Introduction à l'Analyse des Réseaux Sociaux (ARS)

Maria Malek

18 mars 2010

Analyse des réseaux sociaux
- Définition
- Propriétés
- Utilisation & Applications

Les mesures globales

Les mesures locales

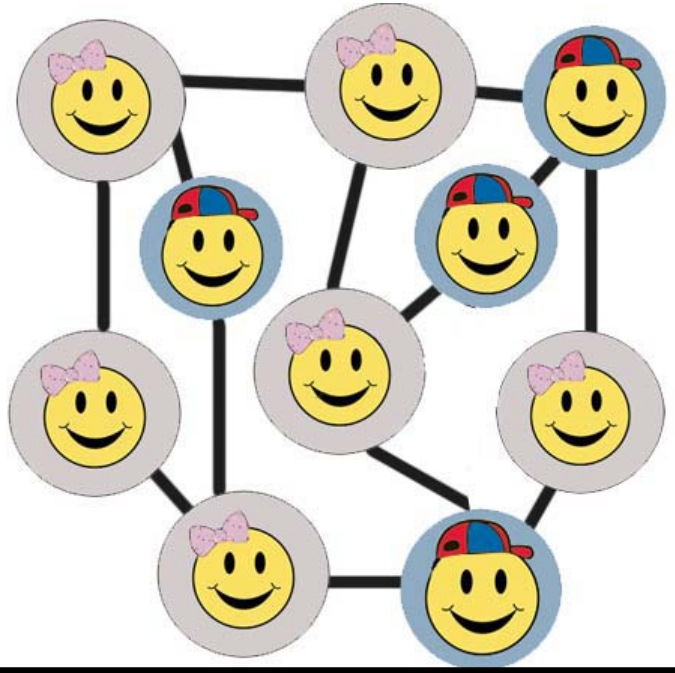
Graphes Bipartis & Réseaux sociaux

Analyse de communautés

Autres Problèmes d'analyse réseaux sociaux

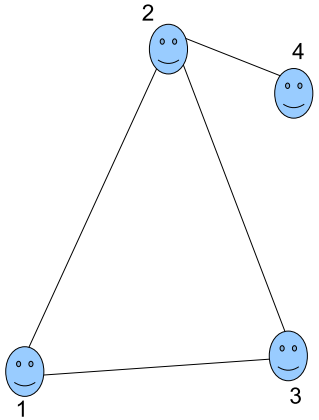
Analyse des réseaux sociaux

- Etude des entités sociales (les personnes dans les organisations qu'on appelle acteurs), leurs interactions et leurs relations ;
- Modélisation par un graphe :
  - un nœud représente un acteur,
  - un lien représente une relation.



Définition et représentation d'un graphe

- Un graphe (non orienté) G est défini par :
  - Un ensemble X de sommets, Un ensemble E d'arêtes.
  - Degré d'un sommet : nombre d'arêtes incidentes à X.
  - Représentation par la matrice d'adjacence :

