Clustering Methods for Detecting Communities in Networks

H

Ademir Cristiano Gabardo
Universidade Tecnológica Federal do Paraná - UTFPR, Brazil

Heitor S. Lopes
Universidade Tecnológica Federal do Paraná - UTFPR, Brazil

INTRODUCTION

Real-world networks, such as social networks, enterprise relationships, and the Internet itself, present large amounts of data that can be represented as networks and organized according to some criteria. Such criteria can be, for instance, a measure of similarity, connectivity or a physical distance. In the last years, many efforts have been spent in graph clustering, so as to develop and apply efficient computational methods to group massive data and find communities in networks (Frank, 1996).

As an example we can address social networks. Social networks are groups of individuals or entities that share one or more types of relationships, these relationships can be of various types, common interests, degrees of kinship, shared services, etc. With the popularization of the Internet, there is an increasing number of connected devices, and even more people and organizations are sharing information. Consequently, social networks are becoming ubiquitous (Kumar, Novak, & Tomkins, 2010). Popular social networks such Twitter, Facebook, Google, etc. are widely known by the general public. There is also a large amount of other social related data among the Internet and other networks forming implicit social networks. For instance, citation networks, e-mail traffic, phone users, coworkers, classmates, etc.

Social networks can reveal several aspects of the social behavior of their users, providing relevant information about relationships, identification of influential groups, spread of information, political behavior or even epidemic diseases. The analysis of complex networks has arisen in many areas, such as sociology, communications, computer science, physics and biology.

In this sense, it is relevant to identify clusters, structural communities where a large number of edges join vertices as a cohesive group, a strongly related group of members which can be described as an independent portion of the network or a subgraph.

Usually, methods for detecting communities in large networks are computationally intensive, demanding high processing power. To achieve good clustering results, efficient methods to discover communities in complex networks are needed.

There are several approaches to group the subjects in complex networks. e.g.: Graph Degree Linkage (Zhang et al., 2012), Hierarchical Clustering Algorithms (Murtagh, 1983), Nearest Neighbor Clustering (Ertoz et al., 2002), Partition Algorithms (Fortunato, 2010), etc. Some clustering methods are mathematically formulated to evaluate the connections between vertices of a graph, instead of being focused on similarity measures. The choice of methods depends on which kind of information the social network analyst is pursuing.

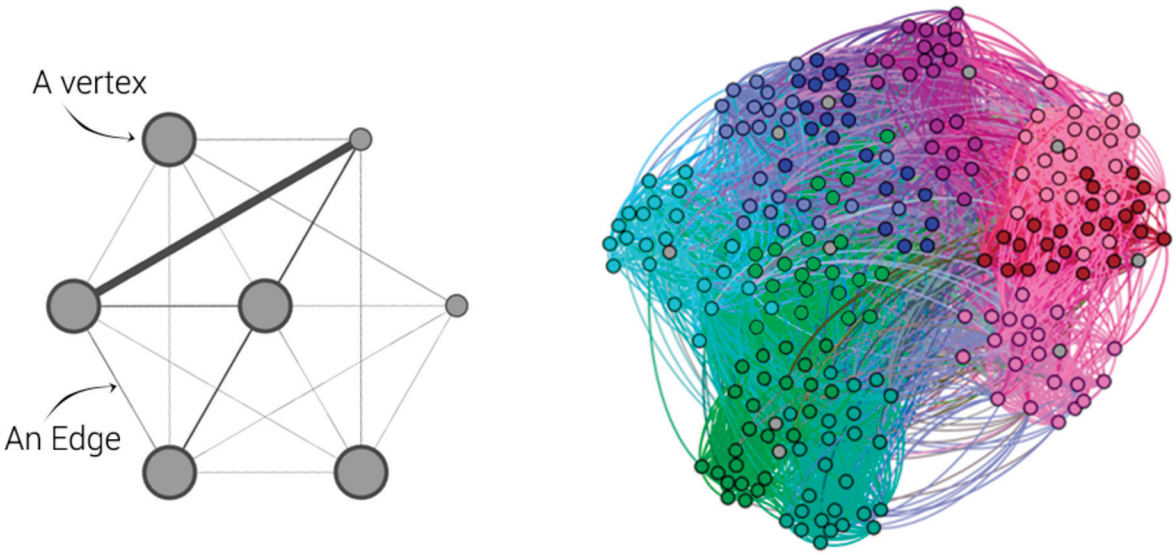
An example is the Girvan and Newman (2002) approach for community detection that focuses on *betweenness*, by removing edges with largest centrality (Freeman, 1977). Another example comprises the Modularity Optimization Methods (Newman, 2006) that uses node degree (how many connections relate to a vertex) as part of the procedure to detect communities.

K-means and its variants, on the other hand, is focused on vertex characteristics. It is more related to data-mining than to community detection, but still can be a powerful tool to group clusters of individuals with high similarity.

This article presents some of the main properties of social networks and complex networks, how the communities and clusters are characterized and the ways used to identify clusters in networks using the

DOI: 10.4018/978-1-4666-5888-2.ch344

Figure 1. (a) An example of a graph, (b) a complex graph



K-means algorithm and its main variants, the fuzzy *c*-means and the weighted *K*-means.

This article is intended for Social Networks analysts, students and researchers in the field of data mining, and for those seeking for agile methods of data arrays in complex networks. Although the K-means algorithm and its variants do not cover the completeness of community detection in complex networks it can be a powerful tool for discovering groups with high similarity. We also show an extended version that weights the data dimensions to be grouped.

BACKGROUND

It is possible to represent a variety of structures by means of complex networks and graphs. A graph is represented as a set of points (vertices) connected by links (edges). More formally, a graph G is an ordered pair of vertices $G = (V,E)$ where: V is a set of nodes (vertices) and E is a set of links (edges). An example of graph is shown in figure 1(a).

Graphs are an abstract mathematical representation of a network. Social networks follow the patterns of complex networks with similar properties. Evolving from purely mathematical models of graphs through the Random Graph of Erdos and Renyi (Erdos & Renyi, 1960) to The Small-World Model of Watts and Strogatz (Watts & Sstrogatz, 1998), complex network analysis have encompassed graph theory and gone so much further to represent real world networks.

A more complex example is shown in the network of figure 1(b). It is from a weighted network of face-to-face proximity between students and teachers. The dataset represents relations of children and teachers from the first to the fifth grade, and it is already grouped in ten clusters. Each cluster represents a particular class of students. This is a good example of how communities can be displayed by a network (Stehlé et al., 2011).

To analyze complex networks, it is mandatory some knowledge about the basic metrics and attributes regarding how the vertices are connected. One of those characteristics is the weight. When connections (edges) between nodes of a graph have weights, it is said a weighted graph. Such weights can represent the strength of a connection or its cost. For instance, to represent a network of cities, the weights could be the distances between them. Both vertices and edges can be weighted. There are also non weighted graphs, in which all connections or vertices receive the same unity value or cost.

Another important property of the connections is the direction; edges can be directed or undirected. In directed graphs, edges have a specific direction, and the relations between pairs of vertices are asymmetric.