Identification of logical websites

Pierre Senellart

OpenSource LaTeX

What is a logical website?

- Simple idea: website = webserver
- But:
  - Some websites span over several webservers (e.g. `www-rocq.inria.fr`, `osage.inria.fr`...).
  - Some webservers contain different websites (e.g. `perso.wanadoo.fr`).
  - Some websites may be split into different websites (e.g. INRIA website).
- Limits of a website: a subjective notion

Talk outline

1. Introduction

2. Flow Simulation
   - ✓ Maximal Flow / Minimal Cut
   - ✓ Preflow-Push algorithm
   - ✓ Adaptation to the Web

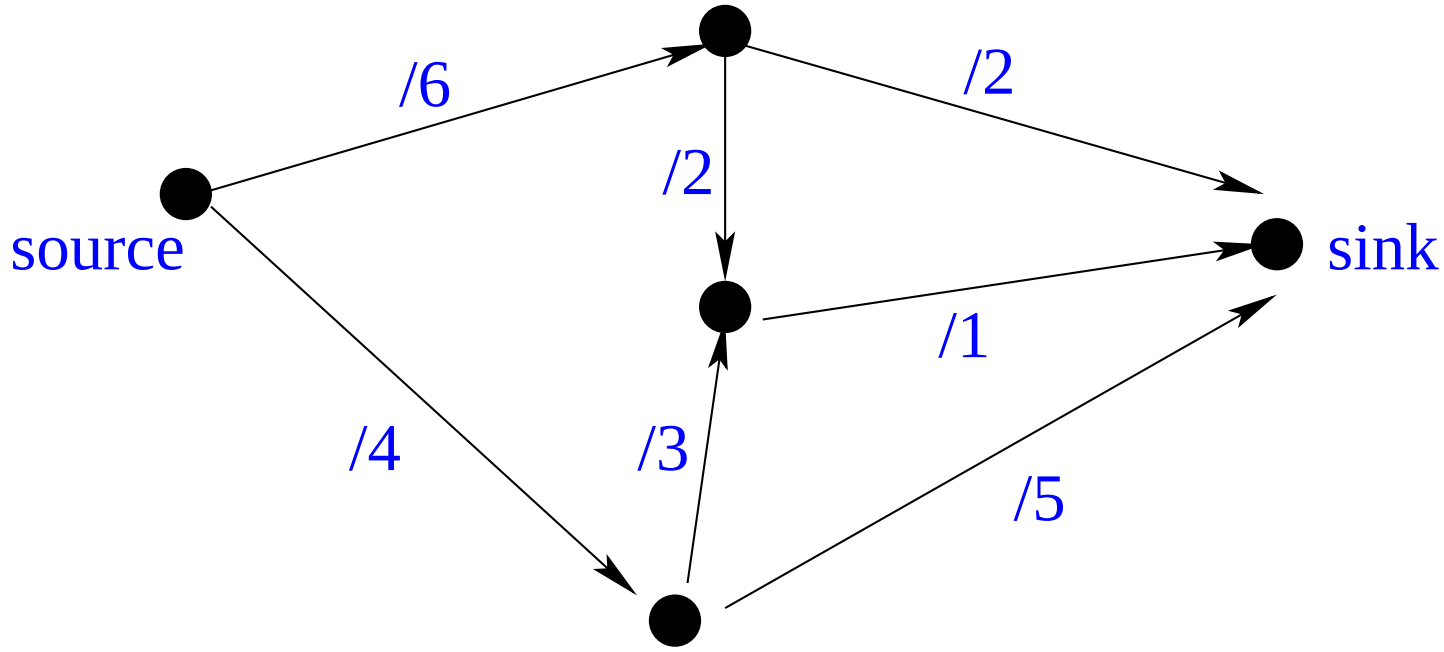
3. Seed Extension
   - ✓ Markov CLustering algorithm
   - ✓ Flow simulation from MCL

