Community detection algorithms: a comparative analysis

[Invited Presentation, Extended Abstract]

Santo Fortunato
Complex Systems and Networks
ISI Foundation
Viale S. Severo 65
10133, Torino, Italy
fortunato@isi.it

Andrea Lancichinetti
Complex Systems and Networks
ISI Foundation
Viale S. Severo 65
10133, Torino, Italy
arg.lanci@gmail.com

ABSTRACT

Uncovering the community structure exhibited by real networks is a crucial step towards an understanding of complex systems that goes beyond the local organization of their constituents. Many algorithms have been proposed so far, but none of them has been subjected to strict tests to evaluate their performance. Most of the sporadic tests performed so far involved small networks with known community structure and/or artificial graphs with a simplified structure, which is very uncommon in real systems. Here we test several methods against a recently introduced class of benchmark graphs, with heterogeneous distributions of degree and community size. The methods are also tested against the benchmark by Girvan and Newman and on random graphs. As a result of our analysis, three recent algorithms introduced by Rosvall and Bergstrom, Blondel et al. and Ronhovde and Nussinov, respectively, have an excellent performance, with the additional advantage of low computational complexity, which enables one to analyze large systems.

Categories and Subject Descriptors

I.5.3 [**Pattern recognition**]: Clustering; J.2 [**Computer applications**]: Physical sciences and engineering—*engineering, physics*

Keywords

Networks, community structure

The modern science of networks is probably the most active field within the new interdisciplinary science of complex systems. Many complex systems can be represented as networks, where the elementary parts of a system and their mutual interactions are nodes and links, respectively [15, 4]. Complex systems are usually organized in compartments, which have their own role and/or function. In the network representation, such compartments appear as sets of nodes with a high density of internal links, whereas links between compartments have a comparatively lower density. These subgraphs are called communities, or modules, and occur in a wide variety of networked systems [10, 9].

Finding compartments may shed light on the organization of complex systems and on their function. Therefore detecting communities in networks has become a fundamental problem in network science. Many methods have been developed, using tools and techniques from disciplines like physics, biology, applied mathematics, computer and social sciences. However, it is still not clear which algorithms are reliable and shall be used in applications. The question of the reliability itself is tricky, as it requires shared definitions of community and partition which are, at present, still missing. This essentially means that, despite the huge literature on the topic, there is still no agreement among scholars on what a network with communities looks like. Nevertheless, there has been a silent acceptance of a simple network model, the *planted ℓ-partition model* [6], which is often used in the literature in various versions. In this model one “plants” a partition, consisting of a certain number of groups of nodes. Each node has a probability p_{in} of being connected to nodes of its group and a probability p_{out} of being connected to nodes of different groups. As long as $p_{in} > p_{out}$ the groups are communities, whereas when $p_{in} \leq p_{out}$ the network is essentially a random graph, without community structure. The most popular version of the planted ℓ-partition model was proposed by Girvan and Newman (GN benchmark) [10]. Here the graph consists of 128 nodes, each with expected degree 16, which are divided into four groups of 32. The GN benchmark is regularly used to test algorithms for community detection. Indeed, algorithms can be compared based on their performance on this benchmark. This has been done by Danon et al. [7]. However, the GN benchmark has two drawbacks: 1) all nodes have the same expected degree; 2) all communities have equal size. These features are unrealistic, as complex networks are known to be characterized by heterogeneous distributions of degree [1, 15, 4] and community sizes [16, 11, 8, 5, 13]. In recent papers [14, 12], we have introduced a new class of benchmark graphs (LFR benchmark), that generalize the GN benchmark by introducing power law distributions of degree and community size. The new graphs are a real generalization, in that the GN benchmark is recovered in the limit case in which the exponents of the distributions of degree and community sizes go to infinity. Most community detection algorithms perform very well on the GN benchmark, due to the simplicity of its structure. The LFR benchmark, instead, poses

VALUETOOLS 2009, October 20-22, Pisa, Italy
Copyright © 2009 ICST 978-963-9799-70-7
DOI 10.4108/ICST.VALUETOOLS2009.8046

a much harder test to algorithms, and makes it easier to disclose their limits. Moreover, the LFR benchmark graphs can be built very quickly: the complexity of the construction algorithms is linear in the number of links of the graph, so one can perform tests on very large systems, provided the method at study is fast enough to analyze them.

