Fouille de Graphes (Dynamiques) Attribués

Marc Plantevit

LIRIS UMR5205

marc.plantevit@univ-lyon1.fr

Systèmes Complexes

Définition [wikipedia]

"Un système complexe est un système composé d'un grand nombre d'entités en interaction locale et simultanée."

Systèmes Complexes

Définition [wikipedia]

"Un système complexe est un système composé d'un grand nombre d'entités en interaction locale et simultanée."

Diversité $\Rightarrow$ étude interdisciplinaire

Deux approches complémentaires utilisées :

- certains scientifiques aux prises avec un système complexe particulier cherchent à le comprendre ;
- d'autres cherchent des méthodes et définitions générales applicables à de nombreux systèmes différents.

Évolution des sciences

- Avant 1600 : empirical science
- 1600-1950 : theoretical science
  - Chaque discipline a acquis des volets théoriques. Les modèles théoriques sont sources d'expériences et généralise notre compréhension.
- Années 50 - Années 90 : computational science
  - La plupart des disciplines ont vu se développer une nouvelle branche : la branche *computationnelle* ou informatique (e.g. l'écologie et la physique empirique, théorique et computationnelle).
  - Simulation $\subset$ computational science. Elle est née de notre incapacité à trouver des solutions analytiques pour des modèles mathématiques complexes.
- 1990 - Aujourd'hui : data science
  - Énorme volumes de données omniprésents (nouveaux instruments scientifiques, simulations)
  - Capacité de stocker et gérer des petaoctets de données en ligne.
  - Internet et les grilles ont rendu ces données "universellement" accessibles.
  - Devant ces données ubiquitaires, la fouille de données : un challenge majeur ! ! !

Fouille de Données et Systèmes Complexes

Données = Pétrole du XXI^{ème} siècle

- Permettre aux experts de tirer partie des données.
  - Données **réelles** ou issues de **simulations**.
- Découverte de connaissances (motifs, modèles, etc.) par des approches complètes.

Données les plus adaptées pour l'étude de systèmes complexes ?

- données transactionnelles,
- séquences,
- **graphes** (dynamiques)

Fouille de Graphes

Pourquoi un tel engouement ?

- De grands graphes sont désormais disponibles (réseaux sociaux, biologie, etc.).
- De nouvelles questions !

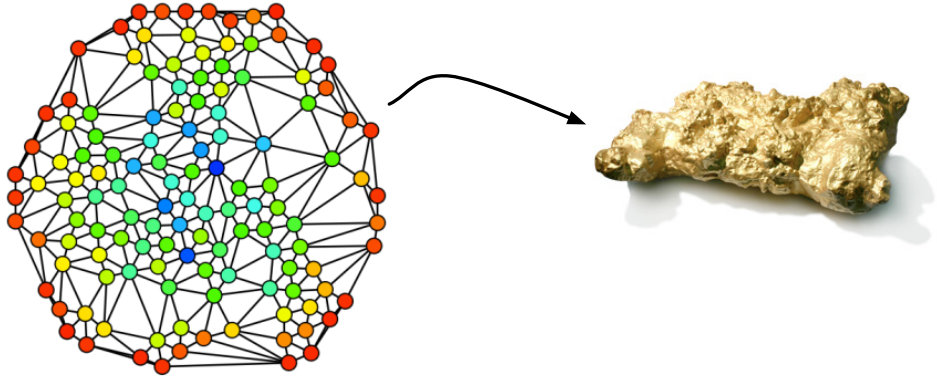
Fouille de Graphes

Pourquoi un tel engouement ?

- De grands graphes sont désormais disponibles (réseaux sociaux, biologie, etc.).
- De nouvelles questions !

Défi : la découverte de connaissances dans des grands graphes.

Extraire des connaissances utiles (motifs, relations, etc.) dans des données structurées représentées sous forme de graphe(s).



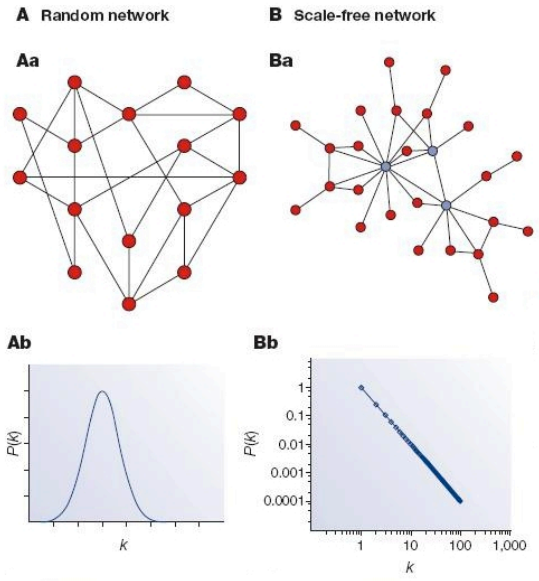
Des données aux connaissances.

Travaux Actuels : 2 Grands Types d'Approches

Étude des propriétés topologiques des graphes

- Diamètres.
- Mise en évidence de power laws.
- Mesures de centralités.
- Vision macroscopique.

