Review of a mixture model for random graphs

Adib Habbou and Teddy Alexandre

December 21, 2023

1 Introduction

The field of mixture models for graphs and clustering aims to enhance graph modeling by incorporating intricate structures, such as clustering, aiming to overcome the limitations of traditional models and to provide a more realistic representation of complex real-world graphs.

The Erdos-Rényi model has long been a cornerstone, providing theoretical insights into properties like subgraphs, degree distribution, connectedness, and clustering coefficients. However, its limitations in accurately representing real-world graphs, particularly in areas such as social, biological, or internet graphs, have led to the exploration of more nuanced models.

Mixture models, especially those incorporating clustering and assortative mixing, have emerged as a promising avenue for addressing these shortcomings. The article dives into the extension of the Erdos-Rényi model through a mixture framework, where vertices are categorized into clusters, and hidden variables guide the degrees distribution, offering a more flexible and realistic depiction of complex graph structures. The focus lies on inferring clustering and mixing parameters solely from the graph topology, highlighting the importance of these models in capturing intricate connectivity patterns that traditional models may overlook.

2 Erdos-Rényi Model

In the Erdos-Rényi model, a graph is generated by connecting pairs of vertices with edges independently and randomly. The probability of any pair of vertices being connected is denoted by p, and these connections are assumed to be mutually independent.

The Erdos-Rényi model generates random graphs by considering a set of n vertices, where each pair of vertices i and j is connected with probability p. This can be expressed using the indicator variable X_{ij}, which equals 1 if vertices i and j are connected and 0 otherwise (no self-connection):

$$X_{ij} = X_{ji} = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases} \tag{1}$$

In Erdos-Rényi model, the probability distribution for the degree follows a binomial distribution:

$$P(K_i = \sum_{j \neq i} X_{ij} = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k} \tag{2}$$

As n becomes large, this binomial distribution approximates a Poisson distribution with mean:

$$P(K_i = k) \approx \frac{e^{-\lambda}\lambda^k}{k!} \quad \text{with} \quad \lambda = (n-1)p \tag{3}$$

This model assumes independence of edges, providing a theoretical foundation for understanding various graph properties. However, it may fall short in representing real-world graphs accurately, motivating the exploration of more sophisticated models.

3 Mixture model for degrees

The mixture model for degrees extends the Erdos-Rényi model by introducing clusters and hidden variables to account for heterogeneities among vertices. In this model, vertices are grouped into clusters, and their degrees follow a mixture of Poisson distributions.

$$P(K_i = \sum_{j \neq i} X_{ij} = k) = \sum_{q=1}^{Q} \alpha_q P(K_i = k | Z_{iq} = 1) \tag{4}$$

Where α_q is the prior probability for vertex i to belong to cluster q, and $P(K_i = k|Z_{iq} = 1)$ is the Poisson distribution for the degree of vertex i given its membership in cluster q.

This approach goes beyond the Erdos-Rényi model's independence assumption, offering a more realistic representation of complex graph structures by considering assortative mixing patterns among clusters. This nuanced degree distribution naturally leads to a mixture model for edges.

4 Erdos-Rényi Mixture for Graphs

The Erdos-Rényi Mixture for Graphs (ERMG) model comes with a new idea where vertices are distributed into Q clusters with prior probabilities $\{\alpha_q\}$. The model introduces symmetric probabilities denoted π_{ql} for vertices from cluster q to connect with vertices from cluster l. The ERMG model directly addresses edge connections, providing a comprehensive representation of both vertex clustering and graph topology using the connectivity matrix $\boldsymbol{\pi} = (\pi_{ql})$.

The probability for an edge between i and j, given their cluster memberships, is expressed as:

$$X_{ij}|\{i \in q, j \in l\} \sim \mathcal{B}(\pi_{ql}). \tag{5}$$

Unlike traditional models, ERMG introduces the concept of vertices belonging to distinct clusters with specified prior probabilities. The model then accounts for the connectivity between these clusters through symmetric probabilities, allowing for a nuanced representation of edge connections. This comprehensive framework not only addresses the limitations of traditional Erdos-Rényi models but also enables a more accurate depiction of the intricate structures.

