Inférence dans le Stochastic Block Model pour les grands graphes

Antoine Channarond, Jean-Jacques Daudin, Stephane Robin

Inférence dans le Stochastic Block Model pour les grands graphes

Antoine Channarond, Jean-Jacques Daudin, Stéphane Robin

AgroParisTech

FGG'10

Présentation du Stochastic Block Model

Les graphes sont non orientés binaires. Ils sont désignés par leur nombre de sommets n et leur matrice d'adjacence $X = (X_{ij})_{1 \leq i,j \leq n}$.

Présentation du Stochastic Block Model

Les graphes sont non orientés binaires. Ils sont désignés par leur nombre de sommets n et leur matrice d'adjacence $X = (X_{ij})_{1 \leq i,j \leq n}$.

$(Z_i)_{1 \leq i \leq n}$ une suite i.i.d. de variables multinomiales à Q états, de paramètre

$$\alpha = (\alpha_1, \ldots, \alpha_Q)$$

Présentation du Stochastic Block Model

Les graphes sont non orientés binaires. Ils sont désignés par leur nombre de sommets n et leur matrice d'adjacence $X = (X_{ij})_{1 \leq i,j \leq n}$.

$(Z_i)_{1 \leq i \leq n}$ une suite i.i.d. de variables multinomiales à Q états, de paramètre

$$\alpha = (\alpha_1, \ldots, \alpha_Q)$$

$(X_{ij})_{1 \leq i,j \leq n}$ une matrice de variables indépendantes conditionnellement à Z. Sachant $Z_i = q$ et $Z_j = r$, X_{ij} suit une loi de Bernoulli de paramètre π_{qr}. On appelle matrice de connectivité la matrice symétrique :

$$\pi = (\pi_{qr})_{1 \leq q,r \leq Q}$$

Présentation du Stochastic Block Model

Les graphes sont non orientés binaires. Ils sont désignés par leur nombre de sommets n et leur matrice d'adjacence $X = (X_{ij})_{1 \leq i,j \leq n}$.

- Variables cachées : $(Z_i)_{1 \leq i \leq n}$ une suite i.i.d. de variables multinomiales à Q états, de paramètre
$$\alpha = (\alpha_1, \ldots, \alpha_Q)$$
- Variables observées : $(X_{ij})_{1 \leq i,j \leq n}$ une matrice de variables indépendantes conditionnellement à Z. Sachant $Z_i = q$ et $Z_j = r$, X_{ij} suit une loi de Bernoulli de paramètre π_{qr}. On appelle matrice de connectivité la matrice symétrique :
$$\pi = (\pi_{qr})_{1 \leq q,r \leq Q}$$

Présentation du Stochastic Block Model

Les graphes sont non orientés binaires. Ils sont désignés par leur nombre de sommets n et leur matrice d'adjacence $X = (X_{ij})_{1 \leq i,j \leq n}$.

- Variables cachées : $(Z_i)_{1 \leq i \leq n}$ une suite i.i.d. de variables multinomiales à Q états, de paramètre
$$\alpha = (\alpha_1, \ldots, \alpha_Q)$$
- Variables observées : $(X_{ij})_{1 \leq i,j \leq n}$ une matrice de variables indépendantes conditionnellement à Z. Sachant $Z_i = q$ et $Z_j = r$, X_{ij} suit une loi de Bernoulli de paramètre π_{qr}. On appelle matrice de connectivité la matrice symétrique :
$$\pi = (\pi_{qr})_{1 \leq q,r \leq Q}$$

Loi des degrés : $D_i | Z_i = q \sim \mathcal{B}(n-1, \overline{\pi}_q)$, où $\overline{\pi}_q = \sum_{r=1}^{Q} \alpha_r \pi_{qr}$.

Cadre statistique

- Modèle génératif statistique à paramètres dont on veut faire l'inférence en déterminant des estimateurs consistants. La classification n'est pas un but en soi, mais un sous-produit de la méthode d'inférence.

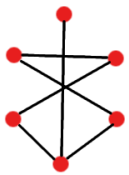
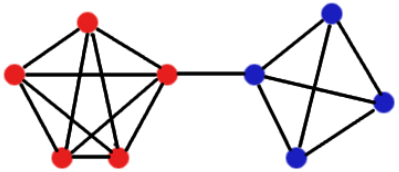
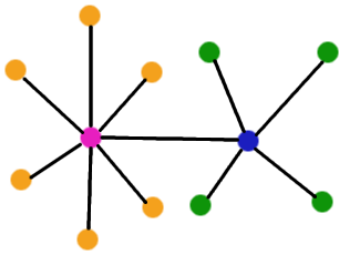
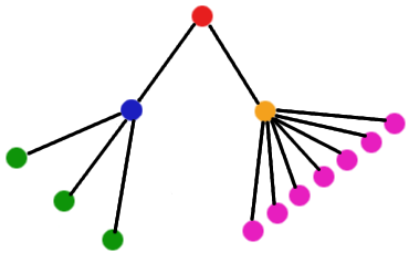
Cadre statistique

- Modèle génératif statistique à paramètres dont on veut faire l'inférence en déterminant des estimateurs consistants. La classification n'est pas un but en soi, mais un sous-produit de la méthode d'inférence.
- Modèle à classes cachées : introduction d'une structure sous-jacente non observée $\Rightarrow$ modèle plus riche qu'Erdős-Rényi par exemple.

Cadre statistique

- Modèle génératif statistique à paramètres dont on veut faire l'inférence en déterminant des estimateurs consistants. La classification n'est pas un but en soi, mais un sous-produit de la méthode d'inférence.
- Modèle à classes cachées : introduction d'une structure sous-jacente non observée $\Rightarrow$ modèle plus riche qu'Erdős-Rényi par exemple.

TABLE: Exemples de modèles

$Q = 1$	$Q = 3$	$Q = 4$	$Q = 5$
			
$\pi = 0.4$	$\pi = \begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}$	$\pi = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$	$\pi = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$

Enjeux de l'inférence dans le SBM

Maximum de vraisemblance trop complexe algorithmiquement. Méthodes d'inférence alternatives existantes :

- Algorithme EM : aussi complexe que le maximum de vraisemblance.

Enjeux de l'inférence dans le SBM

Maximum de vraisemblance trop complexe algorithmiquement. Méthodes d'inférence alternatives existantes :

- Algorithme EM : aussi complexe que le maximum de vraisemblance.
- MCMC (Snijders et Nowicki, 2001) : jusqu'à quelques **centaines** de sommets.

Enjeux de l'inférence dans le SBM

Maximum de vraisemblance trop complexe algorithmiquement. Méthodes d'inférence alternatives existantes :

- Algorithme EM : aussi complexe que le maximum de vraisemblance.
- MCMC (Snijders et Nowicki, 2001) : jusqu'à quelques **centaines** de sommets.
- Variationnel (Daudin, Picard, Robin, 2008) : jusqu'à quelques **milliers** de sommets.

Enjeux de l'inférence dans le SBM

Maximum de vraisemblance trop complexe algorithmiquement. Méthodes d'inférence alternatives existantes :

- Algorithme EM : aussi complexe que le maximum de vraisemblance.
- MCMC (Snijders et Nowicki, 2001) : jusqu'à quelques **centaines** de sommets.
- Variationnel (Daudin, Picard, Robin, 2008) : jusqu'à quelques **milliers** de sommets.