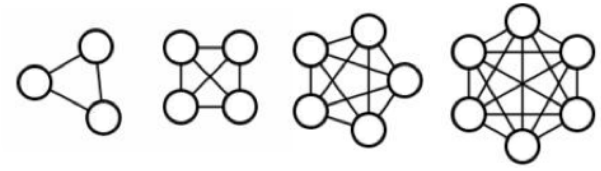
[Barabási, Science , 1999]



Fouille de motifs locaux

- Sous-graphes fréquents.
- Sous-graphes denses.
- Une vision microscopique.

[X. Yan, IEEE ICDM , 2003]



Notre Proposition

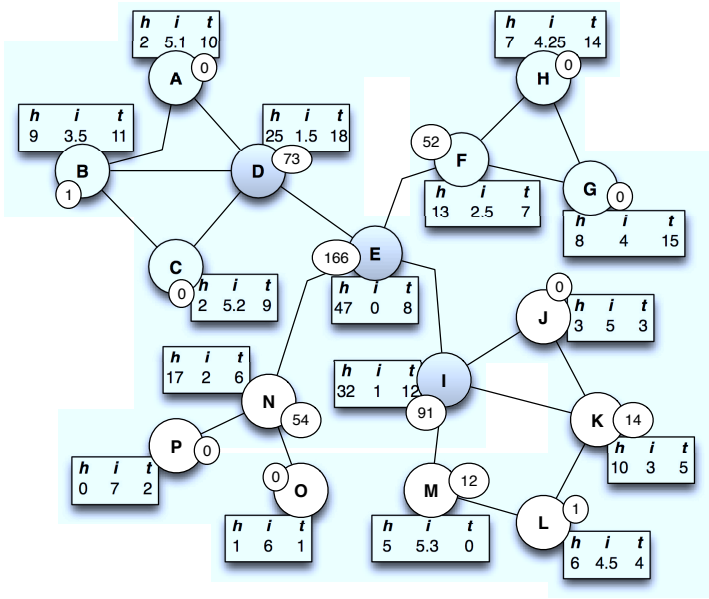
“Graphes attribués” :

- Des attributs *locaux* fournissent des informations sur les entités représentées commes des nœuds.

Notre Proposition

"Graphes attribués" :

- Des attributs *locaux* fournissent des informations sur les entités représentées commes des nœuds.

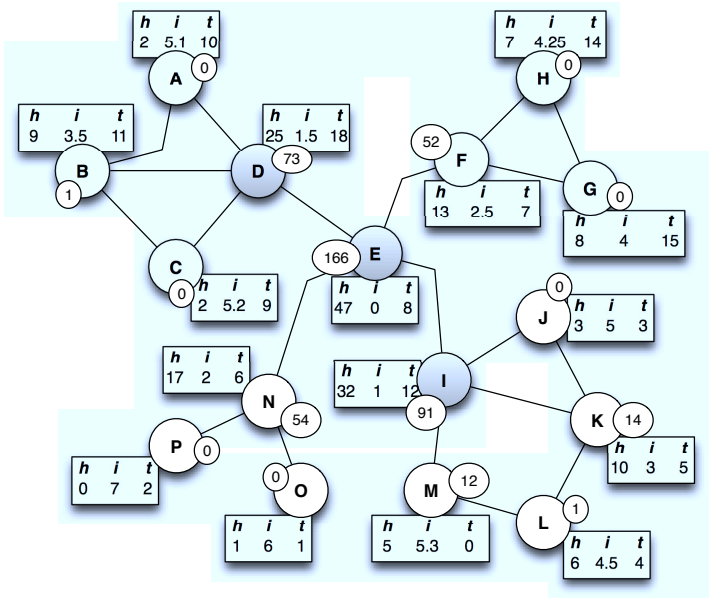


Réseaux de co-auteurs.

Notre Proposition

"Graphes attribués" :

- Des attributs *locaux* fournissent des informations sur les entités représentées commes des nœuds.



Réseaux de co-auteurs.

Macro + Micro ⇒ Vision "Mésoscopique" !

Nous voulons considérer simultanément les deux visions :

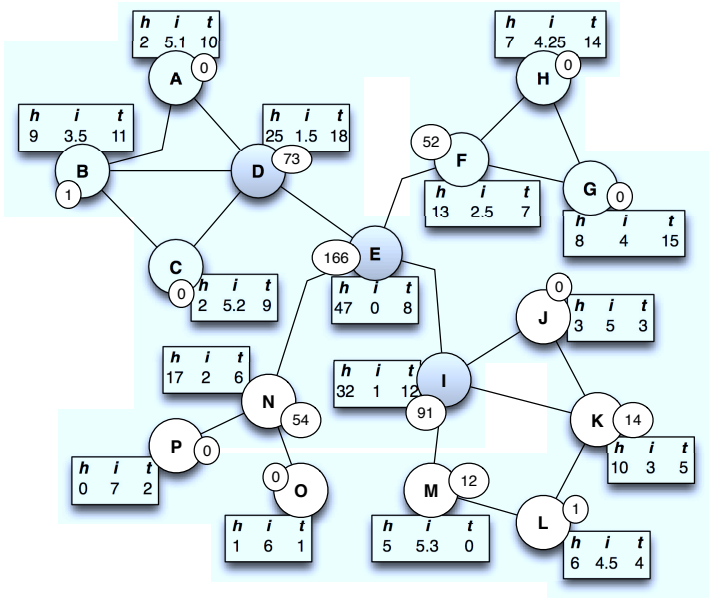
- Idée : Découvrir des motifs fréquents contenant :
  - **propriétés locales** des nœuds,
  - et leurs **propriétés topologiques** dans le graphe (données par des mesures statistiques).

