Graph Clustering an Community Detection

Martin Cuingnet & Matthieu Boyer

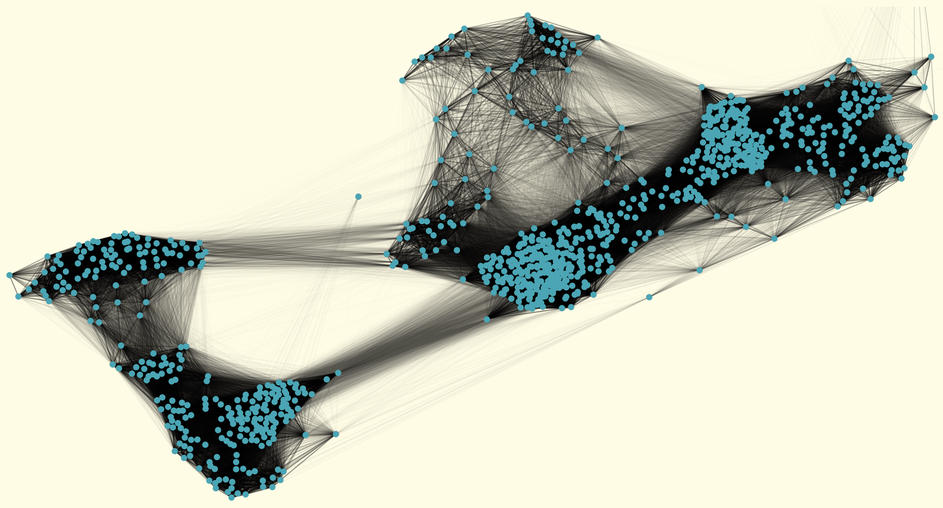
école normale supérieure paris-saclay

ENS 1794

In this report we are interested in extending graph clustering methods designed in [1] 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 of major cities around the world.

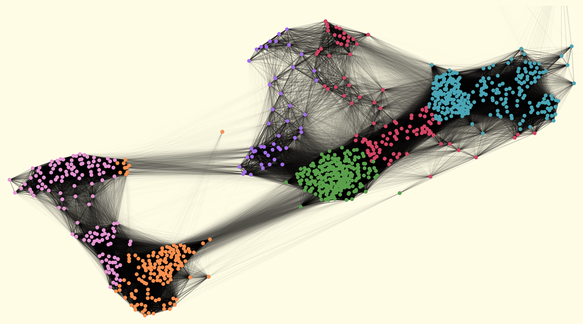


Distance graph of 1000 cities (a node = a city)

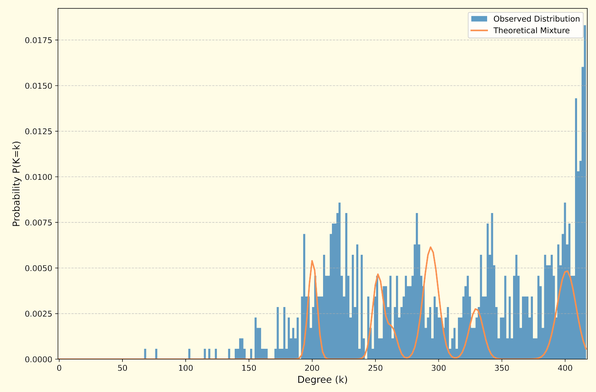
This graph was built by using the Harvesine distance on the world cities data from [2].

Results

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



Clustering of the graph with $Q = 6$

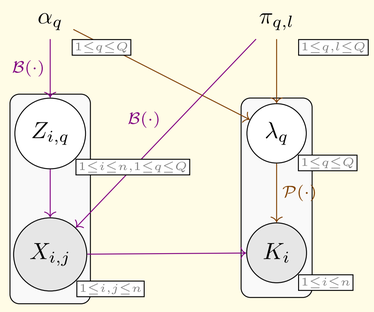
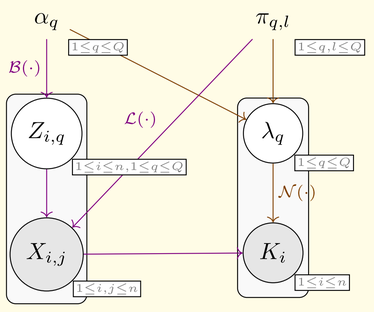


Observed degree distribution in the distance graph and its associated model prediction

The EM algorihtm was able to identify the different true geographic clusters (Europe, United States, Africa, ...).

Comparison

$X_{i,\,j}$: indicates if node i connected to node j
$Z_{i,\,q}$: indicates if node i is member of cluster q
K_i : degree of node i
C_i : clustering Coefficient of node i

