MOMENTS OF UNIFORMLY RANDOM MULTIGRAPHS WITH FIXED DEGREE SEQUENCES

Phil Chodrow

SIAM Workshop on Network Science 2020
July 9–10 · Toronto

Summary

We study the expected adjacency matrix of a uniformly random multigraph with fixed degree sequence $\mathbf{d}$. Although this matrix is central to several standard network analysis techniques, including modularity-maximization and mean-field theories, its general structure is not well understood. We use a dynamical argument to derive an estimator of the expected adjacency matrix and several other moments; supply an algorithm to compute the estimator; demonstrate its accuracy on empirical data; and illustrate the impact of using this estimator on a simple modularity maximization task.

Beyond Sparsity

Let $\mathcal{G}_{\mathbf{d}}$ denote the set of multigraphs without self-loops, and let $\eta_{\mathbf{d}}$ be the uniform distribution on $\mathcal{G}_{\mathbf{d}}$. Let $\mathbf{W}$ be the (random) adjacency matrix of a graph distributed according to $\eta_{\mathbf{d}}$. We are interested in the matrix $\boldsymbol{\omega} = \mathbb{E}[\mathbf{W}]$.

This problem is relatively well-understood in the large, sparse limit. In this case, the moments of $\eta_{\mathbf{d}}$ converge to those of the configuration model [2]. We then have

$$\omega_{ij} \approx \hat{\omega}^0_{ij} = \frac{d_i d_j}{\sum_\ell d_\ell}\ , \tag{1}$$

where the approximation can be made asymptotically precise. Many follow-up works in the network science literature describe the matrix $\hat{\boldsymbol{\omega}}^0$ as the expectation of "a random graph with fixed degree sequence," but this statement is not exactly true for any common such model, and is not usually accompanied by error bounds. Indeed, as Figure 1(a) illustrates, $\hat{\boldsymbol{\omega}}^0$ can behave quite poorly. On this test data set, a subset of `contact-high-school` supplied by [1, 4], $\hat{\boldsymbol{\omega}}^0$ is off by an average of 25% when benchmarked against a Markov Chain Monte Carlo estimate of the ground truth. Worse, there is persistent bias, with $\hat{\boldsymbol{\omega}}^0$ overestimating edges between vertices of dissimilar degrees and underestimating between vertices of similar degrees. This poor performance reflects the fact that this data set, like many of practical interest, does not lie in the large sparse regime. The test data contains 268 nodes and $10,206$ edges, and contains several nodes with degree larger than n. Unfortunately, this density also makes it extremely intensive to estimate $\boldsymbol{\omega}$ by Monte Carlo methods [3]; the estimate in Fig. 1 was computed in approximately one week on a single core of a standard server. We therefore seek tractable ways to estimate $\boldsymbol{\omega}$ with practical accuracy and computational load.

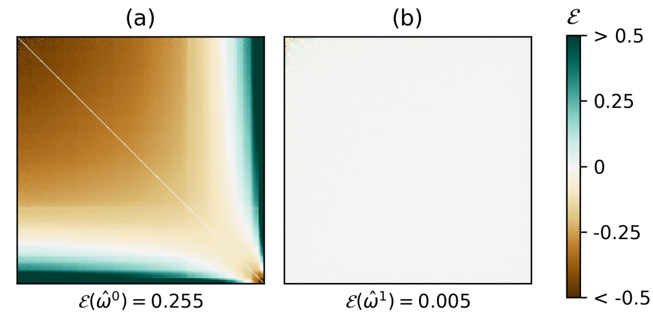

Figure 1: Error comparison of the estimators $\hat{\boldsymbol{\omega}}^0$ (eq. 1) and $\hat{\boldsymbol{\omega}}^1$ (derived in the present work). The degrees of nodes increase top to bottom and left to right. Shading gives the relative error in each entry when using each estimator to estimate $\boldsymbol{\omega}$, which in this case was estimated using 10^7 rounds of Markov Chain Monte Carlo. Data is a subset of `contact-high-school` [1], further described in Fig. 2.

A Dynamical Approach to Model Moments

We will state our results informally; all approximations can be given precise bounds under a simple conjecture on the structure of $\eta_{\mathbf{d}}$. By treating the MCMC sampler as a stochastic dynamical system whose state space is $\mathcal{G}_{\mathbf{d}}$, we derive stationarity conditions describing the moments of $\eta_{\mathbf{d}}$. As we show, there exists a vector $\boldsymbol{\beta} \in \mathbb{R}^n_+$ such that

$$\chi_{ij} \triangleq \eta_{\mathbf{d}}(W_{ij} \geq 1) \approx \frac{\beta_i \beta_j}{\sum_i \beta_i} = f_{ij}(\boldsymbol{\beta}) \tag{2}$$

