ALPHABETIC BIPARTITE NETWORK (α-BiN): THEORY AND CASE STUDY*

A. Maiti†, N. Ganguly‡

Department of Computer Science and Enginering, IIT Kharagpur
Kharagpur 721302, India

(Received January 12, 2010)

Recently, much attention has been paid in analyzing and modeling bipartite network (BNW) due to its importance in many fields like information science, biology, social science, economics. Here we have emphasized on growth of a special type of BNW where the number of nodes in one set is almost fixed. This type of systems can be represented as an Alphabetic Bipartite Network (α-BiN) where there are two kinds of nodes representing the elementary units and their combinations, respectively [5]. There is an edge between a node corresponding to an elementary unit u and a node corresponding to a particular combination v if u is present in v. The partition consisting of the nodes representing elementary units is fixed, while the other partition is allowed to grow unboundedly. In this paper we reveal some characterizations of α-BiN growth and give a real world example of α-BiN. We have done extensive experiments by means of computer simulations of different growth models of α-BiN to characterize them. We present a practical example of this type of networks, *i.e.* protein protein complex network where set of proteins are fixed and set of complexes are growing.

PACS numbers: 89.75.–k, 89.75.Fb, 02.10.Ox

1. Introduction

The Bipartite Networks (BNWs) are a special class of networks whose nodes can be divided into two sets, "top" and "bottom", and edges only connect two nodes from different sets [13]. Formally, a bipartite network or bipartite graph G is a 3-tuple $\langle U, V, E\rangle$, where U and V are mutually exclusive finite sets of nodes (also known as the two partitions) and $E \subseteq U \times V$ is the set of edges that run between these partitions [12]. Without loss of generality, we denote partition V as the top set of nodes while partition U as bottom set.

* Presented at the Summer Solstice 2009 International Conference on Discrete Models of Complex Systems, Gdańsk, Poland, June 22–24, 2009.

† abyaym@cse.iitkgp.ernet.in

‡ niloy@cse.iitkgp.ernet.in

Considerable number of real world systems can be naturally modeled as BNW: The article-author networks [4, 14] have two sets of nodes as the article set and author set, the human sexual network [15] consists of men and women, *etc.* Bipartite network or bipartite graph has been a well studied subject in graph theory and discrete mathematics from the early twentieth century, the studies mainly deal with the static properties of comparatively small graphs. However with the advent of Internet and accessibility to the huge data, researches are directed towards understanding the properties of large real life BNWs. This is done following the overall framework of the complex network theory [2, 7, 18], a new technology that has emerged at the beginning of the current century and helps in analyzing these complex and vast network data.

Considerable amount of research have been done to explain the dynamics of the BNWs. Growth is the most significant dynamics that takes place in BNWs. Most of the studies of growth model in the past assume that both the partitions of the BNW grow with time. Several models have been proposed to synthesize the structure of these BNWs, *i.e.*, when both the partitions grow unboundedly [1, 4, 20, 22]. It has been found that for such growth models, when each incoming *top* node preferentially attaches itself to the *bottom* nodes, the emergent degree distribution of the *bottom* nodes follows a power-law [22]. This result is reminiscent of unipartite networks where preferential attachment results in power law degree distributions [3].

On the other hand, α-BiN where one of the partitions remains fixed (*i.e.*, the number of *bottom* nodes are constant) over time have received comparatively much less attention. However, on inspection, it is found that in many of the BNWs one set of nodes is almost fixed compared to the other much faster growing set. For example, it is reasonable to assume that for the city-people network [11], the city growth rate (emergence of new cities) is close to zero compared with the population growth rate. In biology Codon–Gene network is a very suitable example because the number of codons is fixed to 64 while new genes appear in the network over time. Another example of this type could be the phoneme-language network [6,17], in linguistics. The initial systematic and analytical study of α-BiN has been presented by us in [5, 21], a review of which is presented here.

The organization of this paper is as follows. Sec. 2 gives a short description of the basic growth models of α-BiN with various model parameters. Sec. 2.4 and 2.5 present detail experiments carried out to characterize the growth models. We have enlisted several interesting observations which eventually reveal insights of α-BiN growth and consequently characterize the growth process. We present the protein–protein complex network as our case study in Sec. 3. In the last section we conclude and point to possible future works.

