

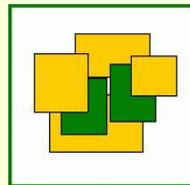
Le Forage des Réseaux Sociaux

Extraction et Gestion des Connaissances (EGC) 2008

Osmar R. Zaïane



**Data Mining Research Group
Database Laboratory**



**Department of Computing Science
University of Alberta, Canada**



1^{er} Février 2008



Edmonton est la 5ème plus grande ville du Canada avec un peu plus d'1 million d'habitants. C'est la capitale de l'Alberta.



- Alberta: 3.5 m d'habitants
- Superficie de la France
- Province riche en pétrole



Production:

Alberta → 1.5 M b/j

3 M b/j en 2010

5 M b/j en 2020

Canada → 2.7 M b/j

Production:

A. Saudite → 8.8 M b/j

Russie → 9.5 M b/j

Iran → 4 M b/j

Norvege → 3 M b/j

Nigeria → 2.3 M b/j

Venezuela → 2.4 M b/j



2ème plus grande réserve mondiale
après l'Arabie Saudite



Plan



Introduction au Réseaux Sociaux



L'analyse de réseaux sociaux (SNA)



Parcours aléatoires pour l'ordonnancement



Communautés dans les réseaux sociaux



Problèmes intéressants pour la recherche





Qu'est ce qu'un réseau social?

- C'est un ensemble d'entités sociales (individus, organisations, ou autres) reliées par des liens créés lors d'interactions sociales (relation ou intérêt commun).
- Ça représente une structure sociale modélisée par des sommets (désignant des individus ou autres) reliés entre eux par des arêtes représentant des relations sociales de divers types (amitié, conflit, partenariat, communication, contact, relation, transmission de maladie, etc.)

● Individual

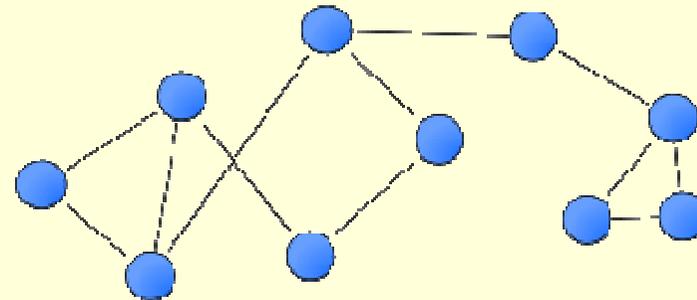


Diagramme d'un réseau social à relation unique.





Analyse des réseaux sociaux

- C'est l'étude des réseaux sociaux pour comprendre et interpréter la structure et le comportement d'un réseau.
- L'étude de réseaux sociaux n'est pas nouvelle. Sciences sociales depuis les années 1970.
- Anthropologie, épidémiologie, criminologie, etc.





Réseaux dans les Sciences Sociales

- Réseaux sociaux [Contractor, 2006]
 - Qui connaît qui
- Réseaux Sociocognitifs
 - Qui croit qui connaît qui
- Réseaux de connaissance
 - Qui connaît quoi
- Réseaux de connaissance cognitive
 - Qui croit qui connaît quoi
- Analyse sociocentrique
 - Surtout en sociologie: quantification d'interaction entre groupes d'individus. Identification de structures globales dans un réseau.
- Analyse egocentrique
 - Surtout en psychologie et anthropologie: quantification d'interaction entre un individu (ego) et d'autres individus (alters).





Vulgarisation

- Six degrés de séparation (Frigyes Karinthy 1929)
- Paradoxe de Milgram: effet du petit monde (Staley Milgram, sociologue 1967)



- PageRank du moteur de recherche  1998 utilise le réseaux des citations des pages Web.
- Outils de réseautage des internautes



le forage des réseaux sociaux

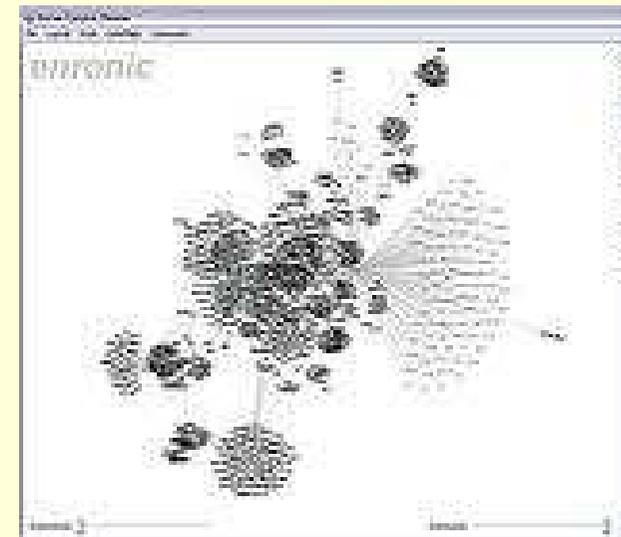




Le fameux cas d'Enron



- Société de courtage pour ressources naturelles
- Faillite en 2001 (scandale financier)
- Données sont publiques
 - <http://www.cs.cmu.edu/~enron>
 - 151 usagers
 - 200 399 messages courriel
 - Classique pour la recherche



Visualisation du réseau d'email d'Enron,
Jeffrey Heer, 2005





Applications

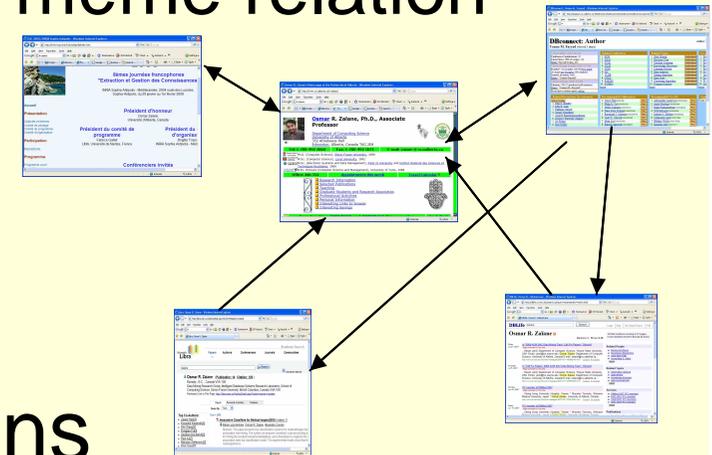
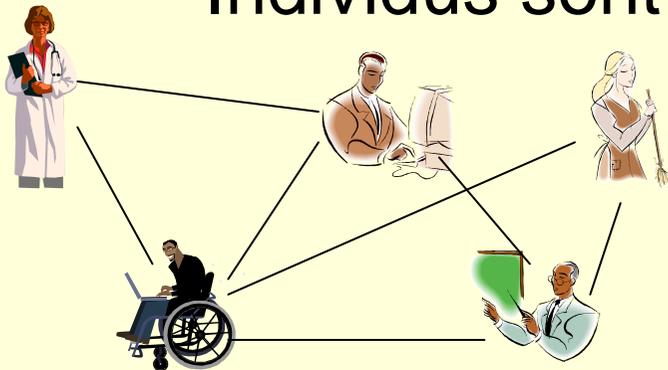
- Terrorisme et crimes
- Médecine – épidémiologie
- Marketing
- Bio-informatique (interaction de protéines...)
- Recommandation de produits
- Ordonnancement par pertinence



