

Fouilles de données et Medias sociaux

Master 2 DAC - FDMS

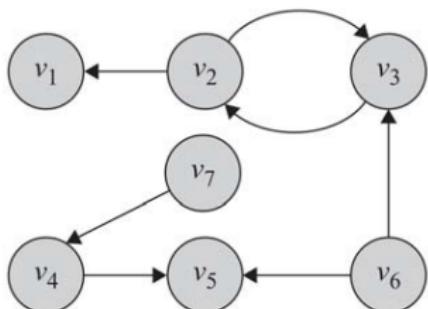
Sylvain Lamprier

UPMC

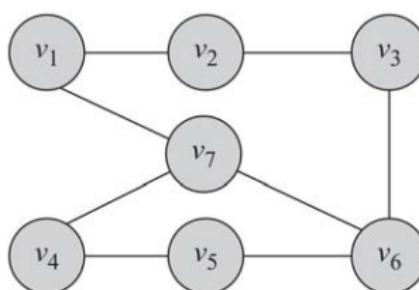
- Réseaux Sociaux
 - Différents types de réseaux
 - Métriques et analyse sur les réseaux
- Fouille de données dans les media sociaux
 - Internet et les Medias Sociaux
 - Nouveaux Challenges
- Applications
 - Détection de communautés
 - Recommendation Sociale
 - Classification collective
 - Diffusion d'information
 - Prédiciton à partir de flux

- Réseau social =
 - Ensemble d'entités
 - +
 - Ensemble de relations
- Entités :
 - Individus, Groupes, Sites, etc...
- Relations
 - Issues d'interactions sociales
 - e.g., amitié, influence, pouvoir, etc...

- Différents types de relations
 - Orientées : influence, pouvoir, etc...
 - Non-orientées : amitié, co-occurrence, etc...



(a) Directed Graph



(b) Undirected Graph

- Le plus grand des réseaux : La société
 - Noeuds = Individus
 - Liens = Relations sociales (famille, travail, amis, etc...)



⇒ S. Milgram (1967) : Six degrés de séparation entre les individus

Différents types de réseaux

- Différents modèles de réseaux:

- Random graphs (modèles Erdös-Rényi)
- Modèles Watts-Strogatz
- Réseaux Scale-free
 - Modèle de Barabási-Albert
 - Modèle Hiérarchique

⇒ Analyse de la structure des réseaux

- Différentes propriétés (degré du graphe, nombre de connexions, densité, diamètre, etc...)
- Étude / Simulation de phénomènes
- Entraînement de modèles (ex: Découverte de sous-structures/communautés)

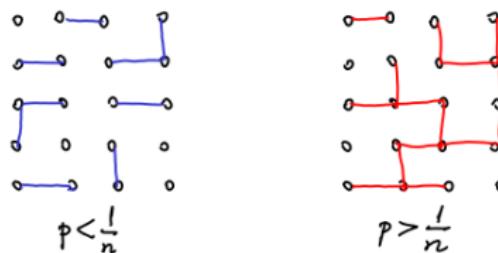
- Deux modèles de graphes non-orientés
 - Graphe aléatoire binomial
 - La présence de chacune des $n(n-1)/2$ arrêtes possibles du graphe est décidée aléatoirement selon une probabilité p .
 - Graphe aléatoire uniforme
 - m arrêtes sont choisies uniformément parmi les $n(n-1)/2$ arrêtes possibles du graphe.
- Modèles les plus étudiés mathématiquement
 - ⇒ Indépendance des arrêtes
- Lorsque $p \approx \frac{2m}{n(n-1)}$, les deux modèles ont des propriétés très proches.

⇒ Faible coefficient de clustering

- Clustering Global = $\frac{3 \times \text{nombre de triangles}}{\text{nombre de triplets connectés}}$
Où triangle est un sous-graphe complet à 3 noeuds et un triplet connecté est un sous-graphe connexe à trois noeuds.
 - ⇒ Probabilité pour un triplet de noeuds d'être entièrement connecté
 - ⇒ Représente le degré de transitivité des relations ("les amis de mes amis sont mes amis").
- Clustering Local Moyen = $\frac{\sum_i c_i}{|V|}$ où : $c_i = \frac{2|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i-1)}$
Où N_i est le voisinage du sommet i , de cardinal k_i .
 - ⇒ c_i : Probabilité pour deux sommets du voisinage de i soient eux même connectés
 - ⇒ Clustering Local Moyen tend à donner plus d'importance aux noeuds de petits degrés.

Random graphs (Modèles Erdös-Rényi)

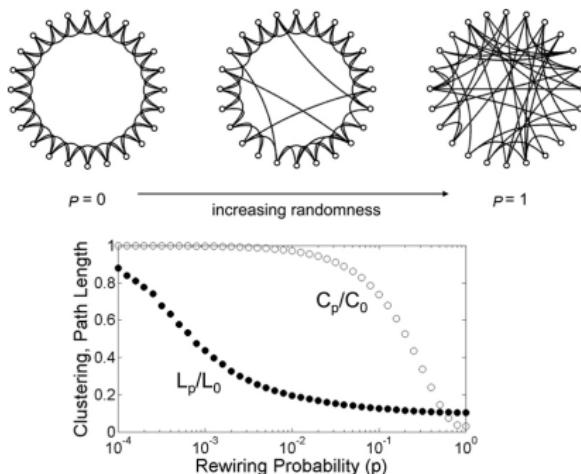
- ⇒ Distribution des degrés des noeuds concentrée autour de sa moyenne
- Converge vers une loi de Poisson, plutôt qu'une loi de puissance observée sur de nombreux réseaux réels



- Graphe non-orienté de degré K construit en 2 étapes:
 - ➊ Construction d'un réseau régulier en anneau: avec les N noeuds du graphe numérotés de n_0 à n_{N-1} , on établit un lien (n_i, n_j) ssi : $0 < |i - j| \bmod (n - \frac{K}{2}) \leq \frac{K}{2}$.
 - ➋ Chaque lien (n_i, n_j) avec $i < j$ est remplacé par (n_i, n_k) avec une probabilité p , où n_k est un noeud du graphe choisi uniformément (en évitant $k = i$ et en évitant de produire un lien déjà existant).