4. GEMO website
   - ✓ Experiments
   - ✓ Results

5. Conclusion

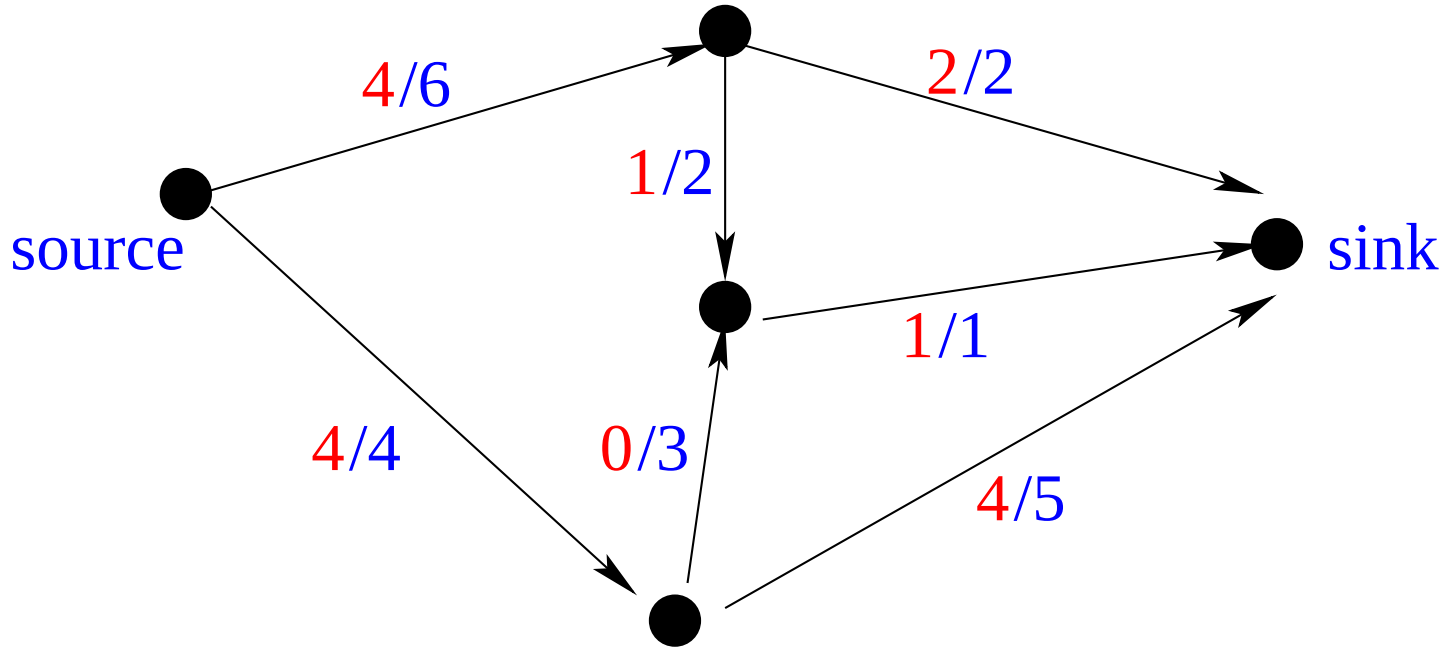
Maximal Flow / Minimal Cut

- Transport network
- Maximum flow ≡ Minimal cut



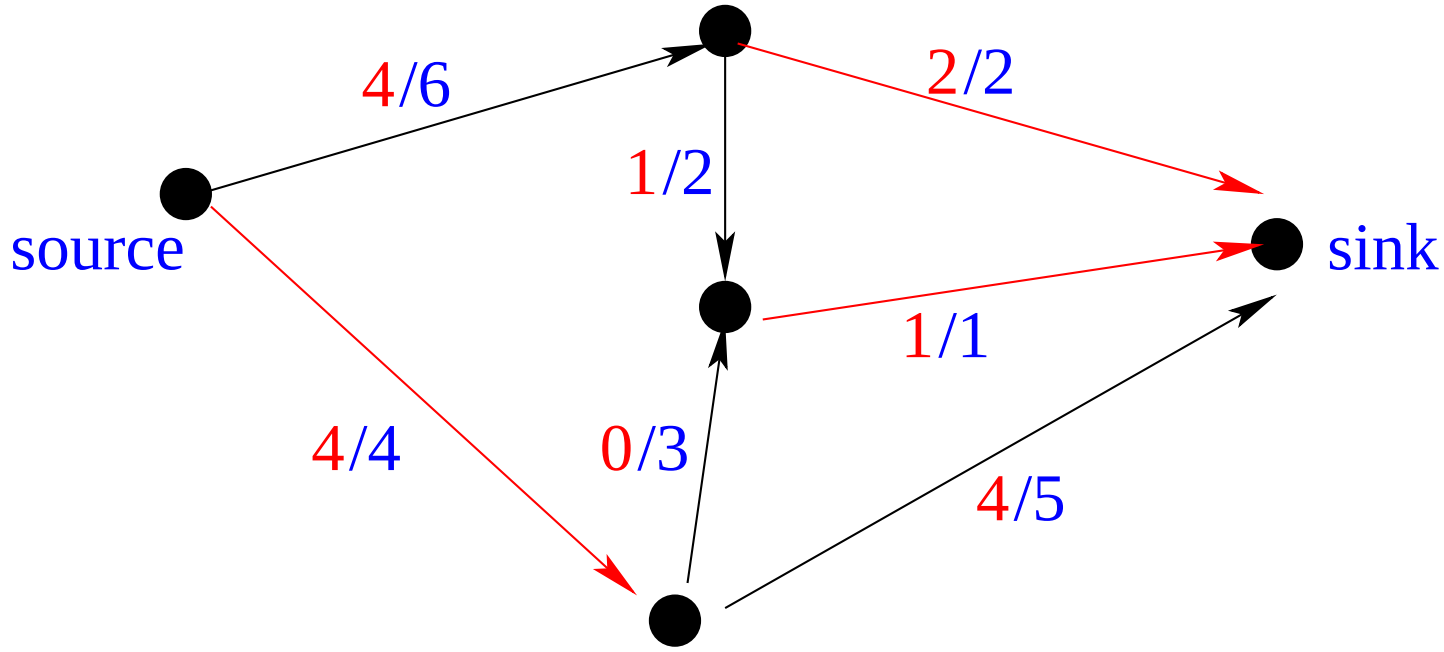
Maximal Flow / Minimal Cut

- Transport network
- Maximum flow ≡ Minimal cut



Maximal Flow / Minimal Cut

- Transport network
- Maximum flow ≡ Minimal cut



Preflow-Push algorithm

1. All nodes are assigned a height h: $h(source) = N$, $\forall k \neq source, h(k) = 0$ (N is the number of nodes)
2. Nodes with an overflow are visited, in any order.
   - If possible, the flow is pushed toward a lower node. Capacities of edges are respected.
   - Otherwise, the node is heigtened.

Theorem
The process converges, whatever the sequence of visited nodes may be. The maximal flow is obtained at the limit.

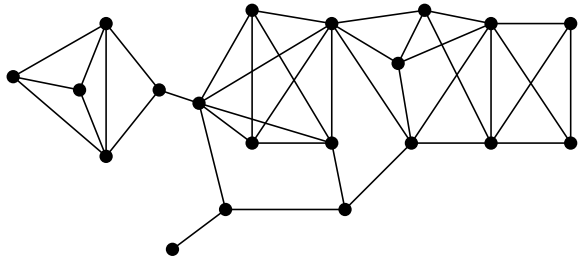
Adaptation to the Web

Website: nodes of a flow network delimited by a MaxFlow / MinCut.

- **Nodes**: webpages, progressively crawled
- **Edges**: hyperlinks
- **Capacities**: edition distance between URLs
- A virtual source, pointing to a seed of pages with infinite capacity edges
- A virtual sink, pointed by all nodes with very low capacity edges

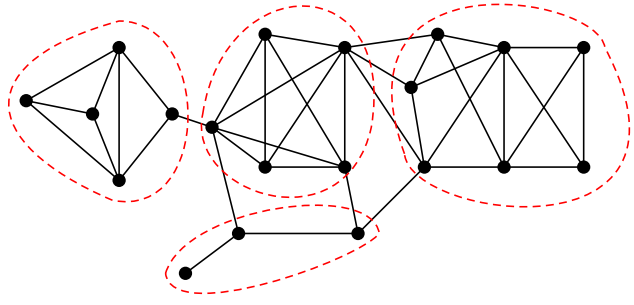
Markov CLustering algorithm

- An off-line Graph Clustering Algorithm based on flow simulation
- Alternation between expansion (multiplication) and inflation (rescaling) of a stochastic matrix
- Work (almost) only on undirected graphs
- **Complexity**: $O(n)$ (but with a high multiplying factor)



Markov CLustering algorithm

- An off-line Graph Clustering Algorithm based on flow simulation
