Graphons pour l'analyse de réseaux politiques sur le web

Soutenance de MODAL

Adélie Benhaim, Victor Mialot, Mathias Ollu

Ecole Polytechnique

13 février 2025

Introduction

Objectif

Analyser des réseaux, sociaux ou politiques

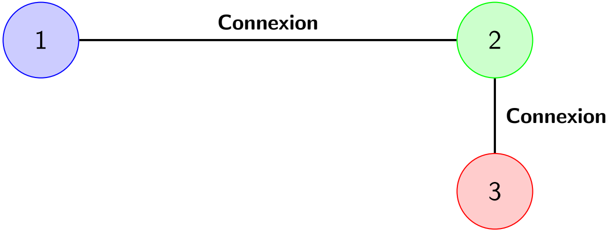
Pourquoi

Prédire des connexions, observer des structures

Comment

Utilisation des graphes aléatoires

Représentation d'un réseau avec un graphe



- Les nœuds représentent des personnes.
- Les arêtes représentent des connexions.

Type de graphe

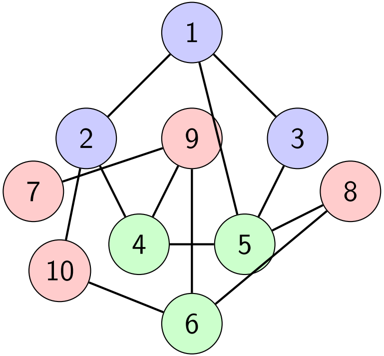
Avec un modèle de graphe trop simple, on perd la spécificité des relations humaines.

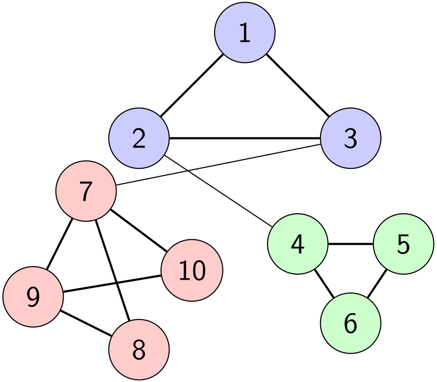
Intuitivement, il est clair qu'un réseau humain est formé de **communautés**.

On se donne par exemple 3 communautés distinctes :

- Les polyetchniciens X23
- Les agriculteurs d'Alaska
- Les gérants de pizzarias parisiens

Type de graphe





Il est clair que le réseau de droite est bien plus réaliste que celui de gauche

Plan

1. W-graphes et blocs stochastiques
2. Estimation du graphon en connaissant les clusters
3. Estimation des clusters avec le nombre de cluster connu
4. Estimation des clusters avec nombre de cluster inconnu

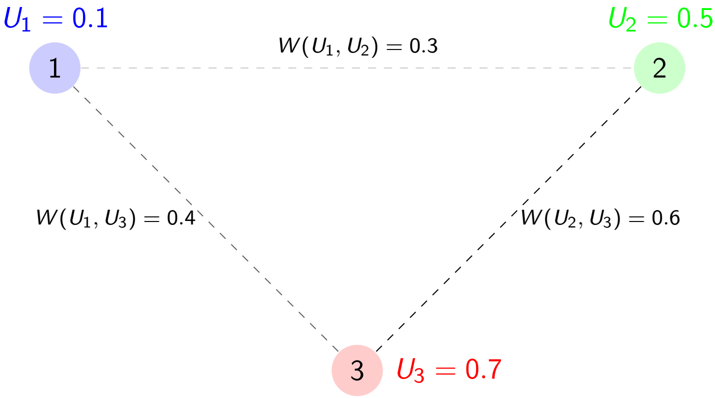
W-graphes

Définition du modèle de W-graphe

- W, le **graphon** : $W : [0,1]^2 \to [0,1]$
- A chaque sommet : $U_i \sim U([0,1])$, i.i.d.
- Arêtes : $X_{ij}|(U_i, U_j) \sim \mathrm{B}(W(U_i, U_j))$

