Inférence de dates d'activité à partir d'un réseau d'interactions datées

Fabrice Rossi & Pierre Latouche

SAMM EA 4543, Université Paris 1 Panthéon-Sorbonne, prénom.nom@univ-paris1.fr

Résumé. Nous proposons dans cet article un nouveau modèle génératif pour les graphes qui s'appuie sur une approche à espace latent pour expliquer un ensemble d'interactions datées. L'objectif du modèle est de fournir des estimations globales pour les dates d'activité d'un ensemble d'acteurs dont les dates d'interaction sont connues avec une précision raisonnable, par opposition à une estimation locale simple. Nous montrons sur des données artificielles que le modèle proposé produit une meilleure estimation des dates d'activité que les moyennes locales si le réseau étudié est suffisamment dense.

Mots-clés. Réseau temporel, dates d'activité, données historiques, modèle à espace latent

Abstract. We propose in this paper a new generative model for graphs that uses a latent space approach to explain timestamped interactions. The model is designed to replace local averages by global estimates of activity dates in historical networks where only the interaction dates between agents are known with reasonable precision. Experimental results show that the model provides better results than local averages in dense enough networks.

Keywords. Temporal network, Activity Date, Historical Data, Latent Space Model

1 Contexte

Nous étudions dans cette communication un ensemble d'interactions entre acteurs enregistrées sur une très longue période, comparativement à la durée de vie des acteurs. Le contexte applicatif typique de cette situation est celui d'archives historiques longues, par exemple celles des notaires : les biens échangés ont une durée de vie potentielle largement supérieure à celles des acquéreurs. Dans Boulet et al. [2008], Rossi et al. [2011], par exemple, les auteurs étudient une base d'actes notariés portant sur plus de 300 ans et faisant intervenir plus de 4 000 acteurs.

Dans ce type de données, il est naturel de supposer que les interactions sont mieux décrites que les agents, et en particulier qu'on connaît de façon sûre (ou au moins très précise) les dates des interactions. L'objectif de notre travail est d'utiliser ces dates d'interactions pour inférer des dates d'activités pour les acteurs, en allant au delà de la solution simple qui consiste à considérer que la date centrale d'activité d'un acteur est simplement la moyenne des dates de ses interactions avec les autres acteurs.

2 Modèle génératif

Pour ce faire, nous introduisons un modèle génératif à variable latente dont l'objectif est d'expliquer le graphe d'interaction. Plus précisément, notons G le graphe non orienté construit sur l'ensemble de sommets $V = \{1, \ldots, n\}$, chaque sommet étant un des acteurs de l'ensemble d'interactions. La matrice d'adjacence du graphe est notée A, avec $A_{ij} = 1$ si et seulement si les acteurs i et j ont interagi. Dans le cas $A_{ij} = 1$, on dispose aussi de D_{ij} la date d'interaction associée (il est facile d'étendre le modèle au cas des multi-graphes, c'est-à-dire quand on autorise plusieurs interactions entre les mêmes acteurs).

Pour engendrer un tel graphe (A, D), on suppose que chaque acteur i est caractérisé par une date centrale d'activité Z_i non observée. On suppose ensuite, que la propension de deux acteurs à interagir est liée à leur distance temporelle $Z_i - Z_j$ et, en cas d'interaction, que la date d'interaction est centrée sur la moyenne des dates d'activité. Enfin, on suppose que les interactions sont indépendantes conditionnellement à la connaissance des dates d'activité. Techniquement, nous choisissons donc un modèle de la forme :

$$p(A, D|Z, \theta) = \prod_{i \neq j, A_{ij}=0} P(A_{ij} = 0|Z_i, Z_j, \theta) \times \prod_{i \neq j, A_{ij}=1} p(D_{ij}|A_{ij} = 1, Z_i, Z_j, \theta) P(A_{ij} = 1|Z_i, Z_j, \theta), \quad (1)$$

où θ désigne l'ensemble des paramètres du modèle. Nous spécialisons le modèle en utilisant une régression logistique pour les interactions (comme dans Hoff et al. [2002]) et une loi normale pour les dates d'interaction, soit :

$$\log \frac{P(A_{ij} = 1|Z_i, Z_j, \alpha, \beta)}{P(A_{ij} = 0|Z_i, Z_j, \alpha, \beta)} = \alpha - \beta(Z_i - Z_j)^2, \quad (2)$$

et

$$D_{ij}|Z_i, Z_j, \sigma \sim \mathcal{N}\left(\frac{Z_i + Z_j}{2}, \sigma^2\right). \quad (3)$$

