Proceedings of Machine Learning Research vol ???:1–6, 2025 38th Annual Conference on Learning Theory

Stochastic block models with many communities and the Kesten–Stigum bound

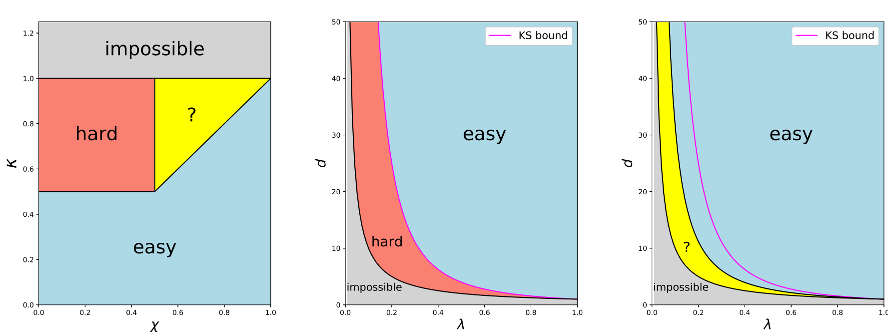
Byron Chin BYRONC@MIT.COM
Elchanan Mossel ELMOS@MIT.EDU
Department of Mathematics, Massachusetts Institute of Technology
Youngtak Sohn YOUNGTAK_SOHN@BROWN.EDU
Division of Applied Mathematics, Brown University
Alexander S. Wein ASWEIN@UCDAVIS.EDU
Department of Mathematics, University of California, Davis

Editors: Nika Haghtalab and Ankur Moitra

We study the inference of communities in stochastic block models with a growing number of communities. For block models with n vertices and a fixed number of communities q, it was predicted in Decelle et al. (2011) that there are computationally efficient algorithms for recovering the communities above the Kesten–Stigum (KS) bound and that efficient recovery is impossible below the KS bound. This conjecture has since stimulated a lot of interest, with the achievability side proven in a line of research culminating in work of Abbe and Sandon (2018). Conversely, the hardness side of the conjecture has been supported by recent progress based on the low-degree paradigm.

In this paper we investigate community recovery in the regime $q \to \infty$ where no such predictions exist. We show that efficient inference of communities remains possible above the KS bound. Furthermore, we show that recovery of block models is low-degree-hard below the KS bound when the number of communities $q \ll \sqrt{n}$. Perhaps surprisingly, we find that when $q \gg \sqrt{n}$, there is an efficient algorithm based on non-backtracking walks for recovery even below the KS bound. We identify a new threshold which we conjecture is the threshold for weak recovery in this regime. Finally, we show that detection is easy and identify (up to a constant) the information-theoretic threshold for community recovery as q diverges. Our low-degree hardness results also naturally have consequences for graphon estimation, improving results of Luo and Gao (2023).

In the figure below, the blue regions represent computationally efficient regimes, the red regions represent low-degree hardness, the gray regions represent information-theoretic impossibility, and the yellow regions are still unknown. The left plot shows κ vs χ where $q \asymp n^{\chi}$ and $\lambda \asymp d^{-\kappa}$. The middle and right plots show d vs λ for $q = n^{1/3}$ and $q = n^{2/3}$ respectively.[1]



1. Extended abstract. Full version can be found at [arXiv:2503.03047, v2].

Acknowledgments

EM and YS thank Allan Sly for helpful discussions. BC is supported by an NSF Graduate Research Fellowship and Simons Investigator award (622132). EM was partially supported by NSF DMS-2031883, Bush Faculty Fellowship ONR-N00014-20-1-2826, Simons Investigator award (622132), and MURI N000142412742. ASW was partially supported by a Sloan Research Fellowship and NSF CAREER Award CCF-2338091.

References

Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: fundamental limits and efficient algorithms for recovery. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science—FOCS 2015*, pages 670–688. IEEE Computer Soc., Los Alamitos, CA, 2015. doi: 10.1109/FOCS.2015.47. URL https://doi.org/10.1109/FOCS.2015.47.

Emmanuel Abbe and Colin Sandon. Proof of the achievability conjectures for the general stochastic block model. *Comm. Pure Appl. Math.*, 71(7):1334–1406, 2018. ISSN 0010-3640. doi: 10.1002/cpa.21719. URL https://doi.org/10.1002/cpa.21719.

Edo M Airoldi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. *Advances in Neural Information Processing Systems*, 26, 2013.