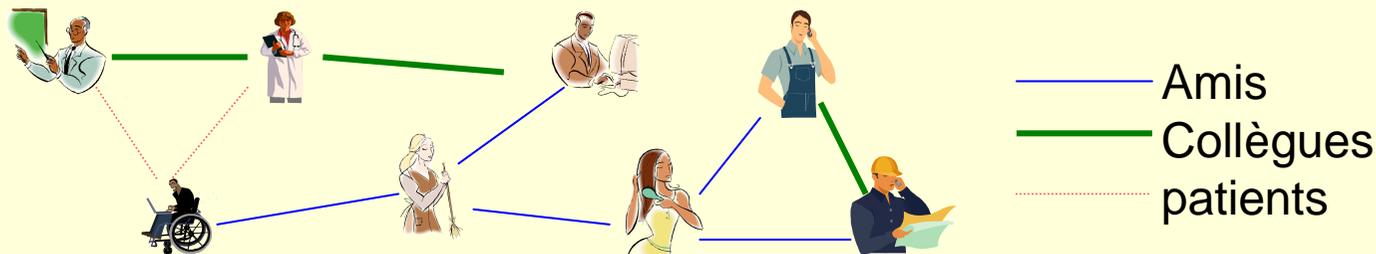


Types de relations et de réseaux (1)

- Réseaux à relation unique
 - Individus sont reliés par une même relation



- Réseaux à multiples relations
 - Individus ont différents types de relations



le forage des réseaux sociaux

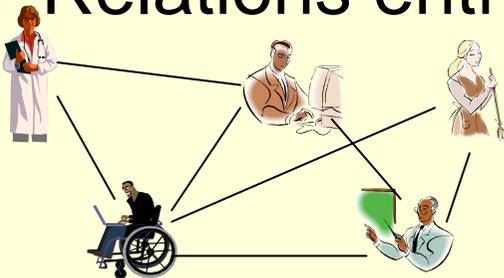


Types de relations et de réseaux (2)



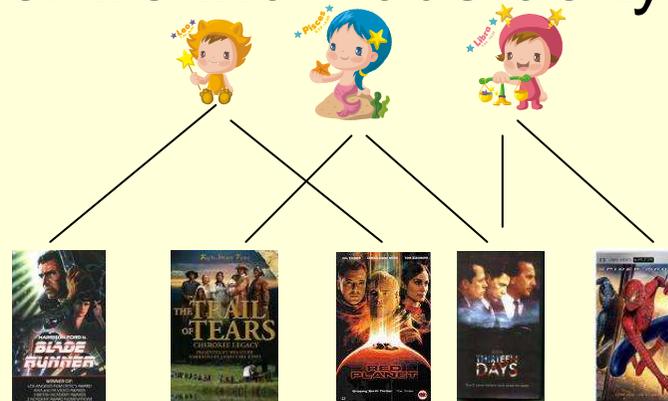
- Relation homogène

- Relations entre individus du même type



- Relation hétérogène

- Relation entre individus de types différents



le forage des réseaux sociaux





D'autres concepts importants

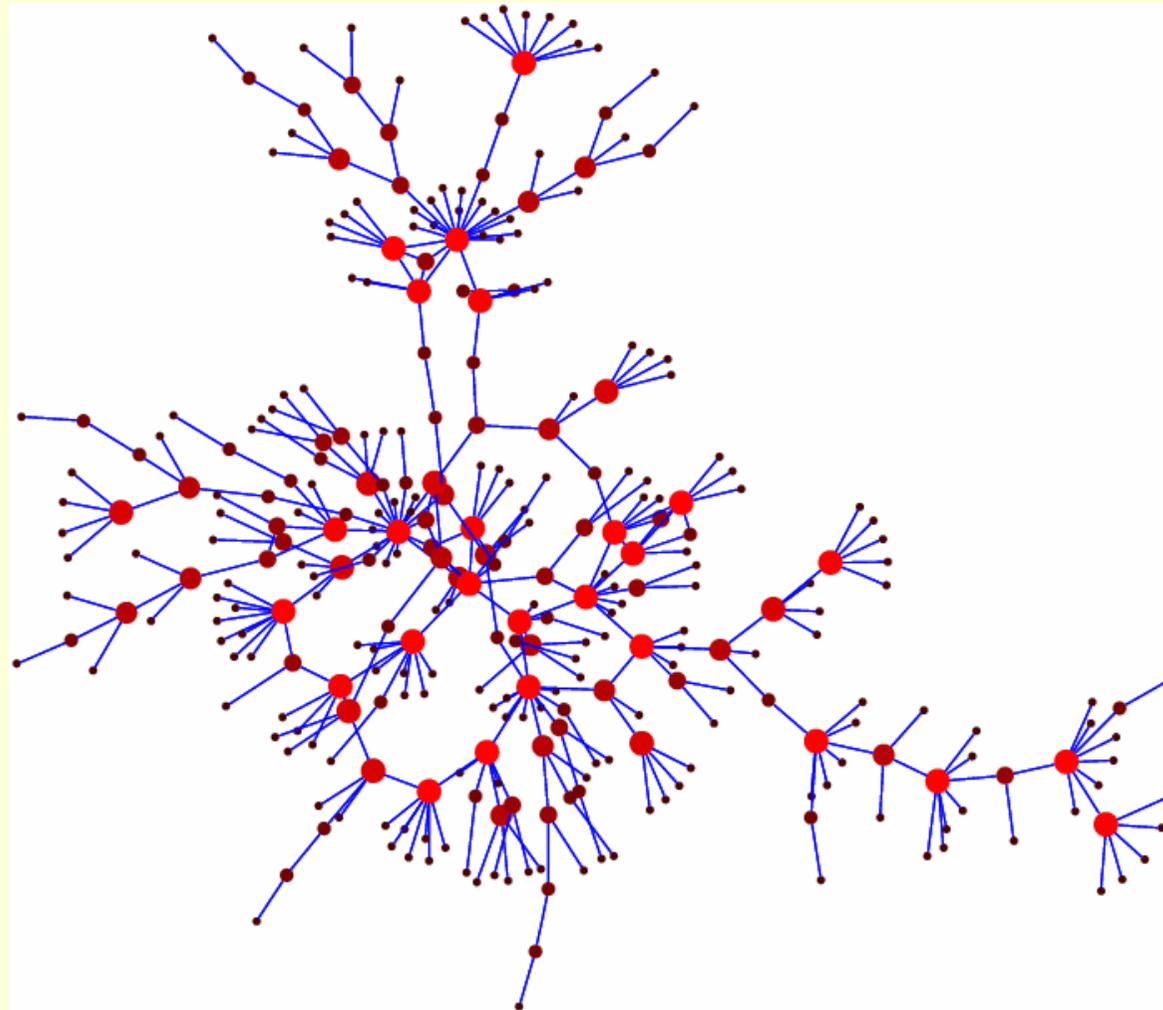
- Densité et connectivité dépendent de la notion de la relation
- Force des liens: fréquence des interactions, importance des échanges d'information, intimité, intensité émotionnelle, etc.
- Relation symétrique ou non (directionnelle)
- Centralité