Point commun de ces méthodes : elles mettent à jour alternativement la classification des sommets et les estimateurs des paramètres.

Enjeux de l'inférence dans le SBM

Maximum de vraisemblance trop complexe algorithmiquement. Méthodes d'inférence alternatives existantes :

- Algorithme EM : aussi complexe que le maximum de vraisemblance.
- MCMC (Snijders et Nowicki, 2001) : jusqu'à quelques **centaines** de sommets.
- Variationnel (Daudin, Picard, Robin, 2008) : jusqu'à quelques **milliers** de sommets.

Point commun de ces méthodes : elles mettent à jour alternativement la classification des sommets et les estimateurs des paramètres.
La classification est un sous-produit de l'inférence, mais reste l'obstacle principal à l'inférence. Si les classes étaient révélées, il suffirait d'utiliser les estimateurs des moments usuels.

Si $(\mathcal{C}_q)_{1\leq q\leq Q}$ est la partition en classes : $\widehat{\alpha}_q = \dfrac{|\mathcal{C}_q|}{n}$ et $\widehat{\pi}_{qr} = \dfrac{1}{|\mathcal{C}_q||\mathcal{C}_r|} \displaystyle\sum_{(i,j)\in\mathcal{C}_q\times\mathcal{C}_r} X_{ij}$

Phénomène de concentration des degrés normalisés

Histogrammes des degrés normalisés $T_i = \frac{D_i}{n}$:

FIGURE: $n = 500$

FIGURE: $n = 5000$

FIGURE: $n = 15000$

Phénomène de concentration des degrés normalisés

Histogrammes des degrés normalisés $T_i = \frac{D_i}{n}$:

FIGURE: $n = 500$

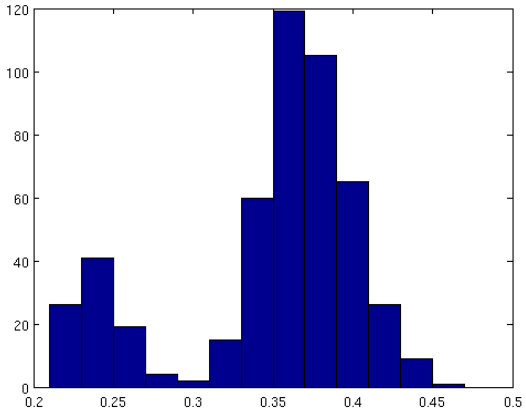


FIGURE: $n = 5000$

FIGURE: $n = 15000$

Phénomène de concentration des degrés normalisés

Histogrammes des degrés normalisés $T_i = \frac{D_i}{n}$:

FIGURE: $n = 500$

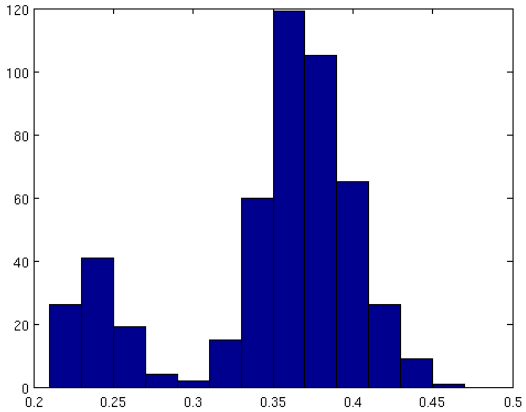


FIGURE: $n = 5000$

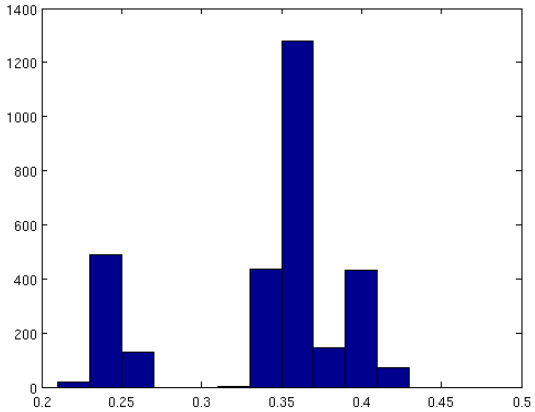


FIGURE: $n = 15000$

Phénomène de concentration des degrés normalisés

Histogrammes des degrés normalisés $T_i = \frac{D_i}{n}$:

FIGURE: $n = 500$

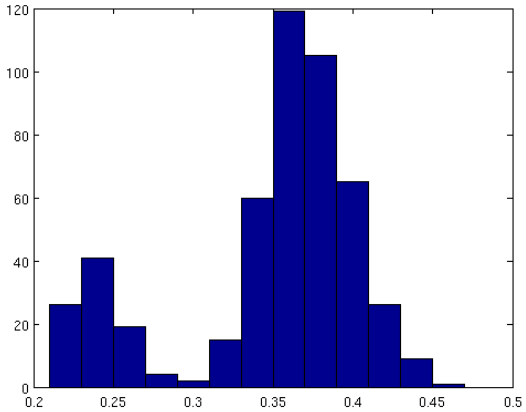


FIGURE: $n = 5000$

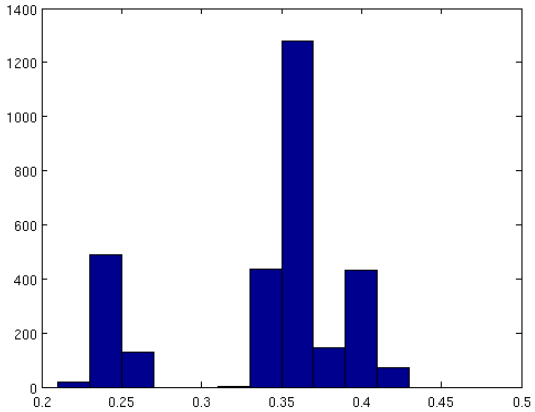
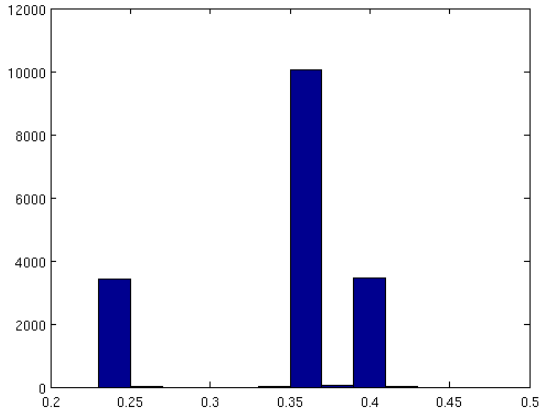


FIGURE: $n = 15000$



- Plus n est grand, plus la structure en classes se révèle d'elle-même dans la distribution des degrés, et donc plus l'inférence sera facile. Cependant la classification n'est pas un but en soi : elle ne servira qu'à mieux inférer.

