Séminaire de Modélisation Statistiques

Projet d'analyse statistique des reseaux : une étude du scandale Enron

Olivier Supplisson et Nicolas Toussaint

5 avril 2018

Résumé

Ce travail propose une implémentation en `Python` (estimation des paramètres du modèle, clustering des nœuds et détermination du nombre de clusters) de la methode d'inférence variationelle Bayesienne appliquée au Stochastic Block Model, développée par Latouche et al. [2012]. L'utilisation du module `Cython` permet d'obtenir des résultats près de 300 fois plus rapidement qu'une implémentation sous `Python` classique. Le code est donc presque aussi performant que le module `Mixer`, développé par C. Ambroise, G. Grasseau, M. Hoebeke, P. Latouche, V. Miele et F. Picard sur `R`. L'algorithme ainsi implementé est utilisé pour étudier le scandale Enron. Le code attaché à ce rapport est disponible sur le GitHub du projet, au lien suivant : `https://github.com/Nicotous1/Enron`

Mots-clés : Reseaux, Algorithme EM, VBEM, Approximations variationnelles, Cadre Bayésien, Scandale Enron.

ENSAE ParisTech

Année Universitaire 2017 - 2018

Introduction

L'analyse des réseaux sert à modéliser les interactions entre différents agents constitutifs d'un système donné. Cette approche est utilisée dans de nombreux champs scientifiques, des sciences sociales comme expérimentales. La nature des agents et du système sont contingents au domaine d'étude. Par exemple, en génétique, les réseaux de co-expression de gènes sont des sujets d'étude fréquents. En sciences sociales, on pourrait citer, par exemple, l'étude des liens amicaux au sein d'une population. Ces réseaux sont canoniquement représentés par des graphes composés de noeuds et d'arrêtes, les premiers représentant les individus et les seconds les relations entre ces individus. Ces relations peuvent être dirigées (la relation va d'un individu vers un autre individu) ou non (il y a alors réciprocité de l'interaction).

La littérature sur les réseaux étant extrêmement vaste, ce travail se concentre sur l'approche développée par Nowicki and Snijder [2001] Cette approche fait l'hypothèse que les liens unissants les individus dépendent de variables non observées. Ces dernières sont alors introduites dans le modèle comme des variables « latentes ». Cette modélisation, prenant son essor dans les années 2000, s'inscrit dans le cadre plus global du clustering à modèle formel. Contrairement aux algorithmes de clustering hiérarchiques ou aux algorithhmes de type K-Means, ces modèles permettent de réaliser de l'inférence paramétrique. Ces modèles peuvent traiter les réseaux en cherchant à mettre en évidence des phénomènes de « cliques », « d'attachement inverse » (« *disassortative mixing* ») ou encore des structures hétérogènes.

Dans cette lignée des modèles de clustering formels, le *Stochastic Block Model* (SBM) proposé par Nowicki and Snijder [2001], se basant sur des travaux antérieurs de Holland et al. [1983] et Wang and Wong [1987], eux même généralisant dans un cadre stochastique l'approche de White et al. [1976], a un caractère plus large que les modèles « canoniques » développés dans les années 1980. En effet, ces derniers ont tendance à ne regrouper les individus que sous forme de « cliques » ; au contraire, le SBM permet de regrouper les individus en fonction d'un comportement de connection similaire, cela incluant les comportements communautaires. La contribution de Nowicki and Snijder [2001] peut être caracterisée de séminale dans la mesure où elle sert de point de départ à une importante littérature s'intéressant aussi bien à l'identifiabilité et à la méthode d'optimisation qu'à la selection de modèle dans des cadres formels s'inspirant de celui du SBM.

Ainsi, les travaux de Allman et al. [2009] et Allman et al. [2011], par exemple, concluent à l'identifiabilité du modèle SBM, pondéré ou non, prenant en compte les directions ou pas. Celisse et al. [2012] montrent que c'est également le cas pour des modèles SBM dirigé et non pondéré. Enfin, Latouche et al. [2011] étendent le résultat de l'identifiabilité au modèle SBM prenant en compte la possibilité qu'un même noeud

appartienne à différents sous-groupes (on parle alors du Modèle à Blocs Stochastiques Chevauchants).

La branche s'intéressant à la résolution du modèle prend comme point de départ l'impossibilité de ce dernier à être résolu en utilisant l'algorithme EM, classiquement utilisé pour contourner les problèmes de variables latentes. Plusieurs techniques différentes ont été retenues pour résoudre ce problème, allant de l'échantillonage de Gibbs (Nowicki and Snijder [2001] (qui ne peut pas être utilisé pour un réseau à plus de 200 noeuds) à l'utilisation de méthodes variationnelles pour l'estimation (inférence variationnelle fréquentiste (Daudin et al. [2008], Picard et al. [2009]) ou inférence variationnelle bayésienne (Latouche et al. [2012]). Celisse et al. [2012] montrent également que la méthode d'inférence variationnelle donne des résultats convergents dans le cadre du SBM. Outre la nécessité d'utiliser des méthodes de maximisation non conventionnelles, l'impossibilité d'écrire la log-vraisemblance du modèle rend impossible le calcul du BIC ou de l'AIC. Une méthode supplétive a été proposée par Daudin et al. [2008] dans un cadre d'inférence variationnelle fréquentiste. Elle a été adaptée dans Latouche et al. [2011] à un cadre bayésien.

De manière générale, concernant les extensions du modèle SBM, on peut par exemple citer l'ajout de la prise en compte de variables explicatives dans le modèle par Zanghi et al. [2010], la prise en compte d'appartenance à plusieurs groupes (Latouche et al. [2011], Airoldi et al. [2008]), déjà évoquée, ou encore l'application de cette approche à des données sous formes de tables de contingence.