Réseau de relations sexuelles

- SIDA
- HIV
- normal



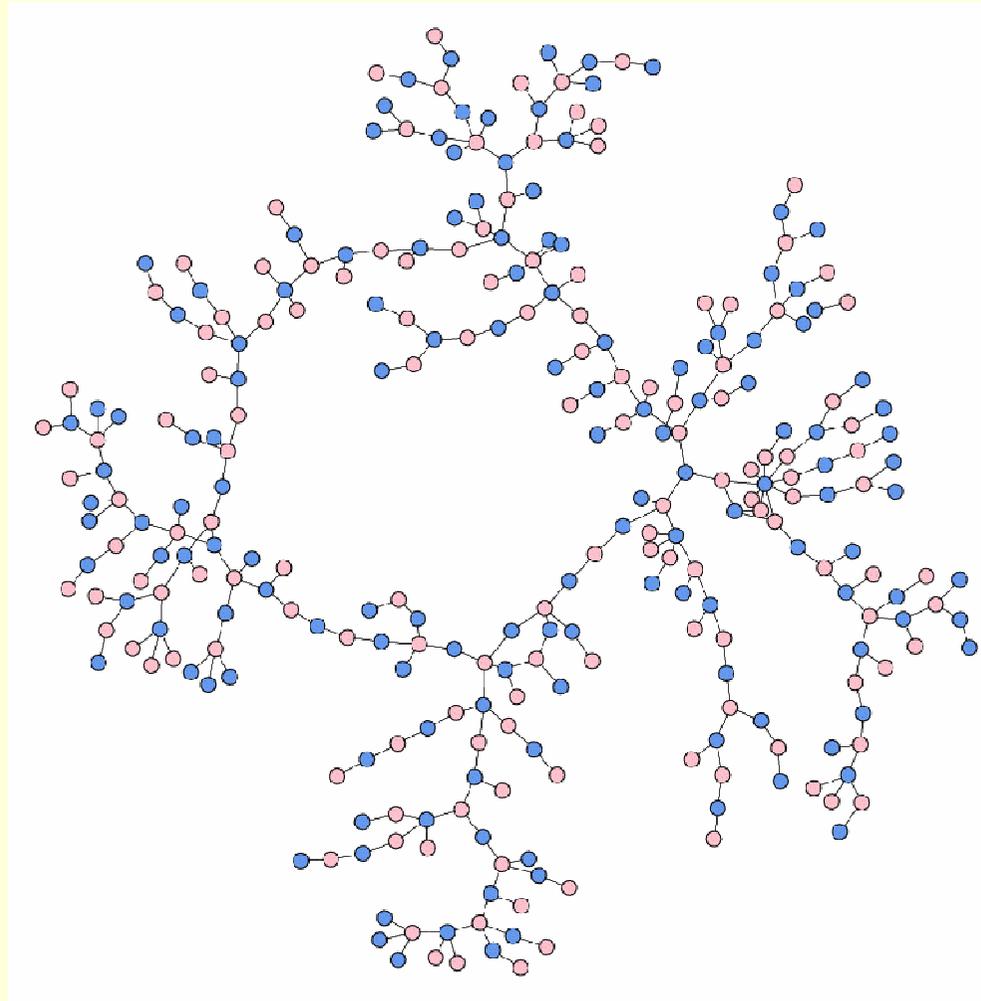
[Newman 2003]





