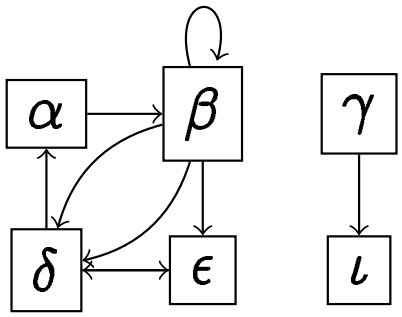
TELECOM ParisTech

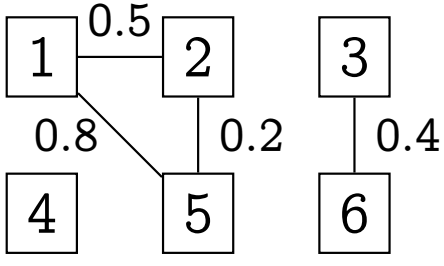
Fouille de graphes et détection d'évènements

Pierre Senellart

2 octobre 2015, *Journée de la Chaire Big Data & Market Insights*

Données graphes





- Un graphe c'est :
  - un ensemble de nœuds (ou sommets)
  - un ensemble d'arêtes (ou liens), chacune reliant deux de ces nœuds
- Variantes :
  - Les arêtes peuvent être orientées ou non orientées
  - Les nœuds ou arêtes peuvent être étiquetées par un identifiant, un nom, une chaîne de caractères
  - Les nœuds ou arêtes peuvent être pondérés par un entier, un nombre décimal
  - Les arêtes multiples entre deux mêmes nœuds ou non

Données modélisées par des graphes

	nœud	arête
réseau social	individu	connexion, amitié, suiveur...
Internet	ordinateur, routeur	connexion filiaire ou sans fil
Web	page Web	hyperlien
Web sémantique	concept, valeur	fait
réseau ferroviaire	gare	connexion
réseau routier	intersection	segment de route
transactions	compte bancaire	transfert
métabolisme	protéine	interaction métabolique
cerveau	neurone	connexion

Défis des données graphes

Les mêmes que les Big Data en général :

Volume : très grand nombre de nœuds ou d'arêtes rendant nécessaire l'utilisation d'algorithmes très efficaces (linéaires ou quasi-linéaires en la taille du graphe), ou la distribution

Vélocité : certains graphes évoluent très rapidement, des nœuds ou arêtes apparaissant ou disparaissant continûment

Variété : hétérogénéité des liens entre nœuds, différences de structure d'un graphe à un autre

Véracité : incertitude sur les pondérations ou les annotations d'un graphe

Types de problèmes à résoudre

- requêtes de chemin ou de distance
  Quel est le chemin le plus court de Pau à Grenoble dans le graphe de la SNCF ?
- identification de communauté ou de nœuds centraux
  Quels sont les individus influents sur Twitter dans le domaine des cosmétiques ?
- fiabilité
  Quelle est la robustesse du sous-réseau Internet du gouvernement face à des attaques par déni de service ?
- recherche de motifs intéressants
  Quels comptes bancaires reçoivent-ils des transferts dont les caractéristiques les rendent remarquables (et suspects) ?
- etc.

Dans cet exposé

- Vue d'ensemble de trois travaux de recherche conduits à Télécom ParisTech :
  - Découverte de motifs dans les graphes hétérogènes
    (avec C. Meng, R. Cheng, S. Maniu, U. Hong Kong ; WWW 2015)
  - Maximisation d'influence en ligne
    (avec S. Lei, S. Maniu, L. Mo, R. Cheng, U. Hong Kong ; KDD 2015)
  - Requêtes efficaces dans des graphes incertains
    (avec M. Monet)

- Zoom sur un quatrième travail d'une doctorante de la chaire :
  - Détection d'événèments dans les graphes
    (O. Balalau ; WSDM 2015)