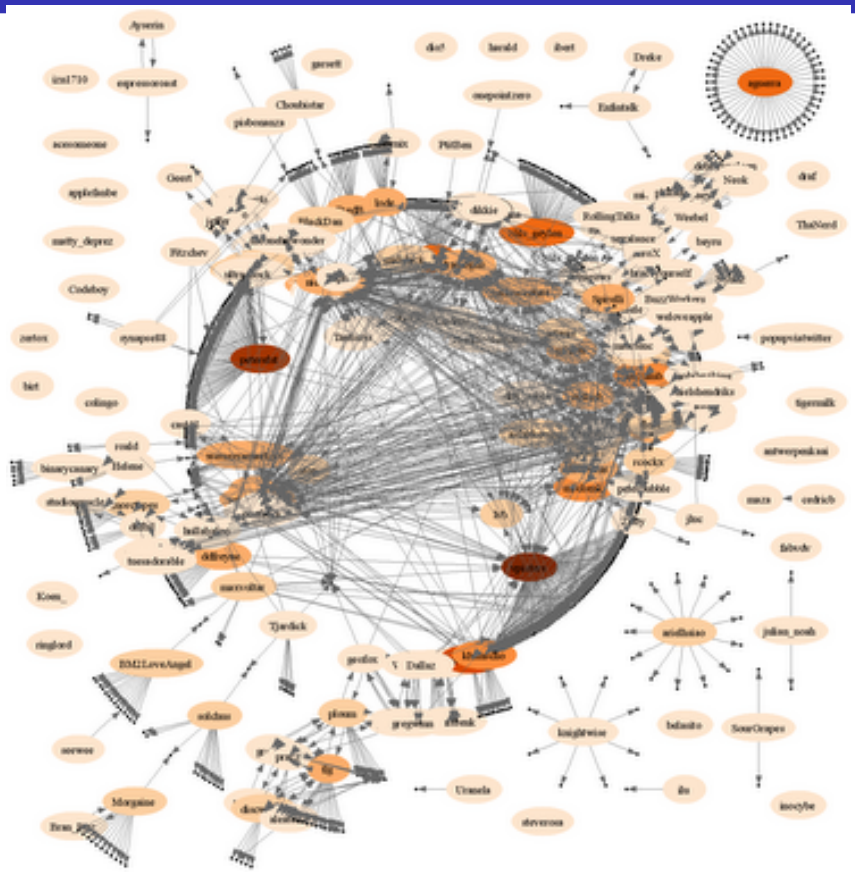


	1	2	3	4
1	0	1	1	0
2	1	0	1	1
3	1	1	0	0
4	0	1	0	0

Réseautage social

- Ensemble des moyens mis en œuvre pour relier des personnes (physiques comme morales) entre elles.
- Applications informatiques qui servent à constituer un réseau social : Web social (facebook, twitter, etc.)
  - Relier des amis, des associés, ou d'autres individus employant ensemble une variété d'outils (entreprise, logiciels, sites Web, etc.).
  - Objectifs :
    - gestion des carrières professionnelles (coaching professionnel),
    - distribution et la visibilité artistique,
    - gestion des contacts personnels (site de rencontres sur Internet, sites d'anciens élèves, site de relations amicales).

Réseautage social - Exemple



Twitter Friends van Belgische Twitteraars

source : http ://datamining.typepad.com/data_mining/2007/04/twitter_social_.html

Propriétés des réseaux sociaux

Grandes tailles : $|X| > 10^3$

Le diamètre du réseau : (le plus long chemin entre deux nœuds) est de longueur bornée.

- Propriété du petit monde issue des études sociologiques 6 degrés de séparation selon l'expérience du petit monde de Stanley Milgram 1967.

La transitivité dans le réseau : *l'ami de mon ami est mon ami.*

- Mesure du taux de clustering suivant :

$$C = \frac{3 * \texttt{nombre de triangles dans le réseau}}{\texttt{nombre de triples connectés de nœuds}}$$

La distribution des degrés : nombre de nœuds connectés à un nœud donné obéit à la distribution de degrés en loi de puissance.

La probabilité pour un nœud d'avoir k voisins est : $P(k) = k^{-\gamma}$

Utilisation des réseaux sociaux

- Décrire les interactions entre les entités sociales.
  - Analyse des interactions dans les mails, les forums de discussions, les réseaux professionnels, etc.
- Modéliser les liens entre un ensemble de pages Web
  - le Web est une société virtuelle (un réseau social virtuel), où chaque page est considérée comme un acteur social et chaque hyperlien est une relation.
- Analyse des citations exprimées dans les références bibliographiques :
  - Découvrir des similarités entre ces documents outre les mesure fondées sur le contenu,
  - Exemple d'application : définir des centres d'affinités entre les auteurs, etc.

Modélisation Sémantique

- Stocker au niveau des acteurs (nœuds) les information concernant le profil utilisateur ;
- Stocker au niveau de lien une plusieurs informations : types de communication (email, etc), nombre ou fréquence d'interaction, la durée, nombre de participants, etc.

Domaines d'application

- Étudier des propriétés de la structure et son rôle ainsi que la position et le prestige de chaque acteur social
  - *le marketing viral*,
  - détection des nœuds faibles (ou ponts) dans un réseau (virus informatiques, terrorisme, etc.),
  - *tri des résultats d'un moteur de recherche (le fameux PageRank)*.
- Recherche des différents types de sous-graphes : les communautés formées par des groupes d'acteurs ayant des intérêts commun.
  - isoler les groupes de densités élevées.
- Élaboration de recommandations (algorithmes d'exploration de chemins, d'analyse de degrés) :
  - Trouver un expert dans un domaine donné,
  - Suggérer des produits à vendre, proposer un ami, etc.

Les mesures globales

- La densité du graphe : permet de décrire la connectivité à l'intérieur du graphe représentant le réseau :

$$D = \frac{m}{\frac{n*(n-1)}{2}}$$

  où m étant le nombre de liens dans le graphe, n étant le nombre de nœuds,
- *La distance géodésique* entre deux nœuds est le plus court chemin entre les deux nœuds.
- La distance moyenne d'un graphe connecté est égale à la moyenne *des distances géodésiques* entre toutes les pairs d'acteurs.
- Le diamètre d'un graphe connecté est égal à *la distance géodésique* maximale au sein d'un groupe.