Phénomène de concentration des degrés normalisés

Histogrammes des degrés normalisés $T_i = \frac{D_i}{n}$:

FIGURE: $n = 500$

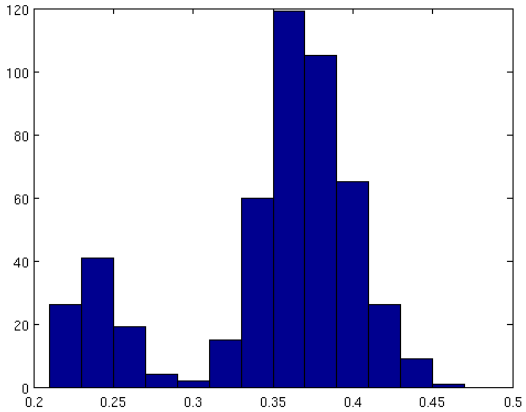


FIGURE: $n = 5000$

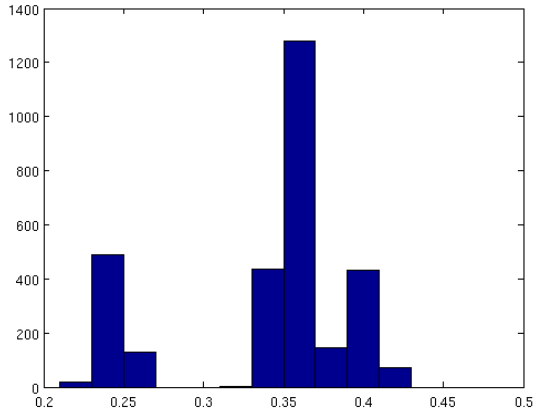
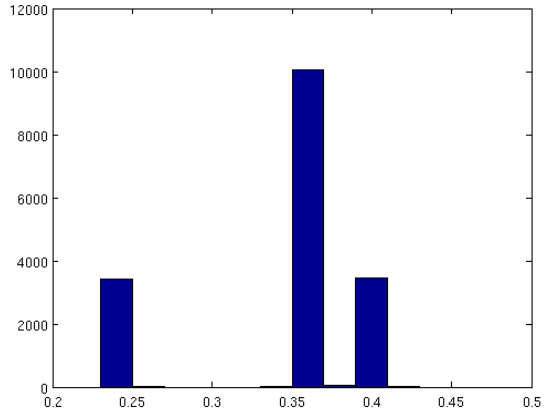


FIGURE: $n = 15000$



- Plus n est grand, plus la structure en classes se révèle d'elle-même dans la distribution des degrés, et donc plus l'inférence sera facile. Cependant la classification n'est pas un but en soi : elle ne servira qu'à mieux inférer.
- Attention, les degrés ne sont pas indépendants, et on ne peut pas utiliser d'algorithme usuel dédié aux modèles de mélanges (type EM) pour faire l'inférence des paramètres.

Algorithme consistant des plus grands écarts (PGE)

- Ordonner la suite $(T_i)_{1 \leq i \leq n}$: $T_{(1)} \leq \cdots \leq T_{(n)}$

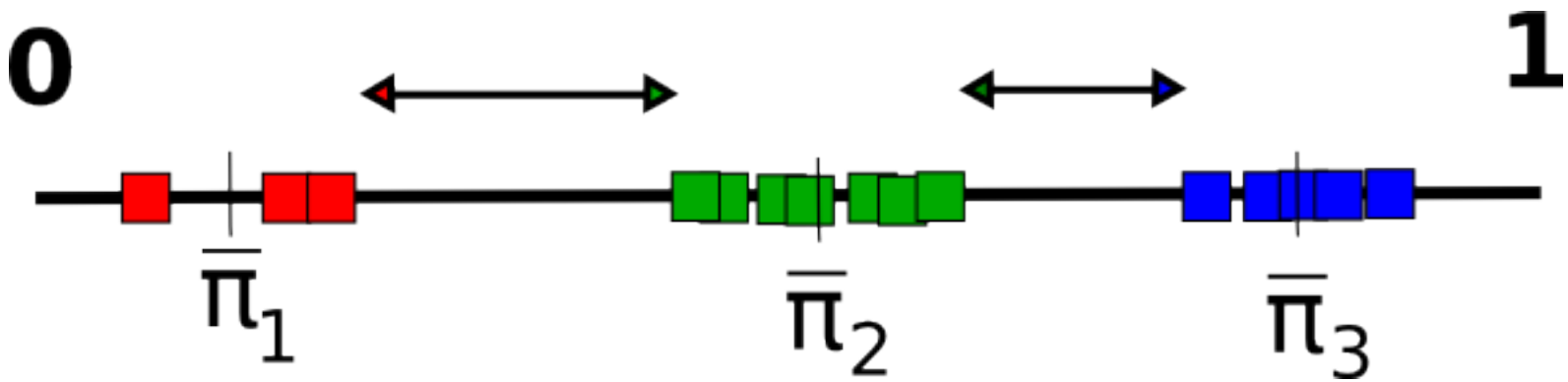


FIGURE: Formation de Q groupes de points séparés par les $Q-1$ plus grands écarts

Algorithme consistant des plus grands écarts (PGE)

- Ordonner la suite $(T_i)_{1 \leq i \leq n}$: $T_{(1)} \leq \cdots \leq T_{(n)}$
- Calculer les écarts $(T_{(i+1)} - T_{(i)})_{1 \leq i \leq n-1}$

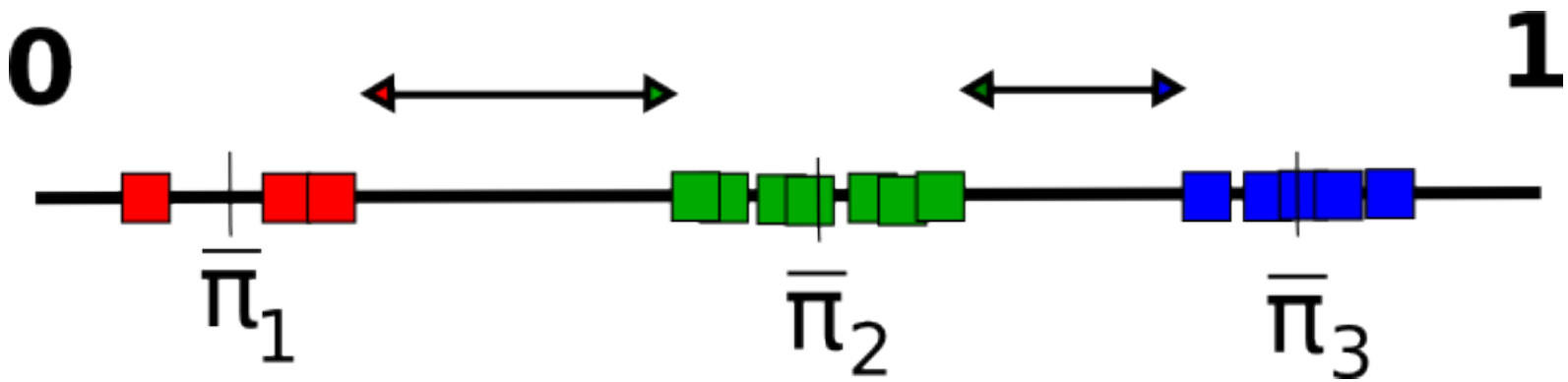


FIGURE: Formation de Q groupes de points séparés par les $Q-1$ plus grands écarts

Algorithme consistant des plus grands écarts (PGE)

- Ordonner la suite $(T_i)_{1 \leq i \leq n}$: $T_{(1)} \leq \cdots \leq T_{(n)}$
- Calculer les écarts $(T_{(i+1)} - T_{(i)})_{1 \leq i \leq n-1}$
- Trouver les $Q-1$ plus grands écarts de sorte à former Q groupes de points.