5 EM algorithm

The Erdos-Rényi mixture model for graphs involves clustering graphs into different groups based on their structural properties. The Expectation-Maximization (EM) algorithm is a method used to estimate the parameters of this mixture model. We denote $\mathcal{X} = \{X_{ij}\}_{i,j=1,\dots,n}$ the set of edges and $\mathcal{Z} = \{Z_{iq}\}_{i\in[|1,n|],q\in[|1,Q|]}$ the set of indicator variables for the vertices.

1. Log-Likelihood:

The log-likelihood function measures the likelihood of observing the given data given the parameters of the model. For the Erdos-Rényi mixture model for graphs, the log-likelihood $\mathcal{L}(\mathcal{X})$ cannot be calculated directly, since the model is incomplete. We use instead a lower bound the log-likelihood :

$$\mathcal{J}(R_{\mathcal{X}}) = log\mathcal{L}(\mathcal{X}) - KL[R_{\mathcal{X}}(.), \mathbb{P}(.|\mathcal{X})] \tag{6}$$

Where $R_{\mathcal{X}}$ is a class of distribution that depends on $\mathcal{X}$, and KL is the Kullback-Leibler divergence. In our case, we restrict $R_{\mathcal{X}}$ to $R_{\mathcal{X}}(\mathcal{Z}) = \prod_i h(\mathcal{Z}_i; \tau_i)$ where h is the distribution of a multinomial distribution of parameter τ.

2. E-step (Expectation step):

In the E-step of the EM algorithm, we compute the expected value of the latent variables, which in this case are the cluster assignments for each graph. We calculate the posterior probabilities or responsibilities of each graph belonging to each cluster given the current parameters.

For each i (graph index) and q (cluster index), we compute the posterior probability :

$$\hat{\tau}_{iq} \propto \alpha_q \prod_{j\neq i} \prod_l b(X_{ij}; \pi_{ql})^{\hat{\tau}_{jl}} \tag{7}$$

Here, $\hat{\tau}_{iq}$ represents the probability that the node i belongs to cluster q given the current parameters. These parameters are updated by computing the Lagrangian of $\mathcal{J}(R_{\mathcal{X}})$.

3. M-step (Maximization step):

In the M-step, the parameters of the model are updated following the E-step update of the $\hat{\tau}_{iq}$'s.

- Update prior probabilities $\hat{\alpha}_q$:

$$\hat{\alpha}_q = \frac{\sum_{i=1}^n \hat{\tau}_{iq}}{n} \tag{8}$$

- Update probabilities of connections $\hat{\pi}_{ql}$ for each pair of clusters:

$$\hat{\pi}_{ql} = \frac{\sum_{i\neq j} \hat{\tau}_{iq}\hat{\tau}_{jl}X_{ij}}{\sum_{i\neq j} \hat{\tau}_{iq}\hat{\tau}_{jl}} \tag{9}$$

The E-step and M-step are iterated until convergence, where the log-likelihood no longer significantly increases or the parameters stabilize. The algorithm aims to find the optimal parameters that best describe the clustering of the observed graphs into different clusters based on the Erdos-Rényi mixture model.

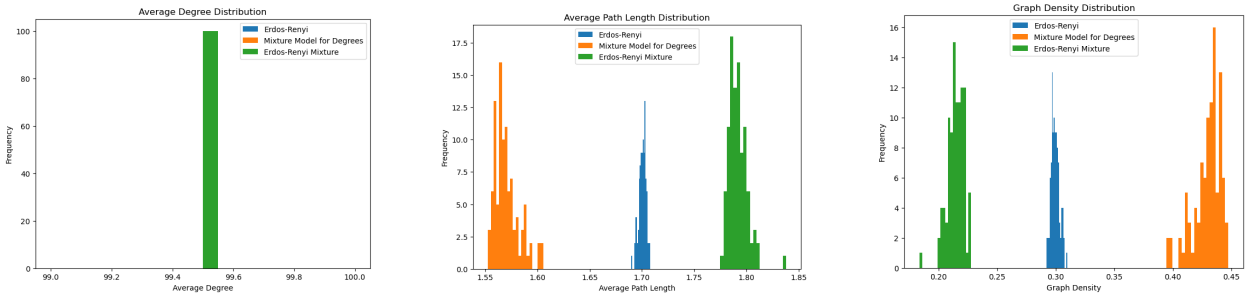
6 Experiments

6.1 Graph Generation: properties comparison between ER, MMD and ERMG

We implemented the three main methods discussed in the article to generate random graphs: Erdos-Rényi (ER), Mixture Model for Degrees (MMD) and Erdos-Rényi Mixture for Graphs (ERMG). For each method, we generated 300 graphs of 200 nodes each time.

