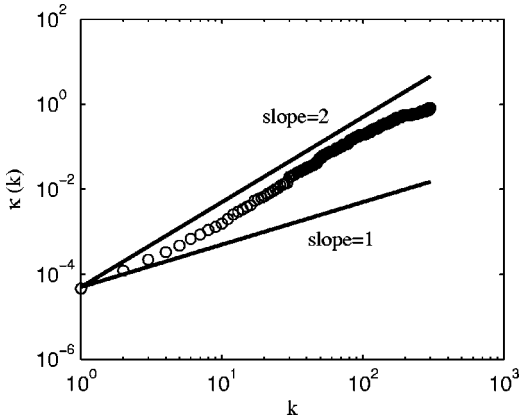

FIG. 6. Cumulated preferential attachment measure $\kappa(k)$ for board THUExpress on logarithmic scales histogram, where k denotes degree of nodes. Here, the articles posted before Nov. 31, 2001 are used to construct the reply network and the rest of the data to measure preferential attachment. The straight lines shown are with slopes of 1 and 2, respectively. The data points are consistent with the line with a slope of 2. So linear preferential attachment seems to exist in reply networks of board THUExpress.

Second, some interest points may be covered by many people. Then when they are involved in discussion, interesting topics that include many articles are probably formed. Then IDs who engage in discussion of the topics may connect largely to each other.

Figure 5 shows a small network constructed from an interesting topic on board Love. There are 167 articles on this topic and 62 IDs who engaged in its discussion forming 91 links. This small network yields a clustering coefficient of 0.202, larger than that of the full board (Love), and a characteristic path length of 2.745, shorter than that of Love, see Table II. So common interests of people may be one factor of small-world features on reply networks.

Different IDs may cover different numbers of interest points. Then when a new reply article is posted, it's interest points are more likely to be covered by IDs with many interest points. On the other hand, more interest points of an ID mean that one may have more chance to connect to others when reply networks are being constructed. So IDs owning more links are more likely to acquire a new link, which is just the definition of preferential attachment.

We use the method in Ref. [32] to measure preferential attachment in reply networks on BBS. Figure 6 shows the cumulated preferential attachment distribution $\kappa(k)$ of board THUExpress on a logarithmic scale histogram, where the data points are consistent with a straight line with a slope of 2. That is to say, $\kappa(k)$ approximates to $k^{\alpha+1}$, where $\alpha \approx 1$. So the preferential attachment function $\Pi(k)$ approximates to $k^{\alpha}(\approx k)$. All other selected boards have the same results. Then there exist linear preferential attachment in reply networks. So evolving and preferential attachment of reply networks on BBS cause scale-free degree distributions of them, which has been stated in Sec. IV A.

V. CONCLUSIONS

In this paper, we have studied reply networks on BBS; in which IDs who have posted articles on boards are nodes, and

reply articles set up links. Using the data downloaded from the biggest BBS of the People's Republic of China SMTH, we have constructed reply networks for several selected boards.

We have investigated some statistics on these reply networks and found that the reply networks are small-world networks with high clustering coefficients and short characteristic path lengths, and their degree distributions are scale-free.

Different from other social networks, reply networks are connected bipartite networks, that is to say, they are composed of two levels of networks. The complexity of the structure forces us to give up providing an evolving model but to put forward a model of interest space to explain the mechanism of small-world and scale-free features.

In our model, the interest space is hierarchical and locally clustering, and linking is motivated by the overlapping interests of different IDs. The structure of interest space, together with overlap of IDs' interests, not only result in small-world topologies of reply networks on BBS, but also give rise to preferential attachment, which is a popular explanation for scale-free characteristics.

The study of reply networks on BBS opens up a good method for exploring people's interests. A BBS also offers potentially useful data for research of other topics, which have been mentioned in Sec. I. We have only just commenced the investigation.

[1] R. Albert and A.L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[2] S.N. Dorogovtsev and J.F.F. Mendes, Adv. Phys. **51**, 1079 (2002).
[3] M. E. J. Newman, Phys. Rev. Lett. (to be published).
[4] J.J. Collins and C.C. Chow, Nature (London) **393**, 409 (1998).
[5] D.J. Watts and S.H. Strogatz, Nature (London) **393**, 440 (1998).
[6] M.E.J. Newman, Phys. Rev. E **64**, 016131 (2001).
[7] M.E.J. Newman, Phys. Rev. E **64**, 016132 (2001).
[8] M.E.J. Newman, Phys. Rev. E **64**, 025102 (2001).
[9] S. Redner, Eur. Phys. J. B **4**, 131 (1998).
[10] A. Vazquez, e-print cond-mat/0105031.
[11] H. Ebel, L.I. Mielsch, and S. Bornholdt, Phys. Rev. E **66**, 035103 (2002).
[12] G.B. Lubkin, Phys. Today **51**(9), 17 (1998).
[13] T. Walsh, in *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, 1999* (Morgan Kaufmann Publishers, San Francisco, 1999).
[14] J. Davidsen, H. Ebel, and S. Bornholdt, Phys. Rev. Lett. **88**, 128701 (2002).
[15] L.A.N. Amaral, A. Scala, M. Barabási, and H.E. Stanley, Proc. Natl. Acad. Sci. U.S.A. **97**, 11149 (2000).
[16] S. Milgram, Psychol. Today **2**, 60 (1967).
[17] R. Kasturirangan, e-print cond-mat/9904055.
[18] M.E.J. Newman, J. Stat. Phys. **101**, 819 (2000).
[19] N. Mathias and V. Gopal, Phys. Rev. E **63**, 021117 (2001).
[20] A.L. Barabási and R. Albert, Science **286**, 509 (1999).
[21] P.L. Krapvisky, S. Redner, and F. Leyvraz, Phys. Rev. Lett. **85**, 4629 (2000).
[22] A.L. Barabási, R. Albert, and H. Jeong, Physica A **281**, 69 (2000).
[23] A. Vazquez, e-print cond-mat/0006132.
[24] J. M. Kleigberg, R. Kumar, P. Raghavan, S. Rajagopalan, A. S. Tomkins, in *Proceedings of the Fifth Annual International Conference on Combinatiorcs and Computin, COCOON*, edited by T. Asano *et al.* (Springer-Verlag, Berlin, 1999), p. 1627.
[25] A.L. Barabási, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek, Physica A **311**, 590 (2002).
[26] K. Klemm and V.M. Eguiluz, Phys. Rev. E **65**, 036123 (2002).
[27] K. Klemm and V.M. Eguiluz, Phys. Rev. E **65**, 057102 (2002).
[28] A.R. Puniyani, R.M. Lukose, and B.A. Huberman, e-print cond-mat/0107212.
[29] E.M. Jin, M. Girvan, and M.E.J. Newman, Phys. Rev. E **64**, 046132 (2001).
[30] J. Laherrere and D. Sornette, Eur. Phys. J. B **2**, 525 (1998).
[31] E. Ravasz and A.L. Barabási, e-print cond-mat/0206130.
[32] H. Jeong, Z. Neda, and A.L. Barabási, e-print cond-mat/0104131.