2. Growth of α-BiN

2.1. Growth model

Basic growth model of α-BiN is as follows: At each time step, one *top* node v is introduced with μ number of edges to attach with the fixed set of *bottom* nodes. Here if the *top* nodes bring single edge ($\mu = 1$), then the model is termed as *sequential attachment model*. If the *top* nodes bring multiple edges ($\mu > 1$), then the corresponding model is named as *parallel attachment model*. Every bottom node $u \in U$ has some attachment probability to get attached to the incoming μ edges. According to the probability distribution, the μ edges will attach to the chosen bottom nodes.

The growth model of [5, 21] incorporates preferential attachment along with a tunable randomness parameter. Suppose that the bottom partition U has fixed N nodes labeled as u_1 to u_N. At each time step, a new node is introduced in the top set V which connects to μ nodes in U by means of μ edges based on a predefined attachment rule. The theoretical analysis assumes that μ is a constant greater than 0. The next task is to derive the attachment probability of each bottom node for attaching itself with μ new incoming edges.

2.2. Attachment probability

Let $\widetilde{A}(k_i^t)$ be the probability of attaching a new edge to a node u_i, where k_i^t refers to the degree of the node u_i at time t. So $\widetilde{A}(k_i^t)$ defines the attachment kernel that takes the form:

$$\widetilde{A}(k_i^t) = \frac{\gamma k_i^t + 1}{\sum_{j=1}^{N}(\gamma k_j^t + 1)}\,, \tag{1}$$

γ is a tunable parameter which specifies the relative weight of preferential to random attachment. The higher value of γ indicate the low randomness in the system and *vice versa*.

We are mainly interested to evaluate the degree distribution of the nodes in the bottom set (U). Degree distribution of U is denoted by $p_{k,t}$ in the rest of the paper. Essentially, $p_{k,t}$ is the probability that a randomly chosen bottom node has degree k after t time steps.

2.3. Characterizing parameters

The parameters which are important to characterize bipartite networks are number of bottom nodes (N), number of top nodes which is equal to the time (t) because at each time step one top node is introduced, number of parallel edges (μ) attached at a time (assumed to be equal in every time

step) and the preferentiality parameter (γ). We have also identified two most prominent measuring parameters in evolved degree distribution which can be used to characterize the growth of α-BiN. These are as follows:
Mode: *Mode* of a distribution means the degree where the distribution reaches its maximum value. Note that mode zero signifies that the number of nodes with zero degree is maximum. Therefore, first occurrence of mode as zero is itself a pointer to the critical transition of the growth process.
Critical γ (γ_c): In our model γ signifies the relative magnitude of randomness and preferentiality, so we term the minimum value of γ at which the distribution shows its mode at $k = 0$ as critical γ or γ_c.

Here we try to observe the behavior of mode and the γ_c of the α-BiN with respect to the model parameters. From these experiments we have obtained many interesting observations. We report two such prominent observations.

1. In the case of $\mu \ll N$, bottom node degree distribution shows $\gamma_c = 1$.
2. For larger μ, the value of γ_c is normally greater than one. When the value of μ is quite large ($\mu > N$ or μ is in the order of N), then degree distribution shows two local maxima. Note that the higher one is actually the mode of the distribution.

2.4. Case 1: $\mu \ll N$

In [21] it has been shown that $p_{k,t}$ can be approximated for $\mu \ll N$ and small values of γ by integrating:

$$p_{k,t+1} = (1 - A_p(k,t))p_{k,t} + A_p(k-1,t)p_{k-1,t}\,, \tag{2}$$

where $A_p(k,t)$ is defined as

$$A_p(k,t) = \begin{cases} \dfrac{(\gamma k + 1)\,\mu}{\gamma\mu t + N}\,, & \text{for } 0 \le k \le \mu t\,, \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{3}$$

for $t > 0$ while for $t = 0$, $A_p(k,t) = (\mu/N)\delta_{k,0}$. The solution of Eq. (2) with the attachment kernel given by Eq. (3) reads:

$$p_{k,t} = \binom{t}{k} \frac{\prod_{i=0}^{k-1}(\gamma i + 1)\prod_{j=0}^{t-1-k}\left(\frac{N}{\mu} - 1 + \gamma j\right)}{\prod_{m=0}^{t-1}\left(\gamma m + \frac{N}{\mu}\right)}\,. \tag{4}$$

