CESi recherche

IRISE : INNOVATION ET SCIENCES DE L'ENTREPRISE

Detection of Communities in Directed Networks based on Strongly p-Connected Components

Vincent Levorato CESI IRISE, LIFO
Coralie Petermann LaISC
FRANCE

Outline

- Context and problematic
- Definitions
- Our method
- Experiments
- Conclusion

CESi recherche

IRISE : INNOVATION ET SCIENCES DE L'ENTREPRISE

Context

- Complex networks $\subset$ Complex systems.
- Analyzing the **structure** and the **dynamic** of a complex network
  $\rightarrow$ search for "equivalent" elements.
- The objective is to classify elements using: clustering, graph partitions methods, ..., and communities detection.

CESi recherche
IRISE : INNOVATION ET SCIENCES DE L'ENTREPRISE

Communities Detection Problematic

- Model of the network $\rightarrow$ graph theory.
- How to detect communitities ?
  Definition of a *community* ?
- Lot of scientific works cover *undirected networks* case.
- Our problematic: detection of communities in directed networks.

CESi recherche
IRISE : INNOVATION ET SCIENCES DE L'ENTREPRISE

Basic Definitions

- A **digraph** $G = (V, A)$ [Berge1970] is composed of:
  - a set $V = \{x_1, x_2, ..., x_n\}$ named *vertices* or *nodes*.
  - a family $A = (a_1, a_2, ..., a_n)$ of elements of the Cartesian product $V \times V = \{(x, y)/x \in V, y \in V\}$ named *arcs*.
- A **path** P is composed of k arcs such as $P = (a_1, a_2, ..., a_i, ..., a_k)$.
- A **chain** is a path without orientation (sometimes called *semi-path*).
- A **circuit** is a path such that the first node of the path corresponds to the last one.

CESi recherche
IRISE : INNOVATION ET SCIENCES DE L'ENTREPRISE

Connected Components in Digraph [Harary1969]

- **weakly connected component**: $\forall x, y \in WCC$, there is a chain between x and y.
- **unilaterally connected component**: $\forall x, y \in UCC$, there is a path between x and y, or there is a path between y and x.
- **strongly connected component**: $\forall x, y \in SCC$, there is a path between x and y, and there is a path between y and x.

CESi recherche
IRISE : INNOVATION ET SCIENCES DE L'ENTREPRISE

Strongly p-Connected Components

- **Definition** [WassermanFaust1994]:
  $\forall x, y \in$ p-SCC, there is a path of length p or less between x and y, and there is a path of length p or less between y and x, with $p \geq 2$.

Communities Definition

- Communities are parts of a network which are relatively similar sets of nodes strongly interrelated and more *weakly associated with the rest of the network*.
- Alternative definition: each element can *strongly* communicate with other elements
  $\rightarrow$ there is always a bidirectional flow that connects two elements.

Our Method

- **Main idea:** detection of communities $\rightarrow$ finding p-SCC.
- How ?
  - Exploration of the graph using a breadth-first search (BFS).
  - Bounded by p (maximum path length).

