Réseaux sociaux et recommandation de contenus non-populaires

Apprentissage Artificiel et Fouille de Données 2010

Cécile Bothorel

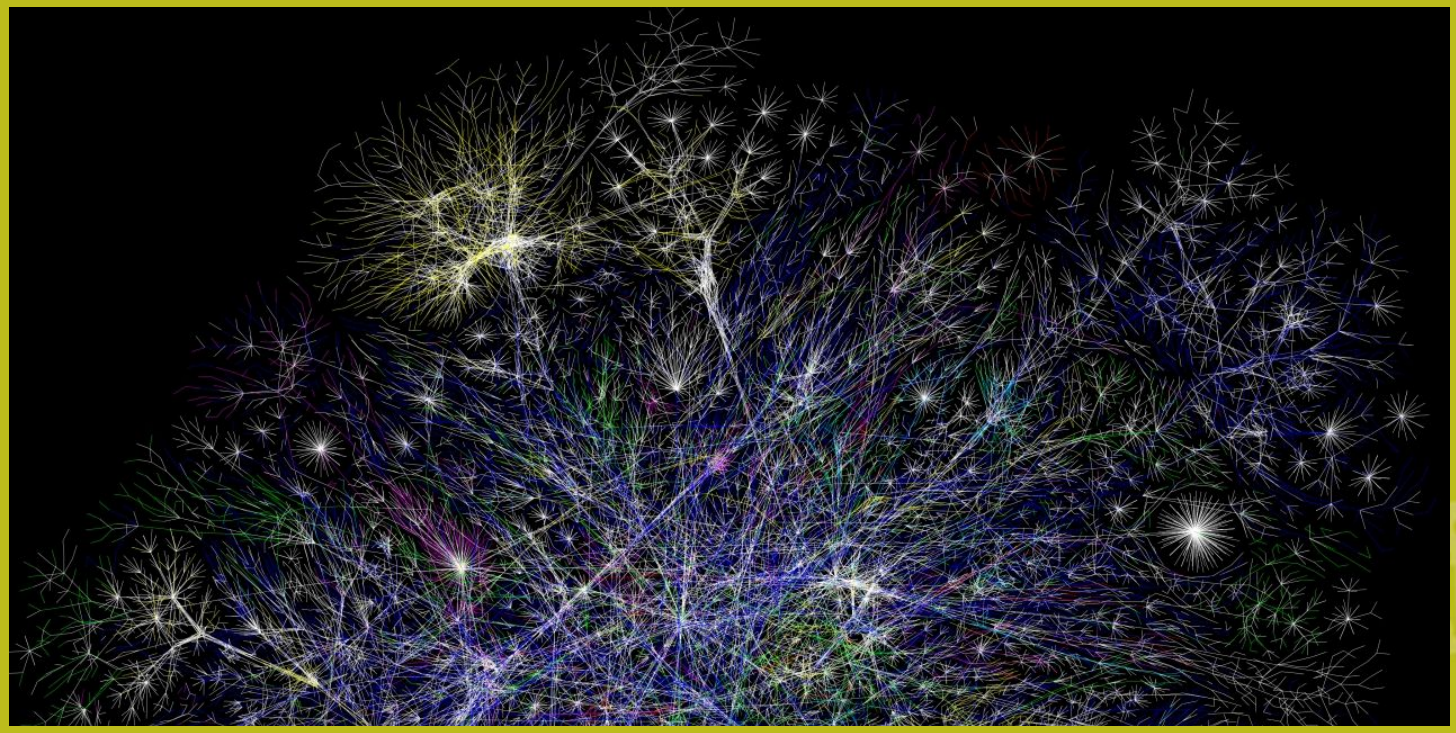
Département LUSSI, Brest

Travail préliminaire

TELECOM Bretagne

« The Web is unfair »

