The Small Community Phenomenon in Networks: Models, Algorithms and Applications

Pan Peng[*]

State Key Laboratory of Computer Science, Institute of Software,
Chinese Academy of Sciences
and
School of Information Science and Engineering,
Graduate University of China Academy of Sciences, Beijing, China
pengpan@ios.ac.cn

Abstract. We survey a recent new line of research on *the small community phenomenon* in networks, which characterizes the intuition and observation that in a broad class of networks, a significant fraction of nodes belong to some small communities. We propose the formal definition of this phenomenon as well as the definition of communities, based on which we are able to both study the community structure of network models, i.e., whether a model exhibits the small community phenomenon or not, and design new models that embrace this phenomenon in a natural way while preserving some other typical network properties such as the small diameter and the power law degree distribution. We also introduce the corresponding community detection algorithms, which not only are used to identify true communities and confirm the existence of the small community phenomenon in real networks but also have found other applications, e.g., the classification of networks and core extraction of networks.

1 Introduction

In recent years, the flourish of real network data that span from various domains including the World Wide Web, the power grid, the friendship network and many others are offering the scientific community new problems and challenging new methodology [29]. One canonical way to study networks is to first empirically explore the network data, extract patterns and properties from the underlying structure, and then design network models that reproduce the observed properties. For example, various networks are observed to share some common phenomena with the best-known two the scale-free property and the small-world phenomenon, which are further simulated by simple random graph models [4,38]. These models not only give us insight how global properties come from local generative rules but also provide us the testbeds to study other problems on networks, e.g. the decentralized search [16] and information diffusion [7].

[*] The author is partially supported by the Grand Project "Network Algorithms and Digital Information" of the Institute of Software, Chinese Academy of Sciences.

M. Agrawal, S.B. Cooper, and A. Li (Eds.): TAMC 2012, LNCS 7287, pp. 40–49, 2012.

We follow this way to study the *community structure* in networks. A community is a group of nodes that share common properties and/or behave similar with each other, which is reflected as a subset of nodes with dense intra-connections and (relatively) sparse inter-connections in the network structure. Communities can be seen as building blocks of networks and play important roles in the spread of information, marketing and searching [9]. There has been an extensive study of the community structure of networks along with plenty of community detection algorithms in the last decade (see e.g., [35,32,12]). The direct impression of the communities of networks is that they are overlapping, hierarchical and that the communities of different scales coexist. In this presentation, we formalize and investigate a new property of the community structure called the small community phenomenon, that a significant fraction of nodes belong to some small communities. This phenomenon is intuitive and has some partial supporting evidence (see Section 5), however, to our knowledge, it has not been seriously proposed and studied before.

We first introduce the definitions of communities, the small community phenomenon and the corresponding community identification algorithms in Section 2. Then we show that some classical network models do exhibit this new phenomenon while some others do not, and we also propose new models that embrace the small community phenomenon as well as the small diameter and the power law degree distribution properties in Sections 3 and 4, respectively. Finally, we provide further evidence of the existence of the small community phenomenon on a set of social networks[1] and give some other applications in Sections 5 and 6.

2 Basic Definitions and Algorithms

2.1 Definition of Community

How to formally characterize the property that a set is community-like, or equivalently, what is the quantitative definition of a community? Though the intuition behind the community seems clear and simple, there is not universal agreement on the formal definition of community (see the surveys [35,32,12]). To give such a definition, we start with a well known concept called *conductance* in the literature of computer science, which captures well the intuition behind a community.

Given an undirected graph $G = (V, E)$ and a vertex subset $S \subseteq V$, let d_v denote the degree of vertex v, and let the volume $\mathrm{vol}(S)$ of a set S be the total degree of vertices in S, that is, $\mathrm{vol}(S) = \sum_{v \in S} d_v$. Let $e(S, \bar{S})$ and $e(S)$ denote the number of edges crossing the boundary of S and the number of edges that lie entirely in S, respectively. For any set $S \subseteq V$, its conductance $\phi(S)$ is defined as $\phi(S) = \frac{e(S,\bar{S})}{\min\{\mathrm{vol}(S), \mathrm{vol}(V-S)\}}$. Thus, roughly speaking, for a subset S of volume smaller than $\mathrm{vol}(V)/2$, the smaller its conductance is, the more likely that it is a community.

