Social mining : Extraction de motifs fréquents dans les réseaux

Erick Stattner

Laboratoire LAMIA
Université des Antilles et de la Guyane, France
erick.stattner@univ-ag.fr

Guadeloupe
Juin 2013

Introduction

Émergence de l'étude réseau:

- Aujourd'hui: émergence de l'étude des réseaux
  Ex. Études des réseaux d'amitiés, de collaboration, d'achats, de communications, d'échanges, ...
- Nait de l'observation que:
  liens sociaux = facteurs déterminants dans l'évolution de nombreux phénomènes
  - *Ex.* Diffusion, Achat, Tabagisme, Obésité, ...
- La "***Nouvelle science des réseaux***" [Newman2006]
  - Famille de méthodes qui s'intéresse aux interactions entre les objets

Introduction

Problématiques:

- Nouvelle façon d'aborder les phénomènes
- Soulève des questions en termes de
  - Modélisation
  - Collecte et stockage
  - **Analyse de données**

Contexte:

- Méthodes traditionnelles de data-mining difficilement applicables
- Nouvelles méthodes de data mining dédiées aux réseaux:
  (Node classification, Link-based Clustering, Link prediction, Search for frequent patterns, etc.)

Introduction

Dans cet exposé:

- **Fouille de réseaux** ou **social mining**
- Recherche de motifs
- Nouvelle approche: *Liens conceptuels fréquents (MFCL)*
  - Combine: **Structure du réseau** et **Attributs des noeuds**
  - Liens fréquents entre groupes de noeuds qui partagent des caractéristiques communes