Quelques graphes de terrain

TABLE: Exemples analysés dans [12]

Graphe	Nœuds	Liens	Jeu de données
Internet	Routeurs	Liaisons de données	http ://www.isi.edu/div7/scan/mercator/maps.html.
Web	Pages	Hyperliens	http ://snap.stanford.edu/data/web-NotreDame.html
Acteurs	Acteurs	Même film	http ://www.imdb.com/.
Co-publication	Auteurs	Co-signature	http ://arxiv.org/.
Co-occurrence	Mots	Même phrase	http ://www.tniv.info/bible/
Protéines	Protéines	Influence	http ://www.nd.edu/ networks

Quelques graphes de terrain : Mesures

TABLE: Exemples analysés dans [12]

Graph	$\|X\|$	$\|E\|$	densité	d	γ	cc
Internet	75885	357317	1.2e-4	5.80	2.5	0.171
Web	325729	1090108	2.1e-5	7	2.3	0.466
Acteurs	392340	15038083	1.9e-4	3.6	2.2	0.785
Co-publication	16401	29552	2.2e-4	7.18	2.4	0.638
Co-occurrence	9297	392066	9.1e-3	2.13	1.8	0.822
Protéines	2113	2203	9.9e-4	6.74	2.4	0.153

Mesures locales : Degré de centralité

- Acteur central est l'acteur le plus actif de point de vue communication,
- Graphe non dirigé, contenant n nœuds, $d(i)$ étant le degré du nœud i :

$$C_D(i) = \frac{d(i)}{n-1}$$

- Graphe dirigé, contenant n nœuds, $d_o(i)$ étant le degré (liens sortants) du nœud i :

$$C'_D(i) = \frac{d_o(i)}{n-1}$$

Mesures locales : Centralité de proximité

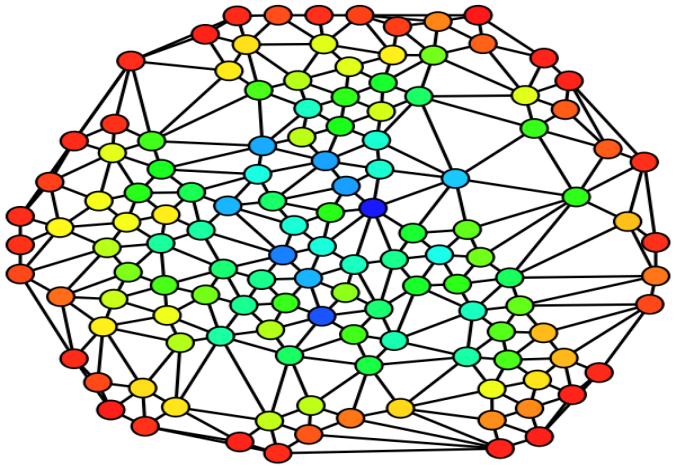
- La notion de centralité utilise la distance.
- Acteur central est l'acteur qui communique *facilement* avec les autres,
- $d(i,j)$ est la distance entre deux acteurs mesurée en nombre minimal de liens.
- Graphe non dirigé, contenant n nœuds :

$$C_C(i) = \frac{n-1}{\sum_{j=1}^{n} d(i,j)}$$

- Graphe dirigé La distance doit prendre en compte les sens des liens.

Centralité d'intermédiarité-1

- Deux nœuds *non adjacents* k & j qui se communiquent et si le nœud i se trouve sur le chemin de communication : i est un acteur itérmédiaire.



