Social Networks 44 (2016) 143–152

Contents lists available at ScienceDirect

Social Networks

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

Dense core model for cohesive subgraph discovery

Sadamori Koujaku*, Ichigaku Takigawa, Mineichi Kudo, Hideyuki Imai

Graduate School of Information Science and Technology, Hokkaido University, Japan

ARTICLE INFO

Keywords:
Graph analysis
Pseudo clique
Core/periphery structure

ABSTRACT

Discovery of cohesive subgraphs is an important issue in social network analysis. As representative cohesive subgraphs, pseudo cliques have been developed by relaxing the perfection of cliques. By enumerating pseudo clique subgraphs, we can find some structures of interest such as a star-like structure. However, a little more complicated structures such as a core/periphery structure is still hard to be found by them. Therefore, we propose a novel pseudo clique called ρ-dense core and show the connection with the other pseudo cliques. Moreover, we show that a set of ρ-dense core subgraphs gives an optimal solution in a graph partitioning problem. Several experiments on real-life networks demonstrated the effectiveness for cohesive subgraph discovery.

1. Introduction

In recent years, much interest has been shown in analysis of networks as a model of relationships between individuals, computers, Web pages, and proteins, and so on. A network/graph consists of nodes and edges in which a node represents an object and an edge between two nodes represents a relationship between the corresponding two objects. One aspect of interest in the analysis of networks is to know what kinds of and how many *cohesive subgraphs* exist in a given graph. Indeed, in many real networks, cohesive subgraphs are connected to special functions/roles, and their discovery is thus of great importance as follows.

(1) *Community Mining*: The goal of community mining is to identify clusters of persons and to understand how one person relates to another (Palla et al., 2005; Zhao et al., 2011). For example, the relations of co-authorships among scientists are modeled as a graph in which two coauthors are connected by an edge. A group of scientists sharing similar interests may publish many coauthored papers and, as a result, form a cohesive subgraph. Reversibly speaking, by identifying cohesive subgraphs, we may identify groups of collaborative scientists and understand how such a group influences other groups of scientists.
(2) *Identification of influential social actors*: Influential social actors can be found as cohesive subgraphs (Borgatti and Everett, 2000; Chakrabarti, 2004; Batagelj and Zaveršnik, 2010) in a graph of human interactions through emails, phone-calls, and social media. For example, bloggers customarily post their likes/dislikes on a Web page as a sign of agreement. Such a blogger network can be modeled as a graph in which a group of influential bloggers may form a cohesive subgraph.
(3) *Query Indexing*: Finding cohesive subgraphs in computer networks and wireless networks helps to find landmark nodes for calculating the approximated shortest distance between two nodes (Cohen et al., 2003; Jin et al., 2009). We may choose one cohesive subgraph to deploy a new landmark node on the network.

The cohesiveness of a graph is typically measured by one of three basic indices: (a) *distance* (number of edges of the shortest path between two nodes), (b) *degree* (numbers of edges a node has) and (c) *density* (ratio of actual number of edges to possible maximum number of edges). In any index, the perfect model of cohesive subgraphs is a *clique* (Luce and Perry, 1949), in which the distance between two nodes is always one (a), the degree is the node set size minus one for every node (b), and the density is the maximum of 100% (c). However, the perfection of cliques is too rigid to use for finding cohesive subgraphs in many real networks, and a number of *pseudo cliques*, therefore, have been developed by relaxing the strength of one of these indices. Such pseudo cliques include *n-clique* (Luce, 1950), *k-plex* (Seidman and Foster, 1978), *k-core* (Seidman, 1983) and ρ*-dense* (Goldberg, 1984), which were proposed in the early stage of research. These pseudo cliques often find a coarse subgraph and are too flexible to find cohesive subgraphs as will be explained later. In order to compromise between overly restrictive cliques and overly relaxed pseudo cliques, other

* Corresponding author. Tel.: +81 11 706 6854.
E-mail addresses: koujaku@main.ist.hokudai.ac.jp (S. Koujaku), takigawa@ist.hokudai.ac.jp (I. Takigawa), mine@main.ist.hokudai.ac.jp (M. Kudo), imai@ist.hokudai.ac.jp (H. Imai).

