ISBN: 978-972-8924-31-7 © 2007 IADIS

IDENTIFYING COMMUNITIES IN SOCIAL NETWORKS: A SURVEY

Anna Puig-Centelles, Oscar Ripolles, Miguel Chover
DLSI, Universitat Jaume I, Castellón (Spain)

ABSTRACT

Many networks of interest now include a variety of social and technological networks, which are naturally divided into communities or modules. How to identify these community structures has considerably attracted the attention of researchers, and it is the aim of this paper. A review of the main algorithms used for finding communities is presented, starting with graph theory and including the latest algorithms for detecting Web communities. The existence of communities that can overlap is also taken into account, and the main commonly used software tools for depicting communities are pointed out.

KEYWORDS

Social networks, clustering, Web-community, community algorithms.

1. INTRODUCTION

In recent years, evidence has grown that very diverse systems in different fields can be described as complex networks. Traditionally, networks have been classified into two large groups: technological, which involve sets of items grouped in human-made networks to share technical resources, and Social, which include groups of people with some pattern of contacts or interaction between them. It is also possible to find the Social-Technological networks, which establishes social connections over a technological network, identifying links over which the information can flow, like the P2P networks overlaid on the Internet. A more precise characterization could include information or biological networks (Newman 2003).

An important property of networks is the existence of a community structure. A community is a set of items that share an environment and are created when one or more entities claim an interest in the same topic. It is interesting to study and identify these groups from the sociological, economical and technological point of view. Extracting communities from a network has applications to study interactions between groups of people or to automate the process of creating Web directories like http://directory.google.com.

Web communities are sets of Web pages that are created by individuals or associations having a common interest. The community Web can be modeled as a graph where vertices are Web pages and hyperlinks are edges. A Web community is defined as a set of sites that have more links between members of the community than between nonmembers, and we can identify a Web community on a topic by extracting densely connected structure in the Web graph. A crawler explores the content offered for each node and extracts its information. A WebCrawler harvests the Web collecting documents referenced by a set of URLs.

1.1 Motivation

Given the existence of a large amount of literature on these topics, the objective of the authors is to give a wide vision of the overall process of finding communities in social networks. This process includes clustering, overlapping communities and software tools that help the user to visualize the communities found.

The outline of this paper is as follows. In Section 2 we review the main algorithms for the extraction of communities. Section 3 describes the need to detect overlapped communities. Section 4 includes a short study of the existing software for community detection. Section 5 concludes and gives future work ideas.

2. IDENTIFICATION OF COMMUNITIES

As we have previously commented, we can represent any social network as a simple graph. In these graphs, the edges used can be directed/undirected edges, implying symmetric/nonsymmetric relationships; weighted/unweighted edges, implying that any edge has been assigned or not with *a priori* strength. The set of groups obtained, as well as their relationships, is what we call the community structure of the network.

A cluster can be seen as a connected dense region of points, surrounded by a less dense region. Finding clusters is thus finding boundaries between density regions. Traditionally the research in graph theory has been concentrated on modeling the spread of information from one vertex to the rest of a graph. This approach has been very useful for modeling some particular types of phenomena like disease spread in a social community or virus infection and error propagation in computer networks.

A basic requirement for a general community finding algorithm is that it should find natural divisions among the vertices without requiring the investigator to specify how many communities there should be, or placing restrictions on their sizes. The authors divide their ideas mainly into two different categories: partitional and hierarchical clustering. The first one relies on the graph theory, offers a unique split and is particularly useful for solving computer science problems. The second one, on the contrary, allows us to obtain a nested series of partitions and it is more suitable for sociology.

2.1 Partitional Approaches

Although these methods are more related to computer science than to social networks, it is important to comment on the main algorithms which are the background of recent approaches.

The Spectral bisection finds the best possible division of the complete graph into two groups, and then further subdivides those two until we have the required number of groups (Pothen, et al. 1990). The Kernighan-Lin is a completely different approach to graph bisection and appears to give good results in practice because it runs moderately quickly. It improves on an initial division of the network by optimization of the number of community edges using a greedy algorithm (Kernigan 1970). The K-means algorithm (Macqueen 1967) begins by randomly placing the nodes into clusters. Then, the normalized center of gravity of these clusters is computed, placing each node into its nearest cluster by using the Euclidean distance.

A recent approach has been proposed by Wu and Huberman (Wu & Huberman 2004), where they apply the properties of resistor networks. An interesting feature is that it can also be used to find the particular community to which a specified vertex belongs, without first having to find all communities in the network.

