PHYSICAL REVIEW E **77**, 065701(R) (2008)

Differential equations as a tool for community identification

Małgorzata J. Krawczyk*
Faculty of Physics and Applied Computer Science, AGH University of Science and Technology,
al. Mickiewicza 30, 30-059 Kraków, Poland

(Received 4 April 2008; published 18 June 2008)

We consider the task of identification of a cluster structure in random networks. The results of two methods are presented: (i) the Newman algorithm [M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004)]; and (ii) our method based on differential equations. A series of computer experiments is performed to check if in applying these methods we are able to determine the structure of the network. The trial networks consist initially of well-defined clusters and are disturbed by introducing noise into their connectivity matrices. Further, we show that an improvement of the previous version of our method is possible by an appropriate choice of the threshold parameter β. With this change, the results obtained by the two methods above are similar, and our method works better, for all the computer experiments we have done.

DOI: 10.1103/PhysRevE.77.065701 PACS number(s): 07.05.Tp, 02.10.Ox, 05.45.−a

I. INTRODUCTION

Relations between elements of many sets in the world can be described with networks. Definition of a network involves the indication of nodes and of relations between them. Nodes of the network are simply elements of the set and edges reflect relations between them. The simplest realization is based on zero-one information, i.e., it is only possible to say whether a given two nodes are mutually connected or not. However, if it is possible to get information on how strongly any two elements are connected to each other, the knowledge about the structure of the so-called weighted network is more complete. Further, networks can be used for a system description at different levels. All of us can be seen as elements of the network of our relatives. We are also connected with people in our workplace, and the companies usually are connected to each other [1,2]. Looking into our bodies, it is clear that their proper functioning is possible thanks to the existence of metabolic, protein, and gene networks [3–6]. Thanks to these complex networks each structure of our bodies fulfils its function. The method of establishing weights of edges between nodes (e.g., companies, people, genes) of any network does depend on the network origin. In the case of social networks the weights can reflect the intensity of contacts between given elements. In the case of biological networks, such as gene or protein networks, weights of edges are calculated on the basis of the measurements of the level of their expression in the cells of organisms. Independently of the origin of the network, weights of edges can be expressed via real values from the range [0,1]. Weight value equal to zero indicates that two nodes are completely disconnected, whereas value equal to one indicates the case of the full connection of two nodes.

Regardless of the type of the network, usually it is possible to distinguish clusters (communities) in the whole network. The question of the definition of the cluster is problematic, but usually it is understood as a community of densely connected nodes, which are only sparsely connected with nodes in other clusters [7–10]. Finding clusters in the network is a very common problem in many different areas. There exist several different methods which allow for the network division into clusters in accordance with some rules. The algorithms can be classified in a different manner depending on the criterion used. Some algorithms identify nodes with the clusters, so at the beginning the number of clusters is equal to the number of nodes of the network. In this case the task is to join some clusters to each other, in a certain manner. Such algorithms are called agglomerative. Some other algorithms run in the opposite direction, i.e., the rules lead to successively erasing the links of the network. This kind of algorithm is named divisive. These two algorithms are known as hierarchical [11–13]. Another classification is based on the analysis of local or global properties of the network. The difference lies in the number of links affected by allocation of nodes into clusters [14,15]. In the local case the change of state of an edge connecting nodes i and j depends on weights of edges connecting those two nodes with a few nodes which are its neighbors. In the global case, the state of edge ij depends on all remaining edges in the network.

The quantity which is usually used for indication of the proper division of the network is the modularity Q, introduced by Girvan and Newman [11]. This quantity is large if there are many edges between nodes within the communities and only a few between nodes from different communities. There are some algorithms which are based on modularity calculations but introduce some modifications to the Newman algorithm, which allow for more efficient calculations. Such modification is an important feature in the case of analysis of large networks [16–18]. Further criteria of the algorithm's quality are the maximal modularity for a given network (e.g., Zachary karate club [8]) or the number of nodes classified correctly to their communities [19]. Some authors apply another approach for identification of communities in the network, based on some local quantities, taking into account internal and external degrees of the nodes [17,19–21], which is motivated by the resolution limit of the modularity [7].