Finally we compared the 300 generated graphs given the following metrics:

- **Average Degree:** average number of edges connected to each vertex in a graph
- **Average Path Length:** average number of edges along the shortest paths between all pairs
- **Global Density:** ratio of the number of actual edges to the maximum possible edges
- **Global Efficiency:** average inverse shortest path length
- **Clustering Coefficient:** measure of how connected a vertex's neighbors are to each other
- **Betweenness Centrality:** extent to which a vertex lies on the shortest paths between others



(a) Average Degree (b) Average Path Length (c) Graph Density

(d) Global Efficiency (e) Clustering Coefficient (f) Betweeness Centrality

Figure 1: Comparison between Erdos-Rényi (ER), Mixture Model for Degrees (MMD) and ERMG

You can look at some of the generated graphs in the appendix on Figures 3, 4 and 5. Meanwhile, Figure 1 displays the comparisons between 300 graphs following the above metrics.

The average degree of 99.5 in ER, MMD, and ERMG graphs indicates a balanced connectivity with similar expected average degrees. MMD has the lowest average path length due to its clustered

structure, while ER has intermediate values with more homogeneous connections, and ERMG records the highest values influenced by inter-cluster connections, leading to longer average paths.

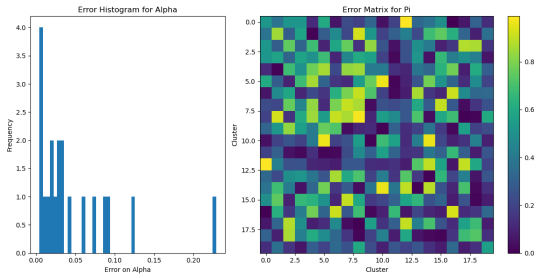
ERMG has the lowest graph density and efficiency, followed by ER and MMD, indicating varying connectivity patterns. MMD's clustering coefficient covers the range of ER and starts just after ERMG's highest values, while betweenness centrality is lower for ER and MMD, with ERMG showing higher values for some graphs due to vertices connecting different clusters.

In summary, ERMG features structured inter-cluster connections, leading to longer paths, lower density and efficiency, and a reduced clustering coefficient compared to ER and MMD. This signifies a more organized connectivity pattern in ERMG, where vertices within clusters are more strongly connected. Contrasting with the randomness in ER and diverse clustering in MMD.

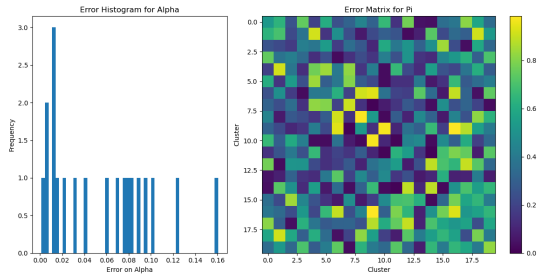
6.2 EM Algorithm

The implementation of the EM algorithm has been quite challenging. Indeed, the E-step was not clear, since we have no closed form expression for the variational parameters $\{\hat{\tau}_{iq}\}_{i,q}$. These parameters are defined up to a multiplicative factor that depends of the Lagrange multiplier $\exp(1+\lambda_i)$. Since, this Lagrange multiplier is not explicitly defined (we don't know what value is given to λ_i), we just assumed there was no multiplicative factor.