2.2 Hierarchical Approaches

This category joins the main efforts for community detection related to social networks. The basic idea of these approaches is to construct a hierarchy of clusters, so that small clusters will be nested into larger ones. These algorithms are considerably more useful than the spectral bisection method because they allow us to split the network into any number of communities

We can find two main families of techniques: agglomerative and divisive. This categorization is based on whether they merge or split the clusters, using a similarity criterion. Every hierarchical clustering algorithm, apart from its own way of extracting communities, has a related metric that represents the similarity between the nodes. There are a variety of ways of defining this similarity: Euclidean distance, Pearson correlation or structural equivalence. A more comprehensive review of these metrics can be found in (Scott 2000).

In an agglomerative method, edges are added to an initially empty network starting with the vertex pairs with highest similarity. In a divisive method, we start with the initial network and find the least similar connected pairs of vertices and then remove the edges between them. The procedure can be halted at any point, and the resulting components in the network become the communities. The entire progress of the algorithm offers a complete graph which can be represented as a dendrogram such as that shown in Figure 1.

Among the hierarchical methods, the algorithm of Girvan and Newman (Girvan & Newman 2002) presents an important improvement. They use the metric called edge betweenness which represents the number of shortest paths between pairs of vertices that run along an edge. It involves simply calculating the betweenness of all edges in the network and removing the one with the highest betweenness, and repeating this process until no edges remain.

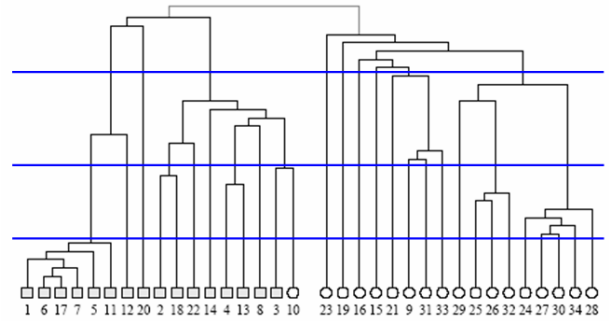

Figure 1. Dendrogram of the communities found in the Zachary Karate Club Network. Horizontal cuts (blue) through the dendrogram represent the communities obtained at different halting points.

The main disadvantages of this method are its long processing time and the fact that there is no guide to how many communities a network should be split into. To approach this problem, Newman and Girvan proposed that the divisions the algorithm generates be evaluated using a metric they call modularity. This proposal is known as Fast GN (Newman 2004), and its main advantage is its speed, even though in general it gives worse results. Several authors presented other approaches to optimizing the GN speed, like the Tyler algorithm (Tyler, et al. 2005), which calculates the betweennes for only a selected subset of the nodes, and the Radicchi algorithm (Radicchi, et al. 2004), which uses a local measure. Finally, it is important to comment on (Brandes 2001), where other algorithms for betweenness were introduced.

More recently, (Son et al. 2006) proposed a method where the breakup is simulated by studying the ferromagnetic random field Ising model (FRFIM). The weak point of the method is the time complexity, so the number of nodes considered is limited. Another efficient method can be found in (Boccaletti, et al. 2006), which is based on the cluster de-synchronization properties of phase oscillators.

Table 1. Clustering algorithms considered in this review. In the cost cells, n represents the nodes of a network, m the links and Θ represents the time complexity. Some cells include the cost in an arbitrary network and in a sparse graph.

Algorithm	Type	Cost	Reference
Spectral bisection	Partitional	$O(n^3)$	(Pothen et al. 1990)
Kernighan-Lin	Partitional	$O(n^2)$	(Kernigan 1970)
K-means	Partitional	$O(n^3)$	(Macqueen 1967)
Wu and Huberman	Partitional	$O(n+m)$	(Wu & Huberman 2004)
GN	Hierarchical Divisive	$O(m^2n)$ or $O(n^3)$	(Girvan & Newman 2002)
Tyler et. al	Hierarchical Divisive	$O(m^2n)$ or $O(n^3)$	(Tyler et al. 2005)
Radicchi et. al	Hierarchical Divisive	$O(m^4/n^2)$ or $O(n^2)$	(Radicchi et al. 2004)
Brandes et. al	Hierarchical Divisive	$O(nm)$	(Brandes 2001)
FRFIM	Hierarchical	$O(n^{2+\Theta})$	(Son et al. 2006)
Dynamical algorithm	Hierarchical	$O(nm)$	(Boccaletti et al. 2006)
Fast GN	Hierarchical Divisive	$O((m+n)n)$ or $O(n^2)$	(Newman 2004)