Another group of algorithms is based on the analysis of the spectral properties of the network, through the eigenvectors of the Laplacian matrix [8,13] or on some physical principles, e.g., Potts model [9] or Kirchhoff equations with

*gos@fatcat.ftj.agh.edu.pl

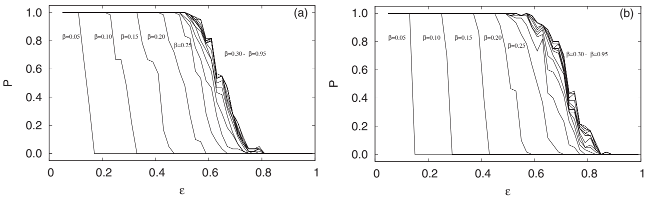

FIG. 1. Results for $N=85$ (a) and $N=110$ (b) nodes, and different values of β.

edges of a network as a resistor [18]. An interesting feature of the Potts model is its ability to detect the overlapping of the communities. The algorithm proposed in Ref. [22], based on the fitness calculation, enables such kinds of analyses, simultaneously rendering hierarchical community structure.

The approach proposed here is different from the algorithms mentioned above because we apply a dynamics continuous in time: the weights of all edges evolve simultaneously. In this way, there is no information loss which is due to the random selection of the order of updating nodes. On the other hand, the program is supposed to work rather slowly in the case of large networks. The advantage is that for most investigated networks, the proper structure is reproduced with larger probability than for other algorithms.

Irrespective of the algorithm used, the most important question is whether the obtained division is correct. In fact, a response to this question is possible if obtained communities can be interpreted thanks to knowledge about analyzed networks. Even if similar results are obtained from different algorithms, it should not be understood as a proof of the correctness of the obtained division. To evaluate the performance of a given method, it should be applied to networks with the structure known *a priori*. Such networks are to be designed for the purposes of computer experiments. If the original, well-defined structure of a network is somehow disturbed, the question is if this initial structure can be reconstructed with the given algorithm.

In this paper we compare the results obtained within two clustering algorithms: the Newman algorithm [11] and our method based on differential equations [23], with respect to reconstructing the original structure of the designed networks. Both algorithms are described in Sec. II. Section III includes the results obtained by these algorithms for different sizes of networks.

II. ALGORITHMS

The choice of the algorithms used was motivated by the results shown in Ref. [23]. The first algorithm used is the well-known Newman algorithm based on the analysis of the modularity Q [8,24]. This approach was adapted for the case of the weighted network, and the formula for Q is as follows:

$$Q=\frac{1}{2m}\sum_{ij}\left[w_{ij}-\frac{k_ik_j}{2m}\right]\delta(c_i,c_j), \qquad (1)$$

where w_{ij}—weight of the link between node i and j ($w_{ij} \in [0,1]$), k_i—weighted node degrees, $m=1/2\Sigma_{ij}w_{ij}$ and

$$\delta(c_i,c_j)=\begin{cases}1 & \text{when } i \text{ and } j \text{ are in the same cluster}\\ 0 & \text{otherwise.}\end{cases}$$

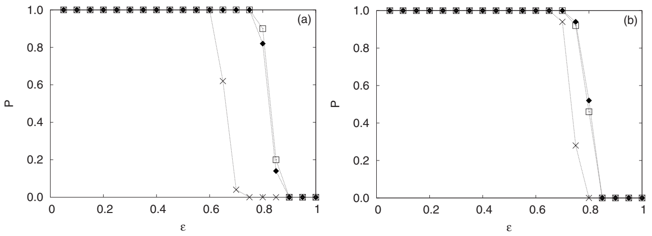

FIG. 2. Results for clusters of equal sizes: (a) $N=68$, clusters of 34 nodes, (b) $N=72$, clusters of 24 nodes [symbols denote: □, the Newman algorithm; ×, Eq. (2) with $\beta=0.25$; ◆, Eq. (2) with $\beta=0.4$].

Apart from the knowledge of the weight of edges of the networks, the calculation of the modularity Q also involves information about the current structure of the network, which depends on the algorithm used. The structure of the network is defined as the clusters identified so far by the clusterization algorithm.

The Newman algorithm is an example of the agglomerative algorithms. Here, the method is to join those clusters in which the obtained value of Q is maximal at each iteration. This algorithm yields to the formation of one large cluster, including the whole set of nodes. The division is established at the partition where the value of the dependency of Q to the iteration step reaches the highest value.

In this algorithm the considered state of the system depends on the nodes belonging to the same cluster, and connections to the rest of the network are neglected.