Réseau de rendez-vous

- garçon
- fille

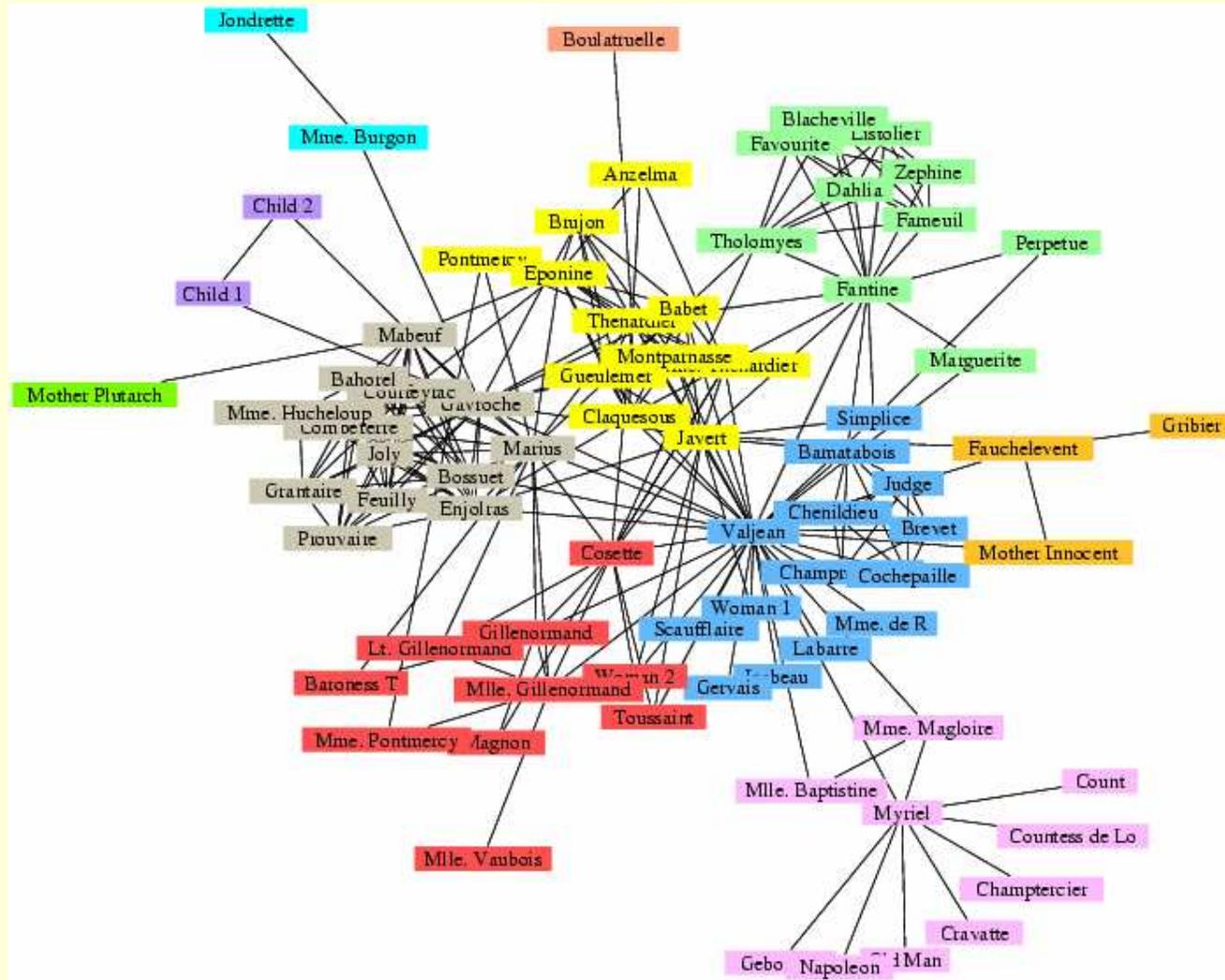


Jefferson High School
Jefferson City, Missouri , USA
[Bearman et al. 2004]





Réseau "Les Misérables" [Newman et al. 2004]



Exploration des réseaux sociaux





Contenu



Introduction au Réseaux Sociaux



L'analyse de réseaux sociaux (SNA)



Parcours aléatoires pour l'ordonnancement



Communautés dans les réseaux sociaux



Problèmes intéressants pour la recherche



Ordonnancement d'entités



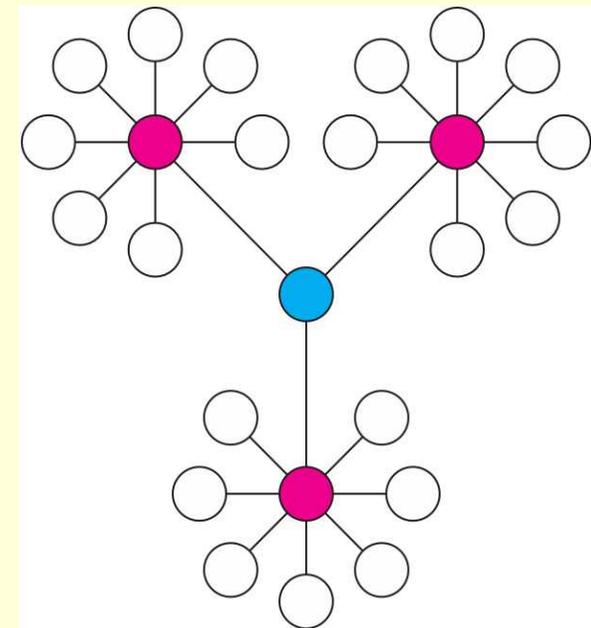
- Comprendre les relations entre entités.
- Trier les ensembles d'objets dans le réseau basé sur les relations entre ces objets et la structure générale des liens.
- Étudié surtout en sociologie (nœud influent) et en informatique



Ordonnancement Sociologique



- Centralité: mesure l'importance des individus dans un réseau sociale
- Degré de centralité
 - + simple
 - mesure seulement la structure locale
- Centralité des vecteurs propres



Centralité des vecteurs propres



- Intuition: Les connections d'un nœud à grande valeur doivent contribuer davantage au nœud en question que les connexions venants de nœuds à faible valeur.
- Pour le nœud i , la centralité se définit comme étant proportionnelle à la somme des valeurs des nœuds qui connectent à i .

$$x_i = \frac{1}{\lambda} \sum_{j=1}^N A_{i,j} x_j \longrightarrow X = \frac{1}{\lambda} AX \longrightarrow AX = \lambda X$$

- La plus grande valeur propre donne la mesure de centralité désirée. (Théorème de Perron–Frobenius)



Approches en informatique



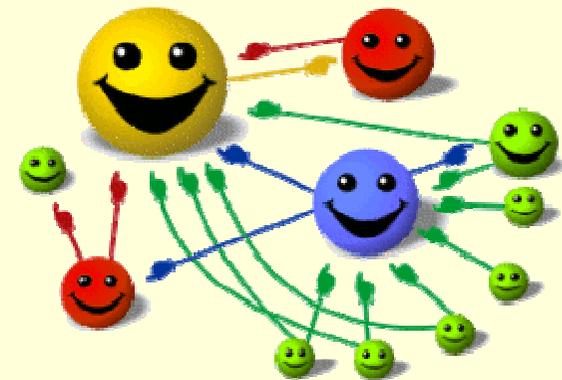
- Trier les documents Web pour un moteur de recherche
- Considérer la structure des hyperliens ainsi que le contenu des pages
- PageRank et HITS





Intuition de PageRank

- Appliquer la centralité des vecteurs propres dans le domaine du Web: un document est placé haut si d'autres documents haut placés sont connectés à celui-ci.
- PageRank peut être interprété comme une probabilité de distribution pour représenter la probabilité qu'un internaute en cliquant aléatoirement les liens d'une page à l'autre revienne au point de départ.



Désavantages de PageRank



- Favorise les anciennes pages
- Souffre du “link spamming”
- Google Bombs: influencer l’ordre d’une page donnée dans les résultats de Google pour des raisons humoristiques, satiriques.
 -  “more evil than Satan himself” → Microsoft
 -  “miserable failure” → George W. Bush





D'autres approches

- **Jeh et Widom : SimRank.** Deux pages sont similaires si elles sont liées ou apparentées à des pages similaires
- **Parcours aléatoires pour ordonner des tuples dans les bases de données** (ObjectRank, RelationalRank, BANKS...)
- **PageRank sur un ensemble de synonymes de WordNet pour trouver des propriétés sémantiques.**
- **Réseaux de courriel, réseaux téléphoniques, attribuer des mots clés aux images...**





