



Prédiction de liens dans les réseaux sociaux

Approches topologiques

Rushed Kanawati, Céline Rouveirol

A3 - LIPN UMR CNRS 7030
université Paris 13

`prénom.nom@lipn.univ-paris13.fr`

Ecole d'été Web Intelligence, 10 juillet 2009, Véraane



Plan

Introduction

Problématique de Prédiction de liens

- Définition

- Applications

- Problèmes similaires

Approches de prédiction de liens

- Classification

- Approches dyadiques / topologiques

 - Apprentissage automatique

- Approches basées sur le contenus des nœuds

- Approches temporelles

- Approches structurelles

Conclusion

Bibliographie

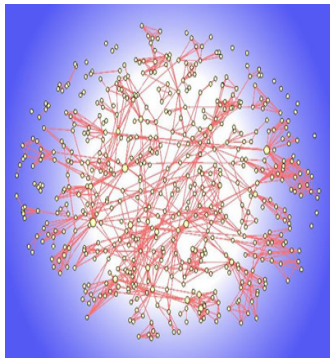
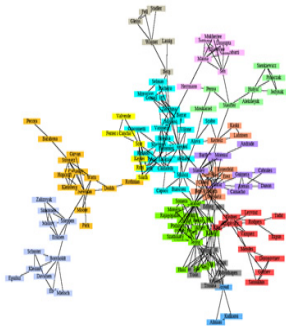


Réseaux sociaux : Notations

- Un réseau social est un graphe $G = \langle V, E \subseteq V \times V \rangle$:
 - V est l'ensemble de nœuds (i.e. acteurs sociaux)
 - E est l'ensemble de liens *sociaux*.
- 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.
 - $\Gamma(v)$ est l'ensemble de voisins de v .
 $\Gamma(v) = \{x \in V : (x, v) \in E\}$.
 - Le degré d'un nœud $d(v) = \|\Gamma(v)\|$



Réseaux cibles



source : www.visulacomplexity.com



Réseaux cibles : Caractéristiques

- **Grand Taille** : $|V| > 10^3$.
- **Sans échelle** : La probabilité pour un nœud $v \in V$ d'avoir k voisins est : $P(k) = k^{-\gamma}$

Beaucoup de nœuds avec peu de connexions et peu de hubs.

- **Coefficient de clustering élevé** :

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

La probabilité que deux voisins d'un nœud choisi aléatoirement soient eux mêmes connectés.

- Réseaux à **relation unique** : tous les liens $e \in E$ sont de même type.
- Réseaux simples : La **matrice** A_G **est binaire et symétrique** ; $\forall i, j \ a_{ij} \in \{0, 1\}$



Réseaux cibles : Exemples

- Réseaux bibliographiques : publications, citation [New01, New04]
- Réseaux sociologiques : Interaction sur forums/blogs [Kan08, MM08], communication [SMS⁺08], site de rencontres [HEL04], terroriste [HZLG08a].
- Réseaux éthologiques : Comportement des Zèbres [LBW07].
- Réseaux biologiques : Interactions entre protéines [TWAK03], Métabolique [NE08], Interactions entre des gènes [OKK03].
- Réseaux technologiques : Internat, Web
- ...

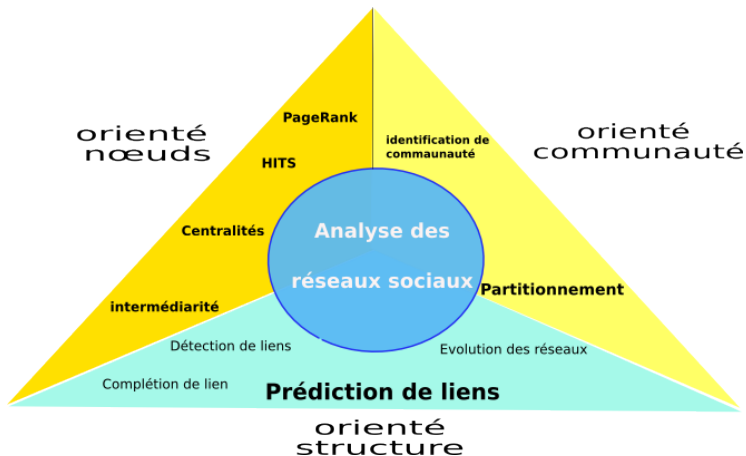


Réseaux cibles : Modélisation

- Beaucoup de réseaux réels sont des graphes **bipartis** (ex. bibliographique, achats, etc.)
- Graphe biparti : $G = (\mathcal{T}, \perp, E \subseteq \mathcal{T} \times \perp)$. \mathcal{T}, \perp sont deux ensembles distincts.
- Projections :
 - $G_{\mathcal{T}}^n = (V_{\mathcal{T}} \subseteq \mathcal{T}, E_{\mathcal{T}} = \{x, y \in \mathcal{T} : |\Gamma_G(x) \cap \Gamma_G(y)| \geq n\})$
 - $G_{\perp}^m = (V_{\perp} \subseteq \perp, E_{\perp} = \{x, y \in \perp : |\Gamma_G(x) \cap \Gamma_G(y)| \geq m\})$
 - n, m : paramètres des projections.
 - La projection augmente, artificiellement, le coefficient de clustering du graphe projeté !!
- **Dans [IGL04] on montre que les réseaux sociaux cibles sont de nature bipartis.** Un algorithme est donné pour construire le graphe biparti à partir d'un graphe unimodal.

