Fouilles de données et Medias sociaux

Master 2 DAC - FDMS

Sylvain Lamprier

UPMC

Diffusion d'Information

Diffusion

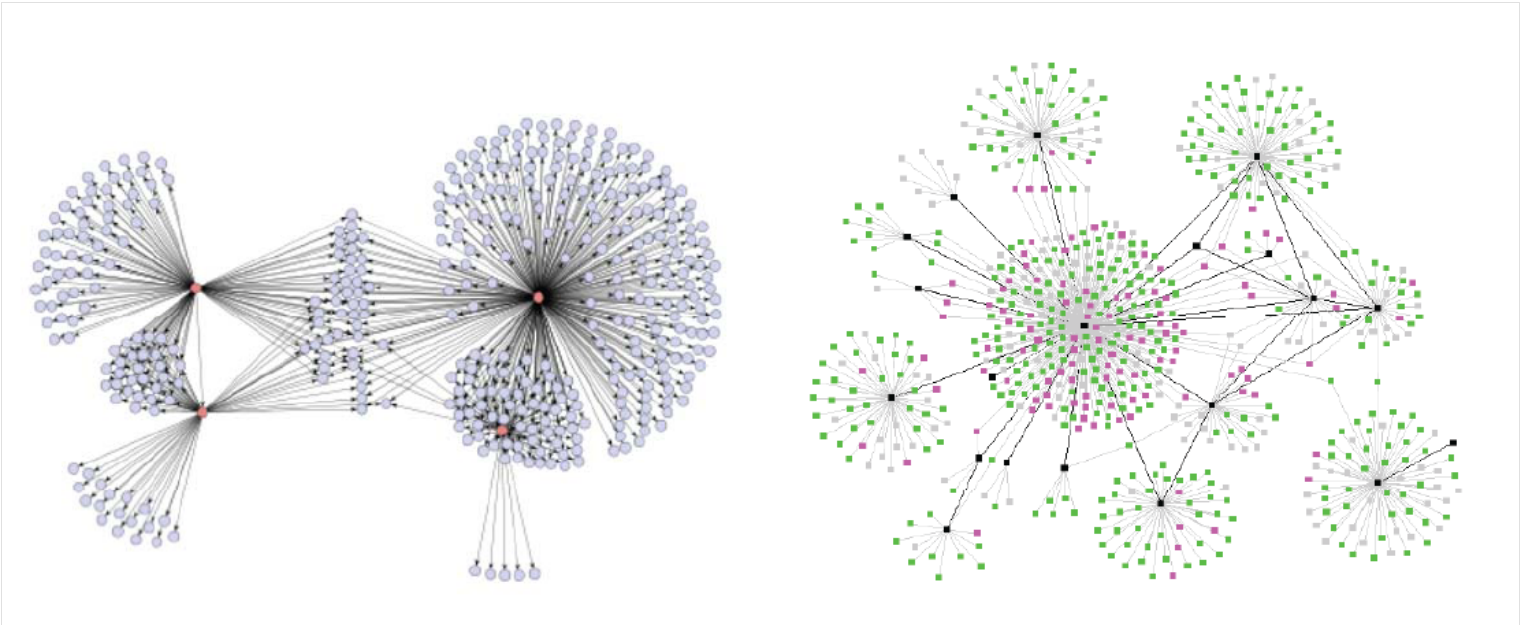
- Diffusion
  - Processus fondamental dans les réseaux : dynamique des échanges
  - Episode de diffusion: Ensemble d'évènements liés qui se propagent sur le réseau au fil du temps
- Objets de la diffusion
  - Virus or propagation de maladies
  - Bouche à oreille / marketing viral
  - News, opinions, rumeurs, ..
  - Thématiques / videos / liens sur les blogs
  - Façons de s'exprimer
  - Relations d'amitié
  - Comportements
  - Propagation d'erreurs / problèmes
  - ...

Diffusion

Les études experimentales autour de la diffusion ont une longue histoire:

- Diffusion de pratiques agricoles (1943)
  - Etude de l'adoption d'un nouveau maïs hybride par 259 fermiers de l'Iowa
  - Conclusion: le réseau des relations entre personnes joue un rôle important pour l'adoption de nouveaux produits
- Diffusion de pratiques médicales (1966)
  - Etude de l'adoption de nouveaux medicaments par des docteurs de l'Illinois
  - Conclusion: Les études cliniques et scientiques ne suffisent pas à convaincre les médecins $\Rightarrow$ **C'est le bouche à oreille qui a permis l'adoption des medicaments par la communauté**
- Effets psychologiques des opinions de l'entourage (1958)
- Contagion de l'obesité (2007)
  - Avoir un ami en surpoids augmente notre probabilité de devenir obèse de 57%.

Diffusion vs. Recommendation



- Un de ces deux réseaux est un réseau de recommendation de produits, l'autre correspond à un réseau de diffusion (e.g., d'un virus)
- Lequel correspond à un réseau de diffusion ?

Diffusion vs. Recommendation

- Recommendation
  - Capture des préférences des utilisateurs
  - Utilisation de leurs ressemblances
- Diffusion
  - Capture de la dynamique des échanges
  - Extraction d'influences
  - Temporalité d'évènements liés

Diffusion vs. Recommendation

- Objet de la diffusion: un Item
  - Noeuds du réseau = Personnes
  - Infection d'un noeud = une personne adopte l'item considéré
  - Réseau d'influences entre personnes
    - $\Rightarrow$ Quand un item est adopté par tel utilisateur, il a ensuite tendance à être adopté par tels autres
- Object de la diffusion: une Personne
  - Noeuds du réseau = Items
  - Infection d'un noeud = un item est adopté par la personne considérée
  - Réseau de recommendation temporelle
    - $\Rightarrow$ Quand un utilisateur adopte tel item, il a ensuite tendance à adopter tels autres

Homophilie vs. Influence

- Homophilie
  - Deux personnes connectées ont tendance à avoir des comportements similaires
- Influence
  - Le comportement d'une personne sur un réseau à un impact sur ceux de son voisinage

$\Rightarrow$ La temporalité des évènements joue un rôle primordial pour distinguer l'influence de l'homophilie

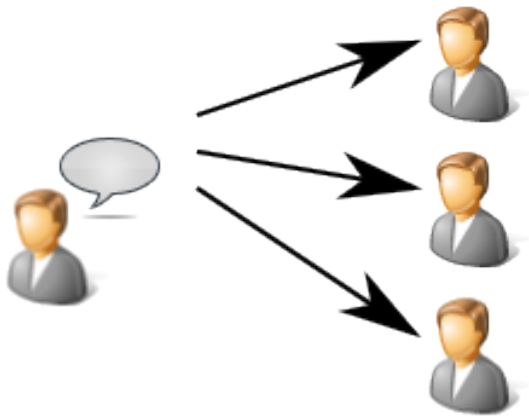
  - Si on observe des relations de précédence entre les évènements : influence

