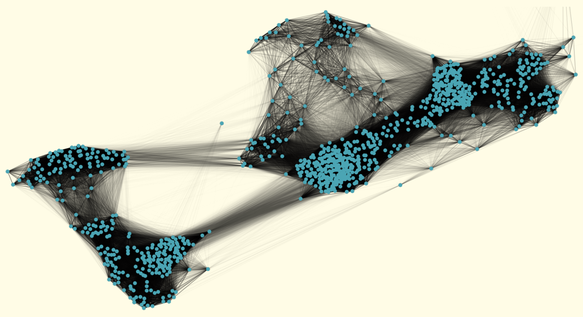
Graph Clustering and Community Detection

Martin Cuingnet & Matthieu Boyer

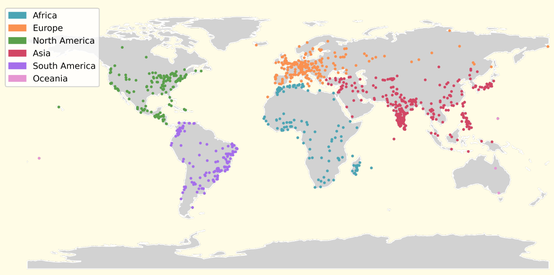
In this report, we are interested in extending graph clustering methods designed in (Daudin et al., 2006) with a probabilistic graphical model, to **weighted graphs**. This will allow us to define a graphical model for clustering on manifold-like graphs, at the cost of a bit of complexity and an added free parameter.

Problem Formulation

We consider a **distance graph** for **major cities** around the world and try to find a good model for the distribution of distances between cities.



Distance graph for 1000 cities (a node = a city)

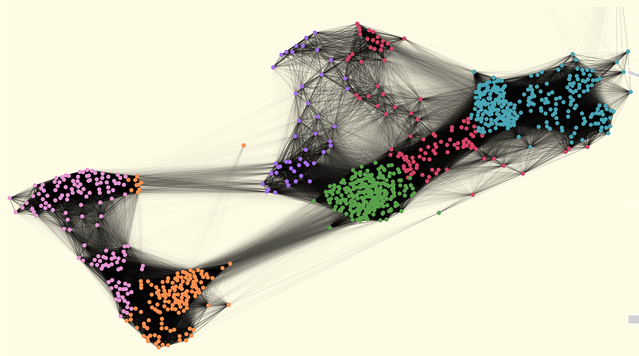


Major cities graph plotted on the world map with a possible continent-based clustering

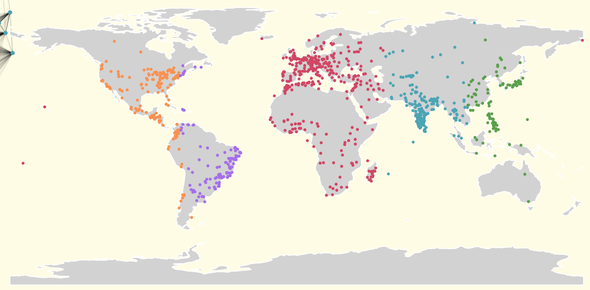
This graph was built using major world cities data from (SimpleMaps, 2025).

Results

We use our new graph model **wERMG** along with an **EM** algorithm to find the best model fitting the distance graph:



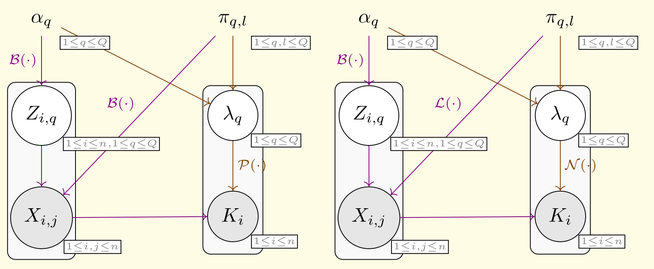
Predicted clusters on the distance graph



Predicted clusters plotted on the world map

The EM algorithm was able to actually identify the different **true geographic clusters** (North and South America, East Asia, India, ...).