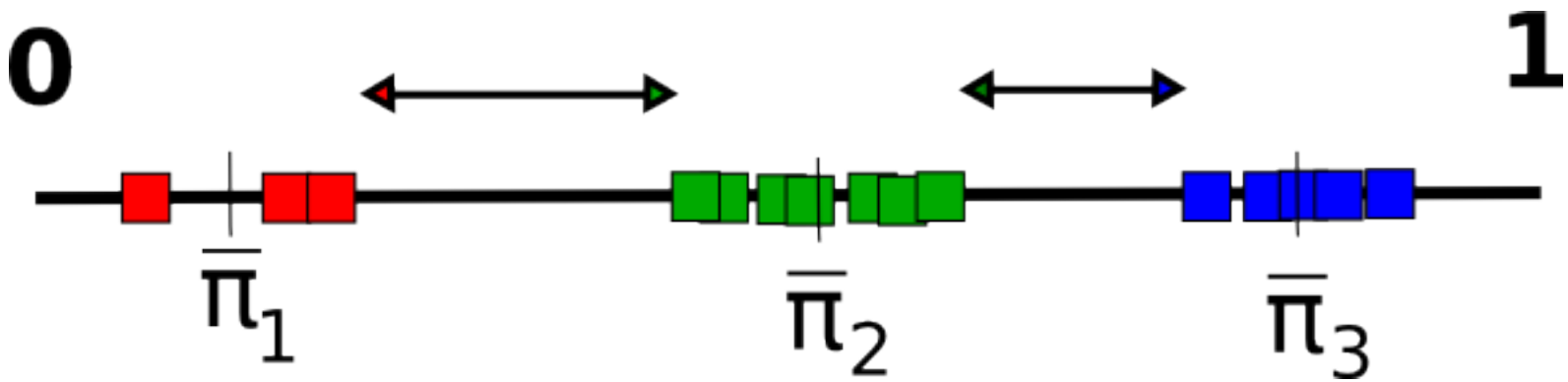


FIGURE: Formation de Q groupes de points séparés par les $Q-1$ plus grands écarts

Hypothèse et définitions préliminaires

Hypothèse

On suppose désormais que pour tout $q \neq r$, $\overline{\pi}_q \neq \overline{\pi}_r$.

Hypothèse et définitions préliminaires

Hypothèse

On suppose désormais que pour tout $q \neq r$, $\overline{\pi}_q \neq \overline{\pi}_r$.

Definition

On appelle distance minimale caractéristique le réel strictement positif δ défini ainsi : $\delta = \min_{q \neq r} |\overline{\pi}_q - \overline{\pi}_r|$

Hypothèse et définitions préliminaires

Hypothèse

On suppose désormais que pour tout $q \neq r$, $\overline{\pi}_q \neq \overline{\pi}_r$.

Definition

On appelle distance minimale caractéristique le réel strictement positif δ défini ainsi : $\delta = \min_{q \neq r} |\overline{\pi}_q - \overline{\pi}_r|$

Definition

On appelle distance maximale intraclasse la variable aléatoire d définie ainsi : $d = \max_{1 \leq q \leq Q} \sup_{i \in \mathcal{C}_q} |T_i - \overline{\pi}_q|$

Hypothèse et définitions préliminaires

Hypothèse

On suppose désormais que pour tout $q \neq r$, $\overline{\pi}_q \neq \overline{\pi}_r$.

Definition

On appelle distance minimale caractéristique le réel strictement positif δ défini ainsi : $\delta = \min_{q \neq r} |\overline{\pi}_q - \overline{\pi}_r|$

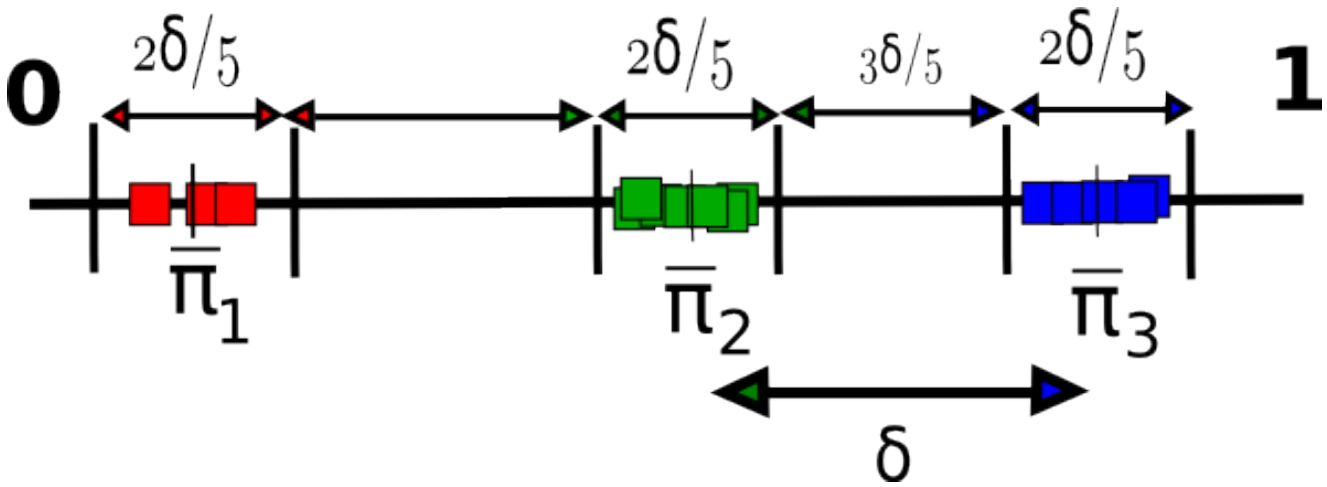
Definition

On appelle distance maximale intraclasse la variable aléatoire d définie ainsi : $d = \max_{1 \leq q \leq Q} \sup_{i \in \mathcal{C}_q} |T_i - \overline{\pi}_q|$

On notera de plus $\alpha_0 = \min_{1 \leq q \leq Q} \alpha_q$ la plus petite des proportions du modèle.

Un heureux événement pour l'algorithme PGE

Si les T_i sont **suffisamment** concentrés autour de leur $\overline{\pi}_q$ respectif, la classification selon les plus grands écarts est sans erreur ; **suffisamment** faisant référence à la distance caractéristique δ.



Si $d \leq \frac{\delta}{5}$, alors le plus grand écart entre des sommets de même classe est strictement plus petit que le plus petit écart entre des sommets de classe distincte, donc les $Q-1$ plus grands écarts classent bien les sommets.