source :http ://en.wikipedia.org/wiki/Centrality

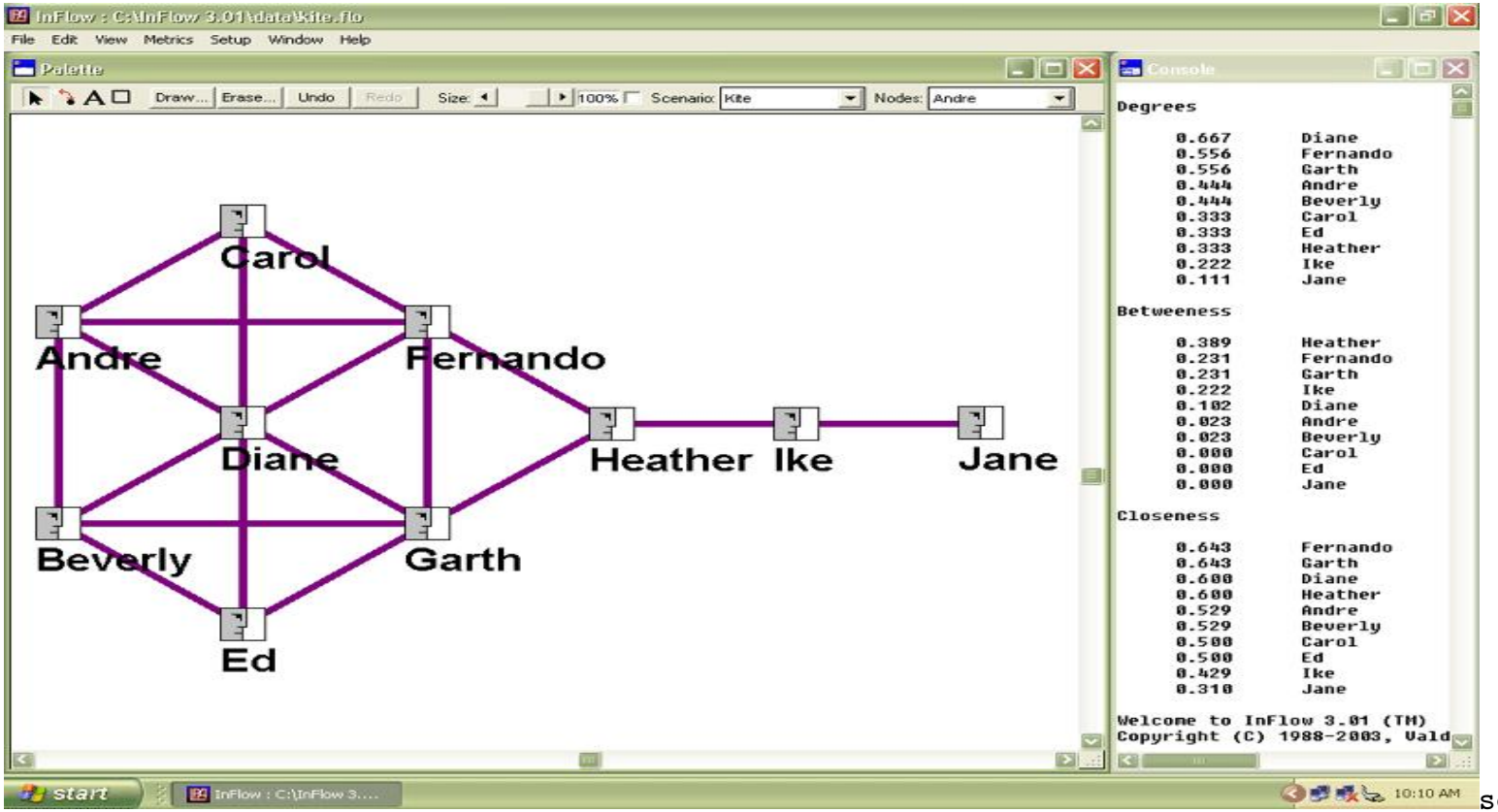
- $C_B(i) = \sum_{j<k} \frac{p_{jk}(i)}{p_{jk}}$
  - p_{jk} le nombre des chemins les plus cours entre j et k,
  - $p_{jk}(i)$ le nombre des chemins les plus cours entre j et k passant par i.

Centralité d'intermédiarité-2

- Graphe dirigé :

$$C'_B(i) = \frac{\sum_{j<k} \frac{p_{jk}(i)}{p_{jk}}}{(n-1)(n-2)}$$

Illustration des mesures de centralité



source :

http ://www.orgnet.com/sna.html

Le prestige

- Le prestige d'un acteur est mesuré par les liens entrants.
- Les mesures :
  - le degré de prestige,
  - le prestige de proximité,
  - le prestige de tri (rank prestige).

Le prestige - mesures

- Le degré :

$$P_D(i) = \frac{d_I(i)}{n-1}$$

- La proximité :
  - I_i est l'ensemble d'acteurs qui atteignent l'acteur i : il existe un chemin de j vers i.
  - $d(j,i)$ le chemin le plus court de j vers i.
  - La distance moyenne est $\frac{\sum_{j \in I_i} d(j,i)}{|I_i|}$
  - Le prestige de proximité : $P_p(i) = \frac{\frac{|I_i|}{n-1}}{\frac{\sum_{j \in I_i} d(j,i)}{|I_i|}}$

"Rank Prestige"