Diffusion sur les réseaux

Diffusion = Processus itératif de passage de message

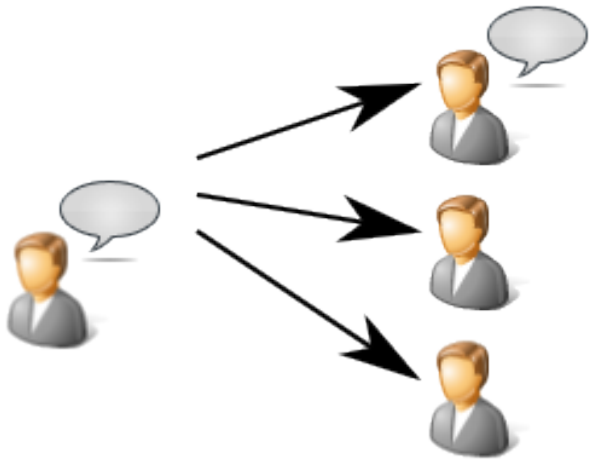
Diffusion sur les réseaux

Diffusion = Processus itératif de passage de message



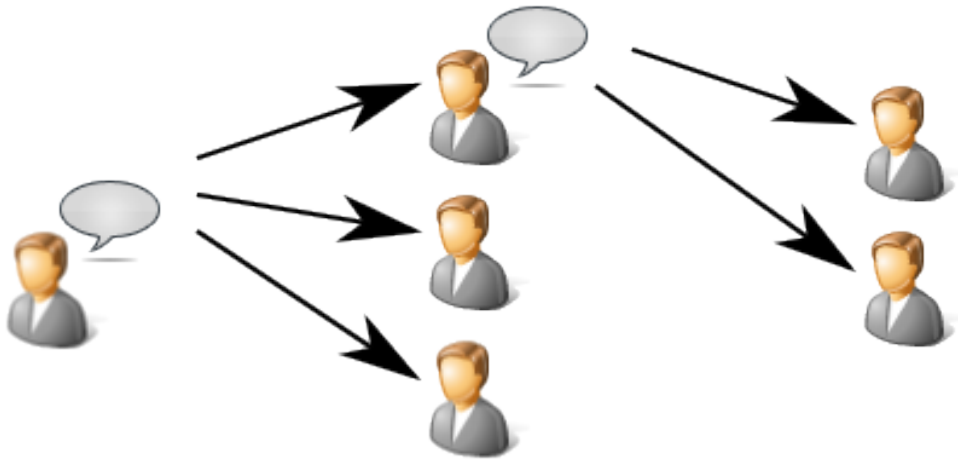
Diffusion sur les réseaux

Diffusion = Processus itératif de passage de message



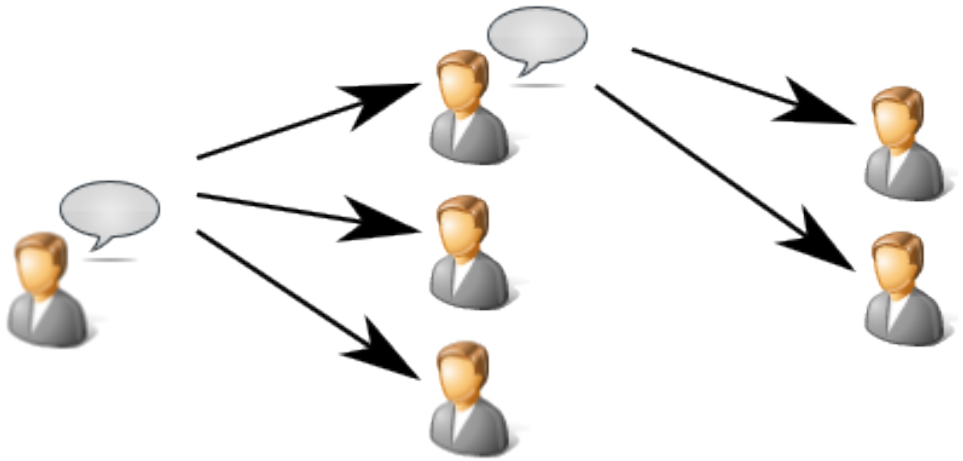
Diffusion sur les réseaux

Diffusion = Processus itératif de passage de message



Diffusion sur les réseaux

Diffusion = Processus itératif de passage de message



$\Rightarrow$ Définit une **Cascade** de diffusion

Diffusion sur les réseaux

- Challenges
  - La cascade de diffusion est généralement cachée
    - On ne sait pas qui a passé le contenu diffusé à qui
  - Ce dont on dispose c'est la participation datée d'utilisateurs à la diffusion (**episode** de diffusion)
    - On sait seulement qui a participé à quoi et quand ils l'ont fait

  $\Rightarrow$ Modélisation des phénomènes de diffusion sur un réseau = Problème d'apprentissage de relations d'influence à partir d'informations incomplètes

Diffusion sur les réseaux

- Challenges
  - Le contenu peut avoir un impact sur les distributions de relations d'influence du réseau
    - On ne se comporte pas de la même manière pour tous les types de contenu
    - *e.g.*, Paul peut avoir une forte influence sur Pierre lorsque l'on s'intéresse à du sport mais très peu quand il s'agit de politique
  - La diffusion ne se cantonne pas au réseau étudié
    - Hypothèse de monde fermé rarement vérifiée
    - Diverses possibilités d'échange d'information: rencontre physique, media traditionnel, divers réseaux, etc...
  - Inter-dépendance / concurrence des diffusions
    - Certains processus de diffusion peuvent être impactés par d'autres processus simultanés qui modifient les comportements sur le réseau
  - Dynamicité du réseau
    - Nouveaux utilisateurs / nouvelles relations
    - Modification des relation d'influence au cours du temps

Diffusion sur les réseaux

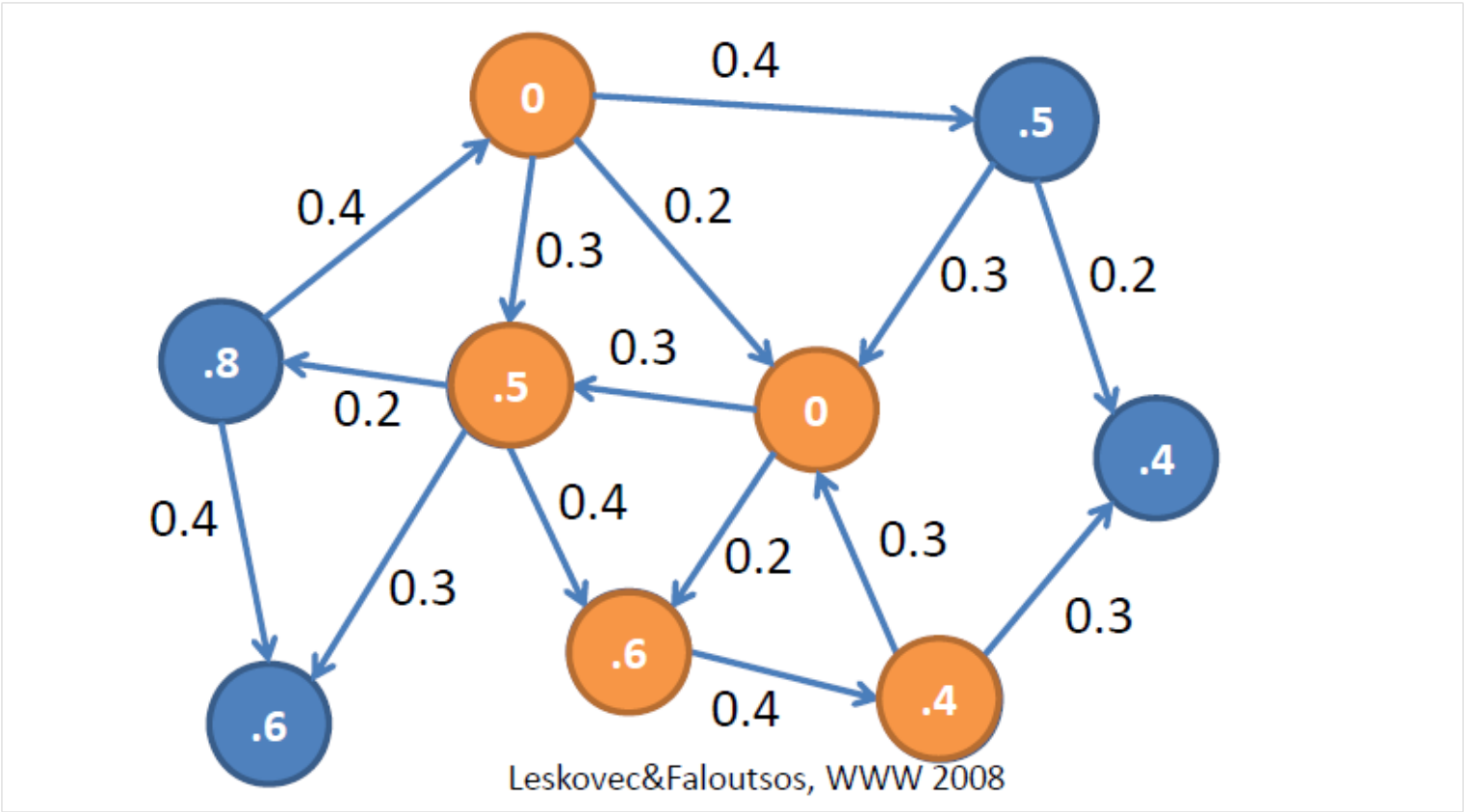
- Diverses Questions
  - Un contenu diffusé par une (ou plusieurs) source(s) donnée(s) va-t-il faire le buzz ?
  - Que va être le volume de personnées infectées / impactées par ce contenu ?
  - Qui va être impacté par une diffusion connaissant les sources ?
  - Qui est la source d'une diffusion ?
  - Quelles sont les personnes les plus influentes d'un réseau ?
  - A qui donner un contenu pour en maximiser la diffusion ?
  - Comment stopper la diffusion d'un contenu ?
  - ...

Modélisation de Diffusion sur les réseaux

- Deux modèles centraux *[Kempe et al., 2003]*
  - Linear Threshold (LT)
    - Modèle centré receveur
  - Independent Cascade (IC)
    - Modèle centré emetteur

Modélisation de Diffusion sur les réseaux

- Modèle Linear Threshold (LT) *[Granovetter, 1973]*
  - Hypothèse: l'influence est additive
  - Si la somme des influences des prédecesseurs infectés d'un utilisateur *u* est supérieure à un seuil γ, alors l'utilisateur *u* est infecté à son tour



Linear Threshold Model

Soient les éléments suivants:

- Un graphe orienté $G = (V, E)$ avec V l'ensemble des noeuds du graphe et E l'ensemble de ses relations;
- Une fonction $Preds : V \rightarrow 2^V$ telle que $Preds(j)$ retourne l'ensemble des prédecesseurs du noeud j dans le graphe : $Preds(j) = \{i \in V | (i, j) \in E\}$;
- Une fonction $Succs : V \rightarrow 2^V$ telle que $Succs(i)$ retourne l'ensemble des successeurs du noeud i dans le graphe : $Succs(i) = \{j \in V | (i, j) \in E\}$;
- Une fonction $I^D : \{1..T\} \rightarrow 2^V$ telle que $I^D(t)$ retourne l'ensemble des noeuds infectés avant l'iteration t de la diffusion D;
- Une fonction $t^D : V \rightarrow \{1..T\} + \infty$ telle que $t^D(v)$ retourne le temps d'infection du noeud v dans l'episode de diffusion D (∞ si pas infecté dans D);
- Un ensemble des poids d'influence θ définis pour tous les couples $(i, j) \in E$, avec $\theta_{i,j} \in \mathbb{R}$ le poids d'influence du noeud i sur le noeud j;
- Un ensemble de seuils d'infection γ définis pour tous les noeuds de V, avec $\gamma_i \in \mathbb{R}$ le seuil d'infection du noeud i.