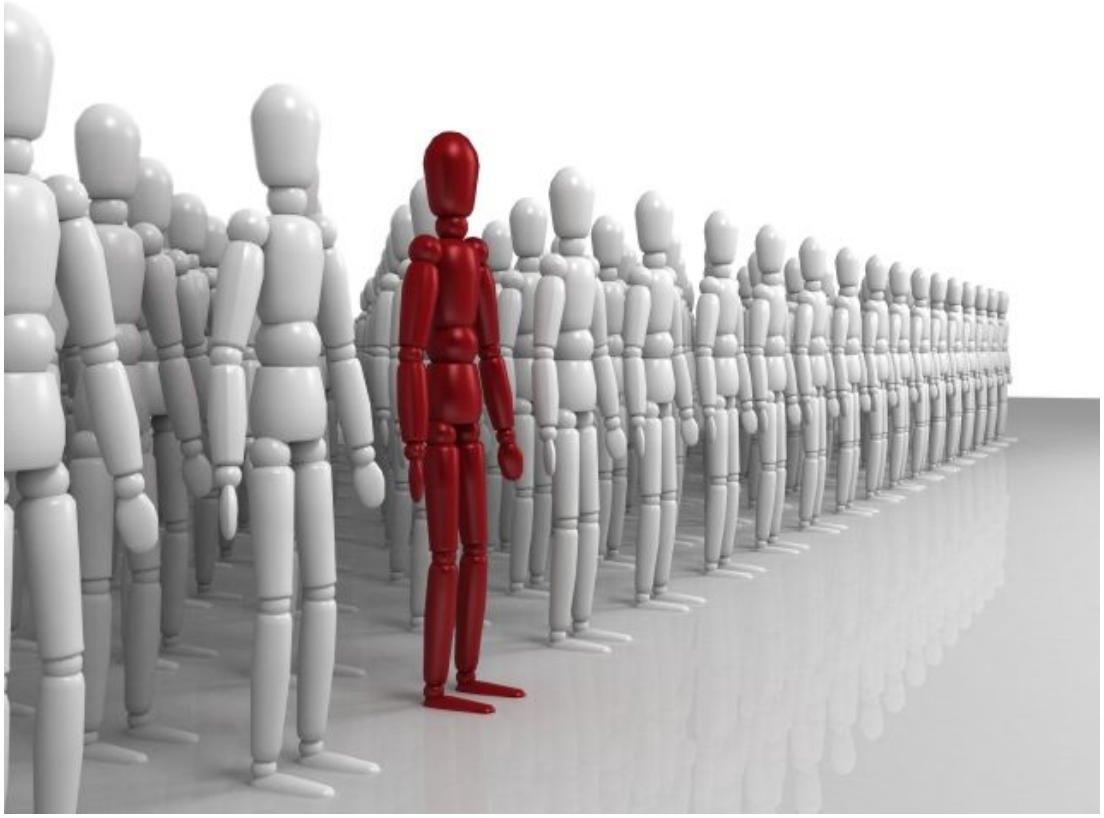
Linked, The New Science of Networks, Albert-Laszlo Barabasi, 2002



Trouver un contenu ?

- **Browsing**
  - Le fil conducteur : mon propre intérêt
  - 1 trillion de pages...
- **Searching**
  - L'algorithmique du moteur de recherche devient le prescripteur

Trouver la personne qui a un contenu ?



What was I Looking For?

- **Wilfing**
  - Partage
  - Souscription
    - web 2.0 + médias sociaux + RSS
  - Navigation sans but, trouvailles surprenantes, sérendipité

