CLUSTERING IN GRAPHS AND HYPERGRAPHS WITH CATEGORICAL EDGE LABELS

Ilya Amburg, Nate Veldt, and Austin R. Benson

SIAM Workshop on Network Science 2020
July 9–10 · Toronto

Summary

Modern network datasets often contain rich structure that goes beyond simple pairwise connections between nodes, such as multiple interaction types of "higher-order interactions" involving more than two nodes at a time. However, developing rigorous methods for analyzing data with such richer models is a chalenge. Here, we develop a computational framework for clustering hypergraphs with categorical edge labels (i.e., different interaction types), where clusters corresponds to groups of nodes that frequently participate in the same type of interaction, and develop algorithms with strong theoretical guarantees.

Background and problem setup

The simple network model of nodes and edges is a powerful and flexible abstraction. Over time, more expressive models have been developed to incorporate richer structure found in data. In one direction, models have more information on the nodes and edges; for example, multilayer networks capture nodes and edges of different types [5]. In another direction, higher-order or multi-way interactions between several nodes — as opposed to pairwise interactions — are paramount to the model [2]. Designing methods that effectively analyze the richer structure encoded by these expressive models is an ongoing challenge.

In this work, we focus on the fundamental problem of clustering, where the general idea is to group nodes based on some similarity score. While graph clustering methods have a long history [4], existing approaches for rich graph data do not naturally handle networks with categorical edge labels. In these settings, a categorical edge label encodes a type of discrete similarity score — two nodes connected by an edge with category label c are similar with respect to c. This structure arises in a variety of settings: brain regions are connected by different types of connectivity patterns, edges in coauthorship networks are categorized by publication venues, and copurchasing data can contain information about the type of shopping trip. In the examples of coauthorship and copurchasing, the interactions are also higher-order — publications can involve multiple authors and purchases can be made up of several items.

Here, we develop a scalable clustering framework for edge-labeled hypergraphs. Given a network with k edge labels, we create k clusters of nodes, each corresponding to one of the labels. As an objective function for cluster quality, we seek to simultaneously minimize the number of edges that cross cluster boundaries and the number of intra-cluster "mistakes", where an edge of one category is placed inside the cluster corresponding to another category. Our methodology is based on a combinatorial objective function related to correlation clustering on graphs but enables the design of much more efficient algorithms that also seamlessly generalize to hypergraphs.

Notation. Let $G = (V, E, C, \ell)$ be an edge-labeled (hyper)graph, where V is a set of nodes, E is a set of (hyper)edges, C is a set of categories (or colors), and $\ell\colon E \to C$ is a function which labels every edge with a category. We use $k = |C|$ to denote the number of categories, $E_c \subseteq E$ for the set of edges having label c, and r for the maximum hyperedge size (i.e., *order*), where the size of a hyperedge is the number of nodes it contains.

Categorical edge clustering objective. Given G, we consider the task of assigning a category (color) to each node in such a way that nodes in category c tend to participate in edges with label c, i.e., we seek to partition the nodes into k clusters with one category per cluster. We encode the objective function as minimizing the number of "mistakes", where a mistake is an edge that either (i) contains nodes assigned to different clusters or (ii) is placed in a cluster corresponding to a category which is not the same as its label. This objective function is related to chromatic correlation clustering [3], which carries an additional penalty for pairs of *unconnected* nodes in the same category. Our adjustment leads to several favorable differences in computational tractability.

Formally, let Y be a categorical clustering, or equivalently, a coloring of the nodes, where $Y[i]$ denotes the color of node i. We define $m_Y\colon E \to \{0, 1\}$ as a *category-*

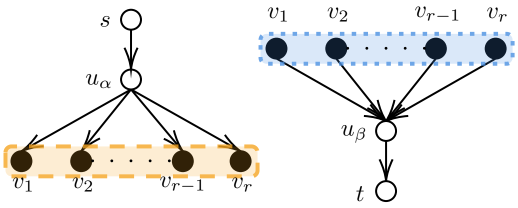

Figure 1: Subgraphs used for the s-t cut reduction of two-color Categorical Edge Clustering in hypergraphs. Here, α and β are hyperedges in the original hypergraph with colors c_1 (orange, left) and c_2 (blue, right).

