Scalable Community Detection in the Heterogeneous Stochastic Block Model

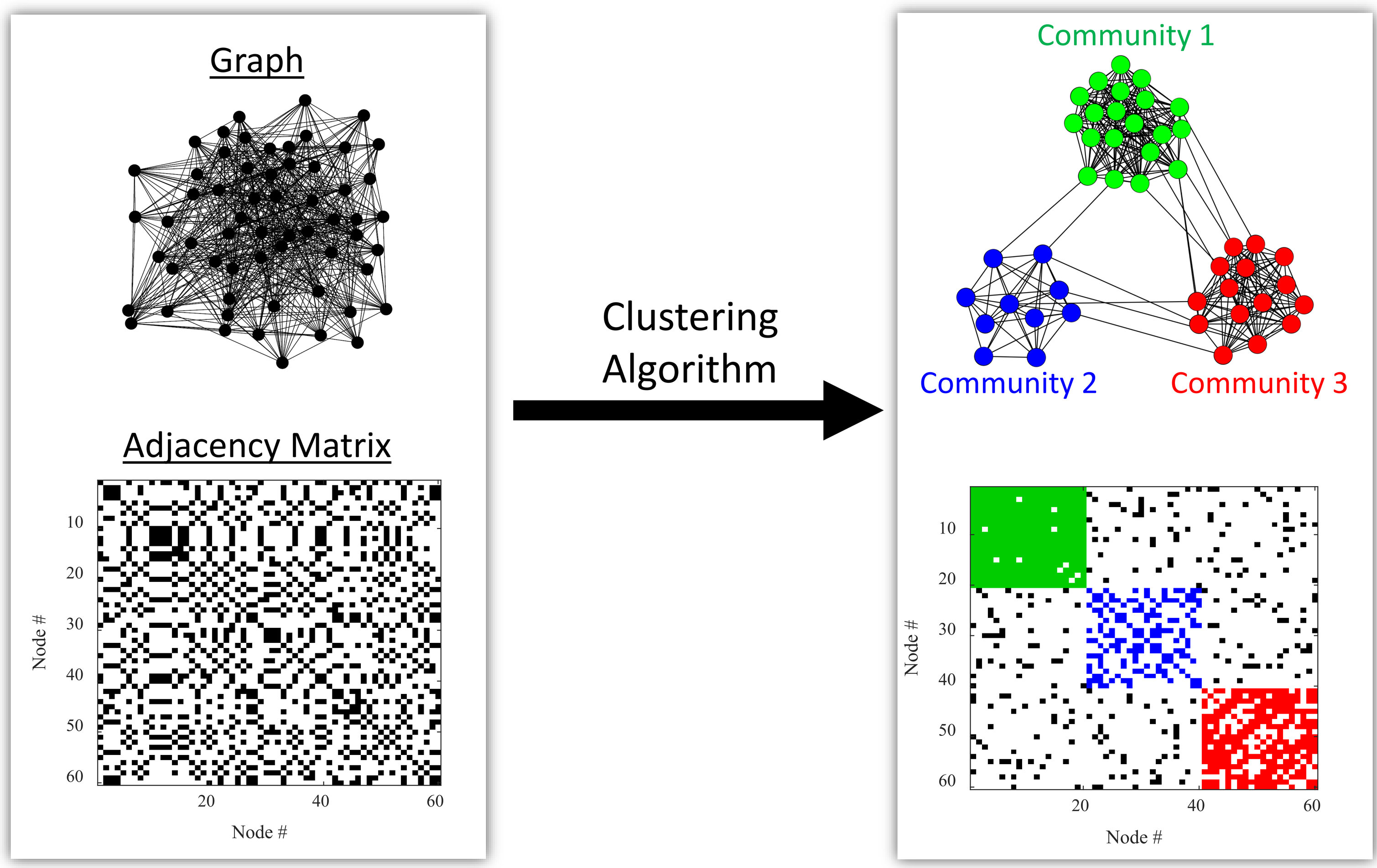
Andre Beckus and George K. Atia

Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL

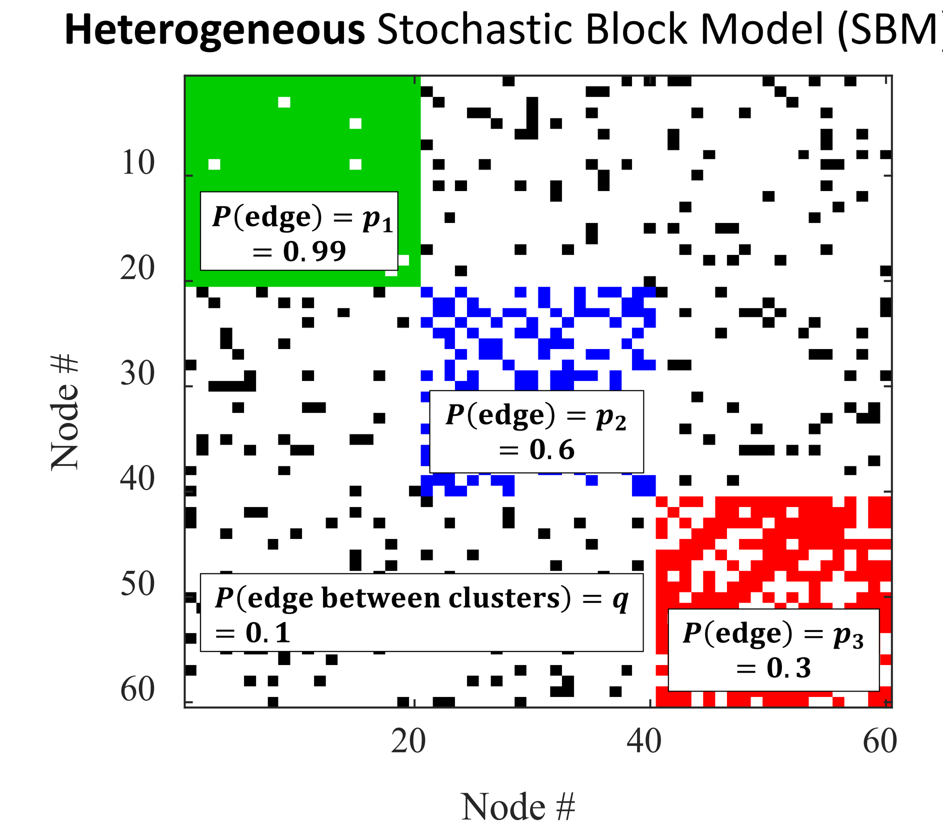
Abstract

This paper studies the unsupervised clustering of large graphs generated from the heterogeneous Stochastic Block Model. We present a sketch-based community detection algorithm, which substantially reduces computational complexity by clustering only a small set of nodes sampled from the full graph followed by a retrieval algorithm. We first show cases where existing algorithms exhibit reduced error rates when all nodes possess the same average number of intra-cluster connections. This behavior is demonstrated for both convex-optimization-based and spectral algorithms. Based on this insight, we develop SPIN, a degree-based sampling method to produce sketches with cluster proportions more favorable for successful clustering. By sampling nodes inversely proportional to their degrees, SPIN can exploit this reduction in error to significantly improve the phase transition as compared to full graph clustering.