Afonso S Bandeira, Jess Banks, Dmitriy Kunisky, Christopher Moore, and Alex Wein. Spectral planting and the hardness of refuting cuts, colorability, and communities in random graphs. In *Conference on Learning Theory*, pages 410–473. PMLR, 2021.

Jess Banks, Cristopher Moore, Joe Neeman, and Praneeth Netrapalli. Information-theoretic thresholds for community detection in sparse networks. In *Conference on Learning Theory*, pages 383–416. PMLR, 2016.

P.J. Bickel and A. Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

P. M. Bleher, J. Ruiz, and V. A. Zagrebnov. On the purity of the limiting Gibbs state for the Ising model on the Bethe lattice. *J. Statist. Phys.*, 79(1-2):473–482, 1995.

Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 1347–1357. IEEE, 2015.

Stanley Chan and Edoardo Airoldi. A consistent histogram estimator for exchangeable graph models. In *International Conference on Machine Learning*, pages 208–216. PMLR, 2014.

Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015. ISSN 00905364. URL http://www.jstor.org/stable/43556512.

Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *J. Mach. Learn. Res.*, 17:Paper No. 27, 57, 2016. ISSN 1532-4435. URL https://mathscinet.ams.org/mathscinet-getitem?mr=3491121.

A. Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(02):227–284, 2010.

A. Condon and R.M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Structures and Algorithms*, 18(2):116–140, 2001.

C. Daskalakis, E. Mossel, and S. Roch. Optimal phylogenetic reconstruction. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing (STOC 2006)*, pages 159–168, 2006. URL http://www.stat.berkeley.edu/~mossel/publications/log_stoc.pdf. See the journal version for proofs. The proofs required an additional assumption of discretized branch length.

A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physics Review E*, 84:066106, Dec 2011. doi: 10.1103/PhysRevE.84.066106. URL http://link.aps.org/doi/10.1103/PhysRevE.84.066106.

M.E. Dyer and A.M. Frieze. The solution of some random NP-hard problems in polynomial expected time. *Journal of Algorithms*, 10(4):451–489, 1989.

Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

W. S. Evans, C. Kenyon, Y. Peres, and L. J. Schulman. Broadcasting on trees and the Ising model. *Ann. Appl. Probab.*, 10(2):410–433, 2000.

Ehud Friedgut, Jean Bourgain, et al. Sharp thresholds of graph properties, and the k-sat problem. *Journal of the American mathematical Society*, 12(4):1017–1054, 1999.

Chao Gao, Yu Lu, and Harrison H. Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624 – 2652, 2015a. doi: 10.1214/15-AOS1354. URL https://doi.org/10.1214/15-AOS1354.

Chao Gao, Yu Lu, and Harrison H. Zhou. Rate-optimal graphon estimation. *Ann. Statist.*, 43(6): 2624–2652, 2015b. ISSN 0090-5364,2168-8966. doi: 10.1214/15-AOS1354. URL https://doi.org/10.1214/15-AOS1354.

Chao Gao, Zongming Ma, Anderson Y. Zhang, and Harrison H. Zhou. Achieving optimal misclassification proportion in stochastic block models. *J. Mach. Learn. Res.*, 18:Paper No. 60, 45, 2017. ISSN 1532-4435. URL https://mathscinet.ams.org/mathscinet-getitem?mr=3687603.

Olivier Guédon and Roman Vershynin. Community detection in sparse networks via Grothendieck's inequality. *Probab. Theory Related Fields*, 165(3-4):1025–1049, 2016. ISSN 0178-8051. doi: 10.1007/s00440-015-0659-z. URL https://doi.org/10.1007/s00440-015-0659-z.

P.W. Holland, K.B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109 – 137, 1983. ISSN 0378-8733. doi: 10.1016/0378-8733(83)90021-7. URL http://www.sciencedirect.com/science/article/pii/0378873383900217.

Samuel Hopkins. *Statistical Inference and the Sum of Squares Method*. PhD thesis, Cornell University, 2018.

Samuel B Hopkins and David Steurer. Efficient bayesian estimation from few samples: community detection and related problems. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 379–390. IEEE, 2017.

Samuel B Hopkins, Pravesh K Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The power of sum-of-squares for detecting hidden structures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 720–731. IEEE, 2017.

S. Janson and E. Mossel. Robust reconstruction on trees is determined by the second eigenvalue. *Ann. Probab.*, 32:2630–2649, 2004. URL http://front.math.ucdavis.edu/0406.5447.

H. Kesten and B. P. Stigum. Additional limit theorems for indecomposable multidimensional Galton-Watson processes. *Ann. Math. Statist.*, 37:1463–1481, 1966.

