Analyse des réseaux sociaux : une introduction

Analyse de graphes de terrain

Rushed Kanawati

LIPN, CNRS UMR 7030
Université Paris 13
http://lipn.fr/∼kanawati
rushed.kanawati@lipn.univ-paris13.fr

October 24, 2016

Objectifs & organisation du cours

- Objectifs : Présentation des techniques d'étude et d'analyse de grands `réseaux complexes`.
- Fouille et analyse des réseaux d'interactions.
- 10 séances : Cours + TP
- Outils :

Python	http://www.python.org/
R	https://www.r-project.org/
igraph	http://igraph.wikidot.com
Gephi	http://gephi.org
Tulip	http://tulip.labri.fr/TulipDrupal/

Thèmes abordées

Graphes de terrains Modèles, caractéristiques & problèmes d'analyse.

Caractéristiques topologiques de nœuds Présentation des différentes mesures de tri de nœuds.

Détection de communautés Définitions, problématique, fonctions de qualité, classification des approches & applications.

Analyse de réseaux dynamiques Problématiques & Application, Prévision de liens,

Modèles d'influence et de de diffusion

Analyse des réseaux Multiplexes

Graphes de terrains

1. Définitions & notations
2. Caractérisation & exemples
3. Modèles
4. Problématiques d'analyse

Graphes de terrain : réseau de collaborations scientifiques

Un réseau social est un ensemble d'acteurs relié par des liens/interactions sociales

Graphe de terrain : Réseau social ?

Twitter Friends van Belgische Twitteraars

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

Notations

- Un graphe $G =< V, E \subseteq V \times V >$:
  - V est l'ensemble de nœudes (i.e. acteurs)
  - E est l'ensemble de liens.
- Notations :
  - A_G est la matrice d'adjacence de G : $a_{ij} \neq 0$ si les nœuds $(v_i, v_j) \in E$, 0 sinon.
  - $n = |V|$
  - $m = |E|$
  - $\Gamma(v)$ est l'ensemble de voisins de v. $\Gamma(v) = \{x \in V : (x, v) \in E\}$.
  - Le degré d'un nœud $d(v) =\| \Gamma(v) \|$

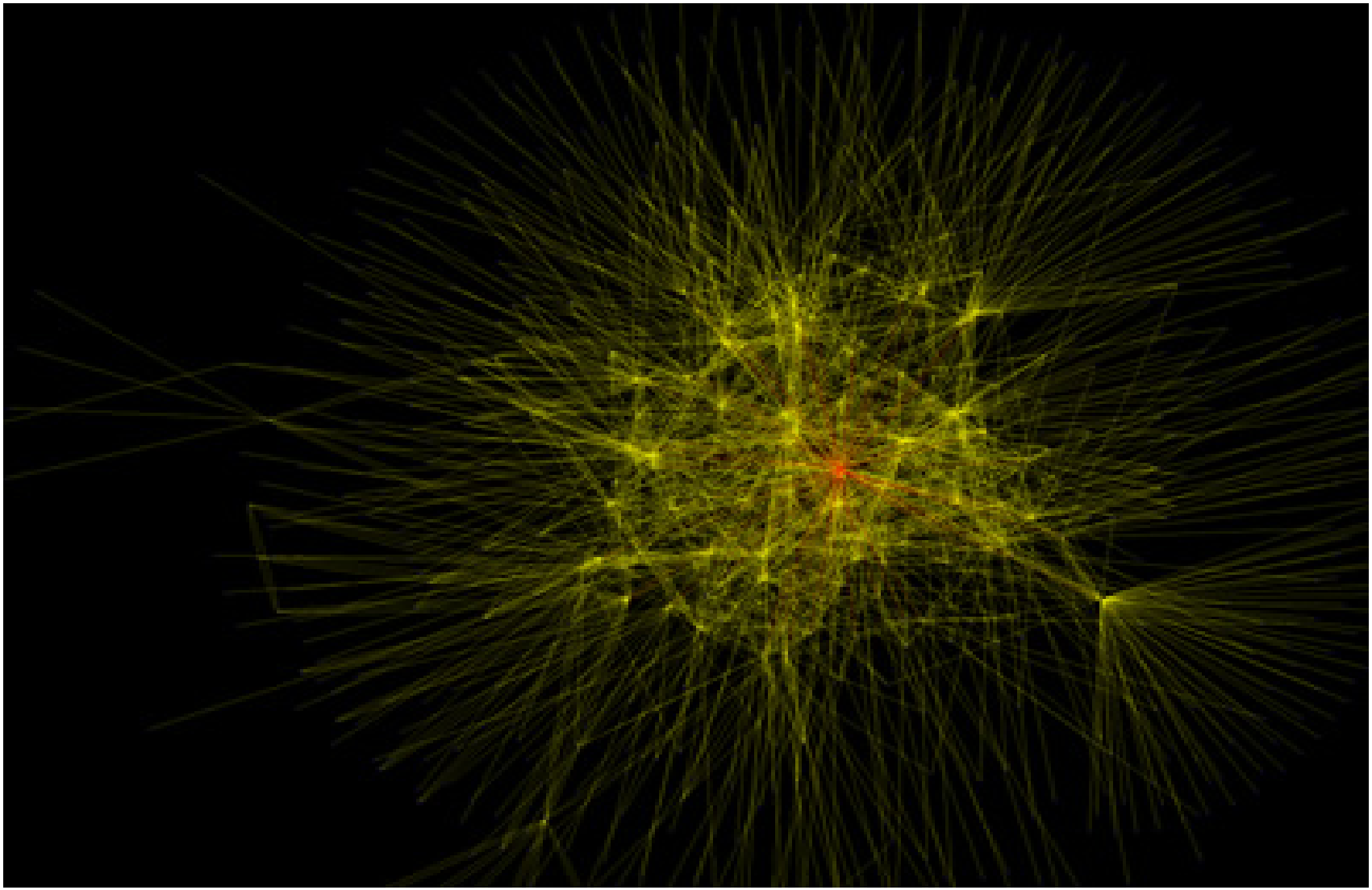
Propriétés de base

- Chemin v_i, v_j : liste de liens à traverser pour aller de v_i à v_j.
- Distance v_i, v_j : la longueur de plus court chemin reliant V_i à v_j.
- Connexité : Un graphe est connexe si il existe un chemin entre n'importe quel pair de nœuds.
- Diamètre d'un graphe : la plus grande distance entre des nœuds du graphe.
- Densité du graphe : Probabilité d'existence de tous liens : $\frac{2 \times m}{n \times (n-1)}$