Traditional Community Detection



Model



Number of nodes = N

Number of clusters = r

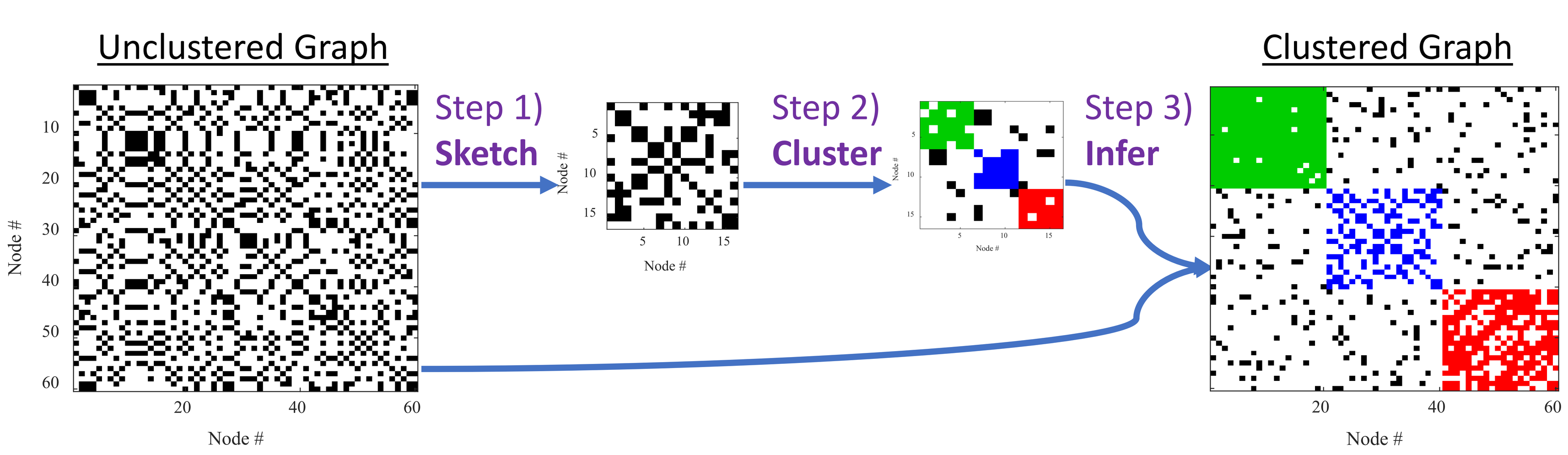
Nodes in cluster $i = n_i$

$n_{min} = \min_{1 \le i \le r} n_i$

Proposed Approach

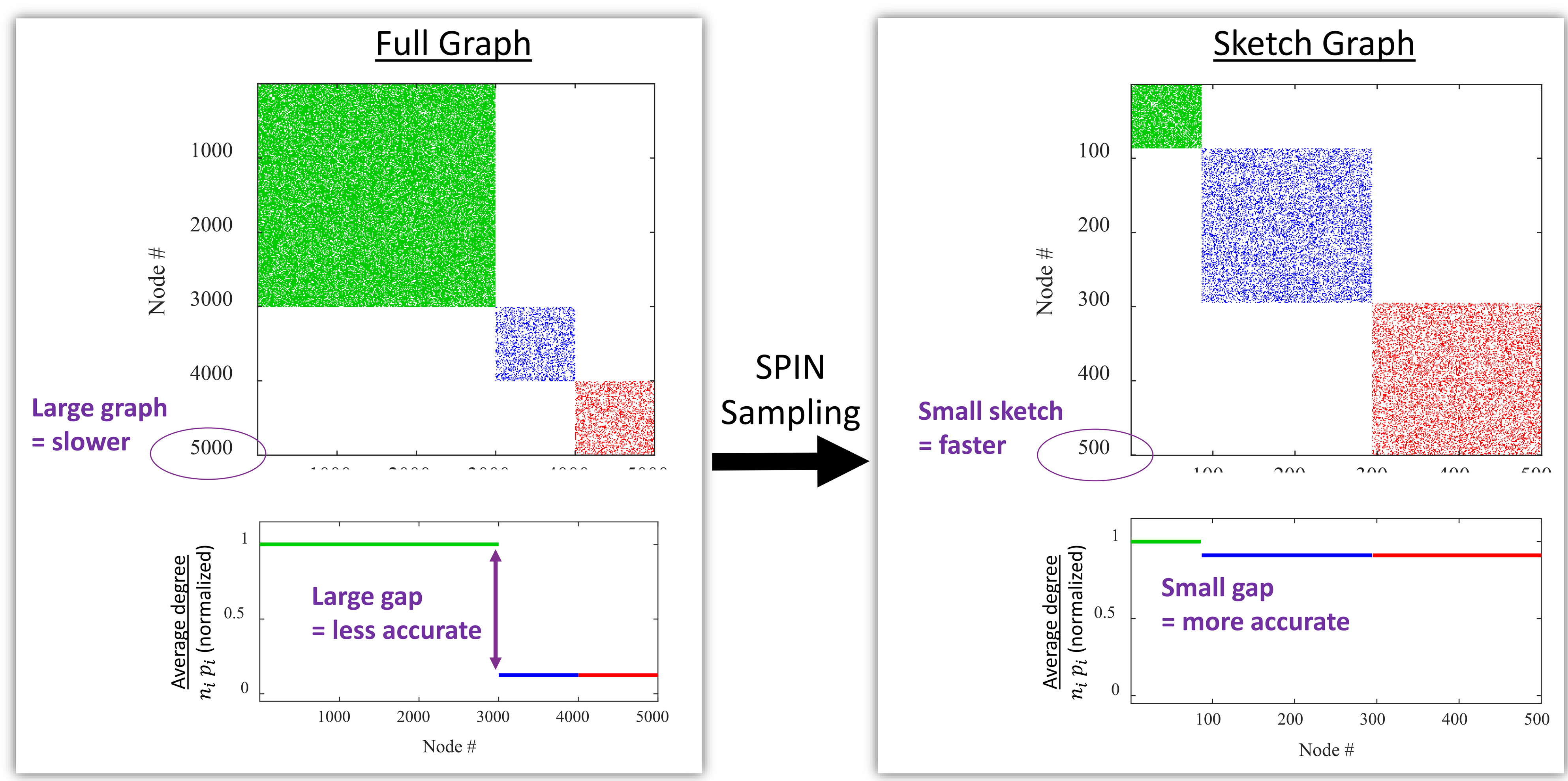
Sketch-based Clustering

We minimize the clustering bottleneck by using a small sketch.
We can incorporate any existing clustering algorithm.



