Spectral signatures of breaking of ensemble equivalence

Dionigi, P.

CHAPTER 5

Sampling random graph models

Abstract

In this Chapter we give a brief introduction to the problem of random graph sampling and we will show simulations that support our findings of the previous Chapters. Simulations were performed using the computational resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University.

In Chapter 1 we spoke about the strong influence the abundance of real-world data had on the flourishing of Network Science. The interplay between models and data validation caused Network Science to emerge as a powerful interdisciplinary field that studies the structure, dynamics and behavior of complex systems represented as networks. As we pointed out, these networks can range from social interactions and biological systems to technological infrastructures. Each network has peculiar features that need to be captured by mathematical models that aim to emulate reality.Typically, the size of the networks of interest is very large, and as a consequence there is no hope to fully reconstruct real-world network structures from the data. Indeed, with the large size of the networks come many problems, such as the impossibility to gather all the data needed, the amount of time and costs that this would take, as well as accuracy and storage problems. Therefore, in the realm of Network Science, sampling random graphs from given distributions plays a crucial role, offering researchers a practical and efficient way to gain insight into large-scale networks after the main features (i.e. the ones that are easily accessible) have been incorporated. Once the model is chosen and is found to recreate the observed datas, it becomes an efficient *null model* that can be used to test whether new gathered data are consistent with it or require more sophisticated models.

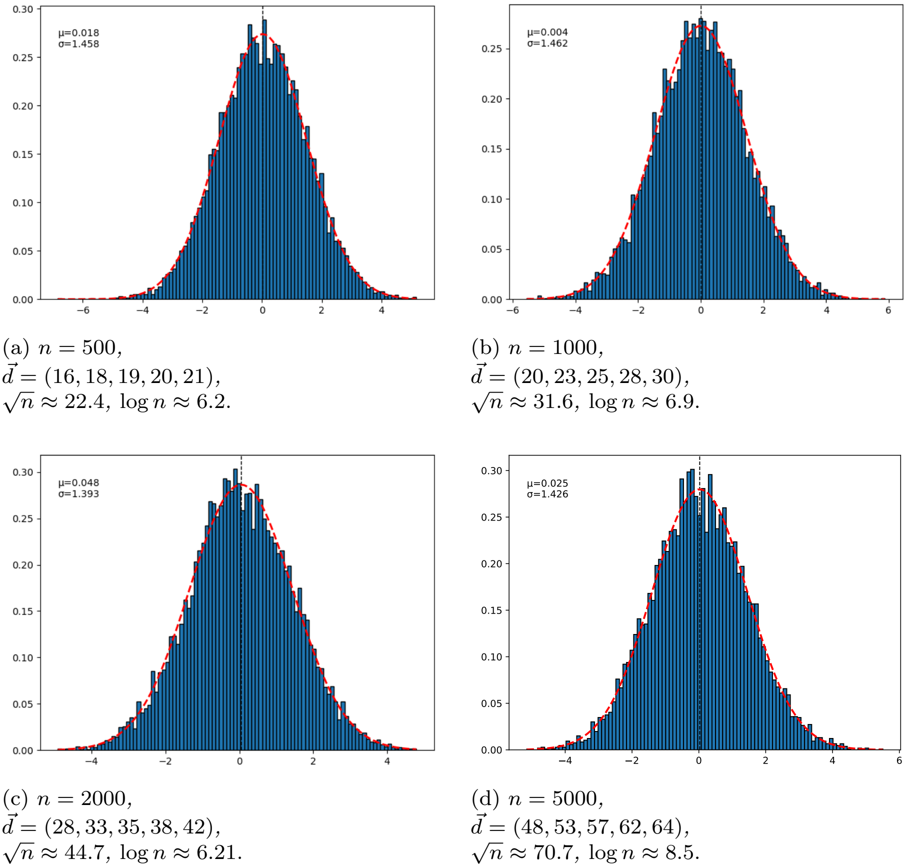
As we already discussed, Network Science provides many versatile models that are able to capture many different features. Sampling-wise a difference needs to be made between sampling from distributions with soft constraints and from distributions with hard constraint. Ultimately, we will specialize our discussion to the type of constraints that we analyzed in this thesis, i.e., constraints on the degree sequence.