- Propriétés

- Diamètre pour $p = 0$ élevé ($N/(2K)$), mais chutant rapidement lorsque p croît (small-world)
- Niveaux élevés de Clustering (surtout avec p proche de 0)
- Mais distributions des degrés fortement concentrées autour de la moyenne K



Modèles "Scale-Free"

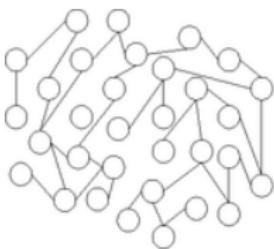
- Algorithme de Barabási-Albert
 - Définir un graphe initial connecté de m_0 noeuds:
 - Pour chaque nouveau noeud n_i à ajouter, on le connecte à m noeuds existants
 - Les noeuds n_j à connecter à n_i sont successivement choisis selon la probabilité $p = k_j / (\sum_{l, (n_i, n_l) \notin E} k_l)$
- ⇒ Tend à choisir les noeuds de degrés supérieurs
- Avec $m_0 = m = 2$:

Modèles "Scale-Free"

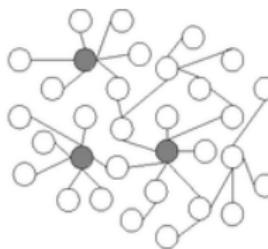
- Propriétés:
 - Diamètre faible
 - Distribution de degrés *heavy-tailed* : loi de puissance

$$\lim_{x \rightarrow \infty} e^{\lambda x} \Pr[X > x] = \infty \quad \text{for all } \lambda > 0.$$

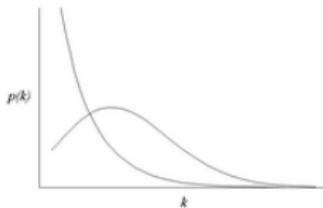
- Faible coefficient de clustering



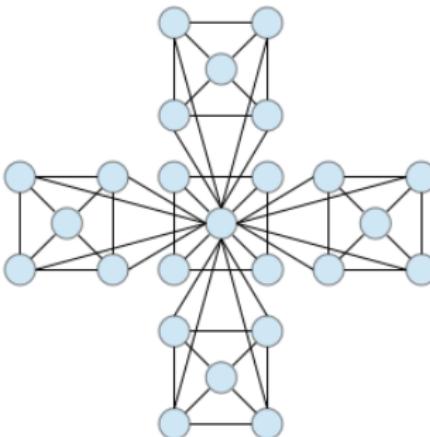
(a) Random network



(b) Scale-free network



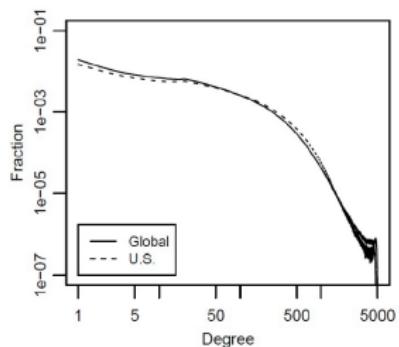
- Création itérative de niveaux hiérarchiques :
 - Replication d'un pattern initial n fois
 - Connexion des nouveaux noeuds avec le pattern initial selon une certaine règle
- Par exemple :
 - ➊ Création d'un cluster initial de 5 noeuds entièrement connecté
 - ➋ Réplication de 4 clusters identiques
 - ➌ Connexion des noeuds périphériques des répliques au noeud central du cluster initial.



- Les modèles hiérarchiques sont des modèles Scale-Free
 - Présence de hubs
 - Heavy tailed degree distribution
- Et possèdent un bon coefficient de clustering
- ⇒ Souvent considérés comme plus réalistes

Modèles "Scale-Free"

- Degree Distribution sur Facebook
 - Quelques utilisateurs ont de très nombreux amis

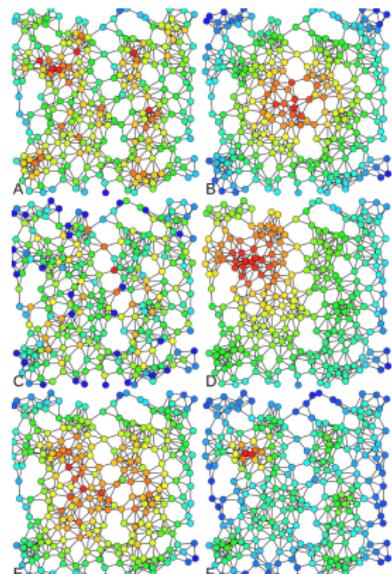


- Exemples de réseaux invariants d'échelle ("Scale-Free") :
 - Le Web, dont l'étude, par Barabási et Albert a donné naissance au terme scale-free network
 - Réseaux de citations dans les articles de recherche scientifique
 - Les réseaux de collaborations, comme celui des collaborations entre scientifiques (publications communes) ou entre entreprises
 - Réseaux de films communs entre acteurs
 - Réseaux de relations sexuelles
 - Les réseaux biologiques dans les cellules.

- Mesures globales pour caractériser la structure d'un réseau:
 - Coefficient de Clustering
 - Distribution de degrés
 - Diamètre
 - Nombre de composantes
 - ...
- Mesures locales pour caractériser un noeud ou un groupe de noeud dans le réseau:
 - Coefficient de Clustering local
 - Degré
 - Centralité
 - Modularité
 - ...

Mesures de centralité

- Mesures de centralité : Importance des noeuds du réseau
- A) Degree centrality : $C(v) = \text{degre}(v)$
- B) Closeness centrality : $C(v) = \sum_{t \in V \setminus v} 2^{-d_G(v,t)}$ avec $d_G(v,t)$ la distance de v à t dans G
- C) Betweenness centrality
$$C(v) = \sum_{s \neq v \neq t \in V} \frac{|\{l \in \sigma_G(s,t), v \in l\}|}{|\sigma_G(s,t)|}$$
 avec $\sigma_G(s,t)$ l'ensemble des plus courts chemins de s à t
- D) Eigenvector centrality : $C(v) = \frac{1}{\lambda} \sum_{(v,t) \in E} C(t)$
- E) Katz centrality : $C(v) = \frac{1}{\lambda} \sum_{(v,t) \in E} (C(t) + 1)$
- F) Alpha Centrality : $C(v) = \frac{1}{\lambda} \sum_{(v,t) \in E} C(t) + e_v$ avec e_i l'influence externe accordée à v à chaque itération
- ...



Une mesure de centralité "centrale" en RI: PageRank

Page & Brin, *The PageRank citation ranking : Bringing Order Into the Web*,
1998

- Motivations :
 - les modèles de recherche documentaire sur le contenu sont facilement manipulables,
 - ils ne sont pas forcément très indicatifs.
- PageRank : Hypothèse
 - une page A pointe vers une page B si B a une information pertinente par rapport à la page A .