mistake function, defined for a (hyper)edge $e \in E$ by

$$m_Y(e) = \begin{cases} 1 & \text{if } Y[i] \neq \ell(e) \text{ for any node } i \in e, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Then, the *Categorical Edge Label Clustering* objective score for the clustering Y is simply the number of mistakes:

$$\textbf{CatEdgeClus}(Y) = \sum_{e \in E} m_Y(e). \quad (2)$$

Overview of theoretical results

When there are only two categories, we can solve the Categorical Edge Clustering problem exactly in polynomial time through a reduction to an s-t cut problem on a modified graph $G' = (V', E')$. The reduction is quite simple: we add terminal nodes $s = v_{c_1}$ and $t = v_{c_2}$ (corresponding to categories c_1 and c_2) as well as all nodes in V to V'. For each hyperedge $e = (v_1, \ldots, v_r)$ of G, we add a node u_e to V' and add *directed* edges to E' (see also Fig. 1): if e has label c_1, add (s, u_e), $(u_e, v_1), \ldots, (u_e, v_r)$ to E'; otherwise, e has label c_2, and add (u_e, t), $(v_1, u_e), \ldots, (v_r, u_e)$ to E'. The minimum s-t cut on G' produces a partition that also minimizes the categorical edge clustering objective.

We establish that Categorical Edge Clustering is NP-hard in the case of more than two categories by a reduction from the maxcut problem. We then present several approximation algorithms with nice approximation guarantees. The first set of algorithms are based on practical linear programming relaxations, achieving an approximation ratio of $\min(2 - 1/k, 2 - 1/(r+1))$. The second approach uses a reduction to multiway cut, where practical algorithms have a $(r+1)/2$ approximation ratio and algorithms of theoretical interest have a $2(1 - 1/k)$ approximation ratio. A final approach optimally solves a modified objective, which runs in linear time and yields an r-approximation.

Table 1: Performance of our linear programming relax-and-round algorithm (LP) compared against baselines of Majority Vote and the *ChromaticBalls* (CB) and *LazyChromaticBalls* (LCB) from chromatic correlation clustering [3]. Performance is listed in terms of the approximation guarantee given by the LP lower bound (lower is better).

					Approx. Guarantee			
Dataset	$\lvert V\rvert$	$\lvert E\rvert$	r	k	LP	MV	CB	LCB
Brain	638	21180	2	2	1.0	1.01	1.56	1.41
MAG-10	80198	51889	25	10	1.0	1.18	1.44	1.35
Cooking	6714	39774	65	20	1.0	1.21	1.23	1.24
DAWN	2109	87104	22	10	1.0	1.09	1.31	1.15
Walmart-Trips	88837	65898	25	44	1.0	1.2	1.26	1.26

Our simple LP algorithm performs well in practice and we use the lower bound provided by the relaxation as a proxy for the performance of Categorical Edge Clustering.

Overview of experimental results

In this brief abstract, we show that our algorithms indeed works well on a broad range of datasets at optimizing our objective function and discover that our LP relaxation tends be extremely effective in practice, often finding an optimal solution (Table 1). Other results demonstrating the efficacy of our method in temporal community detection and data mining appear in our paper [1].

Extensions to fairness problems

We have also developed a regularized Categorical Edge Clustering objective that penalizes overrepresentation of colors within any cluster and have used it to explore both fairness from the concept of disparate impact as well as to solve a novel problem of forming teams based on past experience. We will present preliminary results on these findings as well.

References

[1] I. Amburg, N. Veldt, and A. R. Benson. Clustering in graphs and hypergraphs with categorical edge labels. *arXiv:1708.08436*, 2019.

[2] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg. Simplicial closure and higher-order link prediction. *PNAS*, 2018.

[3] F. Bonchi, A. Gionis, F. Gullo, C. E. Tsourakakis, and A. Ukkonen. Chromatic correlation clustering. *ACM TKDD*, 2015.

[4] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[5] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. Multilayer networks. *J. of Complex Networks*, 2014.