Inférence de réseaux de diffusion latents

Gaël Poux-Médard

13/05/2020

Sommaire

Introduction

Motivations

- Les réseaux sont à la base de nombreux modèles de diffusion (maladies, informations, ...)
- Difficile d'obtenir des réseaux de diffusion (réseaux sociaux, de contacts, ...)

Données disponibles

- Historique des infections au niveau individuel
  - X a eu le rhume à t_1, Y l'a eu à t_2, ...
  - X a retweeté une photo de chat à t_1, Y l'a retweetée à t_2, ...
- Machine learning pour inférer le réseau de diffusion

Acteurs du domaine

- Un nom à retenir : Manuel Gomez-Rodriguez
  - PhD obtenu à Stanford en 2013
  - A lancé toute une littérature d'inférence de réseaux

Sommaire

Petit état de l'art

Travaux fondateurs de M. Gomez-Rodriguez

- **2010 - NetInf** : inférence de réseaux statiques non-pondérés [a]
- **2011 - NetRate** : inférence de réseaux statiques pondérés [b]
- **2013 - InfoPath** : inférence de réseaux dynamiques pondérés [c]
- **2013** : Modèle d'inférence regroupant l'essentiel des travaux existants dans un même cadre mathématique (counting process) [d]
- **Bonus** : toutes les fonctions objectif à maximiser sont convexes + chaque problème se subdivise en N sous-problèmes solubles en parallèle.

[a] *Inferring Networks of Diffusion and Influence*
[b] *Uncovering the Temporal Dynamics of Diffusion Networks*
[c] *Structure and Dynamics of Information Pathways in Online Media*
[d] *Modeling Information Propagation with Survival Theory*

Petit état de l'art

Autres contributions

- **2010 - Connie - S.A. Myers & J. Leskovec** : inférence convexe de réseaux statiques pondérés [a] ; (dans l'article, sorti 5 mois après NetInf, la fonction à optimiser n'est pas convexe selon moi, et les auteurs résolvent d'ailleurs le problème avec un solveur non-convexe. Pour un article qui porte ce nom...)
- **2012 - KernelCascade (NetRate modifié) - N. Du & al.** : inférence multi-kernel (on y reviendra) de réseaux statiques pondérés [b]

[a] *On the Convexity of Latent Social Network Inference*
[b] *Learning Networks of Heterogeneous Influence*

Sommaire

Survival analysis

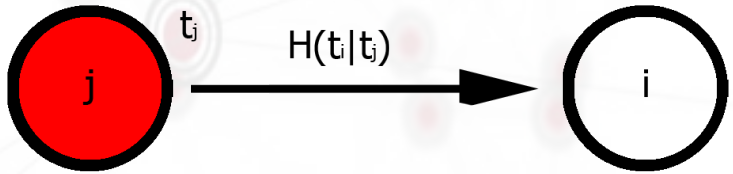
Survival analysis

- Noeud sujet à des attaques/tentatives de contamination dont le nombre varie en fonction du temps. Le noeud ”meurt” lorsqu'une attaque arrive à le contaminer.
  - $H(t_i|t_j)$: Hazard rate, ou la probabilité instantanée de contamination.
  - $H(t_i|t_j)dt$ = probabilité de contamination entre t_i et $t_i + dt$ sachant t_j
- $S(t_i|t_j)$: probabilité de survie (non-contamination) de i au temps t_i sachant que j a été contaminé au temps t_j.
  - S est une fonction non-croissante du temps.
  - Relation : $H(t_i|t_j) = -\frac{S'(t_i|t_j)}{S(t_i|t_j)} = \frac{f(t_i|t_j)}{S(t_i|t_j)}$
  - $f(t_i|t_j)$ est la densité d'attaques.
- Le choix de S(t) définit le *kernel* du modèle.
  - $S(t) \propto e^{-\alpha t} \rightarrow H(t) = \alpha \rightarrow$ exponential kernel
  - $S(t) \propto e^{-\frac{\alpha}{2}t^2} \rightarrow H(t) = \alpha \cdot t \rightarrow$ Rayleigh kernel
  - $S(t) \propto e^{-\alpha \ln(t)} \rightarrow H(t) = \alpha/t \rightarrow$ power-law kernel
- α_{ij} est un élément de la matrice d'adjacence α définie suivant un kernel. De manière générale, étant données les définitions de S(t), $\alpha_{ij} = 0$ signifie l'absence de lien, et $\alpha_{ij} \gg 0$ signifie un lien fort (contamination immédiate).