- Considérons la réputation des personnes choisissant l'acteur i
- $P_R(i) = A_{1i}P_R(1) + A_{2i}P_R(2) + .. + A_{ni}P_R(n)$
  - A_{ij} vaut 1 si i pointe vers j,sinon 0.
  - On pose P le vecteur contenant les valeurs de tri :

$$P = (P_R(1), P_R(2), .., P_R(n))^T$$

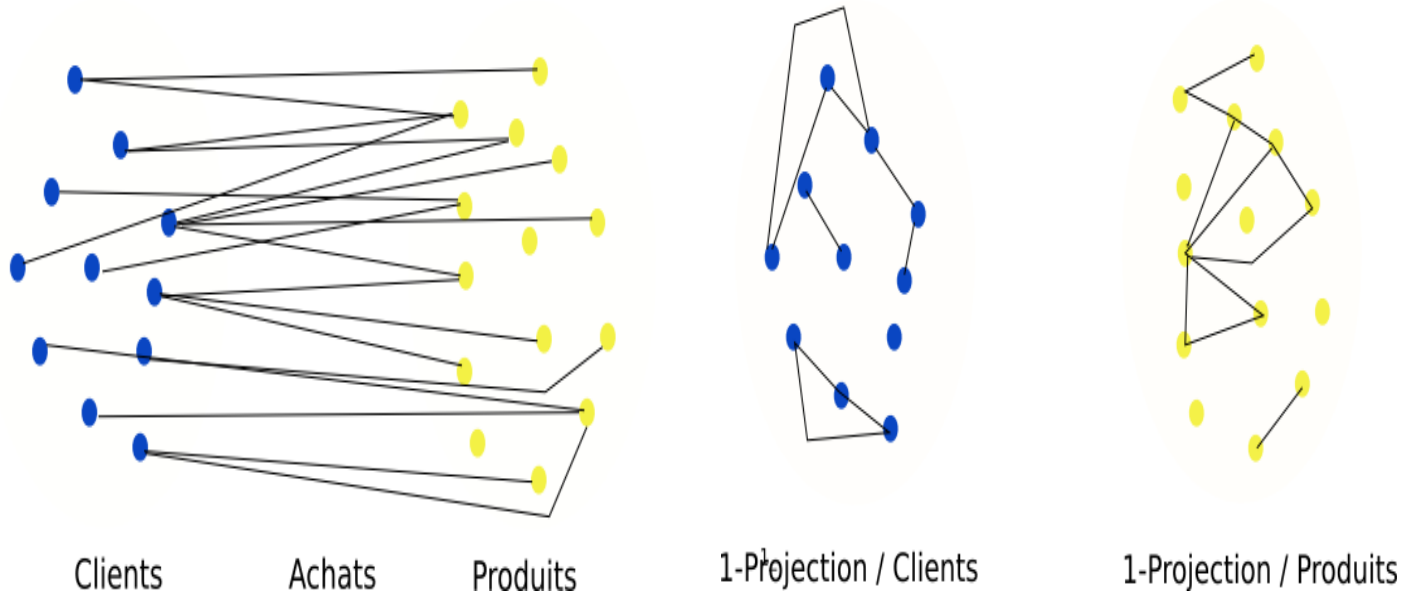
  - A est la matrice exprimant la propriété adjacente

$$P = A^T P$$

.

