Fundamental limits for network community detection

Mung Chiang[1], **Beate Franke**[2], **H. Vincent Poor**[1],
Devavrat Shah[3], **Jacob Shapiro**[1], **Patrick J. Wolfe**[2]
[1]Princeton University, [2]University College London, [3]Massachusetts Institute of Technology
beate.franke.12@ucl.ac.uk

Introduction

THE structure of many networks, including social networks, is strongly influenced by a natural division into communities. With over 4000 citations, one of the the most popular methods to evaluate community structure was suggested by Newman and Girvan [1]. They introduced a quality function called "modularity", which computes for a given community assignment the difference between the observed edges within those communities, and the expected number of edges in the absence of community structure. Finding an optimal community assignment requires maximizing modularity over all possible assignments—an NP-hard problem—but suboptimal approaches have proved practical.

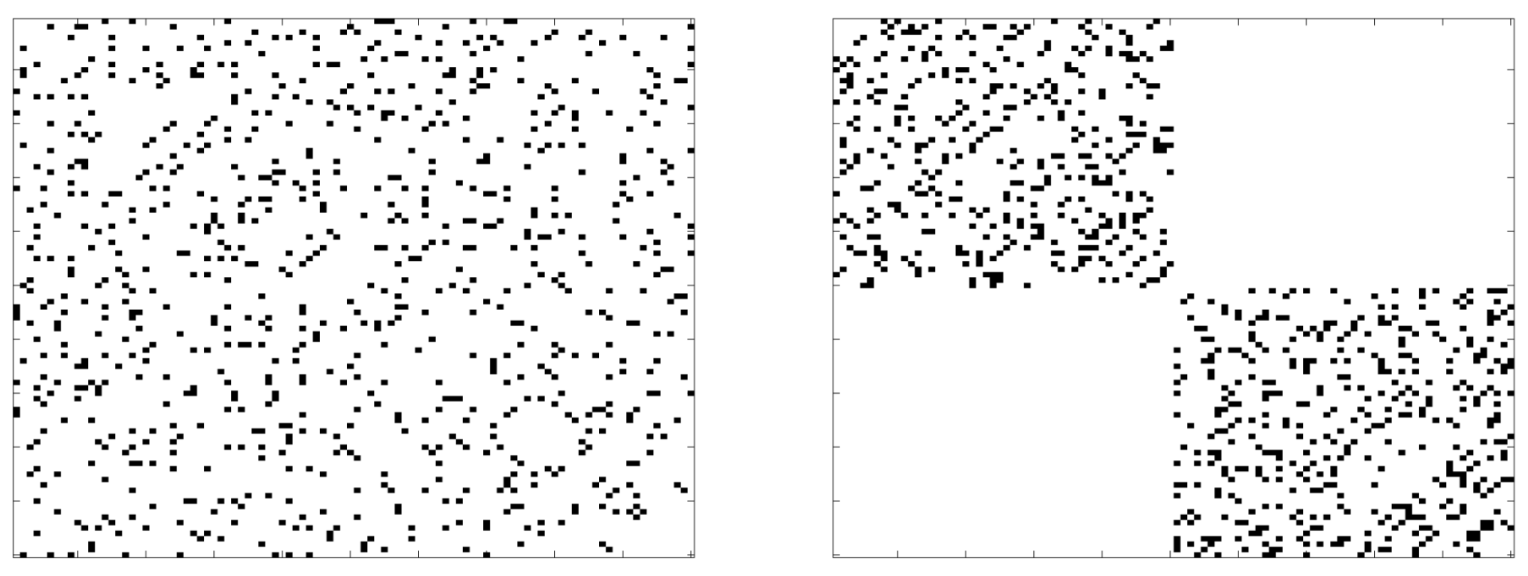

Figure 1: *Two representations of the same two-community network*

Theoretical Background

We model simple graphs on n nodes by an adjacency matrix $\boldsymbol{A}$ as follows:

$$A_{ij} = \mathrm{Bernoulli}(p_{ij}), \quad 1 \le i < j \le n,$$

with $A_{ji} = A_{ij}$ and $A_{ii} = 0$. Letting A_{i+} denote the degree of node i, the Newman–Girvan modularity is then given by

$$Q(g) = \sum_{i<j} \left(A_{ij} - \frac{A_{i+}A_{j+}}{A_{++}} \right) \mathbb{1}(g_i, g_j),$$

where $\mathbb{1}(g_i, g_j) = 1$ when nodes i and j are in the same group, and 0 otherwise.