Un noeud j est alors infecté par la diffusion D à l'iteration t si:

$$\sum_{i \in Preds(j) \cap I^D(t)} \theta_{i,j} \geq \gamma_j$$

Linear Threshold Model

Soient les éléments suivants:

- Un graphe orienté $G = (V, E)$ avec V l'ensemble des noeuds du graphe et E l'ensemble de ses relations;
- Une fonction $Preds : V \rightarrow 2^V$ telle que $Preds(j)$ retourne l'ensemble des prédecesseurs du noeud j dans le graphe : $Preds(j) = \{i \in V | (i, j) \in E\}$;
- Une fonction $Succs : V \rightarrow 2^V$ telle que $Succs(i)$ retourne l'ensemble des successeurs du noeud i dans le graphe : $Succs(i) = \{j \in V | (i, j) \in E\}$;
- Une fonction $I^D : \{1..T\} \rightarrow 2^V$ telle que $I^D(t)$ retourne l'ensemble des noeuds infectés avant l'iteration t de la diffusion D;
- Une fonction $t^D : V \rightarrow \{1..T\} + \infty$ telle que $t^D(v)$ retourne le temps d'infection du noeud v dans l'episode de diffusion D (∞ si pas infecté dans D);
- Un ensemble des poids d'influence θ définis pour tous les couples $(i, j) \in E$, avec $\theta_{i,j} \in \mathbb{R}$ le poids d'influence du noeud i sur le noeud j;
- Un ensemble de seuils d'infection γ définis pour tous les noeuds de V, avec $\gamma_i \in \mathbb{R}$ le seuil d'infection du noeud i.

Un noeud j est alors infecté par la diffusion D à l'iteration t si:

$$\sum_{i \in Preds(j) \cap I^D(t)} \theta_{i,j} \geq \gamma_j$$

$\Rightarrow$ Formaliser le problème d'apprentissage de paramètres θ et γ efficaces pour la prédiction d'infections finales à partir d'une source sur un réseau G, à partir d'un ensemble d'episodes de diffusion observés $\mathcal{D} = \{D_1..D_n\}$

Linear Threshold Model

Formalisation du problème d'apprentissage de paramètres θ et γ efficaces pour la prédiction d'infections finales à partir d'une source sur un réseau G, à partir d'un ensemble d'episodes de diffusion observés $\mathcal{D} = \{D_1..D_n\}$

- Une première proposition:

$$(\theta^*, \gamma^*) = \arg\min_{\theta,\gamma} \sum_{D\in\mathcal{D}} \sum_{v\in I^D(\infty)} max(0, 1 + \gamma_v - \sum_{u\in I^D(t^D(v))\cap Preds(v)} \theta_{u,v})$$
$$max(0, 1 - \gamma_v + \sum_{u\in I^D(t^D(v))\cap Preds(v)\setminus Last^D(v,t^D(v))} \theta_{u,v})$$
$$\sum_{v\notin I^D(\infty)} max(0, 1 - \gamma_v + \sum_{u\in I^D(\infty)\cap Preds(v)} \theta_{u,v})$$

$$Avec\ Last^D(v,t) = \arg\max_{u'\in I^D(t)\cap Preds(v)} t^D(u')$$

Linear Threshold Model

Formalisation du problème d'apprentissage de paramètres θ et γ efficaces pour la prédiction d'infections finales à partir d'une source sur un réseau G, à partir d'un ensemble d'episodes de diffusion observés $\mathcal{D} = \{D_1..D_n\}$

- Une première proposition:

$$(\theta^*, \gamma^*) = \arg\min_{\theta,\gamma} \sum_{D\in\mathcal{D}} \sum_{v\in I^D(\infty)} max(0, 1 + \gamma_v - \sum_{u\in I^D(t^D(v))\cap Preds(v)} \theta_{u,v})$$
$$max(0, 1 - \gamma_v + \sum_{u\in I^D(t^D(v))\cap Preds(v)\setminus Last^D(v,t^D(v))} \theta_{u,v})$$
$$\sum_{v\notin I^D(\infty)} max(0, 1 - \gamma_v + \sum_{u\in I^D(\infty)\cap Preds(v)} \theta_{u,v})$$

$$Avec\ Last^D(v,t) = \arg\max_{u'\in I^D(t)\cap Preds(v)} t^D(u')$$

$\Rightarrow$ Mais les infections sont généralement des évènements rares

Linear Threshold Model

Formalisation du problème d'apprentissage de paramètres θ et γ efficaces pour la prédiction d'infections finales à partir d'une source sur un réseau G, à partir d'un ensemble d'episodes de diffusion observés $\mathcal{D} = \{D_1..D_n\}$

- Une première proposition:

$$(\theta^*, \gamma^*) = \arg\min_{\theta,\gamma} \sum_{D\in\mathcal{D}} \sum_{v\in I^D(\infty)} max(0, 1 + \gamma_v - \sum_{u\in I^D(t^D(v))\cap Preds(v)} \theta_{u,v})$$
$$max(0, 1 - \gamma_v + \sum_{u\in I^D(t^D(v))\cap Preds(v)\setminus Last^D(v,t^D(v))} \theta_{u,v})$$
$$\sum_{v\notin I^D(\infty)} max(0, 1 - \gamma_v + \sum_{u\in I^D(\infty)\cap Preds(v)} \theta_{u,v})$$

$$Avec\ Last^D(v,t) = \arg\max_{u'\in I^D(t)\cap Preds(v)} t^D(u')$$

$\Rightarrow$ Mais les infections sont généralement des évènements rares

$\Rightarrow$ Risque que quasi aucun utilisateur ne soit jamais infecté lors des simulations !

Linear Threshold Model

Formalisation du problème d'apprentissage de paramètres θ et γ efficaces pour la prédiction d'infections finales à partir d'une source sur un réseau G, à partir d'un ensemble d'episodes de diffusion observés $\mathcal{D} = \{D_1..D_n\}$

- Version probabiliste:

$$(\theta^*, \gamma^*) = \arg\min_{\theta,\gamma} \sum_{D \in \mathcal{D}} \sum_{v \in I^D(\infty)} log(P(v|I^D(t^D(v)) \cap Preds(v))) +$$
$$log(1 - P(v|I^D(t^D(v)) \cap Preds(v) \setminus Last^D(v, t^D(v)))) +$$
$$\sum_{v \notin I^D(\infty)} log(1 - P(v|I^D(\infty) \cap Preds(v)))$$

$$Avec\ Last^D(v,t) = \arg\max_{u' \in I^D(t) \cap Preds(v)} t^D(u')$$

$$Et\ P(v|U) = \frac{1}{1 + exp(\gamma_v - \sum_{u \in U} \theta_{u,v})}$$

Linear Threshold Model

Formalisation du problème d'apprentissage de paramètres θ et γ efficaces pour la prédiction d'infections finales à partir d'une source sur un réseau G, à partir d'un ensemble d'episodes de diffusion observés $\mathcal{D} = \{D_1..D_n\}$

- Version probabiliste:

$$(\theta^*, \gamma^*) = \arg\min_{\theta,\gamma} \sum_{D \in \mathcal{D}} \sum_{v \in I^D(\infty)} log(P(v|I^D(t^D(v)) \cap Preds(v))) +$$
$$log(1 - P(v|I^D(t^D(v)) \cap Preds(v) \setminus Last^D(v, t^D(v)))) +$$
$$\sum_{v \notin I^D(\infty)} log(1 - P(v|I^D(\infty) \cap Preds(v)))$$

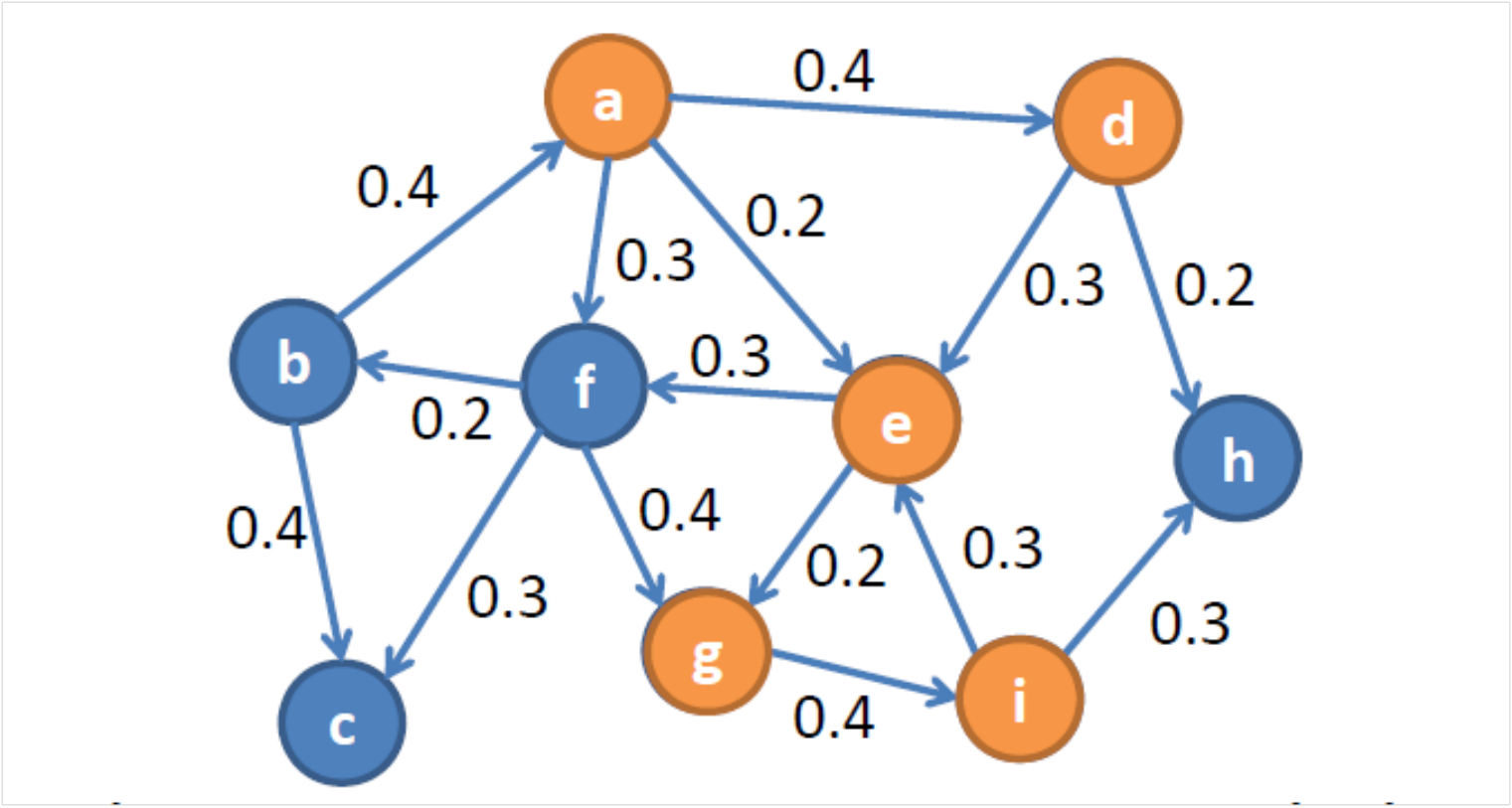
$$Avec\ Last^D(v,t) = \arg\max_{u' \in I^D(t) \cap Preds(v)} t^D(u')$$