Utilisation d'un argument de concentration

Cet heureux événement est en fait de forte probabilité pour n assez grand grâce à la concentration. Illustration : inégalité de concentration issue de l'inégalité de Hoeffding.

$$\forall t > 0 \; P(|T_i - \overline{\pi}_q| > t | Z_i = q) \leq 2e^{-2nt^2}$$

Théorème

Pour tout $t > 0$, $P(d > t) \leq 2ne^{-2nt^2}$

$$\begin{aligned} P(d > t | Z = z) &= P\left(\bigcup_{1 \leq q \leq Q} \bigcup_{i, z_i = q} \{|T_i - \overline{\pi}_q| > t\} | Z = z \right) \\ &\leq \sum_{1 \leq q \leq Q} \sum_{i, z_i = q} P(|T_i - \overline{\pi}_q| > t | Z = z) \\ &\leq \sum_{1 \leq q \leq Q} \sum_{i, z_i = q} P(|T_i - \overline{\pi}_q| > t | Z_i = q) \\ &\leq 2ne^{-2nt^2} \end{aligned}$$

Conclusion

Théorème

Soit E l'événement "il existe une erreur de classification".

$$P(E) \leq 2ne^{-\frac{2}{25}n\delta^2} + Q(1-\alpha_0)^{n+1}$$

L'algorithme PGE est donc consistant. Notons $(\widehat{\mathcal{C}}_q)_{1\leq q\leq Q}$ la partition de l'ensemble des sommets en classes, produite par l'algorithme.

$$\text{Estimateurs de } \alpha \text{ et } \pi : \widehat{\alpha}_q = \frac{|\widehat{\mathcal{C}}_q|}{n} \text{ et } \widehat{\pi}_{qr} = \frac{1}{|\widehat{\mathcal{C}}_q||\widehat{\mathcal{C}}_r|} \sum_{(i,j)\in\widehat{\mathcal{C}}_q\times\widehat{\mathcal{C}}_r} X_{ij}$$

Théorème

$(\widehat{\alpha},\widehat{\pi})$ est un estimateur consistant de (α,π).

Remarque : Efficacité assurée sous la condition $\delta \gg \sqrt{\frac{\ln n}{n}}$.

Plan de simulation

Deux volets de simulations avec $Q = 3$:

- Evaluation de la qualité de classification. Simulation et stockage des degrés normalisés (T_i) uniquement, dans un modèle où $\delta = 0.02$ jusqu'à $n = 50000$.

Plan de simulation

Deux volets de simulations avec $Q = 3$:

- Evaluation de la qualité de classification. Simulation et stockage des degrés normalisés (T_i) uniquement, dans un modèle où $\delta = 0.02$ jusqu'à $n = 50000$.
- Evaluation de la qualité de l'estimation, nécessitant le stockage de la matrice X. Modèle où $\delta = 0.04$, jusqu'à $n = 16000$.

Plan de simulation

Deux volets de simulations avec $Q = 3$:

- Evaluation de la qualité de classification. Simulation et stockage des degrés normalisés (T_i) uniquement, dans un modèle où $\delta = 0.02$ jusqu'à $n = 50000$.
- Evaluation de la qualité de l'estimation, nécessitant le stockage de la matrice X. Modèle où $\delta = 0.04$, jusqu'à $n = 16000$.

Evaluation de l'algorithme sur 500 graphes tirés. Critères d'erreur :

Plan de simulation

Deux volets de simulations avec $Q = 3$:

- Evaluation de la qualité de classification. Simulation et stockage des degrés normalisés (T_i) uniquement, dans un modèle où $\delta = 0.02$ jusqu'à $n = 50000$.
- Evaluation de la qualité de l'estimation, nécessitant le stockage de la matrice X. Modèle où $\delta = 0.04$, jusqu'à $n = 16000$.

Evaluation de l'algorithme sur 500 graphes tirés. Critères d'erreur :

- Taux de mal classés sur chaque classe prédite.

Plan de simulation

Deux volets de simulations avec $Q = 3$:

- Evaluation de la qualité de classification. Simulation et stockage des degrés normalisés (T_i) uniquement, dans un modèle où $\delta = 0.02$ jusqu'à $n = 50000$.
- Evaluation de la qualité de l'estimation, nécessitant le stockage de la matrice X. Modèle où $\delta = 0.04$, jusqu'à $n = 16000$.

Evaluation de l'algorithme sur 500 graphes tirés. Critères d'erreur :

- Taux de mal classés sur chaque classe prédite.
- Taux de manquants dans la classe prédite par rapport à la vraie classe.

Résultats de classification avec $\delta = 0.02$, $n \leq 50000$

FIGURE: Taux de faux sur les classes prédites

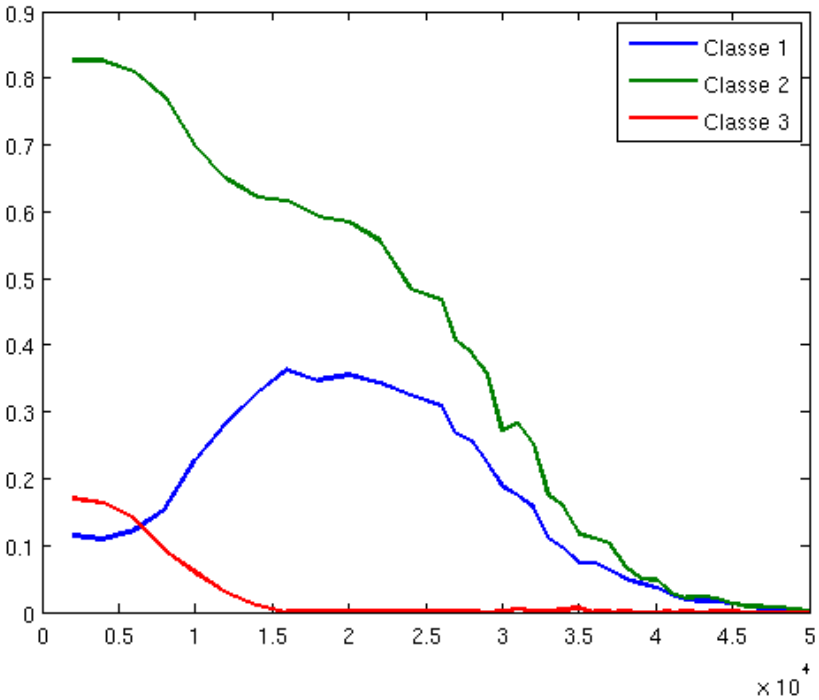
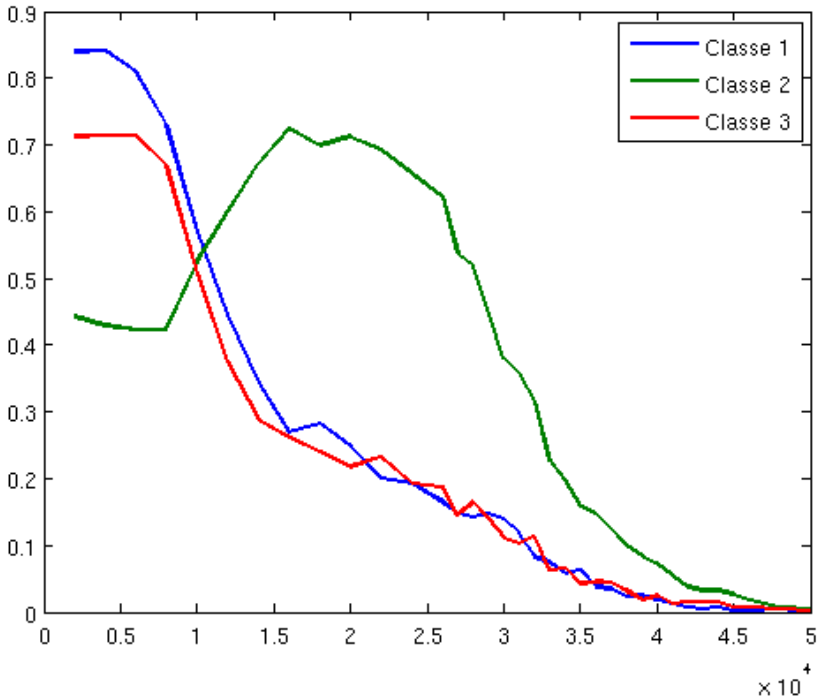


