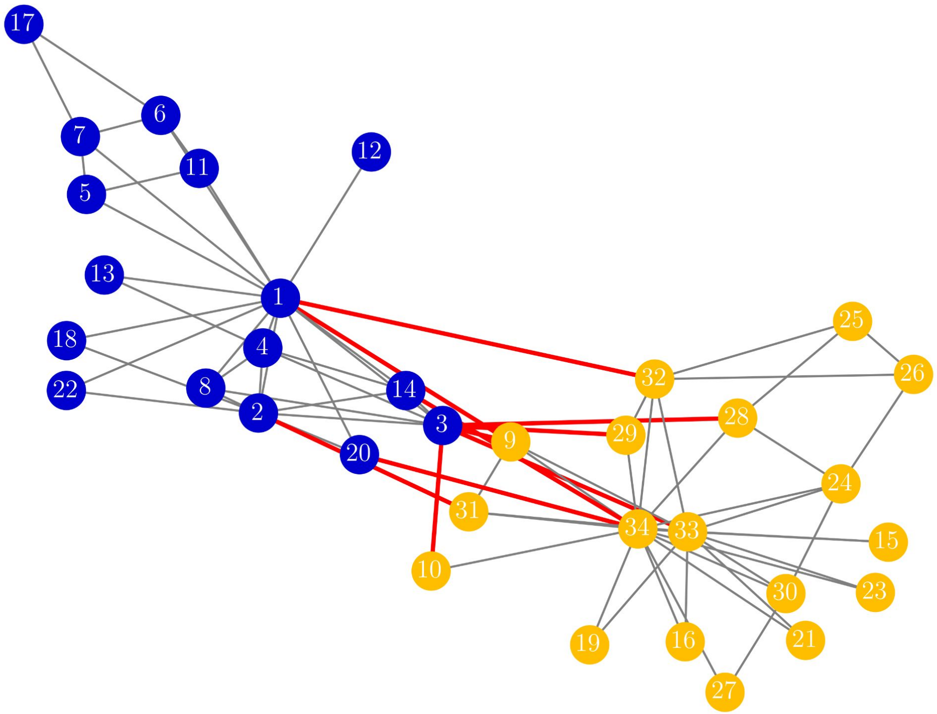
Practical Almost-Linear-Time Approximation Algorithms for Hybrid and Overlapping Graph Clustering

Konstantinos Ameranis, Lorenzo Orecchia, Kunal Talwar, Charalampos Tsourakakis

Ratio Cuts


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34

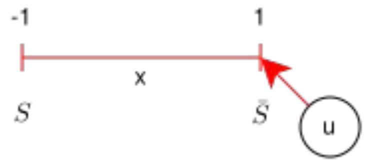
Definitions

$$G(V, E \subseteq V \times V, \mu, w)$$

$\lvert V \rvert = n$	Nodes
$\lvert E \rvert = m$	Edges
$\mu \in \mathbb{R}_+^{\lvert V \rvert}$	Non-negative node weights
$w \in \mathbb{R}_+^{\lvert E \rvert}$	Non-negative edge weights
$W = diag(w)$	Edge weights Matrix
$B \in \mathbb{R}^{m \times n} : \; B_{uv} = e_v - e_u$	Incidence matrix
$L = B^T W B$	Laplacian
$\mathcal{L} = diag(\mu)^{-1/2} \cdot L \cdot diag(\mu)^{-1/2}$	Normalized Laplacian

Ratio-Cut problem definition

$$\min_{x \in [-1,1]^n \perp \vec{1}} \frac{\|Bx\|_{1,w}}{\min_u \|x - u\vec{1}\|_{1,\mu}} = \min_{x \in [-1,1]^n \perp \vec{1}} \frac{\sum_{uv \in E} w_{uv} \cdot |x_u - x_v|}{\min_u \|x - u\vec{1}\|_{1,\mu}}$$



$$\min_u \|x - u\mathbf{1}\|_1 = \min\left(vol(S), vol(\bar{S})\right)$$

$$\min_{S \subset V} \frac{E(S, \bar{S})}{\min\left(\mu(S), \mu(\bar{S})\right)}$$

Ratio-Cut problem definition

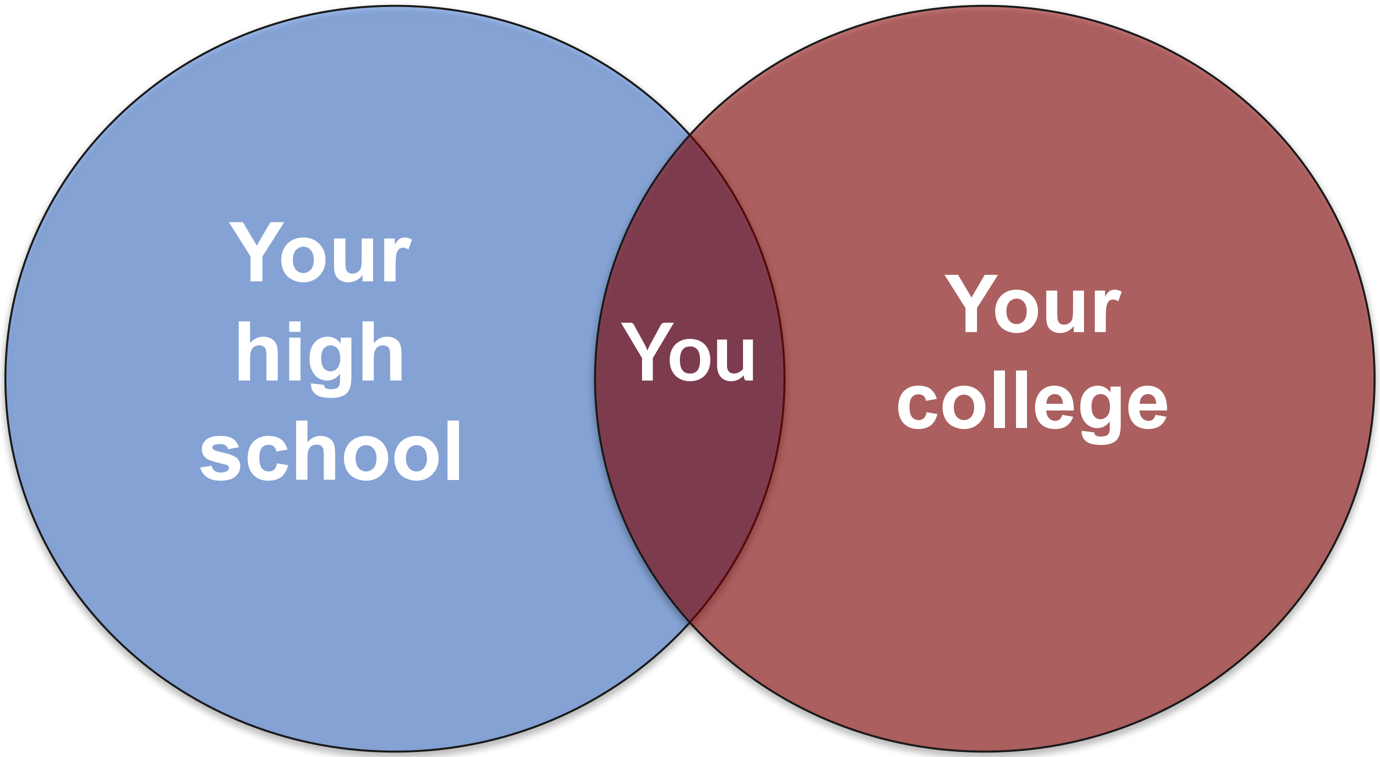
$$\min_{S \subset V} \frac{E(S, \bar{S})}{\min\left(\mu(S), \mu(\bar{S})\right)}$$

