Identification de sites Web logiques

Pierre Senellart

Qu'est-ce qu'un site Web logique ?

- Idée simple : site Web = serveur Web.
- Mais :
  - Certains sites Web couvrent plusieurs serveurs (ex. `www-rocq.inria.fr`, `osage.inria.fr`...).
  - Certains serveurs Web contiennent des sites différents (ex. `perso.wanadoo.fr`).
  - Qu'est-ce qu'un serveur Web en présence d'hôtes virtuels, de miroirs ?
  - Certains sites Web peuvent être décomposés en différents sous-sites (ex. le site Web de l'INRIA).

Plan de l'exposé

1. Introduction

2. Propagation de flots
   - ✓ Flot maximal / Coupe minimale
   - ✓ Algorithme de préflots
   - ✓ Adaptation au Web

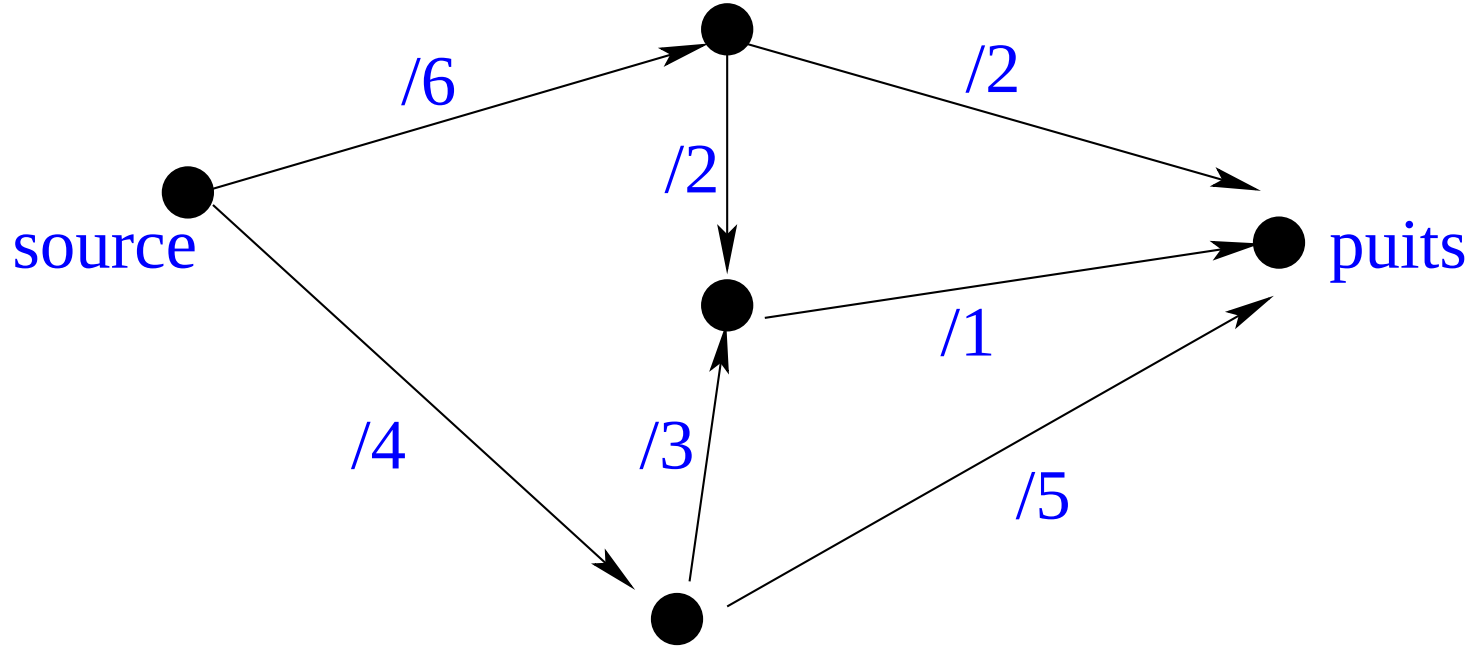
3. Extension de la graine
   - ✓ Markov CLustering algorithm
   - ✓ Propagation de flots à partir de MCL

