Correlation Clustering Generalized

David F. Gleich[1]

Department of Computer Science, Purdue University, West Lafayette, Indiana, USA
dgleich@purdue.edu

Nate Veldt[2]

Department of Mathematics, Purdue University, West Lafayette, Indiana, USA
lveldt@purdue.edu

Anthony Wirth[3]

School of Computing and Information Systems, The University of Melbourne, Parkville,
Victoria, Australia
awirth@unimelb.edu.au
  https://orcid.org/0000-0003-3746-6704

Abstract

We present new results for LambdaCC and MotifCC, two recently introduced variants of the well-studied correlation clustering problem. Both variants are motivated by applications to network analysis and community detection, and have non-trivial approximation algorithms.

We first show that the standard linear programming relaxation of LambdaCC has a $\Theta(\log n)$ integrality gap for a certain choice of the parameter λ. This sheds light on previous challenges encountered in obtaining parameter-independent approximation results for LambdaCC. We generalize a previous constant-factor algorithm to provide the best results, from the LP-rounding approach, for an extended range of λ.

MotifCC generalizes correlation clustering to the hypergraph setting. In the case of hyperedges of degree 3 with weights satisfying probability constraints, we improve the best approximation factor from 9 to 8. We show that in general our algorithm gives a $4(k-1)$ approximation when hyperedges have maximum degree k and probability weights. We additionally present approximation results for LambdaCC and MotifCC where we restrict to forming only two clusters.

1 Introduction

CORRELATION CLUSTERING (CC), introduced by Bansal et al. [3], is often viewed as a partitioning problem on signed graphs. Given n nodes whose edges have so-called positive or negative weights (maybe both), the goal is to find the clustering which *correlates* as much as possible with the edge weights. That is, a positive-weight edge suggests two nodes should be clustered together, while a negative-weight edge suggests separation, and these weights are

in some sense *soft constraints*. There is a variety of settings for CORRELATION CLUSTERING, including different objective functions, and special classes of edge weights, leading to a rich and interesting family of approximation algorithms and hardness results.

In this document, we consider two recent variants of the problem, called LAMBDA CORRELATION CLUSTERING (LAMBDACC) [22] and MOTIF CORRELATION CLUSTERING (MOTIFCC) [17]. Although introduced independently, both problems are motivated by applications to community detection in unsigned graphs, and are interesting to study from a theoretical perspective, each coming with non-trivial approximation guarantees. LAMBDACC is a generalization of the standard unweighted CC in which all positive edges have a common weight, while all negative edges have another (possibly different) common weight. A parameter λ determines these two weights and, implicitly, controls the size and structure of clusters formed by optimizing the objective. MOTIFCC is a generalization of CORRELATION CLUSTERING to hypergraphs, designed to provide a framework for clustering graphs based on higher-order subgraph patterns (i.e., *motifs*). We present new results for LAMBDACC and MOTIFCC, not only where the number of clusters formed is an outcome of minimizing the objective, but also where we (additionally) restrict to forming only two clusters. In summary, we make the following contributions:

1. We show that there exists some small λ such that the LAMBDACC LP relaxation has a $\Theta(\log n)$ integrality gap. This hints at why constant-factor approximations have been developed for $\lambda \geq 1/2$, but no analogous result has been found for *small λ*. We also extend the analysis of our previous algorithm for LAMBDACC [22] to outline the range of $\lambda < 1/2$ values, that admit an approximation factor in $o(\log n)$.

2. We show that when we restrict to two clusters, LAMBDACC reduces to the MIN UNCUT problem, which implies an $O(\sqrt{\log n})$ approximation for this special case [1].

3. We generalize the 4-approximation of Charikar et al. for complete unweighted correlation clustering to obtain a $4(k-1)$ approximation for MOTIFCC on hypergraphs with edges of degree k where edge weights satisfy probability constraints. We consider the same LP relaxation as Li et al. [17], and apply a similar rounding technique. However, we provide an approximation guarantee for arbitrary k that is linear in k, in addition improving the factor for $k = 3$ from 9 to 8.

4. For TWO-CLUSTER MOTIFCC, we design an algorithm that gives an asymptotic $1 + k\,2^{k-2}$ approximation by generalizing the 3-approximation of Bansal et al [3] for 2-CC (which applies when $k = 2$). This is the first combinatorial result for 2-MOTIFCC, and is a 7-approximation for $k = 3$.

2   Background and Previous Results

In the most general formulation of CORRELATION CLUSTERING on (undirected) graphs – excluding, for the moment, the generalization to hypergraphs – each pair of nodes (i, j) is assigned a pair of nonnegative weights (w_{ij}^+, w_{ij}^-), i.e., a similarity score and a dissimilarity score. In many cases, only one of these weights is assumed to be nonzero, to indicate strict similarity or strict dissimilarity between pairs of nodes. We focus on the objective of *minimizing disagreements*, which can be formally expressed as an integer linear program:

$$
\begin{aligned}
\text{minimize} \quad & \sum_{i<j} w_{ij}^+ x_{ij} + w_{ij}^-(1 - x_{ij}) \\
\text{subject to} \quad & x_{ij} \leq x_{ik} + x_{jk} && \text{for all } i, j, k \\
& x_{ij} \in \{0, 1\} && \text{for all } i < j
\end{aligned}
\tag{1}
$$

The variable x_{ij} is 1 if nodes i and j are in separate clusters, and is 0 otherwise. Thus, a clustering that separates i, j incurs a penalty (also called a *mistake*, or a *disagreement*) of

weight w_{ij}^+, while if i, j are together the penalty has weight w_{ij}^-. The objective of *maximizing agreements* has also been extensively considered: it shares the same set of optimal clusterings as minimizing disagreements, but is *easier* from the perspective of approximations. For the general weighted case, correlation clustering is equivalent to MINIMUM MULTICUT [10], which implies an $O(\log n)$ approximation, but also suggests that CORRELATION CLUSTERING (with general weights) is unlikely to be approximated to within a constant factor in polytime [6]. For weights satisfying probability constraints (i.e., $w_{ij}^+ + w_{ij}^- = 1$), Ailon et al. gave a 2.5 approximation [2]. The best approximation factor for the standard unweighted problem (i.e., $(w_{ij}^+, w_{ij}^-) \in \{(0, 1), (1, 0)\}$) is slightly better than 2.06 [7].