Exemples de Motifs

Réseau de co-auteurs

Attributs locaux :

- h : h-index,
- i : le nombre d'heures d'enseignement hebdomadaire,
- t : le nombre de publications dans une revue données (Nature).



Réseaux de co-auteurs.

Motifs que nous pouvons découvrir :

- Plus le h-index est important, plus la centralité des auteurs dans le graphe est forte.
- Plus le nombre d'heures hebdomadaires est important :
  - Plus faible le nombre de publications à Nature.
  - Plus faible la centralité dans le graphe.

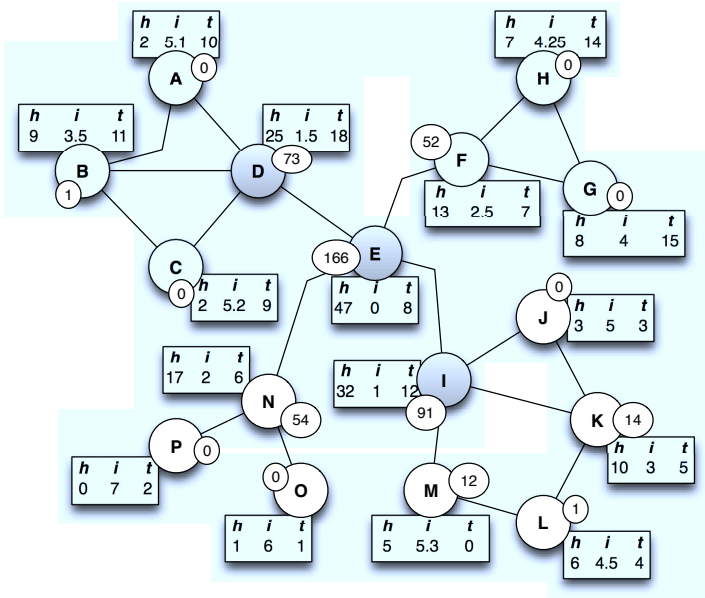
Représentation des Motifs

- *Plus le h-index est important, plus la centralité (BETW) des auteurs dans le graphe est forte.*
  - $\{h^+, BETW^+\}$
- *Plus le nombre d'heures hebdomadaires est important, plus faible le nombre de publications à Nature (t).*
  - $\{i^+, t^-\}$
- $\{i^+, BETW^-\}$

Projection des Motifs : Retour dans le Graphe

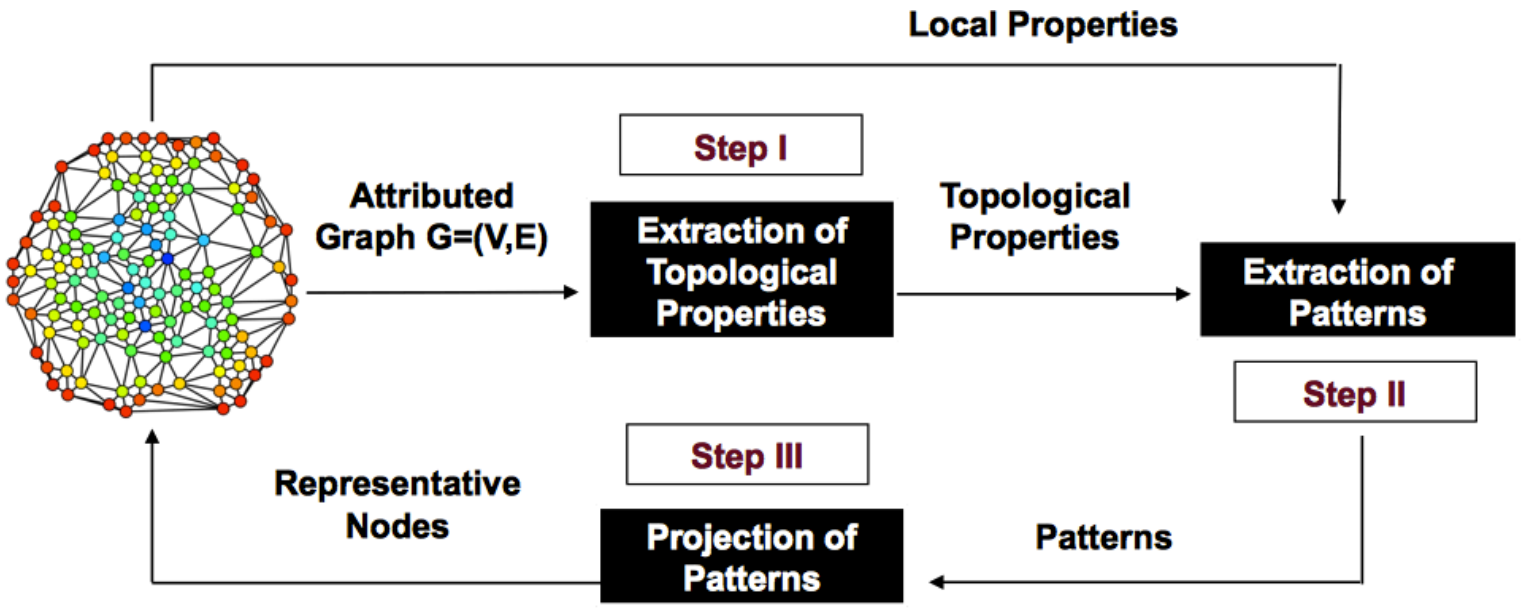
Identifier des nœuds représentatifs du motif :

- Qui sont les auteurs qui représentent le mieux le motif $\{h^+, i^-, BETW^+\}$?