We depict a comparison in Table 1 summarizing the approaches presented in this section and in the previous one. Between the whole set of algorithms presented, we can say that it is possible to find algorithms that run in linear time, but at the expense of being imprecise or of needing *a priori* the number of expected communities. Other methods are more computationally costly, but present better features, mainly better partitions. Thus, it is important to choose a method according to the necessities of the case referred. More comparative studies around this issue can be found in (Gustafsson, et al. 2006) or (Danon, et al. 2005); in (Ino, et al. 2005) we can find a review of two specific methods for finding Web communities.

2.3 Cluster Validity Algorithms

A problem arises when deciding how many groups the algorithm should look for. After performing the clustering process, there is no way of knowing if the partition obtained is relevant or if it would be possible to find a better one. In this sense, a cluster validity algorithm should be able to identify the optimal partition of nodes, according to the measure of quality chosen.

A wide range of validity indexes can be found in the literature, (Davies & Bouldin 1974), (Davies & Bouldin 1979) or (Rousseeuw 1987). Among them, the modularity Q (Newman & Girvan 2004) is considered the most used measure to know the quality of the partitioning of a network into communities.

2.4 Algorithms for the Extraction of Web Communities

A very important application of Web communities discovery is the enhancement of Web searches. Instead of performing the identification of all the communities, these methods only search for the community related to a specified topic. Among the main contributions we can mention: the bibliometric methods like cocitation and bibliographic coupling (Garfield 1979), the Hyperlink Induced Topic Search (HITS) algorithm (Kleinberg 1998), the bipartite subgraph identification (Kumar, et al. 1999), the spreading activation energy (SAE) (Pirolli, et al. 1996) and the famous Google (Brin & Page 1998), based on PageRank.

On the one hand, we can say that the cocitation, the bibliographic coupling and the bipartite subgraph identification are approaches that identify well-defined graph structures that exist inside a narrow region of the Web graph. In these approaches, the structures are correctly identified but they fail to find large related subsets of the Web graph because the localized structures are simply too small.

On the other hand, PageRank, HITS, and SAE are global approaches that iteratively propagate weights through a significant portion of the Web graph, where the weights reflect an estimate of page importance (PageRank), how authoritative or hublike a Web page is (HITS), or how close a candidate page is to a starting region (SAE). They operate on large subsets of the Web graph and can therefore identify large collections of related pages. HITS and PageRank are based on spectral graph partitioning, so sometimes it is difficult to understand why a given page has been included in the results. In practice, it is possible to achieve meaningful results only when using textual content in a pre-processing (HITS) or post-processing (PageRank) step.

3. OVERLAPPING COMMUNITIES

The communities of a network are not always isolated from each other, as an individual can be a member of different social groups at the same time. The previously mentioned methods for dividing the network into smaller pieces do not allow for overlapping communities, although overlaps are generally assumed to be crucial features of communities. In a network, the strategy relies on locating the large complete subgraphs in the network first, and then looking for the connected subsets by studying the overlap between them.

The Clique Percolation Algorithm (CPA) (Gergely Palla & Vicsek 2005) is a novel and efficient approach to the identification of overlapping communities in large real networks. This algorithm uses the communication graph, which analyzes the social behaviour and evolution of the society as a whole, as well as its individual members. Non-overlapping methods inevitably lead to a tree like hierarchy of communities, while this overlapping approach allows the construction of an unconstrained network of communities. As an evolution of the CPA, the way how links between communities are born in an overlapping network is studied in (Pollner, et al. 2006) and a simple model for the dynamics of overlapping communities is introduced.

4. SOFTWARE TOOLS FOR NETWORK COMMUNITIES

Many software tools have been developed to help the users extract, analyze and visualize communities in social networks. In (Huisman & Van Duijn 2003) there is a collection of free and commercial software for network analysis that reviews and compares with examples some of these network tools.

In this paper we will just refer to a couple of free tools for analyzing and visualizing data. CFinder (Eötvös University 2005) is a tool based on the Clique Percolation Method, allowing a node to belong to more than one group. Pajek (Batagelj & Andrej Mrvar 2005) is another freeware tool that can be used to support abstraction by recursive decomposition of a large network into several smaller networks that can be treated with more sophisticated methods. Pajek calculates degree, closeness and betweenness measures for large nets, but does not implement more complex measures such as information or power centrality.