If no community structure is present and we assume the edge probabilities p_{ij} to be small, then the following holds [2]:

The estimator $\hat{p}_{ij} = \frac{A_{i+}A_{j+}}{A_{++}}$ is a near-maximum likelihood estimator of p_{ij} .

Modularity $Q(g)$ thus measures the difference between fitting the Chung–Lu model [3] and a stochastic blockmodel for any given community assignment g.

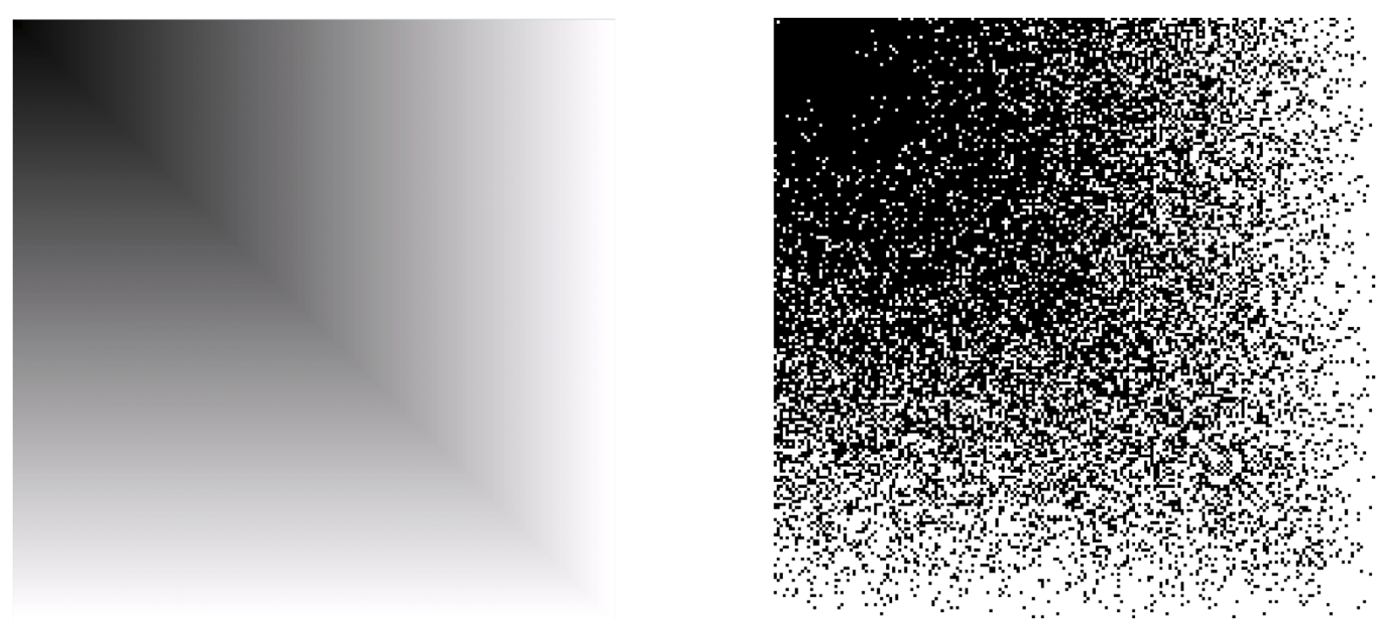

Figure 2: *Expected versus realized edges under the Chung–Lu model [4]*

Aim: Fundamental Limits

Bickel and Chen [5] prove that there are cases where the estimator for community assignment $\hat{g}$ derived by maximizing $Q(g)$ fails to recover the true assignments. A more detailed analysis of the behavior of $Q(g)$ for large n is therefore necessary to clarify the fundamental limits of community detection.