Réseaux cibles : Exemples

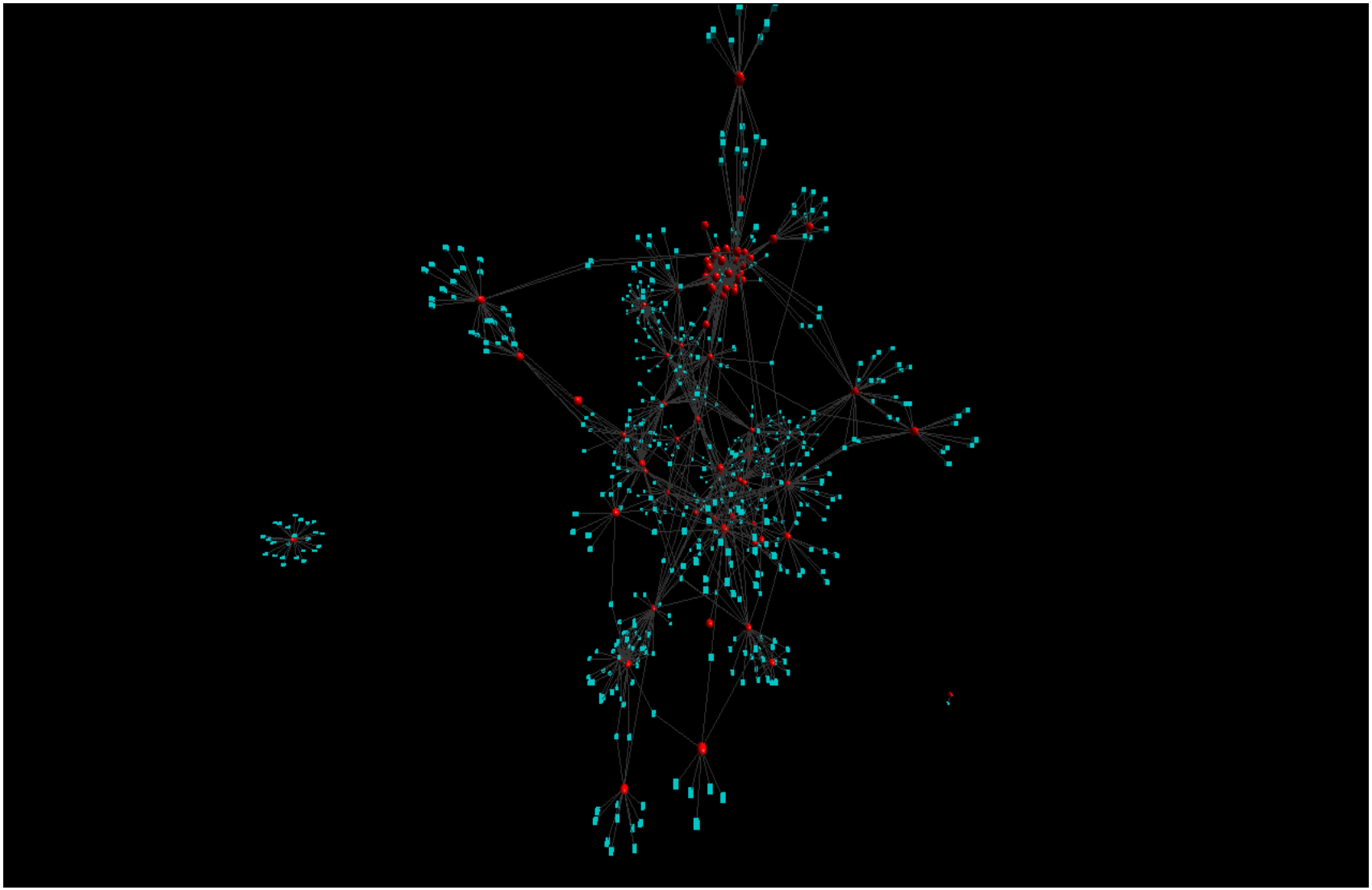
- Réseaux bibliographiques : publications, citation [11, 12]
- Réseaux sociologiques : Interaction sur forums/blogs [6, 10], communication [14], site de rencontres [4], terroriste
- Réseaux éthologiques : Comportement des Zèbres
- Réseaux biologiques : Interactions entre protéines, Métabolique, Interactions entre des gènes.
- Réseaux technologiques : Internat, Web
- ...

Réseaux cibles



source : www.visulacomplexity.com

Réseaux cibles



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

Réseaux cibles : caractéristiques

- **Grande Taille** : $|V| > 10^3$.
- **Faible degré de séparation** : la moyenne des distances géodésiques entre chaque couple de nœuds est *logarithmique* en fonction de $\| V \|$
- **Sans échelle** : majorité de nœuds ayant peu de connexions mais un nombre substantiel de nœuds ont beaucoup de connexions.