$$Et\ P(v|U) = \frac{1}{1 + exp(\gamma_v - \sum_{u \in U} \theta_{u,v})}$$

$\Rightarrow$ Optimisation par montée de gradient stochastique

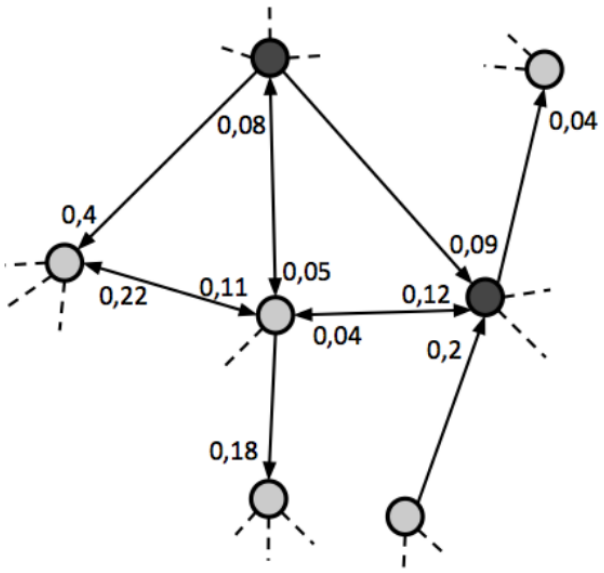
Modélisation de Diffusion sur les réseaux

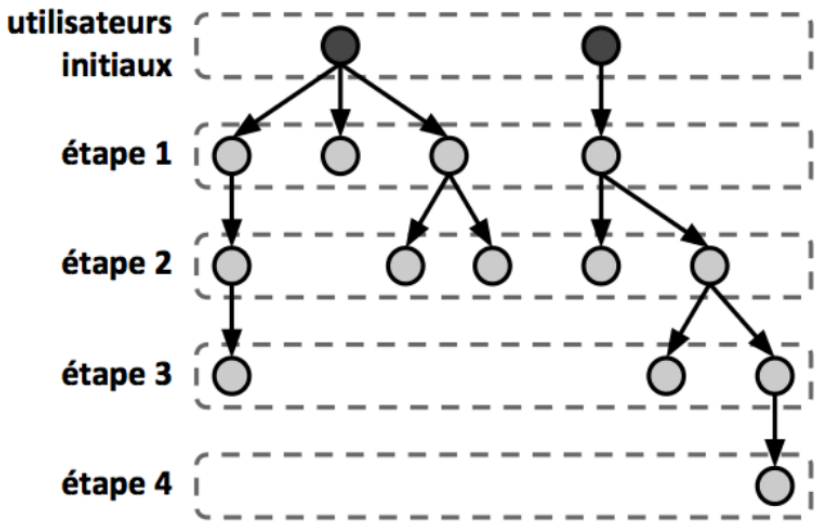
- Modèle Independent Cascade (IC)
  - Hypothèse: chaque influence $u \rightarrow v$ est indépendante
  - Après avoir été infecté, un utilisateur u a une chance unique d'infecter chacun de ses successeurs non encore infectés à l'étape suivante du processus
  - Probabilités d'infection $\theta_{u,v}$ sur les arcs du graphe



Modélisation de Diffusion sur les réseaux

- Modèle Independent Cascade (IC)
  - Inférence à partir d'un graphe d'influence (selon les probabilités $\theta_{u,v}$ définies sur les arcs)





- Probabilité d'être infecté à l'etape t = Probabilité d'être influencé par au moins un de ses prédecesseurs infectés à l'étape $t-1$:

$$P_t(v) = 1 - \prod_{u \in Preds(v) \wedge t_u = t-1} 1 - \theta_{u,v}$$

Modélisation de le Diffusion sur les réseaux

- Modèle Independent Cascade (IC)
  - Extraction d'influences = Apprentissage des $\theta_{u,v}$ qui maximisent la vraissemblance du modèle selon un ensemble d'episodes de diffusion observés $\mathcal{D}$

[Saito et al., 2008] :

$$L(\mathcal{D};\theta) = \prod_{D\in\mathcal{D}} \prod_{u\in D} P_{t_u^D}(u) \prod_{\substack{(u,v),u\in D\wedge v\in Succs(u)\wedge \\ ((v\notin D)\vee(v\in D\wedge t_v^D>t_u^D+1))}} 1-\theta_{u,v}$$

Avec $P_{t_u^D}(u) = 1 - \prod_{v\in Preds(u)\wedge t_v^D=t_u^D-1} 1-\theta_{v,u}$

Ou de manière équivalente :

$$\log\left(L(\mathcal{D};\theta)\right) = \sum_{D\in\mathcal{D}} \sum_{u\in D} \log P_{t_u^D}(u) + \sum_{\substack{(u,v),u\in D\wedge v\in Succs(u)\wedge \\ ((v\notin D)\vee(v\in D\wedge t_v^D>t_u^D+1))}} \log\left(1-\theta_{u,v}\right)$$

$\Rightarrow$ Difficile à maximiser tel quel

$\Rightarrow$ Passage par un algorithme EM

Algorithme Expectation-Maximization (EM)

- L'algorithme EM a été initiallement proposé par [Dempster et al., 1977]
  - $\Rightarrow$ Maximiser la vraisemblance de paramètres θ de modèles probabilistes lorsque le modèle dépend de variables latentes Z non observables

$$L(\theta; \mathbf{X}) = p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)$$

    - Vraissemblance possiblement difficile à maximiser directement
      - Réalisations de $\mathbf{Z}$ trop nombreuses
      - $p(\mathbf{X}, \mathbf{Z}|\theta)$ difficile à caractériser / optimiser
  - $\Rightarrow$ Optimisation itérative en s'appuyant sur les paramètres actuels $\theta^{(t)}$ pour définir les probabilités des réalisations de $\mathbf{Z}$

Algorithme Expectation-Maximization (EM)

Quel que soit $\mathbf{Z}$, on a:

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) - \log p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$$

On peut aussi considérer toutes les réalisations de $\mathbf{Z}$ selon une distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(t)})$:

$$\begin{aligned}\log p(\mathbf{X}|\boldsymbol{\theta}) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(t)}) \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \\ &\quad - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(t)}) \log p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \\ &= Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) + H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\end{aligned}$$

Si on s'intéresse à l'augmentation de la vraissemblance par rapport à la vraissemblance actuelle, on a:

$$\log p(\mathbf{X}|\boldsymbol{\theta}) - \log p(\mathbf{X}|\boldsymbol{\theta}^{(t)}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) + H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$$

Or, selon l'inégalité de Gibbs: $-\sum_{i=1}^{n} p_i log(q_i) \geq -\sum_{i=1}^{n} p_i log(p_i)$
On a donc $H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \geq H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$ et donc:

$$\log p(\mathbf{X}|\boldsymbol{\theta}) - \log p(\mathbf{X}|\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$$

Algorithme Expectation-Maximization (EM)

Quel que soit $\mathbf{Z}$, on a:

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) - \log p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$$

On peut aussi considérer toutes les réalisations de $\mathbf{Z}$ selon une distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(t)})$:

$$\begin{aligned}\log p(\mathbf{X}|\boldsymbol{\theta}) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(t)}) \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \\ &\quad - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(t)}) \log p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \\ &= Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) + H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\end{aligned}$$

Si on s'intéresse à l'augmentation de la vraissemblance par rapport à la vraissemblance actuelle, on a:

$$\log p(\mathbf{X}|\boldsymbol{\theta}) - \log p(\mathbf{X}|\boldsymbol{\theta}^{(t)}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) + H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$$

Or, selon l'inégalité de Gibbs: $-\sum_{i=1}^{n} p_i log(q_i) \geq -\sum_{i=1}^{n} p_i log(p_i)$
On a donc $H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \geq H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$ et donc:

$$\log p(\mathbf{X}|\boldsymbol{\theta}) - \log p(\mathbf{X}|\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$$

$\Rightarrow$ On en déduit donc qu'augmenter $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ par rapport à $Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$ améliore d'au moins autant $p(\mathbf{X}|\boldsymbol{\theta})$ par rapport à $p(\mathbf{X}|\boldsymbol{\theta}^{(t)})$

Algorithme Expectation-Maximization (EM)

L'algorithme EM itère sur les deux étapes suivantes tant que l'on a pas atteint un etat stable :

1. Une étape d'évaluation de l'espérance (E), où l'on définit l'espérance de la vraisemblance en tenant compte des paramètres courants θ^t

$$Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right) = E_{\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(t)}}\left[L\left((\mathbf{X},\mathbf{Z});\boldsymbol{\theta}\right)\right)|\boldsymbol{\theta}^{(t)}\right]$$

2. Une étape de maximisation (M), où l'on estime de nouveaux paramètres θ^{t+1} maximisant l'espérance de vraissemblance définie à l'étape E

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}}\left(Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(\boldsymbol{t})}\right)\right)$$