W-graphes

Par exemple pour $W(x, y) = \frac{x+y}{2}$

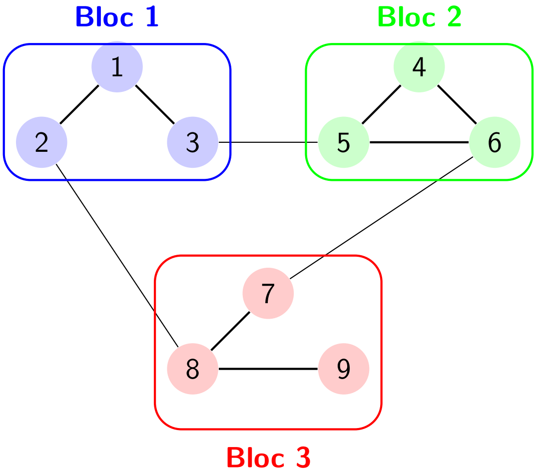


Le modèle à blocs stochastiques

Modèle à **blocs stochastiques** à K blocs :

- Chaque sommet a une probabilité π_i d'appartenir au bloc i.
- Pour des arêtes au sein d'un même bloc k la probabilité de se connecter vaut μ_{kk}
- Pour des arêtes de blocs distincts i et j la probabilité de se connecter vaut μ_{ij}

Le modèle à blocs stochastiques

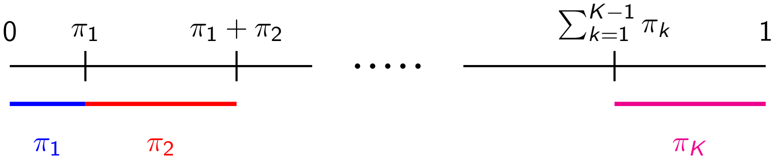


- Connexions fortes à l'intérieur des communautés
- Connexions différentes entre les communautés

Le modèle à blocs stochastiques

Le modèle à blocs stochastiques est un cas particulier de W-graphe.
$\rightarrow$ Trouver W qui réplique le comportement des blocs.

Il faut partitionner l'image d'une v.a. de loi uniforme :



Quelques propriétés

Propriétés

- La loi marginale de X_{ij} suit une loi de Bernouilli de paramètre

$$P(X_{ij}=1)=\int_0^1\int_0^1 P(X_{ij}=1|U_i,U_j)\,dU_i\,dU_j=\int_0^1\int_0^1 W(x,y)\,dx\,dy=w$$

- Les lois uniformes associées aux sommets ne sont plus indépendantes quand on connaît la matrice d'adjacence et le graphon.
- Tout W-graphe est dense ou vide.

$$\mathbb{E}[\sum_{i<j} X_{ij}]=\binom{n}{2}w$$

Identifiabilité du modèle

Il nous fallait montrer ce résultat :

Identifiabilité des W-graphes

Les W-graphes ne sont identifiables que si l'on se restreint à ceux vérifiant

$$g : x \to \int_0^1 W(x, y)\, dy$$

est croissante.

Pour le montrer, on s'est appuyé sur Kallenberg [4], Bickel et Chen [1]

Identifiabilité du modèle

1ère idée :

Si :

- $g_1(x) = \int_{[0,1]} W_1(x,y)\,dy$ et $g_2(x) = \int_{[0,1]} W_2(x,y)\,dy$ sont croissantes,
- $\int\int_{[0,1]^2} W_1(x,y)\,dx\,dy = \int\int_{[0,1]^2} W_2(x,y)\,dx\,dy$,

Alors $W_1 = W_2$.

Toutefois, ce résultat est faux si on prend :

$$W_1(x,y) = xy, \quad W_2(x,y) = \sin\left(\frac{\pi}{2}x\right)\sin\left(\frac{\pi}{2}y\right)\frac{\pi^2}{16}$$

Identifiabilité du modèle

Identifiabilité

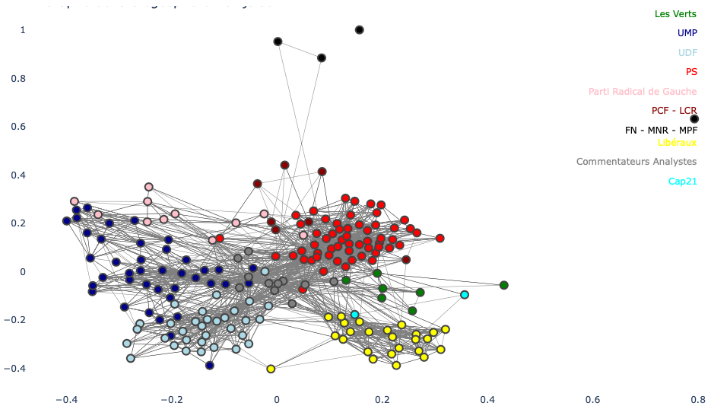
Il faut prendre en compte toutes les données ! Pas seulement les noeuds 2 par 2 en regardant la probabilité d'apparation d'une arête.