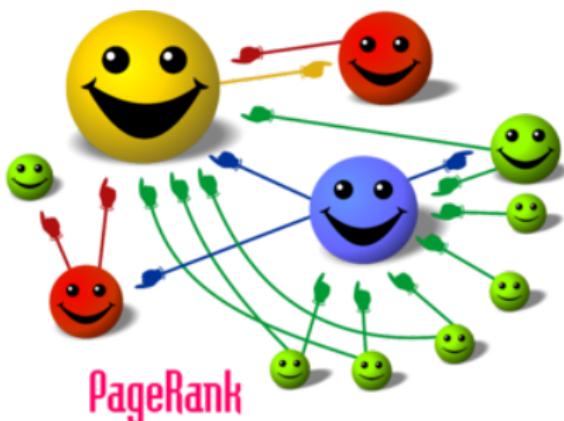
⇒ Une page ayant beaucoup de liens *entrants* est susceptible de contenir une information plus importante qu'une page isolée.

- Une mesure d'importance d'une page doit :
 - Réfléter une notion d'importance *ressentie* par un utilisateur,
 - Être robuste à la *manipulation* :
essayer d'augmenter artificiellement le score doit être coûteux.
Dans l'idéal : coût de la manipulation >> gain...
- Exemple : considérons la mesure d'importance pour la page p :

$$\mu(p) = \text{nb liens pointant vers la page } p$$

Quel est le coût d'obtenir 10000 liens entrants de plus pour un site ? pour 10000 sites se mettant ensemble ?

- Idée de PageRank : définition récursive :
une page A est importante si on a beaucoup de chance
d'arriver sur A en partant de pages importantes



- Idée de PageRank : définition récursive :
une page A est importante si on a beaucoup de chance
d'arriver sur A en partant de pages importantes
- Plus précisément :
 - Soit μ_i l'importance de la page i ,
 - ℓ_i le nombre de liens *sortants* de cette page.
 - Soit P_i l'ensemble des indices j tels qu'il existe un lien de la page j vers la page i .

Posons $\mu_i = \sum_{j \in P_i} \frac{1}{\ell_j} \mu_j$.

- ➊ Que vaut μ_i si aucun lien ne pointe vers i ?
- ➋ Quelle est la contribution d'une page ayant 10000 liens sortants vers une même page d'arrivée ?
- ➌ Que se passe-t'il si une page n'a aucun lien sortant ?

- Surfeur aléatoire :
 - 1 au temps t , le surfeur est sur la page p_t ,
 - 2 avec une probabilité d , il clique aléatoirement sur un lien de p_t . p_{t+1} est la page pointée par ce lien.
 - 3 avec une probabilité $1 - d$, p_{t+1} est une page Web choisie aléatoirement.
 - 4 retour à l'étape 1.
- Soit μ_i^t la probabilité que le surfeur soit sur la page i au temps t :

$$\mu^{t+1} = \frac{1-d}{N} \mathbf{1} + d A \mu^t \text{ avec } A_{ij} = \begin{cases} \frac{1}{N} & \text{si } \ell_j = 0 \\ \frac{1}{\ell_j} & \text{si } j \in P_i \text{ et } N : \# \text{pages Web.} \\ 0 & \text{sinon} \end{cases}$$

Remarques :

- $\frac{1}{1-d} =$
 $\mathbb{E}[\text{nombre de pages visitées avant un choix aléatoire}]$,
- si $d = 1$, on revient à la définition récursive précédente.

- Calcul de μ : méthode de la puissance :

- 1 soit $\mu_i^0 = \frac{1}{N}$ pour tout i
- 2 pour $t = 0..k$:

$$\mu^{t+1} = \frac{1-d}{N} \mathbf{1} + dA\mu^t$$

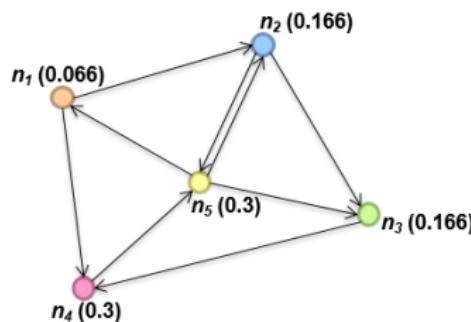
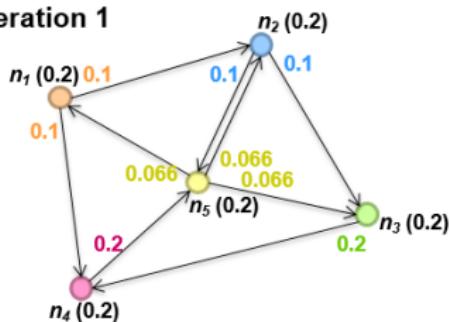
- 3 fin pour

- l'algorithme converge vers une solution unique si $1 \geq d > 0$ (théorème du point fixe),
- chaque itération de l'algorithme s'effectue en visitant une seule fois chaque page Web.

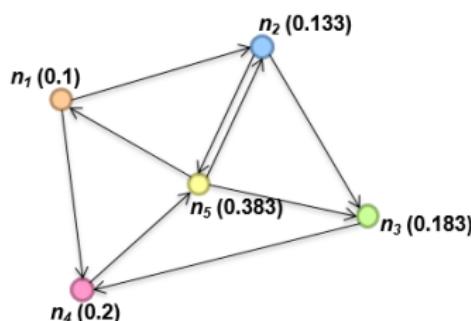
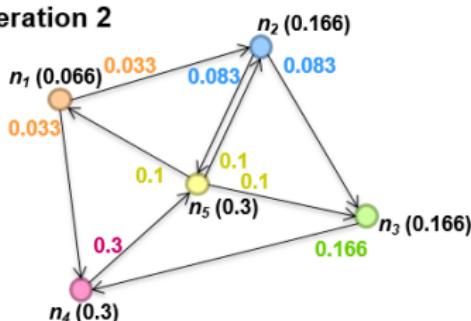
Algorithme PageRank

Avec $d=1$:

Iteration 1

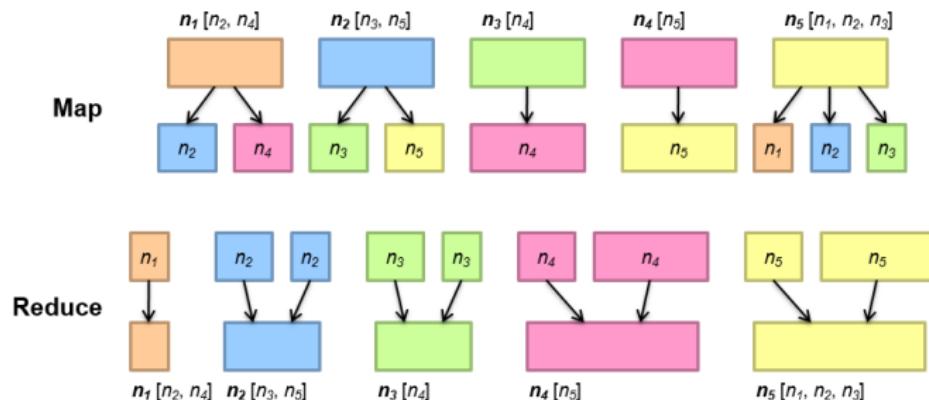


Iteration 2



Algorithme PageRank

Avec $d=1$:



- La modularité est une mesure de la qualité du partitionnement d'un réseau
 - Proportion des poids des arêtes des noeuds d'une classe donnée moins la valeur qu'aurait été cette même proportion si les arêtes étaient disposées au hasard dans le graphe :

$$Q = \sum_{(i,j), c_i = c_j} \left[\frac{A_{ij}}{2m} - \frac{k_i * k_j}{(2m)(2m)} \right]$$

avec :

A_{ij} la valeur de la matrice d'adjacence entre les noeuds i et j,
 c_i le groupe de i ,
 k_i la somme des poids des arêtes adjacentes à i ,
 m la somme des poids des arêtes du graphe.