Balanced & K-Clustering

Balanced Clustering:
$$\mu(S) \geq c \cdot \mu(V)$$
$$\mu(T) \geq c \cdot \mu(V)$$

K-Clusters:
$$S_1 \cup S_2 \cup \cdots \cup S_K = V$$
$$q_\lambda(G, K) = \min_{S_1, \cdots S_K} \max_{S_i} q_{G,\lambda}[S_i, \bigcup_{j \neq i} S_j]$$

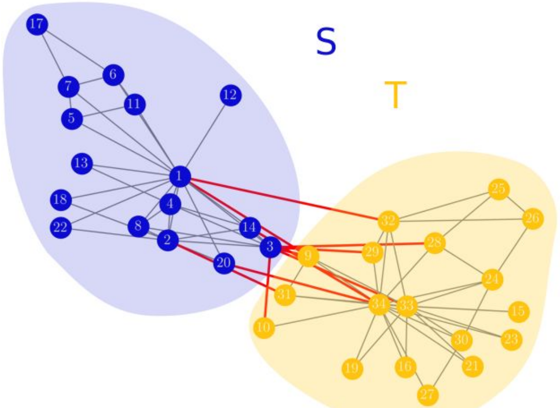
Overlapping Cuts (Arora et. al 2012)



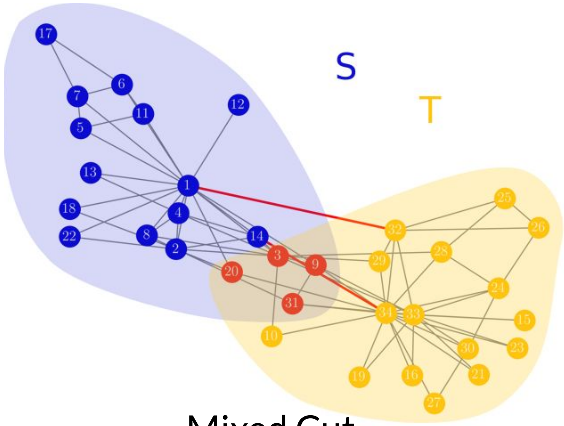
Overlapping Cuts - Introducing nodes into the cut

$$\delta_E[S,T] = E(S \setminus T, T \setminus S) \quad \text{and} \quad \delta_V[S,T] = S \cap T$$

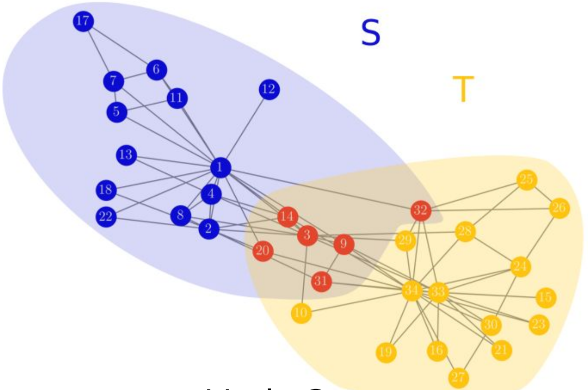
$$q_E[S,T] = \frac{w(\delta_E[S,T])}{\min\{\mu(S), \mu(T)\}} \qquad q_V[S,T] = \frac{\mu(\delta_V[S,T])}{\min\{\mu(S), \mu(T)\}}$$



Edge Cut



Mixed Cut



Node Cut

Overlapping Cuts

$$\delta_E[S,T] = E(S \setminus T, T \setminus S) \quad \text{and} \quad \delta_V[S,T] = S \cap T$$

$$q_E[S,T] = \frac{w(\delta_E[S,T])}{\min\{\mu(S), \mu(T)\}} \qquad q_V[S,T] = \frac{\mu(\delta_V[S,T])}{\min\{\mu(S), \mu(T)\}}$$

λ-Hybrid Cut: λ-HCUT

$$q_{G,\lambda}[S,T] = q_E[S,T] + \lambda \cdot q_V[S,T] = \frac{w(\delta_E[S,T]) + \lambda \cdot \mu(\delta_V[S,T])}{\min\{\mu(S), \mu(T)\}}$$

Overlapping Cuts

$$\delta_E[S,T] = E(S \setminus T, T \setminus S) \quad \text{and} \quad \delta_V[S,T] = S \cap T$$

$$q_E[S,T] = \frac{w(\delta_E[S,T])}{\min\{\mu(S),\mu(T)\}} \qquad q_V[S,T] = \frac{\mu(\delta_V[S,T])}{\min\{\mu(S),\mu(T)\}}$$

ε-Overlapping Ratio Cut: ε-ORC

$$\epsilon - ORC : \min_{S \cup T = V} q_E[S,T] = \min_{S \cup T = V} \frac{w(\delta_E[S,T])}{\min\{\mu(S),\mu(T)\}}$$

$$q_V[S,T] = \frac{\mu(\delta_V[S,T])}{\min\{\mu(S),\mu(T)\}} \leq \epsilon$$

Question: Can we design a framework for overlapping graph partitioning (OGP) that allows for

(i) a principled and intuitive mathematical formulation, together with
(ii) solid worst-case approximation algorithms that
(iii) scale gracefully to large networks?

Previous Work

- Lots of work, but missing at least one of the desired properties (Ahn et al., 2010; Andersen et al., 2012; Arora et al., 2012; Bonchi et al., 2013; Khandekar et al., 2014; Mishra et al., 2007; Airoldi et al., 2008; Yang & Leskovec, 2013; Gopalan & Blei, 2013; Li et al., 2017; Palla et al., 2012; Tsourakakis, 2015; Whang et al., 2016)
- All properties satisfied for non-overlapping ratio-cut objectives (Leighton & Rao, 1999; Arora et al., 2009; Leskovec et al., 2009; Shi & Malik, 2000; Orecchia et al., 2008)
- Scalable NON-OVERLAPPING graph-partitioning heuristics KL (Kernighan & Lin, 1970b), METIS (Karypis & Kumar, 1996; 1998; 1995) Graclus (Dhillon et al., 2007) KaHIP (Sanders & Schulz, 2013).