Ce travail se concentre sur la contribution de Latouche et al. [2012]. Il s'accompagne d'un notebook `Python` permettant d'utiliser leur algorithme pour étudier n'importe quel jeu de données. Dans un premier temps, nous présentons le cadre général du SBM avec lien binaire. Puis, nous présentons de manière générale l'approche variationnelle fréquentiste puis bayésienne. Enfin, nous présentons la contribution de Latouche et al. [2012], c'est-à-dire l'application de l'inférence variationnelle bayésienne au SBM, que nous utilisons ensuite pour étudier le scandale Enron.

1. Le Stochastic Block Model avec lien binaire

Dans cette section, nous présentons le modèle SBM avec lien binaire (*i.e.* absence ou non de relation entre les différents noeuds d'un graph) tel qu'il apparait dans Latouche et al. [2012]. Dans ce cadre, le modèle SBM prend pour entrée deux éléments : une matrice décrivant l'ensemble des liens qui unissent les individus entre eux et une suite de vecteurs indiquant pour chaque individu la classe à laquelle il appartient. On supose qu'il existe Q classes.

Soit une matrice X de dimension $N \times N$ décrivant l'existence ou non d'un lien entre l'individu i et l'individu j. La terme x_{ij} prend alors la valeur 1 si un lien existe et 0 sinon

et le terme x_{ji} prend la valeur 1 si la relation est réciproque et 0 sinon. Lorsque cette matrice est symétrique, les liens considérés sont des liens non directionnels.

Soit une matrice $Z = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{pmatrix}$ avec $(z_i = \begin{pmatrix} z_{i1} \\ z_{i2} \\ \vdots \\ z_{iQ} \end{pmatrix})_{i\in\{1,\cdots,N\}}$ un vecteur de taille $1 \times Q$ associé à l'individu i indiquant la classe q à laquelle l'individu appartient. On a alors $Z_{iq} = 1$ et $Z_{ij} = 0$, $j \in \{1, \cdots, Q\}\backslash\{q\}$. On suppose par ailleurs, que $\forall(i,j) \in \{1,\cdots,N\} \times \{1,\cdots,N\}, i \neq j$, $Z_i \perp\!\!\!\perp Z_j$ et $\forall i \in \{1,\cdots,N\}$:

$Z_i \sim \mathcal{M}(1,\alpha)$, $\alpha = \left(\alpha_1,\cdots,\alpha_Q\right)$ avec $(\alpha_j)_{j\in\{1\cdots Q\}}$ la probabilité d'appartenir à la classe j.

La probabilité que l'individu i et j se connectent, sachant qu'ils appartiennent respectivement à la classe q et l (donc $\Pr\left(X_{ij} = 1|\{Z_{iq}Z_{jl} = 1\right)$) est notée π_{ql}. Comme le lien est bivarié, on a alors :

$$X_{ij}|\{Z_{iq}Z_{jl}\} \sim \mathcal{B}(\pi_{ql}) \tag{1}$$

Tous les liens unissants les individus sont donc par hypothèse indépendants les uns des autres et on peut construire une matrice $N \times N$ qui contient l'ensemble des probabilité que chaque individu se connecte aux autres.

2. Le principe de l'approche variationnelle

Comme dit auparavant, la problématique principale du SBM est l'impossibilité de maximiser la log-vraisemblance en utilisant l'outil canonique permettant de contourner le problème de l'inobservabilité de la variable latente (il s'agit ici de la variable Z) : l'algorithme EM. Comme évoqué dans l'introduction, quelques techniques ont été proposées pour contourner ce problème. Nous présentons ici le principe de l'inférence variationnelle. Elle sera d'abord présentée dans un cadre fréquentiste avant d'être étendue au cadre bayésien.

2.1. Principe général de l'inférence variationnelle

Supposons que l'objectif soit de maximiser la log-vraisemblance, noté $\log\left(p\left(X|\theta\right)\right)$, mais que cela soit impossible, pour une raison ou une autre. L'idée derrière l'approche variationnelle est de maximiser non pas le vrai programme, cela étant impossible, mais un programme alternatif retenant une densité connue et maximisable tout en corrigeant les paramètre successivement obtenus afin de rendre le plus proche possible la distribution de substitution de la vraie distribution.

2.2. Appréciation de la distance entre deux distributions et borne inférieure de la log-vraisemblance

Si on considère que la distance entre deux distributions s'apprécie en fonction de la Divergence de Kullback-Leibler, on a alors, en notant $P(.)$ la distribution de $Z|X$, $p(.)$ la densité associée, $Q(.,\theta)$ une distribution paramétrique à *posterior* simple et dépendant de θ pour $Z|X$ et $q(.)$ la densité associée. Le degré de divergence entre les deux distribution est alors donné par :

$$\begin{aligned}
\mathrm{KL}\left(Q\left(Z|X,\theta\right)||P\left(Z|X\right)\right) &= \sum_{z}\left[q\left(z,\theta\right)\log\left(\frac{q\left(z|x,\theta\right)}{p\left(z|x\right)}\right)\right] \\
&= \sum_{z}\left[q\left(z,\theta\right)\log\left(\frac{q\left(z|x,\theta\right)p\left(x\right)}{p\left(z,x\right)}\right)\right] \\
&= \sum_{z}\left[q\left(z,\theta\right)\log\left(\frac{q\left(z|x,\theta\right)}{p\left(z,x\right)}\right)\right] + \log\left(p\left(x\right)\right) \\
&= \sum_{z}\left[q\left(z,\theta\right)\log\left(\frac{q\left(z|x,\theta\right)}{p\left(x|z\right)p\left(z\right)}\right)\right] + \log\left(p\left(x\right)\right) \\
&= \mathbb{E}_{z}\left[\log\left(\frac{q\left(z|x,\theta\right)}{p\left(x|z\right)p\left(z\right)}\right)\right] + \log\left(p\left(x\right)\right) \\
&\Rightarrow \mathrm{KL}\left(Q\left(Z|X,\theta\right)||P\left(Z|X\right)\right) = \log\left(p\left(x\right)\right) - \underline{\mathcal{B}}\left(q,\theta\right)
\end{aligned}$$