2011 - NetRate[1]

Likelihood d'une infection de i

- $\mathcal{L}_{j \to i} = f(t_i|t_j) \cdot \prod_{t_k < t_i, t_k \neq t_j} S(t_i|t_k)$
- Rappel : $f(t_i|t_j)$ est la densité d'attaques sur i à t_i sachant que j a été infecté à t_j, et $S(t_i|t_j)$ est la probabilité de survie de i à t_i sachant j infecté à t_j.

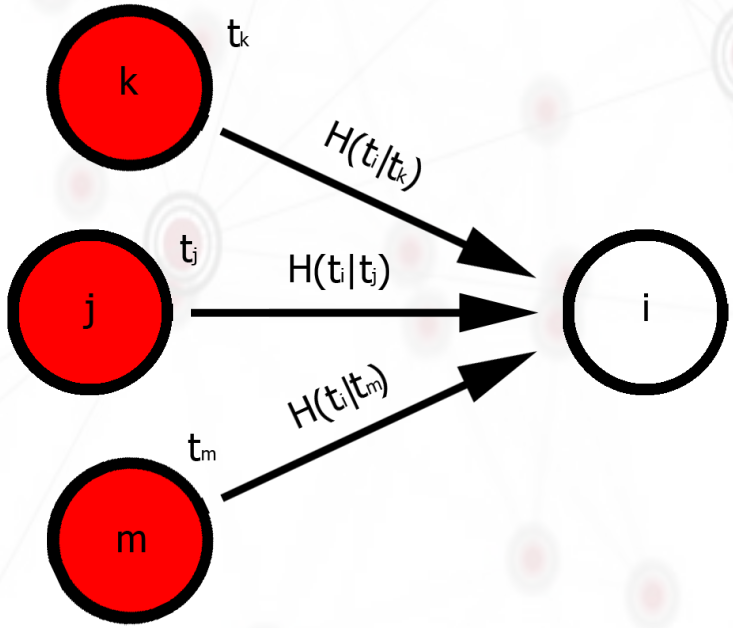


[1] *Uncovering the Temporal Dynamics of Diffusion Networks*

2011 - NetRate

Likelihood des infections de i

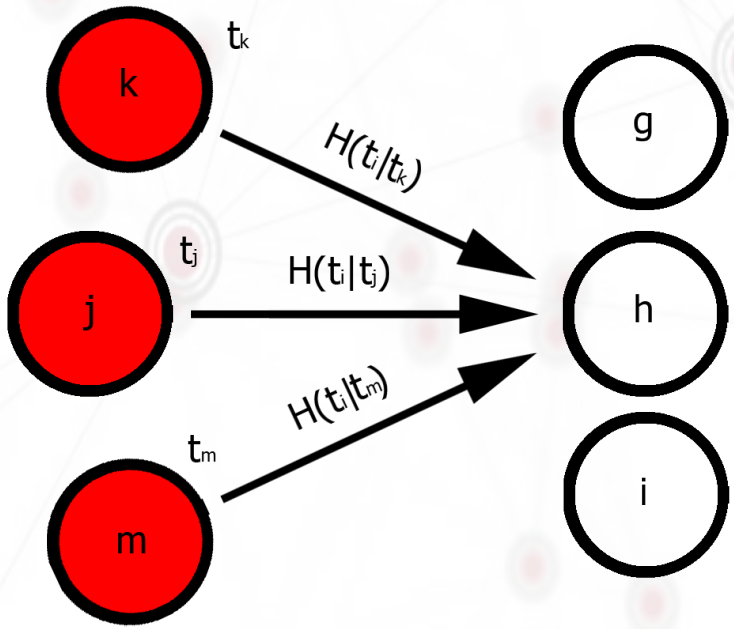
- $\mathcal{L}_c = \sum_{t_j < t_i} f(t_i|t_j) \cdot \prod_{t_k < t_i, t_k \neq t_j} S(t_i|t_k)$
- C'est une somme (et pas un produit) car on considère que le noeud n'est infecté que par un seul voisin. Cette formulation encourage des solutions sparses.