Ratio-Cut problem

$$\min_{x \in [-1,1] \perp \vec{1}} \frac{\|Bx\|_{1,w}}{\min_u \|x - u\vec{1}\|_{1,\mu}}$$

- Global Objective is not convex! (convex over convex)
- But very similar to:

$$\min_{x \in [-1,1] \perp \vec{1}} \frac{\|Bx\|_{2,w}}{\min_u \|x - u\vec{1}\|_{2,\mu}} = \sqrt{\lambda_2(\mathcal{L})}$$

$$u = \frac{\sum_{v \in V} \mu(v) x_v}{\sum_{v \in V} \mu(v)} = mean_\mu(x)$$

Cheeger Inequality (Alon & Milman, 1985)

Guarantee we know for G: $\min_{S,T \subset V} q_E[S,T] \geq \lambda_2(\mathcal{L})/2$

Problem: Eigenvalues of the normalized Laplacian are affected both by the size of the cut but also **from the length of paths**

Solution: Construct certificate graph H where every cut in H is worse that the equivalent in G, but all paths are small

$$q(G) \geq q(H) \geq \lambda_2(\mathcal{L}_H)/2$$

Cut-Matching Game (Khandekar et al., 2014)

$$q(G) \geq q(H) \geq \lambda_2(\mathcal{L}_H)/2$$

$H_0 \leftarrow G$
$\alpha_0 \leftarrow 1$
for $t \leftarrow 1, \cdots, O(\log^2(n))$ **do**
    $(S_t, \bar{S}_t) \leftarrow$ Cut from H_{t-1}
    $M_t, \alpha_t \leftarrow$ matching $(S, \bar{S})$ in G with congestion α_t
    $H_t = H_{t-1} + M_t$
return $best(S_t, \bar{S}_t), \frac{\lambda_2(H_T)}{\sum_{t=1}^{T} \alpha_t}$

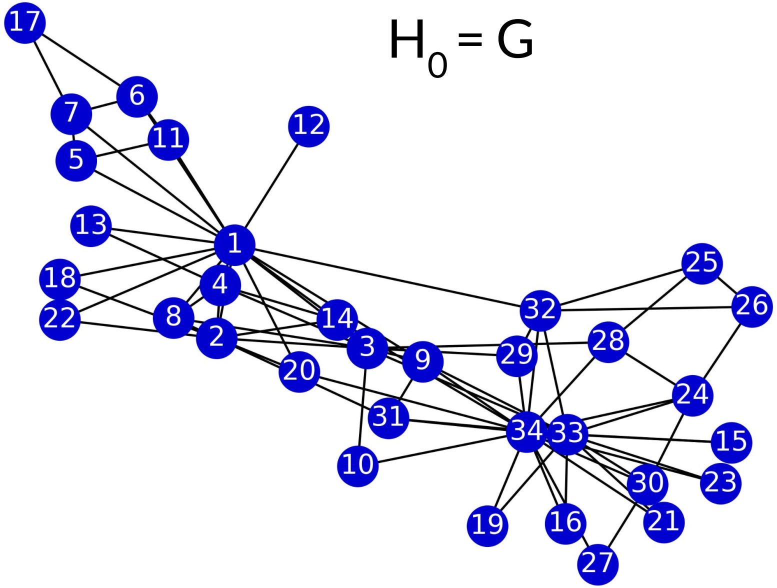
Cut-Matching Game (Khandekar et al., 2014)

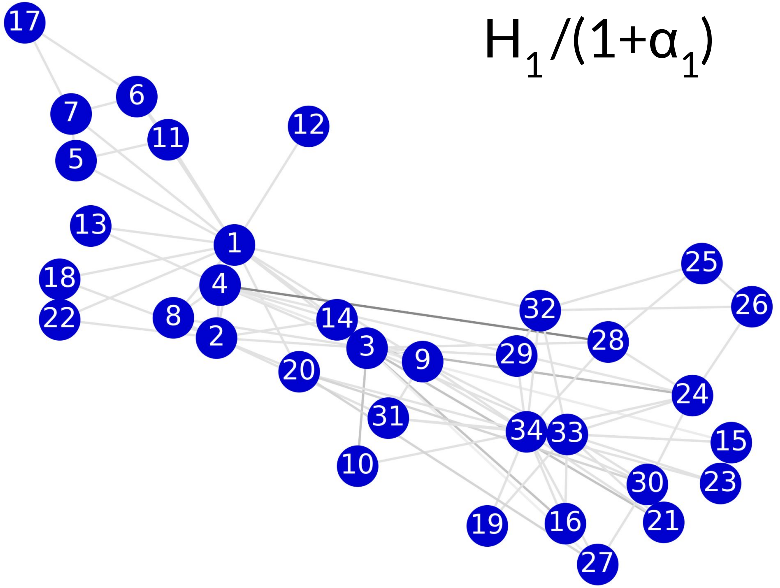
$$q(G) \geq q(H) \geq \lambda_2(\mathcal{L}_H)/2$$

- In every round the smallest eigenvalue increases by 1/log(n).
- In every round we incur constant congestion.
- After O($\log^2$(n)) iterations, H will be O(log(n))-expander with O($\log^2$(n)) congestion.
- H certifies a O(log(n))-approximation.

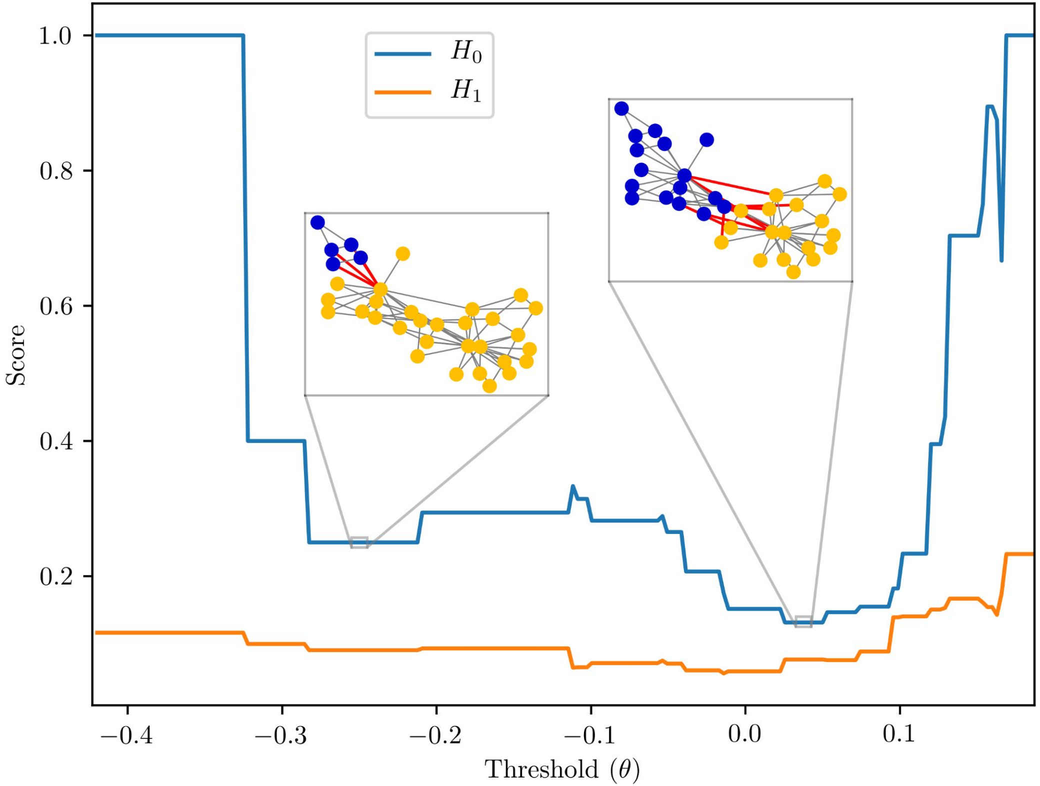
Cut-Matching Game (Khandekar et al., 2014)