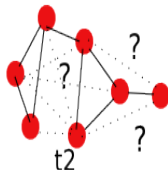
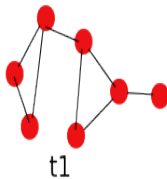
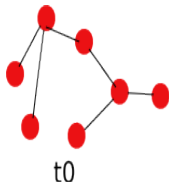


Analyse de réseaux sociaux : les tâches





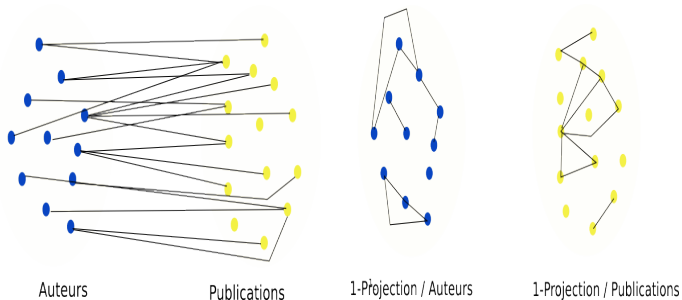
Prédiction de liens : Définition



- Définition informelle : Prédire la formation d'un lien entre deux nœuds jamais connectés auparavant.
- Soit $G = \langle G_1, G_2, \dots, G_T \rangle$ 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 $v, u \in \bigcap_{i=1}^T V_i : \exists E_j : (u, v) \in E_j$ si $(u, v) \in E_{T+1}$.



Application : Recommandation de collaborations académiques



- **Recommandation de collaboration = Prédiction de liens dans le graphe projeté / auteurs [PI07, LN05, KcR09].**



Autres applications

- Aide à la navigation sur le Web [Zhu03]
- Aide à la réponse des questions sur forums [MM08], *help-desk* [Kan08]
- Etude de propagation des virus informatiques par e-mail [LNH03]
- ...



Problème similaire : Liens cachés II

- [Coo06] : 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 le double que pour LF.
 - Pour LC le produit des degrés des nœuds impliqués est la moitié que pour LF.
 - LC : distribution des degrés des nœuds plus étalée.
 - Pas de différence significative en ce qui concerne les distances !



Problème similaire : Complétion de liens

- Dans un graphe $G = \langle V, E \rangle$, 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 [GKK⁺03].
- 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. Trouver l'inconnue en fonction des liens connus.
- Dans [GKK⁺03] on se limite à examiner le cas d'un seul nœud inconnu.



Problème similaire : Liens alarmants

- Etant donné un réseau temporel $G = \langle G_1, G_2, \dots, G_T \rangle$, 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 [RJ05]
- Problème plus facile que la prédiction de liens.



Problème similaire : Evolution des réseaux

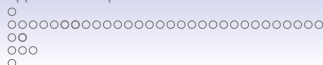
[HZLG08b] : définition de trois niveaux de granularité pour observer et/ou prédire l'évolution.

- Niveau réseau
 - Leskovec et al. [LKF05], Evolution des paramètres des réseaux : diamètre décroissant et densification des degrés dans des réseaux de co-citation.
 - Barabasi et al. [BJN⁺02a], 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



Evolution d'un réseau de co-publications scientifiques

- Niveau “individus” :
 - Définition d'un modèle de poisson stochastique pour modéliser le nombre de collaborations dans le temps, puis d'un arbre d'optimisation pour optimiser le modèle.
 - Pour une collaboration existante au temps t $e_{i,j}(t)$, on extrait un sous-graphe $G(e_{i,j}(t))$ de voisinage de $e_{i,j}(t)$: les auteurs v_i, v_j , leurs voisins immédiats (co-auteurs) et les arcs associés.
 - Un vecteur d'attributs $a_{i,j} = (a_1, \dots, a_p)$ est calculé à partir de ce voisinage
 - Pour un arc spécifique $e_{i,j}(t)$, le taux de collaboration $\lambda(e_{i,j}(t)) = f(a_{i,j}(t))$, apprentissage de $f(a_{i,j}(t))$ par arbre d'optimisation, technique inspirée des arbres de régression [BFOS84]



Expérimentation II

- Tâche d'apprentissage : étant donné une paire d'auteurs v_i et v_j avec une collaboration existante $e_{i,j}$, quelle est la probabilité qu'ils collaborent k fois dans le prochain intervalle de temps Δ .

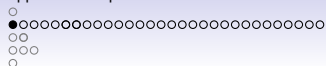
Method	Correlation Coefficient (r)	Root Mean Squared Error (RMSE)
Past Collaboration	0.227	4.74
SVR	0.673	1.53
SPOT	0.782	1.42



Approches de prédiction de liens

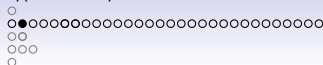
Trois critères de classification :

- **Approche : Dyadiques / Structurelles.**
 - Dyadique : Evaluer un *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) [LBW07]
- **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.**



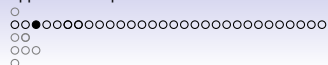
Approches dyadiques topologiques

- score d'un lien (u, v) est calculé par une fonction de *similarité* topologique.
- Deux familles de mesures de similarités topologiques :
 - Mesures basées sur le voisinage des nœuds.
 - Mesures basées sur les distances entre les nœuds.