- Alternation between expansion (multiplication) and inflation (rescaling) of a stochastic matrix
- Work (almost) only on undirected graphs
- **Complexity**: $O(n)$ (but with a high multiplying factor)



Flow Simulation from MCL clusters

Process:

1. MCL Clustering of a large relevant portion of the graph
2. Identification of the most relevant cluster
3. Flow Simulation starting from this cluster

Advantages over pure MCL:

- Dynamic discovery of clusters
- Use the fact that the graph is directed

Experiments

- Crawl of a large part of `*.inria.fr/*` *72 hours*
- MCL clustering of the obtained graph *24 hours*
- Identification of the GEMO cluster *< 1s*
- Flow Simulation from this cluster *3 hours*

$87,140$ webpages in the graph.
276 clusters.

1. Introduction
2. Flow Simulation
3. Seed Extension Results
4. GEMO website
   ✓ Results
5. Conclusion

Results

	Size	Precision	Recall	THESUS
Flow Simulation	8	87.5	1.3	xml web
MCL	320	99.7	33.0	gemo report
MCL + Flow Simulation	788	90.4	86.4	gemo report
`www-rocq.inria.fr/verso/*`	221	100.0	44.4	diapositive texte
`{www-rocq,osage}.inria.fr/verso/*`	683	100.0	68.6	report diapositive

In Brief

- Website: a subjective, not obvious notion
- Flow Simulation used to discover the limits of a website
- Best results given by a combination of an off-line graph clustering and an on-line flow simulation

Future works

- Optimization, scaling
- On-line computing of MCL?
- Hierarchical clustering
- Application of the Preflow-Push algorithm to peer-to-peer networks

Future works

- Optimization, scaling
- On-line computing of MCL?
- Hierarchical clustering
- Application of the Preflow-Push algorithm to peer-to-peer networks

THANK YOU FOR YOUR ATTENTION!