$$q(G) \geq q(H) \geq \lambda_2(\mathcal{L}_H)/2$$





Cut-Matching Game (Khandekar et al., 2014)



Finding the initial cut

$$x = v_2(\mathcal{L}_{H_{t-1}})$$

Non smooth, small changes in the input, lead to big changes in the result

u is a random vector

$$x = \operatorname{soft}\min_{i=2}^{n} \Lambda(\mathcal{L}_{H_{t-1}}) V(H_{t-1}) u = e^{-k\mathcal{L}_{H_{t-1}}} u$$

Smoothness

Every eigenvector is weighted by $\frac{e^{\lambda_i}}{\sum_{j=2}^{n} e^{\lambda_j}}$

Eigenvectors with similar eigenvalues are equally present

Cut Improvement (Andersen & Lang, 2008)

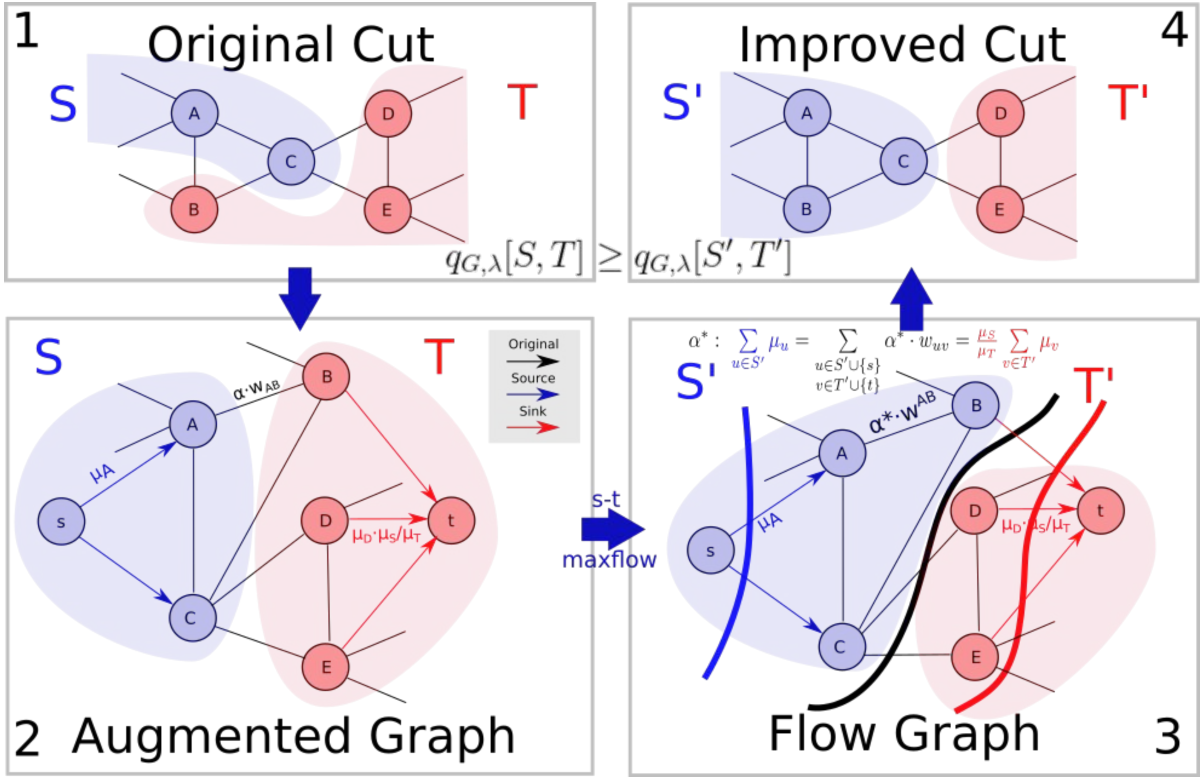
- Andersen & Lang 2008
- Given a seed cut s, find a better cut x

$$\min_{x \in [-1,1]^n \perp 1} \frac{\|Bx\|_{1,w}}{\min_u \|s - u\vec{1}\|_{1,\mu} - \|x - s\|_{1,\mu}}$$

- Convex!!!
- Solution can be found through a small number of s-t maxflow computations

Cut Improvement (Andersen & Lang, 2008)

Augment graph with source s and sink t.



The result has a better ratio cut

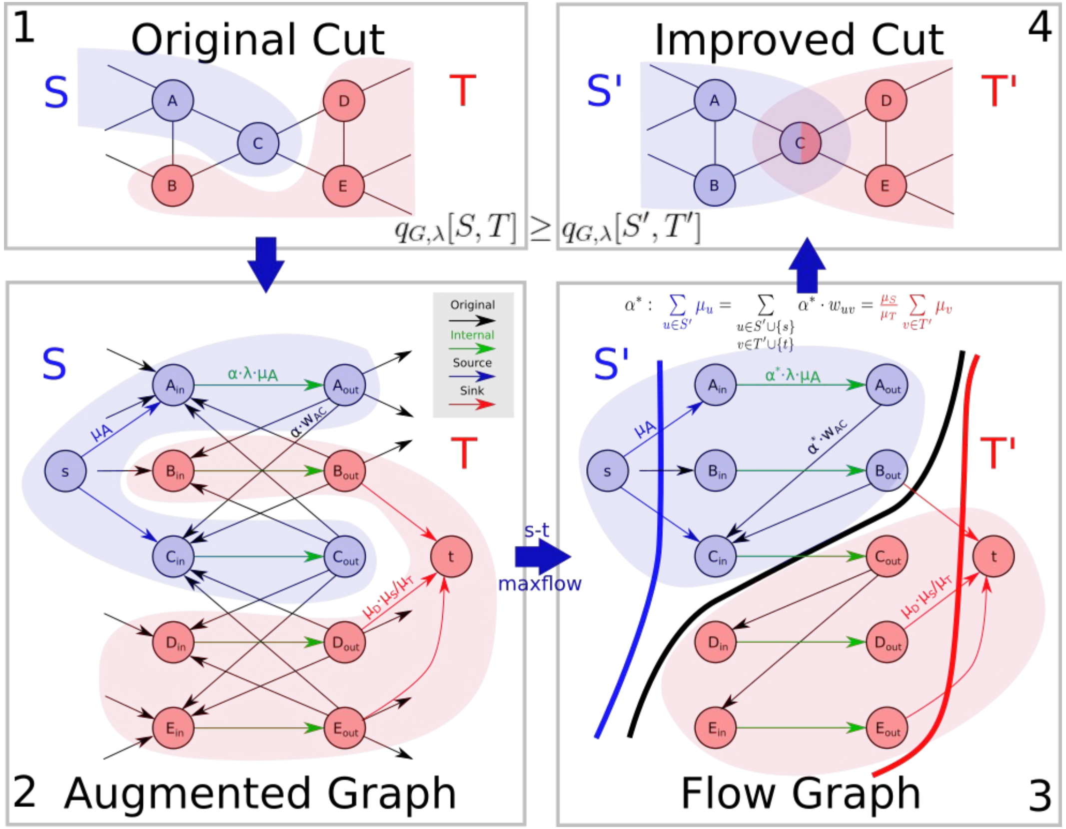
Connect s to all nodes in S and all nodes in T to t. Degree of s and t are the same