In the case of the differential equations method (DEM), proposed by the author in Ref. [23], at each iteration the values of the connectivity matrix elements evolve due to the interactions between nodes in the whole network. The rate of change of the A_{ij} is given by

$$\frac{dA_{ij}}{dt}=G(A_{ij})\sum_{k\neq i,j}(A_{ik}A_{kj}-\beta), \qquad (2)$$

where A_{ij} is an element of the connectivity matrix, $G(x)=\Theta(x)\Theta(1-x)$, β parameter.

The structure of the network at the given iteration step forms the basis for the calculation of the parameter Q [Eq. (1)]. Because we are interested in the finding of the subnetworks in the initial network, and not in the network obtained because of the change of the connectivity matrix, the value of Q is calculated according to the values of the initial connectivity matrix.

The used algorithm, applied to weighted connectivity matrix A, can be summarized as follows: (i) Calculation of the new value of A_{ij} for all $A_{ij} \in (0,1)$, according to the Eq. (2), which means actually that if the weight of the edge reaches value 0 or 1 it is not changed anymore; (ii) Identification of the actual structure of the network, which varies because during evolution some values of A_{ij} decrease to 0; (iii) If the number of clusters differs from the one in the previous simulation step, calculation of the modularity Q with the network structure given by (2) with original weights of edges; (iv) Return to point (i).

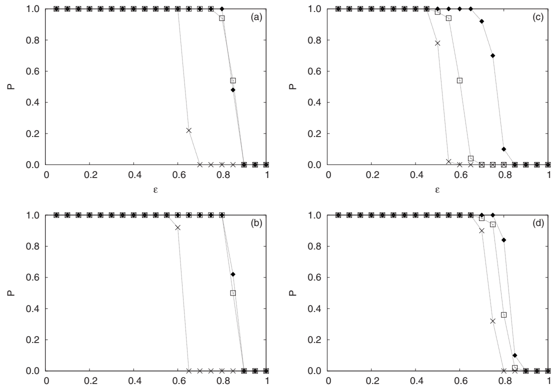

FIG. 3. Results for clusters of different sizes: (a) $N=96$, clusters of 44 and 52 nodes, (b) $N=130$, clusters of 77 and 53 nodes, (c) $N=84$, clusters of 65 and 19 nodes, (d) $N=110$, clusters of 38, 30, and 42 nodes [symbols denote: □, the Newman algorithm; ×, Eq. (2) with $\beta=0.25$; ♦, Eq. (2) with $\beta=0.4$].

At the end of this procedure the dependence of the Q value from the number of clusters (and nodes belonging to each of them) is obtained. The division where the value of Q is maximal is accepted as the optimal division of the network into clusters.

The time of the Girvan-Newman algorithm is $O[(M+N)N]$ [25] (where N is the number of nodes, and M is the number of edges). In the case of DEM the time of each simulation step is $O(N^3)$.

III. RESULTS

We present the results obtained from a series of computer experiments which were performed as follows. At the beginning the network consists of a given number of fully connected clusters of the same size, without connections between them. The connectivity matrix is symmetric, with element A_{ij} equal to 1 if there is a link between nodes i and j, elsewhere A_{ij} is set on 0. The aim of the calculation is to check if the applied clustering algorithm can reproduce the initial structure of the network if all connectivity matrix elements are disturbed by random numbers. In other words, for each connectivity matrix element a random number ε_{ij} is chosen from the range $[0,\epsilon]$. Once an element $A_{ij}=1$, its disturbed value is set to $1-\varepsilon_{ij}$, otherwise $A_{ij}=0$ is changed to ε_{ij}. The disturbed matrix remains symmetric.

In our previous work the value of β in Eq. (2) was set to 0.25 (see [23]). Now we check how the results depend on the value of this parameter. Below we compare the results obtained from the Newman algorithm and from our method with two values of parameter β, for the same networks. The results shown below are checked to be robust for the applied values of the numerical timesteps in Eq. (2).

In the real world the case of equal size of clusters is not generic. So, the results for networks with initial structure composed on randomly chosen number and sizes of clusters are also presented.

All presented results (except the results in Sec. III A, see below) are averaged over 50 realizations of the network. For all figures the notation is N, number of nodes; ε, amplitude of the noise; P, probability of reproduction of the initial structure of the network. The symbols in figures denote: □, the Newman algorithm; ×, Eq. (2) with $\beta=0.25$; ♦, Eq. (2) with $\beta=0.4$.

