9^{th} International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

Heuristic Modularity Maximization Algorithms for Community Detection Rarely Return an Optimal Partition or Anything Similar

Keywords: Network science, community detection, modularity maximization, integer programming, optimization

Extended Abstract

Community detection is a classic problem in network science with extensive applications in various fields (Fortunato and M. E. Newman) including computational social science. The most commonly used methods are the algorithms designed to maximize modularity across different partitions of nodes into communities. We investigate the extent to which current heuristic modularity maximization algorithms succeed in maximizing modularity. We measure their performance in terms of (1) the fraction of their output modularity value for a graph over the maximum modularity of that graph and (2) the similarity between their output partition and a modularity-maximizing (optimal) partition. We obtain optimal partitions by solving an integer programming model of modularity maximization (Dinh and Thai) to global optimality. This evaluation includes eight modularity maximization heuristics known as Clauset-Newman-Moore (CNM) (Clauset et al.), Louvain (Blondel et al.), Leicht-Newman (LN) (Leicht and M. E. J. Newman), Combo (Sobolevsky et al.), Belief (Zhang and Moore), Paris (Bonald et al.), Leiden (Traag et al.), and EdMot (Li et al.). For this evaluation, we include 71 small real networkswith no more than 2812 edges edges as well as 10 Erdős-Rényi graphs and 10 Barabási-Albert graphs.

To quantify proximity to optimality, we define and use the *Global Optimality Percentage* (GOP) as the fraction of the modularity returned by a heuristic method for a network over the globally maximum modularity for that network[1]. We also use a quantitative measure for similarity of a partition to an optimal partition. Normalized Adjusted Mutual Information (AMI) (Vinh et al.) is a measure of similarity of partitions which (unlike normalized mutual information) adjusts the measurement based on the similarity that two partitions may have by pure chance. Values of AMI close to 1 indicate high similarity between the two partitions.

Figure 1 shows GOP on the y-axis and AMI on the x-axis for each of the eight modularity maximization algorithms (represented by colors). For each algorithm, there are 71 data points for the 71 real networks and 2 data points for the average of 10 Erdős-Rényi and the average of 10 Barabási-Albert graphs. The 45-degree line shows cases where the optimality gap ($1-GOP$) is associated with an equal dissimilarity ($1-AMI$) between the sub-optimal partition and an optimal partition. Looking at the y-axis values in Figure 1, we observe that there is a substantial variation in the extent of sub-optimality for the partitions obtained by the heuristic algorithms. Overall, the three algorithms with highest and increasing performance in returning close-to-maximum modularity values are LN, Leiden, and combo respectively. Despite that these instance are graphs with no more than 2812 edges, they are non-trivial instances for these heuristic algorithms. Given that modularity maximization is an NP-hard problem, one can argue that the performance of these heuristic methods in term of proximity to an optimal solution

[1]In cases where a heuristic algorithm returns a partition with a negative modularity value, we set GOP$=0$.

does not improve for larger networks. The x-axis values in Figure 1 show the considerable dissimilarity between the sub-optimal partitions and an optimal partition for these 91 instances.

Focusing on the position of data points, we observe that they are mostly above the 45-degree line. This indicates that near optimal solutions tend to have partitions disproportionally dissimilar to an optimal partition. This result goes against the naive viewpoint that close-to-maximum modularity partitions are also close to an optimal partition. We also report the fraction of networks (out of 91) for which a given heuristic algorithm returns an optimal solution. Combo has the highest success rate, returning an optimal partition for 51.6% of the networks. LN and Leiden maximize modularity for 23% of the networks considered. Louvain has a success rate of 16.4%. The algorithms CNM, EdMot, Paris, and Belief have success rates of 4%, 2%, 1%, and 0% respectively. Taken together, our results indicate a crucial mismatch between the design philosophy of modularity maximization algorithms for community detection and the results they provide: Heuristic modularity maximization algorithms rarely return an optimal partition or a partition resembling an optimal partition. Given these results, developing a community detection algorithm based on exact or approximate maximization of modularity seems to be a justified approach for improving a fundamental method in computational social science.

References

Blondel, Vincent D., et al. "Fast unfolding of communities in large networks". *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, 2008, Publisher: IOP Publishing, P10008.

Bonald, Thomas, et al. "Hierarchical Graph Clustering using Node Pair Sampling". *MLG 2018 - 14th International Workshop on Mining and Learning with Graphs*, 2018.

Clauset, Aaron, et al. "Finding community structure in very large networks". *Physical review E*, vol. 70, no. 6, 2004, Publisher: APS, p. 066111.

Dinh, Thang N. and My T. Thai. "Toward optimal community detection: From trees to general weighted networks". *Internet Mathematics*, vol. 11, no. 3, 2015, Publisher: Taylor & Francis, pp. 181–200.

Fortunato, Santo and Mark EJ Newman. "20 years of network community detection". *Nature Physics*, vol. 18, 2022, pp. 848–850.

Leicht, E. A. and M. E. J. Newman. "Community Structure in Directed Networks". *Physical Review Letters*, vol. 100, no. 11, Mar. 2008, Publisher: American Physical Society, p. 118703. doi:10.1103/PhysRevLett.100.118703.

Li, Pei-Zhen, et al. "Edmot: An edge enhancement approach for motif-aware community detection". *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 479–487.

Sobolevsky, Stanislav, et al. "General optimization technique for high-quality community detection in complex networks". *Physical Review E*, vol. 90, no. 1, 2014, p. 012811.

Traag, V. A., et al. "From Louvain to Leiden: guaranteeing well-connected communities". *Scientific Reports*, vol. 9, no. 1, 2019. doi:10.1038/s41598-019-41695-z.

