Social Networks 46 (2016) 1–10

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Detecting large cohesive subgroups with high clustering coefficients in social networks

Zeynep Ertem [a], Alexander Veremyev [b], Sergiy Butenko [a,*]

[a] Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX 77843-3131, United States
[b] Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611-6595, United States

ARTICLE INFO

Article history:
Available online 1 March 2016

Keywords:
Cohesive subgroups
Clustering coefficient
Clique relaxations
Optimization

ABSTRACT

Clique relaxations are used in classical models of cohesive subgroups in social network analysis. *Clustering coefficient* was introduced more recently as a structural feature characterizing small-world networks. Noting that cohesive subgroups tend to have high clustering coefficients, this paper introduces a new clique relaxation, *α-cluster*, defined by enforcing a lower bound α on the clustering coefficient in the corresponding induced subgraph. Two variations of the clustering coefficient are considered, namely, the local and global clustering coefficient. Certain structural properties of α-clusters are analyzed and mathematical optimization models for determining α-clusters of the largest size in a network are developed and validated using several real-life social networks. In addition, a network clustering algorithm based on local α-clusters is proposed and successfully tested.

1. Introduction

Social interactions have been becoming increasingly observable and measurable with advancements in technology. Interactions using online social media platforms, mobile phones and email are commonly represented as graphs, whose structural analysis may reveal interesting insights about the underlying social networks. For example, community detection techniques have been extensively used to identify groups of people characterized by a high level of interactions and to understand social communication throughout the network by finding *cohesive subgroups* (Scott, 2000; Falzon, 2000; Fortunato, 2010; Schaeffer, 2007).

A cohesive subgroup is a "tightly knit" subset of actors in a social network, which was originally modeled using the graph-theoretic concept of a clique (Luce and Perry, 1949). The notion of a clique embodies a perfect cohesive group, it compels every two individuals in the subgroup to be directly connected to each other. However, requiring every possible pairwise connection between the individuals in a subgroup is often overly restrictive from a practical perspective, as doing so yields only "perfectly cohesive" clusters while ignoring other important cohesive subgroups. In addition, clique-detection algorithms may fail to identify cliques in which a few edges are absent due to imprecisions in collecting the data. To overcome these issues, clique relaxation models have been introduced, including the k-clique (Luce, 1950), relaxing direct interaction between individuals; the k-club (Alba, 1973; Mokken, 1979), relaxing reachability; the k-plex (Seidman and Foster, 1978), allowing at most k non-neighbors; and s-defective clique (Yu et al., 2006), allowing at most s missing edges. Clique relaxation models have been extensively used in social network analysis (Pattillo et al., 2012; Nastos and Gao, 2013; Scott, 2000; Wasserman and Faust, 1994).

This paper proposes a novel clique relaxation based on the notion of *clustering coefficient*. This concept gained popularity in the study of the so-called *small-world networks* (Watts, 1999; Watts and Strogatz, 1998), where it is used to model the hypothesis that two people are more likely to be friends if they have a friend in common. For a given actor (node) with more than one friend (neighbor), its local clustering coefficient measures the local density of direct connections between its friends. Clustering coefficient is equal to one when the node's neighborhood is fully connected (forms a clique). On the other hand, a close to zero clustering coefficient means that there are hardly any connections in the neighborhood. Many real-life networks have been empirically found to have many nodes with rather high clustering coefficients (Newman, 2003), which also appears to be a natural property to expect of cohesive subgroups in social networks. In fact, according to (Jackson, 2008, p. 35), clustering coefficient is "the most common way of measuring some aspect of cliquishness."

* Corresponding author. Tel.: +1 979 458 2333; fax: +1 979 847 9005.
E-mail addresses: zeynepertem@tamu.edu (Z. Ertem), averemyev@ufl.edu (A. Veremyev), butenko@tamu.edu (S. Butenko).

http://dx.doi.org/10.1016/j.socnet.2016.01.001
0378-8733/

Hence, it is reasonable to define a cohesive subgroup by requiring that the corresponding subset of nodes induces a connected subgraph with a desired (high) clustering coefficient α. We will refer to such a structure as an *α-cluster*. If $\alpha = 1$, the connectivity requirement ensures that an α-cluster is a clique. Otherwise, if $\alpha < 1$, an α-cluster can be viewed as a clique relaxation.

Our study focuses on computing α-clusters of the *largest size* in social networks, which is of interest for several reasons. Larger cohesive subgroups tend to have more influence on the overall network structure than their smaller counterparts. In fact, the largest size of a cohesive subgroup of a certain kind can be thought of as a *global measure of cohesiveness* of the whole network with respect to the imposed definition of cohesiveness. The presence of large cohesive subgroups consisting of considerable portions of a network implies a high level of cohesion in the network, whereas their absence indicates the opposite. Nevertheless, smaller cohesive subgroups may also be of interest, and the approaches proposed in this work can be easily modified to compute all α-clusters of a given size by introducing the corresponding constraints in the considered optimization models.

Dropping the connectivity requirement from the α-cluster definition may result in a structure with multiple connected components, each of which is a connected α-cluster. This motivates a novel clustering algorithm, which uses the multiple connected α-clusters as the "seeds," with the remaining nodes assigned to these seed clusters using a certain strategy. The algorithm yields encouraging results on the social networks used in our experiments.

