Review of "A mixture model for random graphs"

Adib Habbou & Teddy Alexandre

Master MVA - Ecole Normale Supérieure Paris-Saclay

10/01/2024

General Context

Mixture models for graphs seek to improve graph modeling by integrating intricate structures like clustering. This approach aims to address the constraints of conventional models, offering a more realistic modelling of complex real-world graphs.

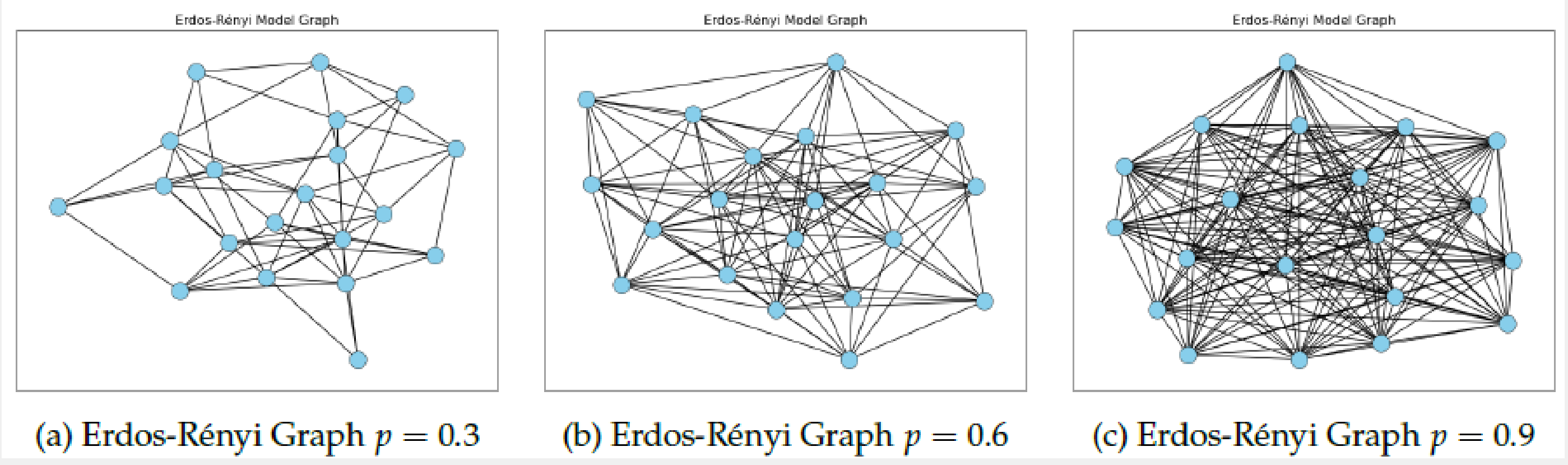
Erdos-Rényi Model (ER)

The Erdos-Rényi model generates random graphs by considering a set of n nodes, where each pair of nodes i and j is connected with probability p.

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

The probability distribution for the degree K_i follows a binomial distribution:

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



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

Figure 1. Erdos-Rényi Graphs

Beyond Erdos-Rényi

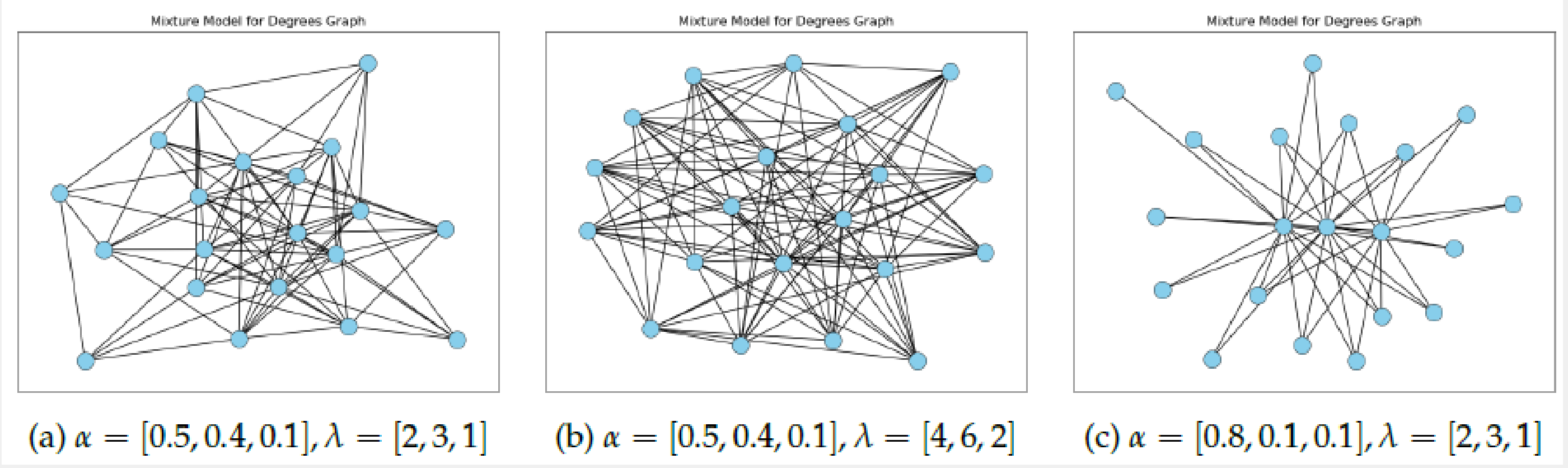
The Erdos-Rényi model offers foundational insights into graph properties but falls short in capturing real-world complexities, prompting exploration of more nuanced models. Mixture models address this issue by categorizing nodes into clusters.

Mixture Model for Degrees (MMD)

The Mixture Model for Degrees extends the Erdos-Rényi model by introducing Q clusters: nodes are grouped and their degrees follow a mixture of Poisson:

$$P(K_i = \sum_{j \neq i} X_{ij} = k) = \sum_{q=1}^{Q} \alpha_q P(K_i = k | Z_{iq} = 1) \qquad (3)$$

- α_q is the prior probability for node i to belong to cluster q
- Z_{iq} is the indicator variable of node i being in the cluster q


Mixture Model for Degrees Graph
Mixture Model for Degrees Graph
Mixture Model for Degrees Graph

(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) $\alpha = [0.8, 0.1, 0.1], \lambda = [2, 3, 1]$

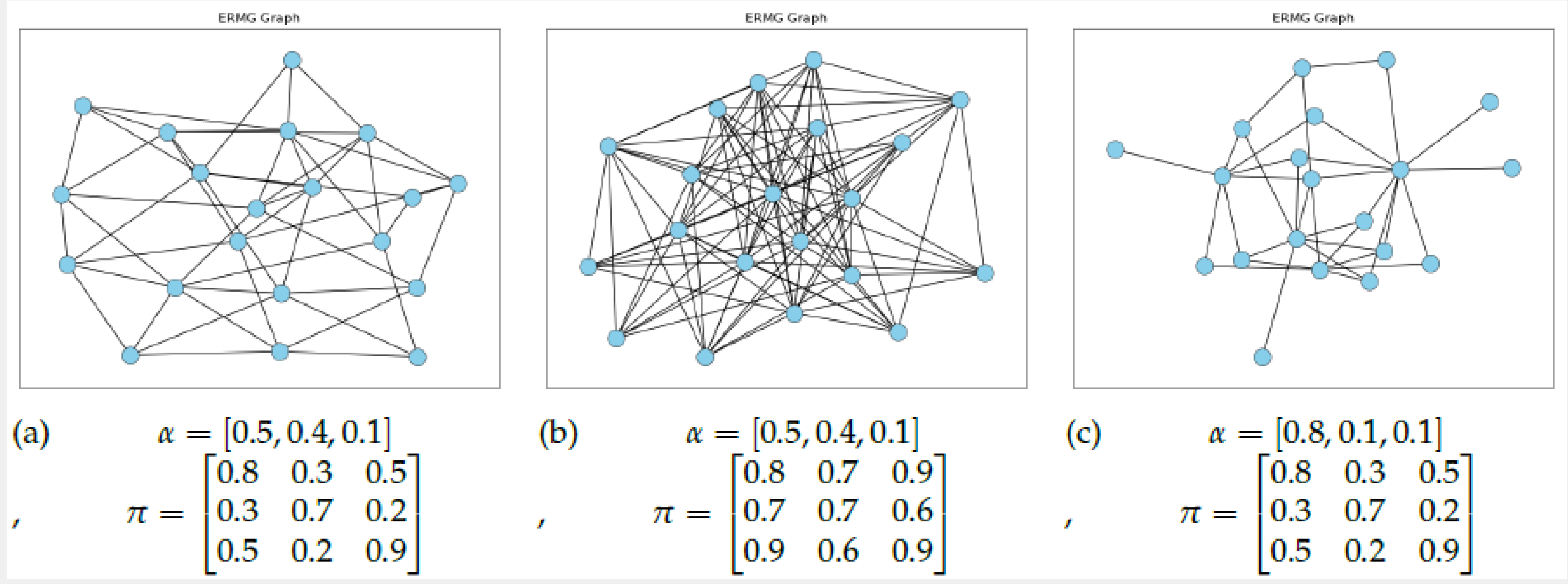
Figure 2. Mixture Model for Degrees Graphs with $Q = 3$