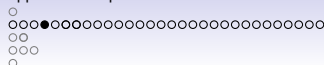
Mesures basées sur le voisinage

- Voisins communs.
- Jaccard.
- Attachement préférentiel
- Adamic Adar
- ...



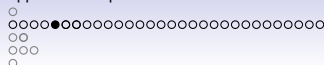
Voisins communs : VC

- Dans un graphe unimodal : $VC(x, y) = \|\Gamma(x) \cap \Gamma(y)\|$
- Dans un graphe biparti G :
 - $VC_{\top}(x, y) = \|\Gamma_{G_{\top}}(x) \cap \Gamma_G(y)\|$ où $x \in \top, y \in \perp$
 - Exemple : Dans un graphe d'achats où \top (resp. \perp) est l'ensemble de clients (resp. produits) : nombre de clients similaires qui ont acheté le même produit y .
 - $VC_{\perp}(x, y) = \|\Gamma_{G_{\perp}}(y) \cap \Gamma_G(x)\|$ où $x \in \top, y \in \perp$
 - Le nombre de produits similaires à y achetés par le client x



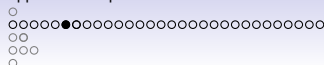
Jaccard : JAC

- Dans un graphe unimodal : $JAC(x, y) = \frac{\|\Gamma(x) \cap \Gamma(y)\|}{\|\Gamma(x) \cup \Gamma(y)\|}$
- Dans un graphe biparti G :
 - $JAC_{\top}(x, y) = \frac{VC_{\top}(x, y)}{\|\Gamma_{\top}^n(x)\|}$
 - $JAC_{\perp}(x, y) = \frac{VC_{\perp}(x, y)}{\|\Gamma_{\perp}^m(y)\|}$



Attachement préférentiel : AP

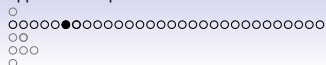
- $AP(x, y) = \|\Gamma(x)\| \times \|\Gamma(y)\|$
- intuition : deux nœuds qui ont beaucoup de relations ont tendance à être liés.
- *Ex. Une fan de shopping achète aussi des produits best-seller.*
- Modèle de construction de réseaux qui génère des réseaux sans échelle [BJN⁺02b]



Adamic Adar : AA I

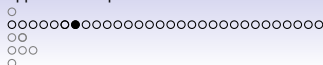
- Mesure proposée dans [AA03].
- Principe : Similarité entre nœuds = somme pondéré des *caractéristiques communes* des deux nœuds.
- Augmenter le poids des caractéristiques faiblement partagées dans le voisinage des deux nœuds.
- Exemple : caractéristique commune = nombre de voisins communs.
 - Dans un graphe unimodal :

$$AA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log(\|\Gamma(z)\|)}$$
 - Un voisin commun lié à 2 nœuds a le poids $\frac{1}{\log(2)} = 1.442$
 - Un voisin commun lié à 4 nœuds a le poids $\frac{1}{\log(4)} = 0.721$



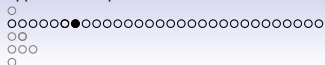
Adamic Adar : AA II

- Dans un graphe biparti :
- $AA_{\top}(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma_{\perp}(y)} \frac{1}{\log(\|\Gamma(z)\|)}$
- $AA_{\perp}(x, y) = \sum_{z \in \Gamma(y) \cap \Gamma_{\top}(x)} \frac{1}{\log(\|\Gamma(z)\|)}$



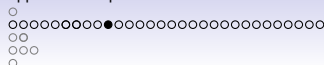
Mesures basées sur la distance I

- Chemin le plus court : $Sp(x, y)$
- Katz [Kat53] :
 - $Katz(x, y) = \sum_{l=1}^{\infty} \beta^l |paths(x, y)^l|$
 - $|paths(x, y)^l|$ est le nombre de chemins de longueur l qui relie x, y .
 - Rappel : $A^l(i, j) = paths(v_i, v_j)^l$
 - Calcul : $(I - \beta A_G)^{-1} - I$ où I est la matrice identité et A_G la matrice d'adjacence.
 - Très couteux pour des grands réseaux. Appliqué à un composant connexe.
- Mesure basée sur les marches aléatoires [FPRS07] :
 - Hit time.
 - temps de commutation .



Mesures basées sur la distance II

- MFA : Matrix-forest-based algorithm.
- ...

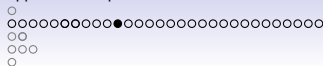


Expérimentations

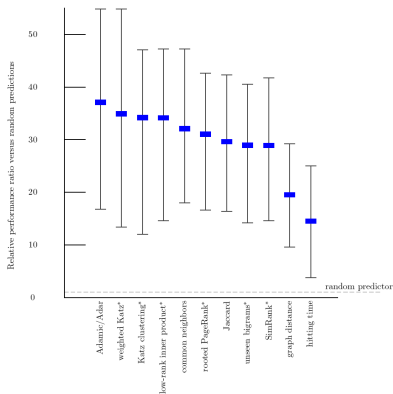
- 5 réseaux de co-auteurs extraits de la base ArXiv

	training period			Core		
	authors	papers	edges	authors	$ E_{old} $	$ E_{new} $
astro-ph	5343	5816	41852	1561	6178	5751
cond-mat	5469	6700	19881	1253	1899	1150
gr-qc	2122	3287	5724	486	519	400
hep-ph	5414	10254	17806	1790	6654	3294
hep-th	5241	9498	15842	1438	2311	1576