- La modularité tend vers 0 lorsque l'on partitionne aléatoirement un graphe dans lequel les arêtes sont distribuées aléatoirement
- Partitionnement considéré bon si $Q > 0.3$

- De Nombreux travaux ont donc concerné la structure des réseaux sociaux...
- ... Mais que faire si :
 - Le graphe n'est que partiellement connu
 - Les relations disponibles ne sont que peu représentatives des vrais canaux d'interaction
 - Les noeuds possèdent des attributs connus (ex: profil utilisateur)
 - Les voies de communication diffèrent selon le message véhiculé / contexte d'interaction
 - Le graphe de relations est amené à évoluer au cours du temps
 - ...

- Les modèles traditionnels de traitement de l'information sur Internet:
 - Le Web considéré comme une très grosse base d'information...
 - ... dont les données **évoluent lentement**
 - Les moteurs de recherche peuvent **crawler et indexer** le Web
 - Les besoins d'information sont formulés par des requêtes définies sur l'index du Web
- Ce modèle traditionnel est train de changer
- Nouveau mode de consommation de l'information sur Internet:
 - L'information en ligne nous atteint via des flux de données temps réels et les réseaux sociaux.
- Changement majeur pour les moteurs de recherche:
 - "Tell me about X" vs "Tell me what is hot now"
 - "Tell me about X" vs "Tell me who to ask"

La part des medias sociaux dans le Web a fortement augmenté ces dernières années

- Web de plus en plus tourné vers les interactions sociales
- Ne constitue plus une bibliothèque statique explorée par les utilisateurs
- Un lieu où les utilisateurs **consomment** de l'information, **créent** du contenu et **interagissent** avec d'autres personnes.
- Exemples:
 - Twitter, Digg, Facebook, Google news, Wikipedia, Flickr, LinkedIn, Blogger
- Lieux où les utilisateurs consomment et produisent du contenu:
 - Forums, Blogs, Social Networks, MicroBloggs, Wikis, Podcasts, Slide Sharing, Social Bookmarks, Product reviews, ...
- Facebook est > 7% du traffic internet des US ! (Mars 2010)

- Medias sociaux = Support de relations sociales
- Les medias sociaux réduisent la frontière entre monde réel et monde virtuel
 - Rendent disponibles des traces complexes d'activités humaines
 - Diverses interactions entre individus/groupes peuvent être enregistrés
- "Social Media Mining" = Representation, Analyse et Exploitation des échanges observés sur les media sociaux
 - Analyse des réseaux
 - Fouille de données
 - Apprentissage automatique / statistique
 - Sociologie

N'importe quel utilisateur peut partager, créer et contribuer à la production de contenu. Il peut également exprimer des opinions, tagguer du contenu et conseiller des liens.

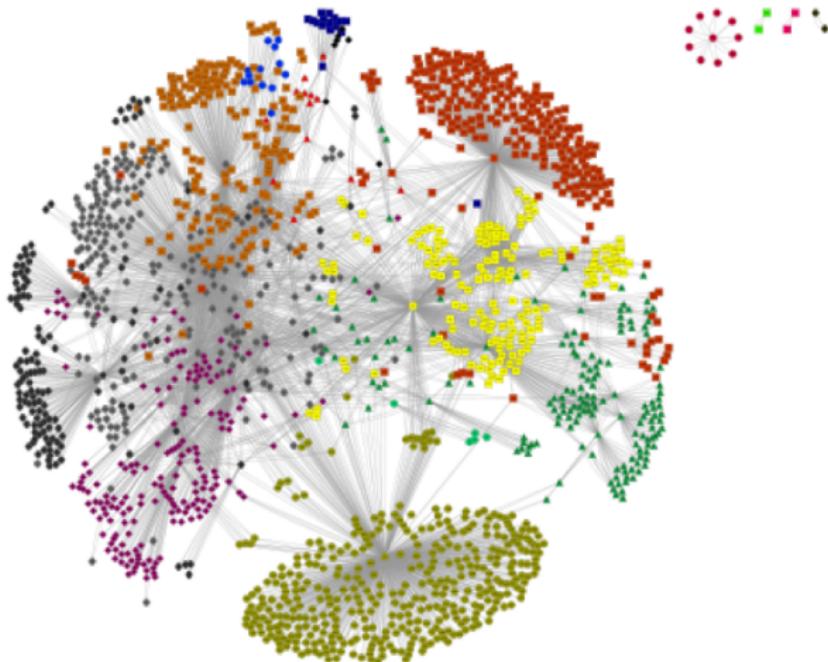
- Consequences:
 - Il est possible d'étudier les opinions et les comportements de millions d'utilisateurs
 - Pour différents objectifs:
 - Analyse et compréhension des comportements d'utilisateurs
 - Analyse du marché / Marketing viral
 - Analyse d'opinion / de sentiment
 - Surveillance / Sécurité
 - Campagnes politiques
 - ...