- Apprentissage des $\theta_{u,v}$ de IC par EM:
  1. $\hat{\theta} = Random$
  2. Tant qu'on n'a pas atteint des paramètres stables (ou que la vraissemblance augmente) :
     1. Pour tout $D \in \mathcal{D}$: calculer la probabilité de tout $u \in D$ d'être infecté au temps t_u^D selon les paramètres courants $\hat{\theta}$:

$$\hat{P}_{t_u^D}(u) = 1 - \prod_{v \in Preds(u) \wedge t_v^D = t_u^D - 1} 1 - \hat{\theta}_{v,u}$$

     2. On pose l'espérance de vraissemblance :

$$Q\left(\theta; \hat{\theta}\right) = \sum_{D \in \mathcal{D}} \Phi^D\left(\theta; \hat{\theta}\right) + \sum_{\substack{(u,v), u \in D \wedge v \in Succs(u) \wedge \\ ((v \notin D) \vee (v \in D \wedge t_v^D > t_u^D + 1))}} \log\left(1 - \theta_{u,v}\right)$$

Avec $\Phi^D\left(\theta; \hat{\theta}\right) =$

$$\sum_{\substack{(u,v) \in D^2, v \in Succs(u) \\ \wedge t_v^D = t_u^D + 1}} \frac{\hat{\theta}_{u,v}}{\hat{P}_{t_v^D}(v)} \log(\theta_{u,v}) + (1 - \frac{\hat{\theta}_{u,v}}{\hat{P}_{t_v^D}(v)}) \log(1 - \theta_{u,v})$$

     3. On maximise :

$$\theta^* = \arg\max_{\theta} Q\left(\theta; \hat{\theta}\right)$$

     4. $\hat{\theta} = \theta^*$

- Apprentissage des $\theta_{u,v}$ de IC par EM:
  - Résolution de l'étape de maximisation ?

- Apprentissage des $\theta_{u,v}$ de IC par EM:
  - Résolution de l'étape de maximisation ?

$$\Rightarrow \text{ Annuler } \frac{\partial Q\left(\theta;\hat{\theta}\right)}{\partial \theta}$$

- Apprentissage des $\theta_{u,v}$ de IC par EM:
  - Résolution de l'étape de maximisation ?

$$\Rightarrow \text{ Annuler } \frac{\partial Q\left(\theta;\hat{\theta}\right)}{\partial \theta}$$

$$\Rightarrow \theta^*_{u,v} = \frac{\sum_{D \in \mathcal{D}^+_{u,v}} \frac{\hat{\theta}_{u,v}}{\hat{P}_{t^D_v}(v)}}{|\mathcal{D}^+_{u,v}| \ + \ |\mathcal{D}^-_{u,v}|}$$

Avec :

$$\mathcal{D}^+_{u,v} = \{D \in \mathcal{D} | (u,v) \in D^2 \wedge t^D_v = t^D_u + 1\}$$

$$\mathcal{D}^-_{u,v} = \{D \in \mathcal{D} | u \in D \wedge ((v \notin D) \vee (v \in D \wedge t^D_v > t^D_u + 1))\}$$

- Limites de IC:
  - Hypothèse de probabilités indépendantes
  - Hypothèse de monde clos
  - Hypothèse que les liens du graphe correspondent bien aux cannaux de diffusion
  - Pas de prise en compte du contenu
  - Discretisation du temps

- Limites de IC:
  - Hypothèse de probabilités indépendantes
  - Hypothèse de monde clos
  - Hypothèse que les liens du graphe correspondent bien aux cannaux de diffusion
  - Pas de prise en compte du contenu
  - **Discretisation du temps**

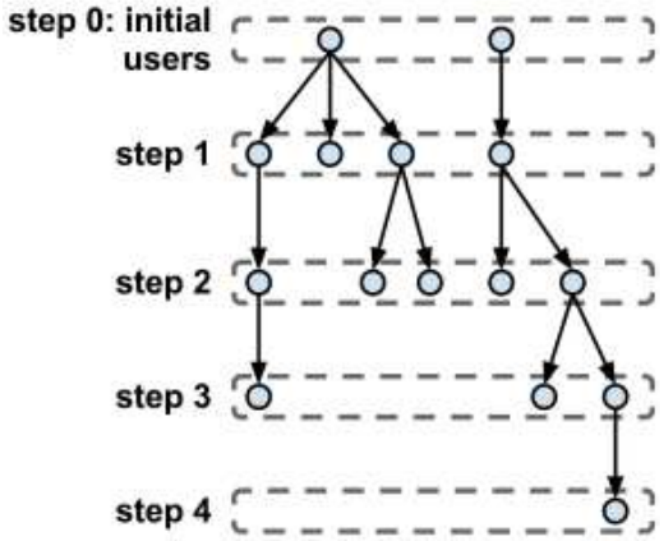
- Limites de IC:
  - Hypothèse de probabilités indépendantes
  - Hypothèse de monde clos
  - Hypothèse que les liens du graphe correspondent bien aux cannaux de diffusion
  - Pas de prise en compte du contenu
  - **Discretisation du temps**
    - Regroupement des évènements par pas de temps
    - Infection de v par u uniquement possible au pas de temps $t_u^D + 1$

- Limites de IC:
  - Hypothèse de probabilités indépendantes
  - Hypothèse de monde clos
  - Hypothèse que les liens du graphe correspondent bien aux cannaux de diffusion
  - Pas de prise en compte du contenu
  - **Discretisation du temps**
    - Regroupement des évènements par pas de temps
    - Infection de v par u uniquement possible au pas de temps $t_u^D + 1$

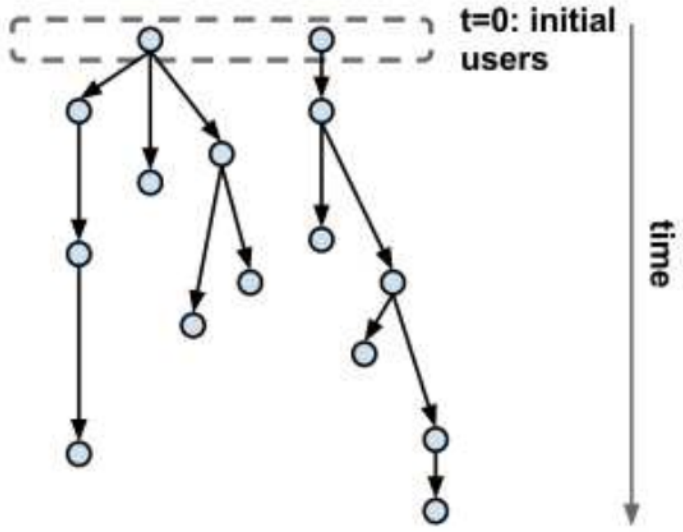
    $\Rightarrow$ Si pas de temps trop long : trop d'évènements dans le même pas de temps

    $\Rightarrow$ Si pas de temps trop court : beaucoup de "trous" dans le processus de diffusion

Modélisation de le Diffusion: Temps continu



Modèle Independent Cascade



Processus de diffusion plus réaliste

- Deux principales variantes d'IC proposent de prendre en compte des délais continus d'infection:
  - NetRate [Gomez-Rodriguez et al., 2011]
  - CTIC [Saito et al., 2009]

Modélisation de Diffusion: Temps continu

NetRate [Gomez-Rodriguez et al., 2011]

- Définition de distributions de probabilités d'influence décroissantes avec le temps
  - Distributions conditionnelles de transmission $f(t_i|t_j;\theta_{j,i})$ Exponentielle, Puissance, Raighley
  - *e.g.*, Distribution exponentielle: $f(t_i|t_j;\theta_{j,i}) = \theta_{j,i} \exp^{-\theta_{j,i}(t_i - t_j)}$
- Vraissemblance d'un épisode de diffusion (avec longueur max diffusion $= T$ et $t_x = \infty$ si x pas infecté): $L(D;\theta) =$

$$\prod_{t_i \leq T} \prod_{\substack{j \in Succs(i), \\ t_j > T}} S(t_j|t_i;\theta_{i,j}) \times \prod_{\substack{k \in Preds(i), \\ t_k < t_i}} S(t_i|t_k;\theta_{k,i}) \times \sum_{\substack{j \in Preds(i), \\ t_j < t_i}} \frac{f(t_i|t_j;\theta_{j,i})}{S(t_i|t_j;\theta_{j,i})}$$

Avec $S(t|t_j;\theta_{j,i})$ la probabilité que j n'ait pas encore infecté i au temps t :

$$S(t|tj;\theta_{j,i}) = 1 - \int_{t_j}^{t} f(x|t_j;\theta_{j,i}) \partial x$$

- $\theta^* = \arg\max_{\theta} \prod_{D \in \mathcal{D}} L(D;\theta)$
- Passage au log $\Rightarrow$ Problème d'optimisation convexe

Modélisation de Diffusion: Temps continu

Continuous-Time Independent Cascade (CTIC) [Saito et al., 2009]

- Contrairement à NetRate : 2 types de paramètres pour tous les couples (u, v)
  - Probabilités d'influence k
  - Paramètres de distribution de delais de transmission r
- Densité de probabilité que u infecte v au temps t_v^D pour l'episode D:

$$A_{D,u,v} = k_{u,v} r_{u,v} \exp^{-r_{u,v}(t_v^D - t_u^D)}$$

- Probabilité que u n'ait pas encore infecté par v au temps t_v^D pour l'episode D:

$$B_{D,u,v} = 1 - k_{u,v} \int_{t_u^D}^{t_v^D} r_{u,v} \exp^{-r_{u,v}(t - t_u^D)} \, \partial t$$

- Vraissemblance $L(\mathcal{D}; k, r) =$

$$\prod_{D \in \mathcal{D}} \prod_{v \in D} \prod_{\substack{x \in Preds(v), \\ t_x^D < t_v^D}} B_{D,x,v} \sum_{\substack{u \in Preds(v), \\ t_u^D < t_v^D}} \frac{A_{D,u,v}}{B_{D,u,v}} \prod_{\substack{w \in Succs(v), \\ w \notin D}} (1 - k_{v,w})$$