Panorama de Notre Approche

Extraction de motifs topologiques dans des graphes attribués :



Étape 1 : Extraction de Propriétés Topologiques

Pour chaque nœud du graphe, des propriétés topologiques sont calculées :

- En s'appuyant sur le voisinage direct :
  - Degree centrality,
  - Clustering coefficient,
  - ...
- ou en prenant en compte tout le graphe :
  - Betweenness centrality,
  - Closeness centrality,
  - PageRank Index,
  - ...

Étape 2 : Extraction des Motifs

Entrée : un tableau !

Une ligne = attributs locaux + propriétés topologiques pour 1 nœud.

But :

Trouver les motifs parmi les attributs locaux et les propriétés topologiques (calculées à l'étape 1).

Comment ?

- Similaire à la fouille d'itemsets :
  - Énumérer tous les ensembles d'attributs signés,
  - Calculer la fréquence des motifs dans les données, et
  - Retourner les motifs fréquents (par rapport à un seuil).

Question :

Comment calculer la fréquence de tels motifs ?

Généralisation de la Mesure de Calders et al.

[Calders et al., ACM SIGKDD, 2006]

- $P = \{h^+, BETW^+\}$
  - $Supp(P)$ = proportion du nombre de couples de nœuds (u, v) t. q. : $u.h < v.h$ et $u.BETW < v.BETW$

- $P = \{\mathrm{i}^+, \mathrm{t}^-\}$
  - $Supp(P)$ = proportion du nombre de couples de nœuds (u, v) t. q. : $u.i < v.i$ et $u.t > v.t$

Plus Formellement :

Le support d'un motif mésoscopique est calculé à partir d'une généralisation de la mesure du τ de Kendall comme suit :

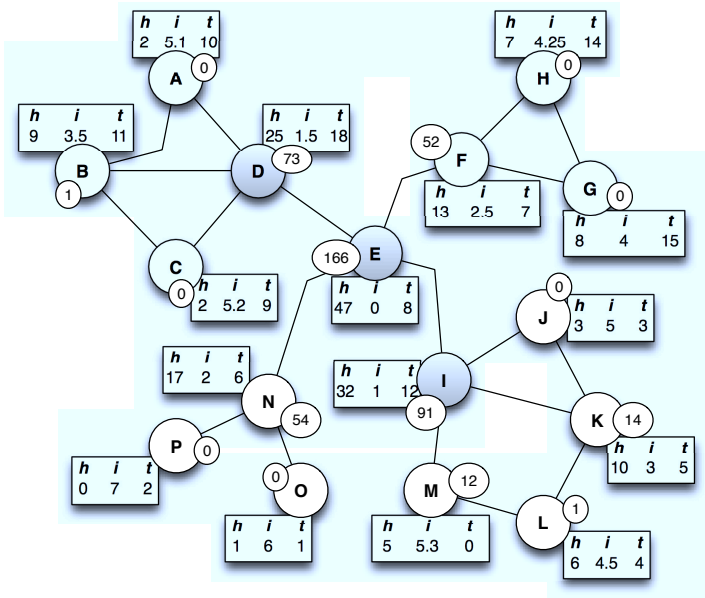
$$Supp_{\tau}(P) = \frac{|\{(u,v) \in V^2 \mid \forall A^s \in P : A(u) \rhd_s A(v)\}|}{\binom{|V|}{2}}$$

Si $s = +$, $\rhd_s$ signifie $<$, autrement $>$.

Exemple

$P = \{h^+, i^-, BETW^+\}$

- Supporté par 89 couples.
- Ex : (D,E)
- $supp_\tau(P) = \frac{89}{\binom{16}{2}} = \frac{89}{120} = 0.74$



Optimisations

Opération quadratique dans le nombre de sommets, mais :

- Le support est “anti-monotone” (monotone décroissant) ;
- Inutiles de considérer les motifs “symétriques” :
  $Supp(\{A^+, B^-\}) = Supp(\{A^-, B^+\})$
- Une borne supérieure sur le support calculable en temps linéaire !

Motifs Émergents (Attribut)

A^+B^+ et A^+B^- tous les deux fréquents ?
$\Rightarrow$ Contradiction ! ! !

Si le support de A^+ est significativement plus élevé dans la "*classe*" B^+ (par rapport à un taux de croissance) ?
$\Rightarrow$ A^+B^+ semble plus intéressant !

Nous définissons les motifs émergents dans ce contexte.