Avec $\underline{\mathcal{B}}\left(q,\theta\right) = \mathbb{E}_{z}\left[\log\left(p\left(z\right)\right)\log\left(\frac{p(x|z)}{q(z|x,\theta)}\right)\right]$. Comme $\mathrm{KL}(.)$ est positif ou nul, il suit que $\log\left(p\left(x\right)\right) \geq \underline{\mathcal{B}}\left(q,\theta\right) = \mathbb{E}_{z}\left[\log\left(p\left(z\right)\right)\log\left(\frac{p(x|z)}{q(z|x,\theta)}\right)\right]$. $\underline{\mathcal{B}}$ est donc la borne inférieure de la vraisemblance de la variable X. C'est ce résultat qui est utilisé dans l'inférence variationnelle.

2.3. Les étapes de l'inférence variationnelle

L'inférence variationnelle est, comme l'algorithme EM, composée de deux étapes :

Etape 1. « E *step* » La première étape de l'inférence variationnelle consiste à chercher la distribution Q la plus proche possible de la distribution P à θ donné. Pour cela on cherche à résoudre le programme suivant :

$$\max_{q}\underline{\mathcal{B}}\left(q,\theta\right) \Longleftrightarrow \min_{q}\mathrm{KL}\left(Q\left(Z|X,\theta\right)||P\left(Z|X\right)\right) \tag{2}$$

Il convient de préciser que pour que cette étape soit faisable, il faut se restreindre aux distributions Q qui sont « conciliantes » (c'est-à-dire factorisables). On obtient alors la distribution $Q^{*}()$ et $q^{*}()$, sa densité.

Etape 2. « M *step* » Lors de la seconde étape, on résout :

$$\max_{\theta} \underline{B}\left(q^{*}, \theta\right) \Longleftrightarrow \min_{\theta} \mathrm{KL}\left(Q^{*}\left(Z|X, \theta\right)||P\left(Z|X\right)\right) \tag{3}$$

2.4. L'inférence variationnelle dans le cadre bayésien

Le cadre fréquentiste correspond au cas spécial du cadre bayésien avec *prior* impropre égal à 1 pour chaque paramètre. On présente ici l'inférence variationnelle dans le cadre plus large à *prior* quelconque, noté $q_{\theta}\left(\theta\right)$.

En utilisant l'approximation $q\left(z, \theta\right) \sim q\left(z\right) q_{\theta}\left(\theta\right)$, on obtient :

$$\underline{B}\left(q\left(z\right), p\left(x\right), \pi\left(\theta\right)\right) = \sum_{\mathrm{z}} \int q\left(z\right) q_{\theta}\left(\theta\right) \log\left[\frac{p\left(x, z, \theta\right)}{q\left(z\right) q_{\theta}\left(\theta\right)}\right] \mathrm{d}\theta \tag{4}$$

3. L'inférence variationnelle bayésienne dans le cadre du SBM avec lien binaire - Latouche et al. [2012]

Dans le cadre développé dans la section 1., $\theta = (\pi, \alpha)$. Le cadre choisi par Latouche et al. [2012] retient par ailleurs l'hypothèse que :

$$q\left(Z, \alpha, \pi\right) = q_{\alpha}\left(\alpha\right) q_{\pi}\left(\pi\right) \prod_{\mathrm{i}=1}^{\mathrm{N}} q\left(Z_{\mathrm{i}}\right) \tag{5}$$

L'algorithme permettant l'inférence variationnelle est alors donné par :

Initialisation : On fournit à l'algorithme $\left\{Q, \left(n_{j}^{(0)}\right)_{\mathrm{j}\in\{1,\cdots,\mathrm{Q}\}}, \left(\tau_{iq}^{(0)}\right)_{(\mathrm{i,q})\in\{1,\cdots,\mathrm{N}\}\times\{1,\cdots,\mathrm{Q}\}}, \left(\eta_{\mathrm{ql}}^{(0)}, \zeta_{\mathrm{ql}}^{(0)}\right)_{(\mathrm{q,l})\in\{1,\cdots,\mathrm{Q}\}^{2},\mathrm{q}\neq\mathrm{l}}, \epsilon\right\}$

Etape k+1 - « E *step* » :

$$\tau_{\mathrm{iq}}^{(\mathrm{k}+1)} \leftarrow \exp\left\{\psi\left(n_{\mathrm{q}}^{(\mathrm{k})}\right) - \psi\left(\sum_{\mathrm{l}=1}^{\mathrm{Q}} n_{\mathrm{l}}^{(\mathrm{k})}\right)\right\} \prod_{\mathrm{j}=1,\mathrm{ji}}^{\mathrm{N}} \prod_{\mathrm{l}=1}^{\mathrm{Q}} \exp\left\{\tau_{\mathrm{jl}}^{(\mathrm{k})}\left[\psi\left(\zeta_{\mathrm{ql}}^{(\mathrm{k})}\right) - \psi\left(\eta_{\mathrm{ql}}^{(\mathrm{k})} + \zeta_{\mathrm{ql}}^{(\mathrm{k})}\right) + X_{\mathrm{ij}}\left(\psi\left(\eta_{\mathrm{ql}}^{(\mathrm{k})}\right) - \psi\left(\zeta_{\mathrm{ql}}^{(\mathrm{k})}\right)\right)\right]\right\}$$

Avec $\psi(.)$ la fonction digamma.