- Core : l'ensemble des auteurs qui ont écrits au moins 3 papiers pendant la période d'apprentissage et 3 papiers pendant la période d'étiquetage
- Prédicteurs : (1) Katz pondéré, (2) Katz classification, (3) low-rank inner product, (4) rooted Pagerank, (5) bigrammes inconnus, voisins communs non pondérés (6) SimRank,

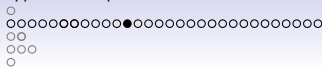


Résultats (1)

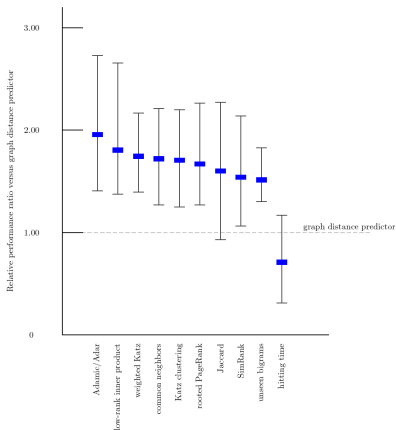


Comparaison / prédicteur aléatoire

Toutes les méthodes prédisent mieux que le prédicteur aléatoire
 → la topologie du réseau contient bien des informations permettant de prédire l'occurrence d'un lien

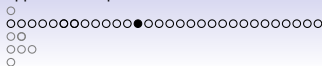


Résultats (2)

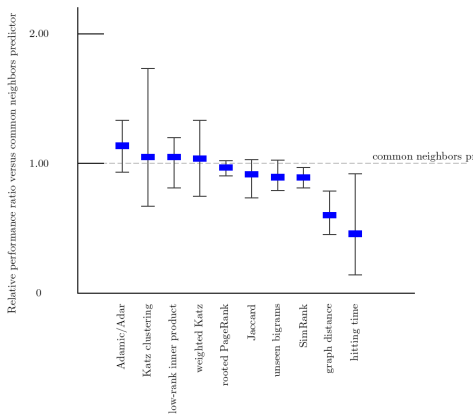


Qualité du prédicteur plus court chemin

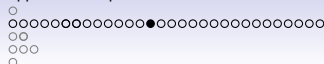
Pas de bons résultats à cause de la propriété *petit monde* (attribut peu discriminant)



Résultats (3)



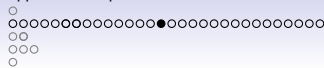
Prédicteur voisins
communs
Un prédicteur simple et
particulièrement efficace



Résultats (4)

	Adamic/Adar	Katz clustering	common neighbors	hitting time	Jaccard's coefficient	weighted Katz	low-rank inner product	rooted Pagerank	SimRank	unseen bigrams
Adamic/Adar	1150	638	520	193	442	1011	905	528	372	486
Katz clustering		1150	411	182	285	630	623	347	245	389
common neighbors			1150	135	506	494	467	305	332	489
hitting time				1150	87	191	192	247	130	156
Jaccard's coefficient					1150	414	382	504	845	458
weighted Katz						1150	1013	488	344	474
low-rank inner product							1150	453	320	448
rooted Pagerank								1150	678	461
SimRank									1150	423
unseen bigrams										1150

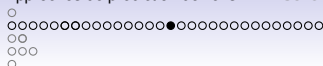
- Corrélation importante des attributs en terme de nombre de prédictions communes



Résultats (5)

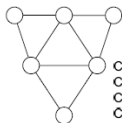
	Adamic/Adar	Katz clustering	common neighbors	hitting time	Jaccard's coefficient	weighted Katz	low-rank inner product	rooted Pagerank	SimRank	unseen bigrams
Adamic/Adar	92	65	53	22	43	87	72	44	36	49
Katz clustering		78	41	20	29	66	60	31	22	37
common neighbors			69	13	43	52	43	27	26	40
hitting time				40	8	22	19	17	9	15
Jaccard's coefficient					71	41	32	39	51	43
weighted Katz						92	75	44	32	51
low-rank inner product							79	39	26	46
rooted Pagerank								69	48	39
SimRank									66	34
unseen bigrams										68

- Corrélation importante des attributs en terme de nombre de prédictions **correctes** communes

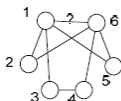


Coefficient de clustering pour prédire les liens

- [Hua06] : utilisation du coefficient de clustering généralisé. La probabilité d'occurrence d'un lien donné est évaluée en fonction du nombre de cycles (de tailles différentes) qui vont être formés en ajoutant ce lien.



$$\begin{aligned} C(2) &= 24/42 \\ C(3) &= 24/66 \\ C(4) &= 30/66 \\ C(5) &= 12/24 \end{aligned}$$



$$\Pr((i, j) \in E \mid |P_{ijk}| = m) = c_k^m / (c_k^m + (1-c_k)^m), \quad k > 1$$

$$P_{m_2, \dots, m_k} = \Pr((i, j) \in E \mid |P_{ij2}| = m_2, \dots, |P_{ijk}| = m_k) =$$

$$c_1 c_2^{m_2} \dots c_k^{m_k} / (c_1 c_2^{m_2} \dots c_k^{m_k} + (1-c_1)(1-c_2)^{m_2} \dots (1-c_k)^{m_k})$$

$$f(c_1, c_2, \dots, c_k) = \sum_i \#(G_i) \Pr(G_i) \Pr((1, k+1) \in E \mid G_i)$$