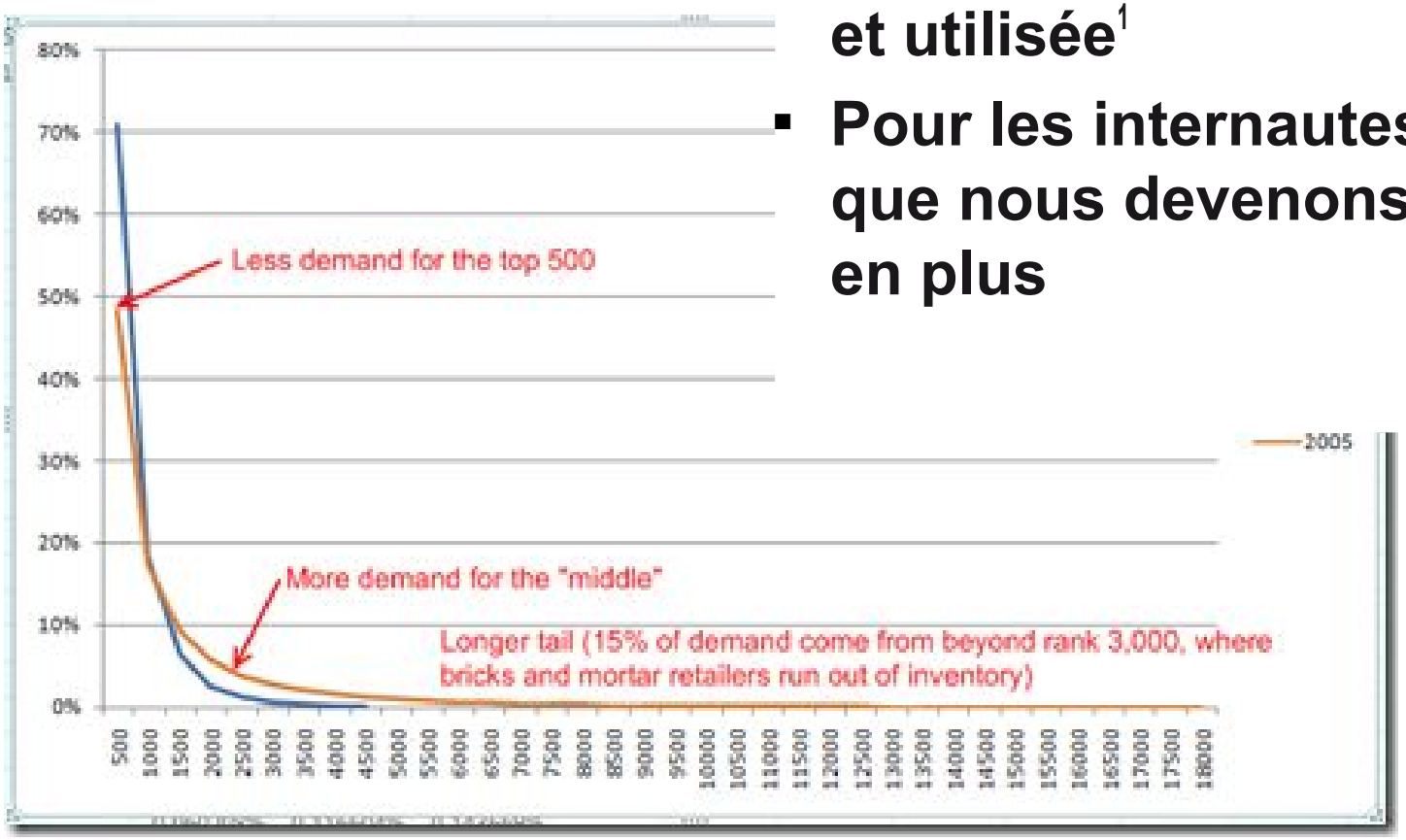
- Internautes producteurs
- Nouvelle problématique : choix de ses internautes « prescripteurs »

Tout est affaire de popularité...

- **… parce que la popularité est un critère efficace de prescription**
  - Popularité = pertinence
  - Popularité = clics
    - Suggestion de vidéos[1], site YouTube
  - Popularité lucrative

[1]Baluja, Shumeet, Rohan Seth, D., Sivakumar, Yushi Jing, Jay Yagnik, Shankar Kumar, Deepak Ravichandran, and Mohamed Aly. Video Suggestion and Discovery for YouTube: Taking Random Walks Through the View Graph. In Proceeding of the 17th International Conference on World Wide Web, Beijing, China, 21–5 April 2008.

