LP based sub-2 approximations in correlated clustering

Anish Jayant

Spring 2025

In the problem of *correlation clustering*, the input is a graph $G = (V, E)$ and a mapping $f : E \to \{\pm 1\}$ where $f(e)$ indicates whether the endpoints of e are 'similar' or 'dissimilar'. Informally, the goal is to find a partition (or *cluster*) $V = [V_1, V_2, .., V_k]$ such maximizing the number of $+1$ edges within clusters and and -1 edges between clusters (max-agree) or, equivalently, minimize the count of -1 edges within a cluster and $+1$ edges between clusters (min-disagree). The max-agree setting is resolved with a PTAS (BBC04), however, note that in application, we care more about the *mistakes* – suboptimality in the sense of min-disagree formulation – which does not follow from approximation of max-agree. Additionally, this problem is accompanied by APX-hardness result; so, what is the best achievable approximation?

More formally, the min-disagree objective is

$$\min_{[V_1, V_2, \ldots, V_k]} \sum_{i=1}^{k} \sum_{u \in V_i} \left[\sum_{v \in V_i} 1[(u,v) \in E^-] + \sum_{v \notin V_i} 1[(u,v) \in E^+] \right], \tag{1}$$

where $E^- = f^{-1}(-1)$ and $E^+ = f^{-1}(+1)$ are the negative and positive edges, respectively. For the rest of this reaction paper, we will try to understand progress and barriers in approximation algorithm design for the min-disagree formulation when $G = K_n$, the complete graph.[1] In class, we concluded with showing the integrality gap of 2 for the ILP method (later, Eq. 2); we spend the first half reviewing methods in this direction, and the second on (very) recent LP-based approaches.

1 Worse than 2 Approximation

A first result in correlation clustering by Bansal, Blum, and Chawla (BBC04) gave a PTAS for max-agree and a constant (17, 433) approximation algorithm for min-disagree. Though the con-

[1] With regard to the page limit, we will focus *strictly* on providing polynomial time running-time algorithms. Unsurprisingly, a major and interesting concern is doing this with little memory/linear time/in parallel.

stant is terrible, their proof strategy is valuable and later sharpened by (ACN08). They effectively count "bad triangles", $\mathcal{T}$: 3-cliques $u, v, w \in V$ with inconsistent edge labelings. Noting that *any* clustering collects at least one error per bad triangle, they show an algorithm with controlled error in this counting of $|\mathcal{T}|$, thus bringing the approximation.

Through an entirely different perspective of the ILP, (CGW03) show a much improved 4-approx. Consider $m = O(n^2)$ variables of the form x_{ij} which is 0 if v_i, v_j are in the same cluster and 1 otherwise, we can frame the min-disagree problem as,

$$\begin{array}{lll} \text{min.} & \sum_{E^+} x_{ij} + \sum_{E^-} (1 - x_{ij}) & \\ \text{subject to} & x_{ik} \leq x_{ij} + x_{jk} & \forall (i,j,k) \in \binom{V}{3} \\ & x_{ij} \in \{0,1\}, & \forall (i,j) \in \binom{V}{2} \end{array} \tag{2}$$

where the inequality constraint enforces cluster membership to be an equivalence. By relaxing Eq. 2 allowing $x_{ij} \in [0,1]$, the solution $\{x_{e_1}, ..., x_{e_m}\}$ defines a semi-metric space on vertices, say S. The rounding approach is intuitive: iterating over unclustered $u \in V$, find $T = B_S(u, 1/2)$ and determine the average distance of $T \backslash \{u\}$ from u; if this exceeds $1/4$ (i.e., 'many' points are peripheral), designate $\{u\}$ as a singleton cluster, and if not let T be the cluster. This work also demonstrates the integrality gap of 2 (meaning that the best possible ILP rounding procedure can only hope to achieve 2-approx.) and claims that even finding a 3-approx. would require a totally different strategy.

In fact, (ACN08) shows that, with a little bit of randomness, a 3-approx. can be achieved by the following simple procedure: pick an un-clustered vertex $v \in V$ uniformly at random, cluster it with all unclustered $u \in V$ such that $(u, v) \in E^+$, and repeat until all vertices are clustered. Continuing this probabilistic line of reasoning, rounding Eq. 2 by adding vertex w to cluster centered at u with probability $1 - x_{uw}$ brings a 2.5 approximation. The final work in this direction (CMSY15) refines the rounding of positive edges to $1 - f(x)$ where f is a carefully clipped quadratic, and shows the current best rounding approximation for Eq. 2 at 2.06.