Un attribut $\longrightarrow$ 3 classes :

- $C^+ = \{(u,v) \mid (C(u) < C(v)) \wedge (u \neq v)\}$
- $C^- = \{(u,v) \mid (C(u) > C(v)) \wedge (u \neq v)\}$
- $C^= = \{(u,v) \mid (C(u) = C(v)) \wedge (u \neq v)\}$

$$gr(P, C^\star) = \frac{Supp(P \cup C^\star)}{Supp(P \cup \overline{C^\star})} \cdot \frac{Supp(\overline{C^\star})}{Supp(C^\star)}$$

Motifs Émergents (par rapport à la structure)

$$\mathcal{C}_E \quad = \quad \{(u,v) \in V^2 \mid \{u,v\} \in E\}$$

Définition ($Supp_E$)

Le support d'un motif topologique P sur les couples de sommets connectés dans G est :

$$Supp_E(P) = \frac{2|\{(u,v) \in \mathcal{C}_E \mid \forall D^s \in P : D(u) \rhd_s D(v)\}|}{|\mathcal{C}_E|}$$

$$Gr(P,E) = \frac{Supp_E(P)}{Supp_{\overline{E}}(P)}$$

Exemple

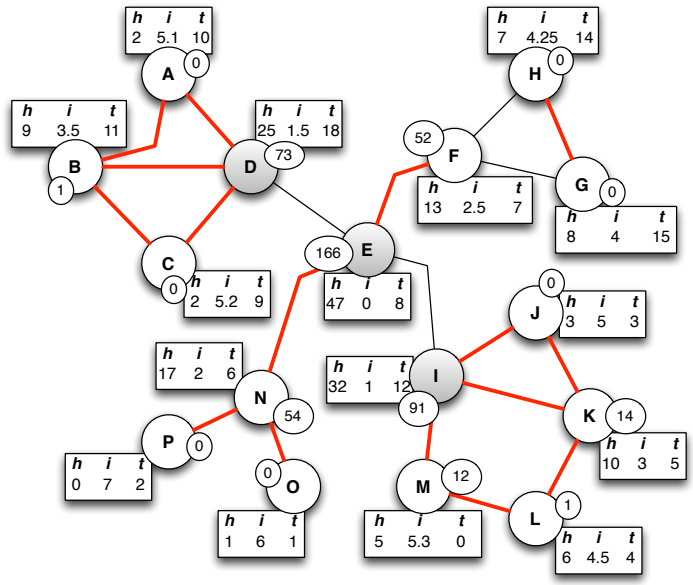
Quid du motif h^+, t^+ ?

Émergence par rapport à t

- $gr(h^+, C^{t^+}) = 2.13$

Émergence par rapport à la structure

- $gr(\{h^+, t^+\}, E) = 1.23$



Expérimentations : Deux graphes d'interactions "faciles" à interpréter

Graphe de films

- Netflix et IMDb (1998-2005).
- Des utilisateurs notent des films (1 à 5 étoiles).
- 5,972 nœuds (films).
- 64 338 arêtes (un acteur en commun).
- 5 attributs locaux :
  - Release year, num_customers (qui ont noté le film), avg_rating, stdev_rating, et num_actors
- 9 propriétés topologiques.

Graphe de co-auteurs (DBLP)

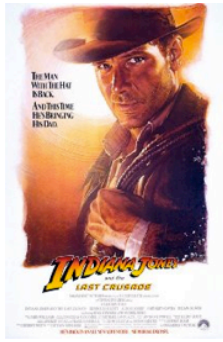
- 42 252 nœuds.
- 210 320 arêtes (un article en commun).
- 29 attributs locaux :
  - Nb publications depuis 1990 [a].
- 9 propriétés topologiques.

a. KDD, ICDM, ECML/PKDD, PAKDD, SIAM DM, AAAI, ICML, IJCAI, IDA, DAS-FAA, VLDB, CIKM, SIGMOD, PODS, ICDE, EDBT, ICDT, SAC, Data Min. Knowl. Discov., IEEE TKDE, IEEE Int. Sys., SIGKDD Exp., Comm. ACM, IDA J., KAIS, SADM, PVLDB, VLDB J., ACM TKDD.

Quelques Résultats

$\{avg\_rating^{+}, num\_customers^{+}\}$

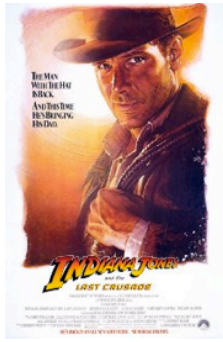
"*Les gens notent les films qu'ils ont aimés*"



Quelques Résultats

$\{avg\_rating^+, num\_customers^+\}$

"*Les gens notent les films qu'ils ont aimés*"



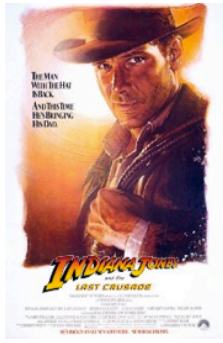
$\{num\_customers^+, Degree^+\}$

"*Les gens notent les films avec acteurs majeurs.*" (e.g., R de Niro, S. Connery, et T. Hanks)

Quelques Résultats

$\{avg\_rating^+, num\_customers^+\}$

"*Les gens notent les films qu'ils ont aimés*"



$\{num\_customers^+, Degree^+\}$

"*Les gens notent les films avec acteurs majeurs.*" (e.g., R de Niro, S. Connery, et T. Hanks)

$\{stdev\_rating^+, PageRank^-\}$

"*Les films controversés sont isolés*"

Résultats sur DBLP

Publier à SAC est il pénalisant ?