  La probabilité pour un nœud $v \in V$ d'avoir k voisins est : $P(k) = k^{-\gamma}$
- **Coefficient de clustering elevé** : la probabilité que deux voisins d'un nœud choisis aléatoirement soient eux-mêmes connectés est assez grande.

$$cc(G) = \sum_{v \in V} \frac{2|E \cap (\Gamma(v) \times \Gamma(v))|}{d(v) \times (d(v) - 1)}$$

Le mythe de six degrés de séparation

Satnley Milgram

(1933-1984)

- Expérience de Milgram (1967)
  - Choisir une personne cible (courtier à Boston).
  - Demander à des groupes de personnes choisies alétoirement (de Boston, de Nebraska, des actionnaires au Nebraska) de lui envoyer une lettre à travers des personnes susceptibles de le connaître.
  - 217 lettres sont envoyées ; 64 sont arrivées (succès : 0.29)
  - La longueur moyenne d'un chemin est de **5.2** intermédiaires.
- Expérience similaire mais au niveau mondial et en utilisant le courrier électronique
  - 24 000 participants de 166 pays : 384 e-mail arrivés à destination (succès 1.6 %)
  - Longueur moyenne des chemins : **5**.

Autre exemple: Le nombre d'Erdös

Paul Erdös

(1913-1996)

- Paul Erdös est un mathématicien qui a publié plus de 1500 articles !
- On mesure l'importance d'un chercheur (en mathématique) par la distance qui le sépare d'Erdös dans le graphe de collaborations
- Site Web : `www.oakland.edu/enp`
- Degré de séparation entre chercheurs 2-Erdös est de 5 !

Quelques graphes de terrain

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

Graph	$\|V\|$	$\|E\|$	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

Modèles de réseaux petit-monde

Approches d'échantillonnage : échantillonner des liens à partir d'un graphe complet

- Erdös-Renyi
- Molloy et Reed

Approches génératives : règles d'évolution de réseau

- Modèle de l'anneau
- Modèle de l'attachement préférentiel
- Le modèle HOT

Erdös-Rényi

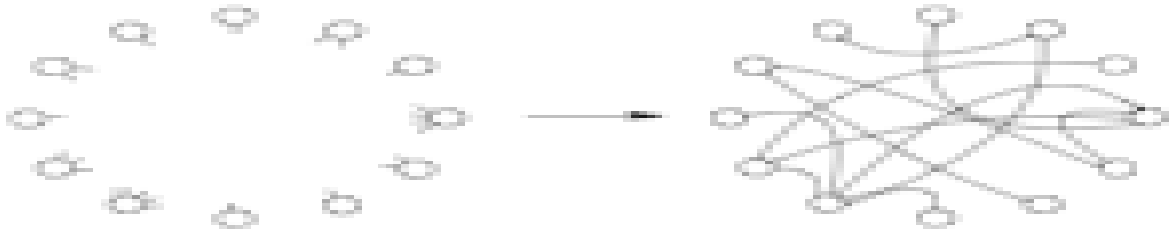
Alfréd Rényi

(1921-1970)

- Principe : pour modéliser un graphe composé de n nœuds et m liens, on choisit aléatoirement m liens, avec une probabilité p parmi l'ensemble de liens possibles ($\frac{n(n-1)}{2}$)
- Le modèle génère des réseaux à faible degré de séparation
- Les graphes générés ne sont pas sans échelle
- Coefficient de clustering $= p$

Modèle de Molley et Reed

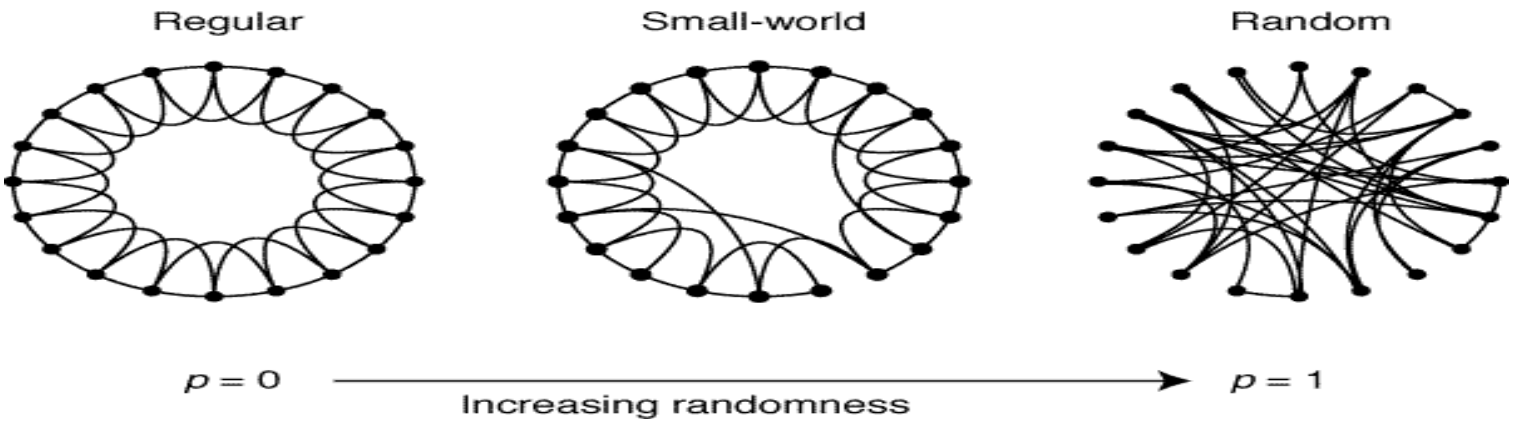
- Idée : forcer l'obtention d'une loi de puissance pour la distribution des degrés.



- Génération de graphes sans échelle et à faible degré de séparation.
- Coefficient de clustering $cc \to 0$

Modèle de l'anneau

- Les nœuds sont placés sur un anneau virtuel où chaque nœud est connecté aux k nœuds les plus proches.
- On reconnecte chaque arête avec une probabilité p à un nœud choisi aléatoirement.