- Fouille de données dans les Médias sociaux
 - Tâche de traitement de contenu dans un cadre social
- Nouveaux Challenges
 - Paradoxe Big Data : Grandes Masses de données sociales vs. Dissémination des données individuelles
 - Collecte des données : APIs limitées, Choix des données à collecter difficile
 - Réduction du bruit : Définition du bruit fortement dépendant de la tâche
 - Évaluation : Vérité terrain très difficile à obtenir
 - Dynamité : Distributions / Structures non stationnaires

- Rich and big data:
 - Millions d'utilisateurs
 - Millions de contenus
 - Multimedia (texte, image, videos, etc...)
 - Millions de connections et relations (de différents types)
 - Préférences, tendances, opinions, ...
- Analyse réseau social ⇒ Traitements à large échelle
 - Taille des données
 - Complexité des données
 - Dynamicité
- Données + Traitements distribués

Détection de communautés

Détection de communautés



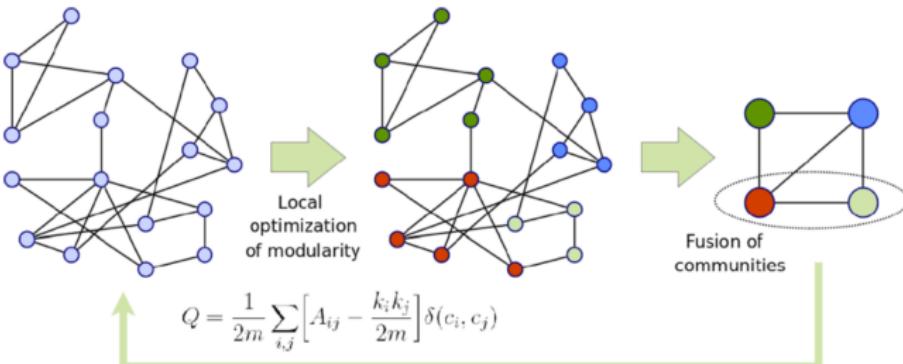
Détection de communautés = Partitionnement du réseau

- Partitionnement en communautés
 - Avec peu de liens inter-communautés
 - Un grand nombre de liens intra-communautés
- Utile pour
 - Comprendre l'organisation générale du réseau
 - Cibler une / des populations spécifiques
 - Simplifier le traitement du réseau
- Apprentissage non supervisé
 - Clustering basé sur les liens du graphe
 - Méthodes statiques / dynamiques
 - Méthode globales / locales

Détection de communautés

- Nombreuses méthodes
 - Minimum-cut
 - Nombre de communautés k prédeterminé
 - Découpage itératif en k communautés de tailles proches minimisant le nombre de liens inter-communautés
 - Méthodes de cliques
 - Recherche des cliques maximales
 - Affectation à la clique la plus proche
 - Méthode Girvan-Newman
 - Intermédiarité des arrêtes (Betweeness centrality)
 - Les arrêtes de forte intermédiaire sont retirées
 - Clustering hiérarchique
 - Similarité entre listes d'adjacence
 - Regroupement itératif des groupes les plus proches
 - Modélisation statistique
 - Modèles génératifs
 - Inférence Bayésienne
$$z^* = \arg \max_z P(z|G) = \arg \max_z P(G|z) \times P(z)$$
 - Optimisation de modularité
 - Modularité = Mesure de qualité d'un partitionnement
 - Différentes méthodes itératives

Détection de communautés : Méthode de Louvain

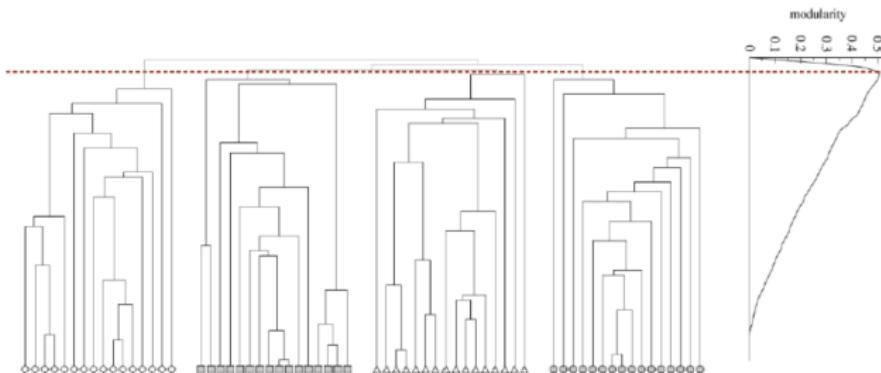


$$\Delta Q = \left[\frac{\Sigma_{in} + k_{i,in}}{2m} - \left(\frac{\Sigma_{tot} + k_i}{2m} \right)^2 \right] - \left[\frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right]$$

Avec :

- Σ_{in} la somme des poids dans la nouvelle communauté de i
- Σ_{tot} la somme des poids des arrêtes vers des noeuds de cette communauté
- k_i la somme des poids des arrêtes partant de i
- $k_{i,in}$ la somme des poids des arrêtes entre i et les noeuds de la communauté
- m la somme des poids de tous les liens du graphe.

Détection de communautés : Méthode de Louvain



D'après Newman & Girvan, 2004

● Limite de Resolution

- Critique à l'encontre des approches basées modularité : leur limite de résolution
- Difficulté à identifier des communautés de tailles très différentes dans un même graphe (ex: réseaux hiérarchiques)

Recommandation / Filtrage collaboratif

- Filtrage Collaboratif
 - Automatisation de la sélection et la recommandation d'items (filtrage)
 - Prédiction des préférences d'un utilisateur à partir des préférences d'autres utilisateurs (collaboratif)
- Approche standard
 - Matrice de notes (utilisateurs x articles)
 - Matrice à "trous"
 - Tâche = prédiction des notes manquantes
 - Factorisation matricielle: $\underset{U,V}{\operatorname{arg\,min}} ||X - UV||^2$
- Dans un contexte social
 - Utilisation de relations sociales explicites

- Corrélation

- La note donnée à un item par un utilisateur a tendance à être proche de celle donnée par des utilisateurs aux jugements similaires
 - ⇒ Deux utilisateurs aux jugements proches auront tendance à émettre des avis proches sur de nouveaux items.

- Influence sociale

- Les notes données par un utilisateur sont influencées par celles données par ses amis
 - ⇒ Deux personnes connectées ont plus de chances d'avoir des avis similaires que deux utilisateurs qui ne se connaissent pas.

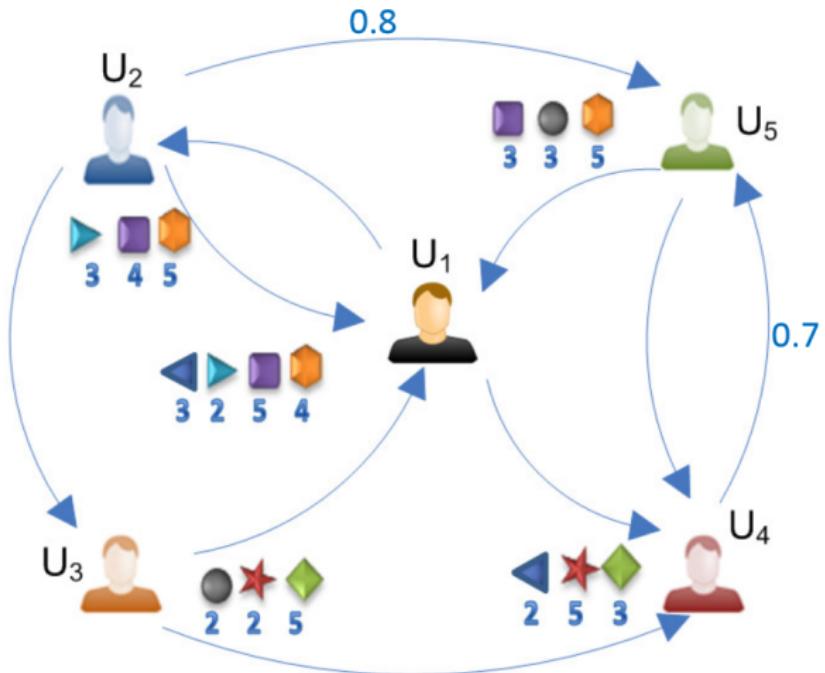
- Homophilie (ou sélection)