§5.1 Sampling from the Canonical Ensemble

Following [14, 16], fast sampling of the canonical model with constraint $\vec{C}(g)$ can be obtained once the Shannon entropy maximization problem has been solved. Indeed, once the functional form of the $p_{ij}\left(\vec{\theta}\right)$ as in (1.6.1) is obtained, it is easy to calculate the value of the Lagrange multipliers $\vec{\theta}$ through maximum likelihood. The precise value of $\vec{\theta}$ needed to express (1.6.1) must be chosen in order to match with what has been measured from data. This is obtained by requiring that the logarithm of the probability of observing $\vec{C}^*$ given $\vec{\theta}$ is maximal, i.e.,

$$\max_{\vec{\theta}} \ln \mathbb{P}\left(\vec{C}(g) = \vec{C}^* \middle| \vec{\theta}\right).$$

This is possible only when the dyadic probabilities p_{ij} can be expressed in closed form from the entropy maximization. When this is not the case, other sampling procedures should be taken into account, most of them based on Monte Carlo approaches (for example, *Hamiltonian Monte Carlo* [1, 2]), or mean-field approaches (for example, the solution for the Strauss model in [13]). See [3] for more examples. Constraints on the degree sequence, i.e., the ones used in this thesis, allow for an explicit form of (1.6.1), even in the directed case with reciprocity or weights. We will therefore use the methodology and the packages developed in [14].



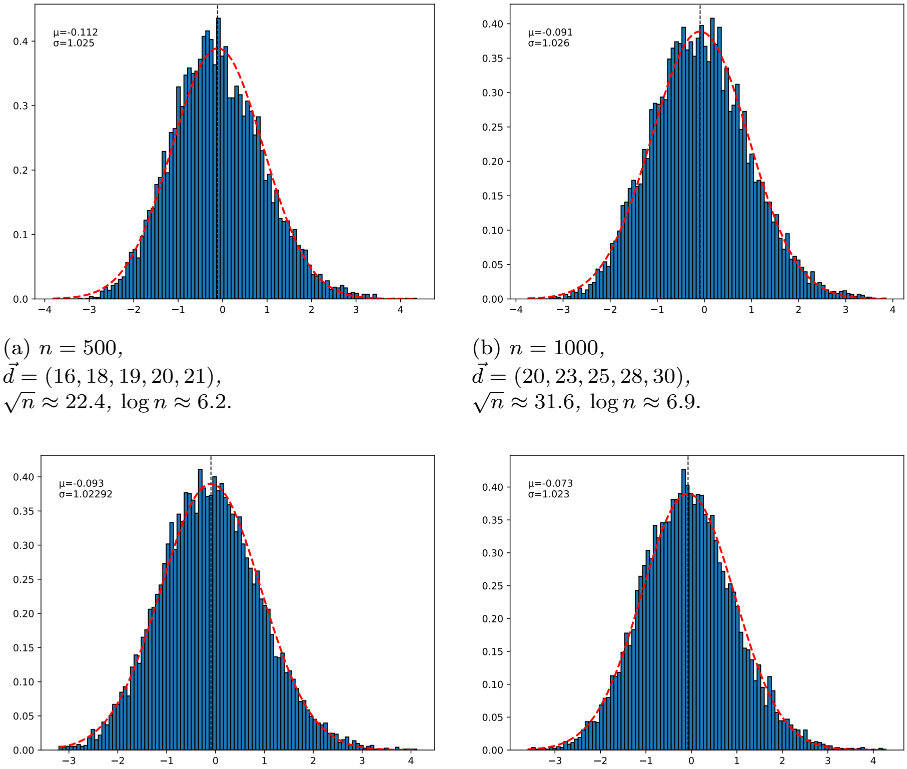
(a) $n = 500$,
$\vec{d} = (16, 18, 19, 20, 21)$,
$\sqrt{n} \approx 22.4,\ \log n \approx 6.2.$

(b) $n = 1000$,
$\vec{d} = (20, 23, 25, 28, 30)$,
$\sqrt{n} \approx 31.6,\ \log n \approx 6.9.$

(c) $n = 2000$,
$\vec{d} = (28, 33, 35, 38, 42)$,
$\sqrt{n} \approx 44.7,\ \log n \approx 6.21.$

(d) $n = 5000$,
$\vec{d} = (48, 53, 57, 62, 64)$,
$\sqrt{n} \approx 70.7,\ \log n \approx 8.5.$

Figure 5.1: Histograms of $\bar{\lambda}_1$ for different graph sizes n and degree sequences $\vec{d}$. The sample size for each regime is 10^4. Each element specified in the degree sequence appears $\frac{n}{5}$ times. In red is plotted the Gaussian fit; μ is the sample mean (represented by a dashed line in the histogram), σ is the sample standard deviation. We expect $\mu \approx 0$ and $\sigma \approx \sqrt{2}$.