- Problème : la distribution des degrés ne suit pas une loi de puissance

L'attachement préferentiel

- Les nœuds sont ajoutés un à un.
- La probabilité de relier un nouveau nœud à un nœud existant dépend du degré du nœud.
- Principe : *enrichir les riches !*
- Corrélation entre le degré d'un nœud et son âge.
- Correction : probabilité d'un lien en fonction du degré des nœuds partenaires et d'une estimation d'utilité.
- Les graphes obtenus sont sans échelle mais $cc \to 0$

Le modèle HOT

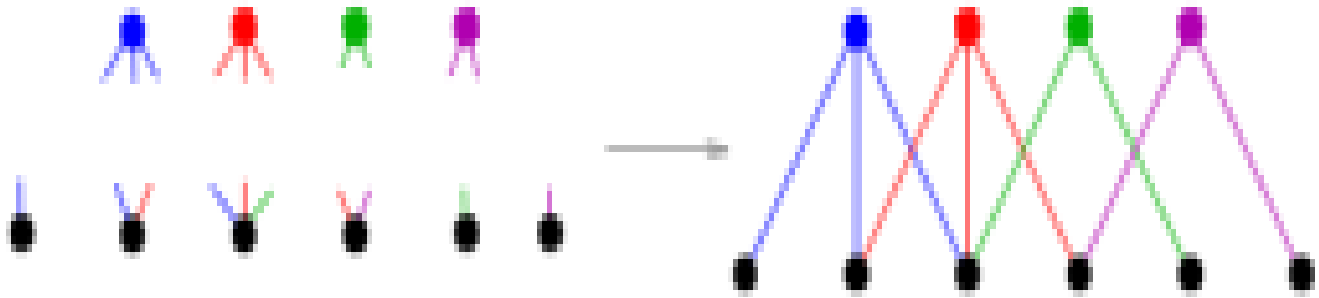
- HOT : Highly Optimized Tolerance
- Placer aléatoirement les nœuds sur une grille.
- Les nœuds sont ajoutés un à un.
- Dans le graphe G, un nouveau nœud n est connecté à un nœud existant a de sorte à :
  - Minimiser $\alpha(\| G \|)d_G(a, n)$
  - Maximiser $c_p(a) = \frac{1}{\sum_{u \in V} d(a,u)}$
- $\alpha <$ *constant* : le graphe est une étoile.
- $\alpha > \sqrt{(\| G \|)}$: la distribution de degrés est exponentielle.
- *constant* $< \alpha < \sqrt{(\| G \|)}$: Graphe sans échelle.
- Même problème : $cc \to 0$

Modèle de graphes bipartis

- Beaucoup de réseaux réels sont des graphes **bipartis** (ex. bibliographique, achats, etc.)
- Graphe biparti: $G = (\top, \bot, E \subseteq \top \times \bot)$. $\top$, $\bot$ sont deux ensembles distincts.
- Projections :
  - $G_{\top}^{n} = (V_{\top} \subseteq \top, E_{\top} = \{x, y \in \top : |\Gamma_G(x) \cap \Gamma_G(y)| \geq n\}$
  - $G_{\bot}^{m} = (V_{\bot} \subseteq \bot, E_{\bot} = \{x, y \in \bot : |\Gamma_G(x) \cap \Gamma_G(y)| \geq m\}$
  - n, m : paramètres des projections.
- La projection augmente artificiellement le coefficient de clustering du graphe projeté !!

Modèle de graphes bipartis

- Utiliser deux lois de puissance définies pour deux ensembles de nœuds.



- Les graphes obtenus sont sans échelle, à faible degré de séparation, avec un coefficient de clustering proche des cas réels.