A. Estimation of the best value of β

In Fig. 1 the results on P as a function of ε are shown (averaged over three different partitions of the network with given size, for each partition results are averaged over 20 realizations). Exact position of the particular curves depend on particular partition and size of the network, but as a rule it can be realized that better results, i.e., higher percentage of the reproduction of initial structure for higher values of noise, are obtained for values of β larger than approximately 0.3. Because of that for further computer experiments the value of β was set to 0.4.

B. Comparison of the results for two different values of β for clusters of equal sizes

As it was shown in [23], for some cases the Newman algorithm based on modularity calculation works better than the algorithm based on differential equations. But, as can be seen from Figs. 2(a) and 2(b) if a higher value of β is used (in this case $\beta=0.4$), the results obtained within two methods are equivalent. For other cases presented in [23], where Eq. (2) worked better, an increase of β does not change the results qualitatively.

C. Clusters with different sizes

In the case when the number of clusters is as small as two, and if the difference between the clusters sizes is not very large, the Newman algorithm works well. Comparable results are obtained from Eq. (2) with $\beta=0.4$. If a smaller value of β is used, the probability of the reproduction of the initial network structure is remarkably smaller [Figs. 3(a) and 3(b)]. The situation changes if sizes of the clusters are significantly different [Fig. 3(c)]. In this case, DEM with $\beta=0.4$ works much better than the Newman algorithm. Also in this case, the enhancement of the parameter β results in the method working well for a higher value of noise.

If the number of clusters increases to three and the network is divided into clusters with approximately equal sizes, the Newman algorithm and Eq. (2) for both tested values of β return results which are similar [Fig. 3(d)]. In the case when the sizes of the clusters are significantly different, DEM with $\beta=0.4$ works much better than the Newman algorithm [Fig. 4(a)]. The information we get from Figs. 4(b)–4(d) is that if the number of clusters is higher than three, the differential equation method works better than the Newman algorithm.

To check if this is actually the rule of used algorithms it is convenient to introduce some quantity which can describe the size distribution—the fragmentation coefficient F:

$$F=\sum_i \left(\frac{n_i}{N}\right)^2, \tag{3}$$

where n_i is the size of the cluster and N is the number of the nodes.

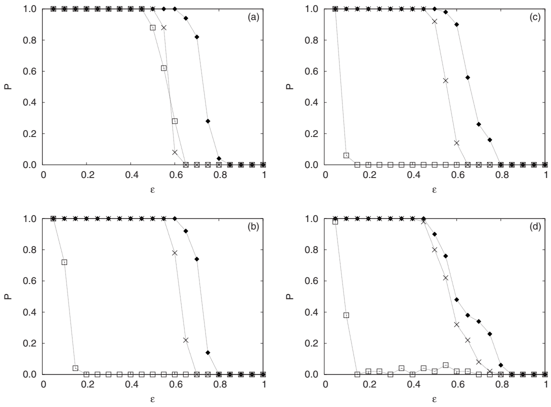

FIG. 4. Results for clusters of different sizes: (a) $N=110$, clusters of 50, 48, and 12 nodes, (b) $N=130$, clusters of 22, 34, 11, 10, and 53 nodes, (c) $N=130$, clusters of 19, 60, 45, and 6 nodes, (d) $N=110$, clusters of 18, 22, 30, 38, and 2 nodes [symbols denote: □, the Newman algorithm; ×, Eq. (2) with $\beta=0.25$; ♦, Eq. (2) with $\beta=0.4$].

Another characteristic quantity which can be useful here is the value of the amplitude of noise ε for which the probability P of the reproduction of the initial cluster structure reaches value 0.5. Below this level of noise is denoted as $\varepsilon(0.5)$. If the network is divided into two clusters of very different sizes, in accordance with results presented above, for some cases the value of $\varepsilon(0.5)$ decreases to zero independently of the algorithm used [Fig. 5(a)]. Similar effects are discussed in [7]. For cases where differences in the sizes of clusters are not very large (smaller values of F) both algorithms seem to work with similar accuracy. This conclusion changes if the number of clusters is higher. As can be seen in Fig. 5(b), the value of $\varepsilon(0.5)$ varies in the range of approximately 0.6–0.9 for all analyzed networks if the differential equations algorithm is applied, whereas for the Newman algorithm this dispersion is much higher.