For the correct value of α the blue, red and black cuts have the same value

Cut Improvement - Overlapping

Break up nodes in v_{in} and v_{out}. All edges exit out-nodes and enter in-nodes.

The flow through node v is limited to $\lambda \cdot \mu_v$



If the internal edge is cut, then the node belongs to the overlap

Extensions

Balanced Clustering:

- In the cut-improve step, starting from a bisection, don't lower α to values that (S', T') are not balanced.
- Bad for theoretical guarantees

K-Clustering:

- Recursive bisectioning K-1 times as described in Kannan et al., 2004
- Also bad for theoretical guarantees

Results

Datasets:

- Synthetic Overlapping Stochastic Block Model (O-SBM)
- Real social networks from SNAP (Leskovec & Krevl, 2014)

Competing algorithms

- cm+improve: Cut matching + cut improvement (this work)
- SweepCut: Best threshold in spectral
- Spectral + Greedy Improve: Start with spectral bisection, use greedy Kernighan-Lin algorithm
- METIS: Contract-Cut-Expand heuristic algorithm

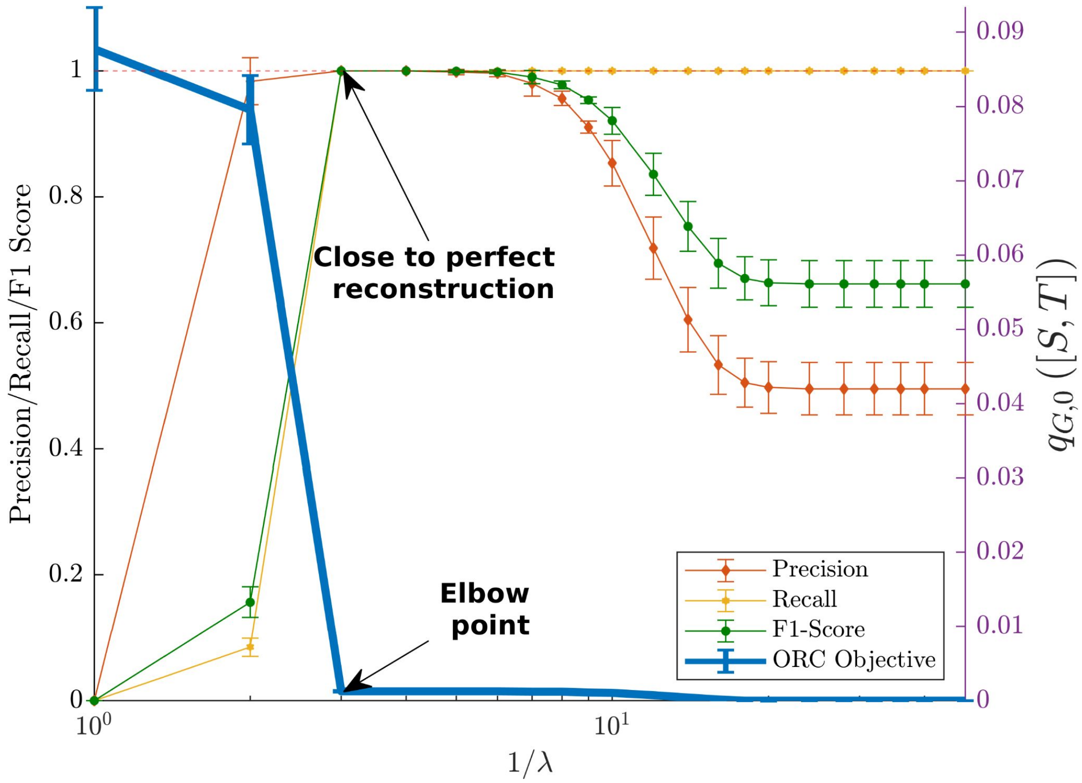
Datasets

	L	R	C
L	p	ε	p
R	ε	p	p
C	p	p	q

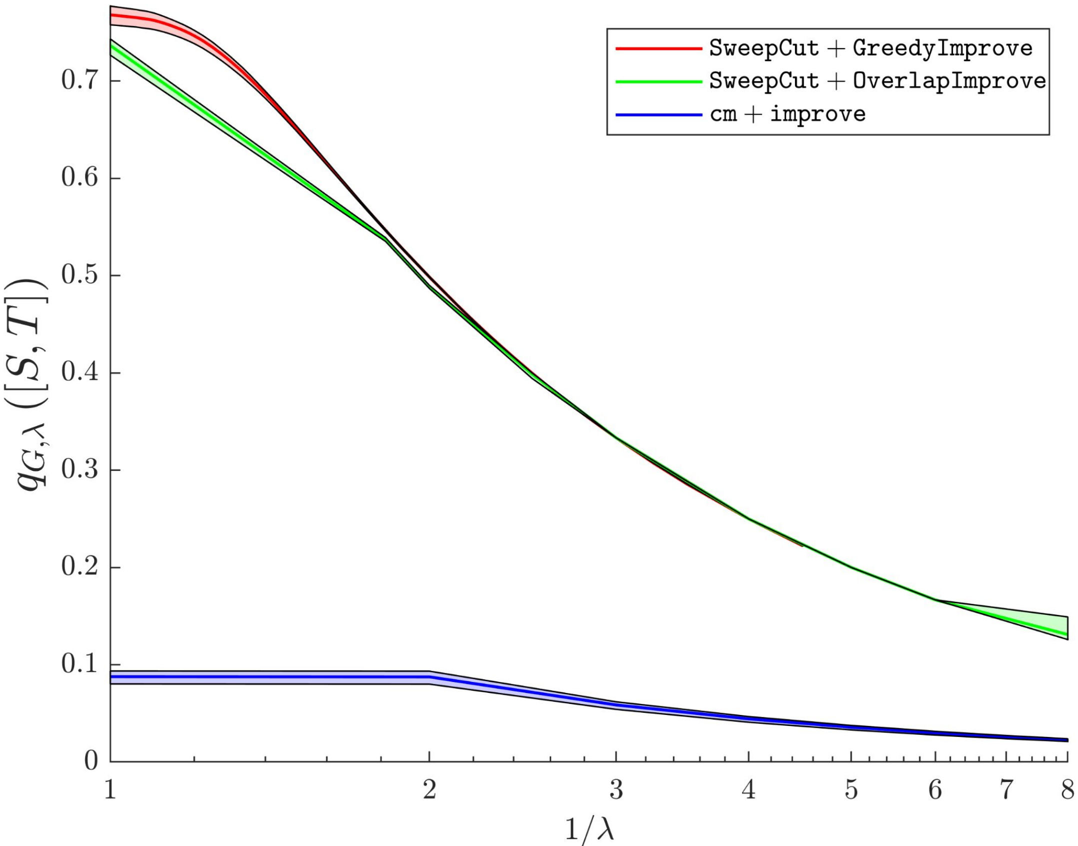
O-SBM: Three blocks (**L**eft, **R**ight, **C**enter) n=10,000
Probability of edge depends only on which blocks the two nodes belong
Center is well connected to both **L**eft and **R**ight

