Efficient Quantitative Heuristics for Graph Clustering

Rafael Santiago
Universidade do Vale do Itajaí
Itajaí, SC, Brazil 88302-901
rsantiago@univali.br

Luís C. Lamb
Federal University of Rio Grande do Sul
Porto Alegre, RS, Brazil 91501-970
lamb@inf.ufrgs.br

ABSTRACT

Modularity Density Maximization (MDM) aims at the identification of modular clusters in graphs. In this paper, we introduce a quantitative heuristic called HLSMDM-λ, which solves the MDM problem in large graphs. We compared it with state-of-the-art results of exact and heuristic methods such as MDB2, GAOD, iMeme-Net, HAIN, divisive BMD-λ, MCN-λ, CNM, and Louvain. The largest tested graphs were executed very efficiently. Our results show that HLSMDM-λ is scalable in terms of time and is able to find partitions with the highest objective value for the largest tested graphs. HLSMDM-λ also produced effective results in a ground truth analysis. These results point out that HLSMDM-λ is a state-of-the-art heuristic for the MDM problem.

CCS CONCEPTS

•**Computing methodologies** → **Search methodologies;**

KEYWORDS

Graph Clustering, Modularity Density Maximization, Large Graphs

ACM Reference format:
Rafael Santiago and Luís C. Lamb. 2017. Efficient Quantitative Heuristics for Graph Clustering. In *Proceedings of GECCO '17 Companion, Berlin, Germany, July 15-19, 2017,* 2 pages.
DOI: http://dx.doi.org/10.1145/3067695.3075995

1 INTRODUCTION

Modularity Density Maximization (MDM) is an optimization graph clustering problem which finds the best partition by using an objective function that measures (or quantifies) the difference between the internal and the external connectivity of each cluster [9]. The MDM objective function is presented in Equation (1), where C is a partition of disjoint clusters, E_c is the set of edges which connects two nodes of the cluster c, and d_v is the degree of a node $v \in V$. This function uses the number of nodes within the cluster, not the total number of edges, so avoiding the resolution limit [5] present in Modularity Maximization (MM) [11]. This function is used to obtain the "ratio association" to find small clusters when $\lambda < 0.5$, and the "ratio cut" to find large clusters when $\lambda > 0.5$. Li et al. [9] suggest that this function can be used to find the appropriate level of topological structure of graphs to find proper partitions.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
GECCO '17 Companion, Berlin, Germany
© 2017 Copyright held by the owner/author(s). 978-1-4503-4939-0/17/07…$15.00
DOI: http://dx.doi.org/10.1145/3067695.3075995

$$D_\lambda(C) = \sum_{c \in C} \left(\frac{4\lambda|E_c| - (2-2\lambda)\left(\sum_{v \in c} d_v - 2|E_c|\right)}{|c|} \right) \quad (1)$$

Some heuristics have been developed for MDM. One can cite five efforts: (i) the metaheuristics genetic algorithm GAOD [10], (ii) the memetic algorithm iMeme-Net [6], (iii) the hybrid artificial immune heuristic HAIN [7]; (iv) the eight divisive heuristics of [4]; and the seven constructive and multilevel heuristics for graphs with more than 300,000 nodes in [13]. Other efforts have been made towards yielding optimal partitions. A non-linear model was proposed in [9]. This model was improved by [7]. By transforming the model of [9], Costa [3] developed a linear model, which solved instances with up to 40 nodes.

In this context, our paper presents one hybrid local search heuristics for MDM that use the quantitative ratio λ of Equation (1) and compares it to the MCN heuristic of Santiago and Lamb [13]. Our hybrid heuristic is the HLSMDM-λ heuristic for the MDM problem that was inspired in the multilevel heuristics reported in [12]. We compared our heuristics with CNM [2] and Louvain [1] heuristics because they are heuristics scalable to hundreds of thousands of nodes for the MM problem.

2 QUANTITATIVE HYBRID LOCAL SEARCH

The Hybrid Local Search heuristic (HLSMDM-λ) for MDM is illustrated in Algorithm 1. The parameter $G(V, E)$ is an undirected, unweighted graph. The initial solution *part* is generated by the constructive search of Coarsening Merger (CM) from [13]. HLSMDM-λ is composed of two local search phases. In the first phase, a local search called MCN-λ from [13] is applied. When the second phase is run, a second local search called LNM-λ from [13] is performed on the partition resulted by the MCN-λ.