Plan

Introduction

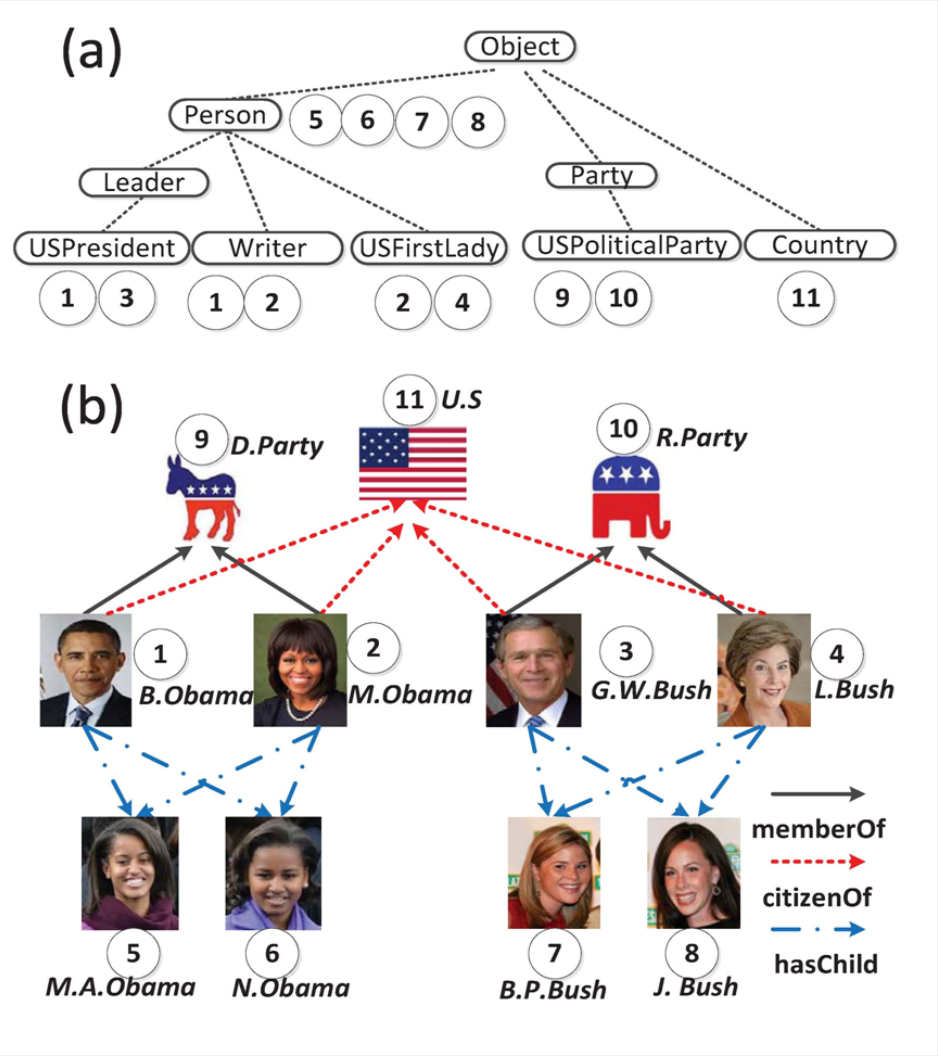
Motifs dans les graphes hétérogènes

Maximisation d'influence en ligne

Requêtes dans les graphes incertains

Détection d'événements

Problème



Trouver des paires de nœuds de ce graphe qui sont similaires à la paire (« B. Obama », « M. Obama »).