Fixing the number of clusters

In general, CORRELATION CLUSTERING does not require a user to specify number of clusters to be formed; the number of clusters arises naturally by optimizing the objective. However, restricting the output of CORRELATION CLUSTERING to a fixed number of clusters has also been studied extensively. In their seminal work, Bansal et al. showed a 3-approximation for minimizing disagreements in the two-cluster unweighted case (2-CORRELATION CLUSTERING) [3]. Later, Giotis and Guruswami showed a *polynomial time approximation scheme* for maximizing agreements and for minimizing disagreements, when the number clusters is a fixed constant [12]. For the maximization version, 2-CORRELATION CLUSTERING is equivalent to MAX CUT; based on this Dasgutpta et al. showed a 0.878-approximation for arbitrary weights [9]. Extending Bansal et al.'s approach, Coleman et al. introduced faster, greedy 2-approximations for minimizing disagreements for unweighted 2-CORRELATION CLUSTERING [8], and gave a more extensive overview of the historical interest in this problem. Given this recurring interest in correlation clustering with a fixed number of clusters, we address several questions involving the two-cluster case in this manuscript.

2.1 Lambda Correlation Clustering

In previous work, we introduced the LAMBDACC objective, which can be viewed as a special case of weighted correlation clustering (1) in which $(w_{ij}^+, w_{ij}^-) \in \{(1 - \lambda, 0), (0, \lambda)\}$ for some user-chosen parameter $\lambda \in (0, 1)$. This provides the following framework for partitioning unsigned networks: given an unsigned graph $G = (V, E)$, treat each edge, in E, as a *positive edge* of weight $(1 - \lambda)$ in a signed graph, and treat each non-edge as a *negative edge* with weight λ. When $\lambda = 1/2$, LAMBDACC amounts to unweighted CORRELATION CLUSTERING; with *small* λ, LAMBDACC amounts to SPARSEST CUT; and when λ is large, LAMBDACC amounts to CLUSTER DELETION. We previously outlined another, similar, edge-weighting scheme [22] that is equivalent to the MODULARITY objective [18]. We do not consider it here, however, as this scheme does not appear to lead to new approximation results.

For $\lambda > 1/2$, we gave a 3-approximation based on the LP-rounding technique of van Zuylen and Williamson [21], and a 2-approximation which holds specifically for $\lambda > |E|/(1 + |E|)$, hence, for CLUSTER DELETION. We also note that when $\lambda > 1/2$, LAMBDACC can be viewed as a specific case of the specially weighted correlation clustering variant considered by Puleo and Milenkovic [19], for which they gave a 5-approximation based on a generalization of the LP rounding scheme of Charikar et al. [5]. However, the proof strategies for all of these algorithms fail when considering arbitrarily small λ.

2.2   Motif Correlation Clustering

Li et al. introduced a higher-order generalization of CORRELATION CLUSTERING, which they call MOTIF CORRELATION CLUSTERING (MOTIFCC), as a means for clustering networks based on higher-order motif patterns shared among nodes [17]. This objective is motivated by previous successful results for motif-based graph clustering (see e.g., [4]). Although a similar higher-order correlation clustering objective was considered by Kim et al. for image segmentation [16], Li et al. were the first to study the objective from a theoretical perspective. In their approach, we let E_k denote the set of all k-tuples of nodes in G, and let each $\mathcal{E} \in E_k$ have a positive weight, $w_{\mathcal{E}}^+$, and a negative weight, $w_{\mathcal{E}}^-$. If a clustering separates at least one pair of nodes in $\mathcal{E}$, this gives a penalty of $w_{\mathcal{E}}^+$; otherwise, there is a penalty of $w_{\mathcal{E}}^-$. MOTIFCC is formally expressed as the following ILP, a generalization of ILP (1):

$$
\begin{array}{llll}
\text{minimize} & \sum_{\mathcal{E} \in E_k} w_{\mathcal{E}}^+ x_{\mathcal{E}} + w_{\mathcal{E}}^-(1 - x_{\mathcal{E}}) & & \\
\text{subject to} & x_{uv} \le x_{uw} + x_{vw} & \text{for all } u, v, w & \\
& x_{uv} \in \{0, 1\} & \text{for all } u < v & \\
& x_{uv} \le x_{\mathcal{E}} & \text{for all } u, v \in \mathcal{E} & (2)\\
& (k-1)x_{\mathcal{E}} \le \sum_{u,v \in \mathcal{E}} x_{uv} & \text{for all } \mathcal{E} \in E_k & \\
& x_{\mathcal{E}} \in \{0, 1\} & \text{for all } \mathcal{E} \in E_k. &
\end{array}
$$

The first two constraints above ensure the variables encode a clustering ($x_{uv} = 1$ if u, v are separated). Since $x_{\mathcal{E}}$ is binary, constraint $x_{\mathcal{E}} \ge x_{uv}$ ensures that if any two nodes u, v in $\mathcal{E}$ are separated, then $x_{\mathcal{E}} = 1$ (i.e., the k-tuple is split). The fourth constraint guarantees that $x_{\mathcal{E}} = 0$ if all pairs of nodes in $\mathcal{E}$ are together. Li et al. considered an even more general objective, which they referred to as MIXED MOTIF CORRELATION CLUSTERING (MMCC), where motifs of multiple sizes are considered at once, and the objective is a positive linear combination of objectives of the form (2) for different values of k. In their analysis they restrict to hyperedges of size 2 and 3, in other words they optimize an objective like this:

$$
\text{minimize} \quad \sum_{u<v} w_{uv}^+ x_{uv} + w_{uv}^-(1 - x_{uv}) + \sum_{\mathcal{E} \in E_3} w_{\mathcal{E}}^+ x_{\mathcal{E}} + w_{\mathcal{E}}^-(1 - x_{\mathcal{E}}).
$$

For this setting, they show a 9-approximation for the problem when hyperedge weights satisfy probability constraints ($w_{\mathcal{E}}^+ + w_{\mathcal{E}}^- = 1$, for every hyperedge $\mathcal{E}$ of size 2 or 3). Recently, Fukunga gave an $O(k \log n)$ approximation for general weighted hypergraphs by rounding the same LP [11].