Etape k+1 - « M *step* » :

$$\begin{cases} \forall q \in \{1, \cdots, Q\} : n_{\mathrm{q}}^{(\mathrm{k}+1)} = n_{\mathrm{q}}^{(0)} + \sum_{\mathrm{i}=1}^{\mathrm{N}} \tau_{\mathrm{iq}}^{(\mathrm{k})} \\ \forall (q, l) \in \{1, \cdots, Q\}^{2}, q \neq l : \eta_{\mathrm{ql}}^{(\mathrm{k}+1)} = \eta_{\mathrm{ql}}^{(0)} + \sum_{(\mathrm{i,j})\in\{1,\cdots,\mathrm{N}\}^{2},\mathrm{j}\neq\mathrm{i}} X_{\mathrm{ij}} \tau_{\mathrm{iq}}^{(\mathrm{k})} \tau_{\mathrm{jl}}^{(\mathrm{k})} \\ \forall q \in \{1, \cdots, Q\} : \eta_{\mathrm{qq}}^{(\mathrm{k}+1)} = \eta_{\mathrm{ql}}^{(0)} + \sum_{(\mathrm{i,j})\in\{1,\cdots,\mathrm{N}\}^{2},\mathrm{i}<\mathrm{j}} X_{\mathrm{ij}} \tau_{\mathrm{iq}}^{(\mathrm{k})} \tau_{\mathrm{jq}}^{(\mathrm{k})} \\ \forall (q, l) \in \{1, \cdots, Q\}^{2}, q \neq l : \zeta_{\mathrm{ql}}^{(\mathrm{k}+1)} = \zeta_{\mathrm{ql}}^{(0)} + \sum_{(\mathrm{i,j})\in\{1,\cdots,\mathrm{N}\}^{2},\mathrm{j}\neq\mathrm{i}} (1 - X_{\mathrm{ij}}) \tau_{\mathrm{iq}}^{(\mathrm{k})} \tau_{\mathrm{jl}}^{(\mathrm{k})} \\ \forall q \in \{1, \cdots, Q\} : \zeta_{\mathrm{qq}}^{(\mathrm{k}+1)} = \zeta_{\mathrm{ql}}^{(0)} + \sum_{(\mathrm{i,j})\in\{1,\cdots,\mathrm{N}\}^{2},\mathrm{i}<\mathrm{j}} (1 - X_{\mathrm{ij}}) \tau_{\mathrm{iq}}^{(\mathrm{k})} \tau_{\mathrm{jq}}^{(\mathrm{k})} \\ \forall (i, q) \in \{1, \cdots, N\} \times \{1, \cdots, Q\},\ \hat{Z}_{\mathrm{iq}}^{(\mathrm{k}+1)} = \mathbb{1}_{\tau_{\mathrm{iq}}^{(\mathrm{k}+1)} > \tau_{\mathrm{il}}^{(\mathrm{k}+1)},\ \forall\, \mathrm{l}\in\{1,\cdots,Q\}, l\neq q} \end{cases} \tag{6}$$

Critère d'arrêt : L'algorithme s'arrête à l'itération $k+1$ si :

$$\sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{q}=1}^{\mathrm{Q}} |\tau_{\mathrm{iq}}^{(\mathrm{k})} - \tau_{\mathrm{iq}}^{(\mathrm{k}+1)}| \leq \epsilon \tag{7}$$

On a alors :

$$\begin{cases} \forall (i,q) \in \{1,\cdots,N\} \times \{1,\cdots,Q\},\ \lim_{\mathrm{k}\to+\infty} \tau_{\mathrm{iq}}^{(\mathrm{k})} = \pi_{iq} \\ \forall q \in \{1,\cdots,Q\},\ \lim_{\mathrm{k}\to+\infty} n_{\mathrm{q}}^{(\mathrm{k})} = \alpha_q \\ \forall (i,q) \in \{1,\cdots,N\} \times \{1,\cdots,Q\}, \lim_{\mathrm{k}\to+\infty} \hat{Z}_{\mathrm{iq}}^{(\mathrm{k})} = Z_{\mathrm{iq}} \end{cases} \tag{8}$$

Nous allons à présent mener l'analyse du scandal Enron en utilisant cet algorithme.

4. Analyse du scandale Enron

Le scandale Enron est l'un des plus gros scandale comptable et financier de ce début de siècle. La société Enron était issue de la fusion, en 1985, de deux companies intervenant sur le marché du gaz : la Houston Natural Gas et la Internorth of Omaha. D'abord simple fournisseur d'électricité, la société s'éloigne ensuite de sa chaine de valeur initiale en réalisant des opérations spéculatives sur le marché de l'énergie, puis de l'eau, de l'électricité, du papier, etc. Cette stratégie de diversification dans des activités spéculatives, risquées a provoquée des pertes. Ces dernières étant en grande partie, grâce à un montage juridique complexe et des malversations comptables, supprimées des comptes (y compris les comptes consolidés) d'Enron et imputées à un ensemble de sociétés filliales, situées tout autour du monde et notamment dans des paradis fiscaux. Une fois exposées publiquement, notamment suite à l'enquête de la Securities and Exchange Commission (SEC), l'ensemble de ces malversations comptables ont entrainé la chute du cabinet d'audit et d'expertise comptable Andersen Consulting, qui s'occupait alors des comptes de la société, ainsi que le démantellement de cette dernière. Ainsi que la condamnation de nombreux hauts managers, la ruine de nombreux employés et du fond de pension chargé s'occuper de leur retraite et détenant essentiellement des actions d'Enron.

Il s'agira dans la suite de ce travail d'étudier les interactions, ici matérialisées uniquement par le fait de recevoir un mail, envoyer un mail ou être parmi les destinataires d'un même mail , entre employés de la société en utilisant une base de donnée contenant près de 149 employés et 500 000 emails. Originellement mis en ligne par la Federal Energy Regulatory Commission pendant son enquête, la version utilisée est celle fournie par William W. Cohen et disponible sur `http://www.cs.cmu.edu/ enron/`, nous utiliserons

une version déjà traitée de la base de données [1].