(a) $n = 500$,
$\vec{d} = (16, 18, 19, 20, 21)$,
$\sqrt{n} \approx 22.4,\ \log n \approx 6.2$.

(b) $n = 1000$,
$\vec{d} = (20, 23, 25, 28, 30)$,
$\sqrt{n} \approx 31.6,\ \log n \approx 6.9$.

(c) $n = 2000$,
$\vec{d} = (28, 33, 35, 38, 42)$,
$\sqrt{n} \approx 44.7,\ \log n \approx 6.21$.

(d) $n = 5000$,
$\vec{d} = (48, 53, 57, 62, 64)$,
$\sqrt{n} \approx 70.7,\ \log n \approx 8.5$.

Figure 5.2: Histograms of $\bar{v}_1(i)$ for different graph sizes n and degree sequences $\vec{d}$. For each of the images, i is chosen to be the last i such that d_i is equal to the 4^{th} element of the corresponding degree sequence (e.g. for $n = 500$, $v_1(400)$ was analysed with $d_{400} = 20$). The sample size for each regime is 10^4. Each element in the degree sequence appears $\frac{n}{5}$ times. In red is plotted the Gaussian fit; μ is the sample mean (represented by a dashed line in the histogram), σ is the sample standard deviation. We expect $\mu \approx 0$ and $\sigma \approx 1$.

.

§5.1.1 Simulations of results of Chapter 3: Largest eigenvalue.

Theorems 3.1.6–3.1.7 show that, after proper scaling and under certain conditions of sparsity and homogeneity, the largest eigenvalue and the components of the largest eigenvector exhibit Gaussian behaviour in the limit as $n \to \infty$. A natural question is how these quantities behave for finite n. Indeed, real-world networks have sizes that range from $n = 10^2$ to $n = 10^9$. Another question is computational feasibility. Indeed, our CLTs require the degrees to lie between $(\log n)^4$ (respectively, $(\log n)^8$) and $\sqrt{n}$. In order to make this possible, n must be at least 10^{11} (respectively, 10^{29}), which is unrealistic. Let us therefore see what simulations have to say[1].

In Figure 5.1 we show histograms for the quantity

$$\bar{\lambda}_1 = \frac{m_2}{m_1 \sigma_1}(\lambda_1 - \mathbb{E}[\lambda_1]),$$

which should be close to normal with mean 0 and variance 2 (for $\mathbb{E}[\lambda_1]$ the correction term o(1) is neglected). The convergence is fast: already for $n = 500$ the Gaussian shape emerges and represents an excellent fit: the sample mean μ is close to 0 and the sample standard deviation σ is close to $\sqrt{2}$.

§5.1.2 Largest eigenvector.

In Figure 5.2 we show histograms for the quantity

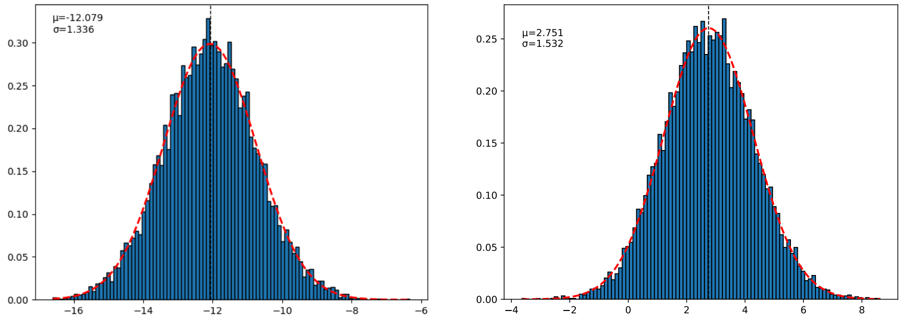
$$\bar{v}_1(i) = \frac{m_2^{3/2}}{m_1 s_1(i)} \left(v_1(i) - d_i/\sqrt{m_2}\,\right),$$

which should be close to normal with mean 0 and variance 1. The fit is again excellent.