3   New Results for LambdaCC

Given a signed graph, G, in which every pair of nodes is part of a negative edge set, E^-, or a positive edge set, E^+, the linear program relaxation of LAMBDACC is

$$
\begin{array}{llll}
\text{minimize} & \sum_{(i,j) \in E^+} (1 - \lambda)x_{ij} + & \sum_{(i,j) \in E^-} \lambda(1 - x_{ij}) & \\
\text{subject to} & x_{ij} \le x_{ik} + x_{jk} & \text{for all } i, j, k & (3)\\
& 0 \le x_{ij} \le 1 & \text{for all } i < j &
\end{array}
$$

Although a constant-factor approximation for LAMBDACC exists for $\lambda \ge 1/2$, by rounding LP (3), we show that there exists some small λ such that the integrality gap is $O(\log n)$. We then give parameter-dependent approximation guarantees for small λ, and consider new results for two-cluster LAMBDACC.

3.1 Integrality Gap for the LambdaCC Linear Program

Demaine et al. prove that the integrality gap for the general weighted Correlation Clustering LP relaxation is $O(\log n)$ [10]. This does not immediately imply anything for our specially weighted case, but adapting some of their ideas, and adding some non-trivial steps, does reveal an $O(\log n)$ integrality gap for the LambdaCC linear program relaxation. The proof takes the following steps.

1. Construct an instance of LambdaCC from an expander graph, G.
2. Prove that, because of the expander properties of G, the optimal LambdaCC clustering must make $\Omega(n)$ mistakes.
3. Demonstrate the LP relaxation has a feasible solution with a score of $O(n/\log n)$.

In order to accomplish third step listed above, we do not (necessarily) produce a feasible solution for the standard LP relaxation of LambdaCC: in particular, in our solution triangle constraints are not guaranteed. Instead, we produce a feasible solution for a related linear program considered by Wirth in his PhD thesis [23]. The fundamental construct of this LP is the Negative Edge with Positive Path Cycle (NEPPC), where, $NEPPC(i_1, i_2, \ldots, i_m)$ represents a sequence (a *path*) of (positive) edges, $(i_1, i_2), (i_2, i_3), \ldots, (i_{m-1}, i_m) \in E$, with a single (negative) non-edge completing the cycle: $(i_1, i_m) \notin E$. For LambdaCC, defined on a graph $G = (V, E)$, with parameter $\lambda \in (0, 1)$, we have the linear program:

$$
\begin{array}{lll}
\text{minimize} & \sum_{(i,j) \in E} (1 - \lambda) x_{ij} + & \sum_{(i,j) \notin E} \lambda (1 - x_{ij}) \\
\text{subject to} & x_{i_1, i_m} \leq \sum_{j=1}^{m-1} x_{i_j, i_{j+1}} & \text{for all } NEPPC(i_1, i_2, \ldots, i_m) \\
& x_{ij} \leq 1 & \text{for all } (i, j) \notin E \\
& 0 \leq x_{ij} & \text{for all } (i, j) .
\end{array}
\tag{4}
$$

Wirth [23] proved that the set of optimal solutions to the NEPPC linear program (4) is exactly the same as the optimal solution set to the Correlation Clustering LP, the relation of ILP (1).[4] Since a feasible solution for the LambdaCC NEPPC linear program (4) is an upper bound on the optimum for (4), which is the same as the optimum for the standard LambdaCC LP, we can bound the optimum of the latter. We now prove our result:

▶ **Theorem 1.** *There exists some λ such that the integrality gap of LP (3) is $O(\log n)$.*

Proof. The expander graph

Let $G = (V, E)$ be a (d, c)-expander graph, where both d and c are constants (Reingold et al. proved that such expanders exist [20]). That is, G is d-regular, and for every $S \subset V$ with $|S| \leq n/2$, we have

$$
\frac{\mathbf{cut}(S)}{|S|} \geq c \implies \frac{\mathbf{cut}(S)}{|S|} + \frac{\mathbf{cut}(S)}{|\bar{S}|} \geq c \implies \frac{\mathbf{cut}(S)}{|S||\bar{S}|} \geq \frac{c}{n}
$$

where $\mathbf{cut}(S)$ denotes the number of edges between S and $\bar{S} = V \backslash S$. Define the *scaled sparsest cut* of a set S to be $\mathbf{cut}(S)/(|S||\bar{S}|)$ and let λ^* minimize this ratio over all possible sets $S \subset V$. In previous work we showed that for any $\lambda \leq \lambda^*$, the optimal LambdaCC clustering places all nodes into one cluster, but there exists a range of λ values slightly larger than λ^* such that the optimum clustering coincides with a partitioning that produces the

---

[4] Although the proof is shown for the unweighted case, we note that all aspects of the proof immediately carry over to the weighted case.

scaled sparsest cut score [22]. For the expander graph we consider, this λ^* is at most the scaled sparsest cut score obtained by setting S to be a single node, so we have these upper and lower bounds on λ^*: $c/n \leq \lambda^* \leq d/(n-1)$.