Outline

Social mining

Introduction

Déroulement classique d'une étude de data-mining

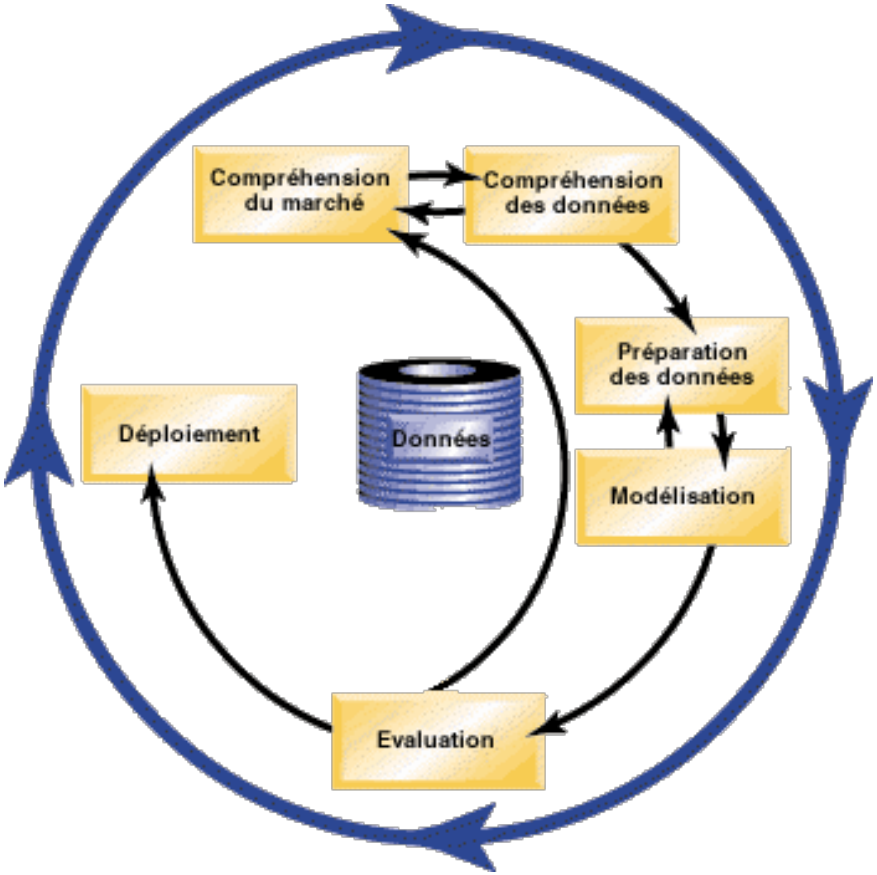


Figure: Approche classique

Extraction de motifs dans les réseaux

Introduction

Differents kinds of patterns from social networks

1. Clustering basé sur les liens
2. Clustering hybride
3. Recherche de sous-graphes fréquents

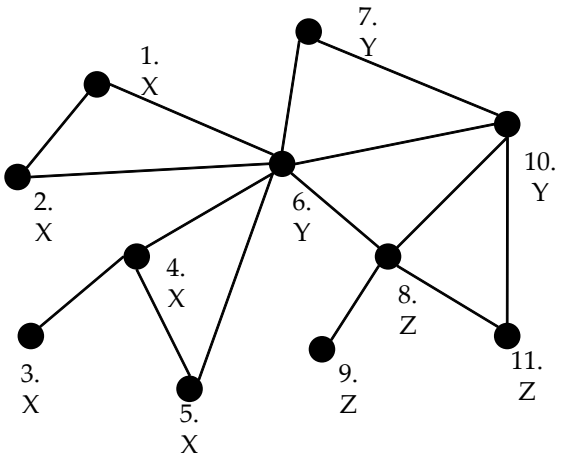


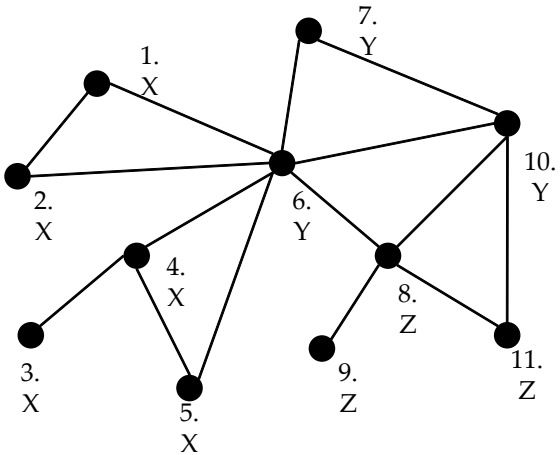
Figure: Reference network

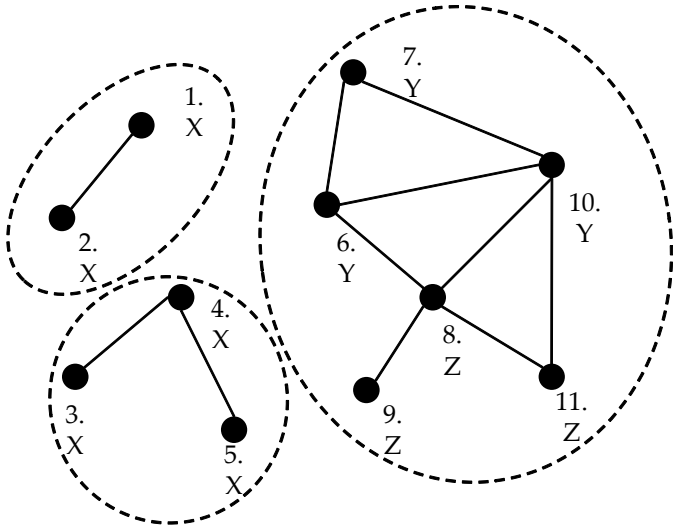
Extraction de motifs dans les réseaux

Clustering basé sur les liens

(1) Clustering basé sur les liens

- Extraction de communautés: groupes de noeuds fortement connectés
  - Algorithmes agrégatifs [Newman2003]
  - Algorithmes séparatifs [Fortunato2009]
  - Algorithmes basés sur des fonctions d'optimisation [Blondel2008]



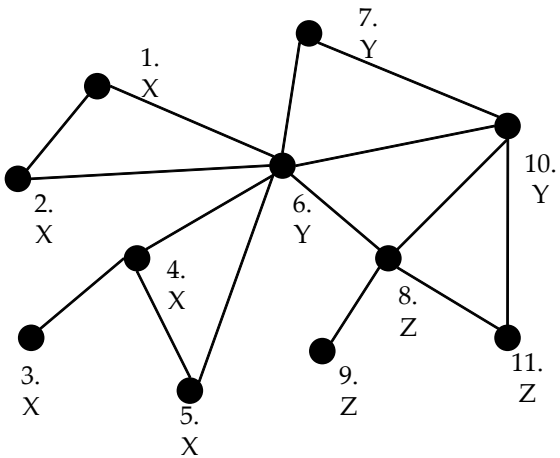


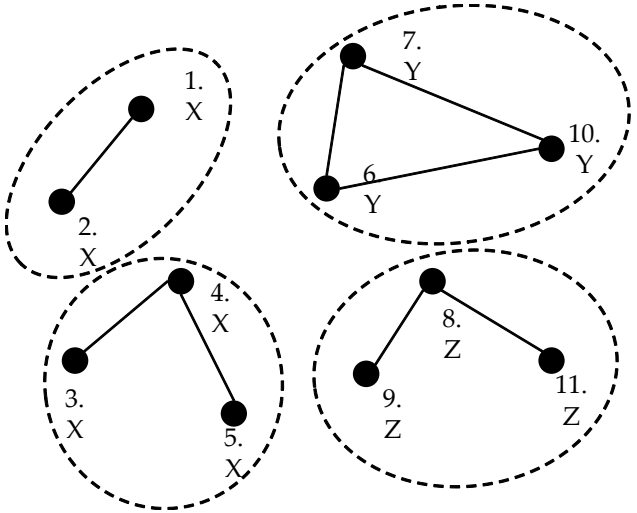
Extraction de motifs dans les réseaux

Clustering hybride

(2) Clustering hybride

- Extraction de communautés dans lesquelles les noeuds partagent des propriétés communes
  - Idem + prend en compte une similarité interne [Zhou2009]



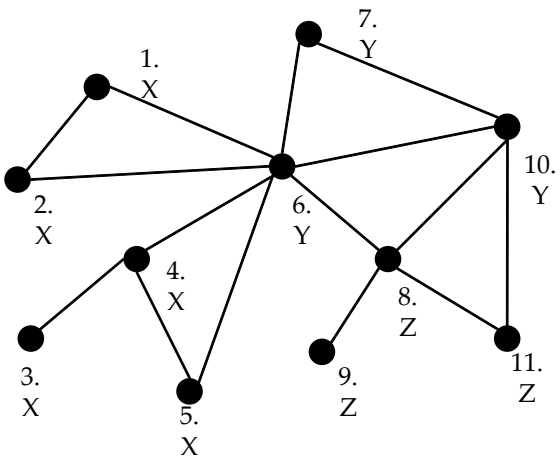


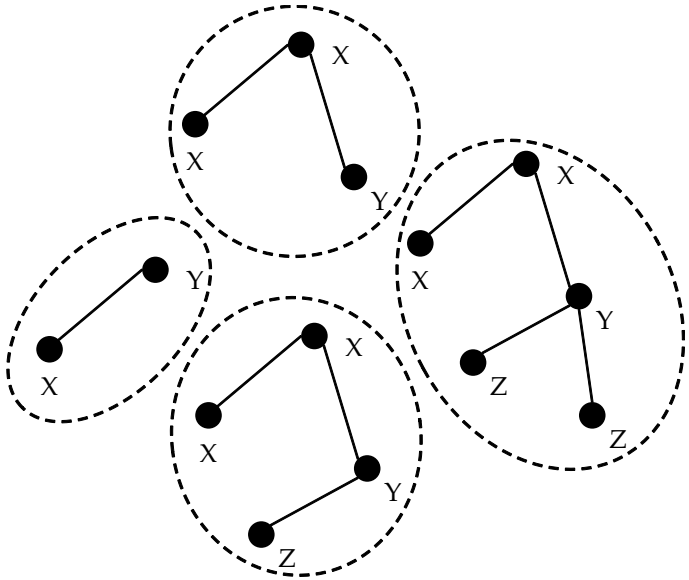
Extraction de motifs dans les réseaux

Sous-graphes fréquents

(3) Recherche de sous-graphes fréquents

- Extraction de sous-graphes retrouvés fréquemment
  - Algorithmes basés sur Apriori [Kuramochi2001]
  - Algorithmes basés sur la croissance de motifs [Nijssen2005]