Comparison



	ERMG	wERMG
$Z_{\mathrm{i,q}} \sim$	$\mathcal{B}(\alpha_q)$	$\mathcal{B}(\alpha_q)$
$X_{\mathrm{i,j}} \sim$	$\sum_{q,l} Z_{\mathrm{i,q}} Z_{\mathrm{j,l}} \mathcal{B}(\pi_{\mathrm{q,l}})$	$\sum_{q,l} Z_{\mathrm{i,q}} Z_{\mathrm{j,l}} \mathcal{L}(\pi_{\mathrm{q,l}})$
$K_{\mathrm{i}} =$	$\sum_j X_{\mathrm{i,j}}$	$\sum_j X_{\mathrm{i,j}}$
$\xrightarrow[n\to+\infty]{\text{law}}$	$\mathcal{P}(\lambda_q)$	$\mathcal{N}(n\mu_q, \sqrt{n}\sigma_q)$
$C_i =$	$\mathbb{E}_{j,k}[\mathbb{P}(X_{\mathrm{j,k}} = 1 \mid X_{\mathrm{i,j}} X_{\mathrm{i,k}} = 1)]$	$\mathbb{E}_{w,j,k}[\mathbb{P}(X_{\mathrm{j,k}} \leq w \mid X_{\mathrm{i,j}}, X_{\mathrm{i,k}} \leq w)]$

Degree law convergence

With $\sigma_n = \sqrt{\sum_j \sum_l \mathbb{V}\left[\alpha_{\mathrm{l}} \mathcal{L}(\pi_{\mathrm{q,l}})\right]} = \sigma_q \sqrt{n}$ and $X_{\mathrm{i,j}}$ independent, the Lyapunov Central Limit theorem then gives:

$$\frac{1}{\sigma_n} \sum_j \sum_l (\alpha \mathcal{L}(\pi) - \mathbb{E}[\alpha \mathcal{L}(\pi)])$$

$$= \frac{1}{\sigma_n} n \left(\sum_l \alpha \mathcal{L}(\pi) - \mu_q \right) = \frac{\sqrt{n}}{\sigma_q} (K_{\mathrm{i}} - n\mu_q)$$

$$\xrightarrow[n\to\infty]{\text{law}} \mathcal{N}(0,1)$$

Note that we retrieve the result for ERMG.

References

Daudin, J.-J., Picard, F., & Robin, S. (2006). *A mixture model for random graphs* (Research Report No. RR-5840; p. 19). https://inria.hal.science/inria-00070186

SimpleMaps. (2025,). *World Cities Database*. Pareto Software, LLC. https://simplemaps.com/data/world-cities

EM Algorithm

- **E-Step**:

$$\hat{\tau}_{\mathrm{i,q}} \propto \alpha_q \prod_{j \neq i} \prod_l \left[\mathcal{L}(\pi_{\mathrm{q,l}})(X_{\mathrm{i,j}})\right]^{Z_{\mathrm{j,l}}}$$

- **M-Step**:

$$\hat{\alpha}_q = \sum_i \frac{\hat{\tau}_{\mathrm{i,q}}}{n}$$

$$\hat{\pi}_{\mathrm{q,l}} = \underset{\pi}{\operatorname{argmax}} \left(\sum_i \sum_j \hat{\tau}_{\mathrm{i,q}} \hat{\tau}_{\mathrm{j,l}} \log(\mathcal{L}(\pi)(X_{\mathrm{i,j}})) \right)$$

When the collection of $\mathcal{L}(\pi)$ form an exponential family, we get:

$$\hat{\pi}_{\mathrm{q,l}} = \sum_i \sum_j \frac{\hat{\tau}_{\mathrm{i,q}} \hat{\tau}_{\mathrm{j,l}} X_{\mathrm{i,j}}}{\sum_i \sum_j \hat{\tau}_{\mathrm{i,q}} \hat{\tau}_{\mathrm{j,l}}}$$

Proof of the M step for wERMG:
Consider the complete-data log-likelihood:

$$Q(X) = \sum_i \sum_q \tau_{\mathrm{i,q}} \log \alpha_q + \sum_{i<j} \sum_{q,l} \theta_{\mathrm{i,q,j,l}} \log(\mathcal{L}(\pi_{\mathrm{q,l}})(X_{\mathrm{i,j}})).$$

With $\mathcal{L}(\pi)(x) = h(x) \exp(\eta T(x) - A(\eta))$, differentiating the expression of the complete-data log-likelihood as a parameter of $\eta = \eta(\pi)$:

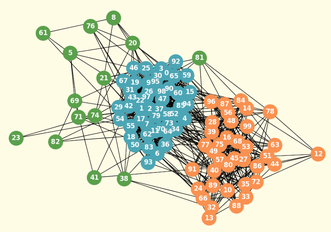
$$\frac{d}{d\eta} Q_{q,l}(\eta) = \sum_{i,j} \theta_{\mathrm{i,q,j,l}} T(X_{\mathrm{i,j}}) - W A'(\eta),$$

which vanishes when

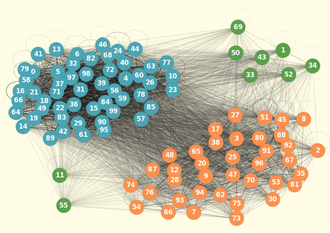
$$A'(\eta) \sum_{i,j} \theta_{\mathrm{i,q,j,l}} = \sum_{i,j} \theta_{\mathrm{i,q,j,l}} T(X_{\mathrm{i,j}}).$$

Experiments I

First we present a comparison of the expressiveness of our two models, as generating systems:



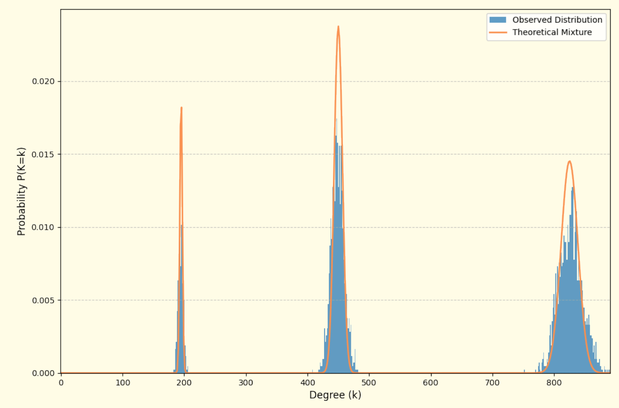
ERMG Graph



wERMG Graph

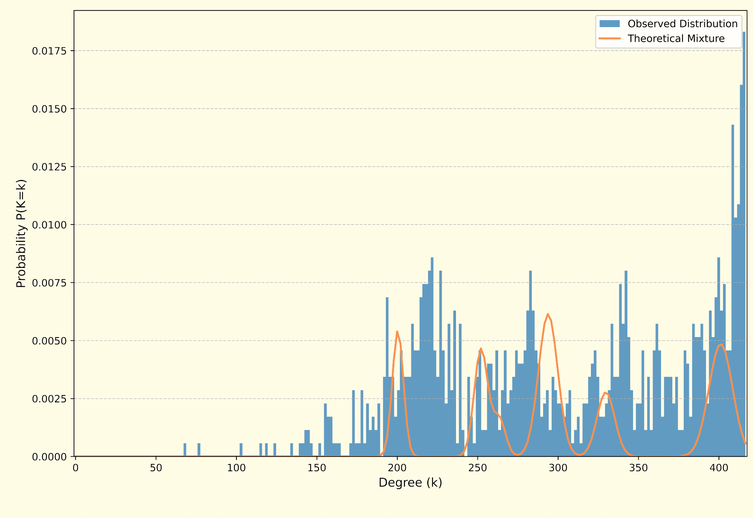
Experiments II

Here, we present how well the normal approximation works for $n = 3000$ vertices on graphs **generated at random** using a wERMG model.

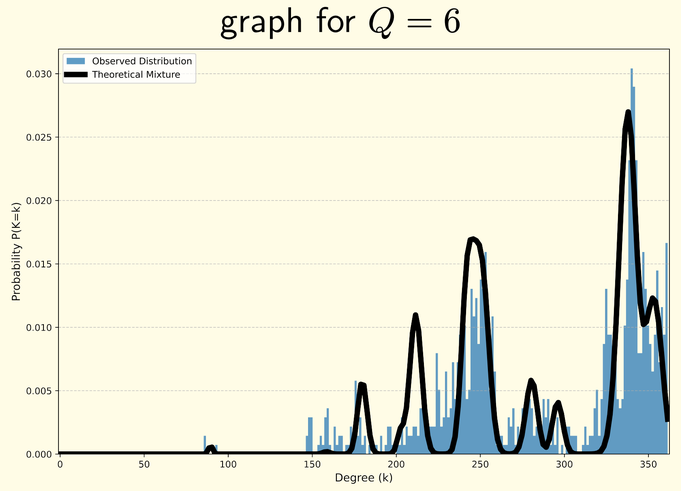


Observed and theoretical wERMG degree distribution

Below we present the **degree distribution** of the world cities distance graph along with their **normal approximation** obtained with wERMG EM for different Q values.



Observed degree distribution in the distance graph for $Q = 6$



Observed degree distribution in the distance graph for $Q = 20$