Technical Report

University of the Basque Country UPV/EHU

Department of Computer Science and Artificial Intelligence

Communities in complex networks

Abdelmalik Moujahid

January 2013

San Sebastian, Spain
www.ccia-kzaa.ehu.es
hdl.handle.net/10810/4562

Communities in complex networks

A. Moujahid

Abstract

The study of complex networks has attracted the attention of the scientific community for many obvious reasons. A vast number of systems, from the brain to ecosystems, power grid, and the Internet, can be represented as large complex networks, i.e, assemblies of many interacting components with nontrivial topological properties. The link between these components can describe a global behaviour such as the Internet traffic, electricity supply service, market trend, etc. One of the most relevant topological feature of graphs representing these complex systems is community structure which aims to identify the modules and, possibly, their hierarchical organization, by only using the information encoded in the graph topology. Deciphering network community structure is not only important in order to characterize the graph topologically, but gives some information both on the formation of the network and on its functionality.

1 Introduction

The science of social networks is one of the pillars upon which the whole field of network science has been built. Since the early works of [1], social networks have been the object of constant analysis and study. Social networks represent the individuals of the population as nodes and the interaction pattern among individuals as links between these nodes. The links therefore may refer to very different attributes such as friendship among classmates, sexual relations among adults, or just the belonging to common institutions or work teams (collaborative interactions). The importance of these networks goes beyond social sciences and affects our understanding of a variety of processes ranging from the spreading of sexually transmitted diseases to the emergence of consensus and knowledge diffusion in different kinds of organizations and social structures.

The study of the underlying laws governing the dynamics and evolution of complex systems and the characterization of their network representations reveals that large-scale networks are generally characterized by complex topologies and heterogeneous structures. The connectivity structure of these networks often features an organization in communities (clusters, modules), revealing the existence of specialized groups of vertices that share common properties and/or play similar roles within the network. Community structure appears in many networked systems, including a variety of biological, social, technological, and information networks [2]. Communities may be groups of related individuals

in social networks, sets of Web pages dealing with the same topic, biochemical pathways in metabolic networks, etc.

A network is a graph composed by a large number of highly interconnected units, and a community is characterized by a large number of edges connecting vertices within individual groups, with only low concentrations of edges between these groups [3, 4]. The aim of community detection is to identify the modules and, possibly, their hierarchical organization, by only using the information encoded in the graph topology. The problem has a long tradition and it has appeared in various forms in several disciplines [6, 7, 8]. The general notion of community structure in complex networks was first pointed out in the physics literature by Girvan and Newman [3].

This review is organized as follows. The next section gives a brief account of the main quantities and measures that are commonly used to characterize the properties of complex networks (see [19] for an extensive review). Section 3 reports some algorithms that have been proposed during the last years to deal with the problem of community detection in complex network. In Section 4 we introduce benchmark graphs commonly used to test community detection algorithms. A brief description of the evolution of social network is discussed in Section 5. Finally, Section 6 gives some useful packages for network analysis and community structure detection.

2 Definitions and measures of complex networks

Formally, a network is represented by a graph. An undirected graph G is defined by a pair of sets G=(V,E), where V is a non-empty countable set of elements, called vertices or nodes, and E is a set of unordered pairs of different vertices, called edges or links. The edge (i,j) joins the vertices i and j, which are said to be adjacent or connected. It is also common to call connected vertices neighbors or nearest neighbors. The total number of vertices in the graph is denoted as N and defines the order of the graph. In many biological and physical contexts, N defines the physical size of the network since it identifies the number of distinct elements composing the system. However, in graph theory, the size of the graph is identified by the total number of edges E. For a graph of size N, the maximum number of edges is $N(N-1)/2$.

There exists an intimate relationship between graph theory and matrix theory with both fields benefiting from insights in the other. A graph can be completely described by giving the adjacency matrix A, a $N \times N$ square matrix whose entry a_{ij} $(i, j = 1, 2, ..., N)$ is equal to 1 when the link l_{ij} exists, and zero otherwise. An important feature of many graphs, which helps in dealing with their structure, is their sparseness. The number of edges E for a connected graph (i.e., with no disconnected parts) ranges from (N-1) to N(N-1)/2. The graph is said to be sparse if the number of edges E scales as N^{α} with $\alpha < 2$, and is considered dense if E scales as N^2.

