Fusion Moves for Correlation Clustering

Thorsten Beier[1], Fred A. Hamprecht[2], Jörg H. Kappes[3]
[1]thorsten.beier@iwr.uni-heidelberg.de [2]fred.hamprecht@iwr.uni-heidelberg.de [3]kappes@math.uni-heidelberg.de

Correlation clustering, or multicut partitioning, is widely used in image segmentation for partitioning an undirected graph or image with positive and negative edge weights such that the sum of cut edge weights is minimized. Due to its NP-hardness, exact solvers do not scale and approximative solvers often give unsatisfactory results. We investigate scalable methods for correlation clustering. To this end we define fusion moves for the correlation clustering problem.

Our algorithm iteratively fuses the current and a proposed partitioning which monotonously improves the partitioning and maintains a valid partitioning at all times. Furthermore, it scales to larger datasets, gives near optimal solutions, and at the same time shows a good anytime performance.

Correlation clustering [6], also known as the multicut problem [10] is a basic primitive in computer vision [2, 3, 4, 16] and data mining [5, 8, 9, 15].

Its merit is, firstly, that it accommodates both positive (attractive) *and* negative (repulsive) edge weights. This allows doing justice to evidence in the data that two nodes or pixels do not wish or do wish to end up in the same cluster or segment, respectively. Secondly, it does not require a specification of the number of clusters beforehand.

In signed social networks, where positive and negative edges encode friend and foe relationships, respectively, correlation clustering is a natural way to detect communities [8, 9]. Correlation clustering can also be used to cluster query refinements in web search [15]. Because social and web-related networks are often huge, heuristic methods, the PIVOT-algorithm [1], are popular [9].

In computer vision applications, unsupervised image segmentation algorithms often start with an over-segmentation into superpixels (superregions), which are then clustered into "perceptually meaningful" regions by correlation clustering. Such an approach has been shown to yield state-of-the-art results on the Berkeley Segmentation Database [2, 3, 12, 16].

While it has a clear mathematical formulation and nice properties, correlation clustering suffers from NP-hardness. Consequently, partition problems on large scale data, huge volume images in computational neuroscience [4] or social networks [14], are not tractable because reasonable solutions cannot be computed in acceptable time.

Contribution. In this work we present novel approaches that are designed for large scale correlation clustering problems. First, we define a novel energy based agglomerative clustering algorithm that monotonically increases the energy. With this at hand we show how to improve the anytime performance of Cut, Clue & Cut [7]. Second, we improve the anytime performance of polyhedral multicut methods [11] by more efficient separation procedures. Third, we introduce cluster-fusion moves, which extend the original fusion moves [13] used in supervised segmentation to the unsupervised case and give a polyhedral interpretation of this algorithm. Finally, we propose two versatile proposal generators, and evaluate the proposed methods on existing and new benchmark problems. Experiments show that we can improve the computation time by one to two magnitudes without worsening the segmentation quality significantly.

[1] Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: Ranking and clustering. *J. ACM*, 55(5):23:1–23:27, November 2008. ISSN 0004-5411. doi: 10.1145/1411509.1411513. URL http://doi.acm.org/10.1145/1411509.1411513.

[2] Amir Alush and Jacob Goldberger. Break and conquer: Efficient correlation clustering for image segmentation. In *2nd International Workshop on Similarity-Based Pattern Analysis and Recognition*, 2013.

[3] Bjoern Andres, Jörg H Kappes, Thorsten Beier, Ullrich Köthe, and Fred A Hamprecht. Probabilistic image segmentation with closedness constraints. In *ICCV*, pages 2611–2618. IEEE, 2011.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

[4] Bjoern Andres, Thorben Kroeger, Kevin L Briggman, Winfried Denk, Natalya Korogod, Graham Knott, Ullrich Koethe, and Fred A Hamprecht. Globally optimal closed-surface segmentation for connectomics. In *ECCV*, pages 778–791. Springer, 2012.

[5] Arvind Arasu, Christopher Re, and Dan Suciu. Large-scale deduplication with constraints using dedupalog. In *Proceedings of the 25th International Conference on Data Engineering, ICDE 2009*. IEEE Computer Society, March 2009. URL http://research.microsoft.com/apps/pubs/default.aspx?id=77606.

[6] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. In *MACHINE LEARNING*, pages 238–247, 2002.

[7] Thorsten Beier, Thorben Kroeger, Jörg Hendrik Kappes, Ullrich Koethe, and Fred A. Hamprecht. Cut, Glue & Cut: A Fast, Approximate Solver for Multicut Partitioning. In *IEEE Conference on Computer Vision and Pattern Recognition 2014*, 2014.

[8] Yudong Chen, Sujay Sanghavi, and Huan Xu. Clustering sparse graphs. In *NIPS*, pages 2213–2221, 2012.

[9] Flavio Chierichetti, Nilesh N. Dalvi, and Ravi Kumar. Correlation clustering in mapreduce. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 641–650, 2014. doi: 10.1145/2623330.2623743. URL http://doi.acm.org/10.1145/2623330.2623743.

[10] Sunil Chopra and M.R. Rao. The partition problem. *Mathematical Programming*, 59(1-3):87–115, 1993. ISSN 0025-5610. doi: 10.1007/BF01581239. URL http://dx.doi.org/10.1007/BF01581239.

[11] Jörg Hendrik Kappes, Markus Speth, Gerhard Reinelt, and Christoph Schnörr. Higher-order segmentation via multicuts. *CoRR*, abs/1305.6387, 2013.

[12] Sungwoong Kim, Sebastian Nowozin, Pushmeet Kohli, and Chang Dong Yoo. Higher-order correlation clustering for image segmentation. In *NIPS*, pages 1530–1538, 2011.

[13] V. Lempitsky, C. Rother, S. Roth, and A. Blake. Fusion moves for Markov random field optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1392–1405, aug 2010.

[14] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1361–1370, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-929-9.

[15] Eldar Sadikov, Jayant Madhavan, Lu Wang, and Alon Halevy. Clustering query refinements by user intent. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 841–850, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772776. URL http://doi.acm.org/10.1145/1772690.1772776.

[16] Julian Yarkony, Alexander Ihler, and Charless C Fowlkes. Fast planar correlation clustering for image segmentation. In *ECCV*. Springer, 2012.