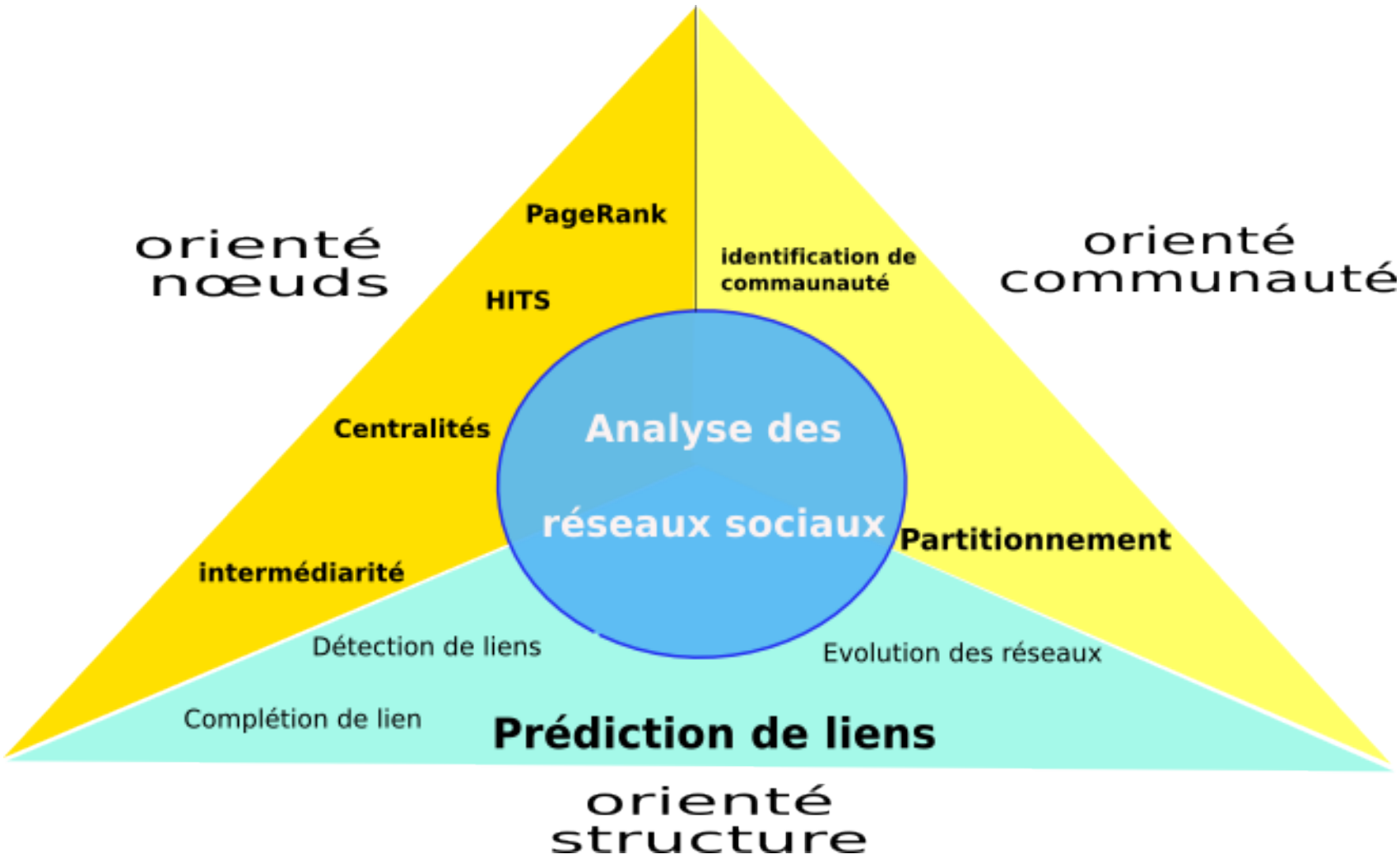
Modèles petit-monde : bilan

Modèle	Faible séparation	Sans échelle	Clustering
Erdös-Renyi	OUI	NON	NON
Molley et Reed	OUI	OUI	NON
Anneau	OUI	NON	OUI
Attachement préférentiel	OUI	OUI	NON
HOT	OUI	OUI	NON
Biparti	OUI	OUI	OUI

Problématique de l'étude des réseaux sociaux

- Extraction et modélisation de réseaux.
- Support sytème : réseaux distribués, protection de vie privée, interopérabilité.
- Navigation et recherche dans les réseaux sociaux
- Visualisation des réseaux
- **Fouille de réseaux (Analyse de réseaux sociaux : SNA)**

Analyse de réseaux sociaux : les tâches



Tâches orientés nœuds

- But : caractériser le rôle et/ou la position d'un nœud dans le réseau.
- Trier les nœuds selon leurs caractéristiques.
- Inférer l'importance d'un acteur dans un réseau :
  - *PageRank* employé par Google pour trier les résultats renvoyés par le moteur de recherche.
  - Marketing Viral,
  - Evaluation des chercheurs !!
  - Détection des nœuds faibles (ou ponts) dans un résesau (virus informatiques, terrorisme, etc.)
- *Les mêmes tâches peuvent être définies pour les liens*.

Tâches orientées communautés

- Communauté : sous-graphe à forte densité locale.
- Deux principales tâches :
  - **Recherche de communautés** : partitionnement d'un réseau en communautés
  - **Identification de communauté** : détection de la communauté d'un nœud donné.
- Approches similaires à la classification de données.

Tâches orientées structure

- Découvrir les lois de l'évolution du réseau.
- Problème phare : prédiction de liens.
- Soit $G =< G_1, G_2, \ldots, G_T >$ un réseau temporel de réseaux sociaux, G_i est le graphe du réseau à l'instant i.
- La tâche de prédiction de liens consiste à prédire G_{T+1} à partir de G.
- Autres problèmes :
  - Détection de liens cachés
  - Complétion de liens
  - Détection de liens alarmants (anormaux)

Caratérisation des nœuds

- Notion de **Centralité** :
  - *de degrés*.
  - *de proximité*.
  - *d'intermédiarité*.
  - *de vecteur*.
- HITS : identifier des *autorités* et des *hubs*.
- PageRank : *prestige* d'un nœud.

Centralité de degré

- Un nœud ayant beaucoup de liens est considéré comme important.
- Un *acteur central* est l'acteur le plus actif de point de vue de communication,
- $C_d(v) = \frac{\|(\Gamma(v)\|}{max_{u \in V} \|\Gamma(u)\|}$

Centralité de proximité

- Le nœud qui communique le plus facilement avec les autres nœuds est important.
- Centralité de proximité : $c_p(v) = \frac{1}{\sum_{u \in V} d(v,u)}$

Centralité d'intermédiarité

- Le nœud qui rapproche le plus les autres nœuds est important.
- Nombre de chemins géodésiques qui passent par v.
- $ch_g(x, y)$: les nœuds qui se situent sur le plus court chemin reliant x à y.
- $C_i(v) = \frac{\|ch(x,y)\|x,y \in V, v \in ch(x,y)}{\|ch(x,y)x,y \in V\|}$

Centralité de vecteur

- La centralité d'un nœud dépend des centralités de ses voisins.
- $C(v) = \frac{1}{\lambda} \sum A_{i,j} C(x_j) : AX = \lambda X.$

HITS

- HITS : Hyperlink Induced Topic Search
- Indicateurs proposés pour implanter un moteur de recherche sur le Web (Clever d'IBM).
- Web : graphe dirigé
- Deux types de pages (i.e. nœuds) :
  - **Autorité** : meilleures sources d'information sur un théme.
  - **Hub** : page fournissant des liens vers des autorités de qualité sur un thème.
- Une autorité de qualité est citée par de nombreux hubs de qualité.
- Un bon hub pointe vers beaucoup d'autorités de qualité.