FIGURE: Taux de manquants sur les vraies classes



Résultats de classification avec $\delta = 0.04$, $n \leq 16000$

FIGURE: Taux de faux sur les classes prédites

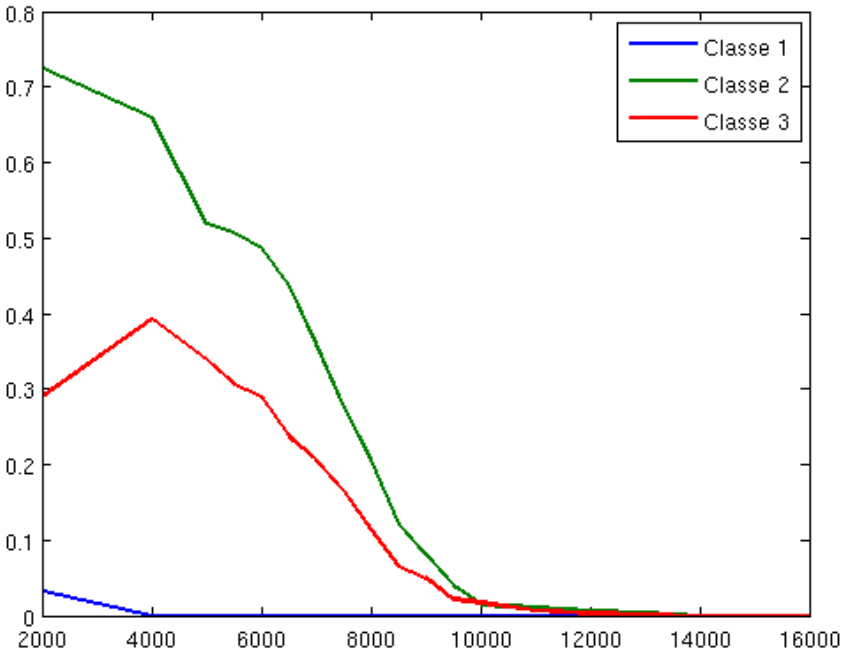
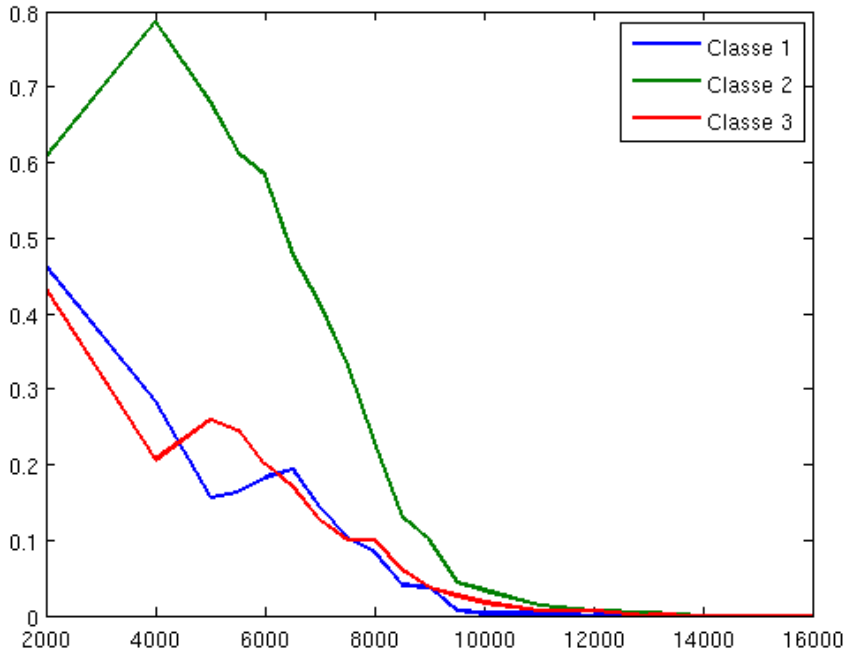
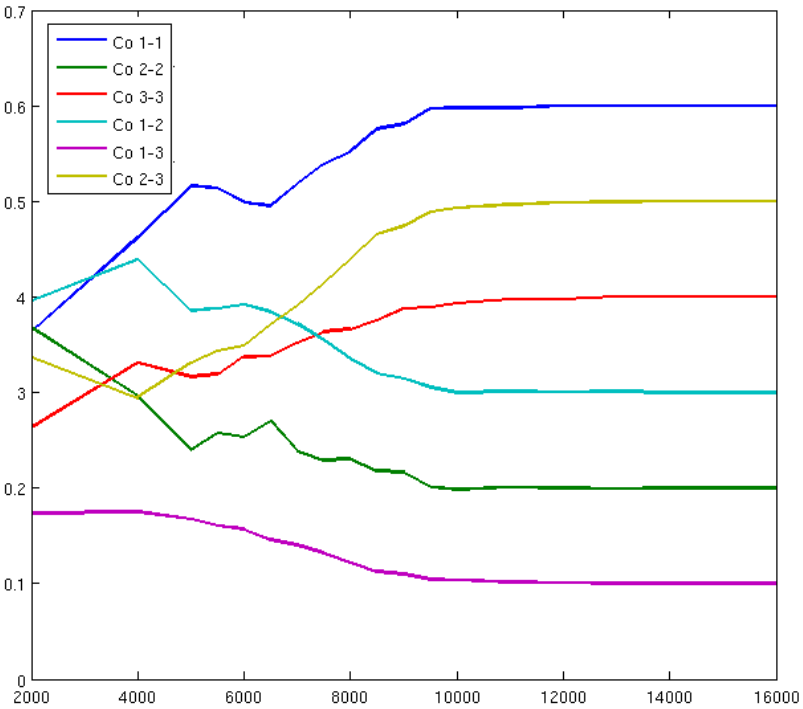