For the null model of no community structure, each $\hat{p}_{ij} = A_{i+}A_{j+}/A_{++}$ can be reparameterized to yield the parameter estimates $\hat{\pi}_i = A_{i+}/\sqrt{A_{++}}$; i.e.,

$$\hat{p}_{ij} = \frac{A_{i+}A_{j+}}{A_{++}} = \frac{A_{i+}}{\sqrt{A_{++}}}\frac{A_{j+}}{\sqrt{A_{++}}} = \hat{\pi}_i\hat{\pi}_j.$$

Thus, to understand the behavior of $Q(g)$, we need to discuss the properties of $\hat{\pi}_i$.

Theoretical Results

Theorem 1 (CLT) *Assume that the ith network degree's variance $\mathrm{Var}(A_{i+})$ grows in n, the number of network nodes. Then as n becomes large, $\hat{\pi}_i$ is approximately distributed as a $\mathrm{Normal}(\pi_i, \frac{\pi_i}{||\pi||_1} - \pi_i^2\frac{||\pi||_2^2}{||\pi||_1^2})$ random variable.*

In general, to apply this central limit theorem (CLT), we require the variance term $\frac{\pi_i}{||\pi||_1} - \pi_i^2\frac{||\pi||_2^2}{||\pi||_1^2}$, which cannot be observed. But if we can derive an estimator $\widehat{\mathrm{Var}}(\hat{\pi}_i)$ that is close to its true value for large n, then Theorem 1 yields a confidence interval (CI) for $\hat{\pi}_i$ of the form

$$\mathrm{CI} = (\hat{\pi}_i - \delta, \hat{\pi}_i + \delta),$$

with $\delta = z_{1-\alpha/2} \cdot \sqrt{\widehat{\mathrm{Var}}(\hat{\pi}_i)}$, and $z_{1-\alpha/2}$ the $1-\frac{\alpha}{2}$ quantile of a $\mathrm{Normal}(0,1)$ distribution.

Power Laws

For example, let us assume a power law setting; i.e., $\pi_i \propto i^{-\gamma}$, $0 < \gamma < 1$:

