Finding Tiny Clusters in Bipartite Graphs (Extended Abstract)

Presentation of work originally published in the Proceedings of the Thirty-second Annual Conference on Neural Information Processing Systems (NeurIPS 2018) under the title *Bipartite Stochastic Block Models With Tiny Clusters* [Ne18]

Stefan Neumann [1]

Abstract: We study the problem of finding clusters in random bipartite graphs. Applications of this problem include online shops in which one wants to find customers who purchase similar products and groups of products which are frequently bought together. We present a simple two-step algorithm which provably finds *tiny* clusters of size $O(n^{\varepsilon})$, where n is the number of vertices in the graph and $\varepsilon > 0$; previous algorithms were only able to identify medium-sized clusters consisting of at least $\Omega(\sqrt{n})$ vertices. We practically evaluate the algorithm on synthetic and on real-world data; the experiments show that the algorithm can find extremely small clusters even when the graphs are very sparse and the data contains a lot of noise.

Keywords: Biclustering; Bipartite Graphs; Random Graphs; Stochastic Block Models

1 Introduction

Finding clusters in bipartite graphs is a fundamental problem and has many applications. In practice, the two sides of the bipartite graph usually correspond to objects from different domains and an edge corresponds to an interaction between the objects. For example, in an online shop setting, the bipartite graph consists of customers (left side of the graph) and products (right side of the graph). An edge indicates that a customer bought a certain product. In this scenario, *customer clusters* consist of customers buying similar items and *product clusters* consist of products which are frequently bought together. Other application domains include paleontology [Fo03], where one wants to find co-occurrences of localities and mammals, and bioinformatics [Er13], where one wants to relate biological samples and gene expression levels.

Note that in many practical scenarios it is important that one can find *tiny* clusters. For example, nowadays online shops sell millions of products, but most product clusters are

[1] Universität Wien, Fakultät für Informatik, Wien, Österreich. stefan.neumann@univie.ac.at. The author gratefully acknowledges the financial support from the Doctoral Programme "Vienna Graduate School on Computational Optimization" which is funded by the Austrian Science Fund (FWF, project no. W1260-N35).

doi:10.18420/inf2019_30

very small compared to the total number of products on sale (e.g., the *Harry Potter* books are often bought together but they are only seven out of more than one million products). Hence, if a clustering algorithm can only detect medium-sized clusters consisting of at least a thousand products, then it is not applicable in this setting.

In this paper, we study this problem under a standard random graph model with a set of planted ground-truth clusters and we propose an algorithm which *provably* recovers extremely small clusters. More formally, we show that the algorithm allows to recover even tiny planted clusters of size $O(n^\varepsilon)$, where n is the number of vertices on the right side of the graph and $\varepsilon > 0$. Previous methods [Xu14, LCX15] could only discover clusters of size $\Omega(\sqrt{n})$. For the formal statement of the random graph model and the results, see [Ne18].

The algorithm consists of a simple two-step procedure: (1) Cluster the vertices on the left side of the graph based on the similarity of their neighborhoods. (2) Infer the right-side clusters based on the previously discovered left clusters using degree-thresholding.

We also implement the algorithm and evaluate it on synthetic and on real-world data. We verify that, in practice, the algorithm can find the small clusters which the theoretical analysis promised. We answer this question affirmatively. On synthetic data, the experiments show that, indeed, the algorithm finds tiny clusters even in the presence of high destructive noise (i.e., when the graphs are sparse and there are few inter-cluster edges). On real-world datasets, the algorithm finds clusters which are interesting and which have natural interpretations; for example, on a dataset consisting of users and books they rated [Zi05], the algorithm finds (among others) one cluster consisting of the *Harry Potter* books by J. K. Rowling and another cluster consisting of books written by John Grisham.

Bibliography

[Er13] Eren, Kemal; Deveci, Mehmet; Küçüktunç, Onur; Çatalyürek, Ümit V.: A comparative analysis of biclustering algorithms for gene expression data. Briefings in Bioinformatics, 14(3):279–292, 2013.

[Fo03] Fortelius, M. (coordinator): , New and Old Worlds Database of Fossil Mammals (NOW). Online. `http://www.helsinki.fi/science/now/`, 2003. Accessed: 2015-09-23.

[LCX15] Lim, Shiau Hong; Chen, Yudong; Xu, Huan: A Convex Optimization Framework for Bi-Clustering. In: ICML. pp. 1679–1688, 2015.

[Ne18] Neumann, Stefan: Bipartite Stochastic Block Models with Tiny Clusters. In: Thirty-second Conference on Neural Information Processing Systems, NeurIPS 2018. pp. 3871–3881, 2018.

[Xu14] Xu, Jiaming; Wu, Rui; Zhu, Kai; Hajek, Bruce E.; Srikant, R.; Ying, Lei: Jointly clustering rows and columns of binary matrices: algorithms and trade-offs. In: SIGMETRICS. pp. 29–41, 2014.

[Zi05] Ziegler, Cai-Nicolas; McNee, Sean M.; Konstan, Joseph A.; Lausen, Georg: Improving recommendation lists through topic diversification. In: WWW. pp. 22–32, 2005.