DBconnect: Topic

Data Mining (viewed 56 times)

| Related Researchers (by relevance) | Related Conferences | Related Topics |
|--|--------------------------------------|--|
| <u>Sort by H-index</u> | 1. KDD | 1. Database System |
| 1. Jiawei Han | 2. SIGMOD Conference | 2. Association Rule |
| 2. Heikki Mannila | 3. VLDB | 3. Information System |
| 3. Rakesh Agrawal | 4. ICDE | 4. Web Service |
| 4. Christos Faloutsos | 5. ICDM | 5. Decision Tree |
| 5. Raymond T. Ng | 6. PKDD | 6. Management System |
| 6. Marek Wojciechowski | 7. PAKDD | 7. Relational Database |
| 7. Osmar R. Zaiane | 8. CIKM | 8. Information Retrieval |
| 8. Hongjun Lu | 9. IJCAI | 9. Knowledge Discovery |
| 9. Maciej Zakrzewicz | 10. SAC | 10. Query Language |
| 10. Philip S. Yu | 11. DEXA | 11. Data Modeling |
| 11. Masaru Kitsuregawa | 12. ICML | 12. Data Stream |
| 12. Carlo Zaniolo | 13. AAAI | 13. Machine Learning |
| 13. Haixun Wang | 14. DaWaK | 14. Data Structure |
| 14. Krzysztof Koperski | 15. PODS | 15. Data Warehousing |
| 15. Hans-Peter Kriegel | 16. ISMIS | 16. Neural Network |
| 16. Usama M. Fayyad | 17. SDM | 17. Query Optimization |
| 17. Huan Liu | 18. INFOCOM | 18. Time Series |
| 18. Mohammed Javeed Zaki | 19. ECML | 19. Semantic Web |
| 19. Richard R. Muntz | 20. SIGIR | 20. Lower Bound |
| 20. Mika Klemettinen | more | more |
| more | | |

DBconnect Beta 0.1 - Based on the [DBLP database](#) (database version of June 2007)
Developed by the [Database Group](#) at the University of Alberta and the [Alberta Ingenuity Centre for Machine Learning](#)



DBconnect: Conference

KDD (12 events)

| Related Researchers | Related Topics | Related Conferences |
|--|--|---------------------------------------|
| 1. Jiawei Han | 1. Data Mining | 1. ICML |
| 2. Philip S. Yu | 2. Association Rule | 2. ICDM |
| 3. Bing Liu | 3. Decision Tree | 3. PAKDD |
| 4. Padhraic Smyth | 4. Knowledge Discovery | 4. PKDD |
| 5. Heikki Mannila | 5. Data Stream | 5. SIGIR |
| 6. Christos Faloutsos | 6. Time Series | 6. VLDB |
| 7. Charu C. Aggarwal | 7. Support Vector Machine | 7. ICDE |
| 8. Wynne Hsu | 8. Database System | 8. SIGMOD Conference |
| 9. Jian Pei | 9. Machine Learning | 9. DaWaK |
| 10. Mohammed Javeed Zaki | 10. Web Service | 10. CIKM |
| 11. Eamonn J. Keogh | 11. Bayesian Network | 11. SDM |
| 12. Rakesh Agrawal | 12. Information Retrieval | 12. UAI |
| 13. Pedro Domingos | 13. Relational Database | 13. IDEAL |
| 14. Hans-Peter Kriegel | 14. Information System | 14. DASFAA |
| 15. Ke Wang | 15. Semantic Web | 15. ECML |
| 16. Martin Ester | 16. Text Classification | 16. DEXA |
| 17. Wei Wang | 17. Feature Selection | 17. Web Intelligence |
| 18. Vipin Kumar | 18. Digit Library | 18. WAIM |
| 19. Andrew W. Moore | 19. Reinforcement Learning | 19. APWeb |
| 20. Salvatore J. Stolfo | 20. Neural Network | 20. Discovery Science |
| all fewer | all fewer | all fewer |





DBconnect: Author

Author: search

Osmar R. Zaiane (viewed 529 times)

From DBLP (2007-06-20)
 Conference Contributions: 60
 Career Since: 1995 (Average: 4.62)
 Query: Za=iuml=ane:Osmar_R=

From Google Scholar (2007-09-27)
 H-index*: 26 (A-index: 62.3077) [See graph](#)
 Average top 10 papers: 110 citations
 Number of entries: 1863
 Query: zaiane

From Citeseer (2008-01-16)
 Citations: 264 (62 predicted self-citations)
 Query: "zaiane"
 If you have a better query, [tell us](#).

| Related Conferences | |
|--------------------------------------|--|
| 1. ICDM | |
| 2. KDD | |
| 3. PAKDD | |
| 4. SIGMOD Conference | |
| 5. ICDE | |
| 6. IDEAS | |
| 7. DEXA Workshops | |
| 8. PKDD | |
| 9. VLDB | |
| 10. DEXA Workshop | |

[more](#)

| Related Topics | |
|--|------------------------------|
| 1. Data Mining | Publications |
| 2. Association Rule | Publications |
| 3. Information System | Publications |
| 4. Relational Database | Publications |
| 5. Frequent Itemset | Publications |
| 6. Knowledge Discovery | Publications |
| 7. Data Warehousing | Publications |
| 8. Query Language | Publications |
| 9. Information Retrieval | Publications |
| 10. Digital Library | Publications |

[more](#)

| Co-Authors (21) | |
|---|----|
| 1. Mohammad El-Hajj | 10 |
| 2. Jiawei Han | 9 |
| 3. Stanley R. M. Oliveira | 5 |
| 4. Randy Goebel | 4 |
| 5. Chi-Hoon Lee | 4 |
| 6. Jenny Chiang | 4 |
| 7. Andrew Foss | 3 |
| 8. Krzysztof Koperski | 3 |
| 9. Hua Zhu | 2 |
| 10. Yongjian Fu | 2 |

[more](#)

| Related Researchers (by relevance) | |
|---|--|
| Sort by H-index | |
| 1. Mohammad El-Hajj | |
| 2. Jiawei Han | |
| 3. Stanley R. M. Oliveira | |
| 4. Chi-Hoon Lee | |
| 5. Randy Goebel | |
| 6. Andrew Foss | |
| 7. Jenny Chiang | |
| 8. Krzysztof Koperski | |
| 9. Wei Wang | |
| 10. Robert C. Holte | |

[more](#)

| Recommended Collaborators | | |
|--|--|----------------------|
| 1. Philip S. Yu (0.000465478) | | Why? |
| 2. Hans-Peter Kriegel (0.000405147) | | Why? |
| 3. Rakesh Agrawal (0.000345893) | | Why? |
| 4. Heikki Mannila (0.000323786) | | Why? |
| 5. Masaru Kitsuregawa (0.000318237) | | Why? |
| 6. Divyakant Agrawal (0.00030855) | | Why? |
| 7. Amr El Abbadi (0.000294673) | | Why? |
| 8. Jian Pei (0.00028927) | | Why? |
| 9. Hongjun Lu (0.000288656) | | Why? |
| 10. Hector Garcia-Molina (0.000285476) | | Why? |

[more](#)

| Recommended To (269) | | |
|--|--|----------------------|
| 1. Krishnamoorthy Sivakumar (0.000408406) | | Why? |
| 2. Amol Ghoting (0.000363388) | | Why? |
| 3. Aleksandar Lazarevic (0.000352352) | | Why? |
| 4. Balaji Padmanabhan (0.0003472) | | Why? |
| 5. Hillol Kargupta (0.000344372) | | Why? |
| 6. Peter Christen (0.000337619) | | Why? |
| 7. Haesun Park (0.000329418) | | Why? |
| 8. Man Leung Wong (0.000325567) | | Why? |
| 9. Jérémy Besson (0.000322607) | | Why? |
| 10. Jia-Yu Pan (0.000321161) | | Why? |

