Information Sciences 294 (2015) 1–14

Contents lists available at ScienceDirect

Information Sciences

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

Finding quasi core with simulated stacked neural networks

Malay Bhattacharyya [a], Sanghamitra Bandyopadhyay [b,*]

[a] Department of Information Technology, Indian Institute of Engineering Science and Technology, Shibpur, Howrah 711103, India
[b] Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India

ARTICLE INFO

Article history:
Received 13 March 2014
Received in revised form 10 September 2014
Accepted 17 September 2014
Available online 28 September 2014

Keywords:
Quasi core
Quasi-clique
Stacked neural network
Real-life network
Bioinformatics
Optimization

ABSTRACT

Studying networks is promising for diverse applications. We are often interested in exploring significant substructures in different types of real-life networks. Finding cliques, which denote a complete subgraph of a graph, is one such important problem in network analysis. Interestingly, many real-life networks often contain a significant number of almost (quasi) complete subgraphs, which are not entirely complete due to the presence of noise. Considering these networks as weighted adds further challenges to the problem. Finding quasi-complete subgraphs in weighted graphs has never been formally addressed. In this paper, we propose a stacked neural network model for finding out the largest quasi-complete module (core) in weighted graphs. We show the effectiveness of the proposed approach on DIMACS graphs. We also highlight its utility in analyzing scientific collaboration networks, social networks and biological networks.

1. Introduction

Finding dense substructures in a network is useful for various applications [31,47]. A dense substructure represents strong co-associativity between its members [4]. The denseness of such a substructure is generally quantified by its connectivity. Unfortunately, in many real-life networks it becomes hard to get well-connected substructures due the noise. Therefore, we often allow a relaxed form of connectivity to define denseness. Moreover, real-life networks are occasionally weighted and such relaxed form of connectivity has never been formally addressed in weighted graphs. In this paper, we address this problem, propose an approach for solving it and demonstrate its effectiveness on several real-life networks.

A graph $G = (V, E)$, with V and E denoting its vertices and edges, respectively, is complete if it has all-inclusive connectivity, i.e. $E = V \times V$. A clique is a complete subgraph of a graph [9,36]. Finding the maximum order clique in an unweighted graph is known to be an NP-hard problem [9]. An important study about a decade ago showed that the approximation of a maximal clique in polynomial time within a factor of $n^{1-\varepsilon}$ (for any $\varepsilon > 0$) is hard, unless NP = ZPP (where n is the number of vertices in the graph) [20]. ZPP stands for zero-error probabilistic polynomial time. The problems in ZPP can be exactly solved in expected polynomial time by a probabilistic algorithm. It is strongly believed that ZPP $\subset$ NP and the hypothesis NP $\neq$ ZPP is almost as strong as P $\neq$ NP [20], where P denotes the decision problems solvable in polynomial time using a deterministic Turing machine. For this reason, many of the recent algorithms to solve the maximum clique problem (MCP) are based on metaheuristics [17,42,46]. However, the initial attempts started with the development of pure heuristics [3,11,44].

* Corresponding author.
E-mail addresses: malaybhattacharyya@it.becs.ac.in (M. Bhattacharyya), sanghami@isical.ac.in (S. Bandyopadhyay).

http://dx.doi.org/10.1016/j.ins.2014.09.032
0020-0255/© 2014 Elsevier Inc. All rights reserved.

A clique symbolizes the maximum connectivity within a subgraph, and therefore, it is occasionally used to characterize a coherent module for many practical applications [7,33,37]. However, the stringent constraint of completeness within a subgraph is often too restrictive in some real-life scenarios [2,23,39]. Often the graph may be noisy or the problem may not demand complete connectivity in the subgraph. Quasi-cliques are often suitable descriptors of coherency in such graphs [8]. As we will be mostly dealing with noisy real-life networks in this paper, mining quasi-cliques is an important motivation for many applications.