[1] All the data mentioned in this paper can be found from the websites: `http://snap.standford.edu`, or `http://www-personal.umich.edu/~mejn/netdata`, and we only consider the corresponding undirected graphs.

Several works directly used the conductance (or related) as a measure of how good a community is [14,2,18,19]. However, we note that the small conductance of a set S may be caused by the fact S just has a large number of nodes inside, and thus fails to reflect the trait of being a community (see for the example in [22]). Here, we introduce a conductance-based definition that characterizes the community in a more refined way [23].

Definition 1. *([23]) Given a graph $G = (V, E)$ and $\alpha, \beta > 0$, a connected set S is an (α, β)-community if $\Phi(S) \leq \frac{\alpha}{|S|^\beta}$. Moreover, if $|S| = O((\ln n)^\gamma)$, where $n = |V|$, then S is called an (α, β, γ)-community.*

Note that under the above definition, 1) only the given set S and its boundary is involved, namely, the definition is local in that it does not require information on other parts of the network; 2) any set of constant size is a trivial (α, β)-community for sufficiently small β. In the following, we are mainly interested in the communities of larger size (i.e., $\omega(1)$), and these communities are called *proper*, in which case β ranges from 0 to 2; 3) if the conductance inequality is changed by $\Phi(S) \leq \frac{\alpha}{(\ln |S|)^\beta}$, then we call the set S a weak (α, β)-community.

2.2 The Small Community Phenomenon

What are the properties of real communities of a given network? Typically, the communities may overlap or nest in other clusters, which in turn lead to the hierarchical organization of the vertices of the network [8,34,5]. Several papers have found the skew distribution of community sizes in many different networks [33,6,28,31]. Leskovec et al. [18,19] find that in many large scale networks, the set of greatest community score (i.e., smallest conductance) is of size about 100 and beyond this size, the community score gradually decreases as the size of the set becomes larger.

We propose a new phenomenon that originates from the daily experience and observation that almost every one in our society belongs to some small communities. (In the following, the term *with high probability* and *almost every* will refer to the probability at least $1 - o_n(1)$ and at least $1 - o_n(1)$ fraction, respectively, where n denotes the size of the graph.)

Definition 2. *([23]) Given a graph G from some network model, if almost every vertex v belongs to some proper (α, β, γ)-community, where $\alpha, \beta, \gamma > 0$ are some universal constants, then G is said to have the small community phenomenon.*

On a real network, we will relax the condition of the small community phenomenon, by requiring that *a significant fraction,* 60% *say, of nodes belong to some small communities*, since the true communities may mix very much with each other so that it is nearly impossible for a structure-based detecting algorithm to extract them. We will corroborate this with a set of social network data the small community phenomenon in Section 5.

2.3 Community Detection Algorithm

How to extract good communities from a given network? The loss of exact definition of community leads to the vastness of community identification algorithms. Concerning on the conductance based clustering, there has been a line of research on local graph partitioning algorithms which may be used as subroutines for clustering [36,1,3]. These algorithms take a graph G and a vertex v as input, only explore parts of the input graph G and with constant probability, output a set of small conductance if v indeed belongs to some sets of small conductance. Such an algorithm is both fast and practical, and has already been used to find communities in real networks (e.g., [18,25,37,19,13]). In particular, Leskovec et al. [18,19] have used the PageRank-based local algorithm to analyze the statistics of the community structures over 100 large real-world networks while they did not test the algorithm on benchmark graphs, which are supposed to have a recognized community structure.

We developed a variant of the local graph partitioning algorithm **Community** which has different stoping conditions from the previous ones, especially the one used in [18,19] (see the details of **Community** in [22]). We further compared the effectiveness of the algorithm (denoted **O_Alg**) used by Leskovec et al. and **Community** (denoted **N_Alg**) on extracting the true communities on several benchmarks. One example on an American college football network is given in Table 1. In this network there are 12 true communities, e.g., Western Athletic, which are expected to be detected by the two algorithms. The numerical value (e.g., 0.663325) in the table denotes the maximum *cosine similarity* of the true community (e.g., Big Ten) and the communities found by the two algorithms. The higher similarity is (which is at most 1), the more accurate that the algorithm identifies the true community, and thus it is easy to see that our algorithm works much better on detecting true communities.