2011 - NetRate

Likelihood d'une cascade

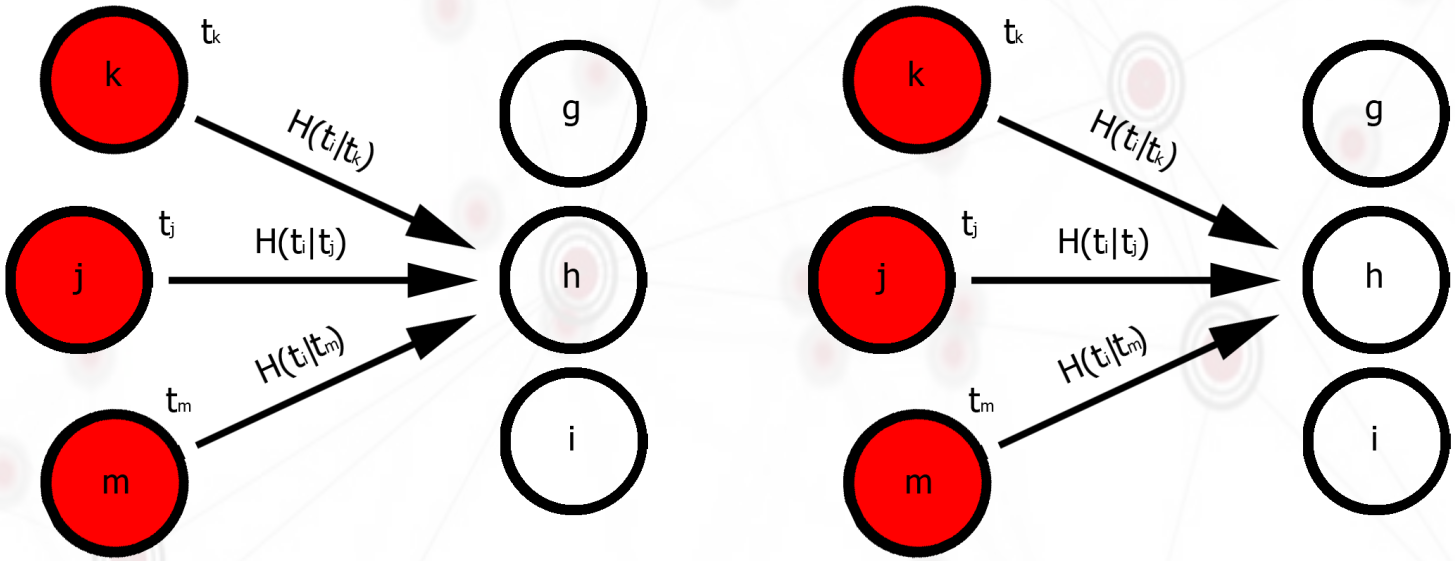
- $\mathcal{L}_i = \prod_{t_i} \sum_{t_j < t_i} f(t_i|t_j) \cdot \prod_{t_k < t_i, t_k \neq t_j} S(t_i|t_k)$
- On considère maintenant le temps d'infection de chacun des noeuds dans une cascade.