4.1. Précision sur l'individu statistique et le type de lien retenu

La base de données contient près de 10 000 adresses électroniques uniques ayant envoyé ou reçu un email d'un de ces 149 employés. Précisons que cette étude considère comme individu statistique non pas une adresse mail mais un salarié d'Enron pour lequel nous disposons de l'ensemble des emails envoyés et reçus.

Considérer une adresse mail comme individu statistique induirait un problème au niveau de la vraisemblance puisque l'on serait susceptible d'attribuer une absence de liens entre des individus qui en fait aurait communiqué entre eux (cette communication étant cachée à cause du fait que l'intégralité des mails envoyés et reçus par cette adresse mail sont inconnus). Cela induirait alors un biais au niveau des coefficients obtenus. De plus, retenir les 149 salariés d'Enron et non les adresses mails comme individu statistique est l'approche également retenue dans les travaux antérieurs étudiants cette base de données.

Les 149 individus retenus possèdent entre 1 et 4 adresses emails connues (d'après la documentation, il est possible que certaines adresses leur ayant envoyé un message ou ayant reçu un message de leur part soient également une de leur adresse mais nous supposons ce nombre négligeable), ce qui représente au total 297 adresses email soit près de 3 % de l'ensemble des adresses emails. Parmis ces 149 employés, on retrouve les 4 CEO de la période (K. Lay, J. Skilling, J. Lavorato et D. Delainey), des Directors (14), des Traders (13), des Presidents (4), des Vice-Presidents (23), des Managers (14), des Managing Directors (3) et un Avocat du services juridiques et des employés (41). Le statut de 32 employés est inconnu.

4.2. Chronologie du scandale Enron

La chronologie du scandale Enron peut se retrouver sur https ://www.nytimes.com/2006/01/18/business/worldbusiness/timeline-a-chronology-of-enron-corp.html. Les dates particulièrement importantes sont :

— 16 octobre 2001 : Enron annonce une perte trimestrielle de 638 millions de dollars et une réduction du 1,2 milliards de dollars en fonds propres. Début du scandale.
— 22 Octore 2001 : Enron révèle que la *Security Exchange Commission* (SEC) commence une enquête sur la société à propos d'un possible conflit d'intérêt.
— 2 décembre 2001 : La société fait faillite.

1. Disponible sur `http://www.ahschulz.de/enron-email-data/`. Nous avons récupéré le fichier R, qui contient lui-même 4 data frames que nous avons exportés en .csv puis traité sous Python.

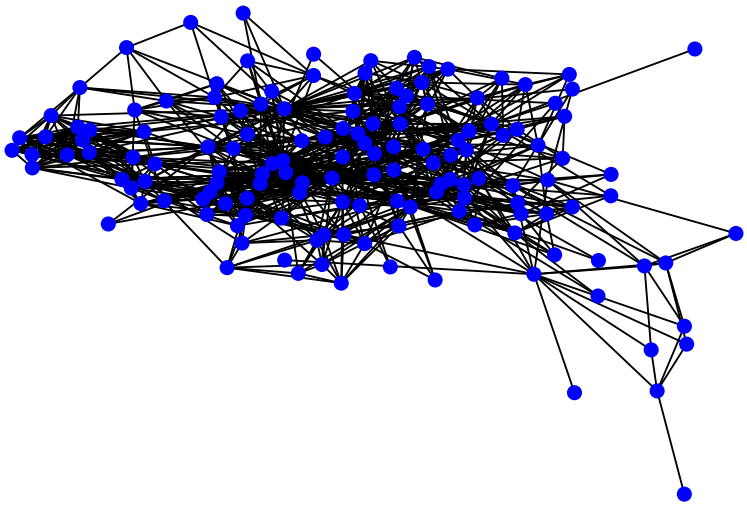
4.3. Construction de la matrice des connexions

Pour chaque employé, on regarde parmi l'ensemble des messages envoyés ceux qui l'ont été avec l'une des adresses emails qui lui sont associées. On récupère l'ensemble des adresses des destinataires de ces emails. Par un procédé similaire, on récupère l'ensemble des adresses mails des individus lui ayant envoyé un mail. On ne conserve que les emails qui correspondent à l'une des adresse des employés pour lesquels nous avons des informations sur l'ensemble de leur correspondance numérique. On obtient une matrice 149×149 avec en colonnes et en lignes la liste des employés. Pour l'individu i, on assigne la valeur 1 à l'élément $(i,j)_{(\mathrm{i,j})\in\{1,\cdots,149\}^2}$ lorsque l'employé j a envoyé un email à l'individu i, a reçu un email de sa part ou si les individus i et j étaient tous les deux parmi les destinaitaires d'un email écrit par une personne z, que cette dernière soit ou non dans la base de données finale. On assigne toujours 1 à l'élément $(i,i)_{\mathrm{i}\in\{1,\cdots,149\}}$. Nous considérons donc uniquement des liens non directionnels.

4.4. Analyse du réseau sur l'ensemble de la période

Lorsqu'on représente graphiquement cette matrice initiale, on voit qu'il existe deux individus pour lesquels il n'existe aucune connexion. Sur l'ensemble de la période, on a 2005 liens au total. On applique le SvBM au réseau sur l'intégralité de la période.