Network	Description	n	m	time
DBLP	Co-authorship network	83,114	409,541	2-4min
Amazon	Co-purchasing network	334,863	925,872	15-18min
Youtube	Group network	1,134,890	2,987,624	55-75min

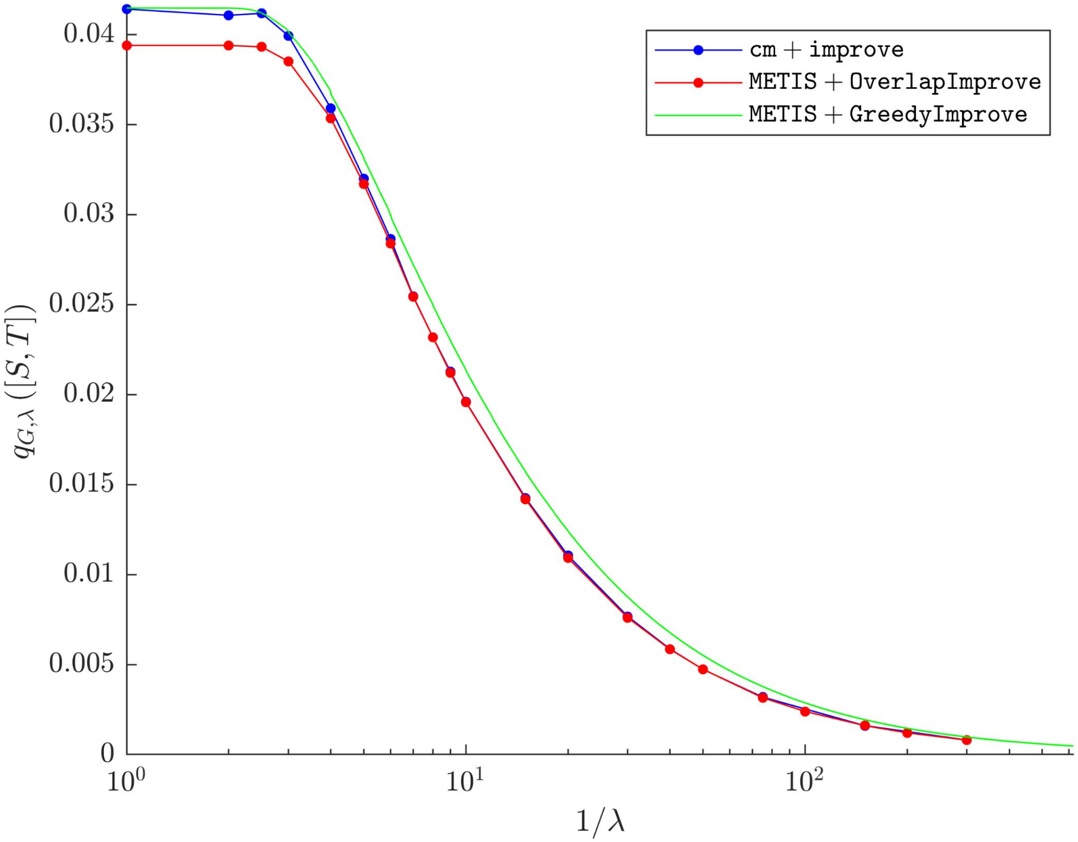
Overlapping Stochastic Block Model



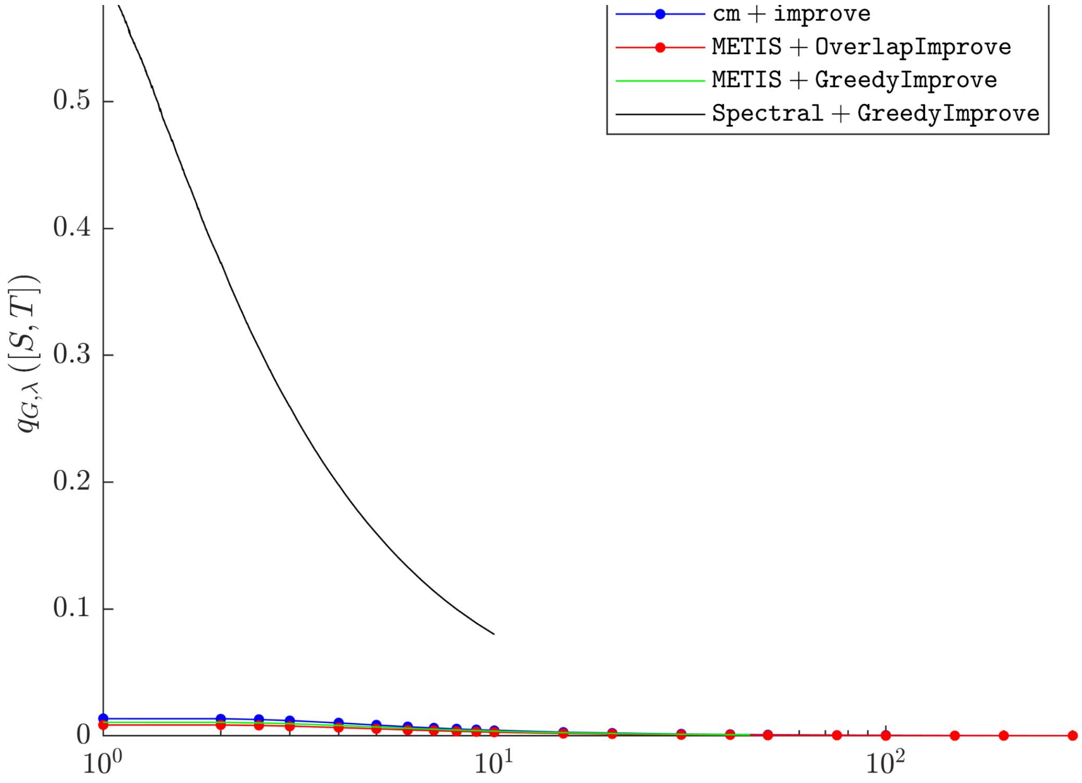
Overlapping Stochastic Block Model



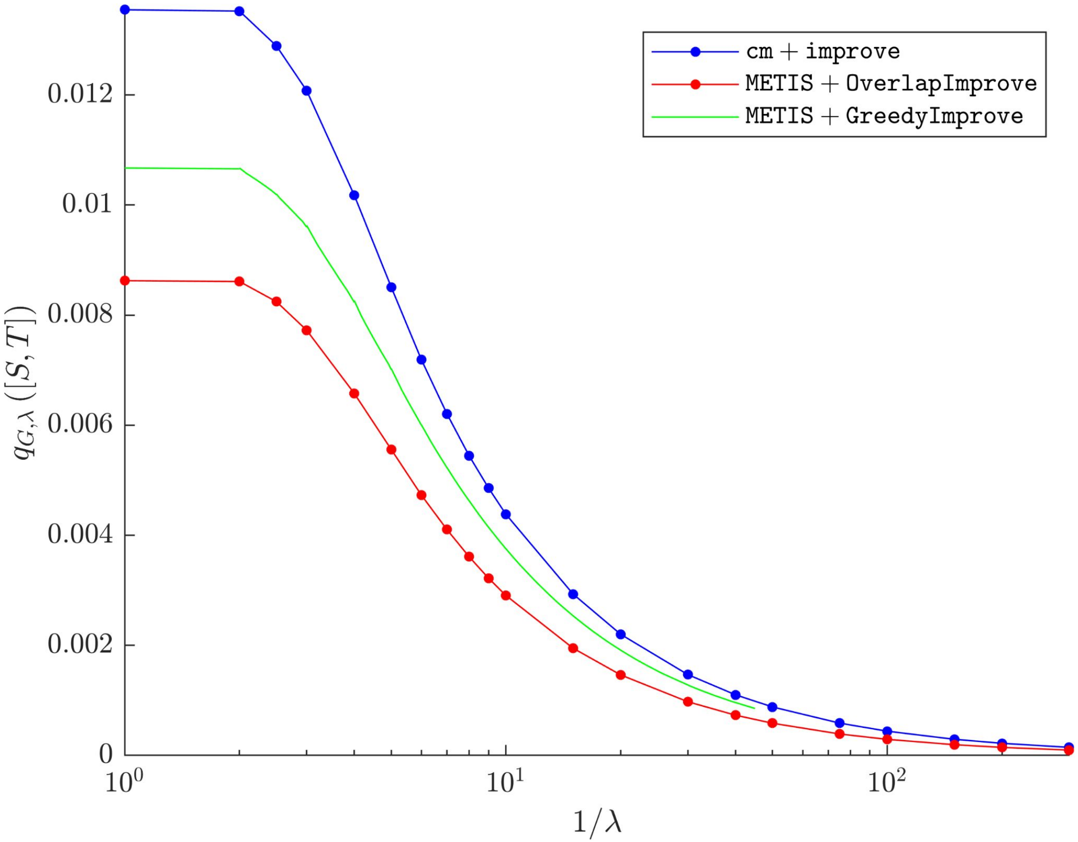
DBLP co-authorship network



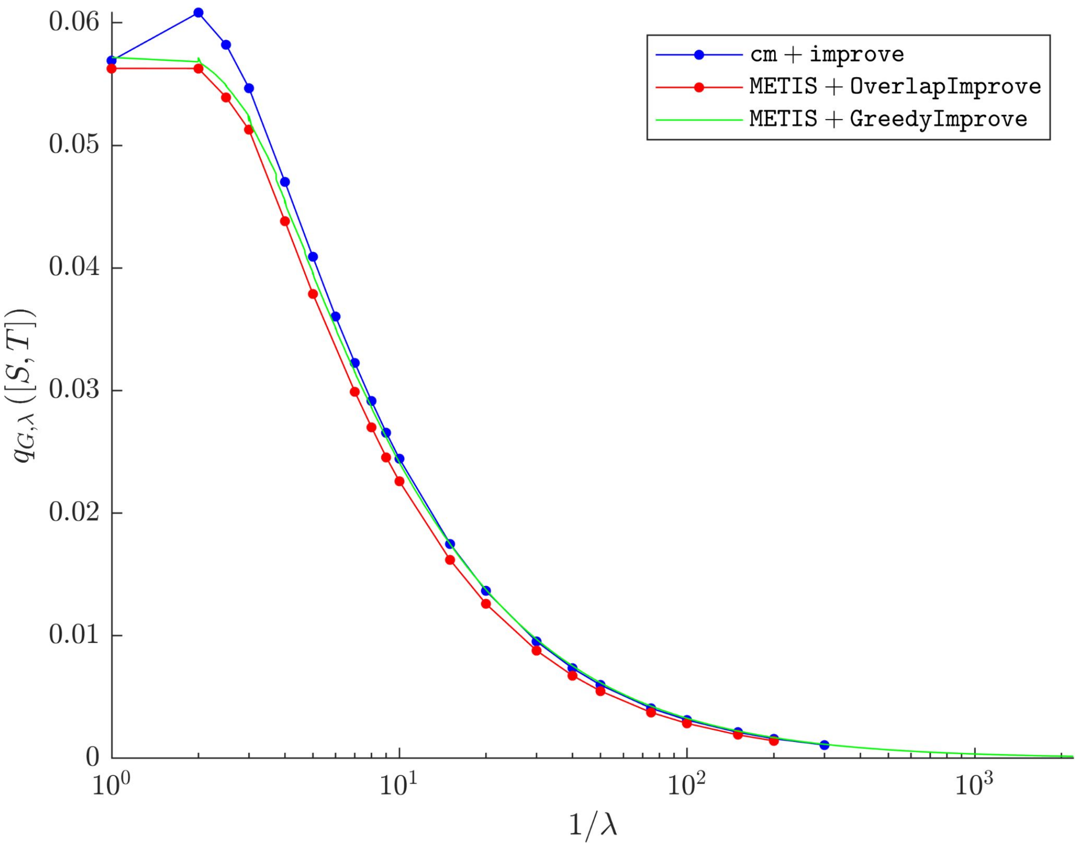
Amazon co-purchasing network



Amazon co-purchasing network

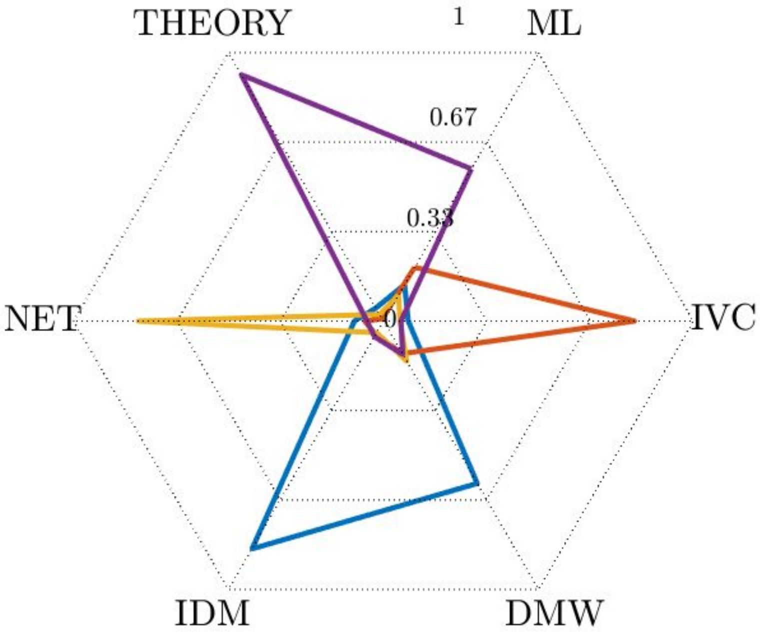


Youtube


cm + improve
METIS + OverlapImprove
METIS + GreedyImprove
0.06
0.05
0.04
0.03
0.02
0.01
0
$q_{G,\lambda}([S,T])$
10^0
10^1
10^2
10^3
$1/\lambda$

K-Clusters in DBLP

Recursive bisectioning!



Future Work

- Extend work to hypergraphs
- Use different initial cut strategies
- Improve runtime

Questions?

References

- Ahn, Y.-Y., Bagrow, J. P., and Lehmann, S. Link communities reveal multiscale complexity in networks. Nature, 466(7307):761–764, 2010.
- Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. Mixed membership stochastic blockmodels. Journal of machine learning research, 9(Sep):1981–2014, 2008.
- Andersen, R. and Lang, K. An algorithm for improving graph partitions. In SODA '08 Proc. 19th ACM-SIAM Symp. Discret. algorithms, pp. 651–660, 2008
- Andersen, R., Gleich, D. F., and Mirrokni, V. Overlapping clusters for distributed computation. In Proceedings of the fifth ACM international conference on Web search and data mining, pp. 273–282. ACM, 2012.
- Arora, S., Rao, S., and Vazirani, U. Expander flows, geometric embeddings and graph partitioning. In STOC '04 Proc. thirty-sixth Annu. ACM Symp. Theory Comput., pp. 222–231, New York, NY, USA, 2004. ACM. ISBN 1-58113-852-0. doi: http://doi.acm.org/10.1145/1007352.1007355.