- Passage au log puis optimisation par EM

Modélisation de Diffusion: Temps continu

- Modélisation de la diffusion par CTIC
  - Efficace lorsque des régularités existent sur les delais de transmission mais...
  - Ces régularités sont rarement observées sur les données réelles

  $\Rightarrow$ Les délais variables peuvent gêner l'extraction d'influences
- Proposition de relaxation de IC: Delay-Agnostic IC [Lamprier et al., 2015]
  - Pas de discretisation du temps
  - Délais d'infection uniformes
  - A chaque etape du EM de IC :

$$\theta^*_{u,v} = \frac{\sum_{D \in \mathcal{D}^+_{u,v}} \frac{\hat{\theta}_{u,v}}{\hat{P}_{t^D_v}(v)}}{|\mathcal{D}^+_{u,v}| + |\mathcal{D}^-_{u,v}|}$$

Avec :
$$\hat{P}_{t^D_u}(u) = 1 - \prod_{v \in Preds(u) \wedge t^D_v < t^D_u} 1 - \hat{\theta}_{v,u}$$

Et :
$$\mathcal{D}^+_{u,v} = \{D \in \mathcal{D} | (u,v) \in D^2 \wedge t^D_v > t^D_u\}$$
$$\mathcal{D}^-_{u,v} = \{D \in \mathcal{D} | u \in D \wedge v \notin D\}$$

\+ D'avantage d'exemples d'apprentissage
\+ Modèle plus réaliste (fonctionne au moins aussi bien que CTIC sur des données réelles)
\+ Modèle bien plus simple que CTIC
\- Temps d'infection ne peuvent pas être prédits

Modélisation de Diffusion: Graphe ou pas Graphe ?

- La plupart des modèles utilisent un graphe de relations explicites
  - Et si pas de relations connues ?
  - Et si relations connues pas représentatives des cannaux de diffusion ? [Ver Steeg et al., 2013]
- Approches de détection de liens de diffusion : e.g., NetInf [Gomez Rodriguez et al., 2010]
  - Recherche de l'arbre couvrant maximum pour chaque cascade
  - Conservation des *n* liens les plus utilisés par les arbres produits
- Mais en pratique la plupart des modèles fonctionnent très bien en considérant le graphe complet des relations possibles
  - On peut se limiter aux liens avec au moins un exemple de possible diffusion dans l'ensemble d'entraînement

Modélisation de Diffusion: Modèles non itératifs

- Modèles itératifs plus précis pour décrire les processus de diffusion mais...
- ... Peu robustes aux modifications du réseau [Najar et al., 2012]
- ... Lourds à apprendre lorsqu'il s'agit de considérer de grandes masses de données
- ... Risque de sur-apprentissage
- ... Généralement inférence complexe (Monte-carlo simulations)

$\Rightarrow$ Certains modèles proposent de s'abstraire de la modélisation du processus itératif de diffusion

- Focus sur les correlations entre les sources de la diffusion et l'etat de contamination final
- $C^D = f_\theta(S^D)$, avec S^D et C^D respectivement l'ensemble des sources et l'etat final de la contamination de l'episode D

Modélisation de Diffusion: Modèles non itératifs

- Modèle naïf avec source unique
  - Une seule source s^D par episode D
  - Vraissemblance :

$$L(\mathcal{D};\theta) = \prod_{D\in\mathcal{D}} \prod_{u\in D} \theta_{s^D,u} \prod_{u\notin D} (1-\theta_{s^D,u})$$

  - Passage au log:

$$L(\mathcal{D};\theta) = \sum_{D\in\mathcal{D}} \sum_{u\in D} \log(\theta_{s^D,u}) + \sum_{u\notin D} \log(1-\theta_{s^D,u})$$

Modélisation de Diffusion: Modèles non itératifs

- Modèle naïf avec source unique
  - Une seule source s^D par episode D
  - Vraissemblance :

$$L(\mathcal{D};\theta) = \prod_{D\in\mathcal{D}} \prod_{u\in D} \theta_{s^D,u} \prod_{u\notin D} (1-\theta_{s^D,u})$$

  - Passage au log:

$$L(\mathcal{D};\theta) = \sum_{D\in\mathcal{D}} \sum_{u\in D} \log(\theta_{s^D,u}) + \sum_{u\notin D} \log(1-\theta_{s^D,u})$$

$$\theta^*_{u,v} = \frac{|\mathcal{D}^+_{u,v}|}{|\mathcal{D}^+_{u,v}| \ + \ |\mathcal{D}^-_{u,v}|}$$

Avec :

$$\mathcal{D}^+_{u,v} = \{D \in \mathcal{D} | u = s^D \wedge v \in D\}$$
$$\mathcal{D}^-_{u,v} = \{D \in \mathcal{D} | u = s^D \wedge v \notin D\}$$

Modélisation de Diffusion: Modèles non itératifs

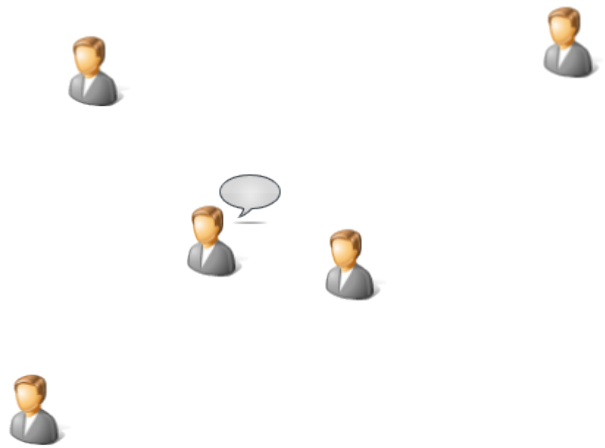
- Modèle discriminant à partir d'un ensemble de sources
  - Entrée = Vecteur binaire initial $S^D \in \{0;1\}^{|U|}$, où $S_i^D = 1$ indique que l'utilisateur i est source de l'episode D
  - Sortie = Vecteur binaire de contaminations $C^D \in \{0;1\}^{|U|}$, où $C_i^D = 1$ indique que l'utilisateur i est contaminé par l'episode D
- Regression Logistique

$$\theta^* = \arg\max_{\theta} \sum_{D \in \mathcal{D}} \sum_{i \in U} C_i^D \log\left(\frac{1}{1 + e^{-f_\theta(i, S^D)}}\right) + (1 - C_i^D) \log\left(1 - \frac{1}{1 + e^{-f_\theta(i, S^D)}}\right)$$

- Diverses fonctions f envisageables (produit scalaire, réseau de neurone, etc...)
- Normalisation éventuelle du vecteur d'entrée S^D

Modélisation de Diffusion: Modèles non itératifs

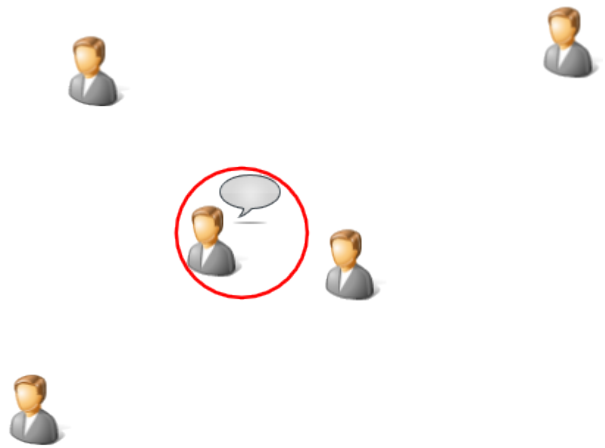
- Projection dans un espace latent [Bourigault et al., 2014]
  - Les utilisateurs sont projetés dans un espace continu
  - La diffusion peut être modélisée comme un processus de diffusion de chaleur dans cet espace



- La chaleur part de la source
- La température de l'utilisateur u_i au temps t, $T(u_i, t)$ = propension à être infecté.

Modélisation de Diffusion: Modèles non itératifs

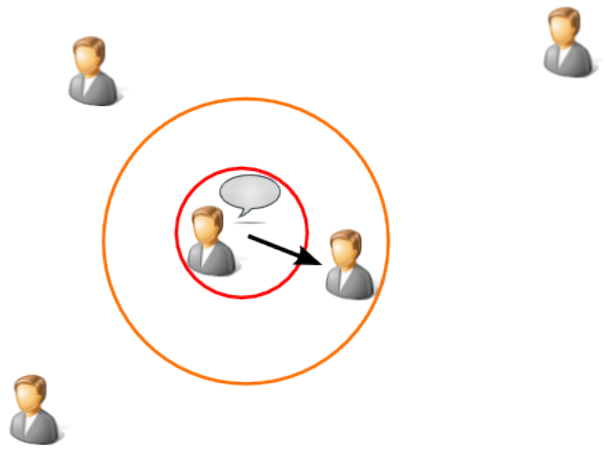
- Projection dans un espace latent [Bourigault et al., 2014]
  - Les utilisateurs sont projetés dans un espace continu
  - La diffusion peut être modélisée comme un processus de diffusion de chaleur dans cet espace



- La chaleur part de la source
- La température de l'utilisateur u_i au temps t, $T(u_i, t)$ = propension à être infecté.