Pourtant...



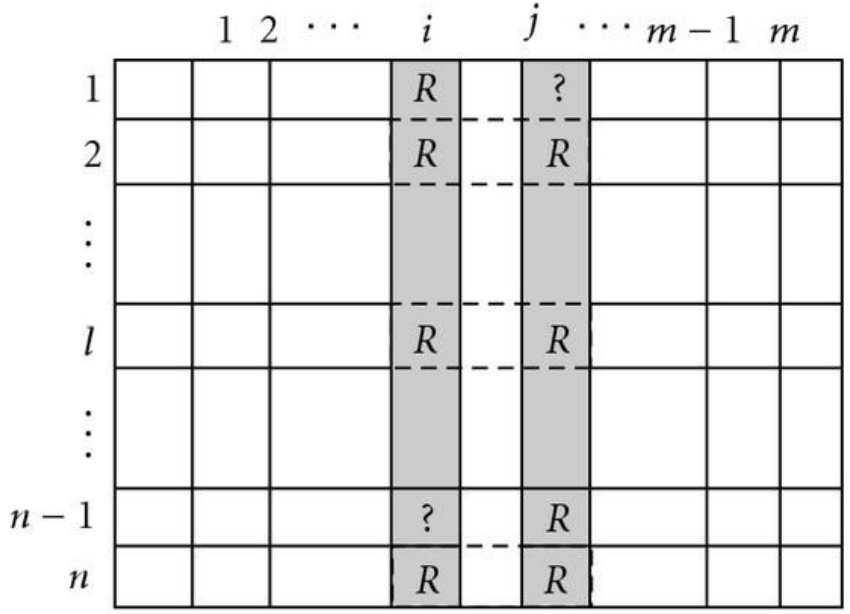
- La recommendation est utile et utilisée[1]
- Pour les internautes avertis que nous devenons de plus en plus

Recommandation

- **Matrice User x Films**
- **Calcul de similarité entre utilisateurs**
- **Prédiction de contenus**

Collaborative filtering

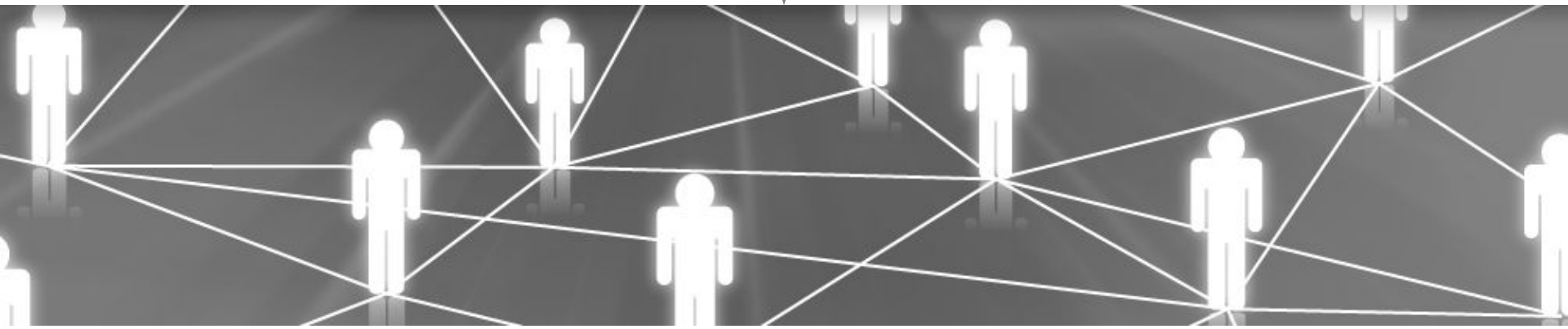
	Shrek	Snow White	Spider-man	Super-man
Alice	Like	Like		Dislike
Bob		Like	Dislike	Like
Chris		Dislike	Like	
Tony	Like		Dislike	?