Approche

- Trouver des méta-chemins de longueur non bornée dans un métagraphe de types qui relient la paire exemple pairs :

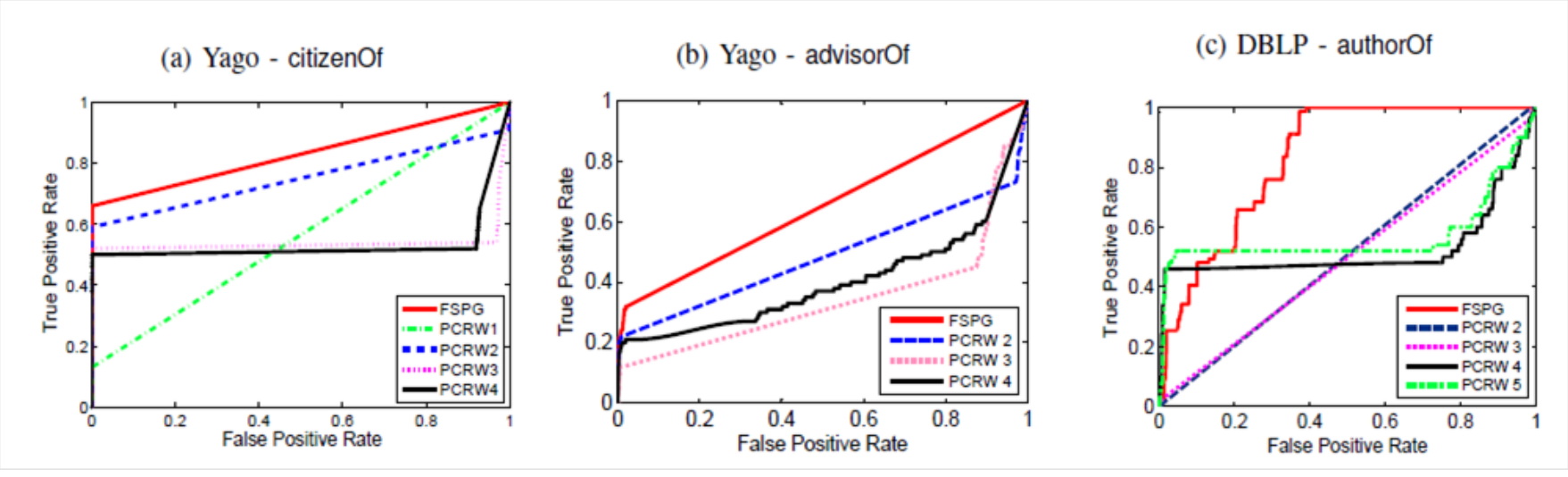
$$\text{Person} \xrightarrow{\text{marriedTo}} \text{Person}$$

$$\text{Person} \xrightarrow{\text{graduateOf}} \text{University} \xrightarrow{\text{graduateOf}^{-1}} \text{Person}$$

- Énumérer efficacement les méta-chemins dans un ordre glouton, les plus prometteurs d'abord, en s'appuyant sur une structure d'index
- Raffiner avec les types de nœuds

Résultats

- Utilisé pour la prédiction de liens dans divers réseaux hétérogènes



- Fonctionne mieux que les approches classiques basées sur le fait de borner la longueur d'un chemin
- Reste efficace