Erdos-Rényi Mixture for Graphs (ERMG)

The Erdos-Rényi Mixture for Graphs model comes with a new idea where nodes are distributed into Q clusters with probabilities α_q. It introduces symmetric probabilities π_{ql} for nodes from cluster q to connect with nodes from cluster l.

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

$$X_{ij} | \{Z_{iq} = 1, Z_{jl} = 1\} \sim \mathcal{B}(\pi_{ql}). \qquad (4)$$

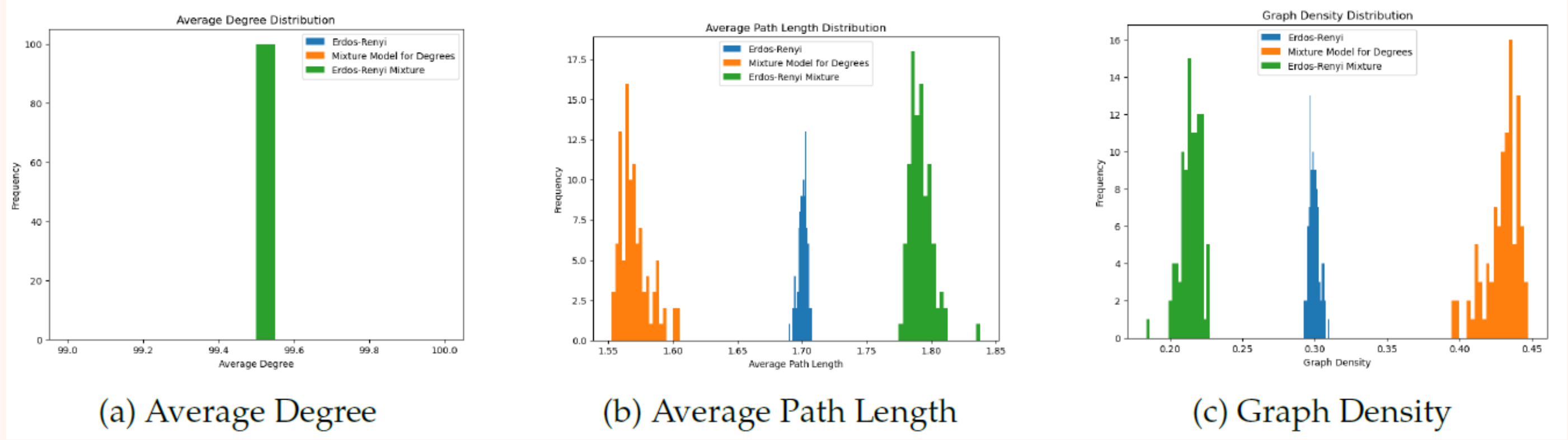


(a) $\alpha = [0.5, 0.4, 0.1]$, $\pi = \begin{bmatrix} 0.8 & 0.3 & 0.5 \\ 0.3 & 0.7 & 0.2 \\ 0.5 & 0.2 & 0.9 \end{bmatrix}$ (b) $\alpha = [0.5, 0.4, 0.1]$, $\pi = \begin{bmatrix} 0.8 & 0.7 & 0.9 \\ 0.7 & 0.7 & 0.6 \\ 0.9 & 0.6 & 0.9 \end{bmatrix}$ (c) $\alpha = [0.8, 0.1, 0.1]$, $\pi = \begin{bmatrix}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

Figure 3. Another figure caption.

