Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Finding disjoint dense clubs in a social network

Peng Zou [a], Hui Li [b], Wencheng Wang [c], Chunlin Xin [d,*], Binhai Zhu [a,*]

[a] Gianforte School of Computing, Montana State University, Bozeman, MT 59717-3880, USA
[b] Department of Computer Science, Beijing University of Chemical Technology, Beijing, China
[c] State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, China
[d] School of Economics and Management, Beijing University of Chemical Technology, Beijing, China

ARTICLE INFO

Article history:
Received 6 August 2016
Received in revised form 16 July 2017
Accepted 20 October 2017
Available online xxxx

Keywords:
Social networks
Clusters
Clubs
FPT algorithms
Kernelization

ABSTRACT

In a social network, the trust among its members usually cannot be carried over many hops. So it is important to find disjoint clusters with a small diameter and with a decent size, formally called *dense clubs*. We focus on handling this NP-complete problem in this paper. First, from the parameterized computational complexity point of view, we show that this problem does not admit a polynomial kernel (implying that it is unlikely to apply some reduction rules to obtain a practically small problem size). Then, we focus on the dual version of the problem, i.e., deleting d vertices to obtain some isolated dense clubs. We show that this dual problem admits a simple FPT algorithm using a bounded search tree method (the running time is still too high for practical datasets). Finally, we combine a simple reduction rule together with two branching rules to obtain a practical solution (verified by extensive testing on practical datasets).

1. Introduction

Social network gives people a new "world" where we can share everything that happens around us and social networks have grown enormously in recent years. It is full of data and has become an indispensable part of our life. Finding cohesive subgroups is vital to understanding the structure of the network. Clique is commonly used to describe a dense subgroup. However, the requirements of a clique are too restrictive in many situations, thus various relaxed cohesive subgroup structures based on clique have been proposed, such as s-club, s-clique, and s-plex, etc. [1,2].

Given a social network, modelled as an undirected graph $G = (V, E)$, a clique is a subset of vertices such that any pair of vertices in this subset form an edge in E. In fact, the maximum clique problem is one of the most widely studied NP-complete problems [3]. Many algorithms for this problem are available in the literature [4–6]. Motivated by practical applications in social (and biological) networks, s-club is a diameter-based graph-theoretic generalization of clique, which was first introduced as an alternative approach to model a cohesive subgroup in the social network area [7,8]. An s-club is a subset of vertices $V' \subseteq V$, such that the diameter of the induced subgraph $G[V']$ is at most s.

With the development of social networks, the trust among its members has become a big issue. In a social network, the trust among its members usually cannot be carried over many hops. So it is important to find disjoint clusters with a small diameter and with a decent size, formally called *dense clubs* in this paper. Secondly, a complex social network is

* Corresponding authors.
E-mail addresses: peng.zou@msu.montana.edu (P. Zou), ray@mail.buct.edu.cn (H. Li), whn@ios.ac.cn (W. Wang), xinchl@buct.edu.cn (C. Xin), bhz@montana.edu (B. Zhu).

https://doi.org/10.1016/j.tcs.2017.10.018
0304-3975/© 2017 Published by Elsevier B.V.

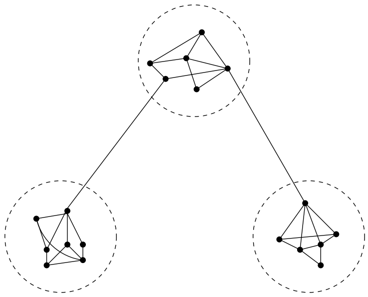

Fig. 1. A small example of the network structure. There are three groups denoted by the circle with many edges and only a small number of edges between the groups.

usually composed of several groups/communities, and this characterization of community structure means the appearance of densely connected groups of vertices, with only sparse connections between groups [9], see Fig. 1.