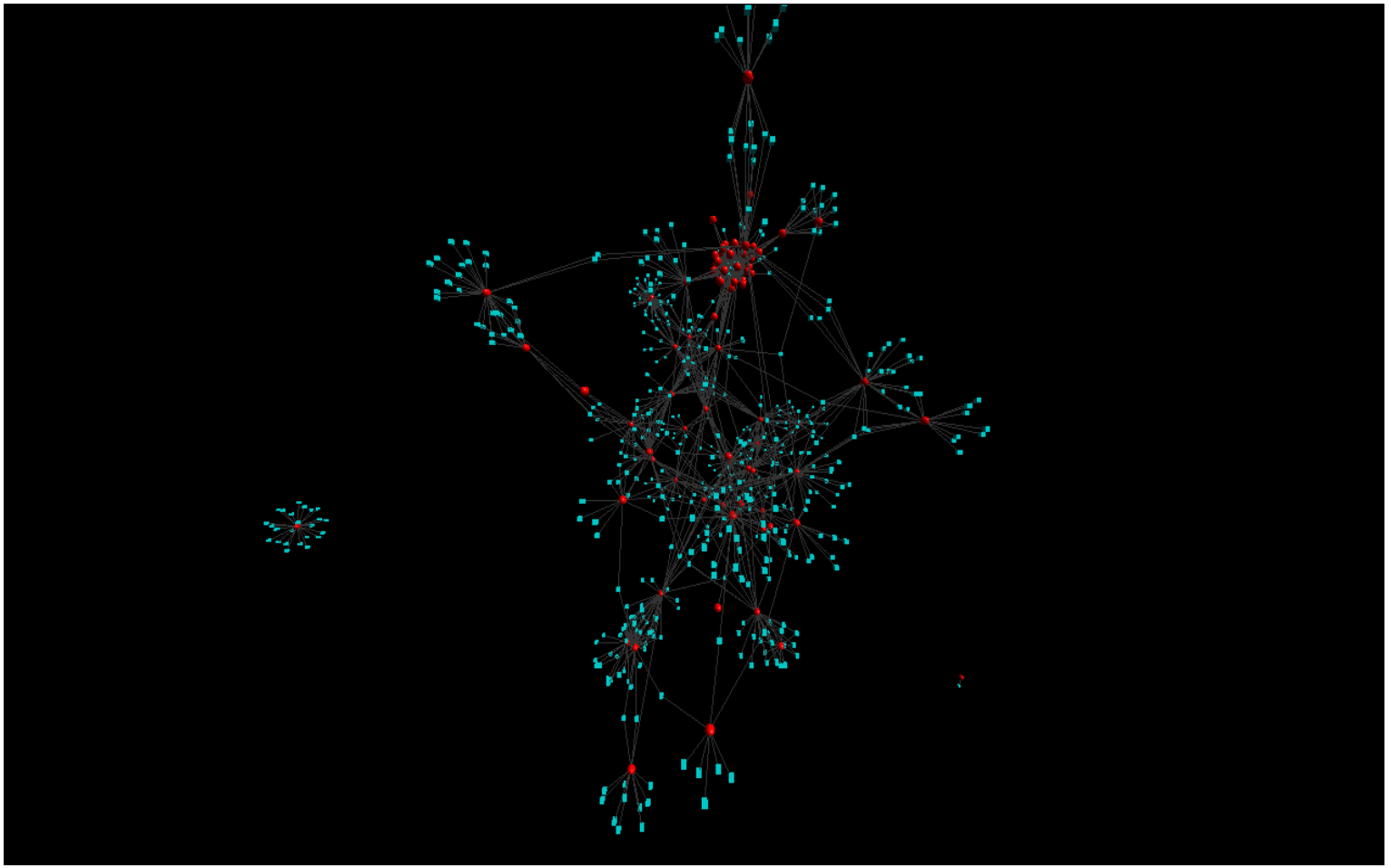
  - Recherche de valeurs propres de A^T.

Graphes Bipartites : système de recommandation



- Recommandation de produits = prédiction de liens dans le graphe biparti [2].