§5.1.3 Degrees of order $\log n$ and $\sqrt{n}$.

What happens when the degrees are of order $\log n$? As can be seen in Figure 5.3, in that range the Gaussian approximation for the largest eigenvalue is visibly worse, especially for the centering. The same happens for the components of the largest eigenvector, as can be seen in Figure 5.4, where the Gaussian shape is lost and two peaks appear.

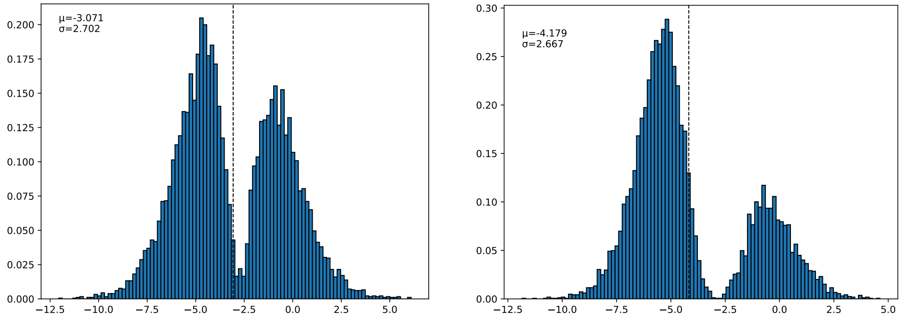
[1] Simulations were performed using the computational resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University.



(a) $n = 500$,
$\vec{d} = (4, 5, 8, 9)$,
$\sqrt{n} \approx 22.4$, $\log n \approx 6.2$.

(b) $n = 5000$,
$\vec{d} = (6, 7, 9, 10)$,
$\sqrt{n} \approx 70.7$, $\log n \approx 8.5$.

Figure 5.3: Histograms of $\bar{\lambda}_1$ for different graph sizes n and degree sequences $\vec{d}$ of order $\log n$*. The sample size for each regime is* 10^4*. Each element specified in the degree sequence appears $\frac{n}{4}$ times. In red is plotted the Gaussian fit; μ is the sample mean (represented by a dashed line in the histogram), σ is the sample standard deviation. If Theorem 3.1.6 would hold, then we would expect $\mu \approx 0$ and $\sigma \approx \sqrt{2}$.*



(a) $n = 500$,
$\vec{d} = (4, 5, 8, 9)$,
$\sqrt{n} \approx 22.4$, $\log n \approx 6.2$.

(b) $n = 5000$,
$\vec{d} = (6, 7, 9, 10)$,
$\sqrt{n} \approx 70.7$, $\log n \approx 8.5$.

Figure 5.4: Histograms of $\bar{v}_1(i)$ for different graph sizes n and degree sequences $\vec{d}$ of order $\log n$*. For each of the images, i has been chosen to be the last i such that d_i is equal to the 3^{rd} element of the specified degree sequence (e.g. for* $n = 500$, $v_1(375)$ *was analysed with* $d_{375} = 8$*). The sample size for each regime is* 10^4*. Each element specified in the degree sequence appears $\frac{n}{4}$; μ is the sample mean (represented by a dashed line in the histogram), σ is the sample standard deviation. If Theorem 3.1.7 would hold, then we would expect $\mu \approx 0$ and $\sigma \approx 1$.*

§5.2 Sampling from the Microcanonical Ensemble

Sampling from uniform distributions is known to be an hard problem. The main reason for this, in graph theory, is the difficulty in estimating the cardinality of the support (1.5.3) of the uniform distribution. Many approximate procedures have been developed over time to overcome this obstacle. In general, there is an interplay between biased sampling and complexity of the algorithm. While most of the procedures to sample with accuracy from $\Gamma_{\vec{C}^*}$ require an exponential complexity time, faster procedures rely on Monte Carlo approaches that suffer from two related types of problems: bias and ergodicity. The latter refers to the fact that, depending on the constraints on the dynamics of the Markov Chain, there can be configurations that are never visited by the MCMC. Biased sampling refers to the fact that our MCMC might sample certain graphs with higher probability, e.g. because of a lack of ergodicity or a high mixing time for the Markov chain. This is usually solved, when possible, by introducing importance sampling. For the case when the constraint is on the degree sequence fast algorithms are available. These algorithm usually are divided into two steps: the first generates an *unbiased* seed which then is fed to the MCMC for the second step. For the case of constraints on the degree sequence the MCMC is shown to mix fast enough to make this method efficient. Usually the shortcoming of these approaches is the limitations on the density and inhomogeneity of our random graphs.