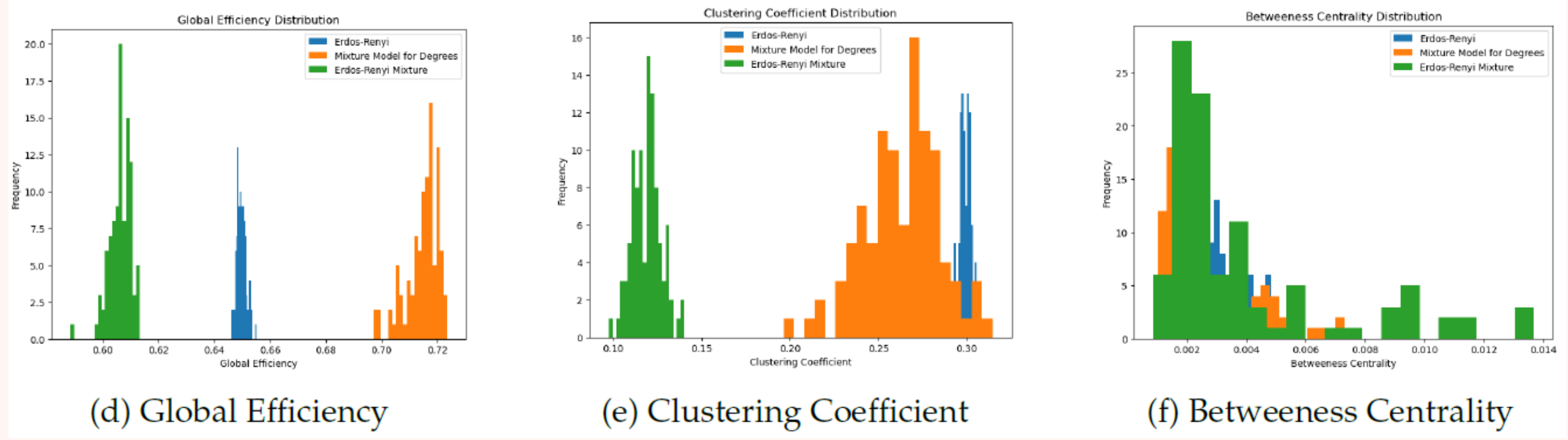
Graph Generation Comparison

We generated 100 graphs with 200 nodes per method, and compared them.



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

Figure 4. Comparison of average degree, average path length and graph density



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

Figure 5. Comparison of global efficiency, clustering coefficient and betweeness centrality

- Average degree of 99.5 across all methods indicates balanced connectivity
- MMD exhibits the lowest average path length due to its clustered structure
- ERMG has the highest average path length due to inter-cluster connections
- ERMG has the lowest graph density, global efficiency and clustering coeff, followed by ER and MMD, indicating varying connectivity patterns
- Betweenness centrality is lower for ER and MMD, with ERMG showing higher values for some graphs due to vertices connecting different clusters

Expectation-Maximization (EM) Algorithm

The Expectation-Maximization (EM) algorithm is employed to estimate the parameters of the mixture model. Here, 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.

The log-likelihood is intractable, we estimate a lower bound of this quantity :

$$\mathcal{J}(R_{\mathcal{X}}) = log\mathcal{L}(\mathcal{X}) - KL[R_{\mathcal{X}}(.), \mathbb{P}(.|\mathcal{X})] \qquad (5)$$

Where $R_{\mathcal{X}}$ is a class of distribution that depends on $\mathcal{X}$, here restricted to a product of multinomial distributions : $R_{\mathcal{X}}(\mathcal{Z}) = \prod_i h(\mathcal{Z}_i; \tau_i)$.