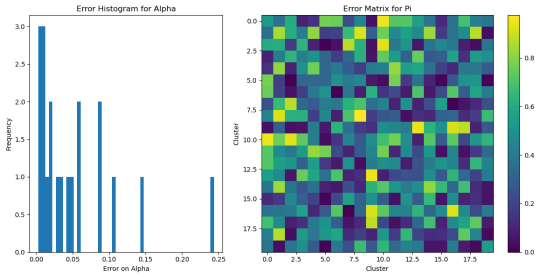
We made several experiments to test the EM algorithm on the same ERMG graph, the following figures shows the errors made when estimating the parameters α and π:



(a) First Run



(b) Second Run



(c) Third Run

Figure 2: Errors made on α (histogram) and π (matrix) compared to the true values

The errors made on the parameters are quite significant, meaning the estimated parameters are far from the true values. This is most likely due to the E-step that could not be done properly, which might have changed the estimated parameters obtained. Since we don't have the Lagrangian multiplier, the exact expression of the variational parameters could not be derived exactly.

7 Limitations

A limitation of ERMG in graph generation is its reliance on specific assumptions, including the structure of inter-cluster connections, symmetric probabilities, and homogeneity within clusters. If these assumptions diverge from a real-world network characteristics, the generated graphs may not accurately capture the complexities of such diverse graph structures.

Real-world networks may have community structures, hierarchical organization, or preferential attachment, aspects not fully addressed by the ERMG framework. Therefore, the effectiveness of graph generation models, including ERMG, depends on the context. Consideration of model limitations is crucial, aligning with the intended application and target network characteristics.

The EM Algorithm has a major limitation: the parameters estimated by the algorithm can be very different from a run to another, since we initialize randomly our algorithm. Indeed, we have no guarantee about the convergence of our algorithm. This was one of the main difficulties encountered in our experiments: the algorithm is not stable enough.

Our goal at first was to implement the EM for ERMG and then test it on some real-word graphs available online to see if they can be accurately represented by an ERMG.

8 Conclusion

In our exploration of graph models, we delved into the Erdos-Rényi Mixture for Graphs (ERMG) as a promising advancement, offering structured connectivity and distinct properties compared to classic models like Erdos-Rényi (ER) and Mixture Model for Degrees (MMD).

ERMG's organized connectivity, visible in longer paths and reduced density, presents a significant step toward realistic network representations. However, its assumptions might limit capturing intricate real-world structures like community formations.

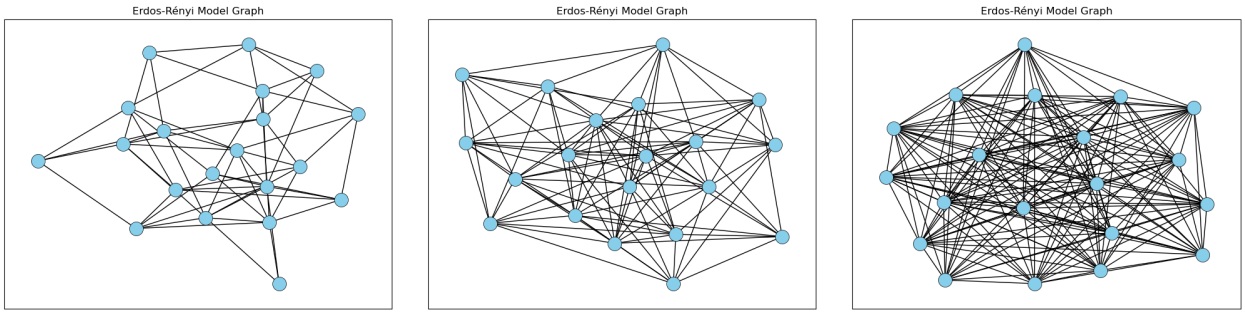
Our implementation challenges with the Expectation-Maximization (EM) algorithm for ERMG underscored stability and convergence issues, signaling the need for further refinement.

Despite limitations, this exploration lays the groundwork for more accurate graph models, emphasizing the importance of context-driven choices and highlighting the ongoing quest for better network representations.