Because of its importance in graph theory, sampling graphs with a given degree sequence, i.e., when the constraint is on the degree sequence, has been studied since the 60s. A good reference is [5]. Three main approaches are used. One is the use of configuration model, which was shown in [8] to be efficient to generate simple graphs only when $m_2 = \mathrm{O}(\frac{m_1}{2}\sqrt{\log n})$ and $\max_i d_i = \mathrm{o}(m_1/2)$ (for example, when $\max_i d_i = \mathrm{O}(\sqrt{\log n})$). A rejection sampling when the degree sequence is above these thresholds will lead to exponential complexity. To overcome this difficulty in [12] Wormald and McKay showed a way to sample from the set of simple graphs with a given degree sequence by implementing switching-based algorithms. In the case of an homogeneous degree sequence, i.e., $d_i = d$ for all i, the microcanonical ensemble coincides with the random d-regular graph model. The problem was solved by implementing the switching algorithm of [12] and was perfectioned in [11, 15, 4]. The algorithm is efficient when $d = \mathrm{O}(n^{-1/3})$. For inhomogeneous degree sequences, it was proved in [9] that when

$$\max_i d_i = \mathrm{o}(\sqrt{n}), \quad m_1 = \Theta(n), \quad m_2 = \mathrm{O}(n),$$

the switching algorithm asymptotically provides a uniform sampling. For the directed case a sequential stub-matching procedure was shown in [17] to lead to asymptotic uniform sampling when $\max_i d_i = \mathrm{O}(m_1^{1/4-\varepsilon})$, provided $m_1 = \sum_i d_i^{\text{in}} = \sum_i d_i^{\text{out}}$.

MCMC methods usually rely on switching chain dynamics performed on a seed graph generated via the *Havel-Hakimi* algorithm [6, 7]. The details of this method and its variations can be found in [5, Chapter 6]. In particular, it was shown that the mixing properties of the Markov chain are linked to *P-stability* of the degree sequence [10].

In our simulations, given the relatively small size of the graphs and the low density and inhomogeneity of the degree sequences taken in account, we opted for a rejection sampling using the configuration model. In Table 5.1 we report the details of the simulated graphs and the rejection rate.

Configuration Model					
Size n	Degree Sequence	Mean Degree	$\sqrt{n}$	$\log n$	Rejection Rate
1000	$\vec{d} = (20, 23, 25, 28, 30)$	25.2	≈ 31.6	≈ 6.9	1.67%
2000	$\vec{d} = (28, 33, 35, 38, 42)$	35.2	≈ 44.7	≈ 7.6	1.35%
5000	$\vec{d} = (48, 53, 57, 62, 64)$	56.8	≈ 70.7	≈ 8.5	0.95%
10000	$\vec{d} = (78, 80, 83, 87, 90)$	83.6	100	≈ 9.2	0.73%

Table 5.1: Configuration model that have been sampled. Each element specified in the degree sequence appears $\frac{n}{5}$ times. The rejection rate has been obtained by sampling 10000 graphs for each different size and degree sequence and counting the non simple realizations.

§5.2.1 Simulations of results of Chapter 4: Largest eigenvalue.

In Theorem 4.1.2 we proved that the expectation of λ_1 in the configuration model conditioned on simplicity satisfies

$$\mathbb{E}[\lambda_1] = \frac{m_2}{m_1} + \frac{m_1 m_3}{m_2^2} - 1 + o(1), \qquad n \to \infty.$$