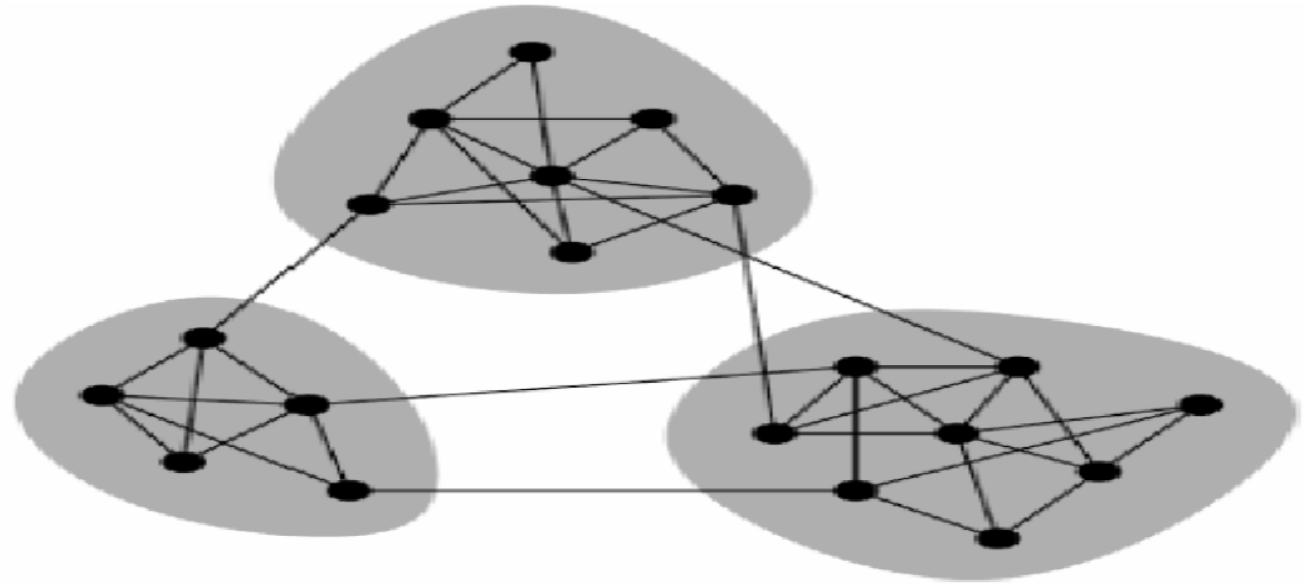
Graphes Bipartis : Exemple



source : Graphe de transactions durant un mois sur un site de musique en-ligne

Communauté : définition

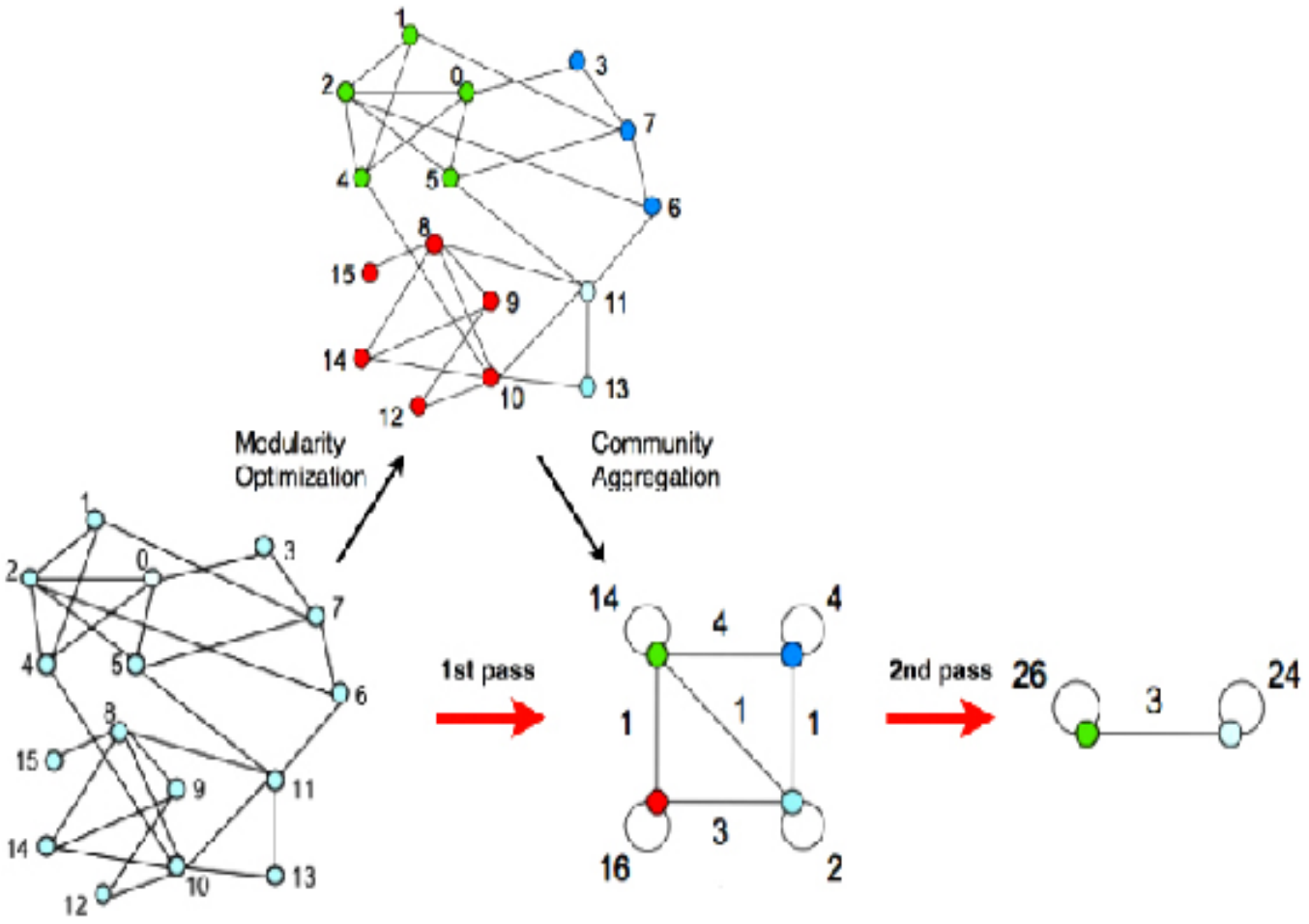
- Un sous-graphe dont les nœuds sont plus liés entre eux qu'avec d'autres nœuds.
- Exemples : un ensemble de pages web traitant un même thème, un ensemble d'amis, un ensemble de gènes dédiés à une même fonction, etc.