- Les utilisateurs aux avis proches ont tendance à devenir amis

- Transitivité

- Les amis des amis d'un utilisateur ont plus de chances de devenir ses propres amis que les autres utilisateurs

Social Rating Networks



- Recommendation sur les réseaux:
 - Filtrage Collaboratif + Liens sociaux
- Bénéfices:
 - Exploitation de l'influence sociale, des correlations de préférences, de transitivité, d'homophilie...
 - Démarrage à froid pour des nouveaux utilisateurs
 - Robustesse à la fraude
- Challenges
 - Avis fortement disséminés sur le réseau
 - Propagation d'avis difficile sur de longues distances
 - Fiabilité des relations très variable
 - Données sensibles (privauté, anonymat)

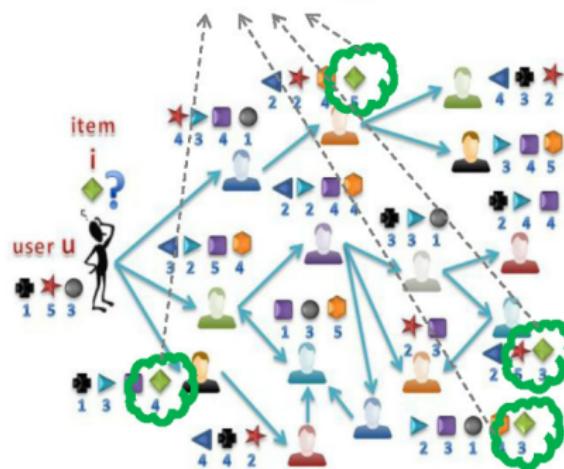
- Deux principales familles d'approches:
 - Modèles de voisinage
 - Marches aléatoires sur des réseaux de confiance
 - Modèles à variables latentes
 - Factorisation matricielle avec prise en compte de la structure

- Deux principales familles d'approches:
 - **Modèles de voisinage**
 - Marches aléatoires sur des réseaux de confiance
 - Modèles à variables latentes
 - Factorisation matricielle avec prise en compte de la structure

Recommendation sur les réseaux sociaux: TrustBased Approaches

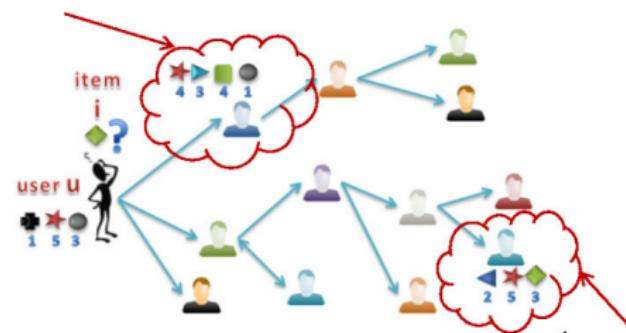
- Approches basées sur la confiance en les utilisateurs du réseau

$$\hat{r}_{u,i} = \frac{\sum_{v \in N_{u,i}} t_{u,v} r_{v,i}}{\sum_{v \in N_{u,i}} t_{u,v}}$$



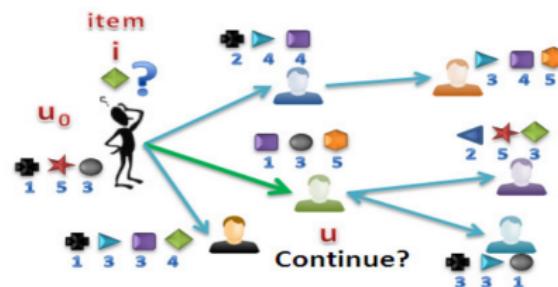
Recommendation sur les réseaux sociaux: TrustWalker

- TrustWalker [Jamali & Ester, 2009]:
 - + confiance en les utilisateurs les plus proches...
 - ... Mais peu de chances d'observer des avis dans le voisinage proche pour un item particulier
 - ⇒ Scores sur items similaires pour utilisateurs proches



- ⇒ Scores sur items identiques pour utilisateurs éloignés

Recommendation sur les réseaux sociaux: TrustWalker



- A chaque iteration k , pour le noeud u :
 - Si u a donné un score $r_{u,i}$ à i alors on le retourne
 - Sinon :
 - Avec une proba $\phi_{u,i,k}$, on sélectionne aléatoirement un item j noté par u selon $p(Y_{u,i} = j)$ et on retourne le score $r_{u,j}$

$$P(Y_{u,i} = j) = \frac{sim(i,j)}{\sum_{l \in RI_u} sim(i,l)}$$

- Avec une proba $1 - \phi_{u,i,k}$, on continue le parcours sur un voisin de confiance de u .

Recommendation sur les réseaux sociaux: TrustWalker

La similarité entre deux items est basée sur un coefficient de correlation positive de Pearson:

$$\text{corr}(i, j) = \frac{\sum_{u \in UC_{i,j}} (r_{u,i} - \bar{r}_u)(r_{u,j} - \bar{r}_u)}{\sqrt{\sum_{u \in UC_{i,j}} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{u \in U} (r_{u,i} - \bar{r}_u)^2}}$$

$$\text{sim}(i, j) = \frac{1}{1 + e^{-\frac{|UC_{i,j}|}{2}}} \times \text{corr}(i, j)$$

La probabilité $\phi_{u,i,k}$ de stopper l'exploration sur l'utilisateur u pour l'item cible i à l'iteration k est donnée par:

$$\phi_{u,i,k} = \max_{j \in RI_u} \text{sim}(i, j) \times \frac{1}{1 + e^{-\frac{k}{2}}}$$