Plan

Introduction

Motifs dans les graphes hétérogènes

Maximisation d'influence en ligne

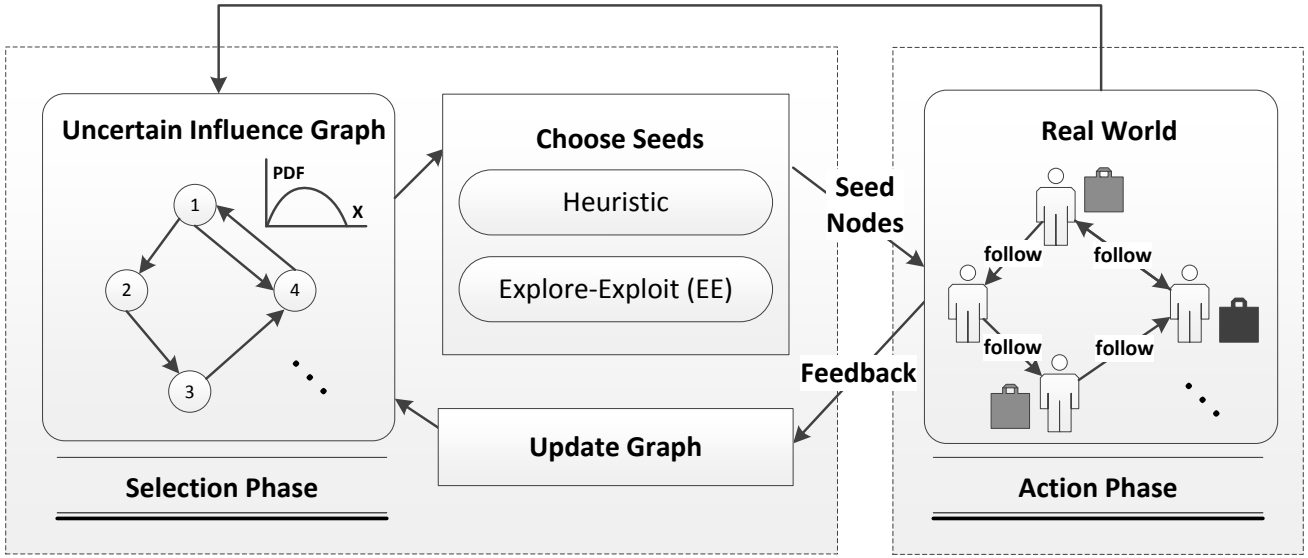
Requêtes dans les graphes incertains

Détection d'événements

Problème

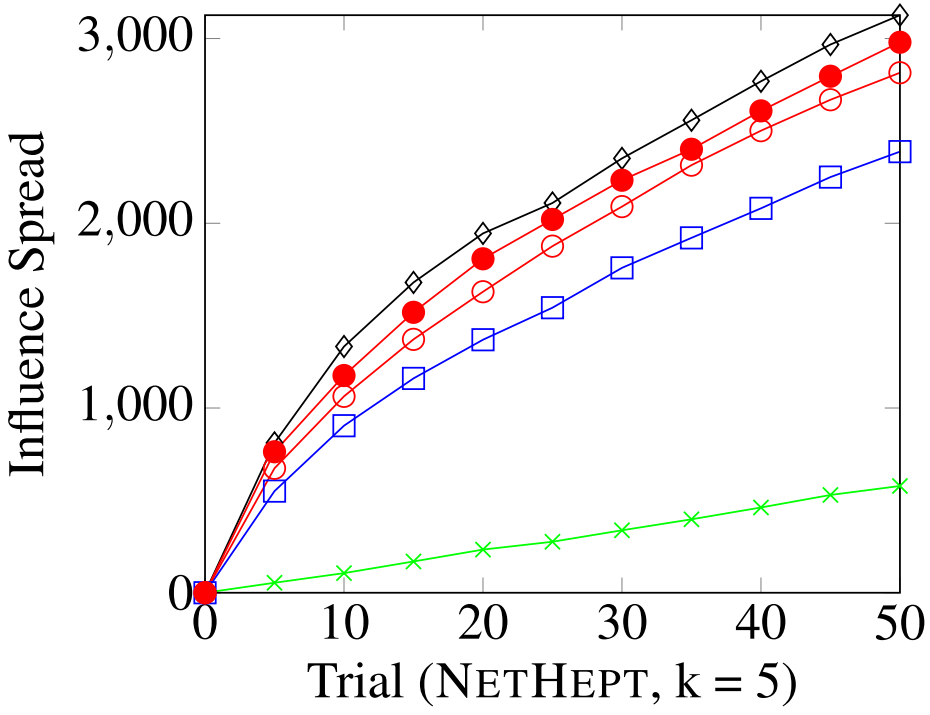
- Trouver les utilisateurs les plus influents dans un réseau social, en lançant des campagnes de marketing et en en observant le résultat
- Le but est d'avoir touché le plus grand nombre d'individu
- On ne connaît pas la probabilité qu'un utilisateur va influencer un autre