Table 1. The comparison of **Community** (**N_Alg**) with a previous one (**O_Alg**) on an American college football network. The numerical value denotes the maximum cosine similarity of the corresponding true community and the communities extracted by the corresponding algorithm.

conference	**O_Alg**	**N_Alg**	**conference**	**O_Alg**	**N_Alg**
Western Athletic	0.471405	0.843274	Independents	0.291111	0.23094
Sun Belt	0.370479	0.412393	Conference USA	0.580948	0.948683
Big East	0.478091	1	Mountain West	0.417029	1
Atlantic Coast	0.480384	1	Mid-American	0.72111	1
Big Twelve	0.561951	1	Southeastern	0.707107	1
Big Ten	0.663325	1	Pacific Ten	0.471405	1

3 Results on Classical Network Models

Are the classical network models exhibit the small community phenomenon? Random network models such as the Erdös-Rényi model (namely, the $G(n, p)$ model) and the preferential attachment (PA) model are not supposed to have communities [28], which is also true under our definition of the community.

Let us take PA model with parameter d for example. This model is a generative model, in which we start with a given graph G_0. Then for each $t \geq 1$, conditioned on G_{t-1}, we form G_t by adding a new vertex x_t together with d edges between x_t and y_i $(1 \leq i \leq d)$, each of which is chosen with probability proportional to the degree of y_i in G_{t-1}. This model has the nice power law degree distribution property that has been observed in many real networks. Mihail et al. [26] have proved that the conductance of a graph from PA (the definition there is slightly different) is larger than some constant, with high probability, which immediately implies that the graph generated from this model has no proper communities.

Theorem 1. *([26,23]) With high probability, there is no proper (α, β)-community in G_n for any $0 < \beta \leq 2$ and $d \geq 2$, where G_n is a random graph in the PA model with parameter d.*

There are also a set of models that have clear community structures, e.g., the geometric preferential attachment (GPA) model [10,11], the hierarchical model [8,34] and Kleinberg's small world (SW) model [16] when proper parameters are chosen.

Let us take the (1-dimensional) SW model with parameter r for example. In this model, we start with a given n-vertex cycle, in which a natural lattice distance can be defined: for any pair of vertices (u, v) , the distance $d(u, v)$ between them is the minimum path length connecting u, v. Then for any vertex v, we connects v to a long-contact u, which is chosen randomly with probability proportional to $(d(u, v))^{-r}$. Kleinberg have proved an interesting threshold result on the delivery time of a decentralized algorithm and thus given a characterization of the conditions under which people can construct short paths when they only have access to partial (local) information. We show that the community structure of this model also exhibit an interesting threshold phenomenon.

Theorem 2. *([23]) In the 1-dimensional small world model G, with high probability,*

1. *if $r < 1$, there is no proper community for an arbitrary node;*
2. *if $r = 1$, there exists proper weak (α_1, β_1)-communities of size $\frac{n}{(\ln n)^{c_1}}$ for every node, where $\beta_1 < 1, c_1 > 0$ and there also exists proper weak $(\alpha_2, 1)$-communities of size $c_2 n$ for every node, where $0 < c_2 \leq \frac{1}{4}$;*
3. *if $r > 1$, every node is contained in some proper (α, β, γ)-communities for some constants α, β, γ.*

4 Two New Models

How to model networks that simultaneously has the power law degree distribution, the small diameter as well as the small community phenomenon? This question is motivated by the fact that there indeed exist real networks that exhibit all the three properties (eg., the network grqc in Figure 1 in Section 5). Here, we briefly introduce two dynamic models that satisfies these good properties. More explanations can be found in [24,20].