A quasi-complete subgraph or quasi-clique is an almost clique in a graph (unweighted in general). The concept of quasi-cliques was first addressed by Abello et al. as a relaxation of cliques in massive graphs [1]. The decision and enumeration versions of the quasi-clique finding problem are known to be hard [39]. In a quasi-clique, every vertex satisfies a minimum fraction of connectivity with respect to the other vertices in it. Formally, for an n-vertex γ-quasi-clique, every vertex should have degree at least $\gamma(n-1)$. However, for an n-vertex subgraph with m edges there are some other terminologies like 'γ-clique', where $m \geqslant \frac{\gamma n(n-1)}{2}$ [2] or 'γ-near-clique', where $m \geqslant \frac{(1-\gamma)n(n-1)}{2}$ [10]. But these are different by definition and map to dissimilar decision problems. These are relaxations on the overall number of edges in a subgraph whereas we are trying to relax the degree of individual vertices here. Quasi-cliques are generalization of the concept of a clique, and therefore, the problem of finding the maximum quasi-complete module is computationally harder than the maximum clique problem. The decision version of the maximum quasi-clique problem is known to be NP-hard [23].

The search for a strong module in a biological graph is of interest in diverse applications [4]. Recently an algorithm has been proposed to find out almost cliques explicitly in scale-free graphs [4] and it is also useful in finding patterns for various bioinformatics [39] and chemoinformatics applications [27]. But the limited earlier approaches are focused over unweighted graphs. In real-life networks, relationships between the vertices are not always binary, rather they may be weighted or probabilistic [28]. Many of the biological networks have weighted edges, indicating the strength of association between the corresponding vertices (biomolecules). Here, we formalize the concept of quasi-completeness on weighted graphs. We aim to find out the maximum order quasi-complete module in a weighted graph. Notably, the clique finding algorithms would have found core subgraphs in real-life networks also, but not the largest ones because they are not fully complete rather quasi-complete.

We propose a stacked neural network model for finding out the core quasi-complete substructure in weighted graphs. We demonstrate the usefulness of the proposed approach on DIMACS graphs, scientific collaboration networks, social networks and biological networks. The rest of this paper is organized as follows: the motivating applications are described in Sections 2 and 3 introduces the preliminaries and basic definitions, Section 4 describes the state-of-the-art, Sections 5 and 6 detail on the proposed approach, Section 7 includes the experimental analyses, and finally Section 8 concludes the paper.

2. Motivations

Before going into the main problem of interest and providing a solution to address that, we discuss why the problem of finding quasi-cliques is significant. As the paper introduces both a novel problem and an algorithm to solve it, it is necessary to have motivation for both the parts. Quasi-cliques are nothing but generalizations of the cliques. The problems associated to cliques are sometimes strict and not flexible. However, quasi-cliques become interesting than the classical cliques because it can relax some constraints, which becomes particularly useful when the data is noisy or we want to explore additional information. Some of such applications and the motivation behind considering the neural network implementation to address the problem is described below.

- **Problem motivation:** Cliques are often used to model dense portions in a real-life network. While finding cliques, which comply with a very strict constraint of exhaustive connectivity, we might loose information about other dense portions. This might happen due to the presence of noise in the networks. On the other hand, we might be interested in finding relaxed form of cliques where the density can be treated as a tunable parameter. We foresee a number of applications based on this in various emerging areas as follows.
  1. Professional network analysis – Professional networks reflect the relations between professionals. Finding quasi-cliques in such networks might give us information about large and strong association of professionals that would be otherwise impossible to track with conventional form of cliques. As an example, we study a scientific collaboration network in this paper.
  2. Social network analysis – Social communication at the age of Internet has enabled the modeling of interactions between numerous people (e.g., students, bloggers, etc.). Analyzing completeness in these networks is interesting to find out core portions of these networks. The formation of large cliques is rare in such large-scale networks. So, finding quasi-cliques might help us to find dense yet large groups in such networks. In this paper, we analyze a Facebook network to establish the utility of finding quasi-cliques.
  3. Analysis of biological networks – Biological systems are often modeled as a network of biomolecules for understanding their cooperative activity. Finding cliques in these networks is important to find out coherent functional modules. Again, it has important applications in chemoinformatics like studying chemical structures and active sites [27]. However, for many biological systems, the entire environment is hardly known. In these cases, finding quasi-cliques