Recent Progress on Correlation Clustering: From Local Algorithms to Better Approximation Algorithms and Back

Vincent Cohen-Addad
Google Research, Paris, France

Abstract

Correlation clustering is a classic model for clustering problems arising in machine learning and data mining. Given a set of data elements represented as vertices of a graph and pairwise similarity represented as edges, the goal is to find a partition of the vertex set so as to minimize the total number of edges across the parts plus the total number of non-edges within the parts.

Introduced in the early 2000s [2], correlation clustering has received a large amount of attention through the years. A natural linear programming relaxation was shown to have an integrality gap of at least 2 and at most 2.5 [1] in 2005, and in 2015 at most 2.06 [4].

In 2021, motivated by large-scale application new structural insights allowed to derive a simple, practical algorithm that achieved an $O(1)$-approximation in a variety of models (Massively Parallel, Sublinear, Streaming or Differentially-private) [6, 5]. These new insights turned out to be a key building block in designing better algorithms: It serves as a pre-clustering of the input graph that enables algorithm with approximation guarantees significantly better than 2 [7, 8]. It is a key component in the new algorithm that achieves a 1.44-approximation [3] and in the new local-search based 1.84-approximation for the Massively Parallel, Sublinear, and Streaming models [9]. This talk will review the above recent development and what are the main open research directions.

A collection of joint works with Nairen Cao, Silvio Lattanzi, Euiwoong Lee, Shi Li, David Rasmussen Lolck, Slobodan Mitrovic, Alantha Newman, Ashkan Norouzi-Fard, Nikos Parotsidis, Marcin Pilipczuk, Jakub Tarnawski, Mikkel Thorup, Lukas Vogl, Shuyi Yan, Hanwen Zhang.

2012 ACM Subject Classification Theory of computation → Design and analysis of algorithms; Theory of computation → Approximation algorithms analysis; Theory of computation → Theory and algorithms for application domains; Theory of computation → Machine learning theory; Theory of computation → Unsupervised learning and clustering

Keywords and phrases Approximation Algorithms, Clustering, Local Model

Digital Object Identifier 10.4230/LIPIcs.ESA.2024.1

Category Invited Talk

References

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

2 Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine learning*, 56(1):89–113, 2004.

3 Nairen Cao, Vincent Cohen-Addad, Euiwoong Lee, Shi Li, Alantha Newman, and Lukas Vogl. Understanding the cluster linear program for correlation clustering. In Bojan Mohar, Igor Shinkar, and Ryan O'Donnell, editors, *Proceedings of the 56th Annual ACM Symposium on Theory of Computing, STOC 2024, Vancouver, BC, Canada, June 24-28, 2024*, pages 1605–1616. ACM, 2024. doi:10.1145/3618260.3649749.

4 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 *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC)*, pages 219–228, 2015.

5 Vincent Cohen-Addad, Chenglin Fan, Silvio Lattanzi, Slobodan Mitrovic, Ashkan Norouzi-Fard, Nikos Parotsidis, and Jakub M Tarnawski. Near-optimal correlation clustering with privacy. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 33702–33715. Curran Associates, Inc., 2022. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/da645920dcd3bd35b0dae329894bad80-Paper-Conference.pdf.

6 Vincent Cohen-Addad, Silvio Lattanzi, Slobodan Mitrovic, Ashkan Norouzi-Fard, Nikos Parotsidis, and Jakub Tarnawski. Correlation clustering in constant many parallel rounds. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 2069–2078, 2021.

7 Vincent Cohen-Addad, Euiwoong Lee, Shi Li, and Alantha Newman. Handling correlated rounding error via preclustering: A 1.73-approximation for correlation clustering. In *64th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2023, Santa Cruz, CA, USA, November 6-9, 2023*, pages 1082–1104. IEEE, 2023. doi:10.1109/FOCS57990.2023.00065.

8 Vincent Cohen-Addad, Euiwoong Lee, and Alantha Newman. Correlation clustering with sherali-adams. In *63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2022, Denver, CO, USA, October 31 - November 3, 2022*, pages 651–661. IEEE, 2022. doi:10.1109/FOCS54457.2022.00068.

9 Vincent Cohen-Addad, David Rasmussen Lolck, Marcin Pilipczuk, Mikkel Thorup, Shuyi Yan, and Hanwen Zhang. Combinatorial correlation clustering. In Bojan Mohar, Igor Shinkar, and Ryan O'Donnell, editors, *Proceedings of the 56th Annual ACM Symposium on Theory of Computing, STOC 2024, Vancouver, BC, Canada, June 24-28, 2024*, pages 1617–1628. ACM, 2024. doi:10.1145/3618260.3649712.