Note that, for $\mu = 1$, *i.e.* for sequential attachment, Eq. (4) is the exact solution of the process. Interestingly, for $\gamma > 0$, Eq. (4) approaches, asymptotically with time, a beta-distribution as follows:

$$p_{k,t} \simeq C^{-1}\left(\frac{k}{t}\right)^{\gamma^{-1}-1}\left(1 - \frac{k}{t}\right)^{\eta-\gamma^{-1}-1}\,. \tag{5}$$

Here, C is the normalization constant and $\eta = N/(\gamma\mu)$. By making use of the properties of beta distributions, it is clear that depending on the value of γ, $p_{k,t}$ can take one of the following distinctive functional forms:
(a) $\gamma = 0$, a binomial distribution whose mode shifts with time,
(b) $0 < \gamma < 1$, a skewed (normal) distribution which exhibits a mode that shifts with time,
(c) $1 \leq \gamma \leq (N/\mu) - 1$, a monotonically decreasing (near exponential) distribution with the mode frozen at $k = 0$, and
(d) $\gamma > (N/\mu) - 1$, a u-shaped distribution with peaks at $k = 0$ and $k = t$.

Fig. 1 illustrates these possible four regimes. Note that in regimes (a) and (b), mode of the distribution is greater than zero. At $\gamma = 1$, distribution becomes monotonically decreasing with mode at zero, *i.e.* $\gamma_c = 1$.

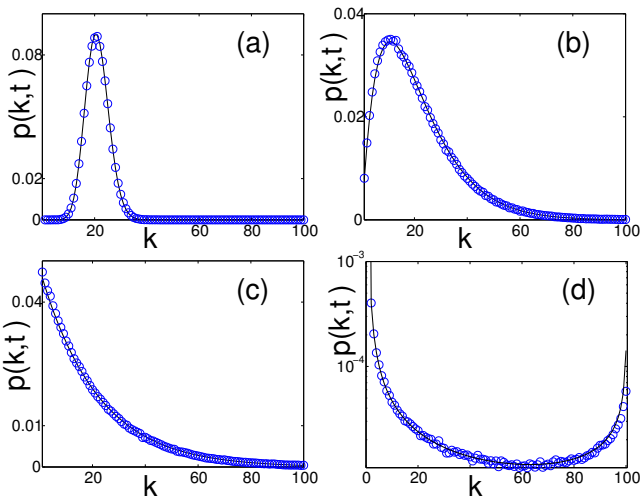

Fig. 1. The four possible degree distributions depending on γ for parallel attachment. Symbols represent average over 5000, in (a)–(c), and 50000, in (d), stochastic simulations. The solid curve is the theory given by Eq. (4). From (a) to (c), $t = 1000$, $N = 1000$ and $\mu = 20$. (a) at $\gamma = 0$, $p(k,t)$ becomes a binomial distribution. (b) $\gamma = 0.5$, the distribution exhibits a maximum which shifts with time for $0 \leq \gamma < 1$. (c) $\gamma = 1$, $p(k,t)$ does no longer exhibit a shifting maximum and it is a monotonically decreasing function of k for $1 \leq \gamma \leq (N/\mu) - 1$. (d) $\gamma = 2500$, $t = 100$, $N = 1000$ and $\mu = 1$. $p(k,t)$ becomes a u-shaped curve for $\gamma > (N/\mu) - 1$.