An s-clique is a subset of vertices $S \subseteq V$ if the shortest path distance $d_G(u, v) \le s$ for all $u, v \in S$. It is pretty obvious that an s-club is also an s-clique, but the converse is not true in general [10]. And for $s = 1$, an s-club is simply a clique. It is known that the maximum clique problem is a classical NP-complete problem [4,5], which is also hard to approximate [11]. The maximum s-club problem [12,13] is NP-complete for any fixed s, even when restricted to graphs of fixed diameter $s + 1$ [12]. In fact, testing whether an s-club is maximal is also NP-complete for any fixed integer s [14].

In reference [15], Bourjolly et al. proposed three heuristic methods which are DROP, CONDTELLATION and s-DLIQUE-DROP for the maximum s-club problem. A variable neighborhood search (VNS) meta-heuristic algorithm was proposed by Shahinpour et al. [10]. In fact, they used the VNS heuristic method as the lower bound to develop an exact algorithm for the maximum s-club problem [10]. Recently, a new heuristic algorithm called IDROP for the largest s-club problem was given in [16].

From the parameterized computational complexity point of view, two fixed-parameter tractable (FPT) algorithms for the maximum s-club problem were obtained [17]. Hartung et al. extended the previous parameterized complexity study for 2-club and provides polynomial-size kernels for 2-club parameterized by "cluster editing set size of G" and "size of a cluster editing set of G" [18]. For the 2-club-editing problem, Liu et al. proposed an improved search tree algorithm with running time $O^*(3.31^k)$ based on two new branching cases, improving the trivial $O^*(4^k)$ bound [21]. (Here $O^*()$ is the major component for measuring the running time of an FPT algorithm, see next section for more details.)

In this paper, we first study the disjoint dense club problem. Specifically, we show that the problem does not have a polynomial kernel (unless the polynomial hierarchy collapses to the third level). This implies that it is unlikely to obtain any efficient FPT algorithm for the problem (and the related ones). Then we consider the dual problem of editing a graph (by deleting vertices) into isolated s-clubs. Since the trivial bounded-degree search method takes $O^*((s+2)^d)$ time, which is not efficient for most real datasets, we propose three rules to design an efficient heuristic method. We then test this method with two real datasets and three social networks.

The rest of this paper is organized as follows. We will introduce some necessary definitions and notations in Section 2. In Section 3, theoretical results are reported. All computational results are shown in Section 4. We conclude the paper in Section 5.

2. Preliminaries

In this section, we present the relevant definitions and review some useful notions.

FPT algorithms and kernels FPT (Fixed-Parameter-Tractable) algorithms are used to study the computational complexity of NP-hard problems [22–24]. Beside the input size n, we also consider a parameter k (or several parameters). An FPT algorithm is one which solves a parameterized problem L in $O(f(k)n^c) = O^*(f(k))$ time (i.e., decide whether $(x, k) \in L$, where $n = |x|$), where $f(-)$ is any computable function and c is a constant not related to n and k.

A parameterized problem L admits a problem *kernel* if there is a polynomial-time transformation of any instance (I, k) to an instance (I', k') such that (1) $(I, k) \in L$ iff $(I', k') \in L$; (2) $|I'| \le g(k)$; and (3) $k' \le k$. L has a polynomial kernel if $g(k)$ is a polynomial function. It is known that L admits an FPT algorithm iff it has a kernel (not necessarily polynomial). But if L has a polynomial kernel then usually it means L is relatively easier to solve.

Polynomial parameter transformations A polynomial parameter reduction is used to reduce a problem known to be without a polynomial kernel to another problem B [20,19]. It is different from the traditional FPT-reductions.

Definition 1. Let P, Q be parameterized problem. P is polynomial parameter reducible to Q, written as $P \le_{pp} Q$, if there exists a polynomial time computable function $f : \Sigma^* \times N \to \Sigma^* \times N$ and a polynomial p, such that for all $(x, k) \in \Sigma^* \times N$,