for all $i \neq j$. Furthermore, the first moment of $\mathbf{W}$ under $\eta_{\mathbf{d}}$ is approximately

$$\omega_{ij} \approx \frac{\chi_{ij}}{1 - \chi_{ij}}\ . \tag{3}$$

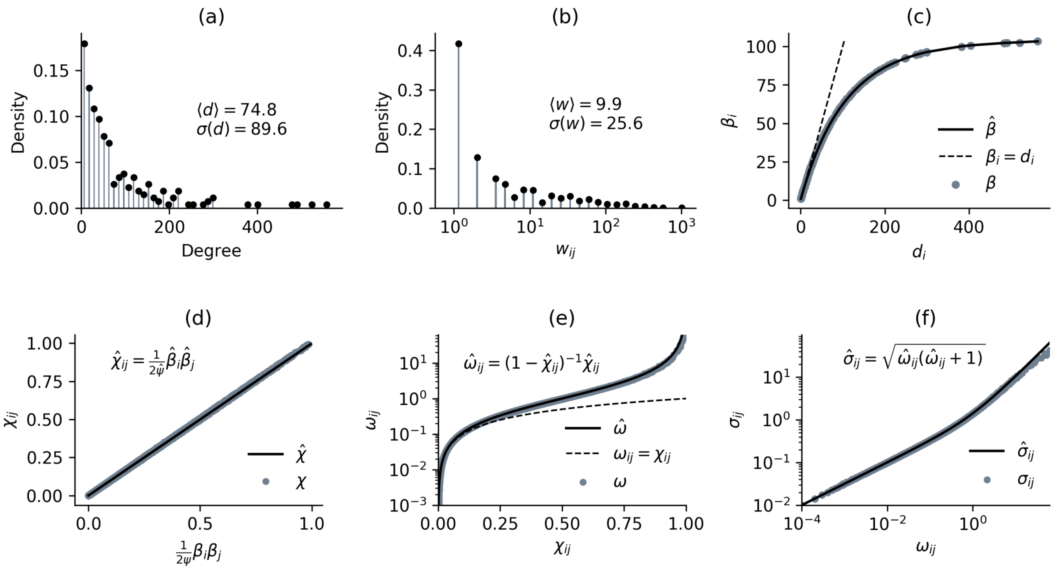

Figure 2: (a): Degree distribution of the `contact-high-school` subnetwork. (b): Distribution of the entries of $\mathbf{w}$. (c): Collapsed degree sequence $\hat{\boldsymbol{\beta}}$ learned from $\mathbf{d}$ by solving (4). (d): Approximation of $\boldsymbol{\chi}$ via (2). (e): Approximation of $\boldsymbol{\omega}$ via (3). (f): Approximation of $\sigma_{ij} = \sigma(W_{ij})$.

Taken to together, these two equations provide a method for computing an estimate of $\boldsymbol{\omega}$ given knowledge of the vector $\boldsymbol{\beta}$. We construct an estimator $\hat{\boldsymbol{\beta}}$ of this vector by solving the system of n equations

$$\sum_j \frac{f_{ij}(\boldsymbol{\beta})}{1 - f_{ij}(\boldsymbol{\beta})} = d_i \,, \quad i = 1, \ldots, n \,. \tag{4}$$

We prove a qualified uniqueness result on solutions of (4), define an estimator $\hat{\boldsymbol{\omega}}^1$ as its solution, and provide an iterative algorithm for computing this estimator. Figure 2 provides some description of the test network and illustrates the construction of the estimator $\hat{\boldsymbol{\omega}}^1$. The accuracy of the derived estimator is highly favorable when compared to the MCMC results. Indeed, the mean relative error of $\hat{\boldsymbol{\omega}}^1$ is less than 2% (Figure 1(a)), with no visible systematic bias.

We illustrate the importance of these results for downstream data analysis with a vignette on spectral modularity maximization [5] when using the estimators $\hat{\boldsymbol{\omega}}^0$ and $\hat{\boldsymbol{\omega}}^1$ for the null matrix. We show that the behavior and performance of this algorithm are sensitive to both the choice of estimator and the data used, highlighting the need for analysts to carefully specify null models when performing null-based data analysis.

References

[1] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg. Simplicial closure and higher-order link prediction. *Proceedings of the National Academy of Sciences*, 115(48):11221–11230, 2018.

[2] B. Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1(4):311–316, 1980.

[3] B. K. Fosdick, D. B. Larremore, J. Nishimura, and J. Ugander. Configuring random graph models with fixed degree sequences. *SIAM Review*, 60(2):315–355, 2018.

[4] R. Mastrandrea, J. Fournet, and A. Barrat. Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLOS ONE*, 10(9), 2015.

[5] X. Zhang and M. E. J. Newman. Multiway spectral community detection in networks. *Phys. Rev. E*, 92:052808, Nov 2015.