Recommandation et Réseaux sociaux

- **Matrice User x Films**
- **Calcul de similarité entre utilisateurs**
- **Prédiction de contenus**

Et si les internautes le faisaient eux-mêmes ?



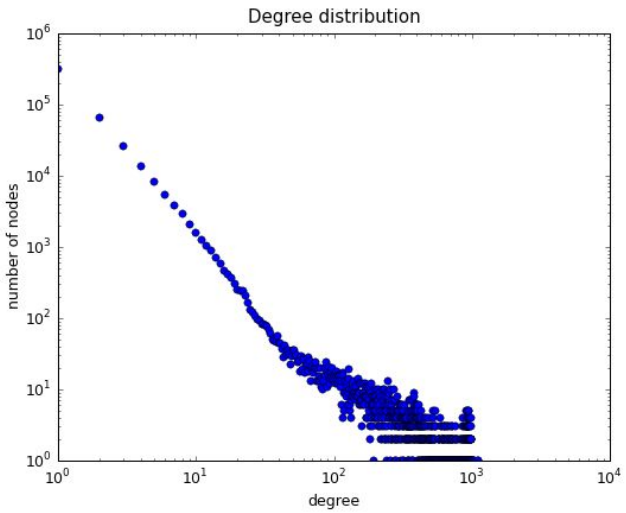
TELECOM Bretagne

Réseau social et recommandation de contenus non populaires

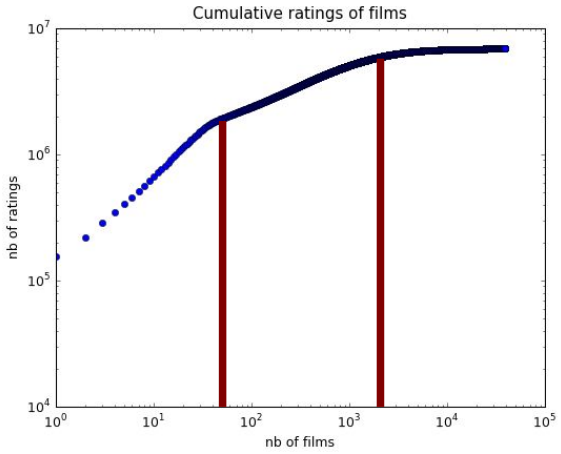
Données

5 millions utilisateurs actifs
2 milliards de reviews

- **Site Flixster**
  - Réseau social
    - Crawl été 2009, à partir de 100 utilisateu
    - 459007 utilisateurs et 895794 liens
    - Environ 10% de liens réciproques
    - Degré max 1108, degré moyen 3.9
  - « reviews » associées
    - Film
    - Note
    - Commentaire textuel
    - 6896205 reviews, 38656 films



Données



- **Reviews :**
  - 3 sous-ensembles de films selon la popularité
    - 40 films populaires
    - 36656 films rares (- 520 reviews)
  - 80% train / 20% test

- **RareReviews :**
  - 789728 reviews
  - 62324 utilisateurs
  - 12.6 notations en moyenne (13 803 pour le plus prolifique !)
  - Le tiers des reviews sont des reviews uniques

Démarche

- **Expérience 0 : Filtrage Collaboratif (référence)**
- **Expérience 1 : local popularity**
  - graphe de films co-annotés
  - calcul de scores entre films
- **Expérience 2 : social popularity**
  - graphe social
  - détection de communautés
  - graphes de films co-annotés communautaires
  - calcul de scores entre films

