

Statistical Inference for Network Data

Advay Bhattacharya
Department of Statistics
Texas A&M University
[advaybhattacharya@tamu.edu]

May 2, 2026

1 Introduction

A network is defined as $G = (V, E)$, consisting of a set of vertices V and edges E connecting pairs of vertices. Networks arise across many domains like the Internet, social platforms, transportation systems, biological systems, and financial networks, among others [Kolaczyk and Csárdi, 2020].

Statistical analysis of network data is distinct from classical i.i.d. settings because the connectivity pattern itself is what we care about and cannot be reduced to a simple vector of real numbers. Standard tools have to be adapted for this setting. Key concepts used throughout this report are:

- **Adjacency matrix.** A graph on n vertices is represented by its $n \times n$ matrix A , where $A_{ij} = 1$ if nodes i and j share an edge and $A_{ij} = 0$ otherwise.
- **Directed vs. undirected graphs.** In an undirected graph, $A_{ij} = A_{ji}$; in a directed graph, A need not be symmetric.
- **Node degree.** The degree of node i is $d_i = \sum_{j=1}^n A_{ij}$, the number of edges incident to it.
- **No self-loops.** We assume $A_{ii} = 0$ for all i .

The goal of this report is to build up the core ideas of inference for network data, starting from a foundational model and working toward real data analysis. We begin with the Erdős–Rényi (ER) model, which treats all edges as independent and identically distributed. While this assumption is strong, it gives a better setting in which to develop point estimation, confidence intervals, and hypothesis tests. After establishing these tools, we examine the limitations of the model and then apply everything to a real biological network. This strategy mirrors the broader pattern of statistical inference in which one would propose a model, derive estimators, assess uncertainty, validate via simulation, and apply to data.

2 The Erdős–Rényi Model

2.1 Model Definition

Let $G = (V, E)$ be an undirected graph with n vertices and no self-loops, represented by its symmetric adjacency matrix A ($A_{ij} = A_{ji}$, $A_{ii} = 0$). The ER model $G(n, p)$ says that each edge is present independently with probability $p \in (0, 1)$:

$$P(A_{ij} = a_{ij}) = p^{a_{ij}}(1-p)^{1-a_{ij}}, \quad a_{ij} \in \{0, 1\}, \quad i < j. \quad (1)$$

So, write $A \sim G(n, p)$ for a random adjacency matrix from this model. The single parameter p controls how dense the graph is, and since all edges are generated independently, this is the simplest possible probabilistic model for a random graph. Understanding how to estimate p and test hypotheses about it is the basis of this report.

2.2 Properties of Node Degree

To understand the structure of a graph drawn from $G(n, p)$, it makes sense to start with the degree of each node. The degree of node i is $D_i = \sum_{j \neq i} A_{ij}$. Since each of the $n - 1$ potential edges is an independent Bernoulli(p) trial,

$$D_i \sim \text{Binomial}(n - 1, p), \quad \mathbb{E}[D_i] = (n - 1)p, \quad \text{Var}(D_i) = (n - 1)p(1 - p). \quad (2)$$

For $i \neq j$, the only edge shared by D_i and D_j is A_{ij} , so using properties of covariance and independence of all other edges:

$$\text{Cov}(D_i, D_j) = \text{Var}(A_{ij}) = p(1 - p), \quad \text{Corr}(D_i, D_j) = \frac{p(1 - p)}{(n - 1)p(1 - p)} = \frac{1}{n - 1}. \quad (3)$$

The correlation is always positive but vanishes as $n \rightarrow \infty$, reflecting the fact that two nodes share only one edge no matter how large the network. This positive dependence is a feature of the graph that would not appear in a standard i.i.d. sample.

2.3 Total Edge Count and Its Distribution

The degree calculation above also means we should look at the total edge count, which is a good sufficient statistic for p . Let Y be the total number of edges. Each edge is counted twice in $\sum_i D_i$, so $Y = \frac{1}{2} \sum_{i=1}^n D_i = \sum_{i < j} A_{ij}$. Since Y is a sum of $m = \binom{n}{2}$ i.i.d. Bernoulli(p) variables:

$$\mathbb{E}[Y] = \binom{n}{2} p = \frac{n(n - 1)}{2} p, \quad \text{Var}(Y) = \binom{n}{2} p(1 - p) = \frac{n(n - 1)}{2} p(1 - p). \quad (4)$$

By the Central Limit Theorem, as $n \rightarrow \infty$:

$$\frac{Y - \frac{n(n - 1)}{2} p}{\sqrt{\frac{n(n - 1)}{2} p(1 - p)}} \xrightarrow{d} \mathcal{N}(0, 1). \quad (5)$$

This asymptotic normality of Y is what will allow us to derive confidence intervals and test statistics in the sections that follow.