http://dx.doi.org/10.1016/j.socnet.2015.06.003
0378-8733/© 2015 Elsevier B.V. All rights reserved.

pseudo cliques have been proposed. An *n-clan* (Mokken, 1979) and an *n-club* (Mokken, 1979) were proposed as intermediate concepts between an *n*-clique and a clique. A *k-truss* (Saito and Yamada, 2006; Cohen, 2009; Wang et al., 2010) was proposed as a variant of *k*-core. These pseudo cliques are useful for finding cohesive subgraphs, in terms of one of three indices, such as a *star* (Harary, 1994) and a *rich club* (Zhou and Mondragon, 2004). However, there are still more cohesive subgraph of interest such as the core of *core/periphery structure* (Borgatti and Everett, 2000). The structure ideally consists of a clique (core) and peripheral nodes that connect to all core nodes. In practice, the ideal definition is relaxed such that the core is a sufficiently dense subgraph whose node are connected by a short path, and have large degree (Borgatti and Everett, 2000; Lee et al., 2014; Rombach et al., 2014). It is still difficult to find such a highly cohesive core with existing pseudo cliques.

In this paper, therefore, we propose one more pseudo clique called a *ρ-dense core* to find more kinds of structures of interest. A subgraph is called ρ-dense core if it cannot be divided into sparsely connected subgraphs; more specifically, no dichotomous division produces two subgraphs with between-density less than ρ. The ρ-dense core is cohesive in terms of all three indices and shares many desirable properties with other pseudo cliques. A ρ-dense core subgraph is a subgraph of equally connected nodes up to the specified density of ρ. At a small value of ρ, we can find a set of cores and peripheries, and at a large value of ρ, we can find the cores. We present two algorithms: one is an enumeration algorithm of all ρ-dense core subgraphs in a given graph and the other is a partitioning algorithm of the graph by ρ-dense core subgraphs. Although the enumeration problem is NP-hard in general, the algorithm terminates in a reasonable time even for problems of a medium size, graphs with 300 nodes and 1000 or fewer edges in the experiment.

In addition, we discuss a hierarchical application of these algorithms with increasing/decreasing the value of ρ.

The rest of paper is organized as follows. In the next section, we briefly review the related works and show what kinds of structures have been found so far. In Section 3, our problems are formally described with notations to be used in this paper. We propose ρ-dense core and analyze its properties in Section 4. In Section 5, we introduce two algorithms for enumerating ρ-dense cores and for graph partitioning. In Section 6, we illustrate the effectiveness of these algorithms on synthetic and real-life data sets. Discussion is presented in Section 7, and the conclusion follows in Section 8.

2. Pseudo cliques

Extraction of cohesive subgraphs can be achieved either implicitly or explicitly. In the implicit approach, cohesive subgraphs are found by dividing a graph into several subgraphs so as to maximize the density within each of the subgraphs and simultaneously to minimize the density between the subgraphs. There are various criteria for division including *ratio cut* (Hagen and Kahng, 1992), *normalized cut* (Malik, 2000), *modularity* (Newman and Girvan, 2004; Newman, 2006), and many others (Schaeffer, 2007; Luxburg, 2007). These criteria often produce "well-balanced" cohesive subgraphs in terms of density, number of nodes and/or number of edges. Other criteria such as *Extraction* (Zhao et al., 2011) and *GraphScan* (Wang et al., 2008) are also used to find the most cohesive subgraph in a given graph. They are often used for identification of influential actors and for detection of anomalies. The implicit approach is especially useful when we do not have in mind an explicit structure to be extracted. Unfortunately, many of the criteria suffer from a *resolution limit* (Fortunato and Barthélemy, 2007) that implies a tendency that the subgraphs to be found become larger as the entire graph becomes larger.