- $\{SAC^+, ECML/PKDD^-\}, \{SAC^+, KDD^-\}, \{SAC^+, VLDB^-\}$
- $\{SAC^+, PageRank^-\}$

Résultats sur DBLP

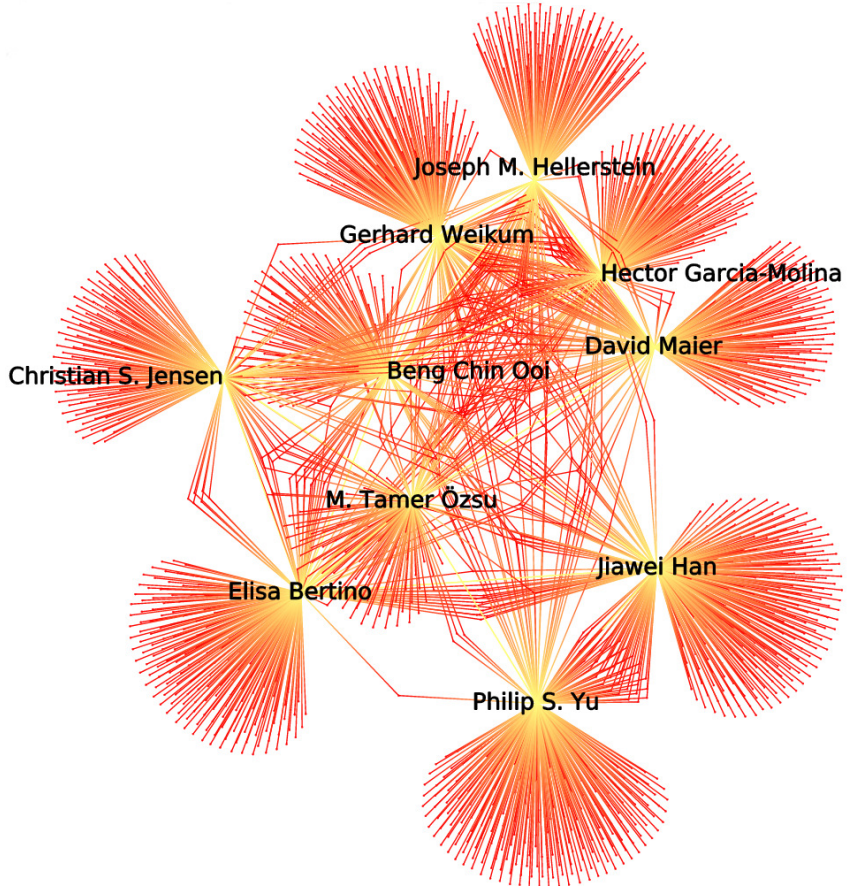
Publier à SAC est il pénalisant ?

- $\{SAC^+, ECML/PKDD^-\}, \{SAC^+, KDD^-\}, \{SAC^+, VLDB^-\}$
- $\{SAC^+, PageRank^-\}$
- Bien sûr que non !
- Biais introduit dans les données (SAC a un spectre bien plus large que les bases de données et la fouille de données).

☞ Choix de la granularité.

$P = \{PVLDB^+, Betw^+\}$

- $Gr(P, E) \simeq 7$



La base de requêtes plus sophistiquée :

- Quelles sont les conférences/journaux les plus impactant par rapport à une mesure de centralité ?
- etc.

Graphe d'interactions géniques

- Attributs locaux : expression du gènes dans 348 situations biologiques.
- 4711 gènes et 6036 arêtes.

Pertinence des motifs par rapport à la littérature ?

- On s'intéresse aux gènes représentatifs pour des motifs combinant tissus sains et tissus cancéreux.
- Calcul des rangs normalisés de deux ensembles spécifiques de gènes connus comme sur-exprimés :
  - Pour le cancer du pancréas : HLADRB4, PAPDC1B, et THBS1 *[D. Campagna et al., Int. J. Clin. Exp. Pathol., 2008]*.
  - Pour le cancer du colon : ANXA1, GJB2, PSMC5, et RPS7 *[V. Orian-Rousseau et al, Int. J. Cancer, 2005]*.

P	Pattern	Measures	PANCREAS AVG RANK	COLON AVG RANK
P_1	PANCREAS NORMAL$^-$ PANCREAS ADENOCARCINOMA$^+$	$Supp_{all} = 0.0125$ $Gr(P, E) = 1.877$	0.378	0.308
P_2	PANCREAS NORMAL$^-$ PANCREAS CARCINOMA$^+$	$Supp_{all} = 0.0097$ $Gr(P, E) = 1.941$	0.510	0.183
P_3	COLON NORMAL$^-$ COLON ADENOCARCINOMA$^+$	$Supp_{all} = 0.0162$ $Gr(P, E) = 1.586$	0.821	0.230
P_4	COLON NORMAL$^-$ COLON CARCINOMA$^+$	$Supp_{all} = 0.0133$ $Gr(P, E) = 1.050$	0.806	0.306

Des résultats préliminaires prometteurs.

Quid des Graphes Dynamiques Attribués ?