Par exemple en prenant la configuration à 3 noeuds en regardant l'appartion d'un "V", nous voyons qu'il faut aussi avoir l'égalité des intégrales doubles car :

$$\int_0^1 \int_0^1 W_1(x,y)\,dy \int_0^1 W_1(x,z)\,dz\,dx = \int_0^1 g_1(x)^2\,dx$$

Présentation des données

Données à disposition : 196 blogs politiques de 2007, annotés par des politologues selon 10 partis politiques.



Estimation pour Z connu

On prend les estimateurs naturels :

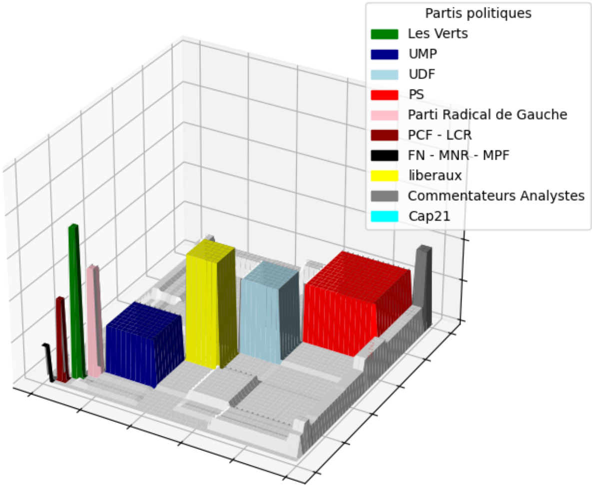
$$n_k = \sum_{i=1}^{n} Z_{ik} \quad \text{et} \quad \hat{\pi}_k = \frac{n_k}{n}, \quad \forall k \in \{1, \dots, K\},$$

$$\mu_{kk} = \frac{1}{n_k(n_k - 1)} \sum_{i \neq j}^{n} Z_{ik} Z_{jk} X_{ij}, \quad \forall k \in \{1, \dots, K\},$$

$$\mu_{kl} = \frac{1}{n_k n_l} \sum_{i \neq j}^{n} Z_{ik} Z_{jl} X_{ij}, \quad \forall k \neq l.$$

Affichage du graphon

A partir de ces estimations on peut retrouver la fonction graphon :



Estimation pour Z inconnu

Approximer :

- Z par τ dont les lignes sont : $\tau_i = (\tau_{i1}, \dots, \tau_{iK})$
- $\mathbb{P}(Z|X)$ par $R_X(Z) = \prod_i h(Z_i; \tau_i)$, h une distrib multinomiale de paramètre τ_i

Maximiser une fonction objectif

$J(R_X) = \log(L(X)) - \mathbb{KL}[R_X(.) \mid Pr(. \mid X)]$, qui devient :

$$\mathcal{J}(R_\mathcal{X}) = \sum_i \sum_q \tau_{iq} \log \pi_q + \frac{1}{2} \sum_{i \neq j} \sum_{q,\ell} \tau_{iq} \tau_{j\ell} \log b(X_{ij}; \mu_{q\ell}) - \sum_i \sum_q \tau_{iq} \log \tau_{iq}$$

Daudin, Picard et Robin [3]

Conditions du premier ordre

Maximisation de $\mathcal{J}(R_{\mathcal{X}})$ via un lagrangien :

Sur $\hat{\pi}$:

$$\hat{\pi}_q = \frac{1}{n}\sum_i \hat{\tau}_{iq}$$

Sur $\hat{\mu}$:

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

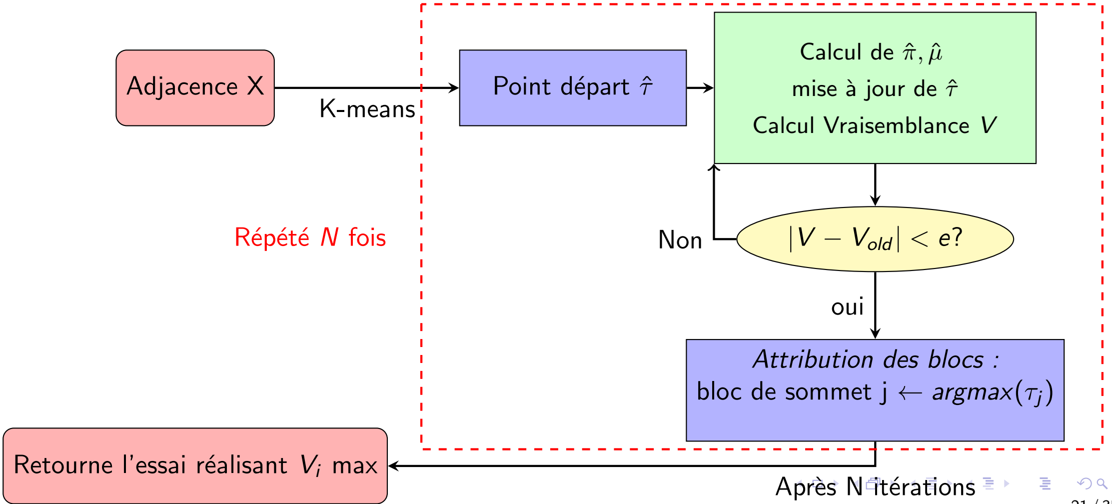
Sur $\hat{\tau}$:

$$\hat{\tau}_{iq} \propto \pi_q \prod_{j\neq i}\prod_{l} b(X_{ij};\mu_{ql})^{\hat{\tau}_{jl}}$$

Astuce utile

Les $\hat{\tau}_i$ sont des mesures de probabilité !

Algorithme d'estimation des blocs



Difficultés rencontrées

1. Temps de calcul long très long. $\rightarrow$ nombres d'itérations maximum ? Quelle valeur ? Compromis entre précision et temps d'exécution.
2. Quel point de départ ? Aléatoire puis K-moyennes
3. Quel critère d'arrêt ? sur paramètres/ vraisemblance. Quelle valeur de e ?
4. Tests difficiles si effectués directement sur les données

Erreur

A priori, pas de raison que nos clusters suivent parfaitement les partis politiques, ce n'est pas une mesure de la qualité de notre algorithme !

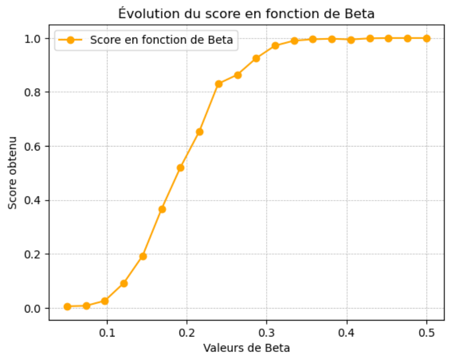
Tester sur des graphes simples : les β-graphes

Intuition : tester la détection de blocs très évidents, selon Côme et Latouche [2]

Générer un graphe aléatoire à blocs stochastiques de paramètres :

$$\pi = \left(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right)$$

$$\mu = \begin{bmatrix} \beta & 0.01 & \cdots & 0.01 \\ 001 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0.01 \\ 0.01 & \cdots & 0.01 & \beta \end{bmatrix}$$



Résultats

	UMP	PS	UDF	liberaux	Commentateurs Analystes	Les Verts	parti Radical de Gauche	PCF - LCR	FN - MNR - MPF	Cap21
cluster1	24	0	1	0	0	0	0	0	0	0
cluster2	11	0	0	0	0	0	0	0	0	0
cluster3	0	21	0	0	0	0	0	0	0	0
cluster4	2	10	1	0	6	2	2	1	0	0
cluster5	2	0	24	0	0	0	0	0	0	0
cluster6	0	0	6	0	0	0	0	0	0	0
cluster7	0	0	6	0	0	0	0	0	0	0
cluster8	0	0	0	24	1	0	0	0	0	0
cluster9	0	0	0	0	3	0	0	0	0	0
cluster10	1	5	0	1	1	5	9	6	4	2

Z inconnu et K inconnu

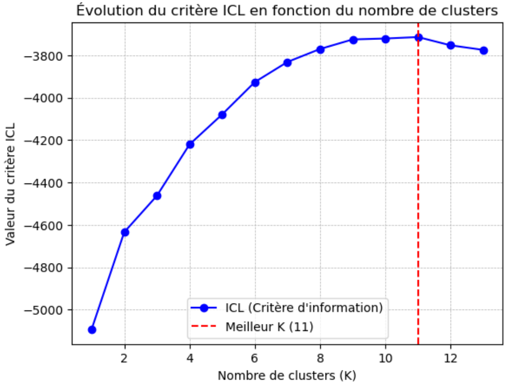
Idée : On lance notre algorithme précédent pour $K : 1 \to N$ et on sélectionne K^* qui donne les meilleurs résultats pour l'ICL.