HITS : l'algorithme

- Soit G un graphe connexe, z le vecteur unité de $\Re^n$
- $x_0 \leftarrow z$
- $y_0 \leftarrow z$
- Répéter jusqu'à convergence ou au max k fois :
  - $x_i^{<p>} = \sum_{\forall q:q \to p} y_{i-1}^{<p>}$
  - $y_i^{<p>} = \sum_{\forall q:p \to q} x_{i-1}^{<q>}$

PageRank

- 1998, par Sergey Brin & Larry Page.
- Tri statique des pages Web.
- Le web est traité comme étant un graphe dirigé(V,E).
- La mesure associée à une page i est :

$$P(i) = \sum_{(j,i) \in E} \frac{P(j)}{O_j}$$

où O_j est le nombre de liens sortant de la page j

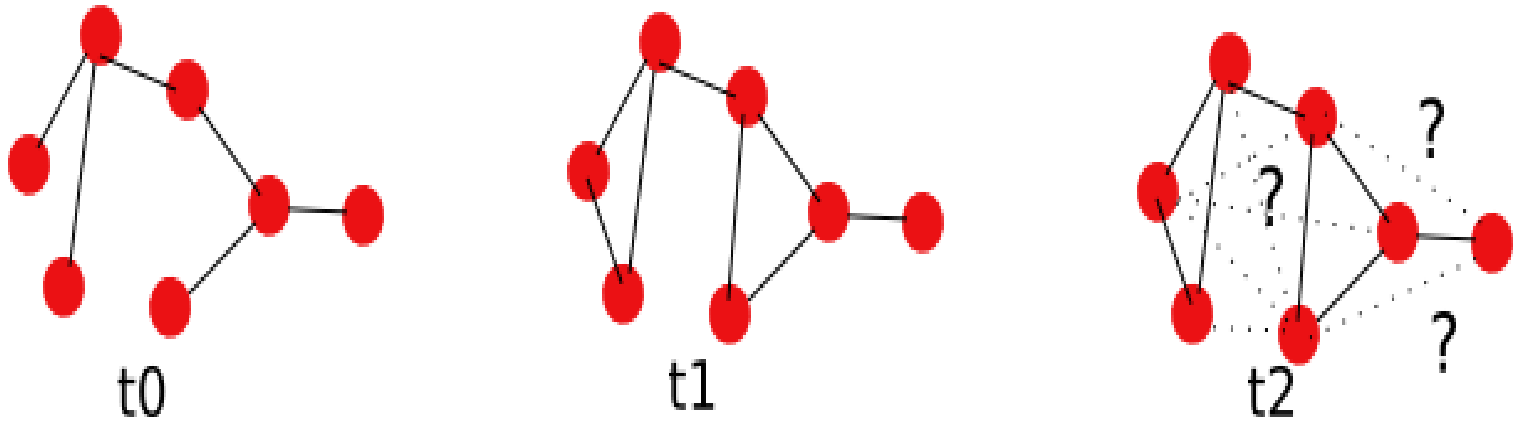
- On pose P le vecteur contenant les valeurs de tri :

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

Calcul de PageRank

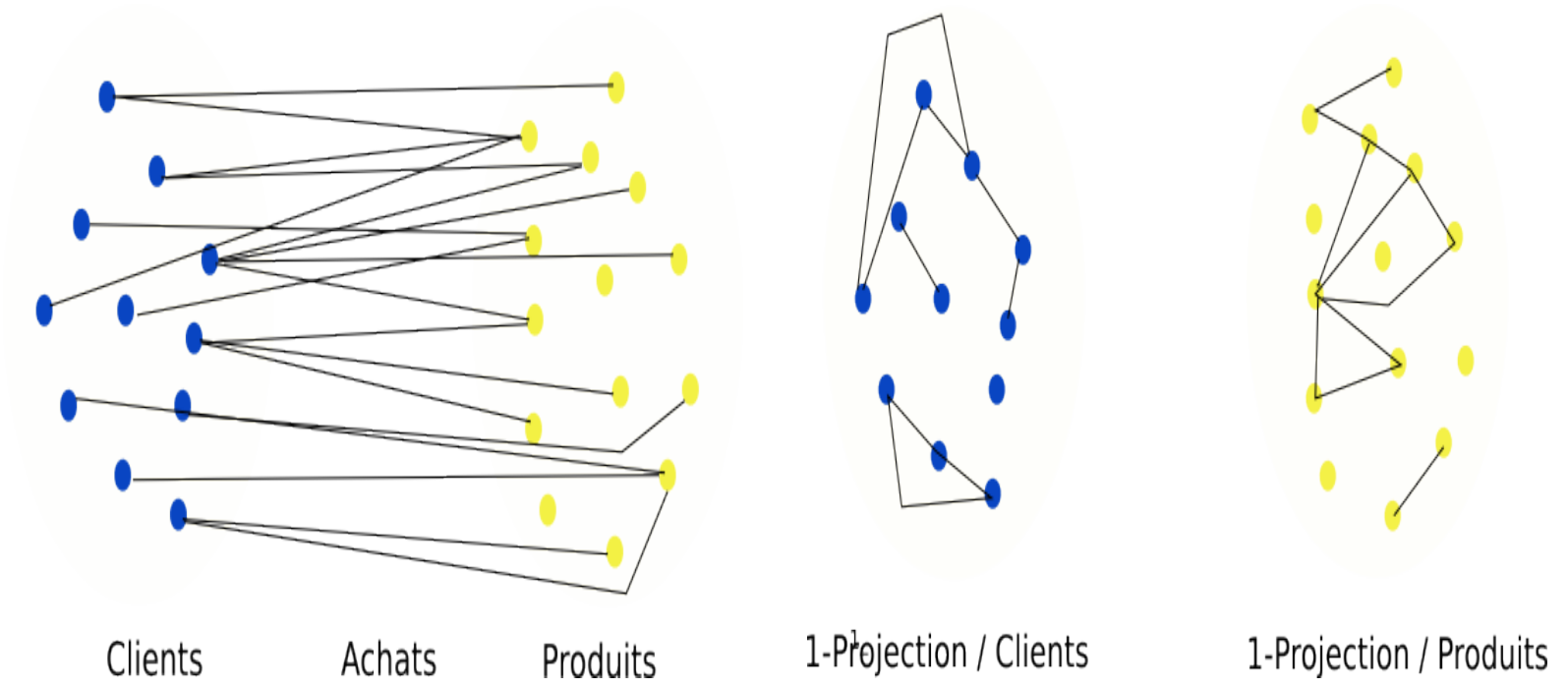
- $s \leftarrow$ vecteur aléatoire.
- $r \leftarrow A^T \times s$
- Tant que $\| r - s \| < \epsilon$ faire :
  - $s \leftarrow r$
  - $r \leftarrow A^T \times s$