FIGURE: Taux de manquants sur les vraies classes



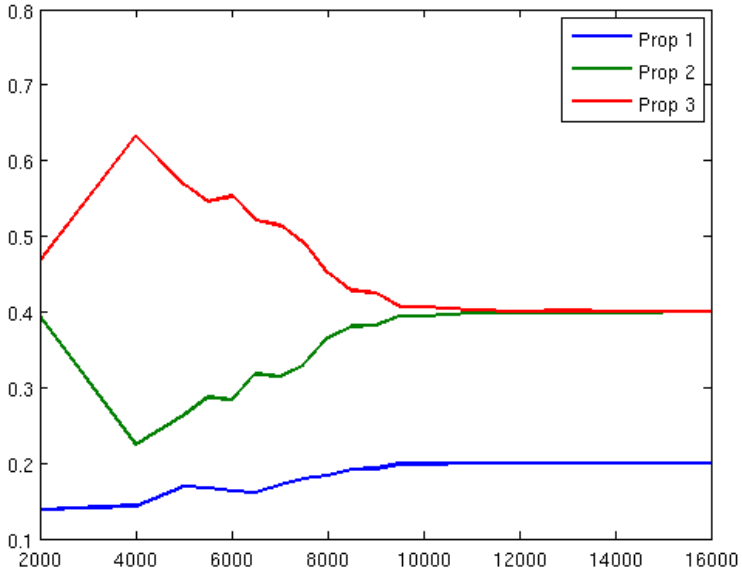
Résultats d'estimation avec $\delta = 0.04$, $n \leq 16000$

FIGURE: Estimation des connectivités



$$\pi = \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.3 & 0.2 & 0.5 \\ 0.1 & 0.5 & 0.4 \end{pmatrix}$$

FIGURE: Estimation des proportions



$$\alpha = (0.2, 0.4, 0.4)$$

Résultats d'estimation avec $\delta = 0.04$, $n \leq 16000$

FIGURE: Ecart-type de l'estimateur des connectivités

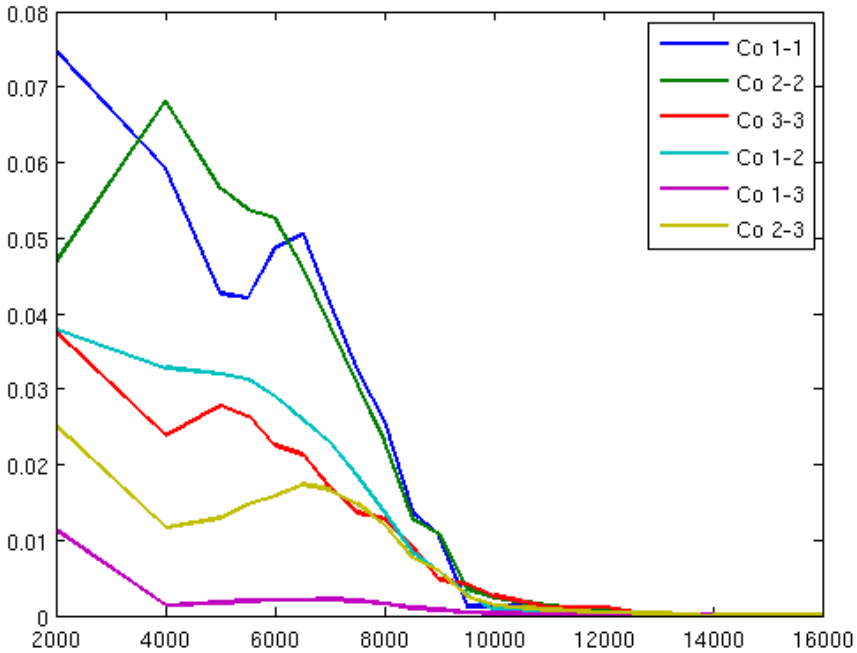
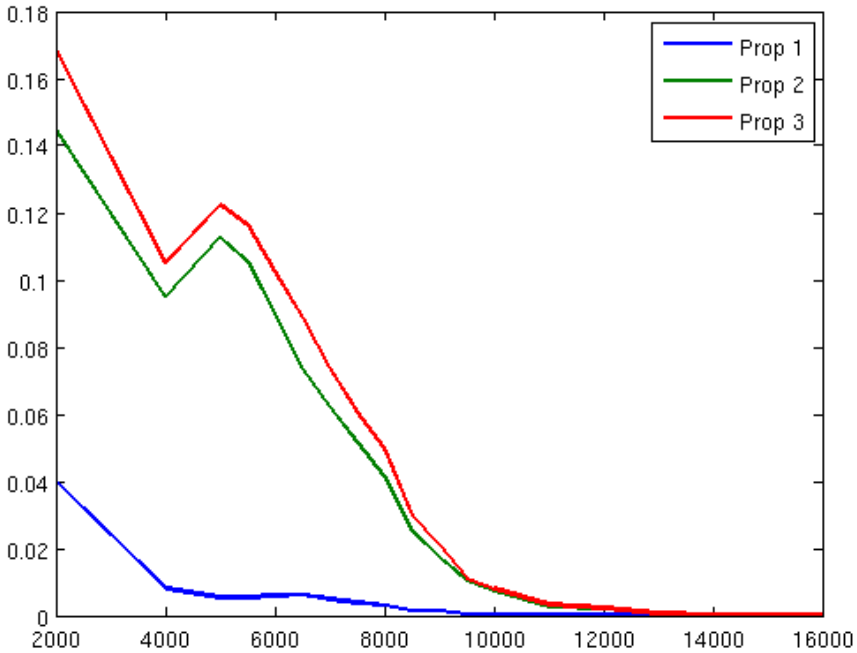


FIGURE: Ecart-type de l'estimateur des proportions



Conclusions et perspectives

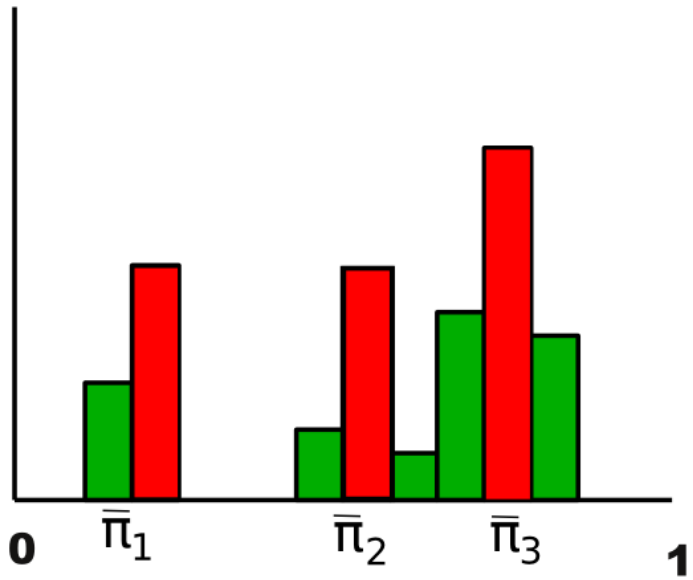
Méthode d'inférence consistante et permettant le traitement de graphes de plusieurs millions de noeuds.

Conclusions et perspectives

Méthode d'inférence consistante et permettant le traitement de graphes de plusieurs millions de noeuds.

- Algorithme à maillage à pas constant (déjà testé) : choisir les Q mailles les plus remplies pour les classes, et jeter les autres sommets pour l'inférence. Meilleur aux n petits, mais perd l'avantage dans les grands graphes.