Définition de l'ICL

Nouvelle quantification d'un "bon essai" qui prend en compte le nombre de blocs en pénalisant quand le nombre devient trop élevé.

$$ICL(X, k) = max_{\pi,\mu} \log(L(X, Z_1|\pi, \mu, K)) - \frac{K(K+1)}{4} \log(\frac{n(n-1)}{2}) - \frac{K-1}{2} \log(n)$$

Résultat de l'ICL



$K^* = 11$, soit un de plus que la réalité. Ce n'est pas inquiétant, l'information de connexion n'équivaut pas à la labellisation politique.

Résultat

	PS	UMP	UDF	liberaux	Commentateurs Analystes	Les Verts	Parti Radical de Gauche	PCF - LCR	FN - MNR - MPF	Cap21
cluster1	22	0	0	0	0	0	0	0	0	0
cluster2	20	0	0	0	0	0	0	0	0	0
cluster3	2	0	0	0	0	0	0	0	0	0
cluster4	0	24	1	0	0	0	0	0	0	0
cluster5	0	11	0	0	0	0	0	0	0	0
cluster6	0	2	24	0	1	0	0	0	0	0
cluster7	0	0	6	0	0	0	0	0	0	0
cluster8	0	0	0	24	1	0	0	0	0	0
cluster9	0	0	0	0	3	0	0	0	0	0
cluster10	8	2	1	0	5	2	2	1	0	0
cluster11	5	1	0	1	1	5	9	6	4	2

Forte correspondance entre clusters et partis.

Anomalies au niveau de l'UMP/UDF.
$\rightarrow$ erreurs d'étiquetages des politologues, l'algorithme arrive donc à repérer des erreurs !

Références

[1] Peter Bickel et Aiyou Chen. « A nonparametric view of network models and Newman-Girvan and other modularities ». In : *Proceedings of the National Academy of Sciences of the United States of America* 106 (nov. 2009), p. 21068-73. DOI : 10.1073/pnas.0907096106.

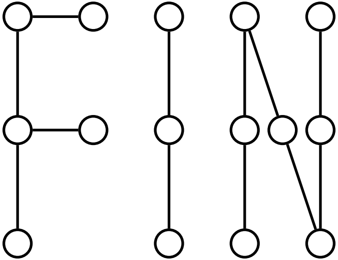
[2] Etienne Côme et Pierre Latouche. « Model selection and clustering in stochastic block models with the exact integrated complete data likelihood ». In : *Statistical Modelling* 15 (mars 2013). DOI : 10.1177/1471082X15577017.

[3] Jean-Jacques Daudin, Franck Picard et Stéphane Robin. « A mixture model for random graph ». In : *Statistics and Computing* 18 (juin 2008), p. 173-183. DOI : 10.1007/s11222-007-9046-7.

[4] O. Kallenberg. « Probabilistic Symmetries and Invariance Principles ». In : (jan. 2005). DOI : 10.1007/0-387-28861-9.

Fin

Merci de votre attention !



Annexe

$$
\begin{aligned}
f_{U|X,W}(u_1,\dots,u_n,x_{11},\dots,x_{nn}) &= \frac{f_{U,X,W}(u_1,\dots,u_n,x_{11},\dots,x_{nn})}{f_{X,W}(x_{11},\dots,x_{nn})} \\
&= \frac{f_{X|U,W}(x_{11},\dots,x_{nn})f_{U,W}(u_1,u_2,\dots,u_n)}{f_{X,W}(x_{11},\dots,x_{nn})} \\
&= \frac{\prod_{i<j} W(u_i,u_j)^{x_{ij}}\left(1-W(u_i,u_j)\right)^{1-x_{ij}} 1_{[0,1]}(u_1)..1_{[0,1]}(u_n)}{\prod_{i<j} w^{x_{ij}}\left(1-w\right)^{1-x_{ij}}} \\
&= \frac{\prod_{i<j} W(u_i,u_j)^{x_{ij}}\left(1-W(u_i,u_j)\right)^{1-x_{ij}}}{\prod_{i<j} w^{x_{ij}}\left(1-w\right)^{1-x_{ij}}}
\end{aligned}
$$

