A Bayesian Framework for Comparing Observed Graphs to Common Generative Models

Marwin Carmo

University of California

Davis

PSC 290

Dr. Jonathan Park

Spring 2025

A Bayesian Framework for Comparing Observed Graphs to Common Generative Models

Real-world networks exhibit complex structures. As such, graph generative models have been proposed to explain their topological properties, each representing a distinct hypothesis about the rules of network formation (Wallis, 2007). For example, some models posit that connections form randomly, while others suggest that new connections are preferentially made to already popular nodes, or that nodes form distinct communities. These structural characteristics affect the distribution of influence and importance among nodes. Therefore, a critical question is: given a real-world network, which generative model provides the most probable explanation for its observed structure?

This project suggests a quantitative method to perform this comparison. The Bayesian paradigm offers an approach for model selection by calculating the posterior probability of each candidate model given the observed data (Rouder et al., 2018). We can weigh the evidence and determine how our belief in each model should be updated after seeing the data. However, they also carry a major challenge: for most graph generative models, the likelihood function (i.e., the probability of observing a specific graph given the model's parameters) is mathematically intractable. This complexity makes standard Bayesian computation impossible.

Approximate Bayesian Computation (ABC) is a class of computational methods designed to circumvent this problem. Instead of evaluating an intractable likelihood function, ABC relies on simulating datasets from a model and comparing them to the observed data. If a simulation closely resembles the real data, the model parameters that produced it are considered plausible. While ABC has been applied to network science for estimating the parameters of a single model and for inference in large growing networks , a user-friendly framework for comparing an observed graph against a diverse suite of standard generative models is still absent.

This project proposes a blueprint for a unified and accessible Bayesian framework

for graph model selection. This framework will allow researchers to obtain a probabilistic measure of whether simple, known mechanisms (e.g., the random edge formation in Erdos-Renyi models) are adequate to explain observed network structure, or if more intricate mechanisms (like preferential attachment in Barabási-Albert models) are required.

Objectives

The first aim is to develop and implement the core Bayesian model selection framework. This involves building an ABC engine capable of estimating posterior model probabilities for graph data. A key part is to integrate simulators for a range of canonical graph generative models, including the Erdos-Renyi (ER) , Watts-Strogatz (WS) , Barabási-Albert (BA) , and Stochastic Block (SBM) models. Then, to formulate and test appropriate prior distributions for the parameters of each model.

The second aim is to establish a methodology for selecting informative summary statistics. The ABC method relies on comparing summary statistics (e.g., average path length, clustering coefficient) rather than the entire, complex graph structures. The choice of these statistics is critical. The proposed framework will evaluate the effectiveness of theory-driven graph metrics and also implement and assess semi-automated, regression-based methods for identifying a reduced set of highly informative summary statistics from a large pool of candidates. The use of boosting techniques will also be explored.

The third aim is to validate the framework's performance and demonstrate its utility. To ensure the framework is accurate, I will develop a validation study using simulated graphs, generated from known models, quantifying the framework's ability to correctly identify the true generating model. Subsequently, this validated framework will be applied to real-world network datasets to demonstrate its practical utility.

Method

Model selection

Given an observed graph G and a set of candidate models $\{M_1, \ldots, M_k\}$, the goal is to compute the posterior probability of each model, $p(M_k|G)$. This is achieved via Bayes' theorem, which states that the posterior probability of a model is proportional to its prior probability multiplied by the marginal likelihood. The marginal likelihood, $p(G|M_k)$, is the probability of observing the data given the model, averaged over all possible parameter values. It is the intractability of this marginal likelihood for graph models that motivates the proposed computational approach.

Approximate Bayesian Computation (ABC)

The ABC algorithm is a simulation-based procedure. It begins by selecting a model and a set of its parameters from their prior distributions. A synthetic graph is then simulated using this model and parameters. Summary statistics are calculated for both the observed and simulated graphs, and if the distance between these sets of statistics is below a small tolerance threshold, the chosen model is 'accepted'. By repeating this process many times, the proportion of acceptances for each model provides an approximation of its posterior probability. This process can be summarised in the following steps:

1. Sample a model M_k from the model prior $p(M)$.
2. Sample a parameter set θ_k from the parameter prior $p(\theta_k|M_k)$.
3. Simulate a graph $\widetilde{G}$ from M_k using θ_k.
4. Calculate a vector of summary statistics $S(\widetilde{G})$ and $S(G)$.
5. If the distance $d(S(G), S(\widetilde{G}))$ is below a tolerance ϵ, accept the model choice M_k.
6. Repeat steps 1-5 many times. The approximate posterior probability $p(M_k|\widetilde{G})$ is the fraction of accepted samples that correspond to model M_k

Comparison with Clauset et al. (2009)

The framework for power-law distributions by Clauset consists of two main parts: a goodness-of-fit test and model comparison via likelihood ratios.