The LambdaCC construction

Let S^* be a set inducing an optimal scaled sparsest cut partition: $\lambda^* = \mathbf{cut}(S^*)/(|S^*||\bar{S}^*|)$. From Theorem 3.2 in our previous work [22], we know that there exists some λ', slightly larger than λ^* whose optimum LAMBDACC solution is the bipartition $\{S^*, \bar{S}^*\}$; let the LAMBDACC score of this solution be OPT, and let $\varepsilon = \lambda' - \lambda^*$. We can choose $\varepsilon > 0$ to be arbitrarily small, so it suffices to assume $\lambda' < 2\lambda^*$.

Bounding OPT from below

With our choice of λ', by definition,

$$
\begin{aligned}
OPT &= \mathbf{cut}(S^*) - \lambda'|S^*||\bar{S}^*| + \lambda'\left(\binom{n}{2} - |E|\right) \\
&= 0 - \varepsilon|S^*||\bar{S}^*| + \lambda^*\left(\binom{n}{2} - |E|\right) + \varepsilon\left(\binom{n}{2} - |E|\right) \\
&= \lambda^*\left(\binom{n}{2} - |E|\right) + \varepsilon\left(\binom{n}{2} - |E| - |S^*||\bar{S}^*|\right) \\
&\geq \lambda^*\left(\frac{n(n-1)}{2} - \frac{nd}{2}\right) + \varepsilon\left(\frac{n(n-1)}{2} - \frac{nd}{2} - \frac{n^2}{4}\right) \\
&\geq \frac{c}{n}\left(\frac{n(n-1)}{2} - \frac{nd}{2}\right) = \Omega(n),
\end{aligned}
$$

relying on the definition of λ^*, the fact that $|E| = nd/2$ in this expander graph, and the bound $|S^*||\bar{S}^*| \leq n^2/4$.

Upper Bounding the NEPPC LP

We now show that a carefully crafted feasible solution for the NEPPC LP (4) has score $O(n/\log n)$. Let $dist(i, j)$ denote the minimum path length between nodes i and j in G, based on unit-weight edges E. We are assuming the graph is connected, so each $dist(i, j)$ is a finite integer. (If the graph is not connected, we ought to solve LAMBDACC on each connected component separately.) Consider the following setting of values x_{ij}:

$$
x_{ij} = \begin{cases}
2/(\log_d n) & \text{if } (i,j) \in E \\
1 & \text{if } (i,j) \notin E \text{ and } dist(i,j) \geq (\log_d n)/2 \\
0 & \text{if } (i,j) \notin E \text{ and } dist(i,j) < (\log_d n)/2 .
\end{cases}
$$

We show that this is feasible for the NEPPC LP (4). Since all (positive) edges are assigned the same LP score, the NEPPC constraints are satisfied at a (negative) non-edge, (i, j), if and only if $x_{ij} \leq dist(i,j) \cdot 2/(\log_d n)$. When $dist(i, j)$ is less than $\log_d(n)/2$, $x_{ij} = 0$, so this inequality is trivially true. When $dist(i, j)$ is at least $\log_d(n)/2$, the NEPPC inequality is true because $dist(i, j) \cdot 2/(\log_d n)$ is at least 1, which is x_{ij}.

For constant d, the contribution from the (positive) edges to LP (4) is:

$$
(1 - \lambda')|E|2/(\log_d n) = (1 - \lambda')(nd)/(\log_d n) = O(n/\log n) .
$$

From the (negative) non-edges, since the factor is $1-x_{ij}$, we only have a non-zero contribution from the set of $(i,j) \notin E$ such that $dist(i,j) < (\log_d n)/2 = \log_d \sqrt{n}$. For each node $v \in V$, there are at most $d^{\log_d \sqrt{n}} = \sqrt{n}$ nodes within this distance; the total number of non-edges that contribute to the LP cost is therefore in $O(n\sqrt{n})$. Each has a weight $\lambda' < 2\lambda^*$, so

$$\text{LP contribution of non-edges} \leq \lambda' n\sqrt{n} \leq (2d/(n-1))\, n\sqrt{n} = O(\sqrt{n}) \leq O(n/\log n).$$

Therefore, the total LP cost corresponding to this feasible solution to NEPPC LP (4) is $O(n/\log n)$. Since the optimal LAMBDACC solution has cost $\Omega(n)$, we have shown that there exists some $\lambda < 1/2$ such that the LP relaxation (3) has an integrality gap of $O(\log n)$. ◄

3.2 Parameter-Dependent Approximation Guarantees

We now describe improved approximation guarantees for ranges of λ below $1/2$, extending the analysis of our previous 3-approximation for $\lambda \geq 1/2$ [22]. This 3-approximation is obtained by solving the LP relaxation, forming a new unweighted signed graph G', and then applying the *pivoting* procedure, which repeatedly selects a node and clusters it with its positive neighbors. The approximation guarantee comes from applying a theorem of van Zuylen and Williamson for deterministic pivoting algorithms for correlation clustering [21]. We give a full proof of the following result in the extended version of the paper [13]

▶ **Theorem 2.** *Let (x_{ij}) be the variables from solving the* LAMBDACC *LP relaxation, and form a new unweighted* CORRELATION CLUSTERING *input G' by putting a positive edge between i and j, if $x_{ij} \leq 1/3$ and a negative edge otherwise. Applying a pivoting algorithm to G' yields a clustering that is a 3-approximation for $\lambda > 1/2$, and an α-approximation otherwise, where $\alpha = \max\{\, 1/\lambda, (6-3\lambda)/(1+\lambda)\}$.*

This theorem implies an approximation better than 4.5 for all $\lambda \in (0.2324, 0.5)$, but shows that the algorithm performs worse and worse as λ decreases. However, for all λ in $\omega(1/\log n)$, this outputs a better result than the standard, $O(\log n)$, rounding scheme.

3.3 Two-Cluster LambdaCC

Before moving on we note an approximation guarantee and a hardness result that holds for the two-cluster variant of LAMBDACC.