2 Breaking 2-Approximation

Using the Sherali-Adams heirarchy, (CALN23) was the first to show a $1.994 + \epsilon$ approximation. The rough intuition here is to add many more constraints to the LP relaxation in the form of variables for *sets of vertices*, and shows a rounding that achieves roughly the integrality gap. Using a pre-clustering approach (CALLN23) is able to bring a 1.73-approximation and lighten the analysis of their earlier work. The most recent result from this group (CCAL+24) is a remarkable 1.437 approximation based on an exponentially large LP which can be rounded close to its integrality gap 4/3. This final result unifies previous LP rounding schemes using the following *cluster LP* and interestingly *does not first pass through an ILP*!

For each set $S \subseteq V$, let variable $z_S \in [0,1]$ roughly indicate whether S appears in the optimal partition (with larger value being more likely) and x_{uv} be the probability u, v are in the same cluster. The Cluster LP is as follows:

$$\begin{aligned} \text{min.} \quad & \sum_{i,j \in E^+} x_{ij} + \sum_{ij \in E^-} (1 - x_{ij}) & \\ \text{s.t.} \quad & \sum_{S \ni u} z_S = 1 & \forall u \in V \\ & \textstyle\sum_{S \ni \{u,v\}} z_S = 1 - x_{uv} & \forall (u,v) \in \tbinom{V}{2} \\ & z_S \geq 0 & \forall S \subseteq V \end{aligned} \tag{3}$$

The objective is familiar from the earliest LP formulation; the first constraint $\sum z_S = 1$ enforces that each $u \in V$ is in exactly one set S (fractionally, think of this as a normalization) and the second constraint ensures that the definition of x_{uv}, the probability u, v are in different clusters, is consistent with z-variables. The main, surprising result is that the cluster LP can be approximated efficiently (for fixed ϵ) due to its nice structure:

Theorem 2.1 (Thm 1, (CCAL$^+$24)). In time $n^{\text{poly}(1/\epsilon)}$ we can output a solution $\left(\{z_S\}_{S \subseteq V}, (x_{uv})_{uv \in \binom{V}{2}}\right)$ to the cluster LP with objective at most $(1+\epsilon)\text{opt}$.

From this approximate solution, there are two simple rounding schemes, focusing on the z or x variables; the algorithm returns the better of these. In the first, *cluster-based approach*, we select a set $S \subseteq V$ with probability given by z_S and set it aside:

```
1. Initialize clusters $\mathcal{C} \leftarrow 0$, $V' \leftarrow V$
2. while $V' \neq \emptyset$,
3.     randomly choose any cluster $S \subseteq V$ with probability $\frac{z_S}{\sum_{S'} z_{S'}}$
4.     if $V' \cap S \neq \emptyset$, then $\mathcal{C} \leftarrow C \cup \{V' \cap S\}$, $V' \leftarrow V' \backslash S$
```

Thanks to the constraints in our cluster LP, it is easy to relate the clusters that form to x values,

Lemma 2.2 (Lemma 6 (CCAL$^+$24)). For any $uv \in \binom{V}{2}$ the probability u, v are separated in the clustering output by cluster-based rounding is $\frac{2x_{uv}}{1+x_{uv}}$.

In the other, *pivot-based rounding*, we include all positive neighbors within a threshold 1/3 and amortize our error by including negative neighbors v with probability $1 - x_{uv}$.

```
1. Initialize clusters $\mathcal{C} \leftarrow 0$, $V' \leftarrow V$
2. while $V' \neq \emptyset$,
3.     randomly choose a pivot $u \in V'$, define cluster $C \leftarrow \{v \in V' \cap N^+(u) : x_{uv} \leq \frac{1}{3}\}$
4.     for each $v \in V' \cap N^-(u)$, add $v$ to $C$ with probability $1 - x_{uv}$
5.     pick some $S$ containing $u$ with probability $z_S$ (note that $\sum_{S \ni u} z_S = 1$, by constraint)
6.     Augment cluster $C \leftarrow C \cup \{S \cap V' \cap N^+(u)\}$
```

The analysis of pivot-based rounding is more detailed, but brings most of the approximation gain: for any triple of unclustered vertices $u, v, w \in V'$, one analyzes the chance that any two v, w will incur cost given that u is selected as the pivot. The peculiarity in step 4 – where we add even negative neighbors with non-zero probability – is exactly what brings better approximation guarantee than the earlier work.