[more](#)

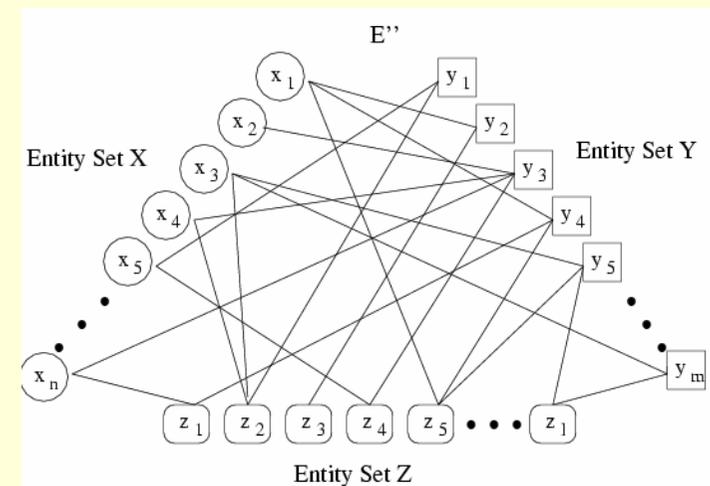
| Symmetric Recommendations | | |
|--|--|----------------------|
| 1. Jian Pei (0.00028927) | | Why? |
| 2. David Wai-Lok Cheung (0.000272961) | | Why? |
| 3. Ke Wang (0.000272037) | | Why? |
| 4. Eamonn J. Keogh (0.000246664) | | Why? |
| 5. Srinivasan Parthasarathy (0.00024336) | | Why? |
| 6. Huan Liu (0.00024207) | | Why? |





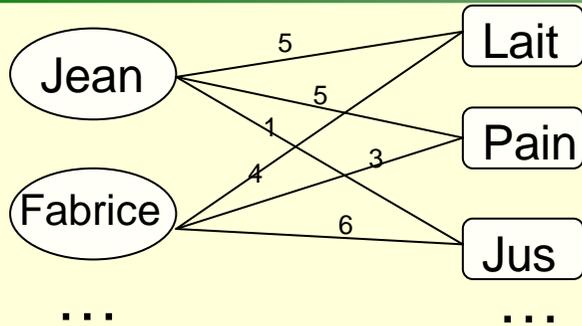
La solution: TupleRank

- Le réseau hétérogène est modélisé par un graphe naire.
- Le graphe est augmenté par des nœuds de remplacement pour modéliser des relations particulières.
- Un parcours aléatoire avec redémarrage est défini sur le graphe.
- La direction du parcours d'une partie du graphe à l'autre est contrôlée.

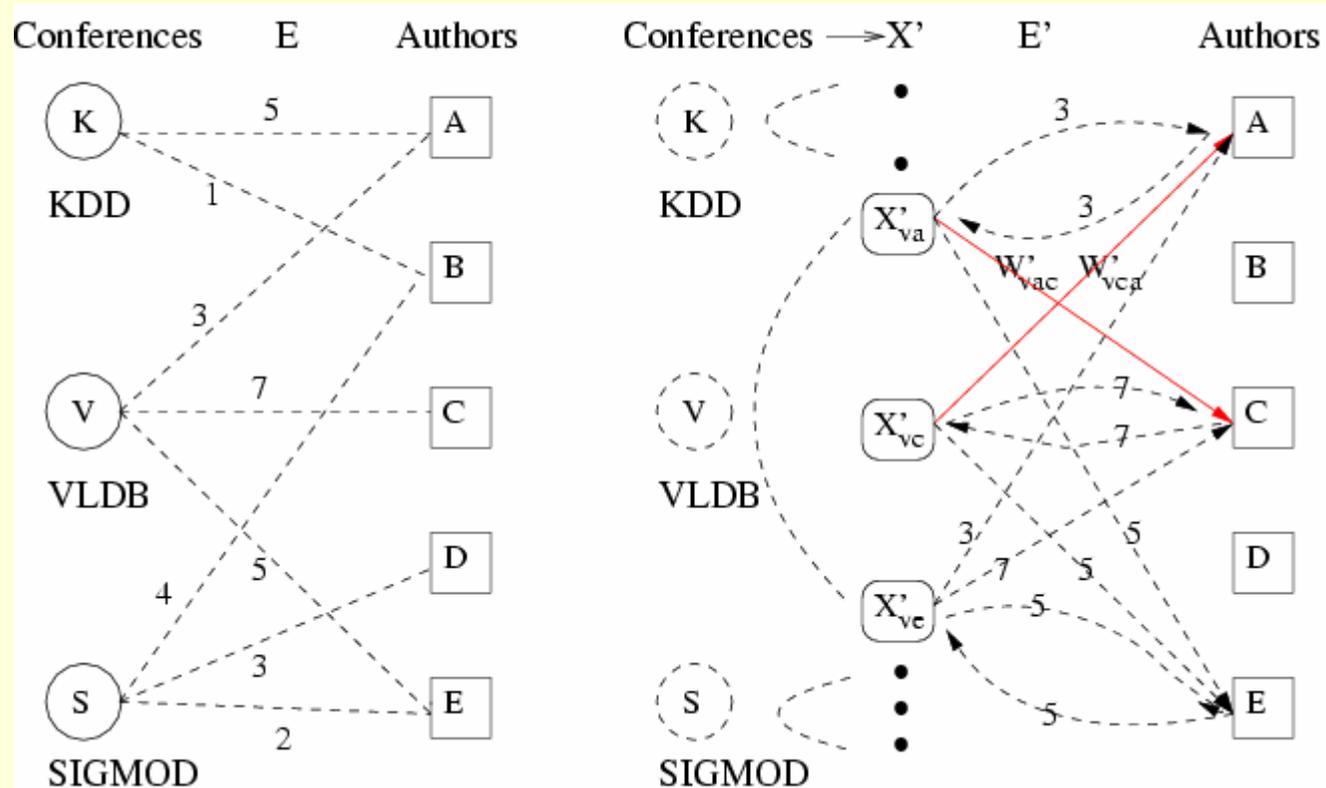




Extension du graphe



Les transactions de Jean ou Fabrice sont en fait perdues. La relation entre produits dans une même transaction n'est plus.