- Il est possible d'étendre l'approche pour prendre en compte la dynamique du graphe :

Cohesive Co-Evolution Patterns

Des ensembles de sommets **similaires** qui **évoluent** de la même façon.
Motif = triplet (N, T, P)

- N : les sommets similaires ;
- T : les pas de temps ;
- P les variations.

Application à l'érosion des sols

6 régions similaires (proches) où le rapport greenred diminue et l'indice de rouge augmente : **caractéristique d'un glissement de terrain**.

Conclusion

Un nouveau type d'analyse de graphes

- Visant à exploiter pleinement la structure du graphe mais aussi les attributs locaux.
- Des approches correctes et complètes visant à répondre à une *requête inductive*.

Conclusion

Un nouveau type d'analyse de graphes

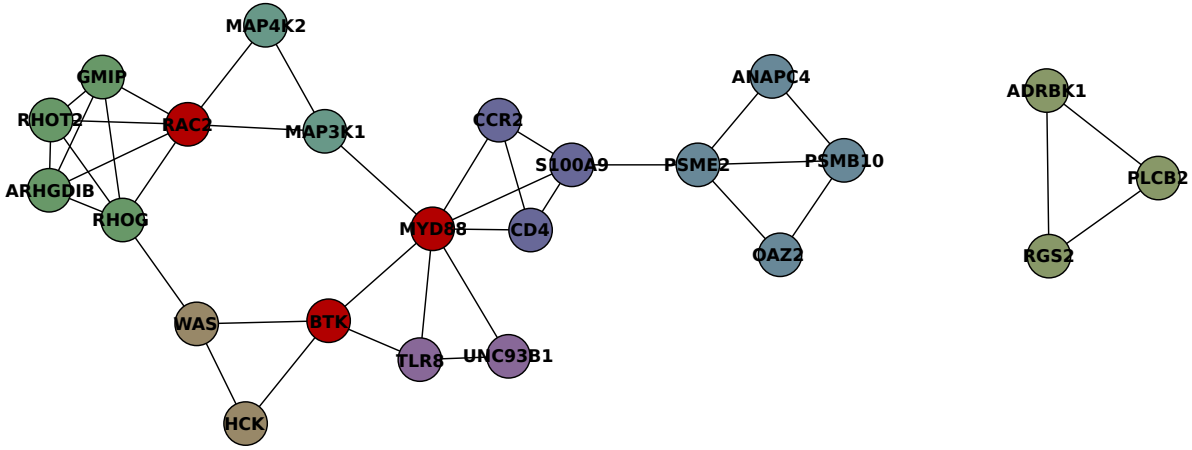
- Visant à exploiter pleinement la structure du graphe mais aussi les attributs locaux.
- Des approches correctes et complètes visant à répondre à une *requête inductive*.

Pourquoi avons nous besoin de pluridisciplinarité (vous) ?

- Besoin permanent de confronter nos propositions sur des données réelles.
- De nouveaux questionnements pour un expert induisent souvent de nouveaux problèmes théoriques en fouille de données.

Percolation de cliques homogènes

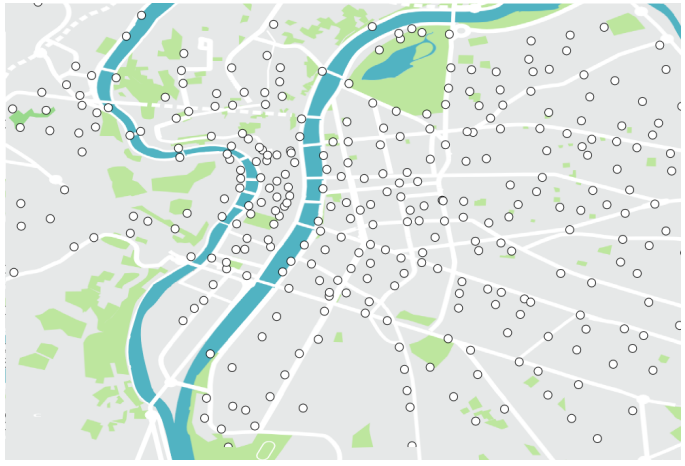
[P.N. Mougel, C. Rigotti, O. Gandrillon, PAKDD, 2012]



A CoHoP with $k = 3$, $\gamma = 4$, and $\alpha = 3$. All genes are over-expressed in 4 situations corresponding to normal white blood cell activities.

- Allows to get a larger picture when analysing protein modules
- Why protein modules are disconnected ?