Extraction de motifs dans les réseaux

Limites

Limites:

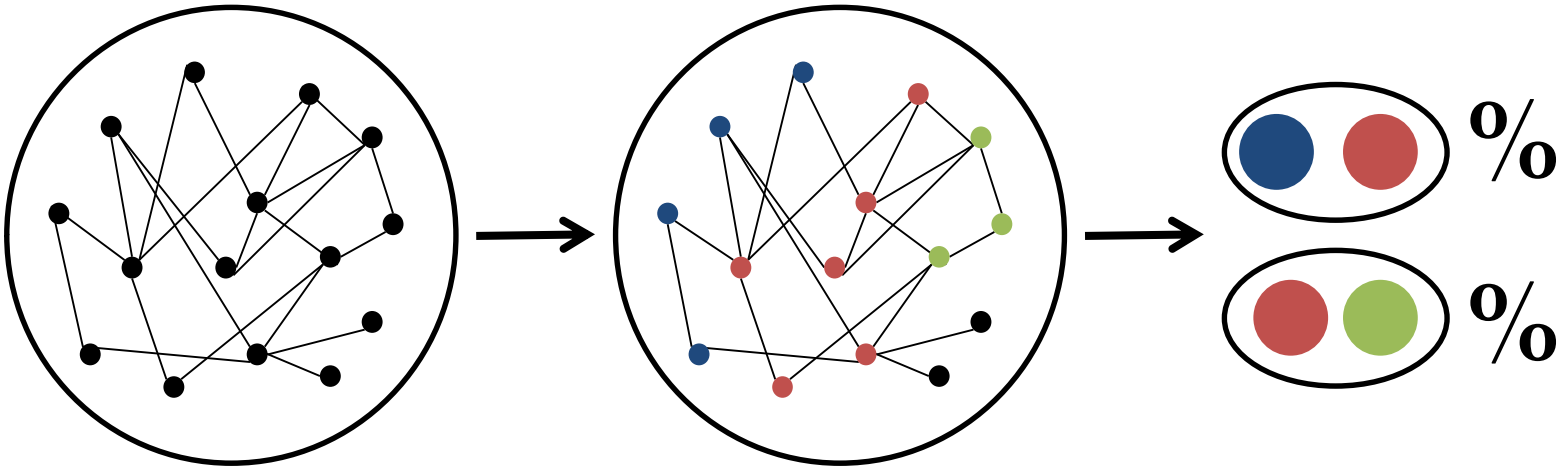
- Les motifs extraits ne permettent pas de répondre à des questions telles que:
  - Quels sont les groupes de noeuds les plus connectés?
  - Quelles sont les caractéristiques les plus fréquemment retrouvées en connexion?

Outline

Liens conceptuels
Introduction

Approche "liens conceptuels"

- Exploite structure et attributs
- Recherche des régularités dans les liens parmi des groupes de noeuds
- Groupe de noeuds (vérifiant certaines propriétés) fréquemment connecté à un autre groupe de noeuds



Liens conceptuels
Définition

Définition:

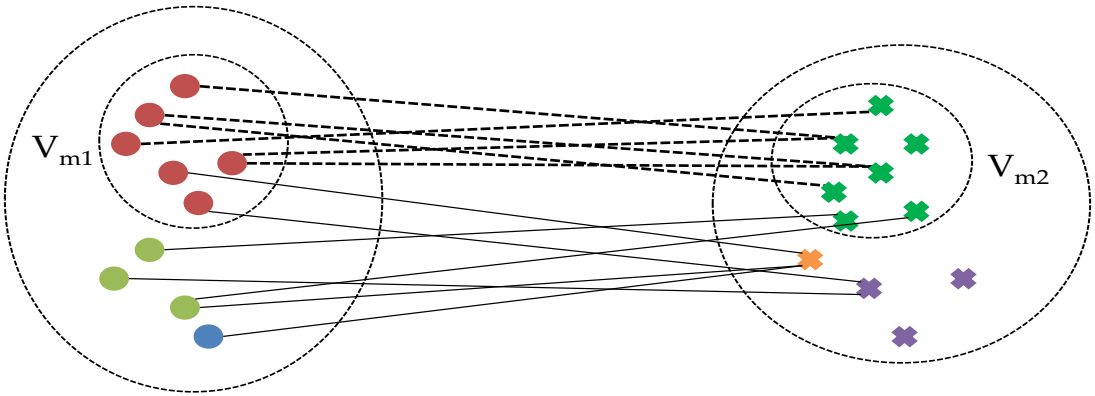
- $G = (V, E)$: Un réseau social
- V défini comme une relation $R(A_1, .., A_p)$ où A_i est un attribut
- Chaque noeud $v \in V$ est défini par un **itemset**
  $(A_1 = a_1$ et ... et $A_p = a_p)$ ou $(a_1, ..., a_p)$
- Soit m **itemset**
  On note V_m l'ensemble des noeuds vérifiant la propriété m

Liens conceptuels
Définition

Définition:

- Soient m_1 et m_2 **deux itemsets**
  (m_1, m_2): **Lien conceptuel**

$$(m_1, m_2) = \{e \in E; \quad e = (a, b) \quad a \in V_{m_1} \text{ et } b \in V_{m_2}\}$$



Lien entre deux concepts

$v_1 \in V_{m_1}, v_2 \in V_{m_1}$ et $v_3 \in V_{m_2}, v_4 \in V_{m_2}, v_5 \in V_{m_2}$

$$(m_1, m_2) \leftrightarrow (((\{v_1, v_2\}, m_1), (\{v_3, v_4, v_5\}, m_2))$$

Liens conceptuels
Définition

Définition:

- (m_1, m_2): **lien conceptuel**
  Support de (m_1, m_2): Pourcentage de liens de type (m_1, m_2)

$$support[(m_1, m_2)] = \frac{|\{e \in E; \quad e = (a, b) \quad a \in V_{m_1} \quad et \quad b \in V_{m_2}\}|}{|E|}$$

- β: **seuil de support des liens**
  (m_1, m_2) est un **lien conceptuel fréquent (FCL)** ssi

$$support[(m_1, m_2)] > \beta$$

Liens conceptuels
Définition

Définition:

- (m'_1, m'_2) est un **sur-lien conceptuel** de (m_1, m_2) ssi

$$m_1 \subseteq m'_1 \quad \text{et} \quad m_2 \subseteq m'_2$$

Ex. (ab, b) *sur-lien conceptuel de* (a, b)

- (m_1, m_2) est un **sous-lien conceptuel** de (m'_1, m'_2)

- (m_1, m_2) **Lien conceptuel fréquent maximal (MFCL)** ssi
$\nexists$ pas de sur-lien conceptuel (m'_1, m'_2) de (m_1, m_2) qui soit fréquent

Liens conceptuels
Définition

Liens conceptuels fréquents maximaux (MFCL):

- Fournissent une connaissance sur les groupes de noeuds les plus connectés au sein du réseau
- $\Rightarrow$ Sur les caractéristiques les plus souvent connectés

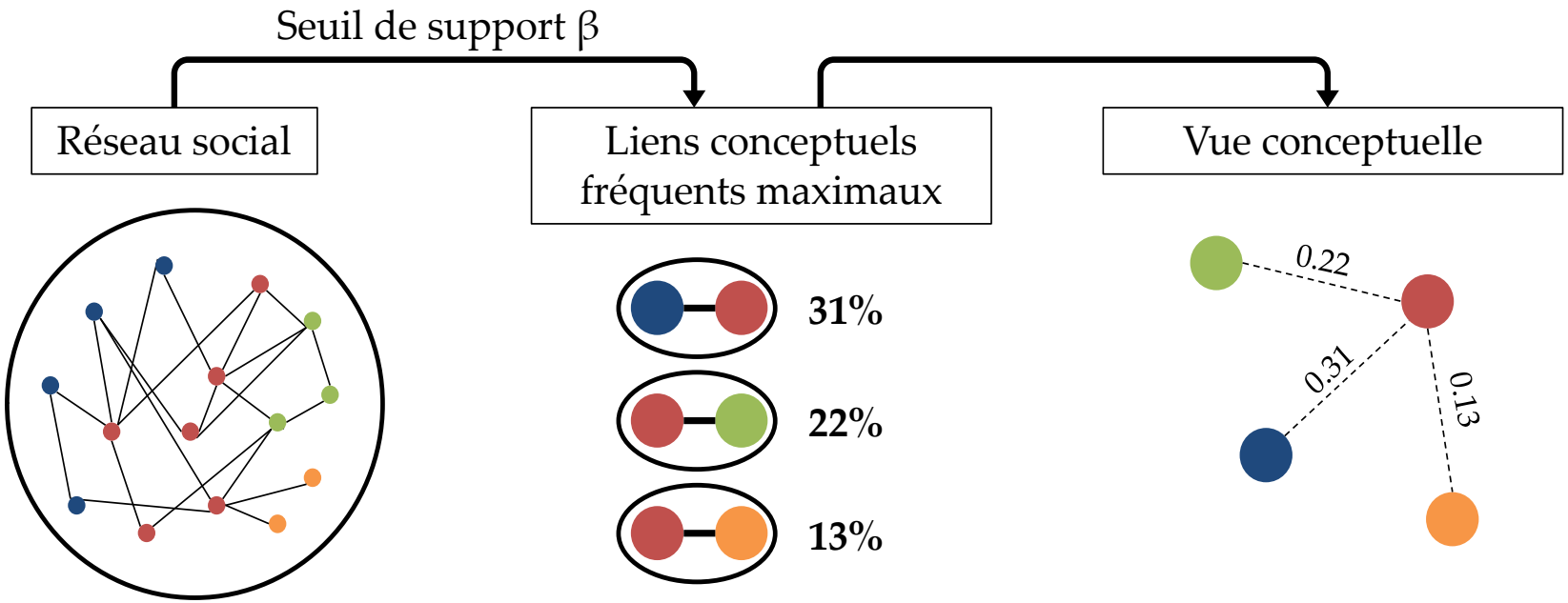
Example:

- **Bipartite purchase network:**
  m_1 = Sex='M' **and** Interest='computer science'
  m_2 = Genre='Science Fiction' **and** Product='book'
  $supp[(m_1, m_2)] = 14\%$

Liens conceptuels
Vue conceptuelle

Vue conceptuelle:

- Connaissance sur les groupes de noeuds les plus connectés
- Fournissent un "**vue conceptuelle**"



Outline

Méthodes d'extraction
Complexité

Complexité:

- Approche naïve
  1. Extraire tous les itemsets à partir de V
  2. Recherche tous les liens conceptuels fréquents
  3. Extraire les maximaux
- Extrêmement couteux si l'espace de recherche est important
- $2^{|R|} \times 2^{|R|} \times |E|$ comparaisons nécessaires

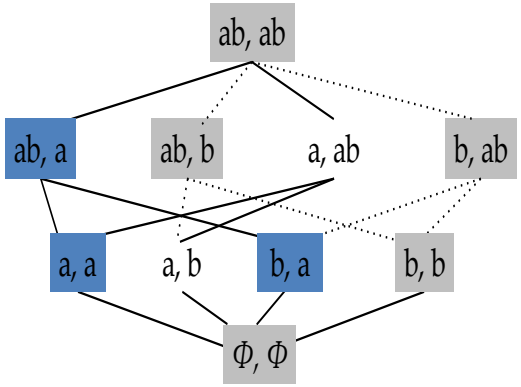
Objectifs:

- Optimiser le processus d'extraction
- Réduire l'espace de recherche

Méthodes d'extraction
Algorithme MFCL-Min

Algorithme MFCL-Min (Maximal Frequent Conceptual Link Mining)

- Recherche ascendante



- Réduit l'espace de recherche selon:
  1. **Propriété de fermeture:**
     Si (m_1, m_2) est non-fréquent, tous ses sur-liens sont non-fréquents
  2. **Propriété de fréquence:**
     Si (m_1, m_2) fréquent, alors $|V_{m_1}| \times |V_{m_2}| \geq \beta \times |E|$, puisque $|V_{m_1}| \times |V_{m_2}| \geq |(m_1, m_2)|$

Méthodes d'extraction
Algorithme MFCL-Min

Testbed:

- Réseau de contacts de proximité
  - 3000 Noeuds
  - 7000 Liens
- Chaque noeud:
  1. classe d'age, i.e. $\lfloor \frac{age}{10} \rfloor$,
  2. sexe,
  3. statut professionnel,
  4. type de relation avec le chef de famille,
  5. classe de contacts, i.e. $\lfloor \frac{degre}{2} \rfloor$
  6. appartenance à une communauté

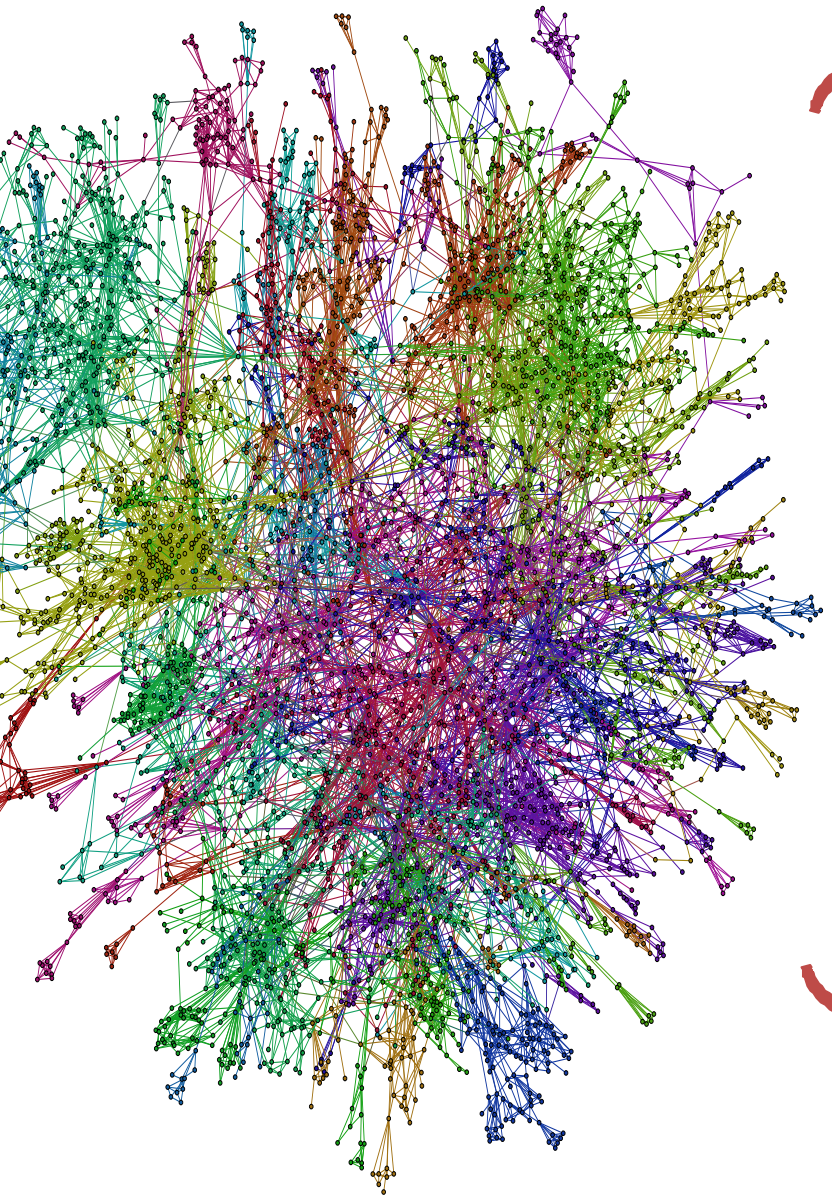
Méthodes d'extraction
Algorithme MFCL-Min

Exemples de motifs extraits:

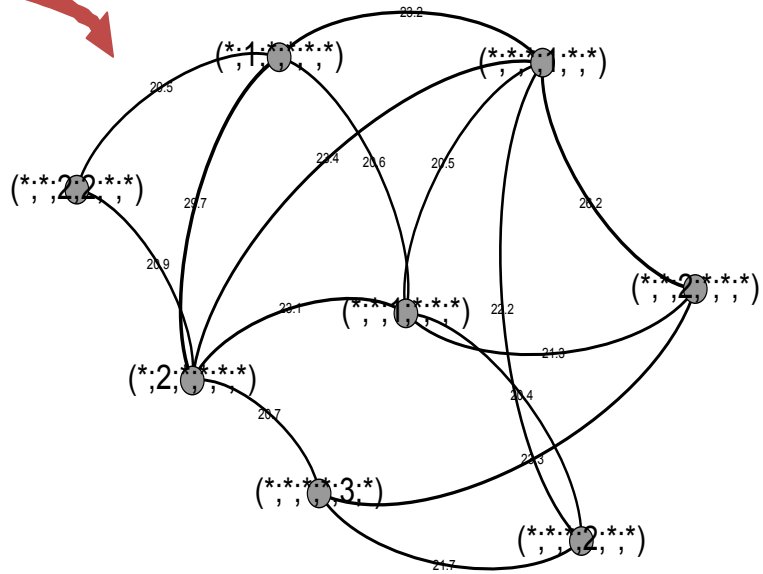
- $\beta = 0.1$

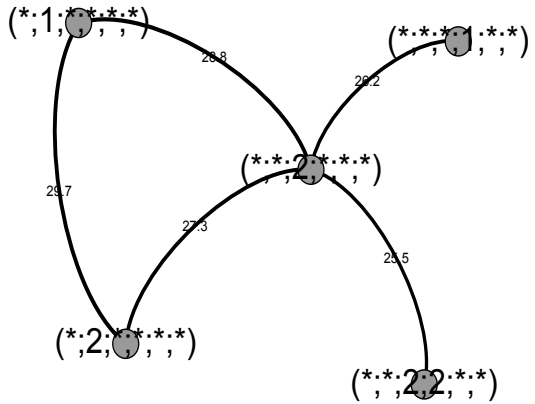
MFCL	Support
$((4;1;1;*;*;*),(*;2;2;*;*))$	0.117
$((2;*;2;*;*;*),(*;*;2;*;*))$	0.113
$((4;*;*;1;*;*),(*;1;*;*;*))$	0.149