▶ **Theorem 3.** *Two-cluster* LAMBDACC *can be reduced to the weighted* MIN UNCUT *problem. An instance of* MIN UNCUT *with non-zero optimum can be reduced to an instance of two-cluster* LAMBDACC *whose objective score for any clustering differs by at most a small constant factor.*

We give a full proof in the full version [13]. The first fact implies the $O(\sqrt{\log n})$ approximation, due to Agarwal et al. [1], extends to 2-LAMBDACC. This has important ramifications even without the restriction on the number of clusters; LAMBDACC is guaranteed to form two clusters for a certain parameter regime near λ^* [22, Theorem 3.2]. The reduction from MIN UNCUT to two-cluster LAMBDACC implies the latter cannot be approximated to within any constant factor [15, 14].

4 Motif Correlation Clustering

We now turn to improved approximations for MOTIFCC. We begin by presenting a $4(k-1)$ approximation algorithm for the problem for hyperedges of degree k with edge weights satisfying probability constraints. We then consider a first step towards algorithms that do not rely on solving an expensive LP relaxation, by showing how to obtain a combinatorial approximation for two-cluster MOTIFCC (2-MOTIFCC) for complete, unweighted instances.

Algorithm 1 Generalized CGW for Minimizing Hyper-Disagreements.

> **Input:** Signed hypergraph $G = (V, E_k)$, and threshold parameters γ and δ
> Solve the LP-relaxation of ILP (2), obtaining *distances* (x_{ij})
> $W \leftarrow V, \mathcal{C} \leftarrow \varnothing$
> **while** $W \neq \varnothing$ **do**
> 5:   Choose $u \in W$ arbitrarily, and define $T_u \leftarrow \{i \in W \backslash \{u\} : x_{ui} \leq \gamma\}$
>   **if** $\sum_{i \in T_u} x_{ui} < \gamma \delta |T_u|$ **then** $S := \{u\} \cup T_u$
>   **else** $S := \{u\}$
> $\mathcal{C} \leftarrow \mathcal{C} \cup \{S\}, W \leftarrow W \backslash S$

4.1 The $4(k-1)$ approximation

Our algorithm for MOTIFCC is closely related to the approach of Li et al. [17] and directly generalizes the LP-rounding technique of Charikar et al. [5], which is itself an instantiation of the more general rounding procedure given in Algorithm 1. The general algorithm forms clusters based on threshold parameters γ and δ, which are part of the input. Charikar et al. proved that for the $k = 2$ unweighted case of MOTIFCC, setting $\gamma = \delta = 1/2$ leads to a 4-approximation. Li et al. generalized this to obtain a 9-approximation for $k = 3$ in the more general probability constrained case, by selecting $\gamma = \delta = 1/3$ [17]. Although they did not provide an analysis for motifs of size $k > 3$, it appears that their strategy of setting $\gamma = \delta = 1/k$ would at best lead to a k^2 approximation. In contrast, we analyze a choice of parameters which leads to an approximation that is *linear* in k.

The result is somewhat detailed, and we begin with some notation. Let the family of k-tuples be E_k, and let $W \subseteq V$ be the subset of nodes in G that remain unclustered after a certain number of rounds of Algorithm 1. When considering a vertex $u \in W$ and a specific k-tuple $\mathcal{E}$, it will be convenient to define a to be the node in $\mathcal{E}$ *closest* to u, i.e., $\arg\min_{i \in \mathcal{E}} x_{ui}$, while z is the *farthest*, $\arg\max_{i \in \mathcal{E}} x_{ui}$. We have T_u similar to Algorithm 1, with $\gamma = 1/(2(k-1))$, while T_u^k are those k-tuples that include u, with all non-u nodes in T_u:

$$T_u = \left\{ i \in W \backslash \{u\} : x_{ui} \leq \tfrac{1}{2(k-1)} \right\} \quad \text{and} \quad T_u^k = \{\mathcal{E} \in E_k : u \in \mathcal{E} \text{ and } (\mathcal{E} - \{u\}) \subset T_u\}. \quad (5)$$

For $z \notin T_u$, we let P_z be those k-tuples in which z is the farthest element from u and some $a \in T_u$ is closest, viz.

$$P_z = \{(a, j_2, j_3, \ldots, j_{k-1}, z) \in E_k : a \in T, x_{ua} \leq x_{u,j_2} \leq x_{u,j_3} \leq \cdots \leq x_{uz}\}. \quad (6)$$

Finally, $\mathrm{LP}(A)$ denotes the LP score associated with a subset A of the set of degree-k hyperedges: $A \subseteq E_k$.

▶ **Theorem 4.** *For constant k, let $G = (V, E_k)$ be a hypergraph in which for all $\mathcal{E} \in E_k$ the weights satisfy probability constraints, $w_{\mathcal{E}}^+ + w_{\mathcal{E}}^- = 1$. Applying Algorithm 1 with $\gamma = 1/(2(k-1))$ and $\delta = 1/2$ outputs a clustering that is a $4(k-1)$-approximation to MOTIFCC.*

We start with a proof outline, establish three lemmata, and then give full details in Section 4.2. At each step the algorithm forms a cluster S_u around an arbitrary $u \in W$. This cluster is associated with a set of hyperedges A_u that have either been cut or placed inside of S_u. If for each S_u individually we can show that mistakes made at A_u are within a fixed factor of the lower bound $\mathrm{LP}(A_u)$, this will imply an overall bound for the entire clustering.