The E-step

Estimation of the posterior probabilities of one node i belonging to a cluster of graphs q, denoted $\hat{\tau}_{iq}$. They are derived with the following fixed-point relation :

$$\hat{\tau}_{iq} \propto \alpha_q \prod_{j \neq i} \prod_{l} b(X_{ij}; \pi_{ql})^{\hat{\tau}_{jl}} \quad \text{where} \quad b(x;\pi) = \pi^x (1-\pi)^{1-x} \qquad (6)$$

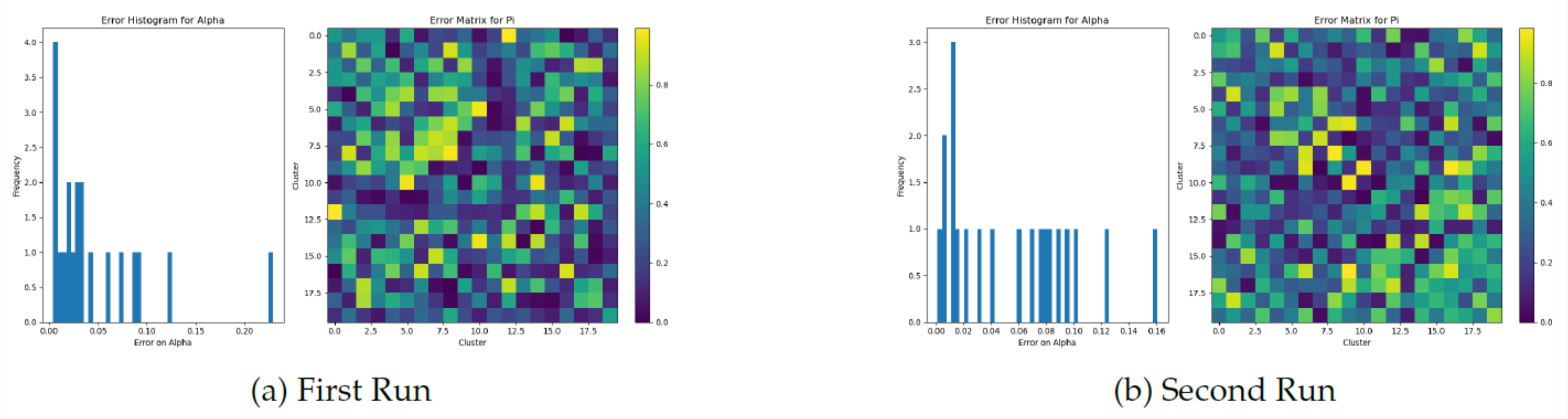
The M-step

Update the parameters of the mixture model: the prior probabilities α and the probabilities of connections between clusters π.

$$\hat{\alpha}_q = \frac{\sum_{i=1}^{n} \hat{\tau}_{iq}}{n} \qquad \hat{\pi}_{ql} = \frac{\sum_{i \neq j} \hat{\tau}_{iq} \hat{\tau}_{jl} X_{ij}}{\sum_{i \neq j} \hat{\tau}_{iq} \hat{\tau}_{jl}} \qquad (7)$$

Using EM for ERMG Parameters Estimation

Challenging experiments due to the $\hat{\tau}_{iq}$'s known up to a multiplicative constant depending on the Lagrange multiplier. We assumed there was no multiplicative factor and we applied EM twice on the exact same graph:



(a) First Run (b) Second Run

Figure 6. Errors made on α (histogram) and π (matrix) compared to the true values

Significant errors due to the miscomputation on the E-step giving estimated parameters which may be quite far from the ground-truth ones.

Challenges and Limitations in ERMG Modelling

- ERMG relies on a lot of specific assumptions (homogeneity between clusters, symmetric probabilities, fixed cluster membership...) $\longrightarrow$ **very restrictive**
- Aspects of real-world graphs are not handled by ERMG: community overlap, hierarchical organization, preferential attachment $\longrightarrow$ **not aligned with reality**
- EM algorithm is very sensitive to the initialization $\longrightarrow$ **hard to get convergence**