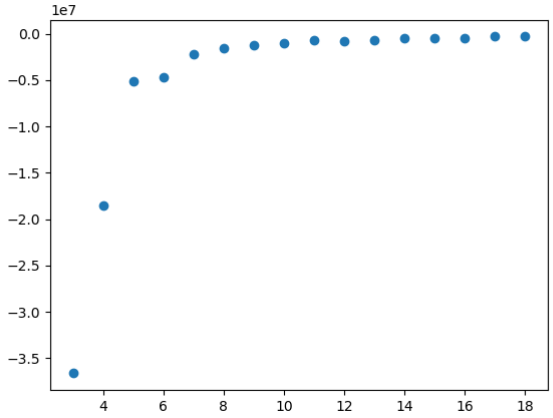
Figure 1 – Représentation graphique de la matrice de connexion (147 individus) sur l'ensemble de la période



Sur l'ensemble de la période, les individus ayant eu le plus de liens avec les autres individus de la base sont, dans l'ordre décroissant, Louise Kitchen (53 liens), President, Kevin Presto (46 liens), Vice-President, Mike Grigsby (38 liens), Manager et enfin Susan

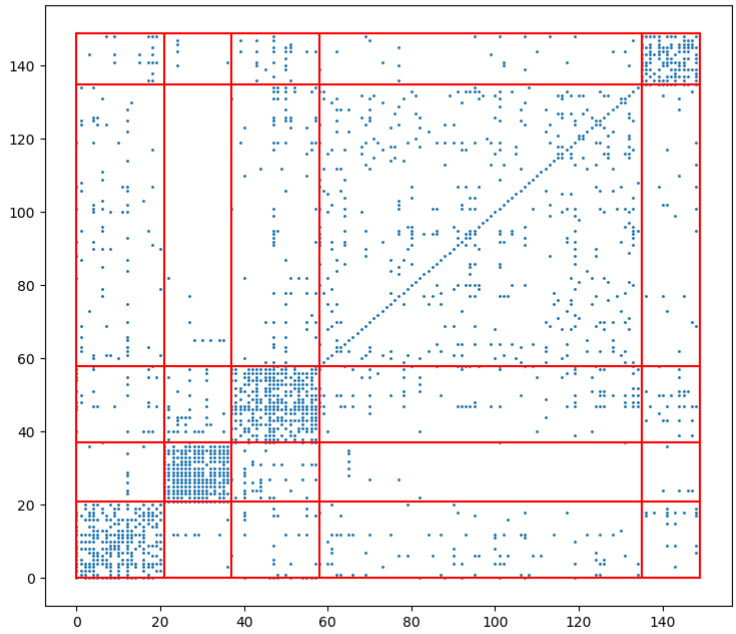
Scott (34 liens), de statut inconnu. Le nombre moyen de lien est 13,6. Donc, en moyenne entre le 21 janvier 2000 et et le 7 février 2002, les 149 individus de notre base se sont connectés au moins une fois à 13,6 autres individus de cette même base.

Figure 2 – Critère Bayésien en fonction du nombre de classe



Afin de conserver une bonne interprétation des résultats tout en obtenant un critère bayésien élevé, on considère 5 classes. Ce nombre a l'avantage d'offrir une bonne interprétabilté tout en offrant un critère bayésien relativement élevé, comme le montre le graphique 2 ci-dessus.

Figure 3 – Matrice adjacente réorganisée par classe



Parmi tous les résultats fournis par notre approche, intéressons-nous tout d'abord à la topologie des groupes latents identifiés. Parmi les 5 classes obtenues, 4 ont une probabilité de connexion intra-classe significativement supérieure à leur probabilité de connexion extra-classe [2]. Ainsi, la classe 2, 3, 4 et 5 correspondent à des communautés différentes avec des individus qui ont une probabilité d'entrer en contact supérieure à la probabilité avec des individus des autres classes. La classe 1 quant à elle contient des individus s'intégrant de manière non structurée dans le réseau.La classe la plus dense du point de vue des échanges est la classe 3 bien que la classe la plus nombreuse soit la classe 1.

Afin d'étudier la dynamique de ce réseau nous allons à présent procéder à une analyse par sous-période. Nous allons en particulier chercher à voir si la structure et l'intensité des liens intraorganisationnels évoluent en fonction de l'avancement de la crise. Pour ce faire nous considérons les 4 événements évoqués dans la section résumant la chronologie du scandale et comparons le réseau durant la semaine qui précède et la semaine qui suit.

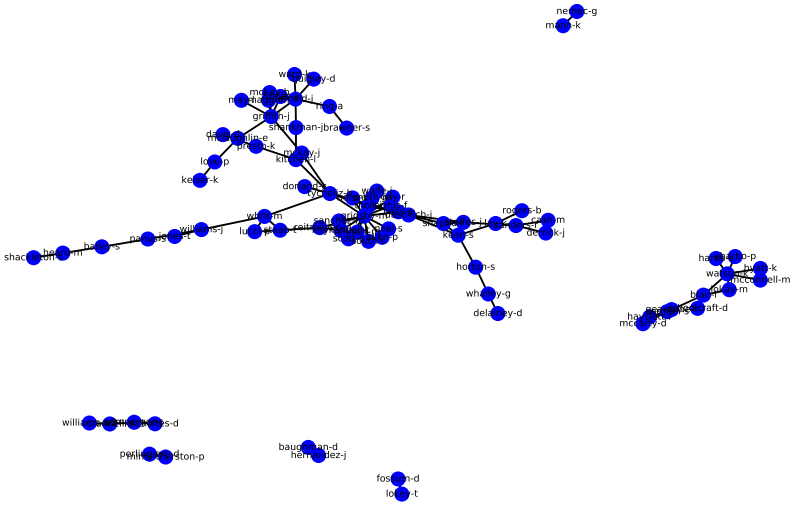
2. Egale à 0,08, 0,59, 0,85, 0,54, 0,57 respectivement pour les classes 1,2,3,4,5