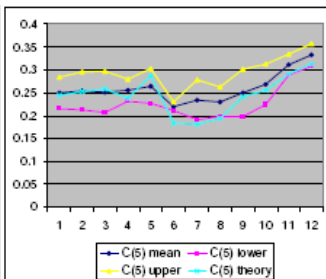
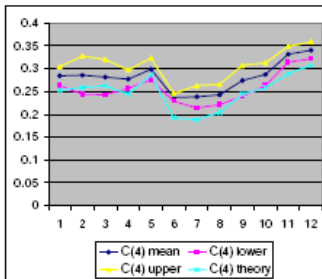
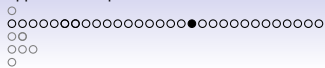
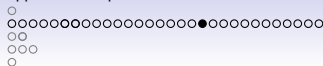


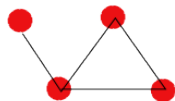
Table 4. Number of new links and AUC measures for the cycle formation and benchmark link prediction algorithms

Month (t)	New Links in G_t	$CF(3)$	PA	SA	GM
1	46	0.76764	0.52091	0.70977	0.67237
2	46	0.76748	0.48449	0.67891	0.69189
3	46	0.81385	0.44235	0.72598	0.73472
4	62	0.82106	0.45057	0.72143	0.75518
5	95	0.73883	0.45141	0.63052	0.69805
6	72	0.71737	0.41585	0.67084	0.70634
7	88	0.73977	0.48104	0.70304	0.69231
8	168	0.74056	0.60122	0.69806	0.65201
9	90	0.72065	0.46032	0.69501	0.70263
10	213	0.74384	0.48877	0.69840	0.74500
11	123	0.74860	0.47632	0.72477	0.75116
12	56	0.70018	0.44078	0.69573	0.76273



Combinaison de prédicteurs

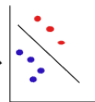
Utilisation des techniques d'apprentissage automatique supervisé.



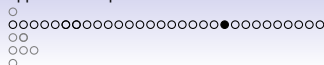
Réseau



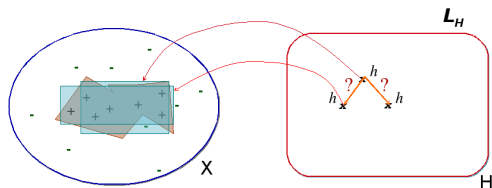
Génération d'attributs
pour chaque lien potentiel



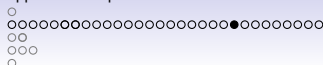
Classification (prédiction)



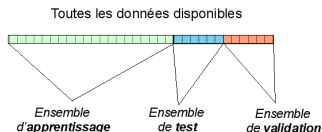
Apprentissage supervisé



- Quel critère inductif ?
- Qu'est-ce qu'une hypothèse optimale étant donné l'échantillon d'apprentissage ?
- Quelle méthode d'exploration de H ?



Evaluation empirique

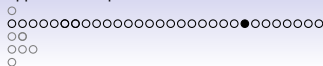


- Erreur réelle :

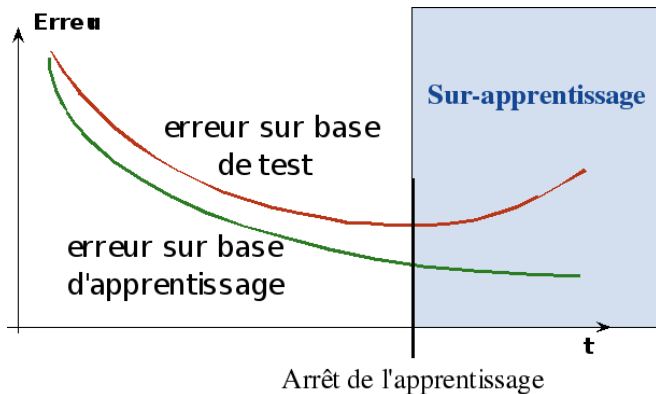
$$e_D = \int_D |y - f(x, \theta)| p(x, y) dx, y$$

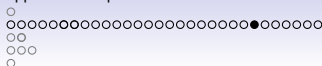
- Erreur estimée :

$$\hat{e}_S = \frac{1}{m} \sum_{\langle x, y \rangle \in S} |y - f(x, \theta)|$$



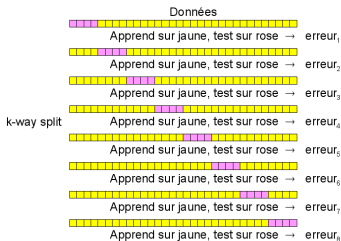
Problème de sur-apprentissage



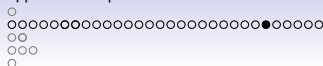


10 validation croisée

- Si peu de données disponibles
- Avantages : simplicité, généralité, temps de calcul raisonnable, robustesse, etc.
- Choix du nombre de plis : compromis entre le biais (le critère de validation croisée "n-fold" surestime d'autant moins l'erreur de prédiction que n est grand), la variabilité du critère, et le temps de calcul.