In Figure 5.5 and Figure 5.6 we plot

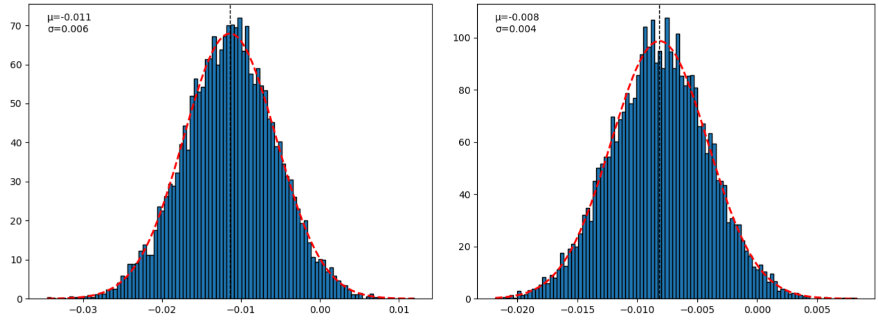
$$\bar{\lambda}_1 = \lambda_1 - \mathbb{E}[\lambda_1]$$

for some degree sequences compatible with the ones studied in Section 5.1.1. It can be seen that, with an increasing size of the graph, the error in the above formula becomes smaller and smaller. Furthermore it can be seen that the empirical standard deviation of λ_1 is much smaller than the one calculated for the Chung-Lu model from the formula for σ^2 in Theorem 3.1.6.

To capture the difference between the largest eigenvalues of the models in Chapter 3 and Chapter 4 we can define the following quantity on the probability space formed by the product measure of the two models

$$\hat{\lambda}_1 = \lambda_1^{\mathrm{CL}} - \lambda_1^{\mathrm{CM}}.$$

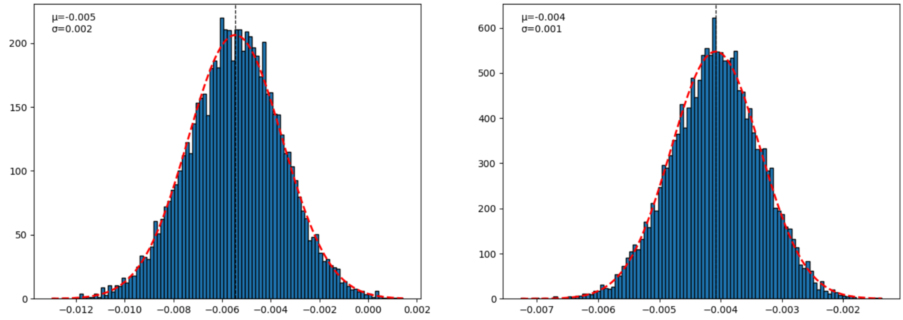
In Figure 5.7 we plot $\hat{\lambda}_1$. The choice of the product measure corresponds to an independent sampling of λ_1^{CL} and λ_1^{CM}. The histogram supports our conjecture formulated in Chapter 2. Remarkably, the difference is 1, like in the homogenous case, irrespective of the degrees.



(a) $n = 1000$,
$\vec{d} = (20, 23, 25, 28, 30)$,
$\sqrt{n} \approx 31.6$, $\log n \approx 6.9$.

(b) $n = 2000$,
$\vec{d} = (28, 33, 35, 38, 42)$,
$\sqrt{n} \approx 44.7$, $\log n \approx 7.6$.

Figure 5.5: Histograms of $\bar{\lambda}_1$ for different graph sizes n and degree sequences $\vec{d}$. The sample size for each regime is 10^4. Each element specified in the degree sequence appears $\frac{n}{5}$ times. In red is plotted the Gaussian fit; μ is the sample mean (represented by a dashed line in the histogram), σ is the sample standard deviation. We expect $\mu \approx 0$.



(a) $n = 5000$,
$\vec{d} = (48, 53, 57, 62, 64)$,
$\sqrt{n} \approx 70.7$, $\log n \approx 8.5$.

(b) $n = 10000$,
$\vec{d} = (78, 80, 83, 87, 90)$,
$\sqrt{n} = 100$, $\log n \approx 9.2$.

Figure 5.6: Histograms of $\bar{\lambda}_1$ for different graph sizes n and degree sequences $\vec{d}$. The sample size for each regime is 10^4. Each element specified in the degree sequence appears $\frac{n}{5}$ times. In red is plotted the Gaussian fit; μ is the sample mean (represented by a dashed line in the histogram), σ is the sample standard deviation. We expect $\mu \approx 0$.