3 Limitations of the Erdős–Rényi Model

Before moving to inference, it is worth asking how realistic the ER model actually is. This matters because the method of inference we develop are only as meaningful as the model they are built on.

The ER model uses two assumptions that rarely hold in practice: edge independence and a single global probability p . In a friendship network, for instance, two people with a mutual friend are far more likely to be connected than two strangers, which violates both assumptions. The ER clustering coefficient equals p , which is not meaningful for large sparse networks, while real social networks show much higher clustering.

Real networks also show that a small number of highly connected hub nodes coexist with many low-degree nodes [Kolaczyk and Csárdi, 2020]. The Binomial($n - 1, p$) degree distribution of the ER model concentrates around its mean, which is inconsistent with distributions observed in, for example, the Internet and citation networks, where $P(\text{degree} = k) \propto k^{-\gamma}$. These observations mean we may depend on better models such as the stochastic block model and exponential random graph models, which we return to in the conclusion. For now, we proceed with ER as a starting point that makes the inferential ideas clear.

4 Estimation of p

4.1 MoM and MLE

Having established the properties of Y , we can now use it to estimate the unknown parameter p . Since $\mathbb{E}[Y] = \binom{n}{2} p$, the method of moments estimator is obtained by setting Y equal to its expectation and solving

for p . This gives $\hat{p} = Y/\binom{n}{2} = 2Y/[n(n-1)]$, the observed edge density. The likelihood given adjacency matrix A is $L(p; A) = p^Y(1-p)^{\binom{n}{2}-Y}$, and maximizing the log-likelihood yields the same estimator:

$$\hat{p}_{\text{MLE}} = \frac{Y}{\binom{n}{2}} = \frac{2Y}{n(n-1)}. \quad (6)$$

The fact that the MoM and MLE are the same is not a coincidence. It reflects the fact that Y is the sufficient statistic for p , so both methods would arrive at the same function of data.

4.2 Properties of $\hat{p}$

Let $m = \binom{n}{2}$. Then $\mathbb{E}[\hat{p}] = p$ (unbiased), $\text{Var}(\hat{p}) = p(1-p)/m = 2p(1-p)/[n(n-1)]$, and $\text{MSE}(\hat{p}) = \text{Var}(\hat{p})$. As $n \rightarrow \infty$, $m \rightarrow \infty$ so $\text{Var}(\hat{p}) \rightarrow 0$; as such, $\hat{p} \xrightarrow{p} p$ (so we have consistency). By the CLT:

$$\frac{\hat{p} - p}{\sqrt{2p(1-p)/[n(n-1)]}} \xrightarrow{d} \mathcal{N}(0, 1). \quad (7)$$

This asymptotic normality is the key result connecting the model properties from Section 2 to the inferential tools in the next section.

5 Confidence Intervals and Hypothesis Testing

With a good estimator in hand, we can now build the full inferential method like confidence intervals and hypothesis tests, first for a single graph and then for comparing two graphs.

5.1 One-Sample Inference for p

From the asymptotic normality of $\hat{p}$, a $(1 - \alpha)$ confidence interval for p is:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{2\hat{p}(1-\hat{p})}{n(n-1)}}. \quad (8)$$

To test $H_0 : p = p_0$ vs. $H_1 : p \neq p_0$, the Wald z -statistic is:

$$Z = \frac{\hat{p} - p_0}{\sqrt{2p_0(1-p_0)/[n(n-1)]}}, \quad (9)$$

which is asymptotically $\mathcal{N}(0, 1)$ under H_0 . We reject when $|Z| > z_{\alpha/2}$; the p -value is $2(1 - \Phi(|Z|))$. Note that p_0 is used in the standard error since we are evaluating the statistic under the null.