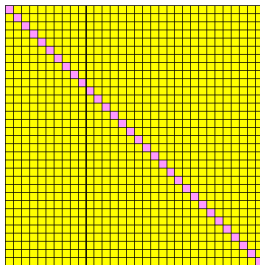
$$\text{erreur} = \sum \text{erreur}_i / k$$

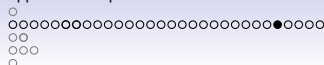


Validation par leave one out

- Si très peu de données disponibles
- Faible biais, haute variance
- Tendance à sous-estimer l'erreur si les données ne sont pas vraiment i.i.d.

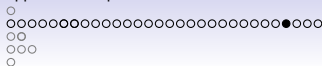
Données





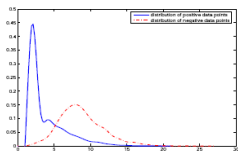
Apprentissage automatique pour la prédiction de liens

- [HCSZ06] : prédiction de liens \rightarrow problème de classification binaire
 - exemple + : apparition des liens dans le graphe G_{i+1}
 - exemple - : non apparition des liens dans le graphe G_{i+1}
- Utilisation des informations topologiques dans le graphe des co-auteurs + information sur les nœuds, si disponible
- (Sous-)exploitation d'un graphe biparti auteurs - mots-clés : un lien entre un auteur et un mot-clé si mot-clé utilisé dans au moins un article, deux mots-clés qui apparaissent dans un même abstract sont connectés.
- Les autres attributs n'apportent pas d'information supplémentaires pour la discrimination

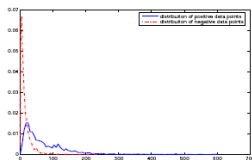


Application aux réseaux de co-publications (1)

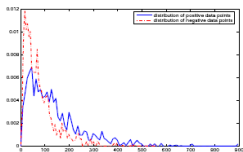
DBLP (1990-2000 : apprentissage, 2001-2004 : étiquetage) et
 BIOBASE (1998-2001 : apprentissage, 2002 : étiquetage)



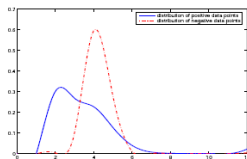
(a) Shortest Distance



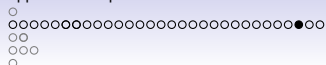
(b) Sum of Neighbors count



(c) Sum of neighbors count



(d) Shortest distance



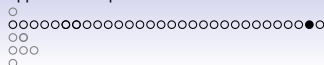
Application aux réseaux de co-publications (2)

Attribute name	Information gain	Gain Ratio	Chi-Square Attribute Eval.	SVM feature evaluator	Avg. Rank
Sum of Papers	4	4	4	2	4
Sum of Neighbors	3	3	3	4	3
Shortest distance	1	1	1	1	1
Second shortest distance	2	2	2	3	2

Attributs utilisés sur la base DBLP

Classification model	Accuracy	Precision	Recall	F-value	Squared Error
Decision Tree	82.56	87.70	79.5	83.40	0.3569
SVM(Linear Kernel)	83.04	85.88	82.92	84.37	0.1818
SVM(RBF Kernel)	83.18	87.66	80.93	84.16	0.1760
K_Nearest Neighbors	82.42	85.10	82.52	83.79	0.2354
Multilayer Perceptron	82.73	87.70	80.20	83.70	0.3481
RBF Network	78.49	78.90	83.40	81.10	0.4041
Naive Bayes	81.24	87.60	76.90	81.90	0.4073
Bagging	82.13	86.70	80.00	83.22	0.3509

Résultats sur la base DBLP

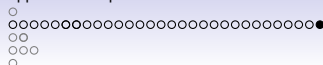


Application aux réseaux de co-publications (3)

Attribute name	Information gain	Gain Ratio	Chi-Square Attribute Eval.	SVM feature evaluator	Avg. Rank
Sum of Papers	3	4	3	4	3
Sum of Neighbors	1	3	1	2	2
Sum of KW count	6	6	6	3	5
Sum of Classification count	5	5	5	6	5
KW match count	2	1	2	1	1
Sum of log of Sec. Neighbor. count	7	7	7	8	7
Shortest distance	4	2	4	5	4
Clustering Index	9	9	9	7	8
Shortest dist. in KW-Author graph	8	8	8	9	8

Attributs utilisés sur la base BIOBASE

Classification model	Accuracy	Precision	Recall	F-value	Squared Error
Decision Tree	90.01	91.60	89.10	90.40	0.1306
SVM(Linear Kernel)	87.78	92.80	83.18	86.82	0.1221
SVM(RBF Kernel)	90.56	92.43	88.66	90.51	0.0945
K_Nearest Neighbors	88.17	92.26	83.63	87.73	0.1826
Multilayer Perceptron	89.78	93.00	87.10	90.00	0.1387
RBF Network	83.31	94.90	72.10	81.90	0.2542
Naive Bayes	83.32	95.10	71.90	81.90	0.1665
Bagging	90.87	92.5	90.00	91.23	0.1288



Apprentissage supervisé : Bilan

- Avantage : Possibilité de combiner les attributs (topologiques) afin d'obtenir un meilleur taux de prédiction
- Inconvénient : problème difficile d'apprentissage
- Classes non équilibrées : Taux d'exemple positifs $\sim 0.1\%$ des exemples négatifs.
- Exemples non i.i.d.
- Solutions à envisager : échantillonnage , apprentissage semi-supervisé ?