Recommendation sur les réseaux sociaux: TrustWalker

- Prediction = Espérance du score retourné

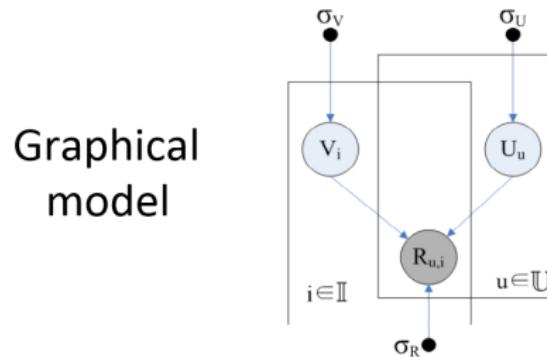
$$\hat{r}_{u,i} = \sum_{\{(v,j) | R_{v,j}\}} P(XY_{u,i} = (v,j)) r_{v,j}$$

- Approximation matricielle des $P(XY_{u,i} = (v,j))$ trop coûteuse
- ⇒ Moyenne des résultats obtenus sur de multiples simulations de marches aléatoires simples

- Deux principales familles d'approches:
 - Modèles de voisinage
 - Marches aléatoires sur des réseaux de confiance
 - **Modèles à variables latentes**
 - Factorisation matricielle avec prise en compte de la structure

Recommendation sur les réseaux sociaux: Factorisation matricielle

- Factorisation matricielle classique
 - Découverte de facteurs latents U et V expliquant les scores observés R :



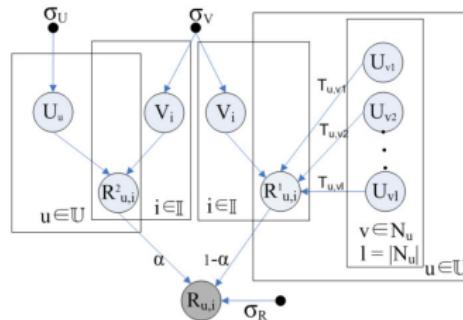
⇒ Avec $\hat{R} = UV$, chercher U et V minimisant :

$$\sum_{\text{observed}(u,i)} (R_{ui} - \hat{R}_{ui})^2 + \lambda(\|U\|^2 + \|V\|^2)$$

- Dans les réseaux ⇒ inclusion de la structure T

Recommendation sur les réseaux sociaux: Factorisation matricielle

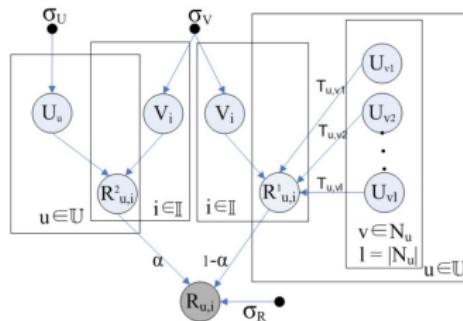
- Recommendation Social Trust Ensemble (RSTE) [Ma et al., 2009]:
 - Combinaison linéaire
 - D'une factorisation matricielle classique sur les scores observés
 - Et de facteurs latents sur les utilisateurs de confiance sur le réseau



$$\hat{R}_{u,i} = \alpha U_u^T V_i + (1 - \alpha) \sum_{v \in N_u} T_{u,v} U_v^T V_i$$

Recommendation sur les réseaux sociaux: Factorisation matricielle

- Recommendation Social Trust Ensemble (RSTE) [Ma et al., 2009]:
 - Combinaison linéaire
 - D'une factorisation matricielle classique sur les scores observés
 - Et de facteurs latents sur les utilisateurs de confiance sur le réseau

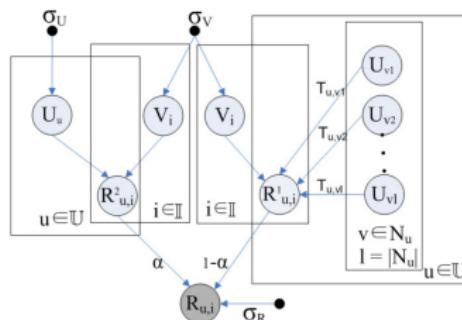


$$\hat{R}_{u,i} = \alpha U_u^T V_i + (1 - \alpha) \sum_{v \in N_u} T_{u,v} U_v^T V_i$$

- ⇒ Scores d'un utilisateur dépendants des scores de son voisinage...

Recommendation sur les réseaux sociaux: Factorisation matricielle

- Recommendation Social Trust Ensemble (RSTE) [Ma et al., 2009]:
 - Combinaison linéaire
 - D'une factorisation matricielle classique sur les scores observés
 - Et de facteurs latents sur les utilisateurs de confiance sur le réseau

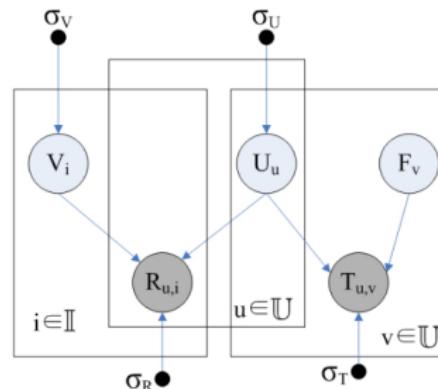


$$\hat{R}_{u,i} = \alpha U_u^T V_i + (1 - \alpha) \sum_{v \in N_u} T_{u,v} U_v^T V_i$$

- ⇒ Scores d'un utilisateur dépendants des scores de son voisinage...
- ⇒ ... Mais pas de réelle propagation des représentations

Recommendation sur les réseaux sociaux: Factorisation matricielle

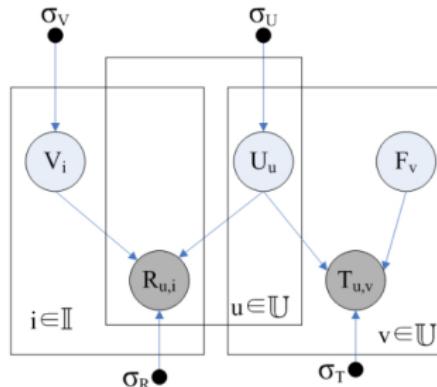
- Modèle SoRec [Ma et al., 2008]:
 - Factorisation conjointe des scores R et du réseau T



- 1 facteur latent pour les items et 2 pour les utilisateurs :
 - 1 pour les sources des relations
 - 1 pour les destinations
- Mêmes facteurs latents U pour les scores et la structure

Recommendation sur les réseaux sociaux: Factorisation matricielle

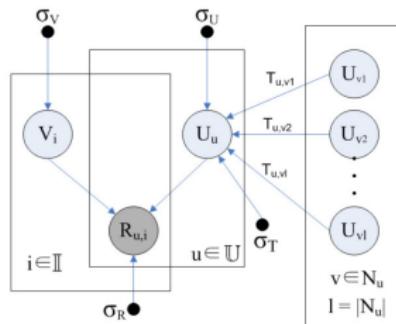
- Modèle SoRec [Ma et al., 2008]:
 - Factorisation conjointe des scores R et du réseau T