ERMG	wERMG
	
$Z_{i,\,q} \sim \mathcal{B}(\alpha_q)$ $X_{i,\,j} \sim$ $\sum_{q,l} Z_{i,\,q} Z_{j,\,l} \mathcal{B}(\pi_{q,\,l})$ $K_i = \sum_j X_{i,j}$	$Z_{i,\,q} \sim \mathcal{B}(\alpha_q)$ $X_{i,\,j} \sim$ $\sum_{q,l} Z_{i,\,q} Z_{j,\,l} \mathcal{L}(\pi_{q,\,l})$ $K_i = \sum_j X_{i,\,j}$
$K_i \underset{n\to+\infty}{\overset{\text{law}}{\to}} \mathcal{P}(\lambda_q)$	$K_i \underset{n\to+\infty}{\overset{\text{law}}{\to}} \mathcal{N}(n\mu_q, \sqrt{n}\sigma_q)$
$C_i = \mathbb{E}_{j,k}\left[\mathbb{P}(X_{j,\,k} = 1 \mid X_{i,\,j} X_{i,\,k} = 1)\right]$	$C_i = \mathbb{E}_{w,j,k}\left[\mathbb{P}(X_{j,\,k} \le w \mid X_{i,\,j} \le w \wedge X_{i,\,k} \le w)\right]$
E : $\hat{\tau}_{i,\,q} \propto \alpha_q \prod_m b(\sum_k Z_{km} X_{i,\,k}; \sum_{j\neq i} Z_{j,m}, \pi_{q,m})$ M : $\hat{\alpha}_q = \sum_i \frac{\hat{\tau}_{i,\,q}}{n}$ $\hat{\pi}_{q,l} = \frac{\sum_i \sum_j \hat{\tau}_{i,\,q} \hat{\tau}_{j,\,l} X_{i,\,j}}{\sum_i \sum_j \hat{\tau}_{i,\,q} \hat{\tau}_{j,\,l}}$	With $\mathcal{L}(\pi)(x) = \frac{1}{\pi} e^{-\frac{x}{\pi}}$ E: $\hat{\tau}_{i,\,q} \propto \alpha_q \prod_{j\neq i} \prod_l \left[\frac{1}{\pi_{q,l}} e^{-\frac{x}{\pi_{q,l}}}\right]^{Z_{j,\,l}}$ M : $\hat{\alpha}_q = \sum_i \frac{\hat{\tau}_{i,\,q}}{n}$ $\hat{\pi}_{q,\,l} = \frac{\sum_i \sum_j \hat{\tau}_{i,\,q} \hat{\tau}_{j,\,l} X_{i,\,j}}{\sum_i \sum_j \hat{\tau}_{i,\,q} \hat{\tau}_{j,\,l}}$

Proofs

Proof of law convergence for K_i:

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

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

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

$$\underset{n\to\infty}{\overset{\text{law}}{\to}} \mathcal{N}(0,1)$$

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

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

The formulas derive from separation of the above in α_q and $\pi_{q,\,l}$.

Proof of the M step for wERMG in the exponential family case:

With

$$\mathcal{L}(\pi)(x) = h(x) \exp(\eta(\pi) T(x) - A(\eta(\pi))),$$

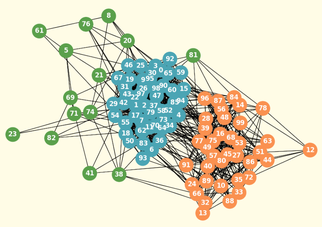
differentiating the expression of the complete-data log-likelihood as a parameter of η:

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

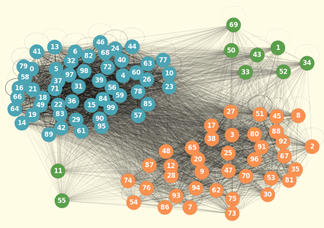
which vanishes when

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

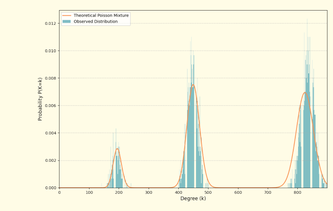
Experiments



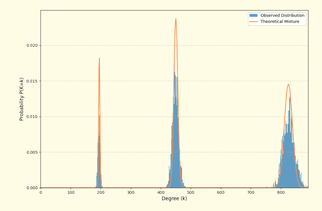
ERMG Graph



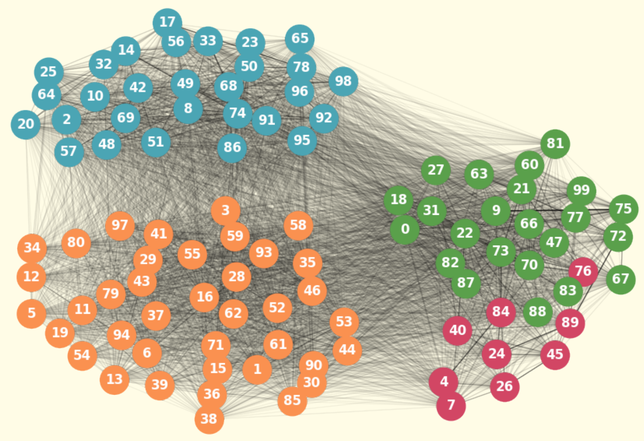
wERMG Graph



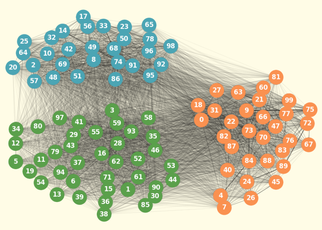
Observed ERMG degree distribution and associated poisson approximation



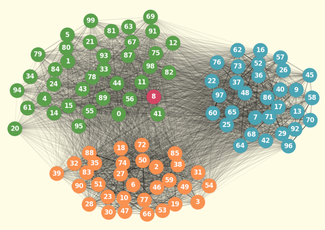
Observed wERMG degree distribution and associated normal approximation



True wERMG



Reconstructed Clusters with EM



Sample Reconstructed wERMG with EM

Visualization of EM on wERMG

References

[1] J.-J. Daudin, F. Picard, and S. Robin, "A mixture model for random graphs," Research Report RR-5840, 2006. [Online]. Available: https://inria.hal.science/inria-00070186

[2] SimpleMaps, "World Cities Database." [Online]. Available: https://simplemaps.com/data/world-cities