9 Contributions

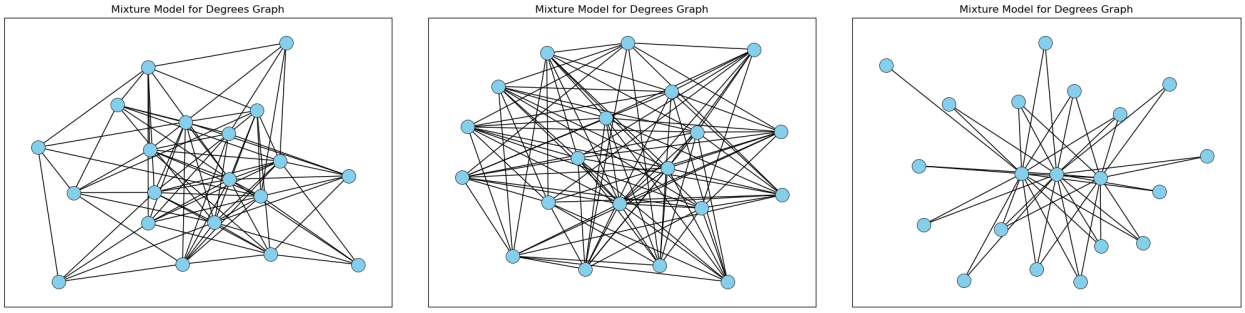
Adib contributed by writing the introduction and providing explanations for the ER, MMD, and ERMG. Additionally, he implemented from scratch the function that generates graphs using the ER, MMD, and ERMG, followed by a comparison of the proprieties of the generated graphs and critique of the graph generation methods. On the other hand, Teddy explained the EM algorithm, re-implemented it in the specific case of ERMG. Finally, we both tried to discover the issues in the EM for ERMG without success, so we applied the EM for ERMG several times on the same graph to show its instability and to highlight its poor performance.

Appendix



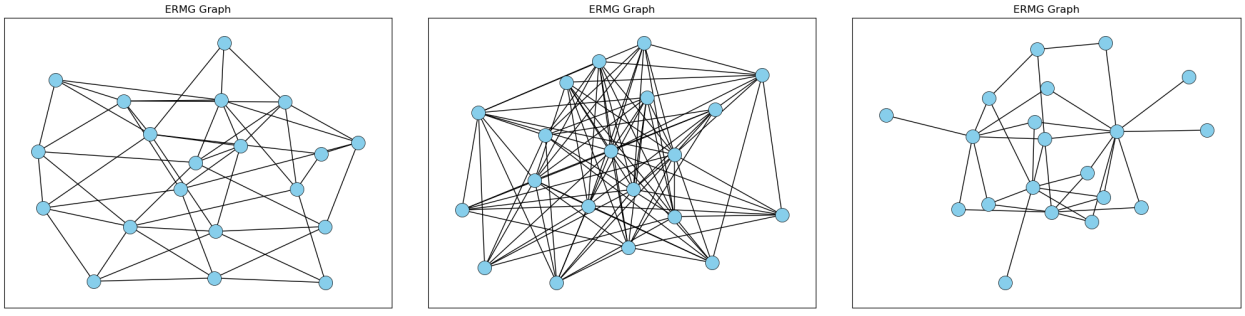
(a) Erdos-Rényi Graph $p = 0.3$ (b) Erdos-Rényi Graph $p = 0.6$ (c) Erdos-Rényi Graph $p = 0.9$

Figure 3: Erdos-Rényi Graphs



(a) $\alpha = [0.5, 0.4, 0.1], \lambda = [2, 3, 1]$ (b) $\alpha = [0.5, 0.4, 0.1], \lambda = [4, 6, 2]$ (c) $\alpha = [0.8, 0.1, 0.1], \lambda = [2, 3, 1]$

Figure 4: Mixture Model for Degrees Graphs with $Q = 3$



(a) $\alpha = [0.5, 0.4, 0.1]$, $\pi = \begin{bmatrix} 0.8 & 0.3 & 0.5 \\ 0.3 & 0.7 & 0.2 \\ 0.5 & 0.2 & 0.9 \end{bmatrix}$ (b) $\alpha = [0.5, 0.4, 0.1]$, $\pi = \begin{bmatrix} 0.8 & 0.7 & 0.9 \\ 0.7 & 0.7 & 0.6 \\ 0.9 & 0.6 & 0.9 \end{bmatrix}$ (c) $\alpha = [0.8, 0.1, 0.1]$, $\pi = \begin{bmatrix} 0.8 & 0.3 & 0.5 \\ 0.3 & 0.7 & 0.2 \\ 0.5 & 0.2 & 0.9 \end{bmatrix}$

Figure 5: Erdos-Rényi Mixture for Graphs with $Q = 3$