- "11.7% des liens du réseau connectent des hommes de 40 ans qui ont un emploi à des femmes qui n'en ont pas"



β=0.20



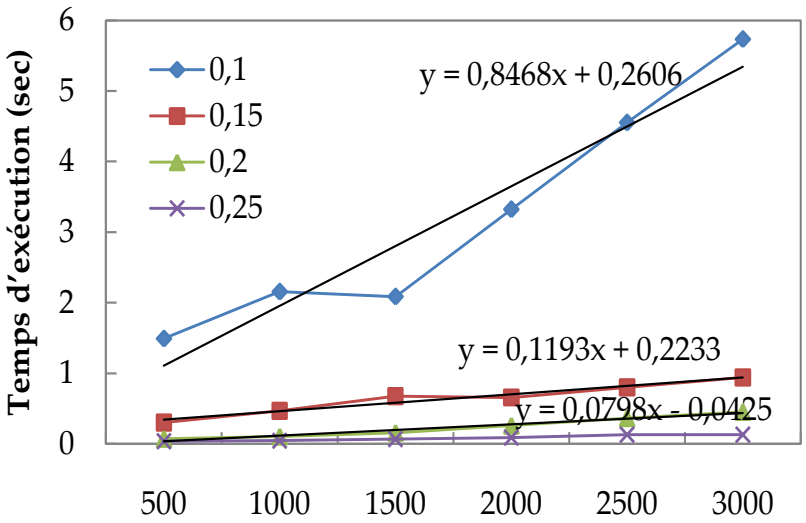


β=0.25

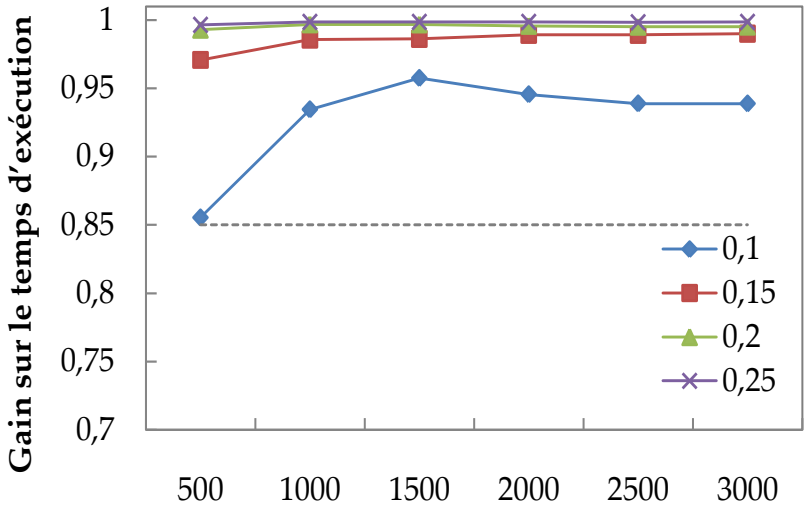
Méthodes d'extraction
Algorithme MFCL-Min

Performances:

- Temps d'exécution et Gain comparé à l'approche naïve selon la **taille du réseau**
  - **Croit linéairement avec** $|V|$
  - **Bonnes performances** $> 85\%$



(1)

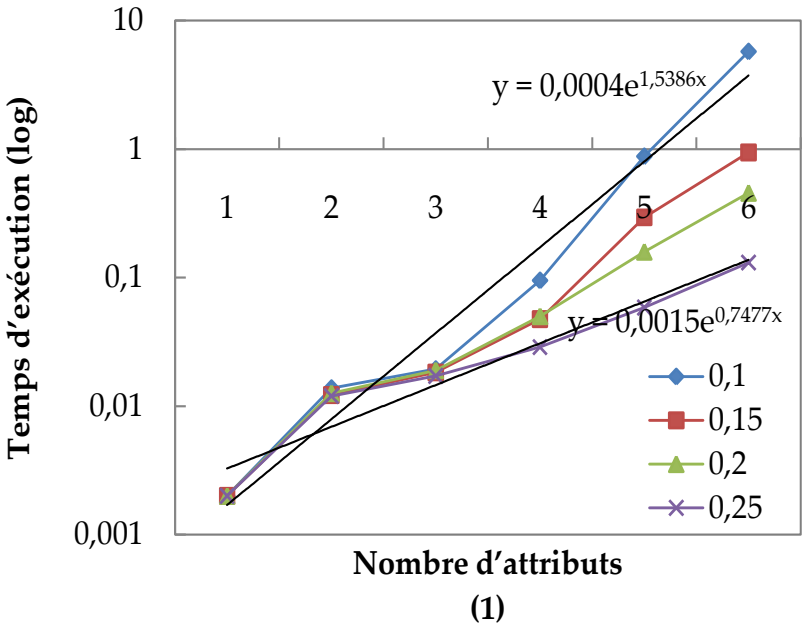


(2)

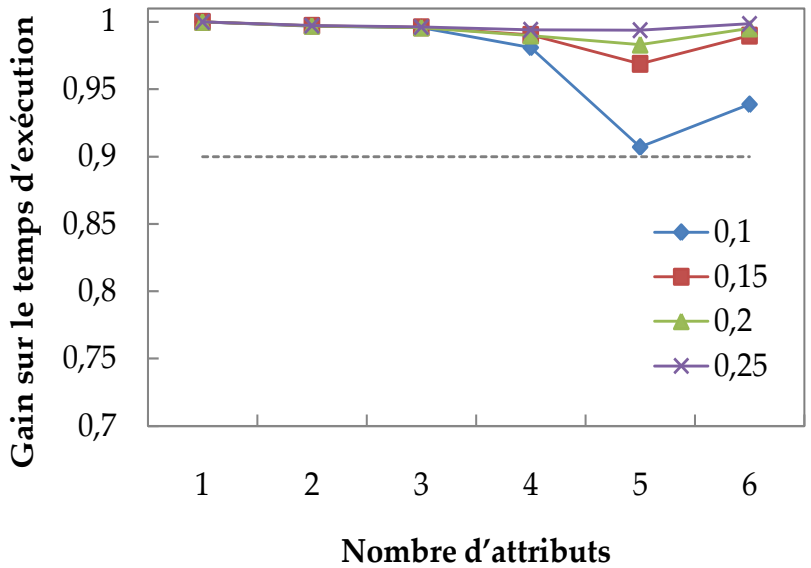
Méthodes d'extraction
Algorithme MFCL-Min