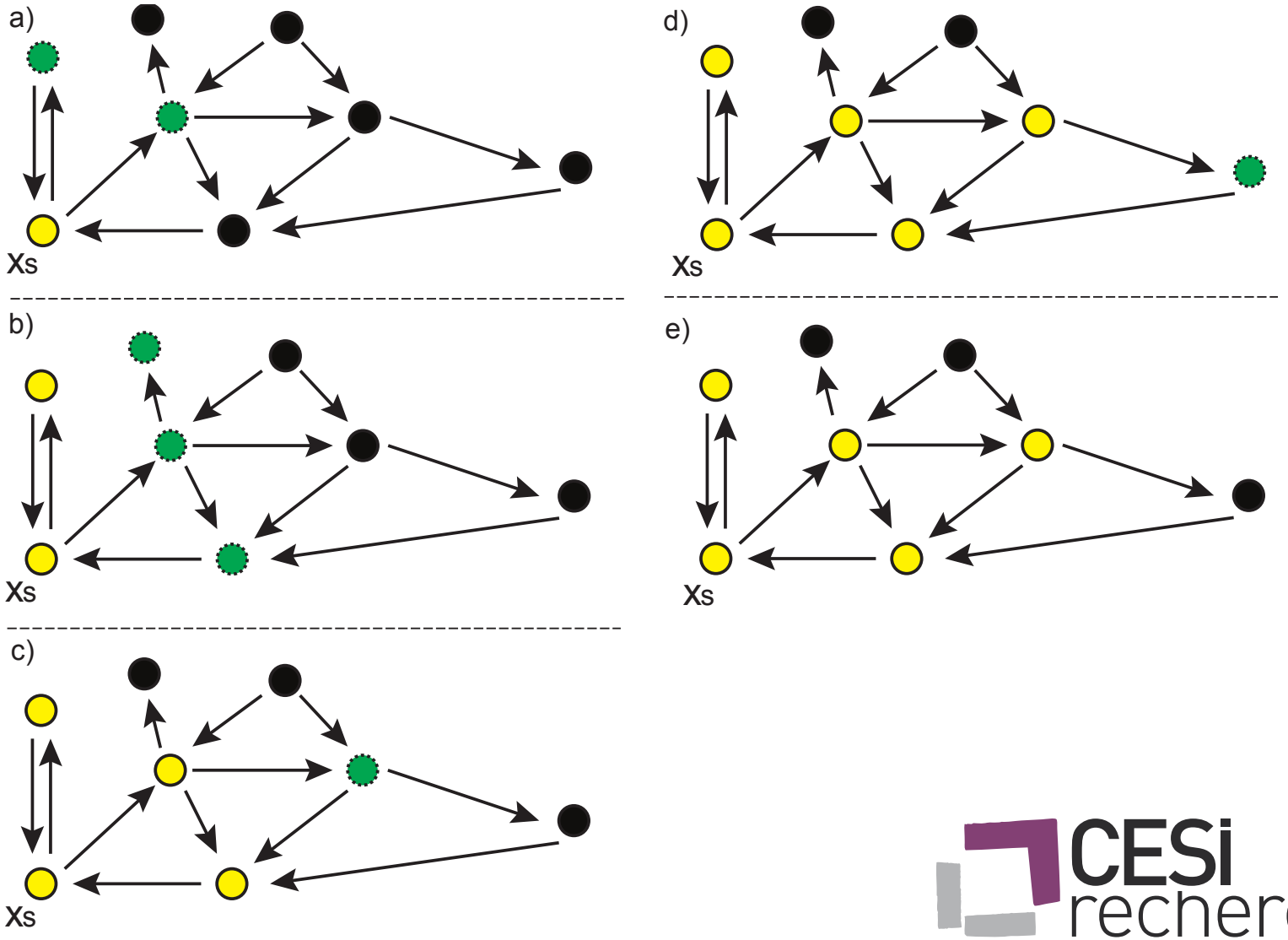
Our Method

1. Choose a random start vertex x_s.
2. Make a BFS beginning at x_s.
3. If, during the BFS, a circuit with x_s is found, all nodes of the circuit are added to the current component.
4. Stop the BFS if current path length reaches p, and add current component to the list of p-SCC.
5. Remove nodes of the last added component of the graph.
6. Go to step 1 if the graph is not empty.

CESi recherche

IRISE : INNOVATION ET SCIENCES DE L'ENTREPRISE

Illustration

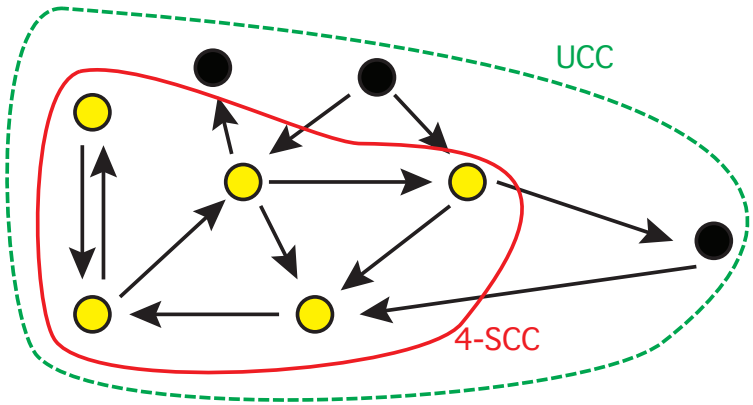


CESi recherche

IRISE : INNOVATION ET SCIENCES DE L'ENTREPRISE

Adjusting phase

- The final phase groups communities having a too *small size* $\rightarrow$ parameter.
- Merging operation: p-SCC $\rightarrow$ UCC.