Approches basées sur l'analyse des attributs des nœuds

- Principe : Apprendre la dépendance entre l'existence d'un lien et les attributs des nœuds reliés par le lien.
- Modèles probabilistes [TWAK03] :
 - Réseaux Bayésiens
 - Réseaux de Markov Relationels
 - Apprentissage statistique relationnel
 - ...
- Nécessite d'avoir accès aux attributs des nœuds.
- Domaine-dependant.



Techniques hybrides I

Principe : Combiner des approches topologiques et des approches basées sur le contenu des nœuds.

Exemple : Alignement de matrices [STCE08]

- soit G un réseau décrit par la matrice d'adjacence A_G
- Soit X est la matrice (de dimension $n \times d$) des attributs des nœuds. n est le nombre de nœuds et d est le nombre d'attributs.
- XX^T est une matrice $n \times n$ de similarités entre les attributs des nœuds.
- Dans le cas idéal on cherche à avoir :

$$\forall i, j A_g[i, j] = XX^T[i, j] \text{ (après normalisation)}$$



Techniques hybrides II

- Approche : Trouver W qui minimise : $\| A_G - XWX^T \|^2$
- Le score d'un lien est donné par $score(x, y) = xWy^T$
- Expérimentations : tester sur DBLP, TakingItGlobal.com, WebKB
- Gain en Précision de ~ 0.2 avec l'utilisation de W



Approches temporelles

Deux grandes approches :

- Utilisation d'attributs temporels
- Résumé des graphes

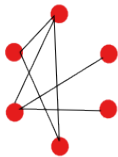


Indicateurs temporels

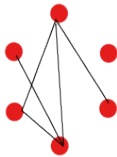
- Indicateurs *financiers* [Coo06]
- *profit* d'un indicateur topologique sur une période de temps.
- Exemple : $\frac{D_{tf}(v) - D_{ti}(v)}{D_{tf}(v)}$
- Moyen temporel : $\sum_{t=ti}^{t=tf} \frac{D_t(v)}{tf-ti}$
- date du dernier lien
- ...



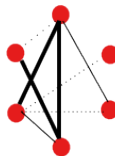
Resumé des graphes



$G(T-1)$



$G(T)$



Graphe resumé [SN07]



Approches structurelles

- Caractérisation des liens par des sous-graphes fréquents dans le réseaux temporels.
- Apprentissage de règle de réécriture des graphes [LBW07]



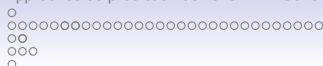
Conclusion I

- Approches topologiques viables pour la prédiction de liens.
- Beaucoup de travaux sur des graphes statiques.
- Approches topologiques généralistes à combiner avec des approches basées sur l'exploration des caractéristiques des nœuds.
- Pas de benchmark pour faciliter la comparaison d'approches.
- Axes de recherche :
 - Analyse des réseaux bipartis, valués, et temporels
 - Exploration simultanée de réseaux hétérogènes, approche de graphe mining.



Conclusion II

- Approches structurelles : Fouille parallèle des grands graphs.
-



Bibliographie I



Lada ADAMIC et Eytan ADAR :

Friends and neighbors on the web.
Social Networks, 25(3):211–230, july 2003.



L. BREIMAN, J. FRIEDMAN, R. OHLSEN et C. STONE :

Classification and regression trees.
Wadsworth International Group, Belmont, CA, 1984.



A. L. BARABASI, H. JEONG, Z. NEDA, E. RAVASZ, A. SCHUBERT et T. VICSEK :

Evolution of the social network of scientific collaborations.
PHYSICA A, 311:3, 2002.



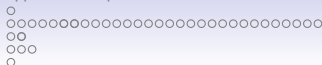
Albert-László Barabási, H. Jeong, Z. Neda, E. Ravasz, A. Schubert et T. Vicsek :

Evolution of the social network of scientific collaboration.
Physica A, 311(3-4):590–614, 2002.



Nesserine BENCHETTARA, Rushed KANAWATI et Céline ROUVEIROL :

Calcul de recommandation par prédiction de liens dans un graphe biparti.
In Actes de l'atelier sur l'apprentissage et graphes pour les systèmes complexes (plate-forme AFIA 2009),
Hammett, Tunisie, Mai 2009.



Bibliographie II



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, November 2006.



Francois FOUSS, Alain PIROTTE, Jean-Michel RENDERS et Marco SARENS :

Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation.
IEEE Transactions on knowledge and data engineering, 19(3):355–369, 2007.



A GOLDENBERG, J. KUBICA, P. KOMAREK, A. MOORE et J. SCHNEIDER :

A comparison 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.



Mohammad Al HASAN, Vineet CHAOJI, Saeed SALEM et Mohammed ZAKI :

Link prediction using supervised learning.
In SIAM Workshop on Link Analysis, Counterterrorism and Security with SIAM Data Mining Conference, Bethesda, MD, 2006.



Petter HOLME, Christofer R. EDLING et Fredrik LILJEROS :

Structure and time-evolution of an internet dating community.
Social Networks, 26:155–174, 2004.



Bibliographie III



Zan HUANG, Xin LI et Hsinchun CHEN :

Link prediction approach to collaborative filtering.

In Mary MARLINO, Tamara SUMNER et Frank M. Shipman III, éditeurs : JCDL, pages 141–142. ACM, 2005.