Démarche

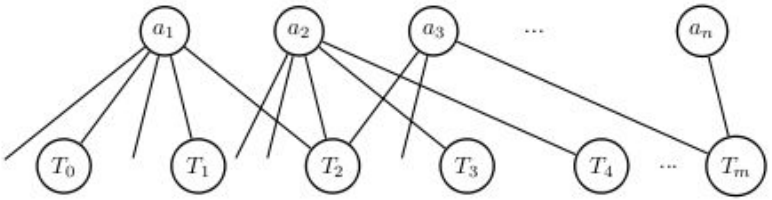
- **Utilisation de matrices de similarité entre films pour faire des prédictions**
  - pour chaque utilisateur présent dans le jeu d'apprentissage et de test
  - en fonction des films déjà notés
  - générer k plus proches prédictions (10 < k <100)
- **Evaluation**
  - calcul de précision/rappel et accuracy
  - comparaison avec le filtrage collaboratif

Local popularity

- **2 graphes de co-annotations**
  - Notations positives
    - 23638 noeuds 7380468 liens
    - Nb max de co-annotations : 160
    - Après filtrage (poids > 5) 6923277 liens retirés
  - Notations négatives
    - 12341 nodes with 1066994 edges
    - Nb max de co-annotations : 66
    - Après filtrage (poids > 2) : 944225 liens retirés
  - Calcul de score
    - Considérant le voisinage de chaque film

Local popularity

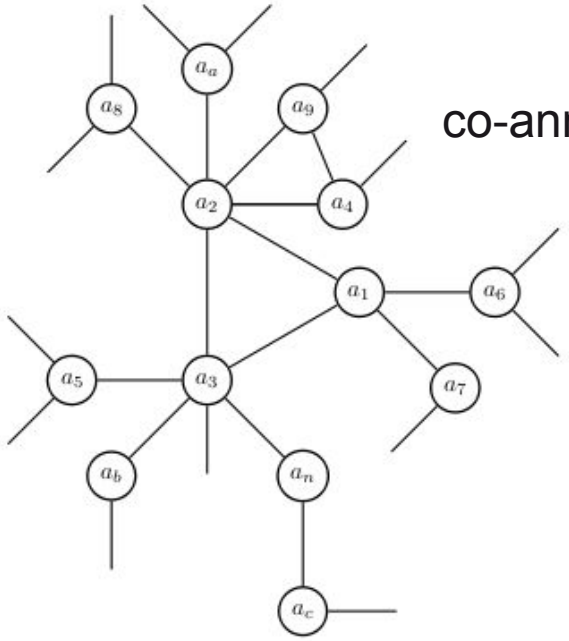
- Méthodes globales
  - Longeur des plus courts chemins
  - Katz
  - Hits, PageRank
  - SimRank



annotation

- Méthodes locales
  - Nombre de voisins communs
  - Coefficient de Jaccard
  - Adamic/Adar Measure[2]

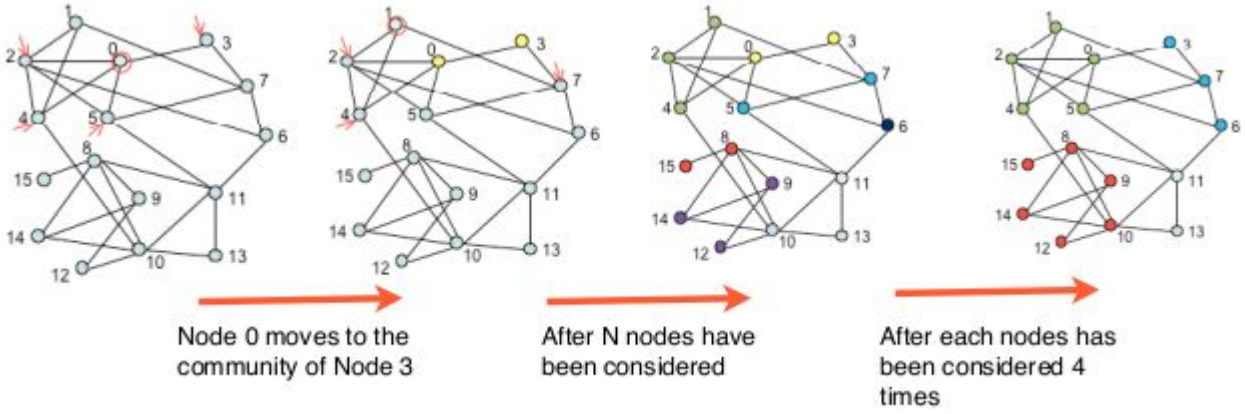
$$\sum_{z\ :\ \text{feature shared by } x,y} \frac{1}{\log(\text{frequency}(z))}$$