Recherche de communautés

- **Recherche de communautés** : partitionnement du réseau social en N communautés.
  - *Approches agglomératives* : au début, chaque nœud est considéré comme une communauté. Puis, on regroupe les petites communautés sur la base de leur *proximité* [4].
    - Modularité : critère de qualité d'un partitionnement du graphe
    - $Q(\pi) = \sum_{s \in \pi} \left(\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right)$
    - L est le nombre de liens total dans le réseau, d_s est la somme des degrés des nœuds dans la partition s, l_s est le nombre de liens dans s.
  - *Approches sépartives* : recherche des composantes connexes du graphe en retirant progressivement des liens.

Algorithme agglomératif



Approche séparative

- Approche basée sur le calcul de *la centralité d'intermédiarité* de liens [20].
- Répéter
  - Calculer les intermédiarités de tous les liens,
  - Couper le lien avec l'intermédiarité maximale
- jusqu'à ce qu'il ait plus de liens.

ARS : Autres Problèmes d'analyse

- Problématique des liens
  - Liens caché : origine : information manquante, dissimulation d'information.
  - Complétion de liens [11].
  - Liens alarmants : lines normaux ou anormaux.
- Evolution du réseau
  - Prédiction de liens.
  - Evolution des paramètres des réseaux : diamètre décroissant et densification des degrés dans des réseaux de co-citations.
  - Evolution des communautés par taille et des communautés de thèmes.

Bibliographie I

Ecole d'été web intelligence.

In *WI09*. Université de Lyon, 2009.

Nesserine Benchettara, Rushed Kanawati, and Céline Rouveirol.

Calcul de recommendation par prédiction de liens dans un graphe biparti.

In *Actes de l'ateiler sur l'apprentissage et graphes pour les syst ?mes complexes (plate-forme AFIA 2009)*, Hammemt, Tunisie, Mai 2009.

C. Berge.

Graphes.

Gauther-Villars, 1983.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre.

Fast unfolding of community hierarchies in large networks.

CoRR, abs/0803.0476, 2008.

Bibliographie II

S. P. Borgatti, M. G. Everett, and L. C. Freeman.

Ucinet for Windows : Software for Social Network Analysis, 2002.

G. Ereteo, F Gando, M. Buffa, and P. Grohan.

Analyse des réseaux sociaux et web sémantique : un état de l'art.

Technical report, Projet ISICIL (ANR), 2009.

M. G. Everett and S. P. Borgatti.

The centrality of groups and classes.

Journal of Mathematical Sociology, 23(3) :181–201, 1999.

J-C. Fournier.

Théorie de Graphes et applications.

Lavoisier, 2006.

Bibliographie III

Linton C. Freeman.

A set of measures of centrality based on betweenness.

Sociometry, 40(1) :35–41, 1977.

Linton C. Freeman.

Centrality in social networks : Conceptual clarification.

Social Networks, 1(3) :215–239, 1979.

A Goldenberg, J. Kubica, P. Komarek, A. Moore, and J. Schneider.

A comparaison of statistical and machine learning algorithms on the task of link completion.

In *Proceedings of the KDD workshop on link analysis for detecting complex Behavior*, August 2003.

Jean-Loup Guillaume and Matthieu Latapy.

Bipartite graphs as models of complex networks. 371, pages, 2006.

Physica A, 37(1) :795–813, 2006.

Bibliographie IV

Rushed Kanawati and Céline Rouveirol.

Prédiction des liens dans les réseaux sociaux.

In *Tutoriel EGC*, 2010.

Bing Liu.

Web Data Mining : Exploring Hyperlinks, Contents, and Usage Data.

Data-Centric Systems and Applications. Springer, 2007.

Jean loup Guillaume and Matthieu Latapy.

Bipartite structure of all complex networks.

Information processing letters, 90 :215–221, 2004.

P. Mika.

Bootstrapping the foaf-web : An experiment in social network mining, 2004.

Bibliographie V

Peter Mika.

Flink : Semantic web technology for the extraction and analysis of social networks.

Web Semantics : Science, Services and Agents on the World Wide Web, 3(2-3) :211–223, October 2005.

Peter Mika.

Ontologies are us : A unified model of social networks and semantics.

In *International Semantic Web Conference*, LNCS, pages 522–536. Springer, 2005.

M. E. J. Newman.

The structure and function of complex networks.

Mar 2003.

M. E. J. Newman.

Coauthorship networks and patterns of scientific collaboration.

Proceedings of the National Academy of Science of the United States (PNAS), 101 :5200–5205, 2004.

Bibliographie VI

Jun Zhang, Mark S. Ackerman, and Lada Adamic.

Expertise networks in online communities : structure and algorithms.

In *WWW '07 : Proceedings of the 16th international conference on World Wide Web*, pages 221–230, New York, NY, USA, 2007. ACM.