We have carried out a comparative analysis of the performances of algorithms for community detection on various graphs: the GN and LFR benchmarks and random graphs. Link direction, weights and the possibility for communities to overlap have been taken into account in dedicated tests. We conclude that the Infomap method by Rosvall and Bergstrom [18] is the best performing on the set of benchmarks we have examined here. In particular, its results on the LFR benchmark graphs, which are much more difficult to examine than the GN benchmark graphs, are encouraging about the reliability of the method in applications to real graphs. Among the other things, the method can be applied to weighted and directed graphs as well, with excellent performances, so it has a large spectrum of potential applications. The algorithms by Blondel et al. [3] and by Ronhovde and Nussinov (RN) [17] also look very good from our analysis and could be used as well. Furthermore, these methods have a low computational complexity, so one could use them on graphs with millions of nodes and links. On the other hand, the algorithms are not able to account for overlapping communities, so they need to be properly refined to deal with this possibility, which is common in many real systems.

One may object that, despite the features planted in the LFR benchmark, i. e. the fat-tailed distributions of degree and community size, which are actually observed in real networks, our artificial graphs are still different from real systems. For instance, the clustering coefficient [19] of the LFR benchmark is very low, due to the very small number of triangles, whereas real networks are characterized by many triangles and consequently a high clustering coefficient. On the one hand the GN benchmark also has very few triangles and low clustering coefficient (the LFR benchmark is just a generalization of the GN benchmark), nevertheless people have used it extensively for testing algorithms. On the other hand, nothing forbids to modify the building mechanism of the LFR benchmark so that it does include triangles. This is actually a potentially interesting improvement of the benchmark, that deserves some attention in the future.

Our whole analysis has made use of graphs with a “flat” community structure, without hierarchy. Many real networks instead have a hierarchical community structure, with communities inside other communities. Good methods must be able to understand when a network has no communities, a flat or a hierarchical community structure. For an analysis of this kind we would need hierarchical benchmarks. There is actually a hierarchical version of the GN benchmark [2], not yet one of the LFR benchmark, which is sorely needed. Methods to find communities in multipartite graphs have yet to be tested as well.

1. REFERENCES

[1] R. Albert, H. Jeong, and A.-L. Barabási. Internet: Diameter of the World-Wide Web. *Nature*, 401:130–131, Sept. 1999.

[2] A. Arenas, A. Díaz-Guilera, and C. J. Pérez-Vicente. Synchronization Reveals Topological Scales in Complex Networks. *Phys. Rev. Lett.*, 96(11):114102, Mar. 2006.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *J. Stat. Mech.*, P10008(10), 2008.

[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang. Complex networks: Structure and dynamics. *Phys. Rep.*, 424(4-5):175–308, February 2006.

[5] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70(6):066111, Dec. 2004.

[6] A. Condon and R. M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Struct. Algor.*, 18:116–140, 2001.

[7] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *J. Stat. Mech.*, 9:8, Sept. 2005.

[8] L. Danon, J. Duch, A. Arenas, and A. Díaz-Guilera. In C. G. and V. A., editors, *Large Scale Structure and Dynamics of Complex Networks: From Information Technology to Finance and Natural Science*, pages 93–114. World Scientific, Singapore, 2007.

[9] S. Fortunato. Community detection in graphs. June 2009.

[10] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99(12):7821–7826, June 2002.

[11] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas. Self-similar community structure in a network of human interactions. *Phys. Rev. E*, 68(6):065103 (R), Dec. 2003.

[12] A. Lancichinetti and S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Phys. Rev. E*, 80(1):016118, 2009.

[13] A. Lancichinetti, S. Fortunato, and J. Kertesz. Detecting the overlapping and hierarchical community structure in complex networks. *New J. Phys.*, 11(3):033015, 2009.

[14] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Phys. Rev. E*, 78(4):046110, 2008.

[15] M. E. J. Newman. The structure and function of complex networks. *SIAM Rev.*, 45(2):167–256, 2003.

[16] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818, June 2005.

[17] P. Ronhovde and Z. Nussinov. Multiresolution community detection for megascale networks by information-based replica correlations. *Phys. Rev. E*, 80(1):016109, 2009.

[18] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. USA*, 105:1118–1123, Jan. 2008.

[19] D. Watts and S. Strogatz. Collective dynamics of ’small-world’ networks. *Nature*, 393:440–442, June 1998.