5.2 Monte Carlo Power Study

To understand how the test behaves in practice, we estimate its power via Monte Carlo simulation. We use $B = 1000$ replicates, $p_0 = 0.1$, and $\alpha = 0.05$. Figure 1 shows the results. Power increases rapidly with n and reaches near one by $n \approx 30$ when $p = 0.2$. The power curve as a function of p achieves its minimum of $\alpha = 0.05$ at $p = p_0$ and increases as $|p - p_0|$ grows, consistent with theory. These results confirm that even at moderate network sizes the test is reliable, which is fine given that some real networks are not enormous.

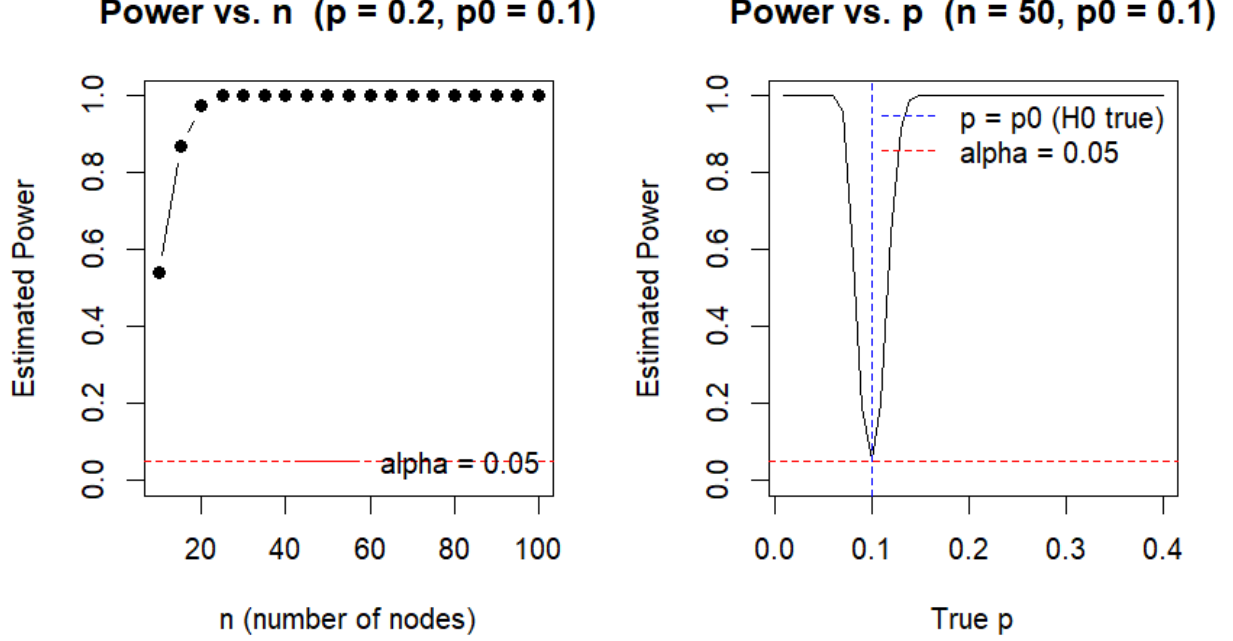


Figure 1: Left: estimated power vs. n with $p = 0.2$, $p_0 = 0.1$. Right: estimated power vs. true p with $n = 50$, $p_0 = 0.1$. Each point is based on $B = 1000$ Monte Carlo replications.

5.3 Two-Sample Inference for $p - q$

Going beyond the one-sample problem, we may ask whether two graphs have the same edge probability. This is relevant whenever we want to compare networks across groups or conditions, which is the question we will face in the *C. elegans* application. Let $A \sim G(n_1, p)$ and $B \sim G(n_2, q)$ be independent, with $m_k = \binom{n_k}{2}$ and edge density estimators $\hat{p} = Y_A/m_1$, $\hat{q} = Y_B/m_2$. By the CLT and independence:

$$\hat{p} - \hat{q} \sim \mathcal{N}\left(p - q, \frac{p(1-p)}{m_1} + \frac{q(1-q)}{m_2}\right). \quad (10)$$