2.5. Case 2: $\mu > N$ or $\mu \approx N$

In the case of $\mu \ll N$, derivation of degree distribution of bottom nodes assumes that at each time step one bottom node can be attached with at most one edge. But if $\mu > N$ or μ is in order of N then there is very high

chance that one bottom node gets multiple edges. To address this issue we extended the work of [21] and introduced the correct expression for the degree distribution of bottom nodes [5]. In any time step t, a bottom node can get any number of edges between zero and μ from the incoming top node. Hence, the correct expression for the evolution of $p_{k,t}$ has the form:

$$p_{k,t+1} = \left(1 - \sum_{i=1}^{\mu} \widehat{A}(k,i,t)\right) p_{k,t} + \sum_{i=1}^{\mu} \widehat{A}(k-i,i,t) p_{k-i,t}\,, \tag{6}$$

where $\widehat{A}(k,i,t)$ represents the probability at time t of a node of degree k of receiving i new edges in the next time step. The term $\sum_{i=1}^{\mu} \widehat{A}(k,i,t)p_{k,t}$ describes the number of nodes of degree k at time t that change their degree due to the attachment of 1, 2, ..., or μ edges.

On the other hand, nodes of degree k will be formed at time $t+1$ by the nodes of degree $k-1$ at time t that receive 1 edge, nodes of degree $k-2$ at time t that receive 2 edges, and so on. This process is described by the term $\sum_{i=1}^{\mu} \widehat{A}(k-i,i,t)p_{k-i,t}$. The expression for $\widehat{A}(k,i,t)$ is derived as

$$\widehat{A}(k,i,t) = \binom{\mu}{i} \left(\frac{\gamma k+1}{\mu\gamma t+N}\right)^{i} \left(1 - \frac{\gamma k+1}{\mu\gamma t+N}\right)^{\mu-i}. \tag{7}$$

The evolution formula of Eq. (6) can be expressed as

$$p_{k,t+1} = \sum_{i=0}^{\mu} \binom{\mu}{i} \left(\frac{\gamma\,(k-i)+1}{\mu\gamma t+N}\right)^{i} \left(1 - \frac{\gamma\,(k-i)+1}{\mu\gamma t+N}\right)^{\mu-i} p_{k-i,t}\,. \tag{8}$$

Interestingly, we can solve the recurrence relation of Eq. (8) for a closed form expression for $p_{k,t}$ when $\gamma = 0$ as

$$p_{k,t} = \binom{\mu t}{k} \left(\frac{1}{N}\right)^{k} \left(1 - \frac{1}{N}\right)^{\mu t-k}. \tag{9}$$

We use Eq. (8) to synthesize the bottom node degree distributions of α-BiN and then analyze those to understand the various characters of α-BiN. We perform exhaustive experiments on synthesis of α-BiN using Eq. (8) for several combinations of values of γ, μ, N and t. We report two distinct observations which are quite interesting.

1. For larger μ, the value of γ_c is normally greater than one. In this case at every time step, a top node comes with almost N number of edges. We observe that when the value of μ is close to N, for $\gamma = 1$, the distribution is still being a skewed (normal) distribution. Fig. 2(a) shows the change in

the pattern of distribution curve over various value of γ for a typical case of $\mu = N$. So in this case γ_c is greater than 1 which departs from the boundary prediction stated in Sec. 2.4.

Fig. 2(b) shows the movement of the values of modes over different γ with $N = 50$, $\mu = 50$ and $t = 100$. Note that when $\gamma \leq 1.05$, modes of the distributions are greater than zero and decrease over γ but after that, mode becomes zero suddenly at $\gamma = 1.05$ and stays here thereafter. As can be seen from Fig. 2(b), there is a sharp fall in the value of mode at the critical junction. This behavior clearly shows that higher value of μ brings some inherent randomness in the model. Hence higher preferentiality is needed to bring in monotonicity.