The first model is a *geometric model*, which combines the preferential attachment scheme and an underlying structure in a natural way. It is defined on a unit sphere S and at each time, a new node is generated uniformly from S and it will connect to some existing nodes within a neighborhood with probability proportional to their degrees. We also require that each new node is born with some *flexible self-loops* which may be eliminated in later steps and are used to make long-distance connections. We note that this model is based on the GPA model which simulates networks that both have the power law degree distribution and small edge expansion [10,11]. We have proved that the coexistence of the small community phenomenon and the power law degree distribution in our geometric model is subtle in that the possible choices of a parameter lies in a very narrow region, beyond which one of the two properties are unlikely to appear [24].

Another model is called the *homophily model*, which combines the preferential attachment scheme and the homophily law in a natural way [20]. In this model, each new node v is born with a color that may be chosen uniformly at random from all the existing colors or totally new, in which case v is called the seed of the color, with some probability. Then node v will connects to some existing nodes that share the same color or all the existing nodes depending on v color, and these neighbors are chosen following the preferential attachment scheme. Long connections may be made between seeds. We have shown that any set of nodes that have the same color is a good small community by choosing appropriate parameters, which indicates that the model naturally characterizes the property that nodes in a community share something in common (the color) and that each community has a representative (its seed). Besides, the whole network as well as the induced subgraphs of small communities is shown to have the power law degree distribution.

Both of these two models have all the three nice properties mentioned above.

Theorem 3. *([24,20]) Under proper parameters, with high probability, the random graphs G_n from geometric model and H_n from homophily model both satisfy that 1) the power law degree distribution; 2) the average node to node distance is $O(\log n)$; 3) almost every node belongs to some proper (α, β, γ)-communities for some global constants α, β, γ.*

5 Empirical Results

Do the real network models exhibit the small community phenomenon? Many different clustering techniques have provide evidence that small communities are abundant, which partially support the thesis of this phenomenon ([33,6,28,31,18,19]). We show that our algorithm **Community** can be used not only to verify that several social networks exhibit the small community phenomenon, but also to give a more elaborate characterization called *local dimension* of the community structure of the networks.

Roughly speaking, given a network G, we will find a triple (α, β, γ) which characterizes best the community structure of G and is called the *local dimension* of G [23,24,22]. A network with local dimension (α, β, γ) has the property that the fraction of nodes that belong to some (α, β, γ) is maximized in some way (we refer to our paper [22] for details). Figures 1 and 2 show the size-fraction curves of several social networks under their local dimensions. A coordinate (x, y) on the size-fraction curve means that at least y fraction of nodes belong to a community of size at most x. Thus, we can see that at least 70% fraction of nodes belong to some communities of size at most 30 in network grqc, which indicates that the network has an obvious small community phenomenon; while in the network wikivote almost no nodes belong to communities of size smaller than 300, which indicates that the network may not have the phenomenon. There are also some networks that lie between these two cases, e.g., the network astro.

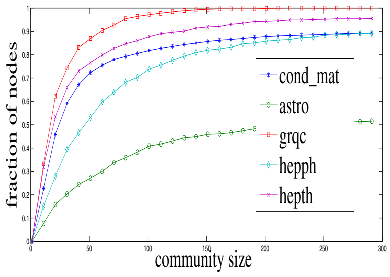

Fig. 1. The size-fraction curve on four collaboration networks

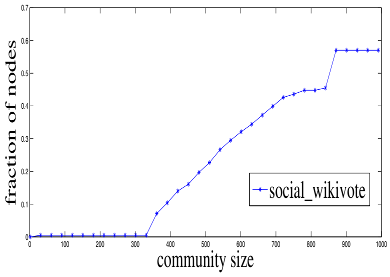

Fig. 2. The size-fraction curve on a Wikivote network

6 Applications

What are the applications of the small community phenomenon and the related clustering algorithm ***Community****?* Besides the potential applications mentioned in the introduction (and many others in [35,32,12]), we give two more examples.