2.1 Degree distribution

The degree k_i of the node i is the number of edges incident with that node, and is obviously defined in terms of the adjacency matrix A as $k_i = \sum_{j \in N} a_{ij}$. The most basic topological characterization of a graph G can be obtained in terms of the degree distribution $P(k)$, defined as the probability that a node chosen uniformly at random has degree k or, equivalently, as the fraction of nodes in the graph having degree k. The n-moment of $P(k)$ is defined as:

$$< k^n >= \sum_k k^n P(k).$$

The first moment $< k >$ is the mean degree of G, while the second moment measures the fluctuations of the connectivity distribution. Graphs are usually said homogeneous (heterogeneous) if the value of the second moment of the degree distribution is small (large) if compared with the value of the first moment of the same distribution.

A first approximation of homogeneous networks is the uncorrelated random graph model proposed by Erdos and Renyi in 1959 with the original purpose of studying, by means of probabilistic methods, the properties of graphs as a function of the increasing number of random connections. This model consists in drawing an undirected edge with a fixed probability p between each possible pair out of N given nodes. The resulting graph shows a binomial degree distribution with average $< k > \simeq Np$, which for large N can be approached by a Poisson distribution. In order to account for degree heterogeneity, other constructions have been proposed for random graphs with arbitrary degree distributions [10, 11].

Since many real networks are not static but evolving, with preferential attachment mechanisms, many models of growing networks have also been introduced. The Barabsi and Albert model [12], has become one of the most famous models for complex heterogeneous networks. The model begins from a small set of m fully interconnected nodes, new nodes are introduced one by one. Each new node is connected to m existing nodes according to the preferential attachment rule, i.e., with probability proportional to their degree, and creates links with them. The procedure stops when the required network size N is reached. The obtained network has average degree $< k >= 2m$, small clustering coefficient (see Section 2.3) and a power law degree distribution $P(k) \propto k^{-\gamma}$, with $\gamma = 3$ (when $\gamma \leq 3$ the graphs are referred to as scal-free networks).

2.2 Characteristic path length

The distribution of geodesic (the shortest path between two nodes) play a crucial role in all processes involving transport of information across the network. It is therefore useful to represent all the shortest path lengths of a graph G as a matrix D in which the entry d_{ij} is the length of the geodesic from node i to node j. The maximum entry of the matrix D is called the diameter of the graph, and gives a measure of the maximal extent of a graph.

A measure of the statistically typical separation between any two nodes in the graph is given by the characteristic path length defined as the mean of geodesic lengths over all couples of nodes:

$$L = \frac{1}{N(N-1)} \sum_{i,j \in V, i \neq j} d_{ij}$$

A graph is said to display the small world property if L scales with the logarithm of N, i.e, an increase of the network size do not affect substantially the mean distance between any pair of nodes of the graph.

2.3 Clustering coefficient

The concept of clustering of a graph refers to the tendency observed in many natural networks to form cliques (A clique is a complete n-subgraph of size $n < N$) in the neighborhood of any given vertex. In this sense, clustering implies the property that, if the vertex i is connected to the vertex j, and at the same time j is connected to l, then with a high probability i is also connected to l. The clustering of an undirected graph can be quantitatively measured by means of the clustering coefficient C which measures the local group cohesiveness [9]. Given a vertex i, the clustering $C(i)$ of a node i is defined as the ratio of the number of links between the neighbors of i and the maximum number of such links. If the degree of node i is k_i and if these nodes have e_i edges between them, we have:

$$C(i) = \frac{2e_i}{k_i(k_i-1)} = \frac{\sum_{j,m} a_{ij} a_{jm} a_{mi}}{k_i(k_i-1)}$$

The average clustering coefficient of a graph is simply given by the average of $C(i)$ over all the nodes in G:

$$C = \frac{1}{N} \sum_{i \in V} C(i)$$

2.4 Node betweenness

The communication of two nonadjacent nodes, say j and k, depends on the nodes belonging to the paths connecting j and k. In order to account quantitatively for the role of vertices which may be crucial for connecting different regions of the network by acting as bridges, the concept of betweenness centrality has been introduced [Newman, 2001]. More precisely, the betweeness b_i of a node i, sometimes referred to also as load, is defined as:

$$b_i = \sum_{j,k \in V, j \neq k} \frac{n_{jk}(i)}{n_{jk}}$$

where n_{jk} is the total number of different shortest paths going from j to k and $n_{jk}(i)$ is the subset of those distances passing through the node i.

