ARTICLE IN PRESS

Discrete Applied Mathematics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Discrete Applied Mathematics

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

Note

Lower bounds on paraclique density

Ronald D. Hagan, Michael A. Langston *, Kai Wang

Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN, 37996, USA

ARTICLE INFO

Article history:
Received 12 August 2015
Accepted 12 November 2015
Available online xxxx

Keywords:
Clique
Paraclique
Graph density
Clustering

ABSTRACT

The scientific literature teems with clique-centric clustering strategies. In this paper we analyze one such method, the paraclique algorithm. Paraclique has found practical utility in a variety of application domains, and has been successfully employed to reduce the effects of noise. Nevertheless, its formal analysis and worst-case guarantees have remained elusive. We address this issue by deriving a series of lower bounds on paraclique densities.

1. Introduction

Clique-centric methods have long played an important role in data science and engineering. Classic techniques include algorithms for $\mathcal{NP}$-hard problems such as maximal clique [3] and maximum clique [2]. The availability of high-throughput data has prompted interest in noise-abatement relaxations, most notably k-clique communities [8] (more recently also called clique percolation) and paraclique [4]. These algorithms have been used for biological data clustering, and been found superior to traditional methods [7]. Although similar in objective, k-clique communities is hampered in practice by its bottom up approach relying on an exhaustive enumeration of maximal cliques. Paraclique, in contrast, applies top down design principles and employs maximum clique, for which there are highly efficient and reasonably scalable algorithms [9], plus viable alternatives based on duality and parameterized complexity [1].

Paraclique can be formulated in a variety of ways. The general idea is to expand a maximum clique by augmenting it with non-clique vertices adjacent to most, but not all, members of the clique. The motivation for deriving dense subgraphs in this fashion is based on the fact that so-called "missing" edges, while relevant, are often lost due to noise, improper thresholding, weak experimental design, and numerous other causes. A classic example of this phenomenon can be found in the use of DNA microarrays for transcriptomic data analysis. In this setting, vertices represent genes, edges signify co-expression, and paracliques denote molecular response networks differentially (in)activated by stimulus [4]. Depending on a variety of factors, most but not all network elements may be highly intercorrelated at any particular time.

Previous paraclique studies have focused mainly on practical results. Representative examples include [5,11,6]. Instead, our primary goal in this paper is to investigate paraclique's theoretical basis. In so doing, we seek to derive bounds on its worst-case behavior, applying density as the classic clustering metric (we compute a subgraph's density in the traditional way, as the number of edges present divided by the maximum number possible). In the original paraclique formulation, the total number of missing edges was left unchecked. Density could, therefore, in principle be driven to zero. By limiting paraclique size to at most twice the maximum clique size, however, and by requiring that a new non-clique vertex be adjacent to all but one vertex in the growing paraclique, it is known that density is maintained at no less than 50% [4]. Here, we greatly expand upon such density results.

* Corresponding author.
E-mail addresses: rhagan@tennessee.edu (R.D. Hagan), langston@tennessee.edu (M.A. Langston), kwang11@tennessee.edu (K. Wang).

http://dx.doi.org/10.1016/j.dam.2015.11.010
0166-218X/

Please cite this article in press as: R.D. Hagan, et al., Lower bounds on paraclique density, Discrete Applied Mathematics (2015), http://dx.doi.org/10.1016/j.dam.2015.11.010

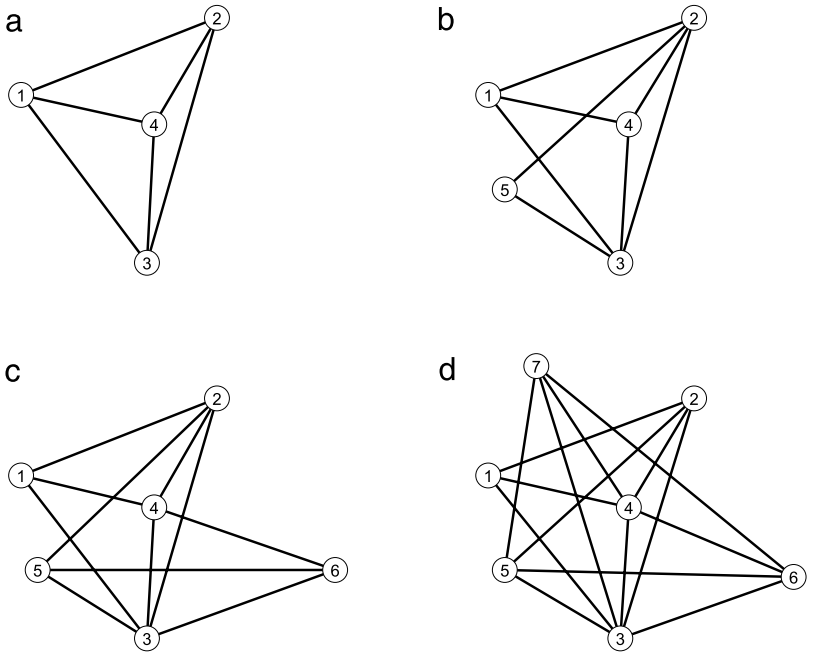

Fig. 1. An example paraclique grown with glom term $g = 2$ from an initial maximum clique of size four $\{1, 2, 3, 4\}$ (a) and successively adding vertices 5 (b), 6 (c) and 7 (d).

In the next section, we formalize definitions, describe relevant background, and establish several helpful preliminary results. In Sections 3 and 4, we derive bounds on general and special cases, respectively. In a final section we draw conclusions and discuss directions for future research.

2. Preliminaries

Let G denote a finite, simple, undirected graph. A clique is a subgraph of G in which every pair of vertices is connected by an edge. A paraclique, P, is constructed by first finding a clique C of maximum size, then glomming onto non-clique vertices in a controlled fashion. An integer glom term, g, is used to accomplish this. In the original algorithmic formulation [4], a non-clique vertex was chosen if and only if it was adjacent to g or more vertices in P. The number of required adjacencies does not scale with the size of P using this approach, however, so we generally invert this comparison. Thus, we glom onto a non-clique vertex if and only if it is adjacent to all but at most g vertices in P. In applications, g is usually some small value. In any case we insist that $0 < g < k$, where k denotes the number of vertices in C.

Pseudocode for the paraclique procedure is displayed in Algorithm 1. A sample paraclique construction is illustrated in Fig. 1. For the reader's convenience, definitions employed in the sequel are summarized in Table 1.

input : graph G, glom term g
output: paraclique P, a subgraph of G

$C \leftarrow$ maximum clique of G
$V \leftarrow$ vertex set of C
while $\overline{V}$ *contains a vertex* v *adjacent to all but at most* g *vertices in* V **do**
  $V \leftarrow V \cup \{v\}$
end
$P \leftarrow$ subgraph induced by V
return P

Algorithm 1: The Paraclique Algorithm

We start by establishing lower and upper bounds on maximum paraclique size.

Lemma 1. *A paraclique may contain as many as* $(g + 1)k$ *vertices.*

Proof. To construct a paraclique P that satisfies this bound, we begin with $g + 1$ disjoint cliques of size k, denoting them $C_0, C_1, \ldots, C_g$, and labeling the vertices of each C_i as $v_{ik}, v_{ik+1}, \ldots, v_{(i+1)k-1}$. To this we add edges connecting vertices v_r

Please cite this article in press as: R.D. Hagan, et al., Lower bounds on paraclique density, Discrete Applied Mathematics (2015), http://dx.doi.org/10.1016/j.dam.2015.11.010