In forming a cluster around u, the algorithm first identifies a set of nodes T_u whose LP distance to u is at most a preliminary threshold $\gamma = 1/(2(k-1))$. To verify if $\{u\} \cup T_u$ will make a good cluster, the algorithm checks whether *on average* the distance from u to T_u is below a tighter threshold $\gamma\delta = 1/(4(k-1))$. If this doesn't hold, we let $\{u\}$ remain a singleton cluster. In forming clusters, we only explicitly consider distance variables x_{ij} for $(i,j) \in V \times V$. However, the MOTIFCC objective and its LP relaxation both depend on the hyperedge variables $x_{\mathcal{E}}$ for $\mathcal{E} \in E_k$. Therefore, in order to bound the weight of hyperedge mistakes we must leverage the LP constraints to understand the relationships between distance and hyperedge variables. Lemma 5 establishes several useful relationships we will need later. Also, because our algorithm makes decisions based on the average distance between u and T_u, we must interpret what this means for the average value of hyperedge variables $x_{\mathcal{E}}$ in certain sets of hyperedges that we are trying to account for (e.g. P_z and T_u^k in (5) and (6)). Lemmata 6 and 7 address this task. We give proofs for these lemmata in the full version of the paper [13]. In the following, we adopt the convention that $x_{ii} = 0$ for every node $i \in V$.

▶ **Lemma 5.** *For all $\mathcal{E} \in E_k$ and any $u \in V$,*

1. $x_{\mathcal{E}} \le \sum_{i \in \mathcal{E}} x_{ui}$,
2. $x_{\mathcal{E}} \le x_{ua} + (k-1)x_{uz}$, *and*
3. $x_{\mathcal{E}} \ge x_{uz} - x_{ua}$.

▶ **Lemma 6.** *For all $u \in W \subseteq V$, if $\sum_{i \in T_u} x_{ui} \ge \beta|T_u|$, then $\sum_{\mathcal{E} \in T_u^k} x_{\mathcal{E}} \ge \beta|T_u^k|$.*

▶ **Lemma 7.** *For all $\mathcal{E} \in P_z$, let $a_{\mathcal{E}}$ denote the node in $\mathcal{E}$ closest to u. If $\sum_{i \in T_u} x_{ui} < \beta|T_u|$, then $\sum_{\mathcal{E} \in P_z} x_{ua_{\mathcal{E}}} < \beta|P_z|$.*

4.2   Proof of Theorem 4

Proof. We must account for the weight of positive mistakes made at singleton clusters, $\{u\}$, and the weight of both positive and negative mistakes made at non-singleton clusters.

Singleton Clusters

Consider a cluster $S = \{u\}$. The algorithm incurs a penalty $w_{\mathcal{E}}^+$ for each $\mathcal{E}$ such that $u \in \mathcal{E}$. If some node $j \in \mathcal{E} - \{u\}$ is not in T_u, then the contribution to the LP score is $w_{\mathcal{E}}^+ x_{\mathcal{E}}$, which is at least $w_{\mathcal{E}}^+ x_{uj}$, and therefore exceeds $w_{\mathcal{E}}^+/(2(k-1))$. Thus the cost of the mistake at most $2(k-1)$ times the LP penalty.

It remains to account for all positive hyperedges in T_u^k. Even if $w_{\mathcal{E}}^+ = 1$ for all $\mathcal{E} \in T_u^k$, $|T_u^k| = \binom{|T|}{k-1}$ is an upper bound on the total weight of mistakes made on hyperedges in T_u^k. By the first observation of Lemma 5, and because $u \in \mathcal{E}$,

$$x_{\mathcal{E}} \le \sum_{i \in \mathcal{E}} x_{ui} \le (k-1)\tfrac{1}{2(k-1)} = \tfrac{1}{2}, \quad \text{hence,} \quad (1 - x_{\mathcal{E}}) \ge x_{\mathcal{E}}.$$

Since $w_{\mathcal{E}}^+ + w_{\mathcal{E}}^- = 1$, we can lower bound the contribution of T_u^k to the LP score:

$$\mathrm{LP}(T_u^k) = \sum_{\mathcal{E} \in T_u^k} w_{\mathcal{E}}^+ x_{\mathcal{E}} + w_{\mathcal{E}}^-(1 - x_{\mathcal{E}}) \ge \sum_{\mathcal{E} \in T_u^k} w_{\mathcal{E}}^+ x_{\mathcal{E}} + w_{\mathcal{E}}^- x_{\mathcal{E}} = \sum_{\mathcal{E} \in T_u^k} x_{\mathcal{E}} \ge |T_u^k|\frac{1}{4(k-1)},$$

by Lemma 6, so we have paid for the mistakes within a factor $4(k-1)$.

Negative Mistakes at Non-Singletons

Next, we account for negative mistakes in clusters of the form $S = \{u\} \cup T$. Charikar et al. showed that, when $k = 2$, these are accounted for within a factor 4; we prove the same for all $k \geq 3$. For each $\mathcal{E} \in E_k$ such that $\mathcal{E} \subset S$, the algorithm makes a mistake of weight $w_\mathcal{E}^-$. On the other hand, the LP pays $w_\mathcal{E}^-(1 - x_\mathcal{E})$. Applying the first observation in Lemma 5,

$$x_\mathcal{E} \leq \sum_{i \in \mathcal{E}} x_{ui} \leq k \frac{1}{2(k-1)} \leq \frac{3}{4}, \quad \text{hence,} \quad w_\mathcal{E}^-(1 - x_\mathcal{E}) \geq \frac{w_\mathcal{E}^-}{4},$$

and we have the desired result for $k \geq 3$.