co-annotation

[2]Lada Adamic and Eytan Adar, Friends and neighbors on the web, Social Networks 25 (2003), no. 3, 211–230

Social popularity



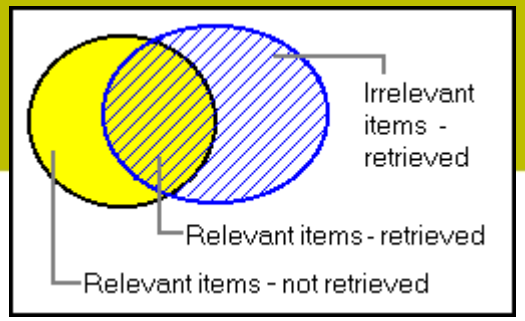
- **Réseau social**
  - 459007 noeuds, 895794 liens
- **Algorithme « Fast Unfolding »**[3]
  - Optimisation locale de modularité
- **85 communautés (Q = 0.66)**
- **41 minutes sur un Core(TM)2 Duo CPU @ 2.66 GHz avec 4 Go RAM, python, Ubuntu 10**

[3] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. 2008 (2008), no. 10, P10008+

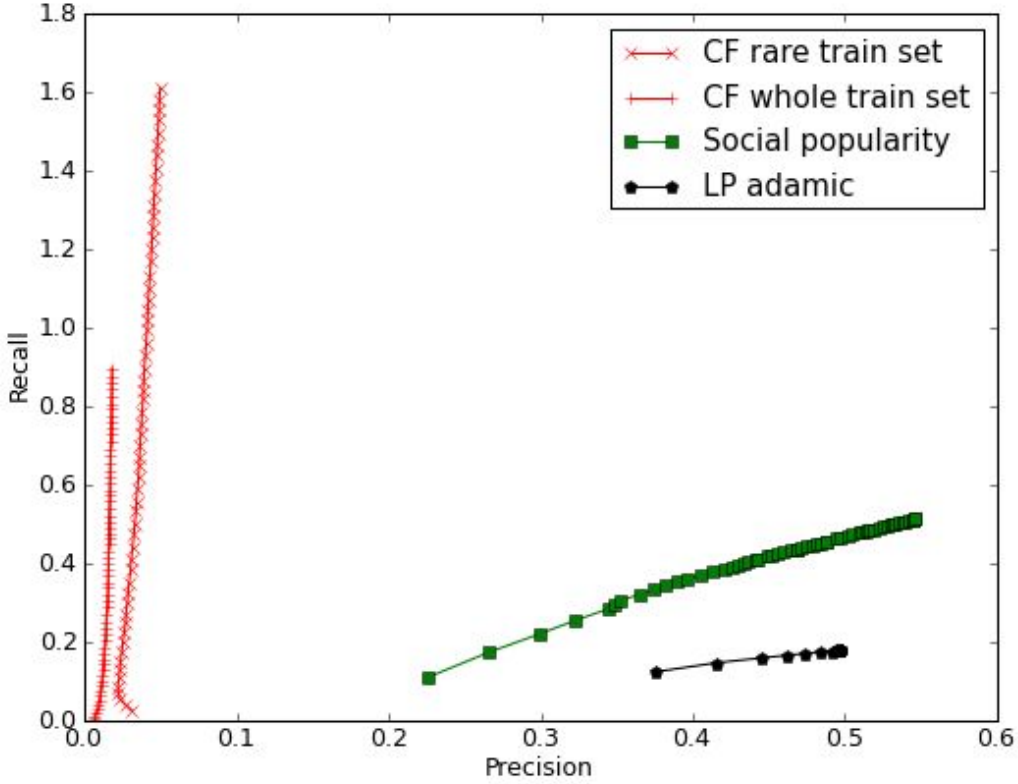
Social popularity

- **Réseau social**
  - 459007 noeuds, 895794 liens
- **Seuls 42541 utilisateurs dans le jeu de données d'apprentissage**
- **Utilisation des reviews de ces 42541 utilisateurs pour générer des graphes de films co-annotés contextuellement aux 85 communautés**
- **Calcul de scores (Adamic 2003)**

Evaluation



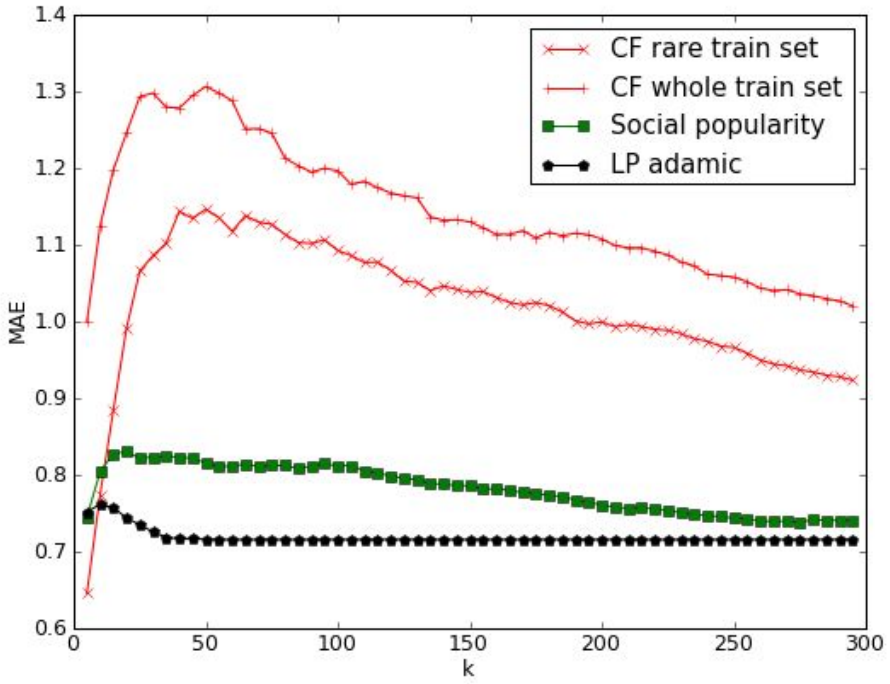
- **Precision/rappel**
  - Très faible précision, surtout CF
  - Très faible rappel : faible pouvoir prédictif (nb de recommendations faible)
    - LP: 2
    - SP : 5
    - CF : 200 pour CFrare et 270 CF



Evaluation

- **Mean Absolute Error**
  - Erreur sur la note prédite pour les films « bien » prédits
  - K = nb de recommendations
  - Plus les calculs sont locaux, et moins il y a de prédictions, mais meilleures elles sont

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N}$$



Conclusion

- Prédiction de contenus rares : rappel et précision faibles
- Travail préliminaire
  - Sont-ce les mêmes prédictions d'une méthode à l'autre ?
  - Travail sur le calcul de score, similarité
  - Calculs de similarité locaux aux communautés mal exploités pour les films qui se trouvent dans plusieurs communautés sociales
- Expérimentations
  - utilisateurs « nouveaux » ou peu prolixes
  - contenus mid-populaires, populaires
- Validation qualitative : découverte de films ?
- Communautés recouvrantes
- Tester les matrices de scores sur des services de VoD hors flixster
- Intégrer l'analyse d'opinions