Zan HUANG :

Link prediction based on graph topology : The predictive value of the generalized clustering coefficient.

In Proceedings of LinkKDD'06, Philadelphia, Pennsylvania, 2006.



Jian HUANG, Ziming ZHUANG, Jia LI et C. Lee GILES :

Collaboration over time : characterizing and modeling network evolution.

In Marc NAJORK, Andrei Z. BRODER et Soumen CHAKRABARTI, éditeurs : WSDM, pages 107–116. ACM, 2008.



Jian HUANG, Ziming ZHUANG, Jia LI et C. Lee GILES :

Collaboration over time : characterizing and modeling network evolution.

In WSDM, pages 107–116, 2008.



Rushed KANAWATI :

On using sna techniques for enhancing performances of on-line help desks.

In Proceedings of IADIS International Conference on E-commerce, Amsterdam, July 2008. IADIS.



L. KATZ. :

A new status index derived from sociometric analysis.

Psychometrika, 18(1), 18(1):39–43, 1953.



Bibliographie IV



Rushed KANAWATI et cÃ©line ROUVEIROL :

Lips : A sna-based system for intelligent management of academic conferences.

In Proceedings of the 4th International conference on application of Social networkk analysis, Zurich, october 2009.



Mayank LAHIRI et Tanya Y. BERGER-WOLF :

Structure prediction in temporal networks using frequent subgraphs.

In CIDM, pages 35–42. IEEE, 2007.



Jean loup GUILLAUME et Matthieur LATAPY :

Bipartite structure of all complex networks.

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



Neal LATHIA, Stephen HAILES et Licia CAPRA :

knn cf : a temporal social network.

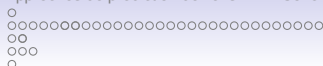
In Pearl PU, Derek G. BRIDGE, Bamshad MOBASHER et Francesco RICCI, Ã©diteurs : RecSys, pages 227–234. ACM, 2008.



David LIBEN-NOWELL et Jon M. KLEINBERG :

The link-prediction problem for social networks.

JASIST, 58(7):1019–1031, 2007.



Bibliographie V



Jure LESKOVEC, Jon KLEINBERG et Christos FALOUTSOS :

Graphs over time : densification laws, shrinking diameters and possible explanations.

In KDD '05 : Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pages 177–187, New York, NY, USA, 2005. ACM.



David LIBEN-NOWELL :

An Algorithmic Approach to Social networks.

Thèse de doctorat, M.I.T., june 2005.



M. LIM, M. NEGNEVITSKY et J. HARTNETT :

Artificial intelligence applications for analysis of e-mail communication activities.

In Proceedings of International Conference On Artificial Intelligence In Science And Technology, pages 109–113, 2003.



Tsuyoshi MURATA et Sakiko MORIYASU :

Link prediction based on structural properties of online social networks.

new Generation Computing, 26:245–257, 2008.



Akira NINAGAWA et Koji EGUCHI :

Link prediction in metabolic networks using topology-based mixture models.

In Proceedings of the 19th International Conference on Genome Informatics, Gold Coast , Australia, December 2008.



Bibliographie VI



M. E. J. NEWMAN :

Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality.
Phys. Rev. E, 64(1):016132, Jun 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.



Jung Hun OHN, Jihoon KIM et Ju Han KIM :

Social network analysis of gene expression data.
In Proceedings of AMIA symposium : Biomedical and health informatics, pages 958–958, Washington D.C., November 2003. AMIA.



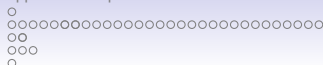
Milen PAVLOV et Ryutarō ICHISE :

Finding experts by link prediction in co-authorship networks.
In Anna V. ZHDANOVA, Lyndon J. B. NIXON, Malgorzata MOCHOL et John G. BRESLIN, éditeurs : FEWS, volume 290 de *CEUR Workshop Proceedings*, pages 42–55. CEUR-WS.org, 2007.



Matthew J. RATTIGAN et David JENSEN :

The case for anomalous link discovery.
SIGKDD Explorations, 7(2):41–47, 2005.



Bibliographie VII



Mukund SESHADRI, Sridhar MACHIRAJU, Ashwin SRIDHARAN, Jean BOLOT, Christos FALOUTSOS et Jure LESKOVEC :
Mobile call graphs : beyond power-law and lognormal distributions.
In Ying LI, Bing LIU et Sunita SARAWAGI, éditeurs : *KDD*, pages 596–604. ACM, 2008.



U. SHARAN et J. NEVILLE :
Exploiting time-varying relationship in staistical relational models.
In *Proceedings of the first SNA-KDD workshop*, 2007.



Jerry SCRIPPS, Pang-Ning TAN, Feilong CHEN et Abdol-Hossein ESFAHANIAN :
A matrix alignment approach for link prediction.
In *ICPR*, pages 1–4. IEEE, 2008.



Benjamin TASKAR, Ming Fai WONG, Pieter ABBEEL et Daphne KOLLER :
Link prediction in relational data.
In Sebastian THRUN, Lawrence K. SAUL et Bernhard SCHÖLKOPF, éditeurs : *NIPS*. MIT Press, 2003.



J. ZHU :
Mining Web Site Link Structures for Adaptive Web Site Navigation and Search.
Thèse de doctorat, University of Ulster, 2003.