Contenu



Introduction au Réseaux Sociaux



L'analyse de réseaux sociaux (SNA)



Parcours aléatoires pour l'ordonnancement



Communautés dans les réseaux sociaux



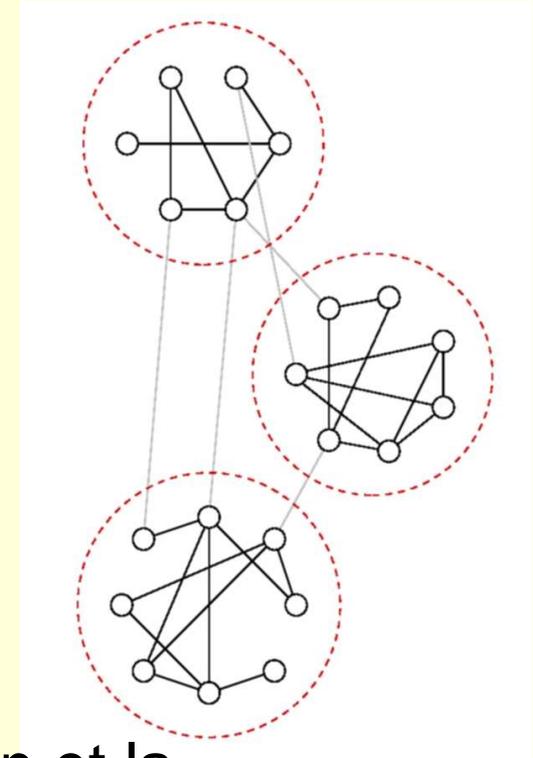
Problèmes intéressants pour la recherche

Qu'est ce qu'une communauté?



Théorie des graphes: groupes de sommets connectés de façons dense, avec quelques rares connexions entre groupes.

Analyse des réseaux sociaux: groupes d'entités qui partagent des propriétés similaires ou connectent entre elles via certaines relations.



Découverte de communautés c'est la détection et la caractérisation de communautés dans la structure d'un réseau. Identification de connexions et entités appartenant à différentes communautés. C'est la classification (clustering) dans le contexte des réseaux sociaux.

Découverte de communautés



- Deux écoles de pensée
 - Informatique → Partitions des graphes: division du réseau en k groupes à taille (à peu près) égale, tout en minimisant le nombre de liens entre groupes
 - Sociologie → Détection des communautés: détection de la structure des communautés
- Différence
 - Partition des graphes doit connaître *a priori* le nombre et la taille désirée des groupes
 - Détection de communautés suppose les sous-groupes sont déterminés par le réseau lui même et non l'utilisateur.





Partition de graphe

- Méthode “Spectral Clustering”: basée sur les vecteurs propres orthogonales du graphe Laplacien.
- Intuition: bisection itérative: trouver la meilleure division du réseau en deux parties, et continuer itérativement jusqu’à ce que le nombre de groupe désiré est atteint.



Avantage: Fondation théorique solide



Désavantage: paramètres tels que le nombre de partitions et la taille sont nécessaires





Détection de communautés

- **Classification hiérarchique:** découvrir les divisions naturelles des réseaux sociaux basées sur des mesures de similarité variées.
- **Méthodes agglomeratives:** les arêtes avec grande similarité sont ajoutées à un réseau initialement vide
- **Méthodes divisives:** les arêtes à faible similarité sont enlevées du réseau



Avantage: Rapide, pas de paramètres requis



Désavantage: précision limitée par les mesures de similarité; découvre le centre intérieur de la communauté correctement mais souvent omet les nœuds périphériques.



Évaluation de communautés



- Comment savoir si les communautés découvertes sont correctes?
- Comment comparer différents résultats ou les résultats de méthodes différentes?
- Il ya rarement des données vraies de référence «ground truth» pour comparaison.
- Une évaluation visuelle n'est pas convaincante pour démontrer une généralisation d'une méthode (difficile quand les résultats sont très semblables)
- Une méthode d'évaluation systématique est nécessaire.



Évaluation par coupure de graphe



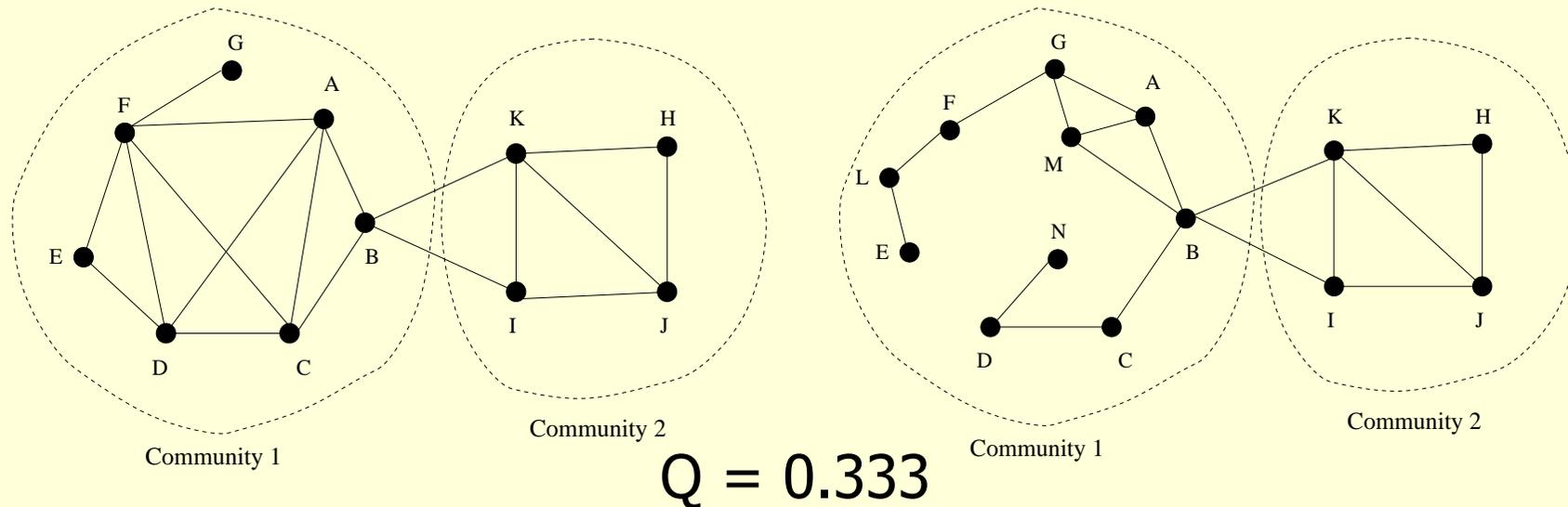
- Critères de coupure: R-cut, N-cut, MinMax-cut, ...
- Une bonne division en communautés n'est pas nécessairement celle où il y a un nombre plus petit d'arêtes entre groupes
- Une bonne division est celle où le nombre d'arêtes entre groupes est plus petit que ce qui est attendu (en supposant une distribution aléatoire uniforme).
- *Évaluation par Modularité Q*
 $Q = (\text{nombre d'arêtes entre communautés}) - (\text{nombre d'arêtes attendu})$