CESi recherche

IRISE : INNOVATION ET SCIENCES DE L'ENTREPRISE

Experiments

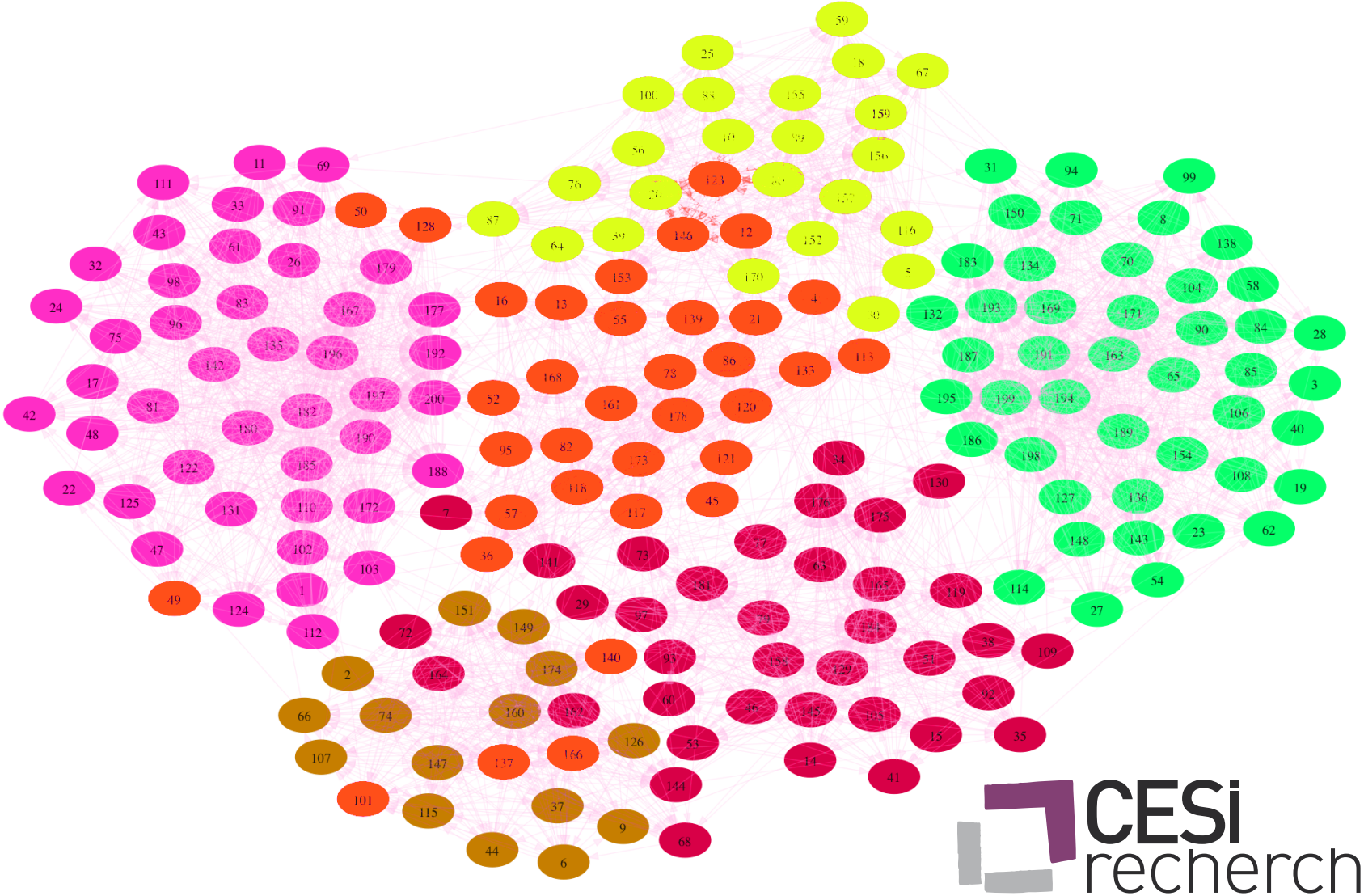
- Generator of (di)graphs [LancichinettiFortunato2009]:
  - n the number of nodes of the graph
  - k_{avg} the average node degree, k_{max} the maximum node degree
  - min_c, max_c min. and max. community size
  - μ the mixing parameter i.e. for each vertex the ratio between the number of outgoing arcs and the number inside its community.