6.1 Classification of Networks

Quantitatively classifying networks which may vary very differently in disciplines and scales will offer us great insights both on the network structures and dynamics. We are able to classify the networks based on their local dimensions and the percentage of nodes belonging to small communities. For example, the network grqc and wikivote in Figures 1 and 2 could be categorized into two classes: networks exhibit the small community phenomenon and those do not. A more refined classification over more social networks is given in [22]. Such a method of course applies to many other networks. We note that recently Lancichinetti et al. [17] and Onnela et al. [30] have also constructed taxonomies of networks based on different clustering algorithms and statistical properties of the resulting communities.

6.2 Core Extraction of Networks

Networks always exhibit the core-periphery structure, in which the core is both densely connected and central in terms of graph distance and may also have an embedded core-periphery structure; and so on [18]. The algorithm **Community** can be used to extract the core of a network [21]. More precisely, we start from the original network (graph) $G = G_0$, and recursively perform the following *reductions*: for $i \geq 0$, run **Community** to find all the communities of G_i corresponding to its local dimension (α, β, γ) and if no community is found, then stop; otherwise, let G_{i+1} be the largest connected component of G_i after deleting all the edges in the communities. The final subgraph G_l is declared to be the core of G.

To test that G_l indeed acts as the core of the original graph G and even that G_{i+1} acts more importantly than G_i in G, we investigate the power of spreading influence of each G_i under a simple threshold diffusion model [27,15], in which we are given a diffusion parameter ϕ, a size parameter s and a graph G whose nodes are all initially inactive. We first choose an initial active set S of size s uniformly at random from the vertices of G and then trigger a diffusion process: an active node v will remain active forever; and an inactive node v will become active if and only if at least ϕd_v of its neighbors are active. The process stops when all nodes are active or the number of active nodes does not increase. We are interested in the expected number of active nodes at the end of the diffusion.

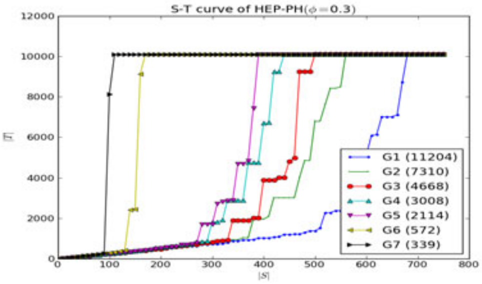

Fig. 3. The curve of diffusion size vs. initial active set size on a collaboration network when $\phi = 0.3$

Our experiments on a set of scientific collaboration networks show that for any i such that $0 \leq i \leq l-1$, by selecting a random set of size s from G_{i+1} as the initial active set always activates more nodes at the end of *the diffusion process in* G than the case by selecting a random set of size s from G_i [21]. In particular, the nodes of the core G_l, which is usually rather small compared to the graph G we start with, are much more influential in the diffusion process than average nodes of G, which indicates that G_l indeed plays a central role and acts as a core in G at least in the sense of diffusion as above. A more refined illustration is given in Figure 3, in which we can see that if the diffusion parameter ϕ is fixed (here, 0.3), the size of the initially active set S selected from G_i required for the diffusion process to reach the limit number (about $10,000$) decreases as i increases.