IV. DISCUSSION

It was shown in our previous work [23] that in the case $\beta=0.25$ and for a small number of clusters, the Newman's algorithm works better than Eq. (2). Here we demonstrate that once the value of β is changed from 0.25 to 0.4 or more, our algorithm is never worse. At present we have no arguments to distinguish between any value of β larger than, say, 0.4. This point remains to be clarified in the near future. Our tests made for networks with different sizes of clusters show that the Newman algorithm enables the reproduction of the real structure of the noisy network if the number of clusters is small (two or three) and the difference in cluster size is not very large. We show that in the cases where there are a few clusters with different sizes, the differential equation method works well, even if the amplitude of the noise is as high as 0.8. We decided to use as a criterion of network division analysis of the value of modularity; because it is used in many models. It would be interesting to make a similar analysis using the fitness parameter [22] instead of modularity.

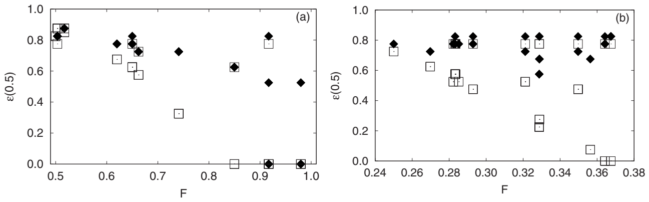

FIG. 5. $\varepsilon(0.5)$ against the fragmentation coefficient: (a) 2 clusters, (b) 4 clusters [symbols denote □, the Newman algorithm; ♦, Eq. (2) with $\beta=0.4$].

In conclusion, for all analyzed networks, the method based on differential equations with an appropriate choice of the value of parameter β works better than, or at least as good as, the Newman algorithm.

ACKNOWLEDGMENT

This work was partially supported by the Polish Ministry of Science and Higher Education Grant No. PBZ-MNiI-2/1/2005.

[1] S. P. Borgatti, Comput. Math. Org. Theory **12**, 21 (2006).
[2] S. Wasserman and K. Faust, *Social Networks Analysis: Methods and Applications* (Cambridge University Press, Cambridge, U.K., 1994).
[3] L. Skrabanek *et al.*, Mol. Biotechnol. **38**, 1 (2008).
[4] V. Cilizza *et al.*, Physica A **352**, 1 (2005).
[5] A. L. Barabasi and Z. N. Oltvai, Nat. Rev. Genet. **5**, 101 (2004).
[6] I. P. Androulakis *et al.*, Annu. Rev. Biomed. Eng. **9**, 3.1 (2007).
[7] S. Fortunato and M. Barthelemy, Proc. Natl. Acad. Sci. U.S.A. **104**, 3641 (2007).
[8] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **103**, 8577 (2006).
[9] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004).
[10] J. M. Kumpula *et al.*, Eur. Phys. J. B **56**, 41 (2007).
[11] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[12] M. E. J. Newman, Eur. Phys. J. B **38**, 321 (2004).
[13] L. Donetti and M. A. Munoz, J. Stat. Mech.: Theory Exp. (2004), 10012.
[14] S. Boccaletti *et al.*, Phys. Rep. **424**, 175 (2006).
[15] S. Fortunato and C. Castellano, *Encyclopedia of Complexity and System Science* (Springer, Berlin, 2008).
[16] A. Clauset *et al.*, Phys. Rev. E **70**, 066111 (2004).
[17] A. Clauset, Phys. Rev. E **72**, 026132 (2005).
[18] F. Wu and B. A. Huberman, Eur. Phys. J. B **38**, 331 (2004).
[19] J. Duch and A. Arenas, Phys. Rev. E **72**, 027104 (2005).
[20] F. Radicchi *et al.*, Proc. Natl. Acad. Sci. U.S.A. **101**, 2658 (2004).
[21] J. P. Bagrow and E. M. Bollt, Phys. Rev. E **72**, 046108 (2005).
[22] A. Lancichinetti *et al.*, e-print arXiv:0802.1218.
[23] M. J. Krawczyk and K. Kulakowski, e-print arXiv:0709.0923.
[24] M. E. J. Newman, Phys. Rev. E **70**, 056131 (2004).
[25] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).