4. Site de GEMO
   - ✓ Expériences
   - ✓ Résultats

5. Conclusion

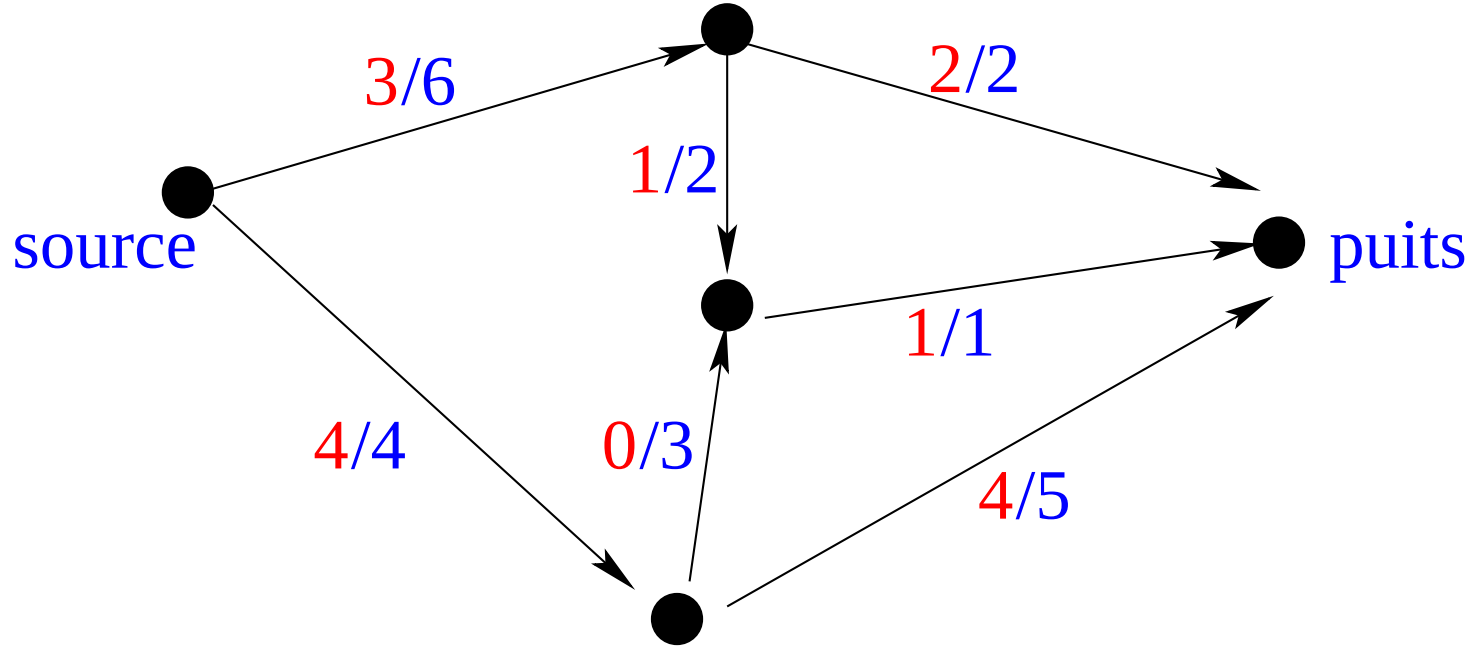
Flot maximal / Coupe minimale

- Réseau de transport.
- Flot maximal ≡ Coupe minimale.