See [1] for application to homogeneous SBM

Key Benefits



Sampling Methods		
	URS Uniform Random Sampling	SPIN Sampling Inversely proportional to Node Degree
Node Sampling Probability	$1/N$	$\propto \frac{1}{node\ degree}$
Advantage	**Speed**	**Speed + Accuracy**

Complexity - Homogeneous case $p_1 = \cdots = p_r$ [1]

$N \to \infty$, $p > 0.5$ constant, $q = O(N/n_{min})$, where n_{min} is size of smallest cluster

	URS Clustering with (Chen,2014)	SPIN (roughly) Clustering with (Chen,2014)	(Chen, 2014)	(Cai, 2015) (state-of-the-art)
Minimum cluster size	$\Omega(\sqrt{N}\log N)$	$\Omega(r\ \log^3 N)$	$\Omega(\sqrt{N}\log N)$	$\Omega(\log N)$
Required # of samples	$\Omega\left(\frac{N^2\log^2 N}{n_{min}^2}\right)$	$\Omega(r^2\log^4 N)$	N	N
Per iteration time complexity (clustering step)	$O\left(\frac{rN^4\log^4 N}{n_{min}^4}\right)$	$O(r^5\log^8 N)$	$O(rN^2)$	$O(rN^2)$

Note: if $n_{min} = \Theta(N)$, then $r = \Theta(1)$

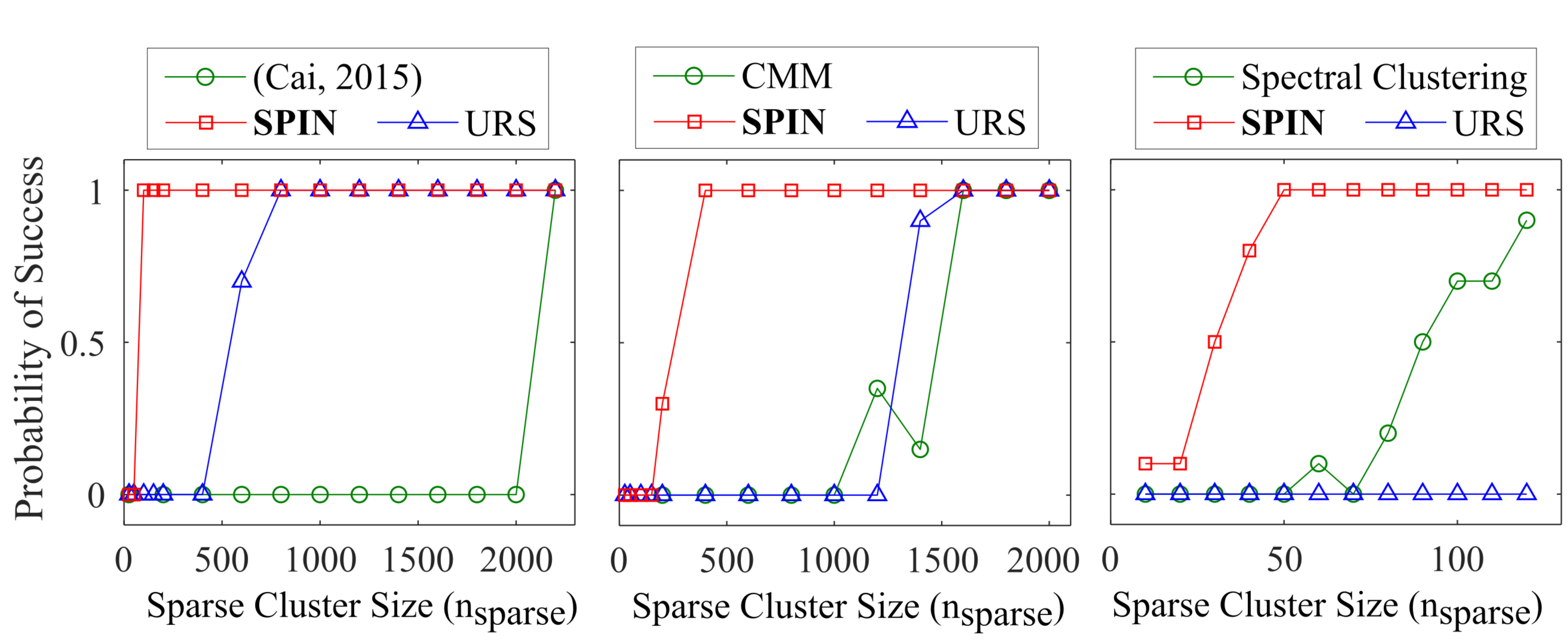
Numerical Results

Accuracy

Full graph has one large dense cluster, two small sparse clusters.

SPIN sampling improves success by producing sketch with smaller dense cluster, and larger sparse clusters

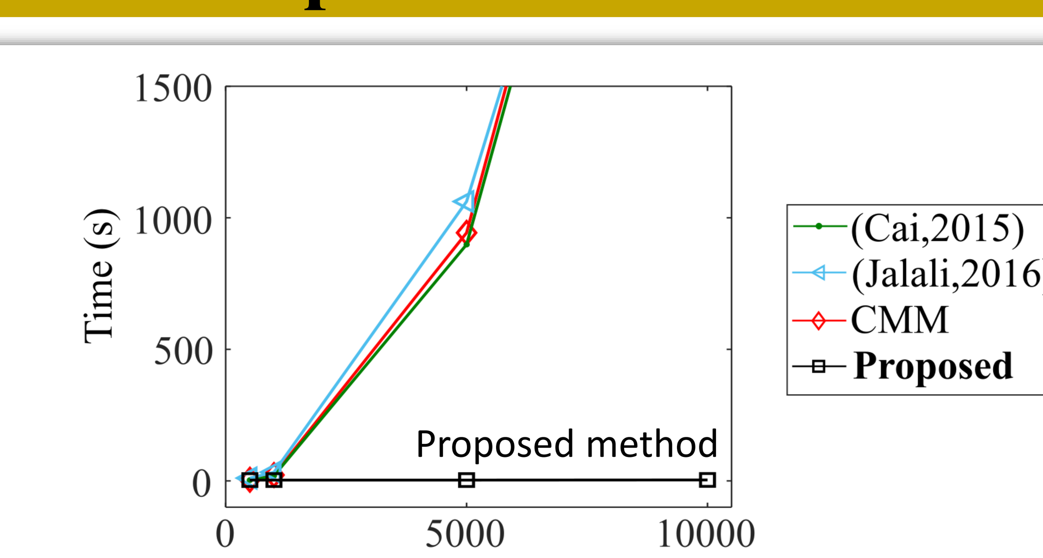
$N = 5000, r = 3,$
$p_1 = 0.8, n_1 = N - 2n_{sparse}$