According to this definition, central nodes are therefore part of more shortest paths within the network than less important nodes. This centrality measure of a node is often used in transport networks to provide an estimate of the traffic handled by the vertices, assuming that the number of shortest paths is a zero-th order approximation to the frequency of use of a given node. Analogously to the node betweenness, the betweenness centrality of edges can be calculated as the number of shortest paths among all possible vertex couples that pass through the given edge. Edges with the maximum score are assumed to be important for the graph to stay interconnected. These high-scoring edges are the "bridges" that inter-connect modules of nodes. Removing them frequently leads to unconnected communities of nodes. These centralized edges are particularly important for decreasing the average path length among nodes in a network, for speeding up the diffusion of information, or for increasing the size of the part of the network at a given distance from a node.

3 Deciphering community structure

A significant step to understand the properties of a network consists in determining its communities. However, the best way to establish the community structure of a network is still disputed. During the last years, many algorithms have been proposed to extract the optimal partition of a network into communities ranging from traditional methods (Graph partitioning, spectral clustering), modularity-based methods [5] and synchronization-based dynamics algorithms [13, 14, 16].

To determine the optimal number of modules or clusters, most of these algorithms adopt the criterion of maximum modularity (Q) [5]. The *modularity* is defined as the fraction of links within communities minus the expected fraction of such links in a random network. This measure provides a way to determine if a certain description of the graph in terms of communities is more or less accurate. High values of modularity should indicate good partitions with many more internal connections than expected at random. For an arbitrary network, and an arbitrary partition of that network into N_c communities We can define a $(N_c \times N_c)$ size matrix e whose entries e_{ij} give the fraction of edges that in the original graph connect subgraph i to subgraph j. The sum of the any row (or column) of e, $a_i = \sum_j e_{ij}$ corresponds to the fraction of links connected to subgraph i. If the network does not exhibit community structure (random graph), the expected value of the fraction of links within partitions can be estimated. It is simply the probability that a link begins at a node in i, a_i, multiplied by the fraction of links that end at a node in i, a_i. So the expected number of intra-community links is just a_i^2. The modularity of a subgraph division is then defined by

$$Q = \sum_i (e_{ii} - a_i^2)$$

Recently, a global criterion called Surprise, which implicitly assumes a more

complex definition of community has been proposed [26]. In this case finding the optimal community structure of a undirected graph is equivalent to maximize the following parameter:

$$S = -\log \sum_{j=p}^{min(M,n)} \frac{\binom{M}{j}\binom{F-M}{n-j}}{\binom{F}{n}}$$

where F is the maximum possible number of links, n is the observed number of links, M is the maximum possible number of intracommunity links for a given partition, and p is the total number on intracommunity links actually observed in that partition. This criterion measures the improbability of finding by chance a partition with the observed enrichment of intracommunity links in a random graph.

Others methods seek the optimal partition by minimizing the compression of the information that best describes the network [20], minimizing the Hamiltonian of a Potts-like spin model that represents the graph [21], or deducing the maximum-likelihood model that best fits the structure of the network [22], to name just a few examples. Figures 1 reports an example of technological complex networks with community structure.

4 Benchmarks to compare the performance of community detection algorithms

As reported early, characterizing the community structure of complex networks is a key challenge in many scientific fields. To this end, many algorithms and methods have been proposed with a performance that varies greatly, depending on the topological parameters of the analyzed network. The main problem is then to estimate the accuracy of a method and to compare it with other methods. This issue of testing is as crucial as devising powerful community detection algorithms.