Analyse du système VELO'V

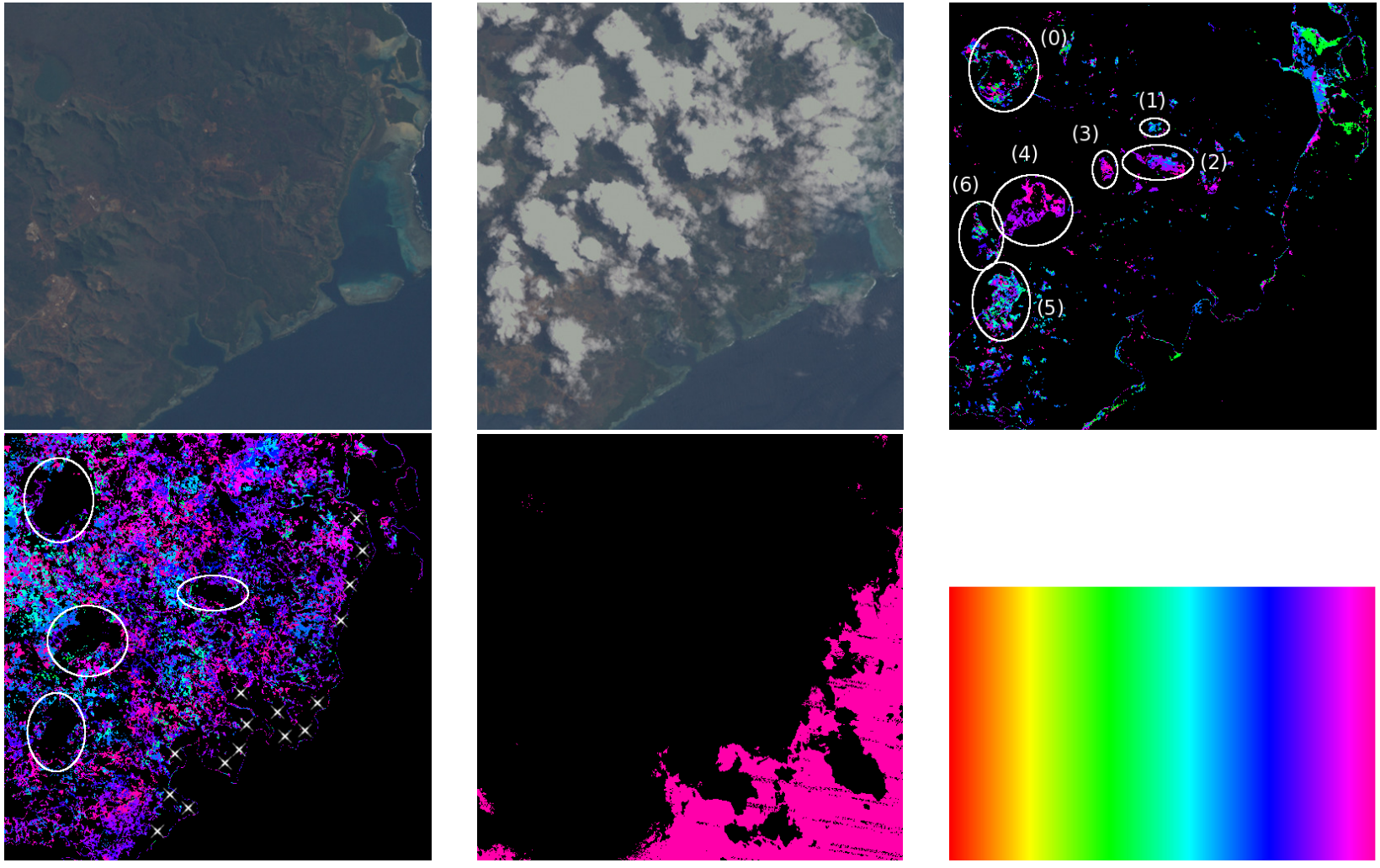


De nombreux travaux :

- (Nguyen et al., IDA Journal, 2012)
- (C. Robardet, IEEE ICDM, 2009)
- (P. Borgnat et al., Advances in Complex Systems, 2011)

Modèles d'Érosion des Sols

ANR Cosinus FOSTER

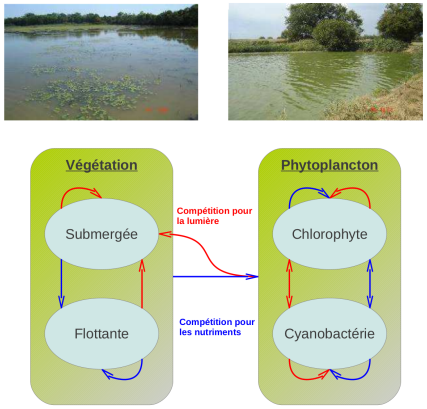


[A. Julea et al,IEEE T. Geoscience and Remote Sensing,2011]

Dynamique végétation/phytoplancton dans les lacs peu profonds

- États alternatifs stables
- Hystérésis
- Modèle dynamique multi-populations (voir poster Magali Vangkeosay)

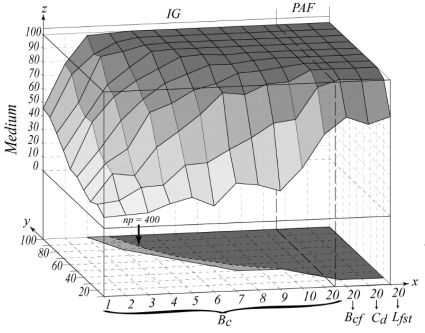
  ⇝Florent Arthaud (LEHNA)



Impact de la fragmentation paysagère sur la répartition de populations de batraciens

- Données populationnelles et génétiques
- Évaluation des impacts des trames vertes et bleues

  ⇝Jean-Paul Lena/Jérôme Prunier (LEHNA)



Collaborateurs “Informaticiens”

Jean-François Boulicaut, Loïc Cerf (UFMG, Brésil), Elise Desmier Serge Fenet, Pierre-Nicolas Mougel, Kim-Ngan Nguyen, Marc Plantevit, Adriana Prado, Christophe Rigotti, Céline Robardet, Julien Salotti, Magali Vangkeosay