Algorithm 1: HLSMDM-λ

```
Input  : G(V, E), λ
part ← CM(G, part, λ) // from [13]
first ← MCN-λ(G, part, λ) // from [13]
second ← LNM-λ(G, first, λ) // from [13]
if D_λ(first) > D_λ(second) then
    return first
else
    return second
```

3 EXPERIMENTS AND RESULTS

The experiments were performed on a PC with an Intel Core i7 64 bits with 3.40GHz with 8192KB of cache memory and 8GB of

RAM, under Linux Ubuntu 14.04.1 LTS operating system. Each experiment was run by using a single thread. The language used was C++, with the GCC compiler.

Experiments with the same 33 instances used in [13] were performed. The heuristic HLSMDM-λ demanded more time than Louvain and MCN-λ after $5,000$ edges. In [7], it was shown that GAOD, iMeme-Net, HAIN found solutions in 663, 78, and 368 seconds for instances with less than $3,000$ edges, respectively. The MCN-λ and HLSMDM-λ found solutions for graphs with at most $100,000$ edges in less than 7 seconds.

In ground truth analysis, we submitted random graphs generated by the LFR framework [8] to the heuristics CNM, Louvain, MCN-λ and HLSMDM-λ. In these graphs, we know the expected clustering, so that we can perform a ground truth analysis. In MCN-λ and HLSMDM-λ heuristics, the λ ratios tested were $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

All graphs were created with $100,000$ nodes, average degree 15, maximum degree 50, minus exponent for the degree sequence equal to 2, minus exponent for the community size distribution equal to 1. The mixing parameter used was $\mu \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$. This parameter was used to generate problems with increasing difficulty because it defines the connection between nodes from different expected clusters. The higher the parameter is; the weaker the modular property of the clusters get.

Table 1 shows the "Matthews Correlation Coefficient" (ϕ value). The larger the coefficient ϕ is, the stronger the correlation between the partition obtained by the heuristic and the correct partition is. The results show the importance of the correct choosing λ parameters. HLSMDM-λ results suggest that this heuristic works better when the clustering have a structure with a strong modular property. MCN-λ was the only heuristic that found the clustering structures for instances with $\mu > 0.3$.

Table 1: Comparisons among ϕ values obtained

μ	CNM	Louvain	λ	MCN-λ	HLSMDM-λ	μ	CNM	Louvain	λ	MCN-λ	HLSMDM-λ
0.1	.274	.362	0.1	.916	**.99**	0.4	.042	.255	0.1	.543	.0
			0.2	.928	**.99**				0.2	.598	.0
			0.3	.937	**.99**				0.3	.656	.0
			0.4	.944	**.99**				0.4	.709	.0
			0.5	.941	**.99**				0.5	.728	.0
			0.6	.929	**.99**				0.6	**.766**	.0
			0.7	.91	**.99**				0.7	.75	.0
			0.8	.889	**.99**				0.8	.704	.0
			0.9	.847	**.99**				0.9	.644	.0
0.2	.096	.324	0.1	.825	**.985**	0.5	.032	.221	0.1	.217	.0
			0.2	.854	**.985**				0.2	.368	.0
			0.3	.87	**.985**				0.3	.502	.0
			0.4	.883	**.985**				0.4	.569	.0
			0.5	.886	.984				0.5	.608	.0
			0.6	.873	.984				0.6	.664	.0
			0.7	.848	.984				0.7	**.691**	.0
			0.8	.812	**.985**				0.8	.647	.0
			0.9	.761	**.985**				0.9	.561	.0
0.3	.056	.288	0.1	.703	**.973**	0.6	.022	.183	0.1	.0	.0
			0.2	.748	**.973**				0.2	.0	.0
			0.3	.785	.972				0.3	.004	.0
			0.4	.804	.972				0.4	.324	.0
			0.5	.826	.972				0.5	.444	.0
			0.6	.826	.972				0.6	.504	.0
			0.7	.798	.972				0.7	**.56**	.0
			0.8	.756	.972				0.8	.554	.0
			0.9	.717	.972				0.9	.45	.0

4 CONCLUSIONS

This paper introduces a hybrid quantitative local search for the Modularity Density Maximization problem, called HLSMDM-λ. The proposed method was compared with GAOD, iMeme-Net, HAIN, and MCN-λ MDM heuristics, and CNM, and Louvain MM heuristics. Our results suggest that MCN-λ and HLSMDM-λ are scalable to hundreds of thousands of nodes. Ground truth analysis showed that they reached the closest expected results when comparing to the correct partitions for the most of the random graphs.