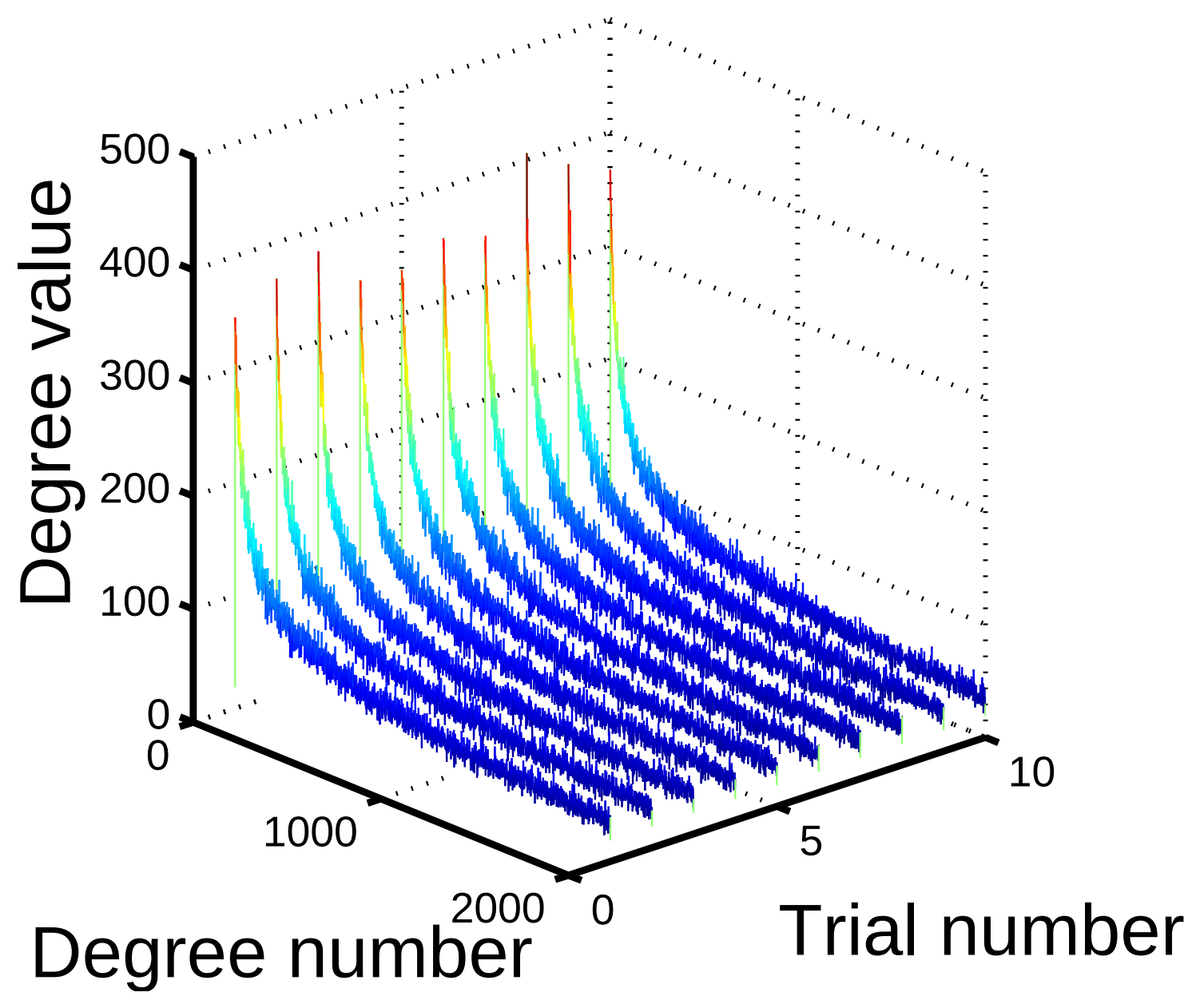

Figure 3: *Network replicates shown for the power law setting in which $\pi_i \propto i^{-1/2}$*

We then have the following theorem:

Theorem 2 (CI) *Whenever $\pi_i \propto i^{-\gamma}$ and the expected value of the ith degree is growing in n, then $\frac{\pi_i}{||\pi||_1} - \pi_i^2\frac{||\pi||_2^2}{||\pi||_1^2}$ becomes close to its empirical version $\frac{\hat{\pi}_i}{||\hat{\pi}||_1} - \hat{\pi}_i^2\frac{||\hat{\pi}||_2^2}{||\hat{\pi}||_1^2}$.*

Figure 4 illustrates how we may approximate $\hat{\pi}_i$ in this setting, using the central limit theorem of Theorem 1.