2.1 Personal Thoughts

A particularly beautiful aspect of correlation clustering algorithms is that both LP rounding and purely combinatorial methods have strong (and continually improving!) results. The field seems to have progressed with the paradigms in theoretical computer science; initially totally deterministic, then introducing randomness in selecting pivots, then using rounding techniques from tractable continuous problems, and finally finding a balance between the methods.

I'm still not certain how certain steps in this current proof can be implemented efficiently, in particular, sampling from a distribution with exponential support (the z_S).[2] To just calculate $\sum_{i'} z_{i'}$ seems $O(2^n)$, so the algorithm together should take $O(2^n)$ if $\epsilon \geq \Omega(\log n/n)$, rather than $n^{O(\text{poly}(1/\epsilon))}$. Of course, if we take ϵ, arbitrarily small, then this issue vanishes; regardless, is there a nicer way of sampling? (Probably not.)

A final remark in the direction of the "qualitative" nature of this problem: there is only 1 bit of information between any two vertices, given by ±1. What if there were instead k bits? A *chromatic* correlation clustering problem has been studied in (BGGU12) – where the edges are colored $\{0, 1, 2\}$ and the objective is again to group similar color edges – but has not received as much attention. Does this problem have applications (purportedly, to protein folding)? How do the methods from vanilla correlation clustering extend?

[2] Assuming you have access to as many coin flips as needed, but the distribution is "hard to describe"; there is no more succinct definition than the tuple of z_i's.

3 Loose Connection from Clustering to RGGs

I think that clustering could also be related to a nice RGG probem, from (DGLU11).[3] In an RGG, each vertex is selected geometrically, like from the surface of the sphere, and edges encode information about pairs of vertices, in particular, having absolute inner-product larger than $\tau \in [0,1]$[4]. To make the problem of distinguishing such an RGG from the Erdös-Renyi $\mathcal{G}(n,p)$ non-trivial, we select τ such that the expected degree np is equal for both graphs. A suite of beautiful combinatorial tests are known using the simple fact of positive correlation induced by geometry:

$$\mathbf{P}\left[(u,w) \in E | (v,u) \in E, (v,w) \in E\right] > p.$$

In simple terms, 3-cliques (triangles) are more likely to form in an RGG than in a purely random graph, so simply counting how many of the $\binom{V}{3}$ vertices form a triangle can give a separation (the optimal test is a centered version called a "signed triangle" (BDER15)). However, $\mathcal{G}(n,p)$ is a strawman for correlation testing – what if *both graphs* have correlated edge structure, particularly:

Problem 3.1. Consider the RGG's induced by $\mathcal{N}(0, I_n)$ and $\mathcal{N}(0, I_n + \theta vv^\top)$ where $v \in \mathbb{R}^n$: what is an (efficient) statistical test for distinguishing?

As far as we can tell, this problem has not been studied in the literature, and intuitively, a triangle-counting approach should be weakened, as both graphs have correlated edges. I claim that perhaps a clustering metric, like the ILP in Eq. 2 (or its relaxation) may be used to distinguish such graphs – a graph drawn from spiked covariance might be more "clusterable" (in the raw data, we might see two hubs around $+kv$ and $-kv$ rather than the even spread), hence have a smaller OPT than the identity covariance. It doesn't seem likely that this would be an information theoretically optimal tester (these are usually very simple, like counting triangles) but one might expect to see some gap.

3.1 Experiment

Building RGG's drawn as described in Problem 3.1 with dimension $n = 25$ and $p = 0.5$ and various signal strengths, we find the following mean and standard deviation for the objectives: As shown in the figure, using the fractional linear program as the objective sadly fails to yield any separation, while the signed triangles still do a spectacular job. However, I still believe that some clustering metric should be effective, to be studied in future projects :).

[3] Chandra and I were discussing this problem and variant a few months ago, which I thought was naturally related to correlation clustering after learning about it. After these experiments, maybe not so much!