4.5. Analyse du réseau par sous-période

4.5.1. Analyse sur la periode pré/post-annonce de la perte de 638 millions de dollars d'Enron

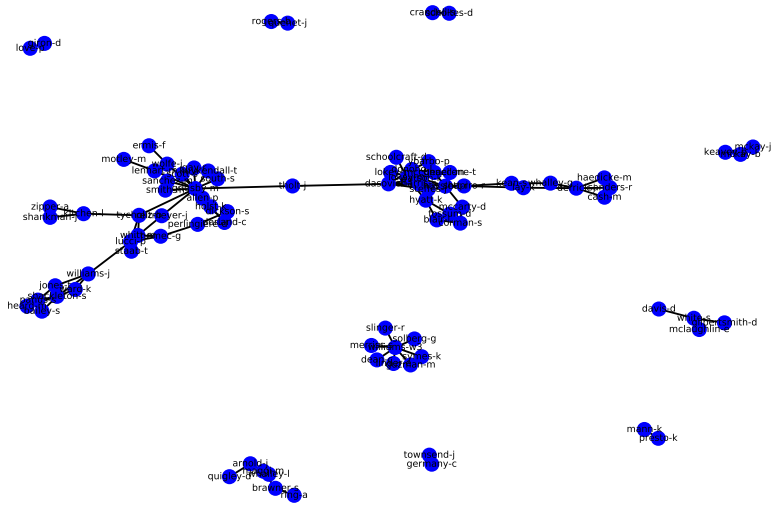
Nous allons à présent étudier le réseau sur la période du 10 au 15 octobre et de la période du 16 au 21 octobre. Comme on peut le voir sur le graphique 4, montrant le réseau à une date antérieure le 10 et le 15 octobre, et le graphique 5, montrant le réseau entre le 16 et le 21 octobre, le nombre de lien total est similaire dans les deux cas (304 contre 307), tout comme le nombre moyen de lien (3,53 contre 3,45). De même dans les deux cas, le nombre de noeuds est quasi-identique (à 3 près).

Figure 4 – Représentation graphique de la matrice de connexion entre le 10 et le 15 octobre 2001 (uniquement individus ayant un lien avec au moins un autre inidividu du réseau)



Néanmoins, il semble que la structure du réseau ne soit plus exactement la même. Pour vérifier cette intuition, nous appliquons à nouveau le SvBM. Sur la période qui précède l'annonce, l'algorithme conclut à l'existence de trois classes latentes. La classe 2 et 3 appartenant en fait à la même communauté. Notons cependant que les individus de la classe 2 ne forment pas une communauté en soit, la probabilité de connexion intra classe étant relativement faible (proche de 0,1). La probabilité de connexion avec la classe 3 étant de plus de 0,9.

Figure 5 – Représentation graphique de la matrice de connexion entre le 16 et le 21 octobre 2001 (uniquement individus ayant un lien avec au moins un autre inidividu du réseau)



Après l'annonce, au contraire, l'algorithme conclut à l'existence de 3 classes ; la classe 2 et 3 étant cette fois deux communautés *per se* et relativement isolée l'une de l'autre, sauf par l'individu "tholt-J", un Vice-Président. Ce dernier est en contact avec l'individu « grisby-m », un « manager », et « Dasovitch-j », un employé. L'annonce de la perte s'est donc accompagnée d'une modification de la structure du réseau de lien intra-firme mais pas de l'intensité des échanges sur une même période.

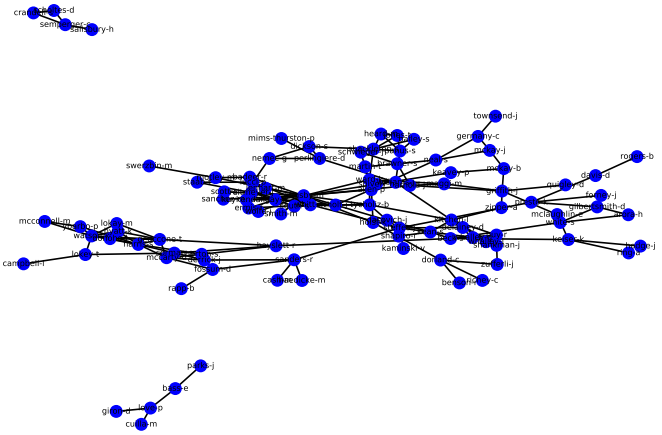
Cependant, l'étude des liens entre le début de la période et chacune des dates de fin considérées révèle la création sur la semaine qui succède de 32 nouveaux liens au sein de la firme. Le nombre de liens maximal restant quant à lui le même. Outre une restructuration du réseau d'échange, l'annonce s'est donc accompagnée de nouvelles prises de contacts.

4.5.2. Etude du réseau sur la période pré-annonce de l'ouverture d'une enquête par la SEC

L'annonce de l'ouverture d'une enquête par le gendarme de la bourse américain, la SEC, a eu le lieu le 22 octobre. Une analyse de réseau durant la semaine qui suit l'annonce montre une hausse significative des liens entre nos individus statistiques. En une semaine, on recense ainsi 447 liens contre 307 la semaine qui précède et une hausse du nombre moyen de lien de 6 (passant de 14 à 20). Sur une semaine, si l'on étudie la prise de contact depuis le début de la période contenue dans la base, le nombre de lien augmente de près de 100 (i.e. une semaine après l'annonce près de 100 contacts jamais observés auparavant se sont créés, contre une dizaine la semaine précédente)

D'un point de vue strictement graphique, on observe aussi une diminution des petits groupes en périphérie du groupe principal ainsi qu'une baisse du nombre d'individus ne prenant part à aucun échange (ils passent de 60 à 46).

Figure 6 – Représentation graphique de la matrice de connexion entre le 22 et le 27 octobre 2001



Sur le seul jour de l'annonce, le 22 octobre, 199 contacts avaient été engagés au total au sein du réseau, avec une moyenne de 3 liens par individus. Sur une période de 2 jours, c'était près de 273 liens qui étaient engagés. L'annonce de l'enquête de la SEC semble donc s'accompagner d'une importante réaction du réseau en terme de contact ainsi que d'une réorganisation de sa structure.