Positive Mistakes at Non-Singletons

A hyperedge $\mathcal{E}$ contained entirely within $S = \{u\} \cup T$ incurs no positive-weight error. So, finally, we account for positive mistakes at hyperedges $\mathcal{E}$ where at least one node of $\mathcal{E}$ is in S and at least one node in $\mathcal{E}$ is $\notin S$. For each such hyperedge, we explicitly label the nodes of $\mathcal{E}$ with indices $a = j_1 < j_2 < \cdots < j_k = z$, with $x_{ua} = x_{u,j_1} \leq x_{u,j_2} \leq \ldots \leq x_{u,j_k} = x_{uz}$ where $a \in T_u$ and $z \notin T_u$. By the second and third observation in Lemma 5 we know that

$$x_{uz} - x_{ua} \leq x_\mathcal{E} \leq x_{ua} + (k-1)x_{uz}, \tag{7}$$

First, if $a = u$, then we know $w_\mathcal{E}^+ x_\mathcal{E} \geq w_\mathcal{E}^+(x_{uz} - x_{uu}) > w_\mathcal{E}^+/(2(k-1))$, and we have individually accounted for each such positive mistake within a factor $2(k-1)$. If $a \neq u$ and $x_{uz} \geq 3/(4(k-1))$, we bound the mistake within factor $4(k-1)$:

$$w_\mathcal{E}^+ x_\mathcal{E} \geq w_\mathcal{E}^+(x_{uz} - x_{ua}) \geq w_\mathcal{E}^+(3/(4(k-1)) - 1/(2(k-1)) = w_\mathcal{E}^+/(4(k-1)).$$

Finally, if $a \neq u$ and $x_{uz} \in \left(\frac{1}{2(k-1)}, \frac{3}{4(k-1)} \right)$, we account for all positive weights associated with edges in the following set, together:

$$P_z = \{\mathcal{E} \in E_k : \mathcal{E} = (a, j_2, \ldots, z), a \in T, x_{ua} \leq x_{u,j_2} \leq x_{u,j_3} \leq \cdots \leq x_{uz}\}.$$

The weight of mistakes made by the algorithm is $W_z^+ = \sum_{p \in P_z} w_p^+$, and we also define $W_z^- = \sum_{p \in P_z} w_p^-$. We start by observing that, since $x_{ua} \leq x_\mathcal{E}$ and $W_z^+ + W_z^- = |P_z|$, due to probability constraints on weights, Lemma 7 tells us that $\sum_{\mathcal{E} \in P_z} x_{ua} < (W_z^+ + W_z^-)/(4(k-1))$.

$$
\begin{aligned}
LP(P_z) &= \sum_{\mathcal{E} \in P_z} w_\mathcal{E}^+ x_\mathcal{E} + w_\mathcal{E}^-(1 - x_\mathcal{E}) \\
&\geq \sum_{\mathcal{E} \in P_z} w_\mathcal{E}^+(x_{uz} - x_{ua}) + w_\mathcal{E}^-(1 - x_{ua} - (k-1)x_{uz}) \quad \text{(by inequalities in (7))} \\
&= \sum_{\mathcal{E} \in P_z} w_\mathcal{E}^+ x_{uz} + w_\mathcal{E}^-(1 - (k-1)x_{uz}) - \sum_{\mathcal{E} \in P_z} x_{ua} \\
&\geq W_z^+ x_{uz} + W_z^-(1 - (k-1)x_{uz}) - \frac{W_z^+ + W_z^-}{4(k-1)} \quad \text{(by the starting observation)} \\
&\geq W_z^+ \left(\frac{1}{2(k-1)} - \frac{1}{4(k-1)} \right) + W_z^- \left(1 - \frac{1}{4(k-1)} - (k-1)\frac{3}{4(k-1)} \right) \geq W_z^+ \frac{1}{4(k-1)},
\end{aligned}
$$

so the mistakes on all hyperedges in P_z are, collectively, accounted for within factor $1/(4(k-1))$, concluding the Proof of Theorem 4. ◄

We outline two immediate extensions of this theorem in the full version [13]. First we note that the same approximation guarantees holds for the MIXED MOTIF CORRELATION CLUSTERING objective, considered by Li et al. We then consider a hybrid LAMBDA-MCC objective in which positive hyperedges have weight $(1 - \lambda)$ and negative hyperedges have weight λ, for which the algorithm is guaranteed to produce the same approximation factor when $\lambda \geq 1/2$.

Algorithm 2 PICK-A-PIVOT-TUPLE.

Input: An instance of 2-MOTIFCC: $G = (V, E_k)$ be a hypergraph where $(w_{\mathcal{E}}^+, w_{\mathcal{E}}^-) \in \{(0, 1), (1, 0)\}$ for every k-tuple.
for $(k-1)$-tuple $\mathcal{K} \subseteq V$ **do**
  $\mathcal{C}_{\mathcal{K}} \leftarrow$ the clustering formed by placing $\mathcal{K}$ in a cluster with all u such that $\mathcal{E} = \mathcal{K} \cup \{u\}$ is positive, and placing all remaining nodes in the other cluster.
Return the $\mathcal{C}_{\mathcal{K}}$ with fewest mistakes.

4.3 Two-Cluster MotifCC

The LP relaxation of MOTIFCC involves $O(n^k)$ variables and $O(n^k)$ constraints for all $k > 2$, and is therefore very expensive to solve in practice. For standard CORRELATION CLUSTERING, only a few of the known approximation algorithms avoid solving an expensive convex relaxation [2, 3]; it is natural to ask whether a similar, combinatorial, approach can be taken for MOTIFCC. We give first steps in this direction, with a constant-factor combinatorial approximation algorithm for MOTIFCC, when the output is restricted to two clusters, generalizing the 3-approximation of Bansal et al. for 2-CORRELATION CLUSTERING [3]. Our method is shown in Algorithm 2. We call this algorithm *Pick-a-Pivot-Tuple*, and show it satisfies the following result:

▶ **Theorem 8.** *For a constant integer $k > 1$, Algorithm 2 returns a $(1 + kc)$-approximation for 2-MOTIFCC, where $c \leq 2^{k-2}$ for $k = 2, 3$, while $\lim_{n \to \infty} c = 2^{k-2}$ for $k > 3$.*

We give a proof of the above result in the full version [13]. Although the exponential dependence on k makes this a poor approximation for large motifs, at least in the case $k = 3$, this is a 7-approximation for all n, not just for large n.