2011 - NetRate

Likelihood des observations

- $\mathcal{L} = \prod_{c} \prod_{t_i} \sum_{t_j < t_i} f(t_i|t_j) \cdot \prod_{t_k < t_i, t_k \neq t_j} S(t_i|t_k)$
- On multiplie la likelihood de chacune des cascades pour avoir la likelihood finale.



2011 - NetRate

Likelihood des observations

- On relaxe maintenant la condition $t_k \neq t_j$ du dernier produit
  - $\mathcal{L} = \prod_c \prod_{t_i} \sum_{t_j<t_i} \cdot \frac{f(t_i|t_j)}{S(t_i|t_j)} \prod_{t_k<t_i} S(t_i|t_k)$
  - Or $\frac{f(t_i|t_j)}{S(t_i|t_j)} \equiv H(t_i|t_j)$

Log-likelihood finale à maximiser avec α

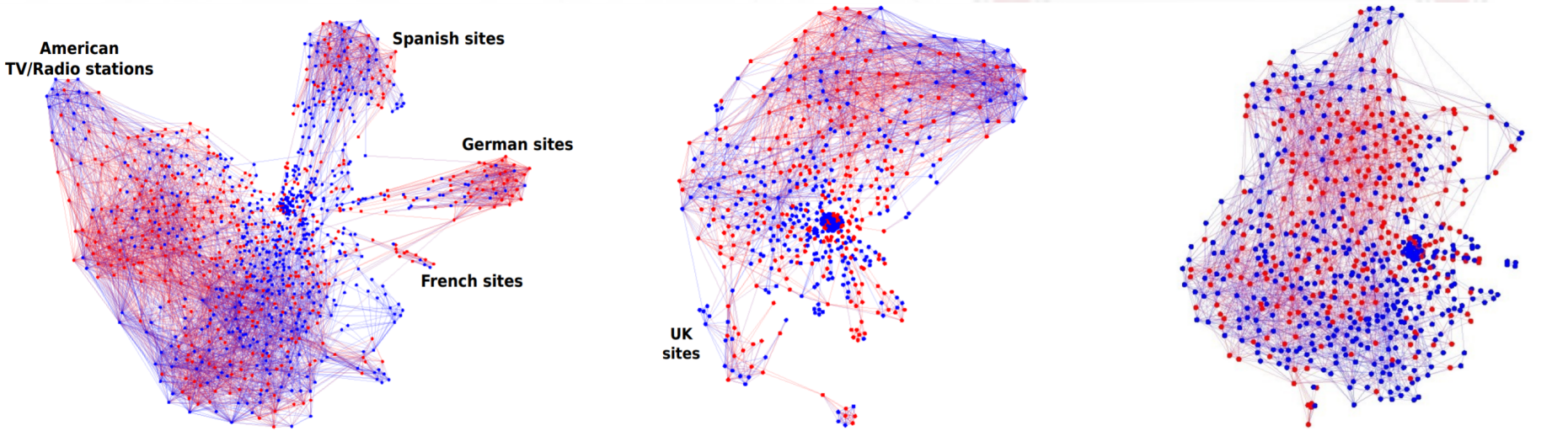
- $\ell = \sum_c \sum_{t_i} (\ln(\sum_{t_j<t_i} H(t_i|t_j, \alpha_{ij})) + \sum_{t_k<t_i} \ln S(t_i|t_k, \alpha_{ik}))$

Convexité

- Par composition, cette fonction est convexe si:
  - S est log-convexe en son argument α (valable pour toutes les fonctions type $S \propto e^{-\alpha f(t)}$)
  - H est convexe en son argument α (valable pour toutes les fonctions type $H \propto \alpha f(t)$)

Exemples NetRate

- Précision sur la présence de liens entre 80% et 95% pour des réseaux synthétiques d'environ 1000 noeuds sur lesquels on a généré ≈5000 cascades indépendantes classiques (algorithme de Gillespie).
- Erreur quadratique moyenne d'environ 0.05 sur les poids/coefficients des liens, dans la même configuration.