For MCN-λ and HLSMDM-λ, the largest tested graphs were executed in less than 10 minutes. Finally, we can state that the reported results suggest that MCN-λ and HLSMDM-λ are state-of-the-art quantitative heuristics for the MDM problem in terms of time.

The results also suggest that the correct usage of the quantitative ratio λ can help to reach results closer to the expected partition than using the standard value $\lambda = 0.5$.

As further investigations, we suggest the use of MCN-λ and HLSMDM-λ as exploitation in other metaheuristics.

ACKNOWLEDGEMENTS

This work is partly supported by the Brazilian Research Council CNPq and the *University of Vale do Itajaí.*

REFERENCES

[1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (Oct. 2008), P10008. DOI: http://dx.doi.org/10.1088/1742-5468/2008/10/P10008

[2] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. 2004. Finding community structure in very large networks. *Physical Review E* 70, 6 (Dec. 2004), 066111-1 – 066111-6. DOI: http://dx.doi.org/10.1103/PhysRevE.70.066111 arXiv:arXiv:cond-mat/0408187v2

[3] Alberto Costa. 2015. MILP formulations for the modularity density maximization problem. *European Journal of Operational Research* 245, 1 (2015), 14–21. DOI: http://dx.doi.org/10.1016/j.ejor.2015.03.012

[4] Alberto Costa, Sergey Kushnarev, Leo Liberti, and Zeyu Sun. 2016. Divisive heuristic for modularity density maximization. *Computers & Operations Research* 71 (jul 2016), 100–109. DOI: http://dx.doi.org/10.1016/j.cor.2016.01.009

[5] Santo Fortunato and Marc Barthélemy. 2007. Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America* 104, 1 (Jan. 2007), 36–41. DOI: http://dx.doi.org/10.1073/pnas.0605965104 arXiv:arXiv:physics/0607100v2

[6] Maoguo Gong, Qing Cai, Yangyang Li, and Jingjing Ma. 2012. An improved memetic algorithm for community detection in complex networks. In *2012 IEEE Congress on Evolutionary Computation*. IEEE, Brisbane, QLD, 1–8. DOI: http://dx.doi.org/10.1109/CEC.2012.6252971

[7] Amir-Mohsen Karimi-Majd, Mohammad Fathian, and Babak Amiri. 2014. A hybrid artificial immune network for detecting communities in complex networks. *Computing* 97, 5 (2014), 483–507. DOI: http://dx.doi.org/10.1007/s00607-014-0433-6

[8] Andrea Lancichinetti, Filippo Radicchi, José J Ramasco, and Santo Fortunato. 2011. Finding statistically significant communities in networks. *PloS one* 6, 4 (Jan. 2011), e18961. DOI: http://dx.doi.org/10.1371/journal.pone.0018961

[9] Zhenping Li, Shihua Zhang, Rui-Sheng Wang, Xiang-Sun Zhang, and Luonan Chen. 2008. Quantitative function for community detection. *Physical Review E* 77, 3 (March 2008), 036109. DOI: http://dx.doi.org/10.1103/PhysRevE.77.036109

[10] Jinxia Liu and Jianchao Zeng. 2010. Community detection based on modularity density and genetic algorithm. *Proceedings - International Conference on Computational Aspects of Social Networks, CASoN'10* (2010), 29–32. DOI: http://dx.doi.org/10.1109/CASoN.2010.14

[11] M. Newman and M. Girvan. 2004. Finding and evaluating community structure in networks. *Physical Review E* 69, 2 (Feb. 2004), 026113. DOI: http://dx.doi.org/10.1103/PhysRevE.69.026113

[12] Randolf Rotta and Andreas Noack. 2011. Multilevel local search algorithms for modularity clustering. *Journal of Experimental Algorithmics* 16, 2 (May 2011), 2.1. DOI: http://dx.doi.org/10.1145/1963190.1970376

[13] Rafael Santiago and Luís C. Lamb. 2017. Efficient modularity density heuristics for large graphs. *European Journal of Operational Research* 258, 3 (May 2017), 844–865. DOI: http://dx.doi.org/10.1016/j.ejor.2016.10.033