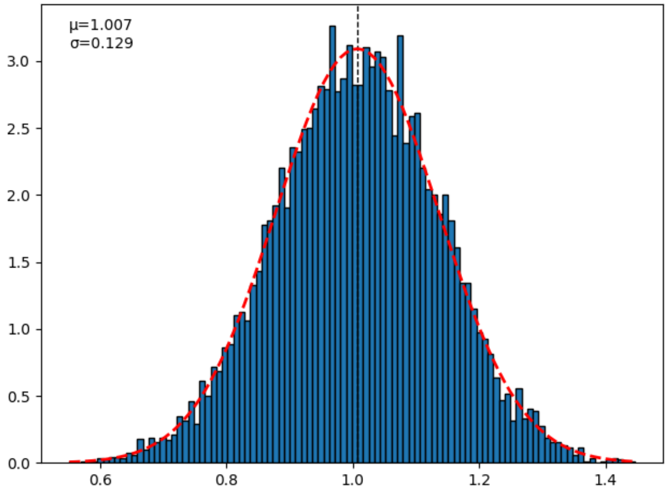


Figure 5.7: Histograms of $\hat{\lambda}_1$ for $n = 10000$ and degree sequences $\vec{d} = (78, 80, 83, 87, 90)$. The sample size is 10^4. Each element specified in the degree sequence appears $\frac{n}{5}$ times. In red is plotted the Gaussian fit; μ is the sample mean (represented by a dashed line in the histogram), σ is the sample standard deviation. We expect $\mu \approx 1$.

Bibliography

[1] M. Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo, July 2018. arXiv:1701.02434 [stat].

[2] M. J. Betancourt, S. Byrne, S. Livingstone, and M. Girolami. The Geometric Foundations of Hamiltonian Monte Carlo, Oct. 2014. arXiv:1410.5110 [stat].

[3] T. Coolen, A. Annibale, and E. Roberts. *Generating Random Networks and Graphs*. Oxford University Press, Mar. 2017.

[4] P. Gao and N. Wormald. Uniform Generation of Random Regular Graphs. *SIAM Journal on Computing*, 46(4):1395–1427, Jan. 2017.

[5] C. Greenhill. Generating graphs randomly. In K. K. Dabrowski, M. Gadouleau, N. Georgiou, M. Johnson, G. B. Mertzios, and D. Paulusma, editors, *Surveys in combinatorics 2021*, London mathematical society lecture note series, pages 133–186. Cambridge University Press, Cambridge, 2021.

[6] S. L. Hakimi. On Realizability of a Set of Integers as Degrees of the Vertices of a Linear Graph. I. *Journal of the Society for Industrial and Applied Mathematics*, 10(3):496–506, Sept. 1962.

[7] V. Havel. Poznámka o existenci konečných grafů. *Časopis pro pěstování matematiky*, 080(4):477–480, 1955. Publisher: Mathematical Institute of the Czechoslovak Academy of Sciences.

[8] S. Janson. The Probability That a Random Multigraph is Simple. *Combinatorics, Probability and Computing*, 18(1-2):205–225, Mar. 2009.

[9] S. Janson. Random graphs with given vertex degrees and switchings. *Random Structures & Algorithms*, 57(1):3–31, Aug. 2020.

[10] M. Jerrum and A. Sinclair. Fast uniform generation of regular graphs. *Theoretical Computer Science*, 73(1):91–100, 1990.

[11] J. H. Kim and V. H. Vu. Generating random regular graphs. *Combinatorica*, 26:683–708, 2006.

[12] B. D. McKay and N. C. Wormald. Uniform generation of random regular graphs of moderate degree. *Journal of Algorithms*, 11(1):52–67, Mar. 1990.

[13] J. Park and M. E. J. Newman. Solution for the properties of a clustered network. *Physical Review E*, 72(2):026136, Aug. 2005. arXiv:cond-mat/0412579.

[14] T. Squartini, R. Mastrandrea, and D. Garlaschelli. Unbiased sampling of network ensembles. *New Journal of Physics*, 17(2):023052, Feb. 2015.

[15] A. Steger and N. C. Wormald. Generating Random Regular Graphs Quickly. *Combinatorics, Probability and Computing*, 8(4):377–396, July 1999.

[16] N. Vallarano, M. Bruno, E. Marchese, G. Trapani, F. Saracco, G. Cimini, M. Zanon, and T. Squartini. Fast and scalable likelihood maximization for Exponential Random Graph Models with local constraints. *Scientific Reports*, 11(1):15227, July 2021.

[17] F. van Ieperen and I. Kryven. Sequential stub matching for uniform generation of directed graphs with a given degree sequence, June 2022. arXiv:2103.15958 [math].