Performances:

- Temps d'exécution et Gain comparé à l'approche naïve selon le **nombre d'attributs**
  - **Croit exponentiellement avec** $|R|$
  - **Bonnes performances** $> 90\%$



(1)



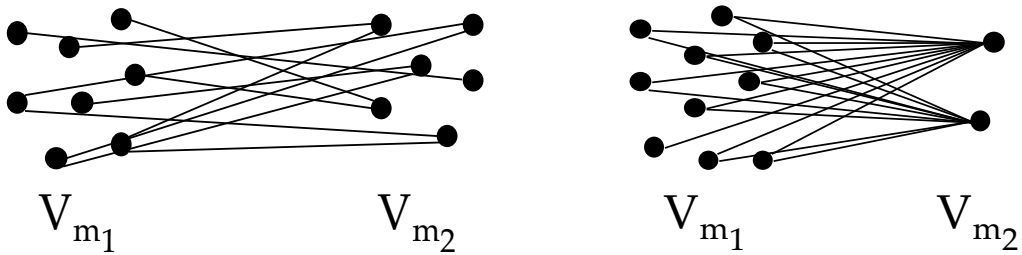
(2)

Méthodes d'extraction

Vers une méthode d'extraction hybride

Property:

- (m_1, m_2) MFCL $\Rightarrow |V_{m_1}| \times |V_{m_2}| \geq \beta \times |E|$
- MFCL peuvent etre identifiés dans 2 types de configurations:
  - (a) Soit m_1 et m_2 sont fréquents
  - (b) Au moins l'un des deux est fréquent



(a) (b)

Hypothèse:

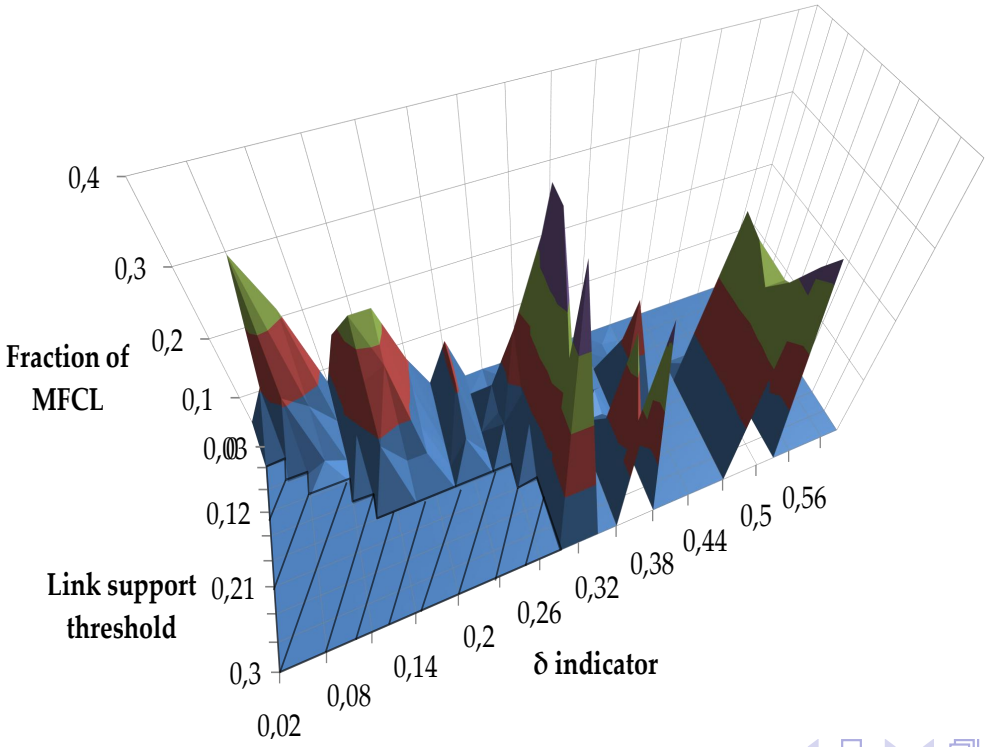
- La situation (b) survient rarement
- MFCL peuvent etre identifiés entre des groupes de noeuds fréquents

Méthodes d'extraction

Vers une méthode d'extraction hybride

Étude de l'espace des solutions

- A partir de quel seuil de fréquence un itemset intervient dans un MFCL?
- Soit δ fréquence de l'itemset, i.e. $\delta = \frac{|V_m|}{|V|}$
- Distribution de δ pour différents seuils de support des liens β



Méthodes d'extraction

Vers une méthode d'extraction hybride

Solution proposed:

- H-MFCLMin Algorithm
  (Hybrid Maximal Frequent Conceptual Link Mining)
- Extends the MFCL-Min algorithm [Stattner-ASONAM'12]
- Introduce α **Itemset support threshold**

Repose sur 3 hypothèses:

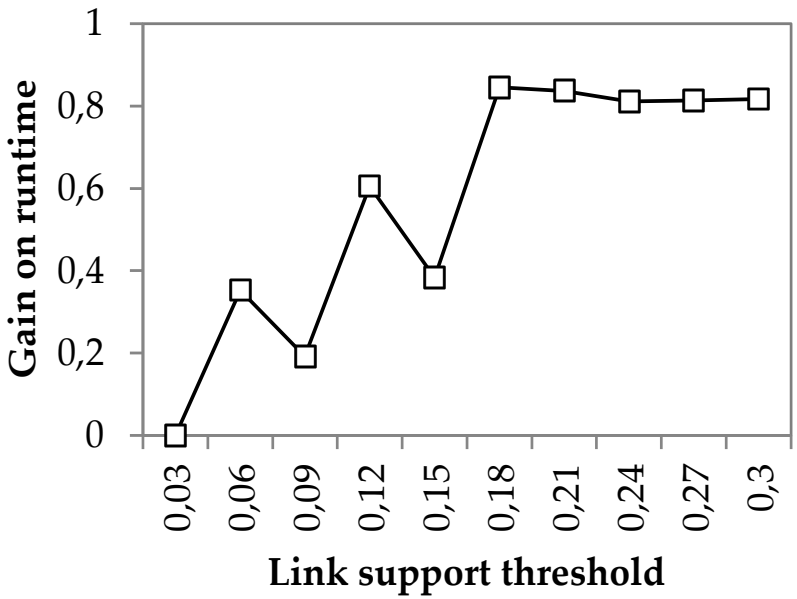
1. Propriété de fermeture:
   Si (m_1, m_2) est non-fréquent, tous ses sur-liens sont non-fréquents
2. Propriété de fréquence: $|V_{m_1}| \times |V_{m_2}| \geq \beta \times |E|$
3. Notre hypothèse: $|V_m| \geq \alpha$