5 Discussion

We have demonstrated a $\Theta(\log n)$ integrality gap for the LAMBDACC LP relaxation, which highlights why previous attempts to obtain a constant-factor approximation via LP rounding have failed. It remains an open question whether better approximation factors exist for small values of λ in $O(1/\log n)$. For minimizing disagreements, there are relatively few techniques that don't rely on the LP relaxation that lead to approximations better than $O(\log n)$ for different variants of correlation clustering. The next step is either to develop an entirely new approach or prove further hardness results for approximating LAMBDACC when λ is small.

For MOTIFCC, we have given an approximation algorithm for arbitrary (constant) hyperedge size k that is linear in k, and provided a first combinatorial approximation result, which avoids solving an LP relaxation, for to the two-cluster case. An interesting open question is whether a pivoting algorithm à la Ailon et al. [2] could be developed for the MOTIFCC objective. For maximizing agreements, the simple strategy of either placing all nodes together or separating all nodes into singletons will still lead to a 1/2-approximation for hypergraphs with arbitrary weights and any k. This leads to open questions about what results for maximizing agreements can be generalized to the hypergraph setting. Another open question is whether an approximation that is independent of k could be developed for minimizing disagreements in hypergraphs.

────── **References** ──────

1 Amit Agarwal, Moses Charikar, Konstantin Makarychev, and Yury Makarychev. $O(\sqrt{\log n})$ Approximation Algorithms for Min UnCut, Min 2CNF Deletion, and Directed Cut Problems. In *STOC 05*, pages 573–581. ACM, 2005. `doi:10.1145/1060590.1060675`.

2 Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM)*, 55(5):23, 2008.

3 Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation Clustering. *Machine Learning*, 56:89–113, 2004.

4 Austin R. Benson, David F. Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016. `doi:10.1126/science.aad9029`.

5 Moses Charikar, Venkatesan Guruswami, and Anthony Wirth. Clustering with qualitative information. *Journal of Computer and System Sciences*, 71(3):360–383, 2005.

6 Shuchi Chawla, Robert Krauthgamer, Ravi Kumar, Yuval Rabani, and D. Sivakumar. On the Hardness of Approximating Multicut and Sparsest-cut. *Computational Complexity*, 15(2):94–114, June 2006. `doi:10.1007/s00037-006-0210-9`.

7 Shuchi Chawla, Konstantin Makarychev, Tselil Schramm, and Grigory Yaroslavtsev. Near optimal LP rounding algorithm for correlation clustering on complete and complete k-partite graphs. In *STOC 15*, pages 219–228. ACM, 2015.

8 Tom Coleman, James Saunderson, and Anthony Wirth. A Local-Search 2-Approximation for 2-Correlation-Clustering. In *ESA 08*, pages 308–319, 2008. `doi:10.1007/978-3-540-87744-8_26`.

9 Bhaskar DasGupta, German A. Enciso, Eduardo Sontag, and Yi Zhang. Algorithmic and Complexity Results for Decompositions of Biological Networks into Monotone Subsystems. In *WEA 06*, pages 253–264, 2006.

10 Erik D Demaine, Dotan Emanuel, Amos Fiat, and Nicole Immorlica. Correlation clustering in general weighted graphs. *Theoretical Computer Science*, 361(2-3):172–187, 2006.

11 Takuro Fukunaga. LP-based pivoting algorithm for higher-order correlation clustering. In *Computing and Combinatorics*, pages 51–62, Cham, 2018. Springer International Publishing.

12 Ioannis Giotis and Venkatesan Guruswami. Correlation Clustering with a Fixed Number of Clusters. *Theory OF Computing*, 2:249–266, 2006.

13 David F. Gleich, Nate Veldt, and Anthony Wirth. Correlation Clustering Generalized. *arXiv*, cs.DS, 2018. `arXiv:1809.09493`.

14 S. A. Khot and N. K. Vishnoi. The unique games conjecture, integrality gap for cut problems and embeddability of negative type metrics into ℓ_1. In *FOCS 05*, pages 53–62, October 2005. `doi:10.1109/SFCS.2005.74`.

15 Subhash Khot. On the Power of Unique 2-prover 1-round Games. In *STOC 02*, pages 767–775, 2002. `doi:10.1145/509907.510017`.

16 Sungwoong Kim, Sebastian Nowozin, Pushmeet Kohli, and Chang D. Yoo. Higher-Order Correlation Clustering for Image Segmentation. In *NIPS 11*, pages 1530–1538, 2011. URL: `http://papers.nips.cc/paper/4406-higher-order-correlation-clustering-for-image-segmentation.pdf`.

17 P. Li, H. Dau, G. Puleo, and O. Milenkovic. Motif clustering and overlapping clustering for social network analysis. In *INFOCOM 17*, pages 1–9, May 2017. `doi:10.1109/INFOCOM.2017.8056956`.

18 Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(026113), 2004.

19 G. Puleo and O. Milenkovic. Correlation Clustering with Constrained Cluster Sizes and Extended Weights Bounds. *SIAM Journal on Optimization*, 25(3):1857–1872, 2015. `doi:10.1137/140994198`.

20 Omer Reingold, Salil Vadhan, and Avi Wigderson. Entropy waves, the zig-zag graph product, and new constant-degree expanders and extractors. In *FOCS 00*, pages 3–13, 2000.

21 Anke van Zuylen and David P. Williamson. Deterministic Pivoting Algorithms for Constrained Ranking and Clustering Problems. *Mathematics of Operations Research*, 34(3):594–620, 2009. `doi:10.1287/moor.1090.0385`.

22 Nate Veldt, David F Gleich, and Anthony Wirth. A Correlation Clustering Framework for Community Detection. In *WWW 18*, pages 439–448. International World Wide Web Conferences Steering Committee, 2018.

23 Anthony Ian Wirth. *Approximation Algorithms for Clustering*. PhD thesis, Princeton University, November 2004. Computer Science Technical Report 716.