Problèmes

- Fonctionne correctement en général
- Ne considère pas l'absence de liens entre nœud d'une même communauté

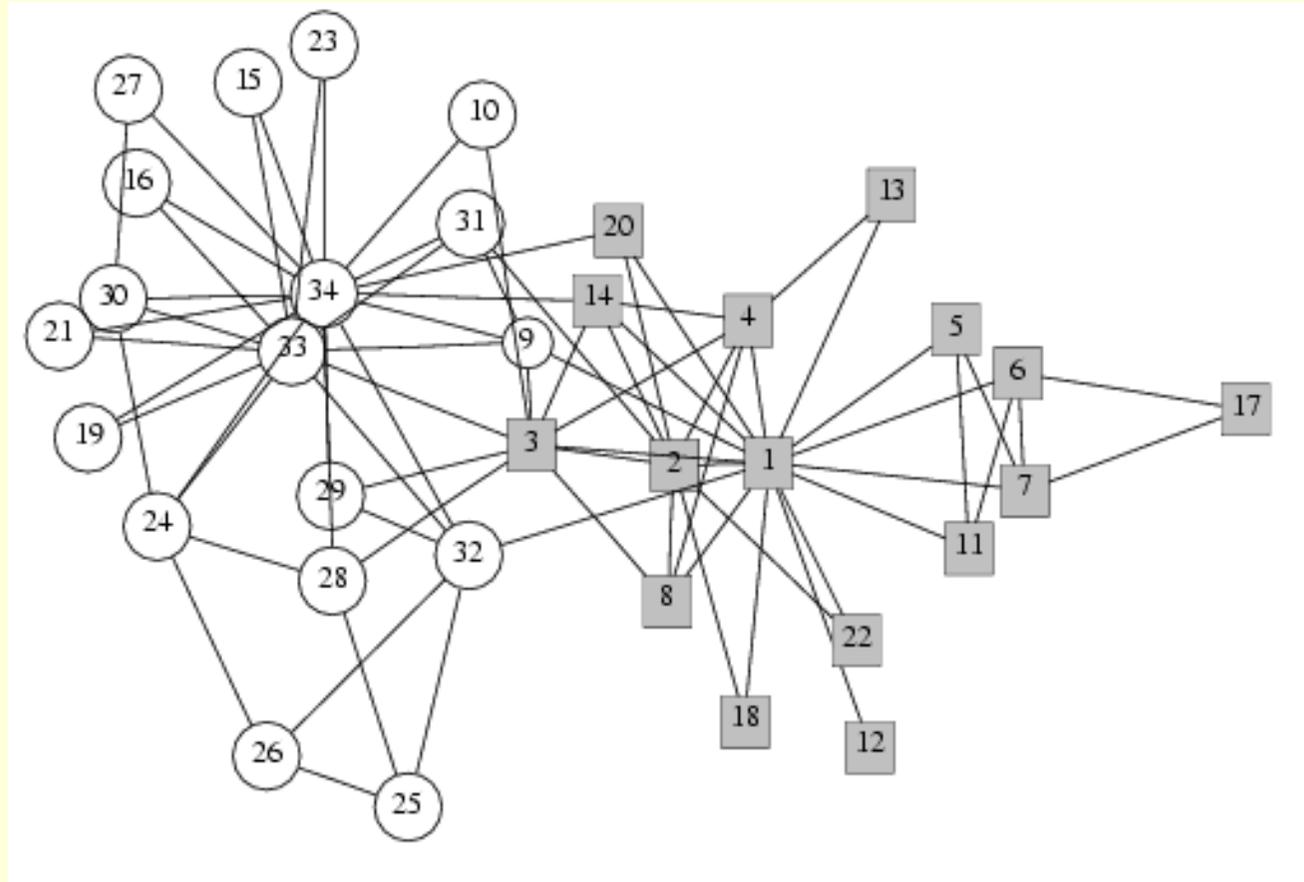


- On pourrait avoir une mesure qui pénalise l'absence de liens au sein d'une communauté





Exemple: forage du réseau “Club Karaté”



Karate Club in an American University [Zachary 1977]





Contenu



Introduction au Réseaux Sociaux



L'analyse de réseaux sociaux (SNA)



Parcours aléatoires pour l'ordonnancement



Communautés dans les réseaux sociaux



Problèmes intéressants pour la recherche



Réseau à relation multiples



- Tous les travaux actuels supposent un réseau social à relation unique.
- La majorité des réseaux sociaux réels sont en fait des réseaux à relations multiples.
- Comment ordonner ou sélectionner les relations les plus importantes? (avec des exemples libellés pour apprentissage)





Premier Scenario

- Une maladie infectieuse se propage d'une façon épidémique. Nous avons un réseau de communautés contenant divers relations: collègue, amis, voisin, ...
- Trouver qui a la plus grande probabilité d'être infecté, en se basant sur une liste de patients.
- Trouver les relations à enquêter en premier lieu pour prédire une propagation (prédire la dynamique du réseaux).



Réseau à communauté “soft”



- Les méthodes courantes supposent qu'une entité appartienne à une seule communauté.
- La plupart des entités dans des applications réelles appartiennent à des communautés multiples.
- Comment identifier et classer les entités dans des communautés avec intersections?



Dynamisme des communautés



- Dans un réseau social, les entités, les relations ainsi que les communautés sont dynamiques
 - Les communautés grandissent ou rétrécissent
 - Les communautés se fusionnent
 - Les communautés se dissolvent
 - De nouvelles communautés se forment.
- Comment étudier et prédire la dynamique d'un réseau social?





Un autre scenario

- Jaguar:



ou



- Michael Jordan:



ou



- Quand on cherche sur un moteur de recherche, quelle communauté est pertinente pour notre requête?



Promiscuité dans les réseaux



- Résolution d'entité dans les données bibliographiques.
- Noms communs:
 - *Wei Wang*: Il y en a 4 ou 5 distincts en informatique (North Carolina, NSW, Chine...).
 - *Jörg Sander*: un en biologie, un en chimie et un en informatique
 - Ils publient dans (et avec) des communautés différentes.
- Y a t il une solution pour lever l'ambiguïté?





Evaluation

- Différents moteurs de recherche présentent leur propre triage des résultats.
- Comment comparer deux tris différents?
- Plusieurs méthodes de découverte de communautés donnent des résultats différents.
- Comment comparer les communautés découvertes?





Merci pour votre attention



UNIVERSITY OF
ALBERTA

Osmar R. Zaïane, Ph.D.
Associate Professor
Department of Computing Science

352 Athabasca Hall
Edmonton, Alberta
Canada T6G 2E8

Telephone: Office +1 (780) 492 2860
Fax +1 (780) 492 1071
E-mail: zaiane@cs.ualberta.ca
<http://www.cs.ualberta.ca/~zaiane/>



Questions

?