- Arora, S., Ge, R., Sachdeva, S., and Schoenebeck, G. Finding overlapping communities in social networks: toward a rigorous approach. In Proceedings of the 13th ACM Conference on Electronic Commerce, pp. 37–54. ACM, 2012.
- Arora, S., Rao, S., and Vazirani, U. Expander flows, geometric embeddings and graph partitioning. Journal of the ACM (JACM), 56(2):5, 2009.
- Bonchi, F., Gionis, A., and Ukkonen, A. Overlapping correlation clustering. Knowledge and information systems, 35(1):1–32, 2013.
- Dhillon, I. S., Guan, Y., and Kulis, B. Weighted graph cuts without eigenvectors a multilevel approach. IEEE transactions on pattern analysis and machine intelligence, 29(11):1944–1957, 2007
- Gopalan, P. K. and Blei, D. M. Efficient discovery of overlapping communities in massive networks. Proceedings of the National Academy of Sciences, 110(36):14534–14539, 2013.
- Karypis, G. and Kumar, V. Metis-unstructured graph partitioning and sparse matrix ordering system, version 2.0. 1995.
- Karypis, G. and Kumar, V. Parallel multilevel graph partitioning. In IPPS, pp. 314–319, 1996.
- Karypis, G. and Kumar, V. A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM J. Sci. Comput., 20(1):359–392, December 1998. ISSN 1064-8275. doi: 10.1137/S1064827595287997. URL http://dx.doi.org/10.1137/S1064827595287997.
- Kernighan, B. W. and Lin, S. An efficient heuristic procedure for partitioning graphs. The Bell system technical journal, 49(2):291–307, 1970b.
- Khandekar, R., Rao, S., and Vazirani, U. Graph partitioning using single commodity flows. Journal of the ACM (JACM), 56(4):19, 2009.
- Khandekar, R., Kortsarz, G., and Mirrokni, V. On the advantage of overlapping clusters for minimizing conductance. Algorithmica, 69(4):844–863, 2014.
- Leighton, T. and Rao, S. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. Journal of the ACM (JACM), 46(6):787–832, 1999.