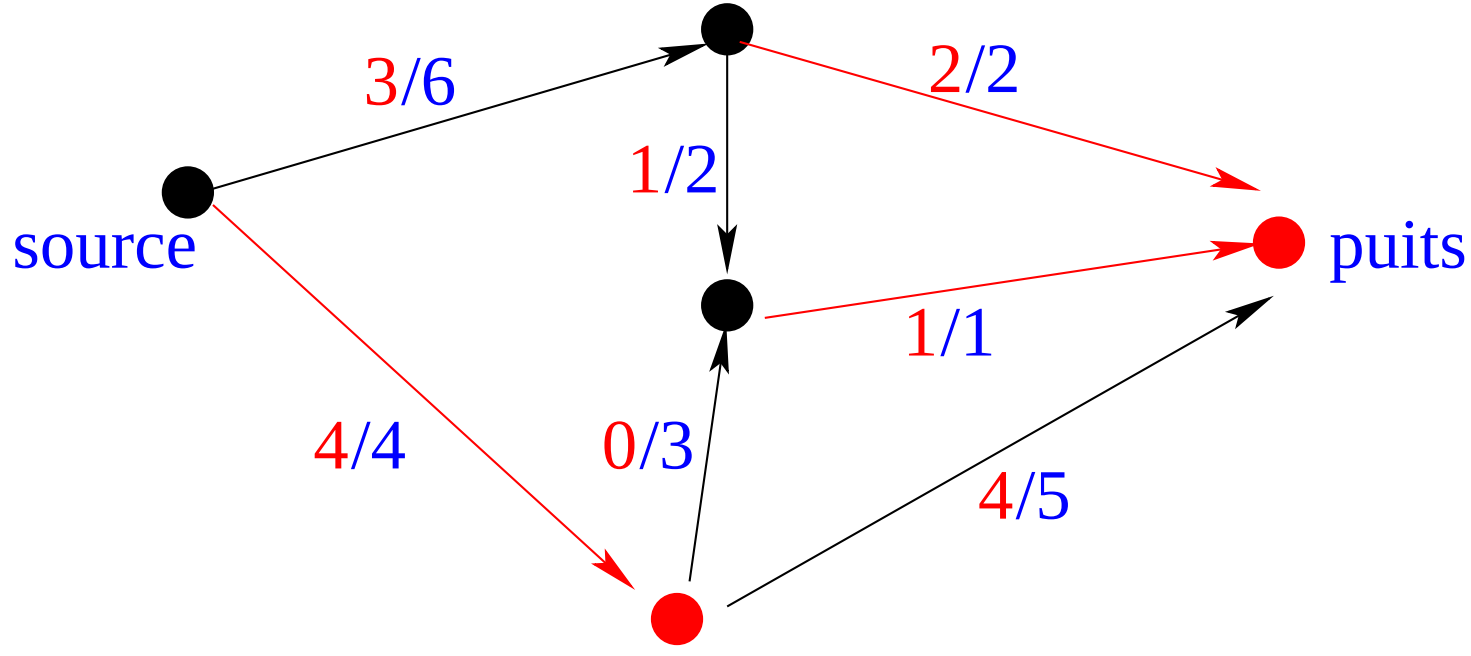
Flot maximal / Coupe minimale

- Réseau de transport.
- Flot maximal ≡ Coupe minimale.



Flot maximal / Coupe minimale

- Réseau de transport.
- Flot maximal ≡ Coupe minimale.



Algorithme de préflots

1. On affecte une hauteur initiale h à tous les nœuds :
   $h(source) = N$ et $\forall k \neq source,\ h(k) = 0$
   (N est le nombre de nœuds).
2. Les nœuds avec excédent de flot sont visités, dans un ordre quelconque.
   - Si possible, le flot est poussé vers un nœud plus bas. Les capacités des arêtes sont respectées.
   - Sinon, on élève le nœud.

Théorème

Le processus converge, quel que soit l'ordre de visite des nœuds. Un flot maximal est obtenu à la limite.

