Efficient Attribute (α, β)-Core Detection in Large Bipartite Graphs (Student Abstract)

Yanping Wu, Renjie Sun, Chen Chen, Xiaoyang Wang[*]

School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou, China
{yanpingw.zjgsu, renjiesun.zjgsu}@gmail.com, {chenc, xiaoyangw}@zjgsu.edu.cn

Abstract

In this paper, we propose a novel problem, named rational (α, β)-core detection in attribute bipartite graphs (RCD-ABG), which retrieves the connected (α, β)-core with the largest rational score. A basic greedy framework with an optimized strategy is developed and extensive experiments are conducted to evaluate the performance of the techniques.

Introduction

The bipartite graph, composed of two disjoint vertex sets and edges connecting vertices from different sets, has numerous applications like fraudsters detection (Chen et al. 2021) and personalized recommendation (Zhu et al. 2020). As a fundamental problem investigated in bipartite graph analysis, community detection (CD) aims to find all or top-k communities by identifying specific models, e.g., (α, β)-core (Chen et al. 2021) and biclique (Lyu et al. 2020). In reality, the relationships between different entities often have properties, which can be modeled as attribute bipartite graphs.

Example 1. *Figure 1 shows a user-movie network, where each vertex in U (resp. V) denotes a user (resp. movie) and each edge associated with a number indicates that the user has a rating for a movie. Notice, the scoring mechanism adopts a five-point system, so there are five scores. Suppose α=2 and β=2 here, then the subgraph induced by set $\{u_2, \ldots, u_8, v_2, \ldots, v_8\}$ is a (2, 2)-core, where each vertex $u \in U$ (resp. $v \in V$) has no less than α=2 (resp. β=2) neighbors. For a movie discussion group, it will have a more harmonious atmosphere if users have a high consistency of preference, i.e., rating the same score for the same movie. Besides, dense groups are more conducive to frequent communication. However, in this $(2, 2)$-core, many users have distinct scoring schemes for the same movie (e.g., u_6, u_7 and u_8 gave three different scores to v_7) and the community size is too large to facilitate communication between users.*

Motivated by these, in this paper, we present a novel problem, namely *Rational (α, β)-Core Detection on Attribute Bipartite Graphs* (RCD-ABG), which retrieves the rational community with cohesiveness (i.e., (α, β)-core) and rationality (i.e., rational score function combining the preference consistency and the community density).

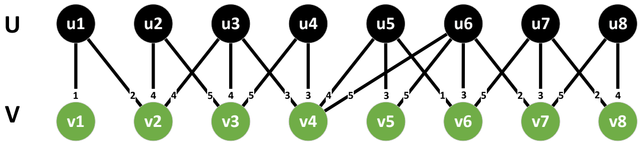


Figure 1: User-movie network with five-point score system

Preliminaries

We consider an attribute bipartite graph $G = (U, L, E, \mathcal{A})$ as an undirected graph without multiple edges and self-loops, where U (resp. L) is the vertex set in the upper (resp. lower) layer, $U \cap L = \emptyset$, E is the edge set, $E \subseteq U \times L$, and $\mathcal{A} = \{a_1, a_2, ..., a_t\}$ is the attribute set. Each edge $e \in E$ is associated with an attribute $a(e) \in \mathcal{A}$. We use m to denote the number of edges in G. Given an attribute bipartite graph G, a subgraph $S = (U_S, L_S, E_S, \mathcal{A})$ is an induced subgraph of G, if $U_S \subseteq U$, $L_S \subseteq L$ and $E_S = E \cap (U_S \times L_S)$. For a vertex $u \in U_S \cup V_S$, the number of u's neighbors is denoted by $d_S(u)$, i.e., the number of the adjacent vertices of u.

Definition 1 ((α, β)-core)**.** *Given a bipartite graph G, a subgraph S is the (α, β)-core of G, denoted by $C_{\alpha,\beta}$, if it satisfies: $1)$ degree constraint, i.e., $d_S(u) \geq \alpha$ for each vertex $u \in U_S$ and $d_S(v) \geq \beta$ for each vertex $v \in L_S$; $2)$ S is maximal, i.e., any supergraph $S' \supset S$ is not a (α, β)-core.*

In the following, we first introduce the consensus score of vertex and community. Then we formally define the rational score function. Note that, we only consider the consensus score of the vertex in the lower layer (e.g., movie layer).

Definition 2 (Consensus score)**.** *Given an attribute bipartite graph G, the consensus score of each vertex $v \in L$, is denoted by $\frac{x_G(v)}{d_G(v)}$, where $x_G(v)$ is the maximum number of its adjacent edges in G with the same attribute. For a subgraph S of G, its consensus score is defined as $(\sum_{v \in L_S} \frac{x_S(v)}{d_S(v)})/|L_S|$, where $|L_S|$ is the number of vertices in lower layer of S.*

To judge a community, our *rational score function* combines the consensus score and the density constraint, which

[*] Corresponding author

is $f(S) = \lambda \frac{\sum_{v \in L_S} \frac{x_S(v)}{d_S(v)}}{|L_S|} + (1-\lambda)\frac{|E_S|}{|U_S||L_S|}$, where λ is a parameter to make the trade-off between the consensus score and the community size. Based on this rational score function, we give the definition of our problem.

Problem Statement. Given an attribute bipartite graph G and two positive integers α and β, we aim to develop efficient algorithms to find the rational (α, β)-core, which is the subgraph S of G meeting the following three criteria: **i) Connectivity**: S is connected; **ii) Cohesiveness**: S is a (α, β)-core; **iii) Rationality**: S has the largest rational score $f(S)$ among subgraphs satisfying the above criteria.