Prvision de liens : définition



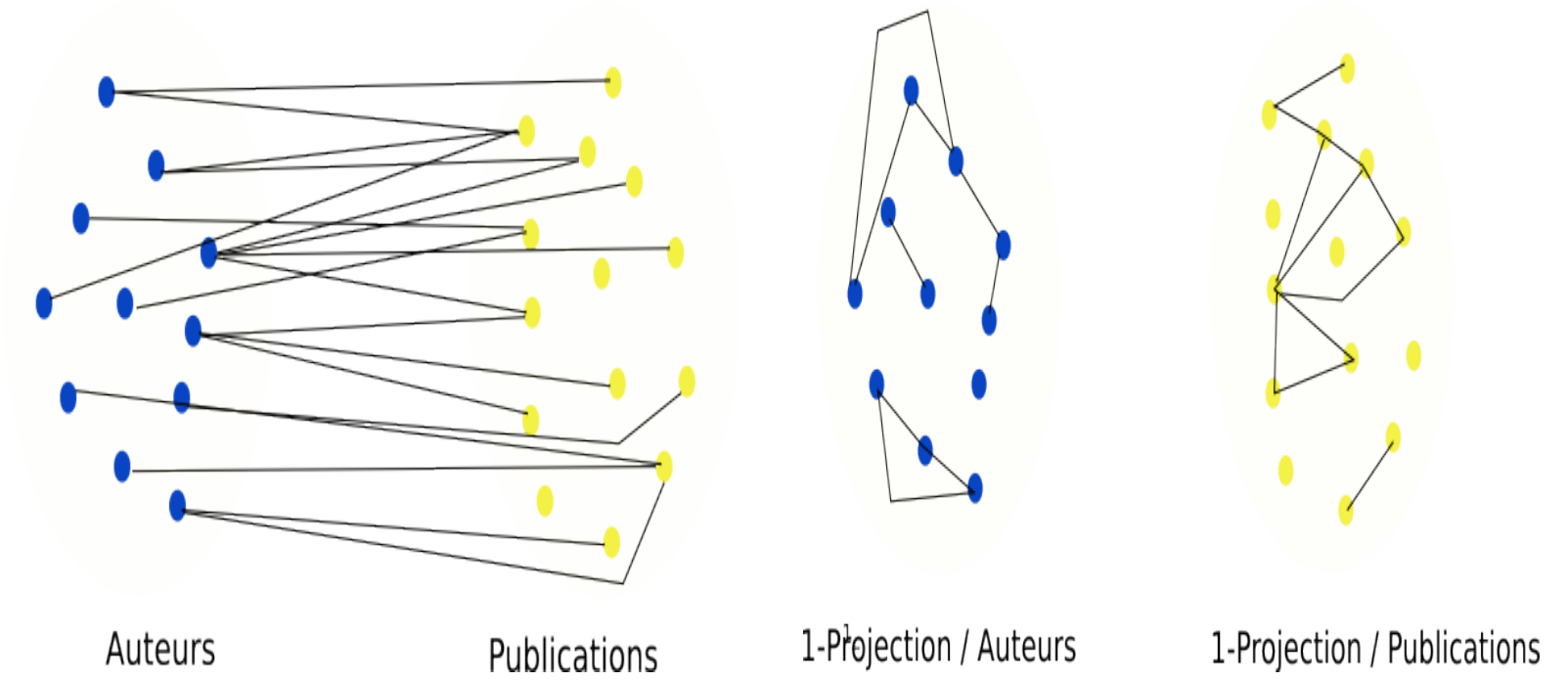
- Définition informelle : prédire la formation d'un lien entre deux nœuds jamais connectés auparavant.
- Soit $G =< G_1, G_2, \ldots, G_T >$ un réseau temporel de réseaux sociaux, G_i est le graphe du réseau social à l'instant i. La tâche de prédiction de liens consiste à prédire pour chaque couple $v, u \in \bigcap_{i=1}^{T} V_i$ tels que $\nexists E_j(u, v) \in E_j$ si $(u, v) \in E_{T+1}$.

Application : système de recommandation



- Recommandation de produits = prédiction de liens dans le graphe biparti [5, 8, 1].
- L'analyse des réseaux sociaux est une solution au problème posé par les matrices de transactions souvent creuses.

Application : recommandation de collaborations académiques

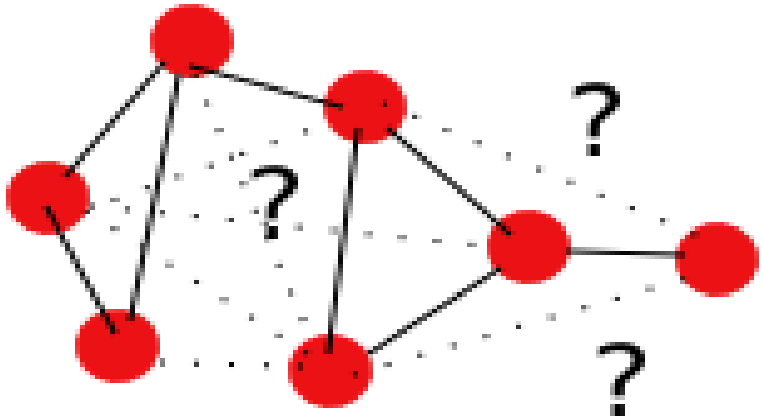


- Recommandation de collaboration = prédiction de liens dans le graphe projeté / auteurs [13, 9].