Plugging in estimates gives $\hat{se} = \sqrt{\hat{p}(1-\hat{p})/m_1 + \hat{q}(1-\hat{q})/m_2}$, and the test statistic $Z = (\hat{p} - \hat{q})/\hat{se}$ is asymptotically $\mathcal{N}(0, 1)$ under $H_0 : p = q$. We reject H_0 at level α when $|Z| > z_{\alpha/2}$; the p -value is $2(1 - \Phi(|Z|))$. A $(1 - \alpha)$ CI for $p - q$ is:

$$\hat{p} - \hat{q} \pm z_{\alpha/2} \hat{se}. \quad (11)$$

5.4 Monte Carlo Coverage Study

Before applying this test to real data, let's validate the CI empirically. We check coverage across 36 scenarios varying $n_1, n_2 \in \{20, 50, 100\}$ and $p, q \in \{0.1, 0.3\}$ with $B = 1000$ replicates each. Coverage ranged from 0.925 to 0.961, clustering around the 0.95 level across all scenarios (see Appendix, Table 1). This shows that the asymptotic CI performs well even at moderate network sizes such as $n = 20$, which gives us confidence in the test's validity for the application in the next section.

6 Application: *C. elegans* Brain Network

6.1 Dataset and Background

Having checked the inferential tools through simulation, we now apply them to a real biological network. The nematode *Caenorhabditis elegans* (*C. elegans*) is a microscopic roundworm whose nervous system has been completely mapped at the level of individual neurons and synapses, making it a widely used standard in both networks and neuroscience [Kolaczyk and Csárdi, 2020]. Its 279 neurons are connected by two types of synapses: electrical synapses called gap junctions, which allow bidirectional ion flow and symmetric signaling, and chemical synapses, which transmit signals via neurotransmitters and are directional. These two synapse types define two separate networks on the same node set. The scientific question is whether they have the same edge probability under the ER model, which is the two-sample testing problem from Section 5. The data is available through the `iGraphMatch` R package and since the ER model is undirected, both adjacency matrices are symmetrized prior to analysis.

6.2 Results

Both networks have $n_1 = n_2 = 279$ nodes and $m = \binom{279}{2} = 38,781$ possible edges. The electrical network has $Y_A = 1,961$ observed edges ($\hat{p} = 0.0506$) and the chemical network has $Y_B = 517$ ($\hat{q} = 0.0133$), so the electrical network is about four times denser. Figure 2 shows the two networks visually.

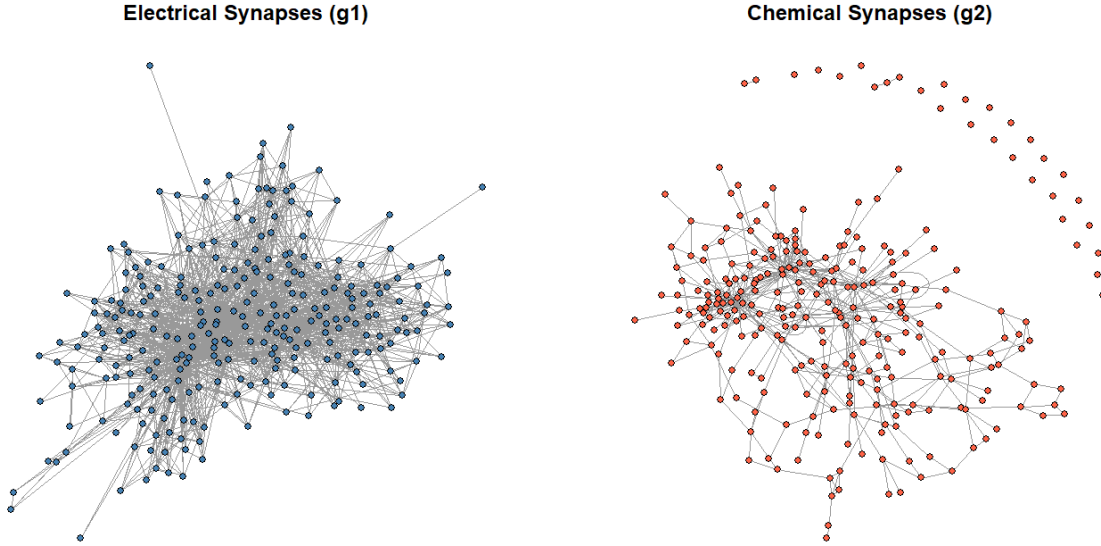


Figure 2: Left: electrical synapse network (g_1), showing a dense core structure with 1,961 edges. Right: chemical synapse network (g_2), sparser (517 edges) with more isolated nodes. Both networks have $n = 279$ nodes; layout via Fruchterman–Reingold.