Testing an algorithm consists in analyzing a network with a well-defined community structure and recovering its communities. Ideally, one would like to have many instances of real networks whose modules are precisely known, but this is unfortunately not the case. Therefore, the most extensive tests are performed on computer generated networks, with a built-in community structure. The most popular benchmark for community detection is a class of networks introduced by Girvan and Newman [3] in which communities are, by definition, Erdos-Renyi subgraphs. This makes this benchmark inappropriate for representing real-world networks since the latter exhibit much more heterogeneous degree distributions [19]. A good benchmark should have a skewed degree distribution, similar to real networks, and should include communities of very different sizes [24]. These benchmarks are characterized by an initial well-defined community structure which is degraded by randomly rewiring links. During this process, the proportion of intercommunity links grows and the original communities gradually disappear.

5 Evolution of social networks

In many social network evolution studies, the underlying process for network change is assumed to be located in the network structure. As real and online social systems grow ever larger, their analysis becomes more complicated, due to their intrinsic dynamic nature, the heterogeneity of the individuals, their interests, behavior etc. In this perspective, revealing the community structures, i.e., the identification of more homogeneous groups of individuals, is a major challenge. In this context, one has to distinguish the communities as typically intended in social network analysis [27] from a broader definition of communities. In a more general context, for e.g. providing recommendation strategies, one is more interested in finding communities of users with homogeneous interests and behavior. Such homogeneity is independent of contacts between the users although in most cases there will be at least a partial overlap between communities defined by the user contacts and those by common interests and behavior.

Recently the modern Information and Communication Technology (ICT) has opened new interaction modes between individuals, like mobile phone communications and online interactions enabled by the Internet. Such new social exchanges can be accurately monitored for very large systems, including millions of individuals, whose study represents a huge opportunity for social science. Social networking services, like Myspace (www.myspace.com), Friendster (www.friendster.com), Facebook (www.facebook. com), etc. have become extremely popular in the last years. They are online platforms that allow people to communicate with friends and other users through private or public messages and a chat feature, and unite people with common interests and/or beliefs through groups and other pages.

6 Network analysis packages

- **Graphviz:** is an open-source software for graph visualization, developed by researchers at AT&T.
- **igraph:** is a package for the generating, manipulating, analyzing, and visualizing network graphs, of sizes up to millions of vertices and edges.
- **Jerarca:** is a suite of hierarchical clustering algorithms that provides a simple and easy way to analyze complex networks. It is designed to efficiently convert unweighted undirected graphs into hierarchical trees by means of iterative hierarchical clustering. Moreover, Jerarca detects and returns the community structure of the network.
- **MultiDendrograms:** is a simple yet powerful program to make the Hierarchical Clustering of real data.
- **Pajek:** is a freely available package for the visualization of large networks. It also has a suite of network analysis tools, mainly oriented towards social

network analysis.

- **Redatools:** is a collection of programs to analyze complex networks, with special emphasis on community detection and mesoscales Search.
- **statnet:** is a suite of software packages for network analysis and modeling, that allows for the estimation, evaluation, and simulation of network models, as well as network analysis and visualization.
- **Workbench:** is a Large-Scale network analysis, modeling and visualization toolkit for biomedical, social science and physics research
- **yEd Graph Editor:** is a powerful desktop application that can be used to quickly and effectively generate high-quality diagrams.

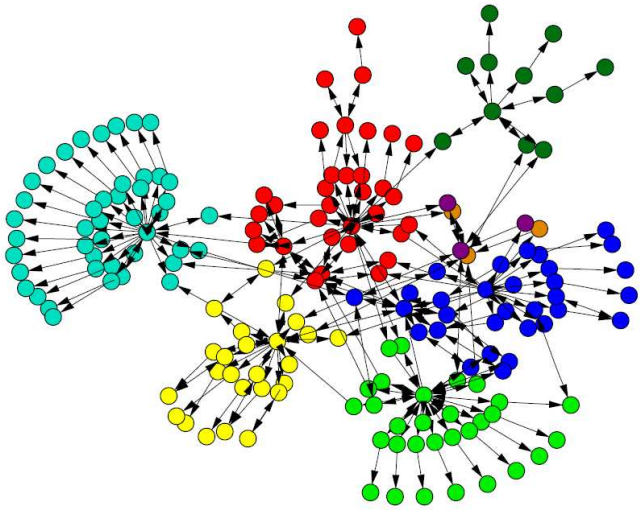

Figure 1: Community structure in technological networks. Sample of the web graph consisting of the pages of a web site and their mutual hyperlinks, which are directed. Communities, indicated by the colors, were detected with the algorithm of Girvan and Newman, by neglecting the directedness of the edges. Reprinted figure with permission from Ref. [5]. ©2004, by the American Physical Society.