FIGURE: Histogramme des degrés normalisés

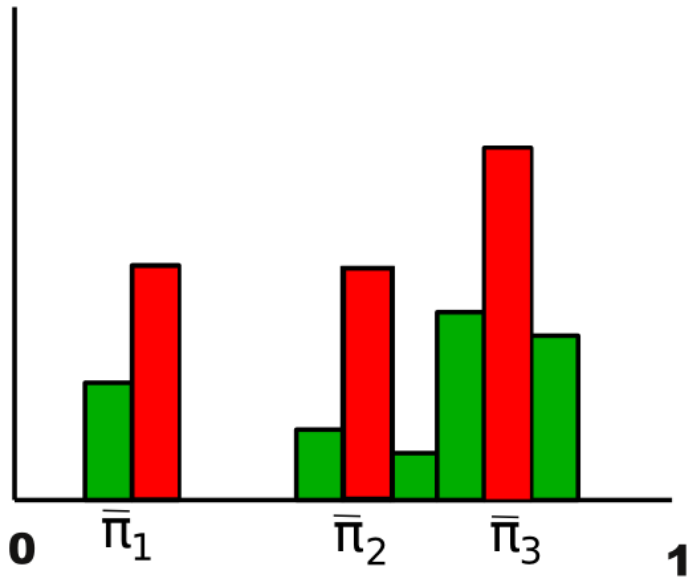


Conclusions et perspectives

Méthode d'inférence consistante et permettant le traitement de graphes de plusieurs millions de noeuds.

- Algorithme à maillage à pas constant (déjà testé) : choisir les Q mailles les plus remplies pour les classes, et jeter les autres sommets pour l'inférence. Meilleur aux n petits, mais perd l'avantage dans les grands graphes.
- Méthode d'histogramme à pas adaptatif

FIGURE: Histogramme des degrés normalisés

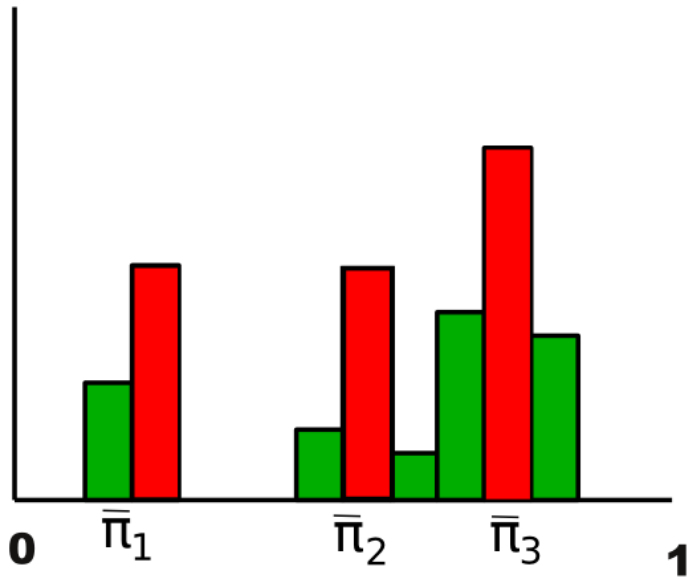


Conclusions et perspectives

Méthode d'inférence consistante et permettant le traitement de graphes de plusieurs millions de noeuds.

- Algorithme à maillage à pas constant (déjà testé) : choisir les Q mailles les plus remplies pour les classes, et jeter les autres sommets pour l'inférence. Meilleur aux n petits, mais perd l'avantage dans les grands graphes.
- Méthode d'histogramme à pas adaptatif
- Utilisation sur des données réelles

FIGURE: Histogramme des degrés normalisés



Merci de votre attention !

Algorithme de maillage à pas constant

FIGURE: Taux de faux sur les classes prédites

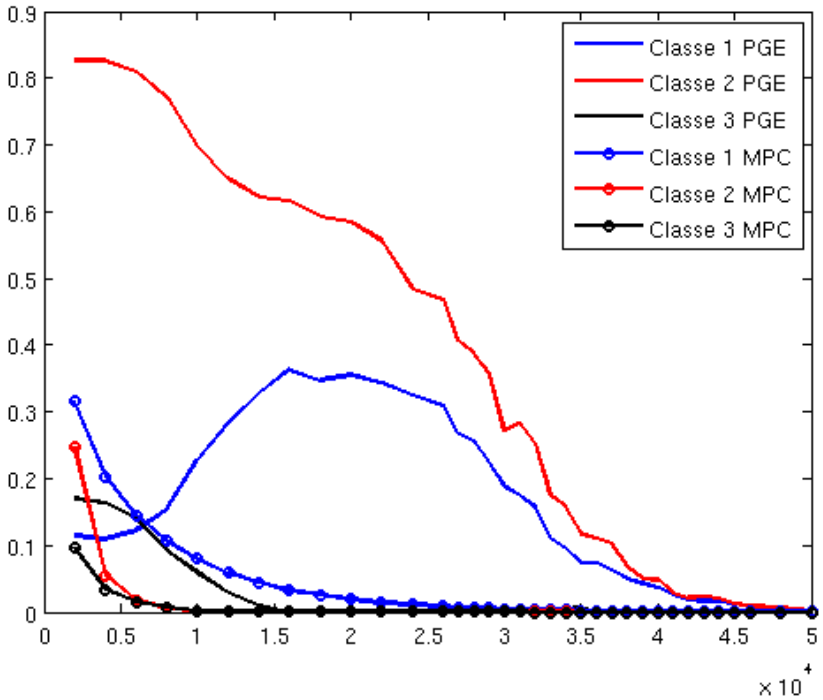


FIGURE: Taux de manquants sur les vraies classes

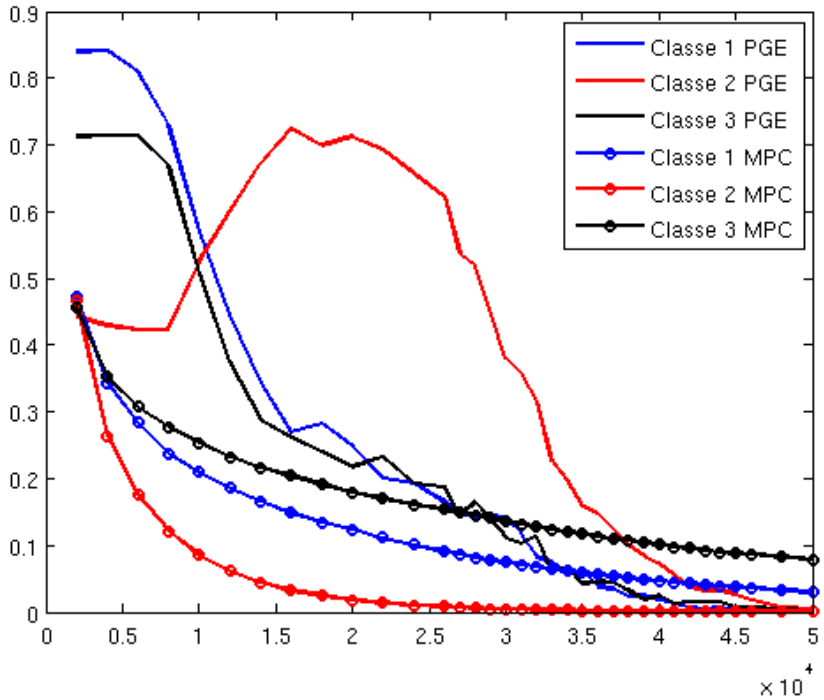


FIGURE: Estimation des proportions