Fukushima (2011-03-18) UK royal wedding (2011-04-02) Syria's uprise (2011-04-05)

Figure 1: Exemple d'application de NetRate sur des données réelles (3 datasets). Les cascades considérées sont des cascades de memes. Le site/blog 1 publie un meme A à un temps t_1, et le site/blog 2 publie le même meme A au temps t_2, etc. Les noeuds bleus sont des blogs, et les noeuds rouges sont les sites de média mainstream.

2013 - InfoPath[2]

Un NetRate amélioré

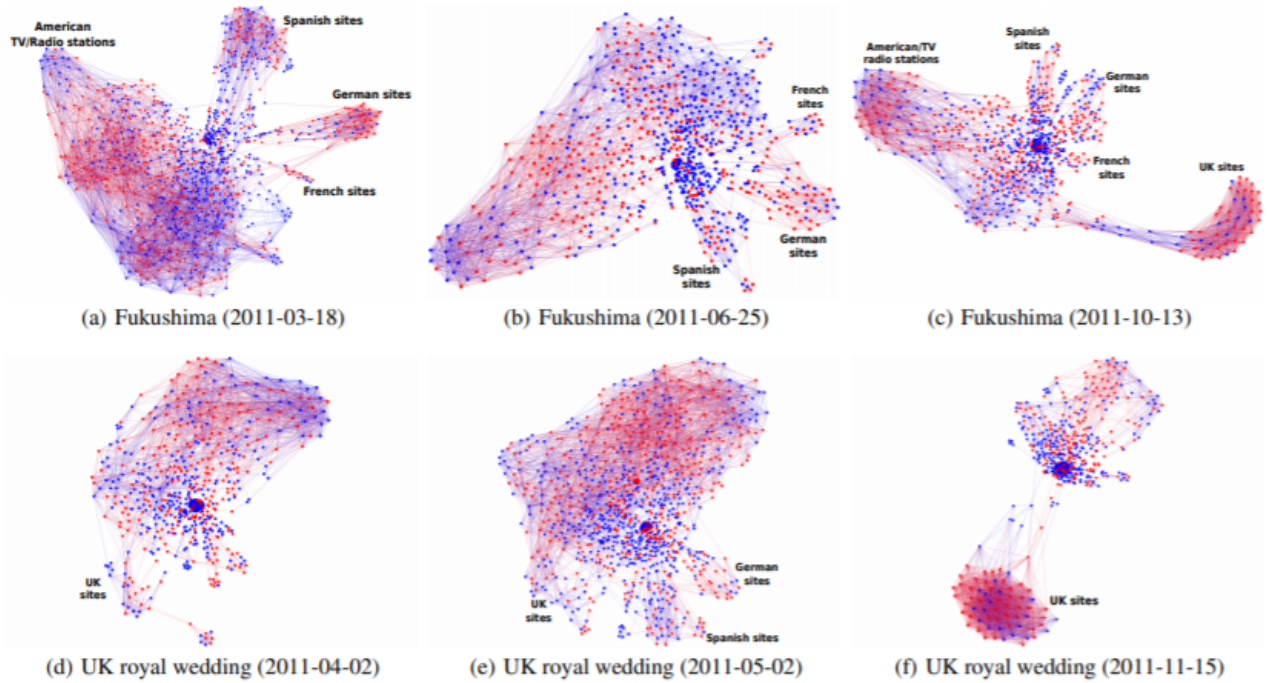
- InfoPath vise à étudier la dynamique du réseau
- NetRate infère α, InfoPath infère $\alpha(t)$
- La likelihood à maximiser est fondamentalement la même. Pour passer de NetRate à InfoPath, il faut modifier les données.

Échantillonage temporel

- Plutôt que de donner à l'algorithme toutes les données disponibles (NetRate), on va lui en donner un sous-ensemble.
- Pour avoir $\alpha(t_1)$, on va échantillonner le corpus suivant une exponentielle décroissante.
  - Une cascade observée à t_c appartiendra au corpus utilisé pour inférer $\alpha(t_1)$ avec une probabilité $p \propto e^{-(t_1 - t_c)}$
- Une fois le corpus pour $\alpha(t_1)$ créé, on lance NetRate, puis on recommence pour $\alpha(t_2)$.

