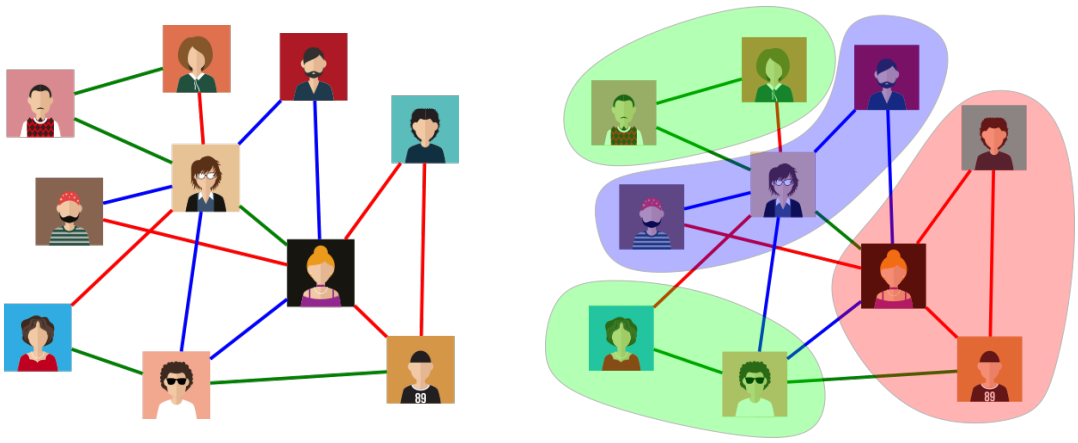
Algorithm Engineering Group
Prof. Dr. Tobias Friedrich
Master Project Winter Term 2020/21

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Clustering in networks with more than one type of similarity

Motivation Clustering is one of the most important computing tasks for data analysis. Generally speaking, the aim is to find a partition such that similar objects are placed in the same cluster and dissimilar objects appear in different clusters. One of the most practically used concrete methods for clustering is correlation clustering [1], where objects are represented as vertices of a graph and edges model similarity. The task then is to find a partition of the vertices that minimizes the sum of the number of non-edges within clusters and the number of edges across clusters (so-called *disagreement*).



A social network example with 3 types of friendships: family, schoolmates and colleagues.

There are many applications where similarity is categorical rather than binary. Consider, for example, a social network where two people have an edge between them if they are *friends*. But there are many categories of *friends* such as *family*, *colleagues*, *schoolmates*, *neighbors* etc. For a good clustering of such data, it is desirable to have mostly edges of similar *type* in a cluster. To capture this categorical similarity, Bonchi et al. [2] introduced *chromatic correlation clustering*, where each edge has a color associated with it, and as output one has to then also assign a color to each cluster.

Chromatic correlation clustering has applications from a variety of fields such as community detection in social network analysis, classifying proteins from protein-protein interaction networks in bioinformatics, and entity de-duplication in data-mining. Since finding the best clustering is NP-hard (already without colors), we focus on developing efficient algorithms that provide good approximate solutions.

The goal of this project In this project, we plan to study chromatic correlation clustering from the viewpoints of both theoretical and applied research. We have already developed a preliminary algorithm that gives a theoretical approximation guarantee that matches the currently best [3] while having a much better runtime. The starting point of the project is to push this further and develop an algorithm that gives even better theoretical guarantees, or to conversely prove lower bounds that show the impossibility of such improvement. Aside from theoretical upper and lower bounds we plan to implement the final version of this algorithm and test it on real-world data-sets. A further interesting step could be to develop heuristics to make the algorithm work better in practice.

Further, there are many interesting variants of correlation clustering for specific applications and we plan to extend the chromatic clustering framework to some of these. Candidates for such extensions are an edge-weighted setting, graphs with non-relevant pairs of vertices, overlapping clusters, edges with multiple colors, maximum agreement instead of minimum disagreement, and more.

What we expect from you An aptitude for doing theoretical research in a relevant topic of Graph Theory. Curiosity to investigate the existing models and creativity to extend the current framework. Some basic coding skills to complement theoretical results with experiments.

What you can expect from us We will gently introduce you to the field and accompany you all along this interesting journey. If circumstances permit, our weekly meeting will be in person, otherwise via Zoom. This will be a team effort, and we aim at publishing our results at a renowned international conference.

How to contact us You're welcome to visit us virtually or on floor A-1:

Prof. Dr. Tobias Friedrich
Dr. Davis Issac
Dr. Katrin Casel
} firstname.lastname@hpi.de

[1] Bansal, N., Blum, A., Chawla, S. (2004). Correlation Clustering. Machine Learning, 56(1-3), 89-113. https://link.springer.com/content/pdf/10.1023/B:MACH.0000033116.57574.95.pdf

[2] Bonchi, F., Gionis, A., Gullo, F., Tsourakakis, C. E., Ukkonen, A. (2015). Chromatic Correlation Clustering. ACM Transactions on Knowledge Discovery from Data (TKDD), 9(4), 1-24. http://people.seas.harvard.edu/~babis/ccc.pdf

[3] Anava, Y., Avigdor-Elgrabli, N., Gamzu, I. (2015). Improved theoretical and practical guarantees for chromatic correlation clustering. In Proceedings of the 24th International Conference on World Wide Web (pp. 55-65). https://dl.acm.org/doi/10.1145/2736277.2741629