N

Clique Size and Centrality Metrics for Analysis of Real-World Network Graphs

Natarajan Meghanathan
Jackson State University, USA

INTRODUCTION

Network Science is a fast-growing discipline in academics and industry. It is the science of analyzing and visualizing complex real-world networks using graph theoretic principles. Several metrics are used to analyze the characteristics of the real-world network graphs; among them "centrality" is a commonly used metric. The centrality of a node is a measure of the topological importance of the node with respect to the other nodes in the network (Newman, 2010). It is purely a link-statistics based measure and not based on any offline information (such as reputation of the node, cost of the node, etc). The commonly used centrality metrics are degree centrality, eigenvector centrality, closeness centrality and betweenness centrality. Degree centrality (DegC) of a node is simply the number of immediate neighbors for the node in the network. The eigenvector centrality (EVC) of a node is a measure of the degree of the node as well as the degree of its neighbor nodes. We refer to DegC and EVC as degree-based centrality metrics. Closeness centrality (ClC) of a node is the inverse of the sum of the shortest path distances of the node to every other node in the network. Betweenness centrality (BWC) of a node is the ratio of the number of shortest paths the node is part of for any source-destination node pair in the network, summed over all possible source-destination pairs that do not involve the particular node. We refer to ClC and BWC as shortest path-based centrality metrics. Computationally efficient polynomial-time algorithms have been proposed in the literature (Brandes, 2001; Strang, 2005; Cormen et. al., 2009; Newman, 2010) to determine exact values for each of the above centrality metrics; hence we categorize centrality as a computationally lightweight metric.

A "clique" is a complete sub graph of a graph (i.e., all the nodes that are part of the sub graph are directly connected to each other). Cliques are used as the basis to identify closely-knit communities in a network as part of studies on homophily and diffusion. Unfortunately, the problem of finding the maximum-sized clique in a graph is an NP-hard problem (Cormen et. al., 2009), prompting several exact algorithms and heuristics to be proposed in the literature (Pattabiraman et. al., 2013; Fortunato, 2010; Palla et. al., 2005; Sadi et. al., 2010; Tomita & Seki, 2003). In this chapter, we choose a recently proposed exact algorithm (Pattabiraman et. al., 2013) to determine the size of the maximum clique for large-scale complex network graphs and extend it to determine the size of the maximal clique that a particular node is part of. We define the maximal clique size for a node as the size of the largest clique (in terms of the number of constituent nodes) the node is part of. Note that the maximal clique for a node need not be the maximum clique for the entire network graph; but, the maximum clique for the entire graph could be the maximal clique for one or more nodes in the network.

Since the maximal clique size problem is a computationally hard problem and exact algorithms run significantly slower on large network graphs, our goal in this chapter is to explore whether the maximal clique size correlates well to one of the commonly studied computationally lightweight metrics, viz., centrality of the vertices, for complex real-world network graphs: if we observe a

DOI: 10.4018/978-1-5225-2255-3.ch565

high positive correlation between maximal clique size and one or more centrality metrics, we could then infer the ranking of the vertices based on the centrality values as the ranking of the vertices based on the maximal clique size in real-world network graphs. Ours will be the first chapter to conduct a correlation study between centrality and maximal clique size for real-world network graphs. To the best of our knowledge, we did not come across such a work that has done correlation study between these two metrics (and in general, a computationally hard metric vis-a-vis a computationally lightweight metric) for real-world network graphs. Throughout the chapter, we use the terms 'node' and 'vertex' as well as 'link' and 'edge' interchangeably. They mean the same.

Background

To the best of our knowledge, ours is the first work to focus on a correlation coefficient analysis between a computationally hard metric (maximal clique size for the individual vertices) with that of a computationally lightweight metric (centrality values of individual vertices) for complex real-world network graphs. The work available in the literature so far considers these two metrics separately. Recently, Li et al (2014) conducted a correlation coefficient analysis study among the centrality metrics for real-world network graphs. Centrality metrics have been widely studied for analysis and visualization of complex networks in several domains, ranging from biological networks to social networks (e.g., Koschutzki & Schreiber, 2008; Opsahl et. al., 2010). The research focus with regards to cliques in the context of complex networks is to come up with efficient heuristics to reduce the run-time complexity in determining the maximum size clique for the entire network graph. Though branch-and-bound has been the common theme among these works, the variation is in the approach used to arrive at the bounds and enforce them in the search space. Strategies used for pruning the search space are typically based on node degree (e.g., Pattabiraman et. al., 2013), vertex ordering (e.g., Carraghan & Pardalos, 1990) and vertex coloring (e.g., Ostergard, 2002). Recently, a parallelized approach (Rossi et. al., 2014) for branch and bound has also been proposed for determining cliques in real-world networks ranging from 1000 to 100 million nodes. Nevertheless, none of the research so far has focused on identifying correlation between the maximal clique size for an individual vertex (the size of the largest clique that a particular vertex is part of) with any of the commonly studied metrics (like centrality) for network analysis. Ours is the first step in this direction. With the problem of determining maximum size clique for the entire network graph and maximal size cliques for the individual vertices being NP-hard and computationally time-consuming for complex real-world networks of larger size, it becomes imperative to analyze the correlations of the maximal clique size values of the individual vertices to that of the network metrics that can be easily computed so that meaningful inferences about maximal clique size values can be made.

Centrality Metrics

This section discusses the four centrality metrics that are studied in this chapter. The highest ranked vertex or set of vertices with respect to each of the centrality metrics is shown shaded in the graphs of these figures. The degree centrality (DegC) of a vertex is the number of neighbors adjacent to it (example illustrated in Figure 1). Eigenvector centrality (EVC) of a vertex is a measure of the degree of the vertex as well as the degree of its neighbors. EVC of the vertices in a graph (example illustrated in Figure 2) are the entries in the principal eigenvector of the adjacency matrix for the graph. The larger the value of the entry for an vertex in the principal eigenvector, the higher is its ranking with respect to EVC. The principal eigenvector is determined by running the power-iteration algorithm (Strang, 2005) on the adjacency matrix of the network graph. The closeness centrality (ClC) of a vertex is the inverse