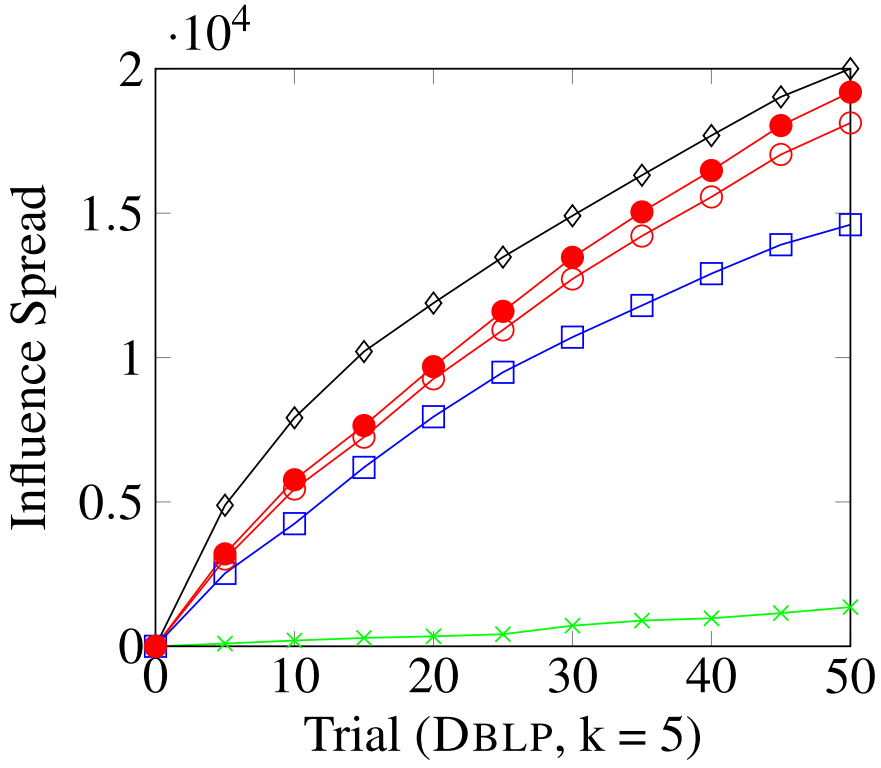
Approche

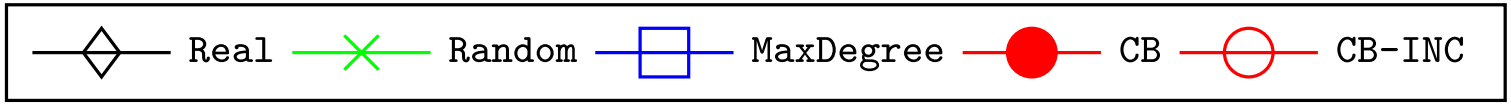


- Maintenir une connaissance partielle du monde sous la forme d'un graphe probabiliste
- Mettre à jour cette connaissance du monde en observant le résultat de la campagne de marketing
- Décider de la prochaine campagne en explorant le monde ou en exploitant la connaissance partielle du monde

Résultats


Influence Spread
3,000
2,000
1,000
0
0
10
20
30
40
50
Trial (NETHEPT, k = 5)


·10^4
Influence Spread
2
1.5
1
0.5
0
0
10
20
30
40
50
Trial (DBLP, k = 5)