Applying the two-sample test from Section 5, we get $Z = 29.65$ and $p\text{-value} = 3.46 \times 10^{-193}$. The 95% CI for $p - q$ is $(0.0348, 0.0397)$, above zero. We therefore reject $H_0 : p = q$ and conclude the two synapse types have significantly different edge probabilities.

6.3 Discussion

The result is consistent with known biology since gap junction connectivity in *C. elegans* is more spread out than chemical synaptic connectivity, which is less frequent but more targeted. However, the ER model has

clear limitations here. Neurons in the same circuit are more likely to be connected, which works against the assumption that edges are independent. Chemical synapses are also directional by nature, and symmetrizing the adjacency matrix loses this information. These issues show us that we may need better models that let edge probabilities vary across node pairs and that can handle directed graphs, both of which are active areas of research in statistical network analysis.

7 Conclusion

This report built tools for inference as it relates to the ER model. We started by deriving the properties of degree and edge count, which gave us a good foundation for the distribution. We then derived the MLE/MoM estimator $\hat{p}$ and established its asymptotic normality, which is what makes confidence intervals and hypothesis tests meaningful. We constructed both one- and two-sample procedures and validated them through Monte Carlo simulation. Finally, we applied the two-sample test to the *C. elegans* brain network, finding strong evidence that electrical and chemical synapse networks differ in edge density where we saw $Z = 29.65$ and $p \approx 0$.

Throughout this process, the ER model served as a useful tool for us to bridge to better models. Its simplicity made the inferential ideas better understood, even though its assumptions turned out to be restrictive for real data. Natural extensions include the stochastic block model, which allows edge probabilities to vary by community, and exponential random graph models, which can have more complex structural dependencies. Both preserve the main inference ideas developed here while expanding the class of networks that can be modeled, and they are a next step beyond what was covered in this report.

Appendix: Coverage Table

Table 1: Coverage of the 95% CI for $p - q$ across 36 scenarios, based on $B = 1000$ Monte Carlo replications.

n_1	n_2	p	q	Coverage
20	20	0.1	0.1	0.950
50	20	0.1	0.1	0.951
100	20	0.1	0.1	0.937
20	50	0.1	0.1	0.934
50	50	0.1	0.1	0.944
100	50	0.1	0.1	0.949
20	100	0.1	0.1	0.941
50	100	0.1	0.1	0.951
100	100	0.1	0.1	0.938
20	20	0.3	0.1	0.958
50	20	0.3	0.1	0.934
100	20	0.3	0.1	0.937
20	50	0.3	0.1	0.940
50	50	0.3	0.1	0.946
100	50	0.3	0.1	0.948
20	100	0.3	0.1	0.937
50	100	0.3	0.1	0.948
100	100	0.3	0.1	0.925
20	20	0.1	0.3	0.950
50	20	0.1	0.3	0.956
100	20	0.1	0.3	0.944
20	50	0.1	0.3	0.947
50	50	0.1	0.3	0.943
100	50	0.1	0.3	0.961
20	100	0.1	0.3	0.934
50	100	0.1	0.3	0.930
100	100	0.1	0.3	0.960
20	20	0.3	0.3	0.947
50	20	0.3	0.3	0.935
100	20	0.3	0.3	0.948
20	50	0.3	0.3	0.935
50	50	0.3	0.3	0.954
100	50	0.3	0.3	0.951
20	100	0.3	0.3	0.958
50	100	0.3	0.3	0.953
100	100	0.3	0.3	0.950

R Code and Simulations

All simulations were conducted in R using the `igraph` and `iGraphMatch` packages [Csardi and Nepusz, 2006]. The full code is available upon request.

References

- Eric D. Kolaczyk and Gábor Csárdi. *Statistical Analysis of Network Data with R*. Springer, New York, 2nd edition, 2020. doi: 10.1007/978-3-030-44129-6.
- Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006.