Annexe

$$W(u,v) = \sum_{l_1=1}^{K} \sum_{l_2=1}^{K} \mu_{l_1,l_2} 1_{\{u \in [\sum_{k=1}^{l_1-1} \pi_k, \sum_{k=1}^{l_1} \pi_k], v \in [\sum_{k=1}^{l_2-1} \pi_k, \sum_{k=1}^{l_2} \pi_k]\}}(u,v)$$

$$\mathbb{P}(U_i \in [\sum_{k=1}^{l_1-1} \pi_k, \sum_{k=1}^{l_1} \pi_k]) = \sum_{k=1}^{l_1} \pi_k - \sum_{k=1}^{l_1-1} \pi_k = \pi_{l_1}$$

et

$$\mathbb{P}(U_j \in [\sum_{k=1}^{l_2-1} \pi_k, \sum_{k=1}^{l_2} \pi_k]) = \sum_{k=1}^{l_2} \pi_k - \sum_{k=1}^{l_2-1} \pi_k = \pi_{l_2}$$

Annexe

$$
\begin{aligned}
J(R_X) &= \log(L(X)) - \mathbb{KL}[R_X(.) \mid Pr(. \mid X)] \\
&= \log(L(X)) - \sum_Z R_X(Z) \log\left(\frac{R_X(Z)}{Pr(Z \mid X)}\right) \\
&= \log(L(X)) - \sum_Z R_X(Z) \log R_X(Z) + \sum_Z R_X(Z) \log Pr(Z \mid X) \\
&= -\sum_Z R_X(Z) \log R_X(Z) + \sum_Z R_X(Z) \log\left(Pr(Z \mid X) Pr(X)\right) \\
&= -\sum_Z R_X(Z) \log R_X(Z) + \sum_Z R_X(Z) \log\left(Pr(Z, X)\right)
\end{aligned}
$$

Annexe

$$\sum_Z R_X(Z)\log R_X(Z) = \sum_i \sum_q \tau_{iq} \log \tau_{iq}$$

$$\log L(Z,X) = \log L(Z) + \log L(Z|X)$$

$$log(L(Z)) = \sum_i \sum_q Z_{iq} log(\pi_q)$$

$$log(L(Z)) = \sum_i \sum_q \tau_{iq} log(\pi_q)$$

$$log(L(X|Z)) = \frac{1}{2} \sum_{i \neq j} \sum_{q,l} Z_{iq} Z_{jl} log(b(X_{ij};\mu_{ql}))$$

avec

$$b(x,\pi) = \pi^x (1-\pi)^{1-x}$$

Annexe

$$log(L(X|Z)) = \frac{1}{2}\sum_{i\neq j}\sum_{q,l}\tau_{iq}\tau_{jl}log(b(X_{ij};\mu_{ql}))$$

$$\mathcal{J}(R_{\mathcal{X}}) = \sum_{i}\sum_{q}\tau_{iq}\log\pi_q + \frac{1}{2}\sum_{i\neq j}\sum_{q,\ell}\tau_{iq}\tau_{j\ell}\log b(X_{ij};\mu_{q\ell}) - \sum_{i}\sum_{q}\tau_{iq}\log\tau_{iq}$$

Pour vérifier l'efficacité de notre programme, nous utilisons l'information mutuelle normalisée (IMN) des matrices Z^e estimée et Z^s simulée pour des modèles simples.

$$IMN(Z^s, Z^e) = \frac{I(Z^s, Z^e)}{max(H(Z^s), H(Z^e))}$$

où

$$I(Z^s, Z^e) = \sum_{k,l}^{K} p_{kl} log\left(\frac{p_{kl}}{p_k^s p_l^s}\right)$$

$$H(Z) = -\sum_{k}^{K} p_k log(p_k)$$

avec $p_{kl} = \frac{1}{N}\sum_{i,j}^{N} Z_{ik}^e Z_{jl}^s$ et $p_k = \frac{1}{N}\sum_{i}^{N} Z_{ik}$