References

[1] Moreno, J. L. (1934),Who Shall Survive? Foundations of Sociometry, Group Psychotherapy, and Sociodram, Beacon House.

[2] Newman M.E.J. , Networks: An Introduction, Oxford University Press, Oxford, 2010.

[3] Girvan M. and M. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. U.S.A. 99, 7821 (2002).

[4] Newman M.E.J. Fast algorithm for detecting community structure in networks. Physical Review E 69, 066133 (2004)

[5] Newman M.E.J. , Girvan M. Finding and evaluating community structure in networks, Phys. Rev. E 69 (2) 026113 (2004)

[6] Chen J., Yuan, B. Detecting functional modules in the yeast protein-protein interaction network. Bioinformatics, 22 (18), pp. 2283-2290 (2006).

[7] Flake G.W., Lawrence S., Lee Giles, C., Coetzee, F.M. Self-organization and identification of web communities. Computer, 35 (3), pp. 66-71 (2002).

[8] Krause A.E., Frank K.A., Mason, D.M., Ulanowicz, R.E., Taylor, W.W. Compartments revealed in food-web structure. Nature, 426 (6964), pp. 282-285 (2003).

[9] Watts. D.J., and S.H. Strogatz, Nature (London) 393, 440 (1998).

[10] Catanzaro, M., M. Boguna, and R. Pastor-Satorras, Phys. Rev. E 71(2), 027103. (2005).

[11] Goh, K.-I., B. Kahng, and D. Kim, Phys. Rev. Lett. 87(27), 278701, (2001)

[12] Barabasi, A.-L., and R. Albert, Science 286, 509 (1999).

[13] Alex Arenas, Albert Daz-Guilera, and Conrad J. Perez-Vicente, Synchronization reveals topological scales in complex networks. Phys. Rev. Lett. 96, 114102 (2006).

[14] Boccaletti S., Ivanchenko M., Latora V., Pluchino A. , and Rapisarda A. Detecting complex network modularity by dynamical clustering. Physical Review E 75, 045102(R) (2007)

[15] Li D., Leyva I., Almendral J.A., Sendia-Nadal I., Buld J.M., Havlin S., and Boccaletti S., Synchronization Interfaces and Overlapping Communities in Complex Networks. Phys. Rev. Lett. 101, 168701 (2008)

[16] A. Moujahid, A. D'Anjou, B. Cases. Chaos Solitons and Fractals, 45, 9-10, pp. 1171-1179 (2012)

[17] Santo Fortunato, Physics Reports 486 (2010) 75-174

[18] Boccaletti S. The synchronized dynamics of complex systems. Monograph series on nonlinear science and complexity. Vol. 6, ISBN: 978-0-444-52743-1 (2008)

[19] S. Boccaletti, V. Latorab, Y. Moreno, M. Chavez, D.-U. Hwang. Physics Reports 424, 175-308 (2006)

[20] M. Rosvall and C. T. Bergstrom,Proc. Natl. Acad. Sci. (USA) 105, 1118 (2008)

[21] P. Ronhovde and Z. Nussinov,Phys. Rev. E80, 016109 (2009)

[22] Newman M. E. J. and Leicht E. A. ,Proc. Natl. Acad. Sci. (USA) 104, 9564 (2007)

[23] P.F. Jonsson, T. Cavanna, D. Zicha, P.A. Bates, Cluster analysis of networks generated through homology: Automatic identification of important protein communities involved in cancer metastasis, BMC Bioinf. 7 (2006) 2.

[24] A. Lancichinetti, S. Fortunato, and F. Radicchi, Phys. Rev. E 78, 046110 (2008).

[25] Rodrigo Aldecoa and Ignacio Marin. PHYSICAL REVIEW E 85, 026109 (2012)

[26] Rodrigo Aldecoa and Ignacio Marin. PLoS ONE 6(9): e24195. doi:10.1371/journal.pone.0024195 (2011)

[27] Freeman, L., 2004, The Development of Social Network Analysis:A Study in the Sociology of Science (BookSurge Publishing, Vancouver, Canada).