En pratique, nous connaissons A et D, et nous estimons Z, α, β et σ par maximum de vraisemblance.

3 Quelques résultats expérimentaux

Nous étudions l'intérêt du modèle proposé sur des données artificielles pour lesquelles nous avons ainsi une vraie date centrale d'activité pour chaque acteur (ce qui n'est pour l'instant pas le cas dans de grandes bases de données historiques). Comme indiqué plus haut, la technique la plus simple pour estimer une date d'activité par acteur consiste à

calculer la moyenne des dates des interactions de l'acteur, c'est-à-dire utiliser

$$\hat{Z}_i = \frac{\sum_{j,A_{ij}=1} D_{ij}}{\sum_j A_{ij}}. \tag{4}$$

Pour évaluer le modèle, nous calculons l'amélioration entre terme d'erreur quadratique moyenne (EQM) obtenue en utilisant l'estimation des Z_i par maximum de vraisemblance global, par rapport l'erreur obtenue par moyenne locale ($\hat{Z}_i$).

Nous étudions dans un premier temps des réseaux comportant 100 acteurs avec des dates d'activité choisies uniformément dans $[1200, 1400]$. Le réseau (A, D) est engendré en utilisant le modèle génératif choisi, avec un réglage des paramètres β et σ réalistes, c'est-à-dire des valeurs qui conduisent à une durée de vie maximale pour les acteurs de l'ordre de 80 ans. La valeur de α varie afin d'obtenir des graphes plus ou moins denses. La figure 1 résume les résultats obtenus sur un ensemble de 2150 réseaux : quand le nombre d'interactions par acteur dépasse 2 en moyenne, l'estimateur basé sur le modèle est presque constamment meilleur que l'estimateur local.

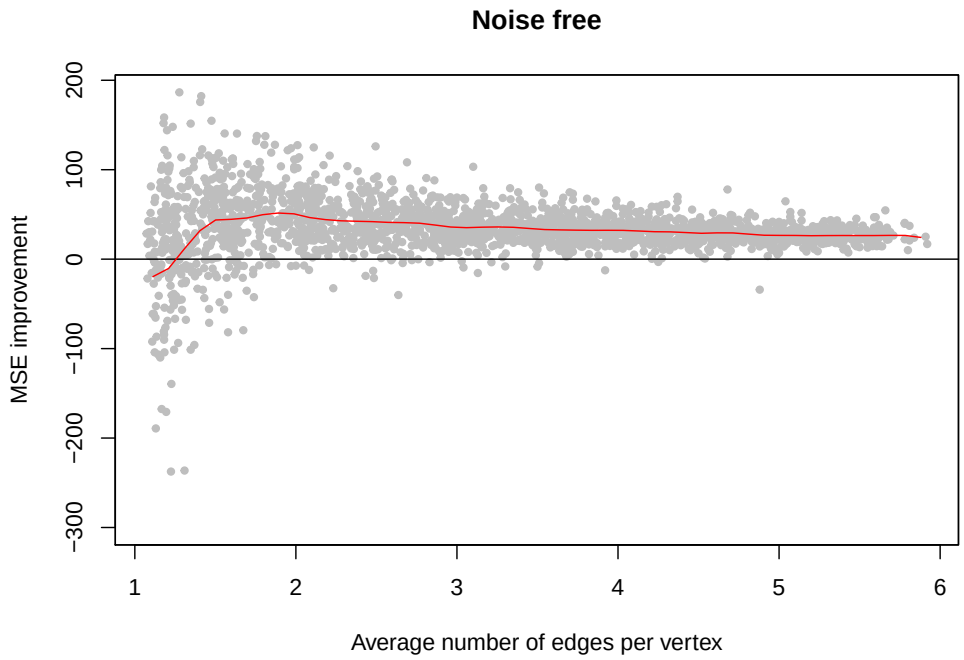

FIGURE 1 – Situation idéale : les données sont simulées selon le modèle. Chaque point correspond à un réseau et indique l'amélioration en erreur quadratique moyenne (eqm).