Méthodes d'extraction

Vers une méthode d'extraction hybride

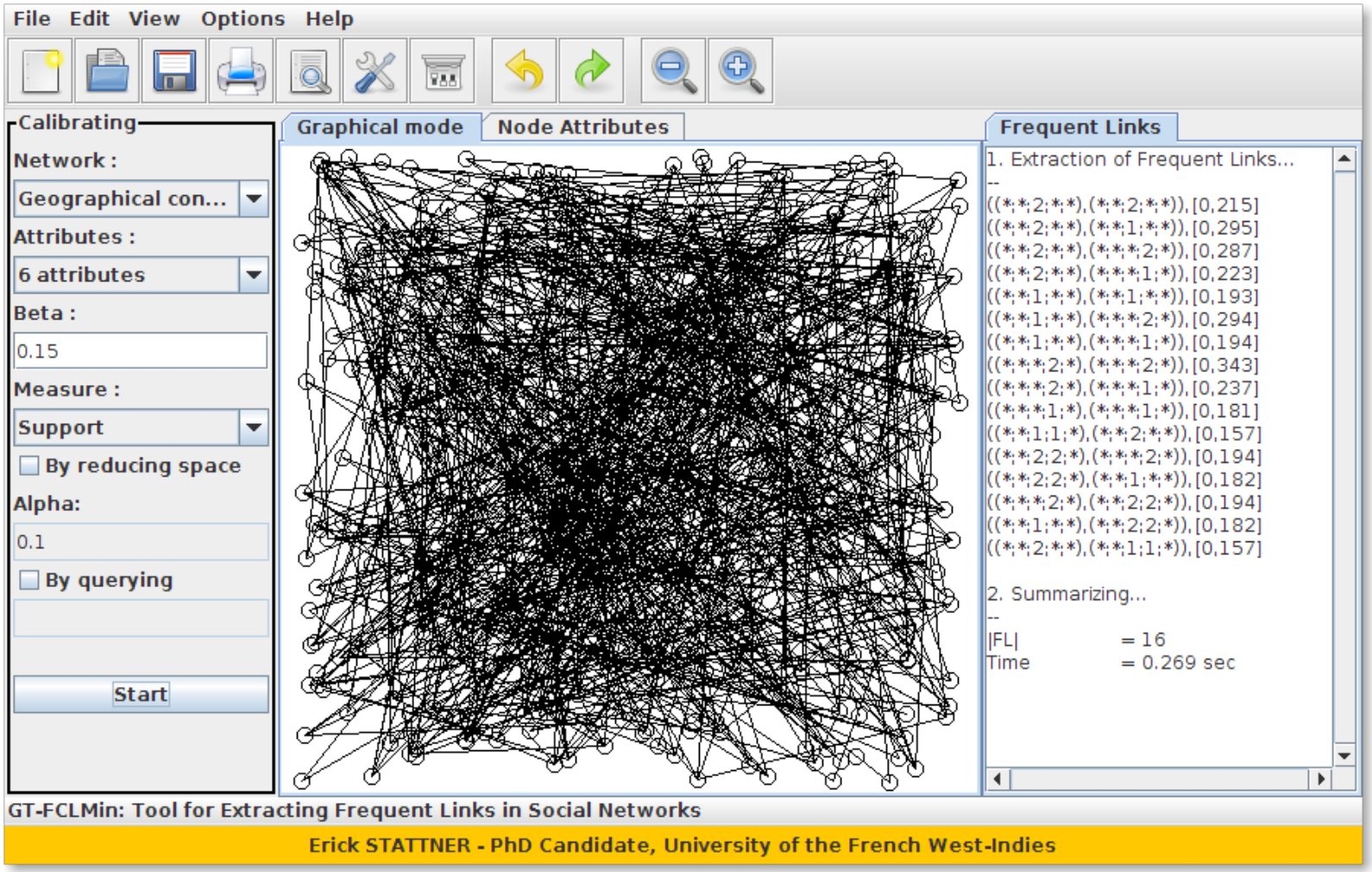
Gain comparé a notre premier algo. MFCL-Min

- Gain croit avec β
- Bonnes performances pour β élevé



Méthodes d'extraction
Outil GT-FCLMin

Outil GT-FCLMin



Outline

Conclusion

Conclusion

- Nouvelle approche pour l'extraction de motifs fréquents
- Exploite structure et attributs
- Double intérêt
  - Extrait des informations pertinents des réseaux
  - Fournit une représentation synthétique des réseaux
- Deux algorithmes d'extraction
- Outil graphique GT-FCLMin

Perspectives

- Optimiser le processus d'extraction *[IJISMD'13]*
- Prise en compte de la mixité des connaissances *[ASONAM'13]*
- Nouveaux modèles de génération

Conclusion

Nos travaux sur les réseaux:

- **4 Revues internationales:**
  IJISMD'12, ProcediaCS'12, IJISMD'13, CHB'13
- **18 Conférences internationales:**
  PE-WASUN'09, RCIS'10, MoMM'10, WOWMOM'10, LCN'11, LCN'12, RCIS'11, DEXA'11, WSKS'11, ASONAM'12, ASONAM'12, NDT'12, RCIS'12, ADBIS'12, DEXA'12, RCIS'13, ASONAM'13, ASONAM'13
- **3 Conférences nationales:**
  MARAMI'12, EGC'12, EGC'12
- **4 Outils développés:**
  DynSpread, ER-Net, GT-FCLMin, Lypus

Conclusion

Merci de votre attention !

Questions?