- 1 facteur latent pour les items et 2 pour les utilisateurs :
 - 1 pour les sources des relations
 - 1 pour les destinations
- Mêmes facteurs latents U pour les scores et la structure
 - ⇒ Utilisateurs avec des amis communs ont une représentation U proche

Recommendation sur les réseaux sociaux: Factorisation matricielle

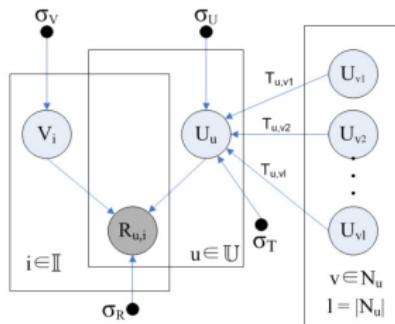
- Social MF [Jamali & Ester, 2010]:
 - Facteurs d'un utilisateur dépendent de ceux de son voisinage



$$\sum_{\text{observed}(u,i)} (R_{ui} - \hat{R}_{ui})^2 + \lambda(\|U\|^2 + \|V\|^2) + \beta \left(\sum_u ((U_u - \sum_v T_{u,v} U_v)(U_u - \sum_v T_{u,v} U_v)^T) \right)$$

Recommendation sur les réseaux sociaux: Factorisation matricielle

- Social MF [Jamali & Ester, 2010]:
 - Facteurs d'un utilisateur dépendent de ceux de son voisinage

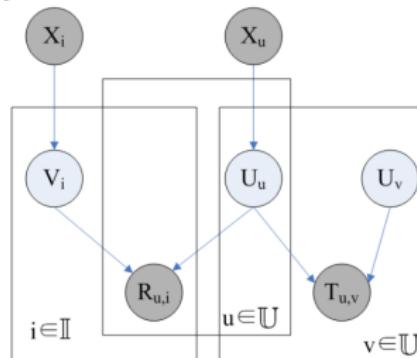


$$\sum_{\text{observed}(u,i)} (R_{ui} - \hat{R}_{ui})^2 + \lambda(\|U\|^2 + \|V\|^2) + \beta \left(\sum_u ((U_u - \sum_v T_{u,v} U_v)(U_u - \sum_v T_{u,v} U_v)^T) \right)$$

- Proche de SoREC
 - ⇒ Utilisateurs avec des amis communs ont une représentation U proche

Recommendation sur les réseaux sociaux: Factorisation matricielle

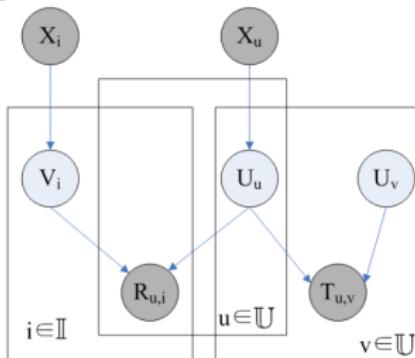
- Modèle FIP [Yang et al., 2011]:
 - Factorisation conjointe des scores R et du réseau T



- 1 facteur latent pour les items et 1 pour les utilisateurs
⇒ Graphe non-orienté
- Mêmes facteurs latents U pour les scores et la structure
- Profils items et utilisateurs en tant que priors des facteurs latents

Recommendation sur les réseaux sociaux: Factorisation matricielle

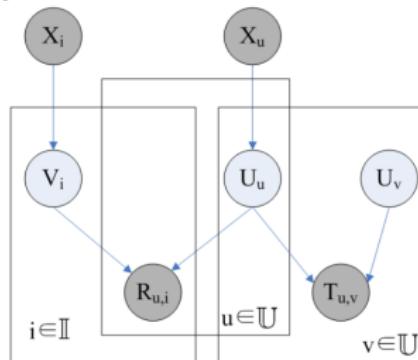
- Modèle FIP [Yang et al., 2011]:
 - Factorisation conjointe des scores R et du réseau T



- 1 facteur latent pour les items et 1 pour les utilisateurs
⇒ Graphe non-orienté
- Mêmes facteurs latents U pour les scores et la structure
⇒ Utilisateurs avec des amis communs ont une représentation U proche
- Profils items et utilisateurs en tant que priors des facteurs latents

Recommendation sur les réseaux sociaux: Factorisation matricielle

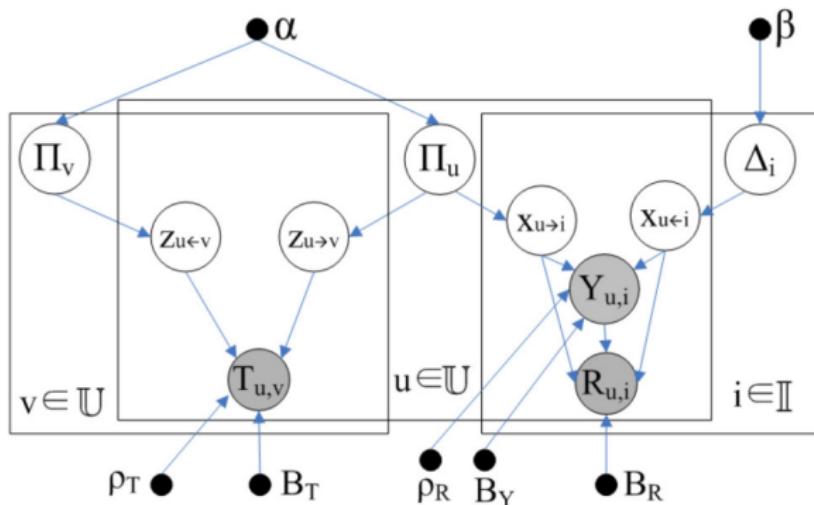
- Modèle FIP [Yang et al., 2011]:
 - Factorisation conjointe des scores R et du réseau T



- 1 facteur latent pour les items et 1 pour les utilisateurs
⇒ Graphe non-orienté
- Mêmes facteurs latents U pour les scores et la structure
⇒ Utilisateurs avec des amis communs ont une représentation U proche
- Profils items et utilisateurs en tant que priors des facteurs latents
⇒ Utilisateurs et items avec des profils similaires ont des représentations proches

Recommendation sur les réseaux sociaux: Factorisation matricielle

- Generalized Stochastic Block Model [Jamali et al. 2011]:
 - Appartenance à des groupes / communautés latents



- Les relations entre utilisateurs et items sont gouvernées par les relations entre leurs groupes d'appartenance