- Leskovec, J., Lang, K., Dasgupta, A., and Mahoney, M. W. Statistical properties of community structure in large social and information networks. In Proceeding of the 17th international conference on World Wide Web, pp. 695–704. ACM, 2008.
- Li, P., Dau, H., Puleo, G., and Milenkovic, O. Motif clustering and overlapping clustering for social network analysis. In IEEE INFOCOM 2017-IEEE Conference on Computer Communications, pp. 1–9. IEEE, 2017.
- Mishra, N., Schreiber, R., Stanton, I., and Tarjan, R. E. Clustering social networks. In International Workshop on Algorithms and Models for the Web-Graph, pp. 56–67. Springer, 2007.
- Orecchia, L. Fast Approximation Algorithms for Graph Partitioning using Spectral and Semidefinite-Programming Techniques. PhD thesis, EECS Department, University of California, Berkeley, May 2011. URL http://www.eecs.berkeley.edu/Pubs/TechRpts/2011/EECS-2011-56.html.
- Palla, K., Knowles, D., and Ghahramani, Z. An infinite latent attribute model for network data. arXiv preprint arXiv:1206.6416, 2012.
- Sanders, P. and Schulz, C. Think Locally, Act Globally: Highly Balanced Graph Partitioning. In Proceedings of the 12th International Symposium on Experimental Algorithms (SEA'13), volume 7933 of LNCS, pp. 164–175. Springer, 2013.
- Shi, J. and Malik, J. Normalized cuts and image segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(8):888–905, 2000.
- Tsourakakis, C. Provably fast inference of latent features from networks: with applications to learning social circles and multilabel classification. In Proceedings of the 24th International Conference on World Wide Web, WWW 2015, pp. 1111–1121, 2015.
- Whang, J. J., Gleich, D. F., and Dhillon, I. S. Overlapping community detection using neighborhood-inflated seed expansion. IEEE Transactions on Knowledge and Data Engineering, 28(5):1272–1284, 2016.