[2] *Structure and Dynamics of Information Pathways in Online Media*

Exemples InfoPath



(a) Fukushima (2011-03-18) (b) Fukushima (2011-06-25) (c) Fukushima (2011-10-13)

(d) UK royal wedding (2011-04-02) (e) UK royal wedding (2011-05-02) (f) UK royal wedding (2011-11-15)

Figure 2: Exemple d'application d'InfoPath sur des données réelles (2 datasets). Les cascades considérées sont encore des cascades de memes. Les noeuds bleus sont des blogs, et les noeuds rouges sont les sites de média mainstream.

- Vidéo exemple : Amy Winehouse post-mortem, 20/01/2012-28/02/2012. Note : elle a gagné un grammy le 12/02/2012.
- Autres vidéos : Site de Stanford

Sommaire

2012 - KernelCascade[3]

Kernels multiples

- Nouvelle définition de S(t).
- Dans NetRate et InfoPath, $S(t) \propto \{e^{-\alpha t}, e^{-\alpha t^2}, e^{-\alpha \ln t}\}$
- Dans KernelCascade, $S(t) \propto e^{-\sum_i \alpha_i f_i(t)} \rightarrow H(t) = \sum_i \alpha_i f_i'(t)$
- On vérifie trivialement que ce modèle est toujours convexe en $\alpha_i \ \forall i$.

Application

- Dans l'article, les auteurs choisissent un kernel gaussien.
- La probabilité de contamination instantanée H(t) est une combinaison linéaire de gaussiennes centrées en t={0, 1, 2, ...}.
- Suivant $H(t) = -\frac{S'(t)}{S(t)}$, on peut calculer S(t).
- Très bons résultats sur des réseaux synthétiques (F1-score $\approx$ 1), et permet d'inférer le kernel expliquant le mieux une observation.

[3] *Learning Networks of Heterogeneous Influence*

2013 - Généralisation[4]

Diffusion comme un processus de comptage

- Gomez-Rodriguez repart de la base de la théorie de la survie pour dériver un modèle général d'inférence de réseaux latents.
- Pas tout compris encore...

Likelihood généralisée

$$\log f(\mathbf{t};\mathbf{A}) = \sum_{i:t_i<T} \log\left(\sum_{j:t_j<t_i} \alpha_{ji}\gamma(t_j;t_i)\right) - \sum_{i:t_i<T}\sum_{k:t_k<t_i} \alpha_{ki}\int_{t_k}^{t_i}\gamma(t_k;t)\,dt - \sum_{n:t_n>T}\sum_{m:t_m<T} \alpha_{m,n}\int_{t_m}^{T}\gamma(t_m;t)\,dt,$$

(a) Likelihood

Network Inference Method	$\gamma(\mathbf{t_j};\mathbf{t_i})$
NetRate, InfoPath (Exp)	$I(t_j < t_i)$
NetRate, InfoPath (Pow)	$\max(0, 1/(t_i - t_j))$
NetRate, InfoPath (Ray)	$\max(0, t_i - t_j)$
KernelCascade	$\{k(\tau_l, t_i - t_j)\}_1^m$
moNet	$I(t_j < t_i)\gamma e^{-d(\mathbf{f}_j,\mathbf{f}_i)}$

(b) Modèles inclus dans cette formulation

[4] *Modeling Information Propagation with Survival Theory*

Sommaire

Résumé

	Weighted	Dynamic	Convex	Kernel
NetInf			x	EXP, P-L
NetRate	x		x	EXP, RAY, P-L
InfoPath	x	x	x	EXP, RAY, P-L
KernelCascade	x		x	Multiple
Connie	x		?	EXP, P-L, Weibull, ...