Autres applications

- Aide à la navigation sur le Web
- Aide à la réponse des questions sur forums, *help-desk*
- Etude de propagation des virus informatiques par e-mail.
- ...

Problème similaire : liens cachés (1)



- Etant donné un réseau $G_t =< V, E >$ observé à l'instant t, la tâche de détection de liens cachés consiste à trouver les liens manquants à l'instant t.
- Origine : information manquante, dissimulation d'information.

Problème similaire : liens cachés (2)

- [2] : Comparer les caractéristiques topologiques des nœuds impliqués en liens cachés (LC) / liens en formation (LF).
  - Méthodologie : éliminer aléatoirement des liens dans un réseau temporel de réseaux sociaux.
  - Pour LC, le nombre de voisins communs est deux fois plus grand que pour LF.
  - Pour LC, le produit des degrés des nœuds impliqués est deux fois plus petit que pour LF.
  - LC : distribution des degrés des nœuds plus étalée.
  - Pas de différence significative en ce qui concne les distances !

Problème similaire: complétion de liens

- Dans un graphe $G =< V, E >$, on considère un nœud $v \in V$ dont on connaît le degré réel $d(v) > d_G(v)$. Le problème consiste à trouver les nœuds $u_i \in V$ auxquels v est susceptible d'être lié [3].
- Exemple : Un client achète 4 livres sur un site de e-commerce mais le nom d'un des livres achetés est perdu dans la transaction. Quel est ce livre ?
- Alice, Bob et une troisième personne ont fait une réunion. Identifier cette personne inconnue en fonction des liens connus.
- Dans [3], on se limite à examiner le cas d'un seul nœud inconnu.

Problème similaire : liens alarmants

- Etant donné un réseau temporel $G =< G_1, G_2, \ldots, G_T >$, le problème est de classifier les nouveaux liens qui apparaissent dans G_T en deux classes : $\{$normal, anormal$\}$, en fonction de l'évolution du réseau
- Problème plus facile que la prédiction de liens.

Problème similaire : évolution des réseaux

Définition de trois niveaux de granularité pour observer et/ou prédire l'évolution d'un réseau.

- Niveau réseau
  - Leskovec et al. , évolution des paramètres des réseaux: diamètre décroissant et densification des degrés dans des réseaux de co-citations.
  - Barabasi et al. , réseaux de co-auteurs, domaine des maths et des neurosciences. Evolution gouvernée par l'attachement préférentiel (lien internes et externes), augmentation du degré moyen et décroissance de la séparation entre les nœuds.
- Niveau communauté : évolution des communautés par taille et des communautés de thèmes

Approches de prédiction de liens

Trois critères de classification :

- **Approche : Dyadiques / Structurelles.**
  - Dyadique : Evaluer le *score* d'un lien entre deux nœuds v_i, v_j
  - Structurelle : prédire l'évolution de sous-graphes (prédiction de plusieurs liens en même temps) [7]
- **Type d'attributs : topologiques / caractéristiques des nœuds.**
  - Approche topologique : utiliser seulement le graphe du réseau.
  - L'emploi des approches fondées sur l'analyse du contenu des nœuds nécessite une expertise dans le domaine de l'application.
- **Prise en compte du temps : Oui / non.**

Bibliographie I

Nesserine Benchettara, Rushed Kanawati, and Céline Rouveirol.

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

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

Richard J E Cooke.

Link prediction and link detection in sequences of large social networks using temporal and local metrics.

Master thesis, University of cape Town, 2006.

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

A comparaison of statistical and Machine learning Algorithms on the task of linl completion.

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

Petter Holme, Christofer R Edling, and Fredrik Liljeros.

Structure and Time-Evolution of an Internet Dating Community,.

Social Networks, 26:155–174, 2004.

Zan Huang, Xin Li, and Hsinchun Chen.

Link prediction approach to collaborative filtering.

In Mary Marlino, Tamara Sumner, and Frank M Shipman III, editors, *JCDL*, pages 141–142. ACM, 2005.

Bibliographie II

Rushed Kanawati.

On Using {SNA} techniques for enhancing performances of On-line help-desks.
In *Proceedings of {IADIS} International Conference on E-commerce (E-commerce'08)*, pages 286–291, Amsterdam, 2008.

Mayank Lahiri and Tanya Y Berger-Wolf.

Structure Prediction in Temporal Networks using Frequent Subgraphs.
In *CIDM*, pages 35–42. IEEE, 2007.

Neal Lathia, Stephen Hailes, and Licia Capra.

kNN CF: a temporal social network.
In Pearl Pu, Derek G Bridge, Bamshad Mobasher, and Francesco Ricci, editors, *RecSys*, pages 227–234. ACM, 2008.

David Liben-Nowell.

An Algorithmic Approach to Social networks.
PhD thesis, M.I.T., 2005.

Tsuyoshi Murata and Sakiko Moriyasu.

Link Prediction based on Structural Properties of On-line Social Networks.
New Generation Computing, 26:245–257, 2008.

Bibliographie III

M E J Newman.

Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality.
Phys. Rev. E, 64(1):16132, 2001.

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.

Milen Pavlov and Ryutaro Ichise.

Finding Experts by Link Prediction in Co-authorship Networks.
In Anna V Zhdanova, Lyndon J B Nixon, Malgorzata Mochol, and John G Breslin, editors, *FEWS*, volume 290 of *CEUR Workshop Proceedings*, pages 42–55. CEUR-WS.org, 2007.

Mukund Seshadri, Sridhar Machiraju, Ashwin Sridharan, Jean Bolot, Christos Faloutsos, and Jure Leskovec.

Mobile call graphs: beyond power-law and lognormal distributions.
In Ying Li, Bing Liu, and Sunita Sarawagi, editors, *KDD*, pages 596–604. ACM, 2008.