- Goodness-of-fit: The goodness-of-fit test based on the Kolmogorov-Smirnov (KS) statistic to generate a p-value. This p-value represents the probability of seeing data as extreme as, or more extreme than, the observed data, assuming the fitted power-law model is the true generating process. A small p-value is taken as evidence to reject the power-law hypothesis.

- Model comparison: For data that passes the goodness-of-fit test, likelihood ratio tests are used to compare the power law against alternative distributions (e.g., log-normal, exponential). This test determines which model has a higher likelihood at its best-fit point and provides a p-value for the significance of that preference.

  The proposed Bayesian framework differs in its fundamental inferential philosophy and the questions it answers:

- Probability statements: The most significant difference lies in the output. The Clauset et al. method provides a statement about the probability of the data given the model. It allows for the rejection of a hypothesis but does not quantify the probability of the hypothesis itself. The Bayesian framework, in contrast, aims to compute the posterior model probability, $p(\text{mode}|\text{data})$. This is a statement about the probability of the model being correct, given the data.

- Model comparison: The maximum likelihood estimate (MLE) compares models at their single best-performing parameter settings. The Bayesian approach, through the marginal likelihood, compares models by averaging their performance over their entire parameter space, weighted by the prior.

- Handling uncertainty: The MLE provides a point estimate for model parameters and

standard errors. Bayesian inference provides a full posterior distribution for each parameter, offering a more complete view of uncertainty.

Analysis plan

Validation with simulated data

The first phase will focus on a simulation study to validate the proposed method's accuracy and characterize its performance. This involves generating a large number of simulated graphs where the generative model and its parameters are known.

- Graphs will be generated from three primary models: Erdos-Renyi (ER) , Watts-Strogatz (WS), and Barabási-Albert (BA). For each model, I will vary its key parameters across a range of values (e.g., connection probability for ER; rewiring probability for WS). A range of network sizes (e.g., n = 30, 50, 100, 500, 1000 nodes) will also be tested.

- For each simulated graph, the posterior probabilities of all candidate models and their parameters will be computed and performance will be assessed using metrics such as:

    1. Model recovery: The primary metric will be the percentage of simulations where the true generating model is correctly assigned the highest posterior probability.
    2. Posterior concentration: The average posterior probability assigned to the true model across all simulations. A high average value indicates strong evidence for the correct model.
    3. Parameter recovery: For simulations where the model is correctly identified, the accuracy of the parameter estimates will be assessed by comparing the posterior mean or median to the known true parameter values.

- Sensitivity Analysis: I will also analyze how these performance metrics are affected by (a) the choice of summary statistics (comparing a manually-selected set against a

semi-automatically selected set 1), (b) the ABC tolerance parameter ϵ, and (c) the degree of ambiguity between models (e.g., testing a WS model with high rewiring against an ER model to test discriminability).

Validation with simulated data

Upon successful validation, the proposed method will be applied to open network datasets, possibly from different domains but with a greater emphasis to psychometric networks. The analysis will focus on interpreting the resulting posterior model probabilities. For instance, finding that a Barabási-Albert model has a much higher posterior probability than other models for a given network would provide quantitative evidence that growth and preferential attachment are key organizing principles for that system.

Challenges

Despite the promising alternative this Bayesian framework proposes, there are many aspects inherent to it that could make its implementation challenging:

1. ABC is inherently simulation-heavy and can be computationally expensive, especially for large networks.

2. The choice of summary statistics is crucial for the accuracy of ABC but finding a method that works well across diverse graph types is not a simple task. Automated methods exist but have their own complexities.

3. Bayesian model comparison can be sensitive to prior distributions, both for the models and parameters within each model. Each generative model has their own set of parameters that requires different prior specifications. Therefore, it is not trivial to specify priors that make these comparisons “fair”.

4. Only a limited set of models will be tested. If the true generative process is not represented, the best fit among the options might still be a poor absolute fit.

References

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, *25*(1), 102–113. https://doi.org/10.3758/s13423-017-1420-7

Wallis, W. D. (2007). *A beginner's guide to graph theory* (2nd ed). Birkhäuser.