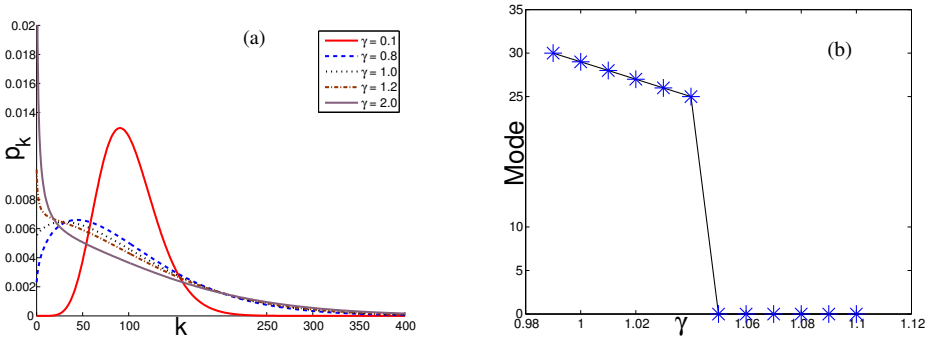

Fig. 2. (a) For the case of $\mu = N$: six distributions given by Eq. (8) for $N = 50$, $\mu = 50$ and $t = 100$ with different values of the γ where corresponding curve for $\gamma = 1$ does not show monotonically decreasing pattern. (b) Mode of the distribution for various γ ranging from 0.99 to 1 with 0.01 increment.

2. When the value of μ is quite large, then degree distribution shows two local maxima. In some real life system, each element enters the system with μ ($> N$) links. For example, in Codon–Gene bipartite network usually number of codons in a gene is of the order of hundred to thousand but the number of codons is only 64. In Fig. 3(a), we have depicted six different distributions to show the change in the nature of the degree distribution as γ is increased (N, μ and t are constant). It is seen that the distribution violates the conclusion of [21] *i.e.* at $\gamma = 1$ it will show monotonically decreasing nature. Another interesting observation is that at some values of γ higher than 1 (in Fig. 3(a) at $\gamma = 1.2$) it shows another local maxima. We have examined the nature of this second maxima and found that it persists even after considerable large value of time (Fig. 3(b)).

In the next section we apply our growth models to understand real world systems. We have taken one important α-BiN from biology, *i.e.* protein–protein complex network. We present a comparative study of real bottom node degree distribution of protein–(protein complex) network with the simulated one obtained from our models.

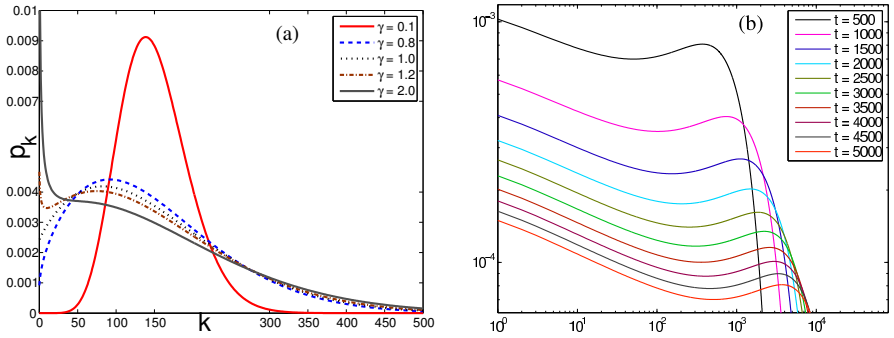

Fig. 3. (a) For the case of $\mu > N$: Six distributions given by Eq. (8) for $N = 10$, $\mu = 15$ and $t = 100$ with different values of γ where corresponding curve for $\gamma = 1$ does not show monotonically decreasing pattern. At higher γ curves show another mode other than $k = 0$. (b) For $\mu = 15$ and $\gamma = 1.2$ the second mode persists for large t. Here times are taken from 500 to 5000 with the interval of 500. Plots are drawn in log–log scale to fit in the same screen.

3. Protein–(protein complex) network (ProComp)

A protein complex (or multimeric protein) is a group of two or more proteins. Protein complexes are a form of quaternary structure[1]. Although basic set of proteins is fixed, the complexes arising from them are continuously evolved. The protein complex set of a particular organism shows an instance of ongoing evolution process. Hence we model the data set of an organism into α-BiN and try to infer the nature of the evolution process.