Real
Random
MaxDegree
CB
CB-INC

Plan

Introduction

Motifs dans les graphes hétérogènes

Maximisation d'influence en ligne

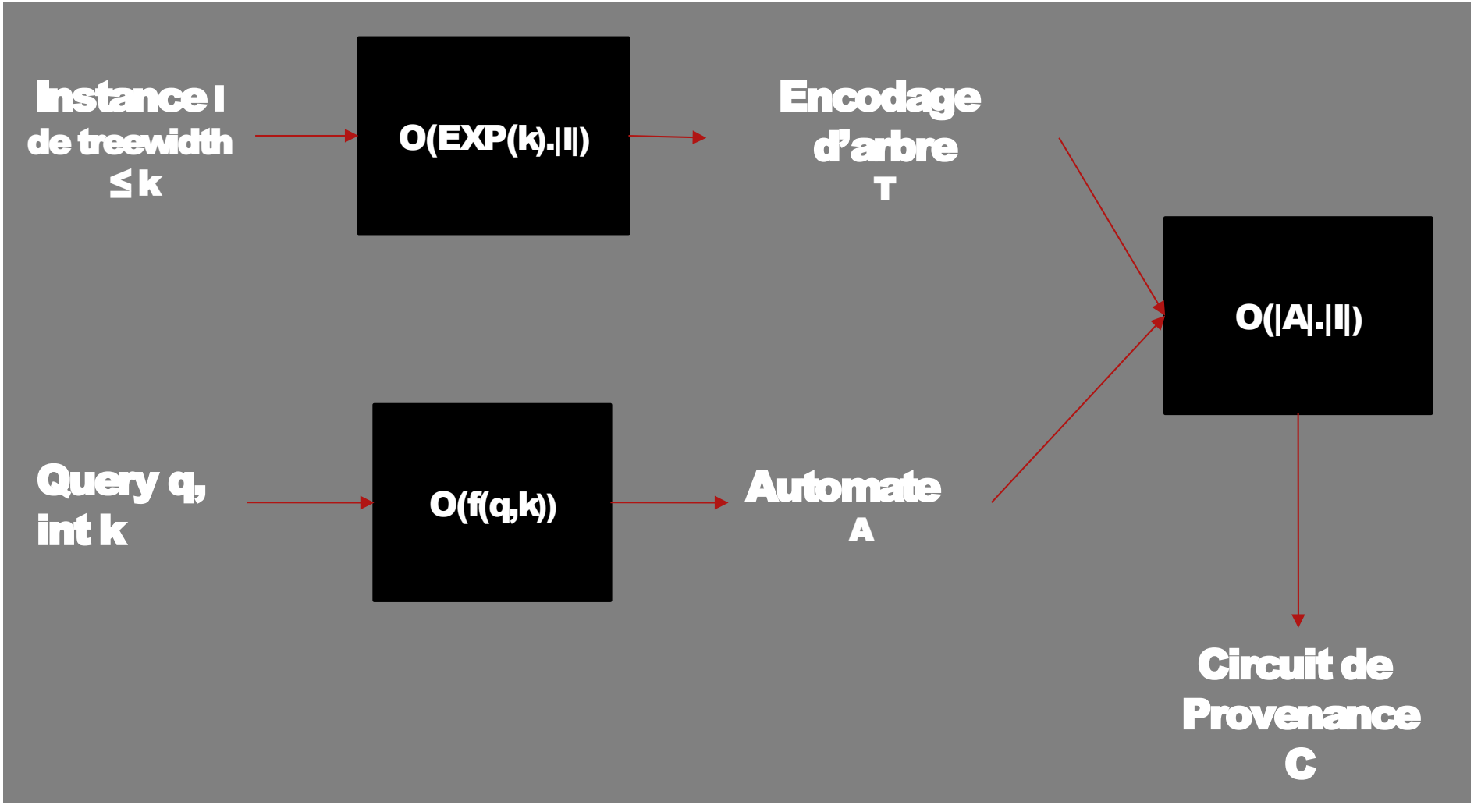
Requêtes dans les graphes incertains

Détection d'événements

Problème

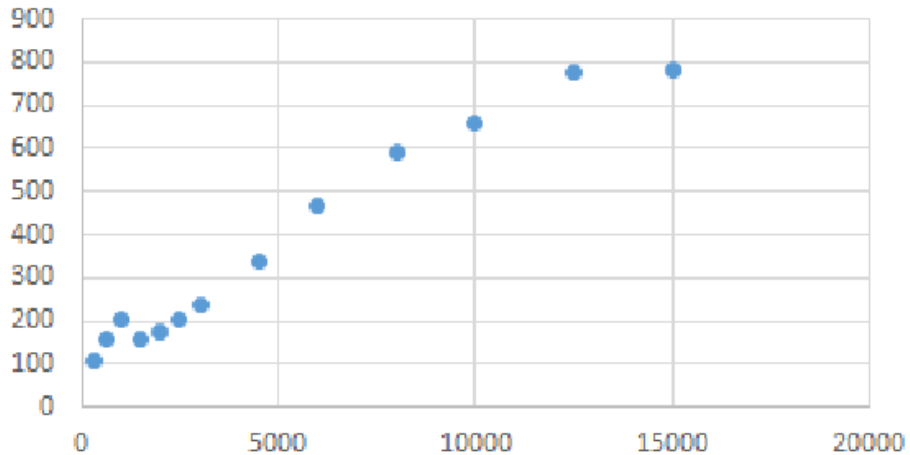
- On se donne un graphe avec des probabilités sur les arêtes (probabilité qu'une connexion Internet soit fonctionnelle, distribution de probabilité sur les temps de trajet dans un réseau de transport, etc.)
- On pose une requête sur ce graphe probabiliste (quelle est la probabilité que ce graphe soit connexe ? quelle est la probabilité que j'arrive à mon domicile en moins d'une heure ?)
- Même pour une modélisation très simple et des requêtes très simple, ce problème est #P-difficile : aucun espoir de le résoudre en temps raisonnable