The remainder of this paper is organized as follows. In the next section, we introduce the necessary definitions and study some basic structural properties of α-clusters. Section 3 provides optimization models for finding the largest α-clusters in a network. In Section 4, the proposed models are used to analyze several well-known social networks. The proposed local α-clustering algorithm is outlined and tested in Section 5. The paper concludes with a summary of findings and suggestions for future research in Section 6.

2. Definitions and properties

This section presents basic graph-theoretic definitions and notations used throughout the paper. Let $G = (V, E)$ be a simple graph with set V of n vertices (nodes) and set E of edges (links), $E \subset \{\{i, j\} : i, j \in V\}$. Let $N_G(i)$ and $d_G(i)$ denote the *neighborhood* and *degree* of i in G, respectively, and let A_G be the *adjacency matrix* of G. The *distance* between vertices i and j in G is denoted $d_G(i, j)$; $d_G(i, j) = \infty$ if i and j are not connected. The diameter of G is denoted $diam(G)$. Given a subset $V' \subset V$, the corresponding induced subgraph $G[V']$ is defined as $G[V'] = (V', E')$, where E' is the subset of edges of G connecting pairs of vertices from V'.

Watts and Strogatz (1998) define the *local clustering coefficient* for a node of degree at least 2 as the proportion of links between the vertices within its neighborhood divided by the number of links that could possibly exist between the neighbors. The following notations will be used in the definitions below:

$$D_i = \binom{d_G(i)}{2} \quad \text{and} \quad \mathcal{D} = \sum_{i \in V} D_i.$$

Definition 1 *(Local clustering coefficient).* The local clustering coefficient C_i of node i of degree $d_G(i) \geq 2$ in G is given by

$$C_i = \frac{1}{D_i} \sum_{j,k \in N_G(i), j<k} a_{jk}. \tag{1}$$

The global clustering coefficient $\mathcal{C}$ of G can be thought of as the proportion of triangles among the triplets, where a triplet is defined as an ordered subset of three vertices that induces a subgraph with at least two edges. It can be expressed mathematically as follows.

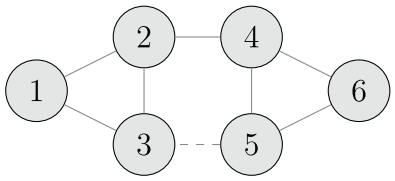

Fig. 1. Addition of an edge decreases global clustering coefficient.

Definition 2 *(Global clustering coefficient).* The global clustering coefficient $\mathcal{C}$ of graph G that has at least one connected component with more than 2 vertices is given by

$$\mathcal{C} = \frac{1}{\mathcal{D}} \sum_{i \in V} \sum_{j,k \in N_G(i), j<k} a_{jk}. \tag{2}$$

It is interesting to note that both local and global clustering coefficients of a graph can decrease with an increase in edge density. Indeed, Fig. 1 shows a graph where adding an edge between nodes 3 and 5 decreases the clustering coefficients. Before the dashed edge is added, local clustering coefficients of the corresponding nodes are $\{1, \frac{1}{3}, 1, \frac{1}{3}, 1, 1\}$, and the global clustering coefficient is $\mathcal{C} = \frac{6}{10}$. After the addition, local clustering coefficients change to $\{1, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 1\}$, and the global clustering coefficient value is $\frac{6}{14}$.

We define a *local α-cluster* as a subset of vertices that induces a subgraph in which each node's local clustering coefficient is at least α.

Definition 3 *(Local α-cluster).* Given a graph $G = (V, E)$, a subset of vertices $C \subseteq V$ is called a local α-cluster if $G[C]$ is connected and every node in C has the local clustering coefficient at least α in $G[C]$, that is,

$$\sum_{j,k \in N_{G[C]}(i), j<k} a_{jk} \geq \alpha \binom{d_{G[C]}(i)}{2} \quad \forall i \in C. \tag{3}$$

Note that the definition of local clustering coefficient implies that for an α-cluster C the degree of each node in $G[C]$ is at least 2. Also, it is easy to see that the edge density of the subgraph induced by any local α-cluster is at least α. However, the set of vertices inducing a subgraph with the edge density α may not be a local α-cluster.

Similarly, we can define a *global α-cluster* as follows.

Definition 4 *(Global α-cluster).* Given a graph $G = (V, E)$, a subset of at least three vertices $C \subseteq V$ is called a global α-cluster if $G[C]$ is connected and $G[C]$ has the global clustering coefficient at least α, that is,

$$\sum_{i \in C} \sum_{j,k \in N_{G[C]}(i), j<k} a_{jk} \geq \alpha \sum_{i \in C} \binom{d_{G[C]}(i)}{2}. \tag{4}$$

It is obvious that a local α-cluster is also a global α-cluster, whereas the converse does not hold in general. Hence, the definition of a local α-cluster guarantees stronger cohesiveness properties than those enforced by the definition of a global α-cluster. This is also evident from the experiments with real-life networks reported in Section 4, which led us to focus primarily on local α-clusters in this study.