Modélisation de Diffusion: Modèles non itératifs

- Projection dans un espace latent [Bourigault et al., 2014]
  - Les utilisateurs sont projetés dans un espace continu
  - La diffusion peut être modélisée comme un processus de diffusion de chaleur dans cet espace



Equation de la chaleur :

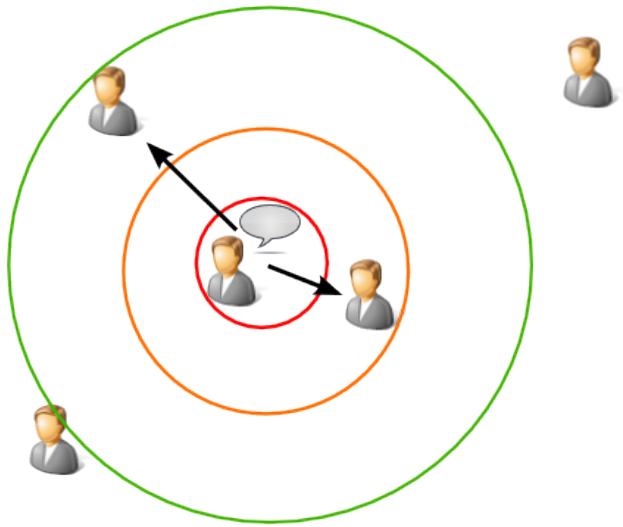
$$\begin{cases} \frac{\partial T}{\partial t} = \Delta_x T \\ f(x,0) = f_0(x) \end{cases}$$

Solution lorsque la source est en x_0 :

$$T_{x_0}(x,t) = (4\pi t)^{-\frac{n}{2}} e^{-\frac{||x_0 - x||^2}{4t}}$$

Modélisation de Diffusion: Modèles non itératifs

- Projection dans un espace latent [Bourigault et al., 2014]
  - Les utilisateurs sont projetés dans un espace continu
  - La diffusion peut être modélisée comme un processus de diffusion de chaleur dans cet espace



Equation de la chaleur :

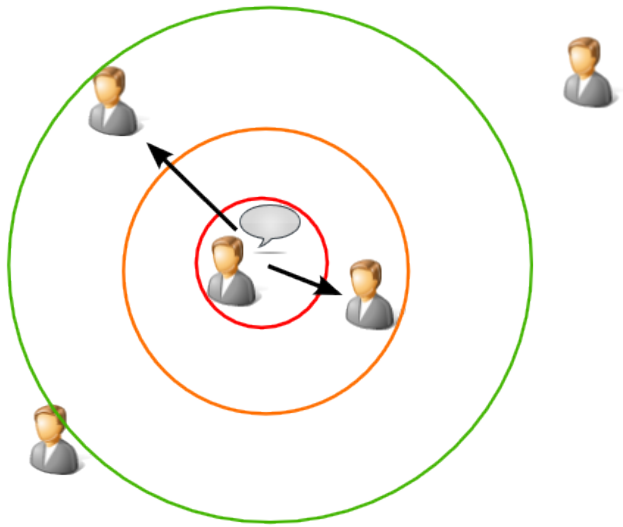
$$\begin{cases} \frac{\partial T}{\partial t} = \Delta_x T \\ f(x,0) = f_0(x) \end{cases}$$

Solution lorsque la source est en x_0 :

$$T_{x_0}(x,t) = (4\pi t)^{-\frac{n}{2}} e^{-\frac{||x_0 - x||^2}{4t}}$$

Modélisation de Diffusion: Modèles non itératifs

- Projection dans un espace latent [Bourigault et al., 2014]
  - Les utilisateurs sont projetés dans un espace continu
  - La diffusion peut être modélisée comme un processus de diffusion de chaleur dans cet espace



→ Trouver une representation $\mathcal{Z}$ des utilisateurs permettant d'expliquer les cascades observées selon un noyau de chaleur se propageant à partir de la source

$$\mathcal{U} = (u_1, ..., u_N) \rightarrow \mathcal{Z} = (z_1, ..., z_N) \subset \mathbb{R}^D$$

Modélisation de Diffusion: Modèles non itératifs

- Projection dans un espace latent [Bourigault et al., 2014]
  - Pour un episode D dont la source est s^D, 2 constraintes définies sur T:
    - $\forall (u,v) \in D^2, t_u^D < t_v^D \Rightarrow \forall t\ T_{s^D}(u,t) > T_{s^D}(v,t)$
    - $\forall u \in D\ \forall v \in U, v \notin D \Rightarrow \forall t\ T_{s^D}(u,t) > T_{s^D}(v,t)$
  - Dans l'espace latent $\rightarrow$ contraintes géometriques :
    - $\forall (u,v) \in D^2, t_u^D < t_v^D \Rightarrow ||z_{s^D} - z_u|| < ||z_{s^D} - z_v||$
    - $\forall u \in D\ \forall v \in U, v \notin D \Rightarrow ||z_{s^D} - z_u|| < ||z_{s^D} - z_v||$

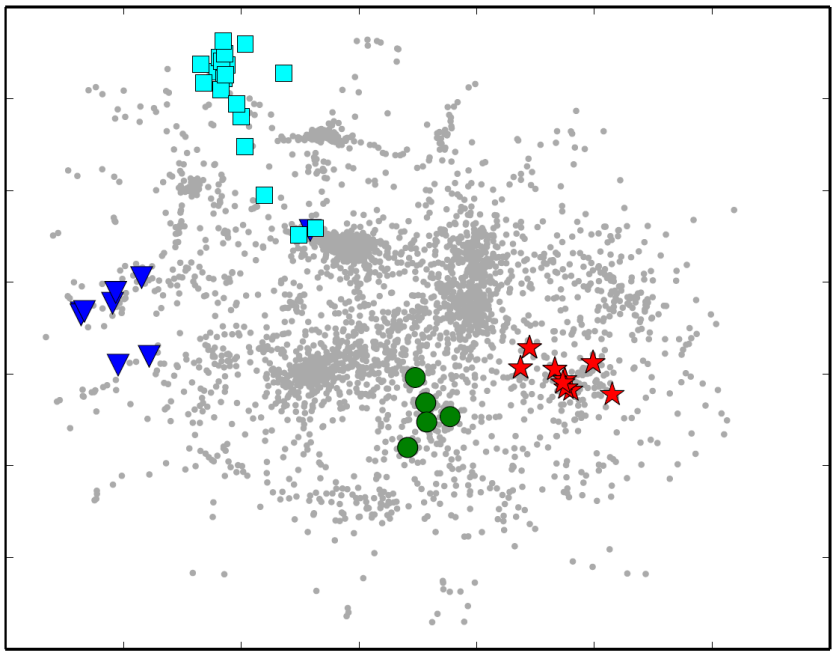
$\Rightarrow$ Fonction de coût :

$$\Delta_{rank}(\mathcal{Z}, \mathcal{D}) = \sum_{D \in \mathcal{D}} \sum_{\substack{u,v \\ t^c(u) < t^c(v)}} max(0, 1 - (||z_{s^D} - z_v||^2 - ||z_{s^D} - z_u||^2))$$
$$+ \sum_{u,v \in D \times \bar{D}} max(0, 1 - (||z_{s^D} - z_v||^2 - ||z_{s^D} - z_u||^2))$$

Optimisation par descente de gradient stochastique sur les représentations des utilisateurs $\mathcal{Z}$.

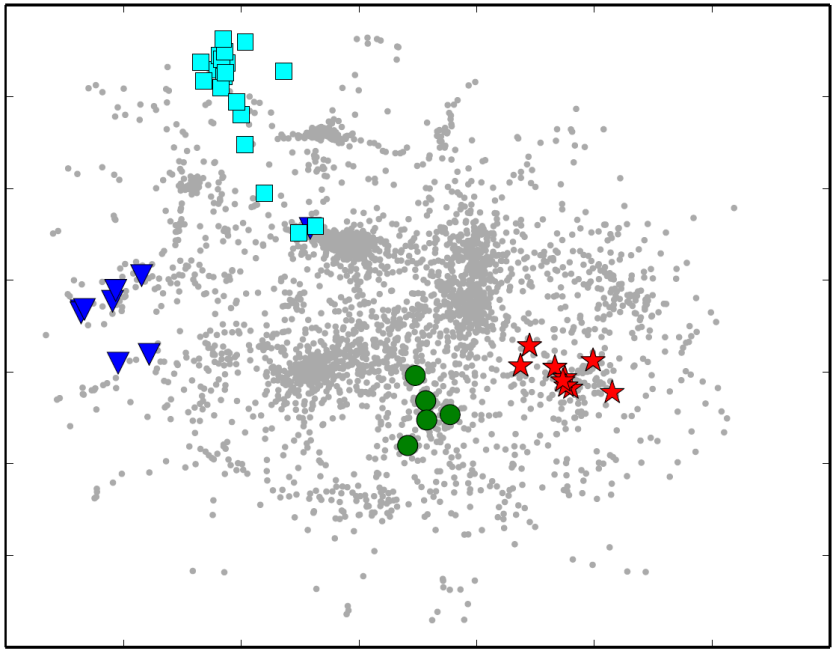
Modélisation de Diffusion: Modèles non itératifs

- Projection d'utilisateurs de **Digg** en 2 dimensions :



Modélisation de Diffusion: Modèles non itératifs

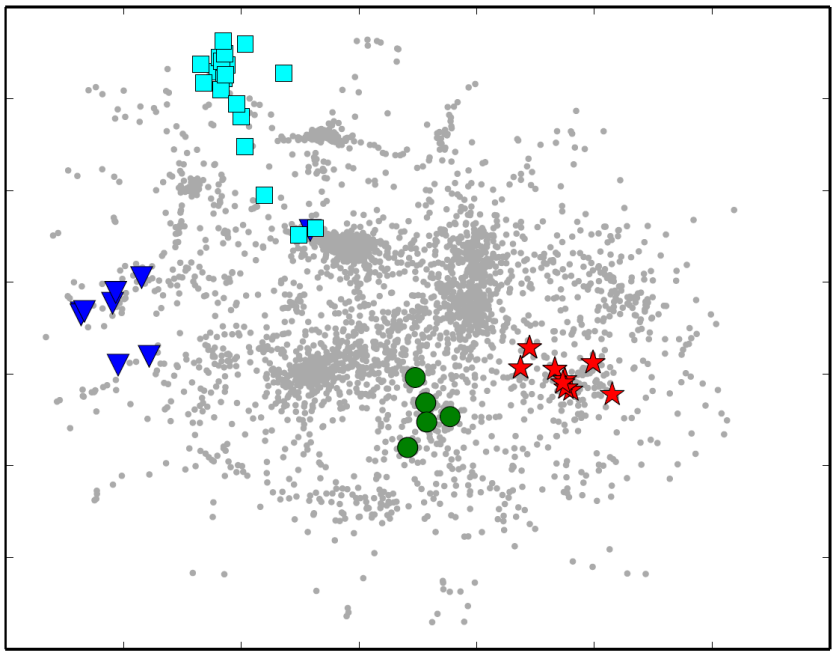
- Projection d'utilisateurs de **Digg** en 2 dimensions :