- Mais les graphes du monde réel ne sont pas arbitraires, certains ont une faible largeur d'arbre et peuvent êtres « décomposés en arbres »

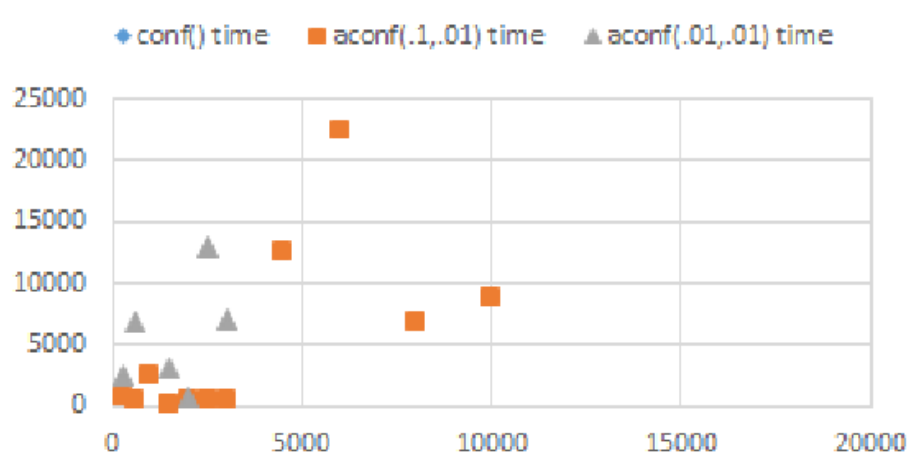
Approche


Instance I
de treewidth
≤ k
O(EXP(k).|I|)
Encodage
d'arbre
T
O(|A|.|I|)
Query q,
int k
O(f(q,k))
Automate
A
Circuit de
Provenance
C

Résultats

Significativement plus rapide que MayBMS, un système de gestion de données probabilistes, pour certaines requêtes (celles qui ne se prêtent pas à des optimisations) et certains jeux de données (ceux avec faible largeur d'arbre)





Plan

Introduction

Motifs dans les graphes hétérogènes

Maximisation d'influence en ligne

Requêtes dans les graphes incertains

Détection d'événements