Vinh, Nguyen Xuan, et al. "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance". *Journal of Machine Learning Research*, vol. 11, no. 95, 2010, pp. 2837–2854. jmlr.org/papers/v11/vinh10a.html.

Zhang, Pan and Cristopher Moore. "Scalable detection of statistically significant communities and hierarchies, using message passing for modularity". *Proceedings of the National Academy of Sciences*, vol. 111, no. 51, 2014, pp. 18144–18149.

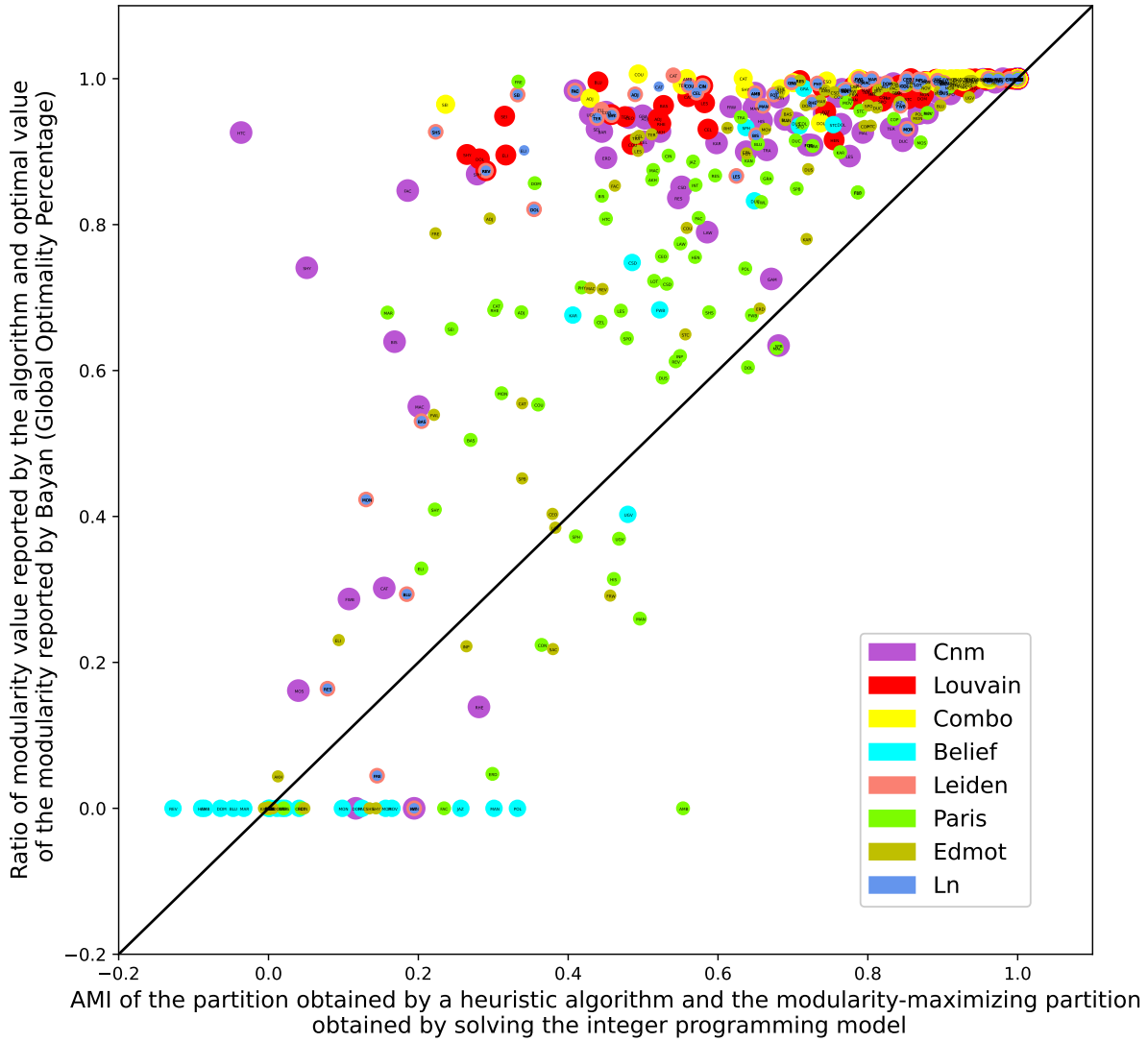

Figure 1: Global Optimality percentage and normalized adjusted mutual information measured for eight modularity maximization heuristic methods in comparison with globally optimal partitions. Magnify the high-resolution figure on screen for names of networks an other details. The 71 real networks used in this figure are publicly available from the network repository Netzschleuder with the following names: dom, packet_delays, sa_companies, ambassador, florentine_families, rhesus_monkey, kangaroo, internet_top_pop, high_tech_company, moviegalaxies, november17, moreno_taro, sp_baboons, bison, dutch_school, zebras, cattle, moreno_sheep, 7th_graders, college_freshmen, hens, freshmen, karate, dutch_criticism, montreal, ceo_club, windsurfers, elite, macaque_neural, sp_kenyan_households, contiguous_usa, cs_department, dolphins, macaques, terrorists_911, train_terrorists, highschool, law_firm, baseball, blumenau_drug, sp_hospital, lesmis, fresh_webs, sp_office, swingers, polbooks, game_thrones, adjnoun, sp_hypertext, football, football_tsevans, sp_high_school_new, foodweb_baywet, revolution, foodweb_little_rock, student_cooperation, jazz_collab, ugandan_village, cintestinalis, faculty_hiring, interactome_pdz, residence_hall, physician_trust, contact, celegansneural, malaria_genes, marvel_partnerships, facebook_friends, london_transport, copenhagen, netscience.