Solution

Intuitively, to find the rational (α, β)-core, we can iteratively delete the vertex in the lower layer whose deletion will increase the score directly. Based on this, we define the rational marginal score of each vertex and propose our basic greedy framework as follows.

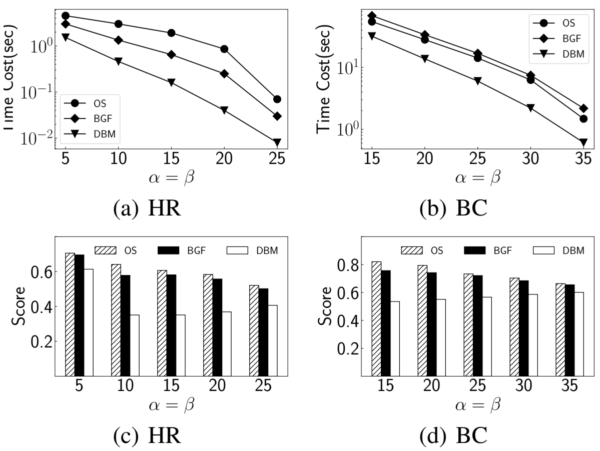
Definition 3 (Rational Marginal Score)**.** *Given an attribute bipartite graph G and a vertex $u \in L$, the rational marginal score is defined as $\triangle_G(u) = f(G \backslash \{u\}) - f(G)$*

Basic Greedy Framework (BGF). The details of BGF are illustrated as three main steps. $Step$ 1, we obtain the (α, β)-core of G and store into $\mathcal{G}$. Note that, in the following steps, we process each connected (α, β)-core of $\mathcal{G}$, iteratively. We use $\mathcal{G}_i$ to denote the current processing (α, β)-core and calculate its corresponding rational score. $Step$ 2, we greedily peel the vertex v in graph $\mathcal{G}_i$ providing the largest marginal score, i.e., $v = \arg\max_{u \in \mathcal{G}_i} f(\mathcal{G}_i \backslash \{u\}) - f(\mathcal{G}_i)$. After removing this vertex, we obtain the (α, β)-core in the remaining graph, and we store all the connected (α, β)-core into $\mathcal{G}$, separately. We continue this process until $\mathcal{G} = \emptyset$. $Step$ 3, the connected (α, β)-core with the largest score is the result. The time complexity of BGF is $O(2m + \min(\frac{|U|}{\beta}, \frac{|L|}{\alpha})|U|m)$.

An Optimized Strategy (OS). The basic greedy framework is simple but may severely limit the effectiveness of the algorithm. Removing a vertex may lead other vertices drop from the community, which are not considered in the marginal score. Hence, it may cannot reflect the real change of score especially for abundant removed vertices. To optimize it, we modify the rational marginal score by removing a vertex and its neighbors from two layers. Specifically, the marginal score is adapted as $\triangle'_G(u) = f(G \backslash (H2_G(u) \cup \{u\})) - f(G)$, where $H2_G(u)$ is the 2-hop neighbors of u in G that will be removed due to the deletion of u. The time complexity of OS is $O(2m + \min(\frac{|U|}{\beta}, \frac{|L|}{\alpha})(|U| + |L|)d_{max}m)$, where d_{max} is the maximum degree of vertex in G.

Experiments

To our best knowledge, there is no existing work for RCD-ABG. We implement three algorithms: **DBM**: the degree-based method choosing the vertex with the largest degree in each iteration, **BGF**: the basic greed framework, **OS**: BGF modified by optimized strategy. We employ 2 real-world bipartite graphs, i.e., HetRec (HR) and BookCrossing (BC), whose details can be referred to Grouplens (https://grouplens.org/datasets/). λ is set as 0.7 because the density of community will strengthen with the continuous deletion of vertices, thus we focus on consensus score.



(a) HR

(b) BC

(c) HR

(d) BC

Figure 2: Efficiency evaluation by varying α and β

To evaluate the efficiency, we report the response time of algorithms by varying α and β in Figures 2(a)-2(b). As observed, OS entails the similar time cost as BGF although OS needs to consider 2-hop neighbors. When α and β increase, the response time decreases for all methods since the community size decreases. To evaluate the effectiveness, we report the rational scores of returned communities in Figures 2(c)-2(d). OS and BFG can find the communities with higher scores than DBM. The score returned by OS is 0.08 higher than the one by BGF in HR. Note that, the consensus score is a fraction no more than 1, so the improvement of OS is already significant for the overall score. The score decreases when α and β increase because of tighter support constraint.

Future Work

In the future, we plan to investigate the multiple properties in our problem. We want to assess the performance of our algorithm from the experimental point of view and to propose some heuristics with the aim of improving the efficiency of the algorithm. We will conduct more experiments over lager datasets and different parameters.

Acknowledgements

This work is supported by NSFC 61802345, ZJNSF LQ20F020007, ZJNSF LY21F020012.

References

Chen, C.; Zhu, Q.; Wu, Y.; Sun, R.; Wang, X.; and Liu, X. 2021. Efficient critical relationships identification in bipartite networks. *World Wide Web Journal*.

Lyu, B.; Qin, L.; Lin, X.; Zhang, Y.; Qian, Z.; and Zhou, J. 2020. Maximum Biclique Search at Billion Scale. *VLDB*.

Zhu, Q.; Zheng, J.; Yang, H.; Chen, C.; Wang, X.; and Zhang, Y. 2020. Hurricane in Bipartite Graphs: The Lethal Nodes of Butterflies. In *SSDBM*.