References

1. Andersen, R., Chung, F., Lang, K.: Local graph partitioning using pagerank vectors. In: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science, pp. 475–486. IEEE Computer Society, Washington, DC, USA (2006)
2. Andersen, R., Lang, K.J.: Communities from seed sets. In: Proceedings of the 15th International Conference on World Wide Web, WWW 2006, pp. 223–232. ACM, New York (2006)
3. Andersen, R., Peres, Y.: Finding sparse cuts locally using evolving sets. In: Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, pp. 235–244. ACM, New York (2009)
4. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science 286, 509–512 (1999)
5. Clauset, A., Moore, C., Newman, M.E.: Hierarchical structure and the prediction of missing links in networks. Nature 453(7191), 98–101 (2008)
6. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Physical Review E, 1–6 (2004)
7. Doerr, B., Fouz, M., Friedrich, T.: Social networks spread rumors in sublogarithmic time. In: Proceedings of the 43rd Annual ACM Symposium on Theory of Computing, STOC 2011, pp. 21–30 (2011)
8. Ravasz, E., Somera, A.L., D.M.Z.O., Barabási, A.L.: Hierarchical organization of modularity in metabolic networks. Science 297, 1551 (2002)
9. Easley, D., Kleinberg, J.: Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press (July 2010)
10. Flaxman, A.D., Frieze, A., Vera, J.: A geometric preferential attachment model of networks. Internet Mathematics 3(2) (2007)
11. Flaxman, A.D., Frieze, A.M., Vera, J.: A geometric preferential attachment model of networks II. Internet Mathematics 4(1), 87–111 (2007)
12. Fortunato, S.: Community detection in graphs. Physics Reports 486 (2010)
13. Hodgkinson, L., Karp, R.M.: Algorithms to Detect Multiprotein Modularity Conserved during Evolution. In: Chen, J., Wang, J., Zelikovsky, A. (eds.) ISBRA 2011. LNCS, vol. 6674, pp. 111–122. Springer, Heidelberg (2011)
14. Kannan, R., Vempala, S., Vetta, A.: On clusterings: Good, bad and spectral. J. ACM 51(3), 497–515 (2004)
15. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: KDD 2003: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 137–146 (2003)
16. Kleinberg, J.: The small-world phenomenon: an algorithmic perspective. In: Proceedings of the 32nd ACM Symposium on the Theory of Computing (2000)
17. Lancichinetti, A., Kivelä, M., Saramäki, J., Fortunato, S.: Characterizing the community structure of complex networks. PLoS ONE 5(8), e11976 (2010)
18. Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. CoRR abs/0810.1355 (2008)

19. Leskovec, J., Lang, K.J., Mahoney, M.: Empirical comparison of algorithms for network community detection. In: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, pp. 631–640 (2010)
20. Li, A., Li, J., Pan, Y., Peng, P.: Homophily law of networks: Principles, methods and experiments (2011) (to appear)
21. Li, A., Li, J., Pan, Y., Peng, P., Zhang, W.: Small core phenomenon of networks: Global influence core of the collaboration networks (2011) (to appear)
22. Li, A., Li, J., Peng, P.: Small community phenomenon in social networks: Local dimension (2011) (to appear)
23. Li, A., Peng, P.: Communities structures in classical network models. Internet Mathematics 7(2), 81–106 (2011)
24. Li, A., Peng, P.: The small-community phenomenon in networks. Mathematical Structures in Computer Science, Available on CJO doi:10.1017/S0960129511000570
25. Liao, C.S., Lu, K., Baym, M., Singh, R., Berger, B.: IsoRankN: spectral methods for global alignment of multiple protein networks. Bioinformatics 25(12), i253–i258 (2009)
26. Mihail, M., Papadimitriou, C., Saberi, A.: On certain connectivity properties of the internet topology. J. Comput. Syst. Sci. 72(2), 239–251 (2006)
27. Morris, S.: Contagion. The Review of Economic Studies 67(1), 57–78 (2000)
28. Newman, M.E.J.: Detecting community structure in networks. The European Physical Journal B 38 (2004)
29. Newman, M.E.J., Barabási, A.L., Watts, D.J. (eds.): The Structure and Dynamics of Networks. Princeton University Press (2006)
30. Onnela, J.P., Fenn, D.J., Reid, S., Porter, M.A., Mucha, P.J., Fricker, M.D., Jones, N.S.: A Taxonomy of Networks. CoRR abs/1006.5731 (June 2010)
31. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043), 814–818 (2005)
32. Porter, M.A., Onnela, J.P., Mucha, P.J.: Communities in networks. Notices of the American Mathematical Society 56, 1082–1097 (2009)
33. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. Proceedings of the National Academy of Sciences 101(9), 2658 (2004)
34. Ravasz, E., Barabási, A.L.: Hierarchical organization in complex networks. Physical Review E 67, 026112 (2003)
35. Schaeffer, S.: Graph clustering. Computer Science Review (1), 27–64
36. Spielman, D.A., Teng, S.H.: Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In: Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing, STOC 2004, pp. 81–90. ACM, New York (2004)
37. Voevodski, K., Teng, S.H., Xia, Y.: Finding local communities in protein networks. BMC Bioinformatics 10(1), 297 (2009)
38. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393, 440–442 (1998)