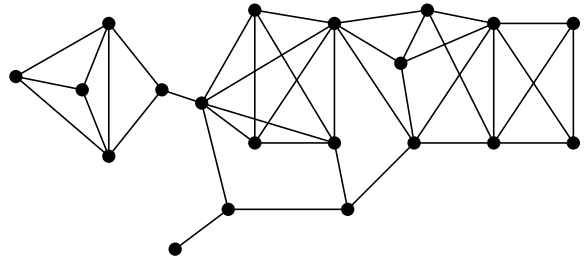
Adaptation au Web

Site Web : Nœuds d'un réseau de transport délimité par un Flot max. / Coupe min.

- **Nœuds** : pages Web, parcourues progressivement.
- **Arêtes** : hyperliens.
- **Capacité** : fonction décroissante des distances d'édition entre les URLs.
- Une source virtuelle, pointant vers une graine de pages avec des arêtes de capacité infinie.
- Un puits virtuel, pointé par tous les nœuds avec des arêtes de très faible capacité.

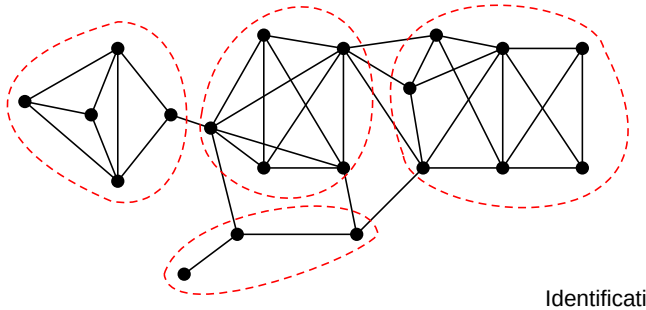
Markov CLustering algorithm

- Un algorithme hors-ligne de segmentation de graphe, basé sur la propagation de flots.
- Alternativement, expansion (multiplication) et inflation (renormalisation) d'une matrice stochastique.
- Fonctionne (presque) seulement sur des graphes non dirigés.
- **Complexité** : $O(n)$ (mais avec un important facteur).



Markov CLustering algorithm

- Un algorithme hors-ligne de segmentation de graphe, basé sur la propagation de flots.
- Alternativement, expansion (multiplication) et inflation (renormalisation) d'une matrice stochastique.
- Fonctionne (presque) seulement sur des graphes non dirigés.
- **Complexité** : $O(n)$ (mais avec un important facteur).



Propagation de flots à partir des classes de MCL

Processus :

1. Segmentation MCL d'une grande partie (appropriée) du graphe.
2. Identification de la classe la plus pertinente.
3. Propagation de flots à partir de cette classe.

Avantages par rapport à MCL seul :

- Découverte dynamique de classes.
- Utilise le fait que le graphe est dirigé.

Expériences

- Parcours d'une large partie de `*.inria.fr/*` — *72 heures*
- Segmentation par MCL du graphe obtenu — *24 heures*
- Identification de la classe GEMO — *< 1s*
- Propagation de flots à partir de cette classe — *3 heures*

$87\,140$ pages Web dans le graphe.
276 classes.
($5\,247$ en segmentant à nouveau la classe la plus grande).

Résultats

	Taille	Précision	Rappel
Propagation de flots	8	87.5%	1.3%
MCL	320	99.7%	33.0%
MCL + Propagation de flots	788	90.4%	86.4%
`www-rocq.inria.fr/verso/*`	221	100.0%	44.4%
`{www-rocq,osage}.inria.fr/verso/*`	683	100.0%	68.6%

En bref

- Site Web : une notion subjective, non évidente
- Propagation de flots utilisée pour découvrir les limites d'un site Web
- Meilleurs résultats obtenus par la combinaison d'une segmentation de graphe hors-ligne et d'une propagation de flots en-ligne

Perspectives

- Optimisation, passage à l'échelle.
- Calcul en-ligne de MCL ?
- Choix d'une stratégie de parcours efficace.
- Combinaison avec une méthode sémantique.

Perspectives

- Optimisation, passage à l'échelle.
- Calcul en-ligne de MCL ?
- Choix d'une stratégie de parcours efficace.
- Combinaison avec une méthode sémantique.

MERCI POUR VOTRE ATTENTION !