The first eukaryote[2] organism whose genome has been completely sequenced is budding yeast (*Saccharomyces cerevisiae*). It is also the first eukaryotic cell whose proteome (the set of all proteins) and interactome (the network of all mutual interactions between proteins) have been well established. Lots of research work on yeast genome has been done in the field of system biology and molecular biology. The genome of yeast is well understood from biological point of view. That is why we have chosen yeast as our model organism to build protein–protein complex network or ProComp.

[1] In biochemistry, quaternary structure is the arrangement of multiple folded protein molecules in a multi-subunit complex.

[2] A eukaryote is an organism whose cells contain complex structures enclosed within membranes including nucleus.

Researchers from diverse fields have been studying protein complex as protein–protein interaction network where nodes are proteins and edges represent occurrence of two proteins in the same protein complex [9, 10, 16, 19]. Contrary to that, in this section we analyze protein interactions from bipartite network point of view. Here we study the structure of the yeast protein–(protein complex) network in which edges between a protein complex and a protein represent the involvement of the protein in that complex.

3.1. Construction of ProComp

Protein–(protein complex) network or ProComp is a bipartite network where the nodes in U correspond to distinct proteins and those in V correspond to unique protein complexes. There is an edge $(u, v) \in E$ if a protein u is a part of a protein complex v.

In order to construct ProComp of yeast (*Saccharomyces cerevisiae*) we collected data from `http://yeast-complexes.embl.de/complexview.pl?rm=home` that contains information about the protein–protein interactions and protein complexes found in the yeast *Saccharomyces cerevisiae* [8]. There are 959 distinct proteins and 488 unique protein complexes in this database. We construct ProComp from this data where $\mid U \mid = 959$ and $\mid V \mid = 488$. The total number of edges running between the partitions U and V is 3653.

3.2. Experiments

From the protein complex data we can see that every protein complex is a compound of several proteins or every top node is connected with multiple edges. So, definitely ProComp is an instance of parallel attachment growth model. In simulation we use attachment kernel specified by Eq. (1).

For our simulations, we assume that t is the number of protein complexes and μ is the average number of proteins that a protein complex usually contains. We explore various values of γ and for each individual γ the results are averaged over 100 simulation runs. In experiments, the values of γ are varied in steps of 0.01. The best fitted value of γ will be associated with the closest distributions obtained from the simulation to the empirical data of yeast ProComp.

As a general technique to estimate good fits, we measured the mean square errors between the degree distributions of the real networks and those produced by the models. The closeness of the simulated distributions is determined by this error, E, and defined as follows.

$$E = \sum_{k=0}^{\infty} (p_k(\gamma) - p_k^*)^2 \,, \tag{10}$$

where p_k^* represents the empirical distribution.

3.3. Results

In this case, the initial settings for the simulation are as follows: $N = 959$, $t = 488$ and $\mu = 9$. We find that good fits emerge in the range $\gamma \in [0.55, 0.66]$ (in steps of 0.01) with the best being at $\gamma = 0.58$ (see Fig. 4). The plots in Fig. 4 clearly indicate that our growth model explains the degree distribution of ProComp quite accurately.

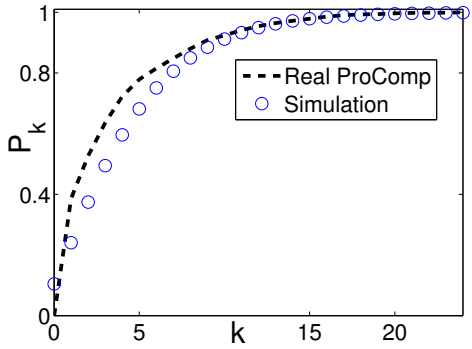

Fig. 4. Protein distribution in ProComp with $N = 959$, $t = 488$ and $\mu = 9$. Dashed black line shows the distribution related to the empirical data, blue circles are for the best fitted simulated degree distribution of our growth model.

4. Conclusion and future works

In this paper we have considered attachment probability as linearly dependent on the node's degree while analysing the growth of the α-BiN. An exhaustive study with various model parameters has been done. Interesting observations about the effect of various parameters on the final shape of α-BiN has been reported. We have extended the parallel attachment growth analysis presented in our earlier works [5, 21] and introduced a real world example of α-BiN.