[4] In the sphere example, connectivity is exactly equivalent to being close in ℓ_2.

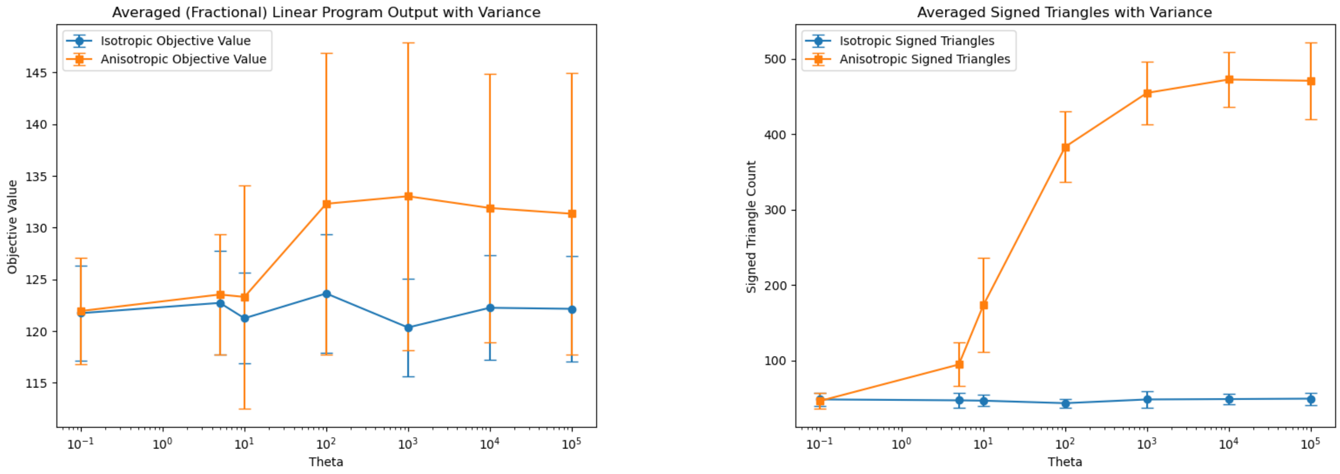

Figure 1: Plot of the Fractional LP vs. Signed Triangle testers, std. deviation plotted as error bars.

References

[ACN08] Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: Ranking and clustering. *J. ACM*, 55(5):23:1–23:27, 2008.

[BBC04] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Mach. Learn.*, 56(1–3):89–113, June 2004.

[BDER15] Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Rácz. Testing for high-dimensional geometry in random graphs, 2015.

[BGGU12] Francesco Bonchi, Aristides Gionis, Francesco Gullo, and Antti Ukkonen. Chromatic correlation clustering. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 1321–1329, New York, NY, USA, 2012. Association for Computing Machinery.

[CALLN23] Vincent Cohen-Addad, Euiwoong Lee, Shi Li, and Alantha Newman. Handling correlated rounding error via preclustering: A 1.73-approximation for correlation clustering, 2023.

[CALN23] Vincent Cohen-Addad, Euiwoong Lee, and Alantha Newman. Correlation clustering with sherali-adams, 2023.

[CCAL^{+}24] Nairen Cao, Vincent Cohen-Addad, Euiwoong Lee, Shi Li, Alantha Newman, and Lukas Vogl. Understanding the cluster linear program for correlation clustering. In

Proceedings of the 56th Annual ACM Symposium on Theory of Computing, STOC '24, page 1605–1616. ACM, June 2024.

[CGW03] Moses Charikar, Venkatesan Guruswami, and Anthony Wirth. Clustering with qualitative information. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '03, page 524, USA, 2003. IEEE Computer Society.

[CMSY15] Shuchi Chawla, Konstantin Makarychev, Tselil Schramm, and Grigory Yaroslavtsev. Near optimal lp rounding algorithm for correlationclustering on complete and complete k-partite graphs. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 219–228, New York, NY, USA, 2015. Association for Computing Machinery.

[DGLU11] Luc Devroye, András György, Gábor Lugosi, and Frederic Udina. High-Dimensional Random Geometric Graphs and their Clique Number. *Electronic Journal of Probability*, 16(none):2481 – 2508, 2011.