+ Capture de régularités entre les relations d'influence extraites $\Rightarrow$ meilleure généralisation

Modélisation de Diffusion: Modèles non itératifs

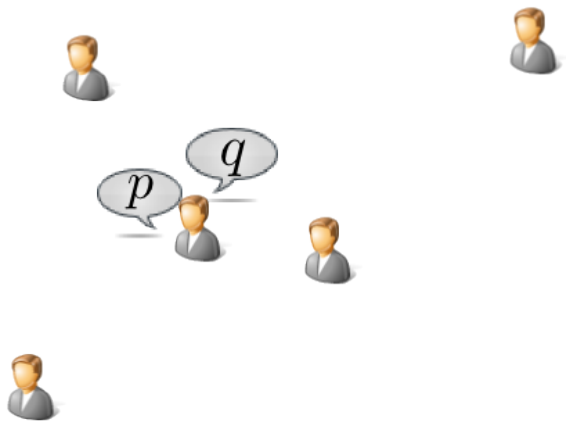
- Projection d'utilisateurs de **Digg** en 2 dimensions :



+ Capture de régularités entre les relations d'influence extraites $\Rightarrow$ meilleure généralisation
+ Possibilité de prendre en compte le contenu diffusé

Modélisation de Diffusion: Prise en compte du contenu

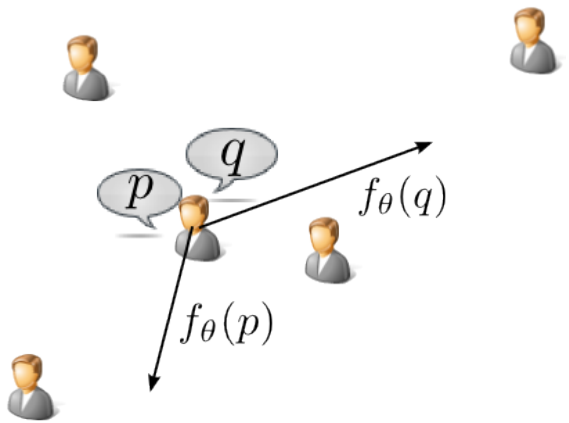
- Hypothèse : Le contenu diffusé influe sur les dynamiques de diffusion
- Possibilité d'intégrer le contenu diffusé dans le modèle de projection
  - Définition d'une fonction de translation de la source de chaleur f_θ en fonction du contenu



- Algorithme d'apprentissage similaire :
  - Apprentissage conjoint de la fonction de translation f_θ et des projections $\mathcal{Z}$
  - Optimisation par descente de gradient stochastique

Modélisation de Diffusion: Prise en compte du contenu

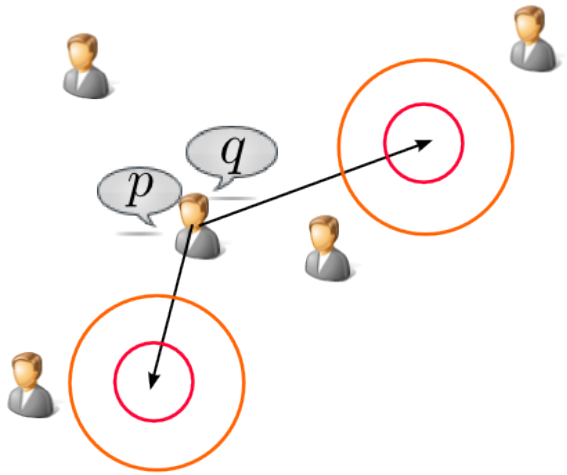
- Hypothèse : Le contenu diffusé influe sur les dynamiques de diffusion
- Possibilité d'intégrer le contenu diffusé dans le modèle de projection
  - Définition d'une fonction de translation de la source de chaleur f_θ en fonction du contenu



- Algorithme d'apprentissage similaire :
  - Apprentissage conjoint de la fonction de translation f_θ et des projections $\mathcal{Z}$
  - Optimisation par descente de gradient stochastique

Modélisation de Diffusion: Prise en compte du contenu

- Hypothèse : Le contenu diffusé influe sur les dynamiques de diffusion
- Possibilité d'intégrer le contenu diffusé dans le modèle de projection
  - Définition d'une fonction de translation de la source de chaleur f_θ en fonction du contenu



- Algorithme d'apprentissage similaire :
  - Apprentissage conjoint de la fonction de translation f_θ et des projections $\mathcal{Z}$
  - Optimisation par descente de gradient stochastique

Modélisation de Diffusion: Prise en compte de l'utilisateur

- Modèles itératifs
  - [Lagnier et al., 2013] Similaire à LT avec probabilité d'être contaminé au temps t dépendant de :
    - Nombre de prédecesseurs contaminés avant t
    - Similarité entre le profil de l'utilisateur et le contenu diffusé
    - Taux d'activité de l'utilisateur
  - [Saito et al., 2011] Similaire à CTIC en prenant en compte un vecteur d'attributs $x_{u,v}$ pour chaque lien
    - Probabilité de transmission $k = \frac{1}{1 + \exp^{-\theta . x_{u,v}}}$
    - Paramètre de delai $r = \exp^{\Theta . x_{u,v}}$
  - [Guille et al., 2012] Similaire à CTIC mais simple regression logistique sur probas paramètriques à partir d'exemples de diffusion. Prise en compte de :
    - Activité de la source et du destinataire du lien
    - La similarité sociale de la source et du destinataire du lien
    - Popularité de la source et du destinataire
    - Leur fréquence d'utilisation des mots-clé du contenu diffusé
    - L'heure/le jour de l'infection de la source
    - ...

Références

- [Bourigault et al., 2014] Simon Bourigault, Cédric Lagnier, Sylvain Lamprier, Ludovic Denoyer, Patrick Gallinari: Apprentissage de représentation pour la diffusion d'Information dans les réseaux sociaux. CORIA-CIFED 2014: 155-170
- [Dempster et al., 1977] A.P. Dempster, N.M. Laird et Donald Rubin, " Maximum Likelihood from Incomplete Data via the EM Algorithm ", Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, no 1, 1977, p. 1–38
- [Kempe et al., 2003] D. Kempe, J. Kleinberg, and E. Tardos, Maximizing the spread of influence in a social network, In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge
- [Gomez Rodriguez et al., 2010] Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause. 2010. Inferring networks of diffusion and influence. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10). ACM, New York, NY, USA, 1019–1028.
- [Gomez-Rodriguez et al., 2011] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the temporal dynamics of diffusion networks. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), ICML'11, pages 56–568. ACM, 2011.
- [Granovetter, 1973] Granovetter, M.S. The Strength of Weak Ties. The American Journal of Sociology 78 (6): 1360–1380. Discovery and Data Mining (KDD), 2003.
- [Guille et al., 2012] Adrien Guille, Hakim Hacid: A predictive model for the temporal dynamics of information diffusion in online social networks. WWW (Companion Volume) 2012: 1145-1152
- [Lagnier et al., 2013] Cédric Lagnier, Ludovic Denoyer, Éric Gaussier, Patrick Gallinari: Predicting Information Diffusion in Social Networks Using Content and User's Profiles. ECIR 2013: 74-85

Références

- [Lamprier et al., 2015] Sylvain Lamprier, Simon Bouriguault, Patrick Gallinari: Extracting Diffusion Channels from Real-World Social Data: a Delay-Agnostic Learning of Transmission Probabilities. ASONAM 2015
- [Najar et al., 2012] Anis Najar, Ludovic Denoyer, Patrick Gallinari: Predicting information diffusion on social networks with partial knowledge. WWW (Companion Volume) 2012: 1197-1204
- [Saito et al., 2008] K. Saito, R. Nakano, and M. Kimura. Prediction of information diffusion probabilities for independent cascade model. In Proceedings of the 12th international conference on Knowledge-Based Intelligent Information and Engineering Systems, Part III, KES '08, pages 67–75. Springer-Verlag, 2008.
- [Saito et al., 2009] K. Saito, M. Kimura, K. Ohara, and H. Motoda. Learning continuous-time information diffusion model for social behavioral data analysis. In Proceedings of the 1st Asian Conference on Machine Learning: Advances in Machine Learning, ACML '09, pages 322–337, Berlin, Heidelberg, 2009. Springer-Verlag.
- [Saito et al., 2011] Kazumi Saito, Kouzou Ohara, Yuki Yamagishi, Masahiro Kimura, and Hiroshi Motoda. Learning diffusion probability based on node attributes in social networks. In Marzena Kryszkiewicz, Henryk Rybinski, Andrzej Skowron, and Zbigniew W. Ras, editors, ISMIS, volume 6804 of Lecture Notes in Computer Science, pages 153–162. Springer, 2011.
- [Ver Steeg et al., 2013] G. Ver Steeg and A. Galstyan. Information-theoretic measures of influence based on content dynamics. In Proceedings of the sixth ACM international conference on Web search and data mining, WSDM '13, pages 3–12, New York, NY, USA, 2013. ACM.