Detail simulation results have been presented to validate our model. The analytical models mostly consist of recurrence expression of evolution of bottom node degree distribution. Developing the corresponding closed form solution of those recurrence expressions are rather very tough. Nevertheless, the closed form solution may give more accurate and clear explanations of all of our observations. The main difficulty in developing closed form equations arises from the fact that the nature of the solution is different from most of the previous works. Most of the previous works on network growth gave the closed form solution in asymptotic region. Here asymptotic degree distribution of these models does not converge because the number of bottom nodes is fixed. However, we have performed extensive simulations to understand the growth of α-BiN.

We have also identified many new problems. For example ,considerable untouched part is the superlinear attachment growth. Another interesting topic regarding α-BiN study is its one mode projection. One mode projection of a α-BiN is the network of any one set of nodes of the α-BiN where two nodes are connected if they are both neighbors of same node from other set of that α-BiN. One mode projection analysis is a common study in collaboration network research. A part of any future work will be directed towards understanding the one mode projection.

This work was partially supported by the DST, Goverment of India project grant no. SR/S3/EECE/059/2006. The authors would like to extend their gratitude to Fernando Peruani of SPEC/CEA, France, Monojit Choudhury of MSR, India and Animesh Mukherjee, Mozaffar Afaque and Maj. Vikas Yadav of IIT Kharagpur, India for their valuable comments and suggestions.

REFERENCES

[1] R. Albert, A.-L. Barabási, *Phys. Rev. Lett.* **85**, 5234 (2000).

[2] R. Albert, A.-L. Barabási, *Rev. Mod. Phys.* **74**, 47 (2002).

[3] A.-L. Barabási, R. Albert, *Science* **286**, 509 (1999).

[4] A.-L. Barabási, H. Jeong, R. Ravasz, Z. Néda, T. Vicsek, A. Schubert, *Physica A* **311**, 590 (2002).

[5] M. Choudhury *et al.*, `arXiv:0811.0499v1[physics.data-an]`, submitted to *Phys. Rev.* **E**.

[6] M. Choudhury, A. Mukherjee, A. Basu, N. Ganguly, Proceedings of COLING–ACL, **P06**, 128 (2006).

[7] S.N. Dorogovtsev, J.F.F. Mendes, *Adv. Phys.* **51**, 1079 (2002).

[8] A. Gavin *et al., Nature* **440**, 631 (2006).

[9] P. Uetz *et al., Nature* **403**, 623 (2000).

[10] T. Ito *et al., PNAS* **98**, 4569 (2001).

[11] S. Eubank, H. Guclu, V.S.A. Kumar, M.V. Marate, A. Srinivasan, Z. Toroczkai, N. Wang, *Nature* **429**, 180 (2004).

[12] J.-L. Guillaume, M. Latapy, *Information Processing Letters* **90**, 215 (2004).

[13] P. Holme, F. Liljeros, C.R. Edling, B.J. Kim, *Phys. Rev.* **E68**, 056107 (2003).

[14] R. Lambiotte, M. Ausloos, *Phys. Rev.* **E72**, 066107 (2005).

[15] F. Liljeros, C.R. Edling, L.A.N. Amaral, H.E. Stanley, Y. Aberg, *Nature (London)* **411**, 907 (2001).

[16] S. Maslov, K. Sneppen, *Science* **296(5569)**, 910 (2002).

[17] A. Mukherjee, M. Choudhury, A. Basu, N. Ganguly, *Int. J. Mod. Phys.* **C18**, 281 (2006).

[18] M.E.J. Newman, *SIAM Review* **45**, 167 (2003).

[19] R. Pastor-Satorras, E. Smith, R.V. Solé, *J. Theor. Biology* **222**, 199 (2003).

[20] M. Peltomäki, M. Alava, *J. Stat. Mech.* **1**, 01010 (2006).

[21] F. Peruani, M. Choudhury, A. Mukherjee, N. Ganguly, *Europhys. Lett.* **79**, 28001 (2007).

[22] J.J. Ramasco, S.N. Dorogovstev, R. Pastor-Satorras, *Phys. Rev.* **E70**, 036106 (2004).