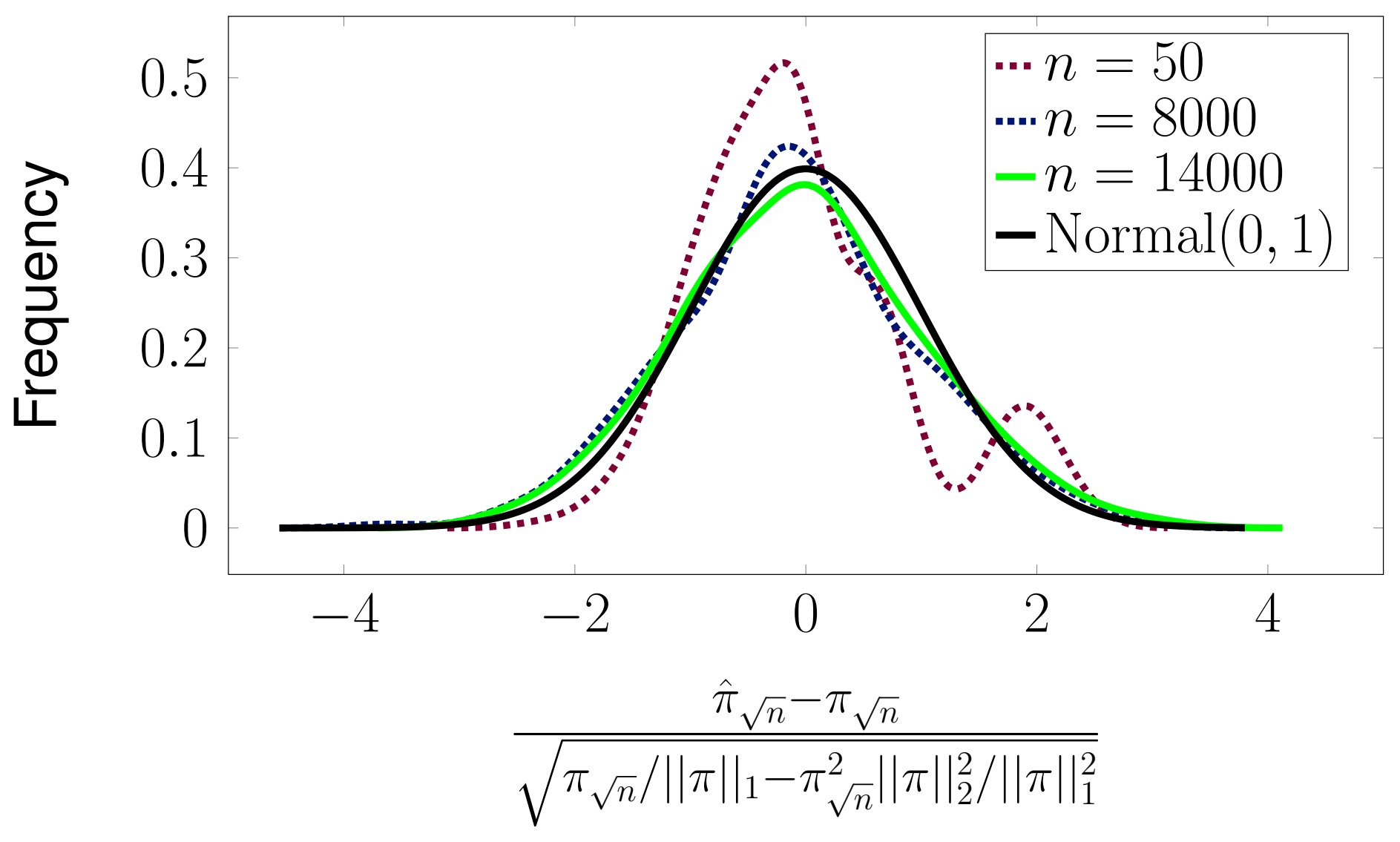

Figure 4: *Empirical densities of standardized estimators and a unit Normal, for the power law setting of [6] ($\gamma = 1/100$)*

It follows from Theorem 2 that in this case, we obtain reliable confidence intervals for the estimator $\hat{\pi}_i$ as n grows large:

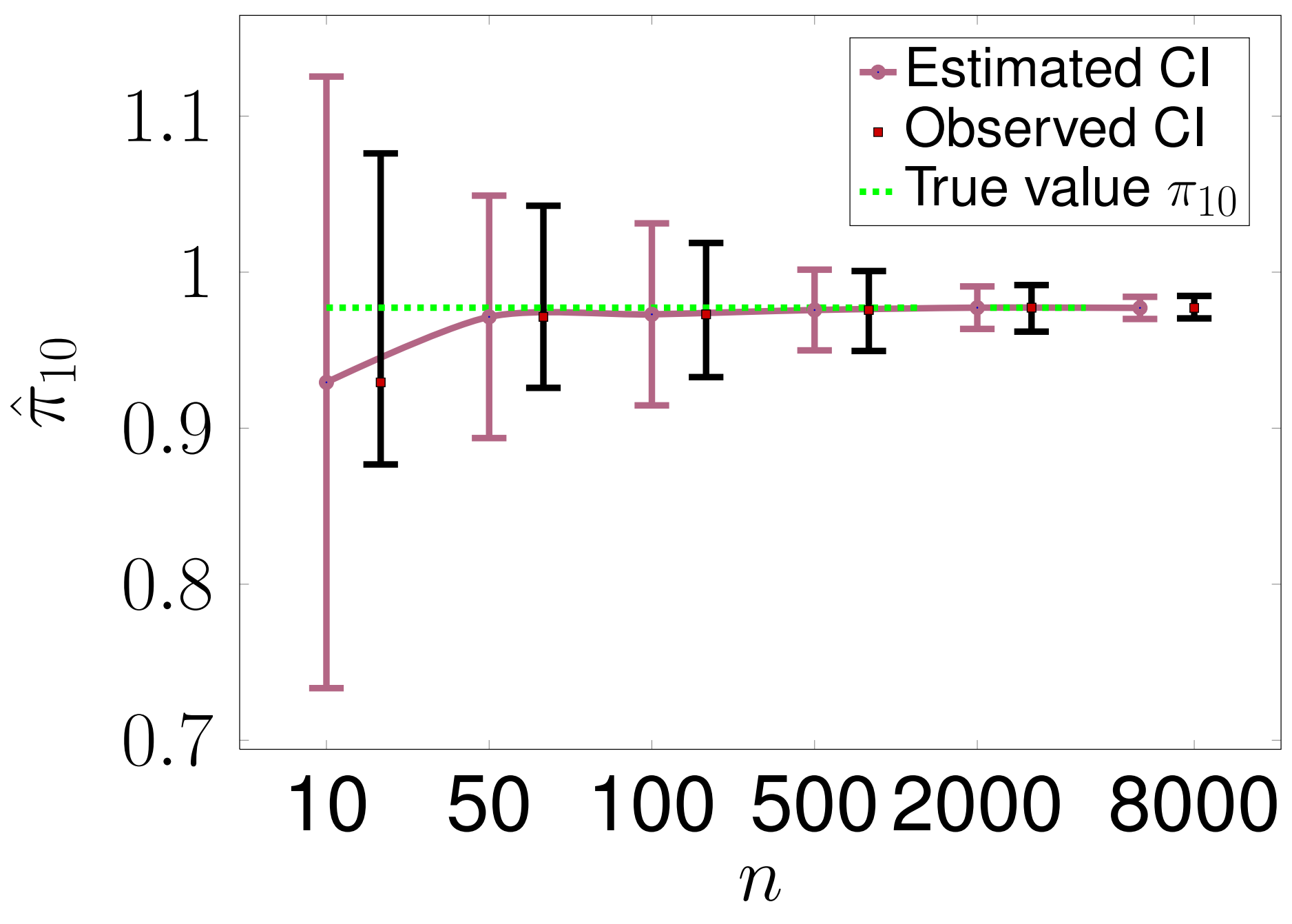

Figure 5: *The estimator $\hat{\pi}_{10}$, shown along with its actual and empirical large-sample confidence intervals ($\gamma = 1/100$)*

Future Directions

1. Study asymptotic distribution of $Q(g)$.
2. Use side information to identify communities (→ Chiang et al. [7]).
3. Optimize Newman–Girvan community detection algorithms by optimizing modularity in each iteration only inside local neighborhoods (→ Shah et al.).

References

[1] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, pp. 1–15, 2004.

[2] P. O. Perry and P. J. Wolfe, "Null models for network data," *arXiv preprint arXiv:1201.5871*, 2012.

[3] F. Chung, L. Lu, and V. Vu, "Spectra of random graphs with given expected degrees," *Proc. Natl. Acad. Sci. USA*, vol. 100, pp. 6313–6318, 2003.

[4] L. Lovász, *Large Networks and Graph Limits*, American Mathematical Society, Providence, RI, 2012.

[5] P. J. Bickel and A. Chen, "A nonparametric view of network models and Newman–Girvan and other modularities," *Proc. Natl. Acad. Sci. USA*, vol. 106, pp. 21068–21073, 2009.

[6] S. C. Olhede and P. J. Wolfe, "What is a degree distribution?," *arXiv preprint arXiv:1211.6537*, 2012.

[7] C. S. Leberknight, A. Tajer, M. Chiang, and H. V. Poor, "Identifying online communities of interest using side Information," *IEEE Statistical Signal Processing Workshop*, pp. 137–140, 2012.

Acknowledgements: *Acknowledgements: Work supported in part by the US Army Research Office under PECASE Award W911NF-09-1-0555 and MURI Award 58153-MA-MUR; by the UK EPSRC under Mathematical Sciences Established Career Fellowship EP/K005413/1 and Institutional Sponsorship Award EP/K503459/1; by the UK Royal Society under a Wolfson Research Merit Award; and by Marie Curie FP7 Integration Grant PCIG12-GA-2012-334622 within the 7th European Union Framework Program.*

SUNBELT 2013, 2013 Social Networks Conference of the International Network for Social Network Analysis, May 24, 2013, Hamburg, Germany