$p_2 = p_3 = 0.2, n_2 = n_3 = n_{sparse}$
$N' = 600$, 10 trials
$q = 0.02$



Speed

$r = 3$
$p_1 = 0.4, n_1 = \frac{N}{2}$
$p_2 = p_3 = 0.8, n_2 = n_3 = \frac{N}{4}$
$N' = 200$, 3 trials
$q = 0.08$,

All algorithms achieve 100% accuracy



References

[1] Andre Beckus*, Mostafa Rahmani*, Adel Karimian, and George K. Atia, "**Scalable and Robust Community Detection with Randomized Sketching**," , arXiv:1805.10927 , 2019. *Authors contributed equally.

(Chen, 2014) Y. Chen, A. Jalali, S. Sanghavi, and H. Xu, "**Clustering partially observed graphs via convex optimization**," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2213–2238, Jan. 2014

(Cai, 2015) T. T. Cai and X. Li, "**Robust and computationally feasible community detection in the presence of arbitrary outlier nodes**," Ann. Statist., vol. 43, 2015.

(Jalali, 2016) A. Jalali, Q. Han, I. Dumitriu, and M. Fazel, "**Exploiting tradeoffs for exact recovery in heterogeneous stochastic block models**," in Proc. Adv. Neural Inf. Process. Syst., 2016.

CMM Y. Chen, X. Li, and J. Xu, "**Convexifed modularity maximization for degree-corrected stochastic block models**," Ann. Statist., vol. 46, no. 4, 2018.

Acknowledgement: This work was supported by NSF CAREER Award CCF-1552497