L'étude par SvBM des liens engagés durant la semaine suivant l'annonce de l'enquête de la SEC conclut à l'existence de 4 groupes latents : deux ayant des tendances communautaires (probabilité de connexion intraclasse égale à 0,5 et 0,26). La classe 1 a quant à elle une probabilité de connexion intraclasse relativement faible (0,1) mais a une forte probabilité de connexion à la classe 4 (0.78) ce qui suggère un certain lien communautaire entre les individus qui constituent ces deux classes. La recomposition structurelle du réseau de communication s'est donc accompagnée d'une recomposition des groupes latents au sein de la structure.

4.5.3. Analyse sur la periode pré/post-annonce de la faillite d'Enron

La faillite d'Enron est annoncée le 2 décembre 2001. Dans un article publié le lendemain sur son site, le New York Times titre « ENRON'S COLLAPSE : THE OVERVIEW ; ENRON CORP. FILES LARGEST U.S. CLAIM FOR BANKRUPTCY ». Nous allons à présent étudier la répercution de cette annonce sur le réseau en suivant la même méthodologie que précedemment.

La veille de l'annonce, le nombre total de lien était de 1763, avec un nombre moyen de lien de 12 par individu. Une semaine après, 24 nouveaux liens avaient été opérés. Entre le 25 novembre et le 1er décembre inclus, un peu plus de 300 contacts intra-réseau ont été initiés contre un peu plus de 245 entre le 2 décembre et le 8 décembre inclus. Contrairement à l'événement précédent, les prises de contact entre les individus du réseau ont donc été moins nombreuse durant la semaine qui suit la faillite que durant celle qui précède. Cela a une répercussion substantielle sur la conclusion du SvBM : de trois classes durant la période du 25 novembre au 1er décembre, le modèle conclut à l'existence de 2 classes sur la période qui suit l'annonce de la faillite, l'une des deux correspondant à une communauté.

La faible réaction du réseau à cette nouvelle pourrait provenir de l'absence d'un effet de suprise de cette annonce. En effet, les deux événements marquants précédents étaient :

1. Nov. 19 - Enron déclare devoir rembourser 690 millions de dollars de dette d'ici le 27novembre ;
2. Nov. 28 - Les actions d'Enron plongent sous la barre des 1 dollar ;

4.6. Conclusion de l'étude Enron

Nous avons donc vu que la structure et la densité des liens au sein du réseau évoluent en fonction des événments du scandale : ils augmentent après chaque événement notable et inattendu. Comme ces événements, notamment les deux premiers ne pouvaient pas être connus des salariés d'Enron en avance (du moins pas par tous), on peut se risquer à voir dans cette observation un effet causal de l'annonce sur la structure.

L'étude ci-dessous pourrait être étendue en considérant non plus l'existence ou non d'une prises de contact mais le nombre de prises de contact. Notons que l'étude par statut des employés parait compliquée compte tenu du faible effectif de chaque groupe.

5. Conclusion générale

Le SvBM est donc une alternative aux algorithmes de clustering à modèle non-formel ainsi qu'aux modèles de clustering formels fréquentistes. Contrairement aux premiers, il permet de mettre en évidence des groupes d'individus latents en partant d'une modélisation statistique, il permet ainsi l'inférence paramétrique ; par rapport aux seconds, il permet d'inscrire le modèle dans un cadre bayésien et ainsi de potentiellement mieux coller aux données.

Néanmoins, il convient de faire attention au temps de calcul nécessaire pour obtenir les résusltats, notamment sur des réseaux importants et lorsque l'algorithme est codé

dans un langage relativement lent (comme c'est le cas de `Python` par rapport au `C`, par exemple).

D'un point de vue pratique, nous avons vu que le SvBM, et l'analyse en réseau de manière plus générale, pouvait servir à analyser les changements survenant dans un réseau organisationnel. Il peuvent ainsi servir à détecter en amont l'existence d'un potentiel problème qui serait caché par les individus du réseau étudié ; les signes étant l'augmentation brusque du nombre total de prise de contact, du nombre d'individu et du nombre de prises de contact moyen ainsi que le nombre de classe détecté par le SvBM.

Références

E. Airoldi, D. Blei, S. Fienberg, and E. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn.*, 9 :1981–2014, 2008.

E. Allman, C. Matias, and J. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 6A(37) :3099–3132, 2009.

E. Allman, C. Matias, and J. Rhodes. Parameter identifiability in a class of random graph mixture models. *J. Statist. Planning and Inference*, 5(141) :1719–1736, 2011.

A. Celisse, J.-J. Daudin, and L. Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Statist.*, 6 :1847–1899, 2012.

J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Annals of Applied Statistics*, 18(12) :173–183, 2008.

P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels : First steps. *Social Networks*, 5(2) :109–137, June 1983.

P. Latouche, E. Birmelé, and C. Ambroise. Overlapping stochastic block models with application to the french political blogosphere. *Annals of Applied Statistics*, 5(1) : 309–336, 2011.

P. Latouche, E. Birmele, and C. Ambroise. Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, 12(1) :93–115, August 2012.

K. Nowicki and T. A. B. Snijder. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455) :1077–1087, September 2001.

F. Picard, V. Miele, J.-J. D. A. Cottret, and S. Robin. Deciphering the connectivity structure of biological networks using mixnet. *BMC Bioinformatics*, 10 :1–11, 2009.

Y. J. Wang and G. Y. Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397) :8–19, March 1987.

H. White, S. Boorman, and R. Breiger. Social structure from multiple networks. *American journal of Sociology*, (81) :730–780, June 1976.

H. Zanghi, S. Volant, and C. Ambroise. Clustering based on random graph model embedding vertex features. *Pattern Recognition Letters*, 31(9) :830–836, 2010.