La figure 2 résume les résultats obtenus quand on simule les données de façon plus réaliste. On commence par engendrer un réseau avec le modèle proposé, puis, pour tenir compte des erreurs de transcriptions fréquentes dans les archives anciennes, on reconnecte

aléatoirement un certain pourcentage des arêtes (ici 5 %). L'idée sous-jacente est qu'en raison des nombreuses homonymies dans la période médiévale étudiée, il arrive qu'une personne soit affectée de façon erronée à une interaction (sans que la date de celle-ci soit incorrecte). On constate que le modèle reste meilleur que la solution locale, mais à condition d'avoir plus d'information (un minimum de 3 interactions par acteur).

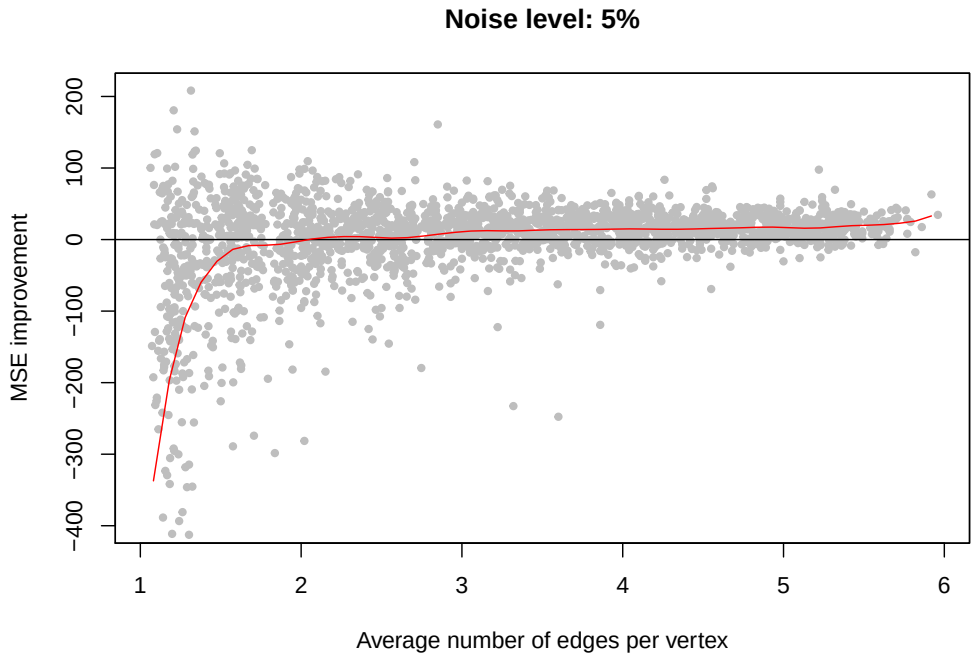

FIGURE 2 – Situation bruitée : les données sont simulées selon le modèle mais 5 % des arêtes sont reconnectées aléatoirement.

Références

R. Boulet, B. Jouve, F. Rossi, and N. Villa. Batch kernel SOM and related Laplacian methods for social network analysis. *Neurocomputing*, 71(7–9) :1257–1273, March 2008. doi : 10.1016/j.neucom.2007.12.026.

P. Hoff, A. Raftery, and M. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97 :1090–1098, 2002.

F. Rossi, N. Villa-Vialaneix, and F. Hautefeuille. Exploration of a large database of french charters with social network methods. In *International Medieval Congress (IMC 2011)*, number 1607-c, Leeds (United Kingdom), July 2011. Institute for Medieval Studies.