5. CONCLUSIONS

In this paper we have reviewed the different aspects of community detection. We have started with the identification algorithms and the validation metrics, which give us a measure of the quality of the existent communities and also help us to determine the ideal sizes of the groups found. Nevertheless, the fact that the communities can overlap leads to different algorithms. Finally, there is a wealth of related software which includes these algorithms and also offers some tools for analyzing the information from the communities. Nevertheless, we went more deeply into Web-communities because of the raising interest in the interactions between people through the Internet.

The number of close relationships a person may have within a community is necessarily limited to quite a small number, independently of the type. This may be the consequence of the fact that establishing close relationships is normally very time consuming, and time is a limited resource for every node. Therefore, it is reasonable to think that each member devotes time to his neighbors proportionally to the information obtained. An important aspect to consider is the temporal component, how these linked networks evolve.

REFERENCES

V. Batagelj & S. A. Mrvar, 2005. Pajek, Program for Analysis and visualization of large Networks. Reference Manual.

S. Boccaletti, et al., 2006. Detection of Complex Networks Modularity by Dynamical Clustering.

U. Brandes, 2001. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2), pp 163-177.

S. Brin & L. Page, 1998. The anatomy of a large-scale hypertextual Web search engine. Comp. Net. and ISDN Systems.

L. Danon, et al., 2005. Comparing community structure identification. *J. Stat. Mech*.

D. L. Davies & D. W. Bouldin, 1974. Well separated clusters and optimal fuzzy partitions. *In J.Cybernetics*, Vol. 4.

D. L. Davies & D. W. Bouldin, 1979. A cluster separation measure. *In IEEE Trans. Patt. Anal. Machine Intell.*

Department of Biological Physics, Eötvös University, Budapest, Hungary, 2005. CFinder's User Manual.

E. Garfield, 1979. *Citation Indexing: Its Theory and Application in Science*. J. Wiley and Sons, NY.

I. Palla, et al.,2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*.

M. Girvan & M. E. J. Newman, 2002. Community structure in social and biological networks. *Proc. Natl. Acad. Sci.*

M. Gustafsson, et al., 2006. Comparison and validation of community structures in complex networks. *Physica A: Statistical Mechanics and its Application,*.367, pp 559-576.

M. Huisman & M. Van Duijn, 2003. *Software for social network analysis*. NY: Cambridge University Press.

H. Ino, et al., 2005. A Comparative Study of Algorithms for Finding Web Communities. *In ICDEW '05*.

B. Kernigan, 1970. An Efficient Heuristic Procedure for Partitioning Graphs. *Bell System Technical Journal*.

J. M. Kleinberg, 1998. Authoritative sources in a hyperlinked environment. *In SODA '98*.

R. Kumar, et al., 1999. Trawling the Web for emerging cyber-communities. *Computer Netowrks*, vol 31, pp 1481-1493.

J. B. Macqueen, 1967. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symp. on Mathematical Statistics and Probability*, pp 281-297.

M. E. J. Newman, 2003. The structure and function of complex networks. *SIAM Review*, 45:167.

M. E. J. Newman, 2004. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69, 66–133.

M. E. J. Newman & M. Girvan, 2004. Finding and evaluating community structure in networks. *Ph. Rev. E*, 69:026113.

P. Pirolli, et al., 1996. Silk from a Sow's Ear: Extracting Usable Structures from the Web. *In Proceedings ACM Press*.

P. Pollner, et al., 2006. Preferential attachment of communities: the same principle, but a higher level. *Europhy. Letters*.

A. Pothen, et al., 1990. Partitioning sparse matrices with eigenvectors of graphs. *SIAM J. Matrix Anal. Appl.,* vol 11(3).

F. Radicchi, et al., 2004. Defining and identifying communities in networks. *Pr Nat Acad Sci USA*, 101(9), 2658–2663.

P. Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J.C.Appl. Math*.

J. Scott, 2000. *Social Networks Analysis: a handbook*. SAGE Publications, London.

S.W. Son, et al., 2006. Random field Ising model and community structure in complex networks. *The Euro. Phys. Journal*

J. R. Tyler, et al., 2005. E-Mail as Spectroscopy: Automated Discovery of Community Structure within Organizations. *The Information Society*, vol 21(2), pp 143-153.

F. Wu & B. Huberman, 2004. Finding communities in linear time: a physics approach. *The European Physical Journal*.