Olga Klopp and Nicolas Verzelen. Optimal graphon estimation in cut distance. *Probability Theory and Related Fields*, 174(3):1033–1090, 2019. doi: 10.1007/s00440-018-0878-1. URL https://doi.org/10.1007/s00440-018-0878-1.

Olga Klopp, Alexandre B. Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. *Ann. Statist.*, 45(1):316–354, 2017a. ISSN 0090-5364,2168-8966. doi: 10.1214/16-AOS1454. URL https://doi.org/10.1214/16-AOS1454.

Olga Klopp, Alexandre B. Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics*, 45(1):316 – 354, 2017b. doi: 10.1214/16-AOS1454. URL https://doi.org/10.1214/16-AOS1454.

Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. In *ISAAC Congress (International Society for Analysis, its Applications and Computation)*, pages 1–50. Springer, 2019.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1):215–237, 2015. ISSN 0090-5364. doi: 10.1214/14-AOS1274. URL https://doi.org/10.1214/14-AOS1274.

László Lovász. *Large networks and graph limits*, volume 60 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2012. ISBN 978-0-8218-9085-1. doi: 10.1090/coll/060. URL https://doi.org/10.1090/coll/060.

Yuetian Luo and Chao Gao. Computational lower bounds for graphon estimation via low-degree polynomials. *arXiv preprint arXiv:2308.15728*, 2023.

Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703. ACM, 2014.

Frank McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.

Andrea Montanari. Finding one community in a sparse graph. *Journal of Statistical Physics*, 161 (2):273–299, 2015. doi: 10.1007/s10955-015-1338-2. URL https://doi.org/10.1007/s10955-015-1338-2.

E. Mossel. Survey: Information flow on trees. In J. Nestril and P. Winkler, editors, *Graphs, Morphisms and Statistical Physics. DIMACS series in discrete mathematics and theoretical computer science*, pages 155–170. 2004. URL http://front.math.ucdavis.edu/0406.5446.

E. Mossel and Y. Peres. Information flow on trees. *Ann. Appl. Probab.*, 13(3):817–844, 2003. ISSN 1050-5164. URL http://front.math.ucdavis.edu/math.PR/0107033.

E. Mossel, S. Roch, and A. Sly. On the inference of large phylogenies with long branches: How long is too long? *Bull. Math. Bio.*, 73(7):1627–1644, 2011.

Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probab. Theory Related Fields*, 162(3-4):431–461, 2015. ISSN 0178-8051. doi: 10.1007/s00440-014-0576-6. URL https://doi.org/10.1007/s00440-014-0576-6.

Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708, 2018. ISSN 0209-9683. doi: 10.1007/s00493-016-3238-8. URL https://doi.org/10.1007/s00493-016-3238-8.

Elchanan Mossel, Allan Sly, and Youngtak Sohn. Exact phase transitions for stochastic block models and reconstruction on trees. 12 2022. URL https://arxiv.org/pdf/2212.03362.pdf.

Sebastien Roch and Allan Sly. Phase transition in the sample complexity of likelihood-based phylogeny inference. *Probability Theory and Related Fields*, 169(1):3–62, 2017.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

Andrzej Ruciński. When are small subgraphs of a random graph normally distributed? *Probab. Theory Related Fields*, 78(1):1–10, 1988. ISSN 0178-8051. doi: 10.1007/BF00718031. URL https://doi.org/10.1007/BF00718031.

Tselil Schramm and Alexander S. Wein. Computational barriers to estimation from low-degree polynomials. *Ann. Statist.*, 50(3):1833–1858, 2022. ISSN 0090-5364. doi: 10.1214/22-aos2179. URL https://doi.org/10.1214/22-aos2179.

Allan Sly. Reconstruction for the Potts model. *Ann. Probab.*, 39(4):1365–1406, 2011. ISSN 0091-1798. doi: 10.1214/10-AOP584. URL https://doi.org/10.1214/10-AOP584.

T.A.B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100, 1997.

Youngtak Sohn and Alexander S. Wein. Sharp phase transitions in estimation with low-degree polynomials. *Symposium on Theory of Computing*, 2025.

Ludovic Stephan and Yizhe Zhu. Community detection with the bethe-hessian. 11 2024. URL https://arxiv.org/pdf/2411.02835.pdf.

Patrick J. Wolfe and Sofia C. Olhede. Nonparametric graphon estimation, 2013. URL https://arxiv.org/abs/1309.5936.

Jiaming Xu. Rates of convergence of spectral methods for graphon estimation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5433–5442. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/xu18a.html.