CESi recherche
IRISE : INNOVATION ET SCIENCES DE L'ENTREPRISE

Evaluation Methods

- Quality of the found clutering Vs. the reference clustering.
- We use the Normalized Mutual Information measure [Danon and al. 2005] but also:
  - Jaccard Index
  - Adjusted Rand Index
  - F-measure

Results



CESi recherche

IRISE : INNOVATION ET SCIENCES DE L'ENTREPRISE

Results

- We compare our method (LP) to InfoMap method [RosvallBergstrom2008] using different profiles:
  - "Common" graph
    $k_{avg} = 15, k_{max} = 50, min_c = 20, max_c = 50, \mu = 0.1$
  - Graph with Bad Defined Communities
    $\mu = 0.1$ to $\mu = 0.5$.
  - Undense Graph
    $k_{avg} = 7$
- The parameters used for our method are:
  $p = 4$ (maximum path length)
  $minCsize = 3$ (minimum community size).

CESi recherche
IRISE : INNOVATION ET SCIENCES DE L'ENTREPRISE

Common Graphs

		Amount of nodes					
Algorithm	**Measures**	1000	2000	3000	4000	5000	10000
LP Method	ARI	0.95	0.97	0.98	0.98	0.98	0.99
	Jaccard	0.91	0.95	0.96	0.96	0.97	0.98
	F-Measure	0.97	0.98	0.99	0.99	0.99	0.99
	NMI	0.93	0.95	0.97	0.97	0.98	0.99
InfoMap	ARI	0.62	0.64	0.65	0.67	0.67	0.69
	Jaccard	0.47	0.48	0.49	0.50	0.50	0.53
	F-Measure	0.69	0.71	0.72	0.74	0.75	0.77
	NMI	0.41	0.46	0.48	0.50	0.51	0.52

Graphs with Bad Defined Communities

		Mixing parameter				
Algorithm	**Measures**	0.1	0.2	0.3	0.4	0.5
LP Method	ARI	0.98	0.92	0.79	0.59	0.33
	Jaccard	0.97	0.86	0.66	0.43	0.20
	F-Measure	0.99	0.96	0.89	0.78	0.62
	NMI	0.98	0.92	0.81	0.66	0.45
InfoMap	ARI	0.67	0.48	0.33	0.19	0.13
	Jaccard	0.50	0.32	0.20	0.11	0.07
	F-Measure	0.75	0.62	0.51	0.38	0.31
	NMI	0.51	0.30	0.15	0.03	0.00

CESi recherche

IRISE : INNOVATION ET SCIENCES DE L'ENTREPRISE

Undense Graphs

	ARI	Jaccard	F-Measure	NMI	$\lvert C \rvert$
LP ($minCsize = 3$)	0.68	0.51	0.76	0.56	$\simeq 260$
LP ($minCsize = 5$)	0.60	0.43	0.73	0.57	$\simeq 195$
LP ($minCsize = 10$)	0.48	0.39	0.69	0.62	$\simeq 155$
LP ($minCsize = 15$)	0.23	0.13	0.62	0.61	$\simeq 130$
InfoMap	0.59	0.42	0.64	0.37	$\simeq 265$

CESi recherche

IRISE : INNOVATION ET SCIENCES DE L'ENTREPRISE

Complexity

- Theoritical complexity: exponential (bounded by p parameter).
- Practical complexity (generated graphs):

Nodes	1000	5000	10 000	50 000	100 000	250 000
Time	2s	4s	7s	36s	1min 10s	3min 20s

- Practical complexity (real networks):
  Epinions social network (75 879 nodes, 508 837 arcs).
  $\Rightarrow$ 5 mins
  EU email communication network (265 214 nodes, 420 045 arcs).
  $\Rightarrow$ 1 hour

CESi recherche
IRISE : INNOVATION ET SCIENCES DE L'ENTREPRISE

Conclusion

- Our work: We propose a simple algorithm which gives good results for detecting communities in directed networks.
- Future work:
  - Optimize the algorithm.
  - Find a way to automatically set the *minCsize* parameter.
  - Apply method to "real" cases.
  - Think about overlapping communities...

CESi recherche
IRISE : INNOVATION ET SCIENCES DE L'ENTREPRISE

Contact

LEVORATO Vincent
Researcher at IRISE (CESI)
Institute de Recherche en Innovation et Sciences de l'Entreprise
vlevorato@cesi.fr
France
959 rue de la Bergeresse
45160 Olivet