In the explicit approach, we define first what a desired graph is and then find subgraphs having such desired properties. Typically, this is achieved with an index showing how strongly a graph is cohesive. We enumerate (maximal) subgraphs that are sufficiently cohesive in the index. The three basic indices of cohesiveness are *distance*, *degree* and *density*: (a) *distance* is the number of edges of the shortest path between two nodes, (b) *degree* is the number of edges a node has, and (c) *density* is the ratio of the number of edges to the maximum possible number of edges. In general, a graph is more cohesive when every two nodes have a shorter distance, every node has a larger degree, and the graph itself is denser. The strongest model in all the three indices is a *clique* (Luce and Perry, 1949): the distance is one, the degree of each node is the node set size minus one, and the density is one. However, the concept of a clique is too strong to use for finding cohesive subgraphs in many real networks, and a number of pseudo cliques have therefore been developed by relaxing the perfection in these indices. An *n-clique* (Luce, 1950) is developed by relaxing the distance. An *n-clique* is a maximal subgraph in which the distance of two nodes is less than or equal to *n* (the path may step out the subgraph). A clique is a 1-clique since any two nodes are connected by an edge. If one edge is removed from a 1-clique, then it becomes a 2-clique. A *k-plex* (Seidman and Foster, 1978) and a *k-core* (Seidman, 1983) are obtained by relaxing the degree. A *k*-plex (Seidman and Foster, 1978) is a maximal subgraph in which a node is nonadjacent to at most *k* nodes in the subgraph. As a complement model of the *k*-plex, a *k-core* (Seidman, 1983) is a maximal subgraph in which each node is adjacent to at least *k* nodes in the subgraph. For example, a 4-node clique subgraph is a 3-core if no node of the clique is linked to three or more nodes outside the clique. By relaxing the density, *ρ-dense* (Abello et al., 2002) is developed. A connected subgraph is a ρ-dense if the density is greater than ρ. Hence, a 1-dense (rather 100%-dense) subgraph is a clique. In this paper, we focus on the implicit approach because our main concern is extraction of subgraphs with special structures of interest.

One problem of these pseudo cliques is that they are too relaxed from cliques so that they find subgraphs with excessively coarse structures. For example, being an *n*-clique ($n > 1$) does not mean that the subgraph is connected, because the definition allows that two separated components of the subgraph are connected through a single outside node. In order to compromise between overly restrictive cliques and overly relaxed pseudo cliques, many variants have been proposed. An *n-clan* (Mokken, 1979) and an *n-club* (Mokken, 1979) were proposed as intermediate concepts between an *n*-clique and a clique. They adopt another measure of distance, *diameter*, i.e., maximum distance over all pairs of nodes. An *n*-club is a maximal subgraph with diameter *n*. An *n*-clan is an *n-clique* (Luce, 1950) with diameter *n*, so that every *n*-clan is an *n*-club. The *n*-club and *n*-clan are always connected if $n \geq 1$. As a stronger variant of *k*-core, *k-truss* (Cohen, 2009) (also known as *k-dense* (Saito and Yamada, 2006), or *Dense Neighborhood graph* (Wang et al., 2010)) is introduced. A *k*-truss is a maximal subgraph in which any pair of nodes have at least $k - 2$ common neighbors (nodes linked to the node by an edge). For example, a 4-node clique subgraph is a 4-truss subgraph if it is maximal in the set of subgraphs sharing the same property. A remarkable nature of a *k*-truss is that enumeration of all *k*-truss subgraphs needs only a polynomial time in the number of nodes. Each of the previously proposed pseudo cliques is useful for finding a special type of cohesive subgraphs such as a *maximal clique*, a *star* (Harary, 1994), and a *rich club* (Zhou and Mondragon, 2004). For example, maximal clique subgraphs that have diameter one and density one can be found with distance-based pseudo cliques such as *n*-clan and density-based pseudo cliques such as ρ-dense. Similarly, a star consisting of a "hub node" and many surrounding nodes can be found with distance-based pseudo cliques such as *n*-clan. A rich club consisting of a "hub set" of nodes surrounded by many peripheral nodes can be found with degree-based pseudo cliques such as *k*-core. Finding of the hub set is carried out