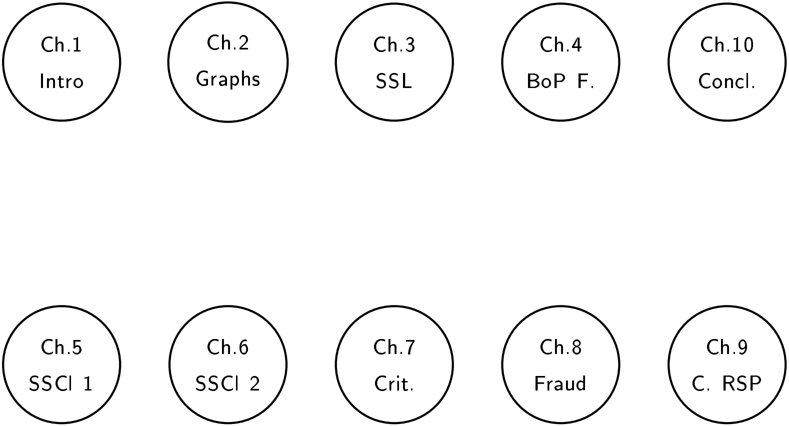
Network analysis based on bag-of-paths: semi-supervised classification, betweenness, criticality and Markov decision processes

B. Lebichot

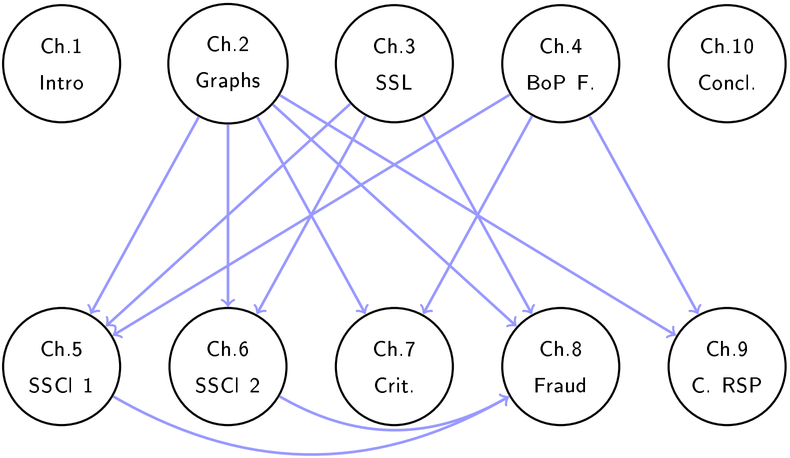
Université catholique de Louvain

21 février 2018

Introduction 1/2



Introduction 1/2



5 pages per chapter :
Intuition - Motivation(s) - Contribution(s) - Methodology - Results

Important concept of Chapter 2 (Graph and networks) :
graphs, paths, (killed) Markov chain, shortest path and commute time (CT) distance, cost of a path

Important concept of Chapter 3 (Semi-supervised learning) :
supervised/semi-supervised/unsupervised learning, consistency assumption, transductive/inductive learning, graph-based classification

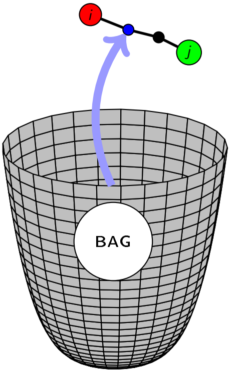
Intuition :
A bag containing ALL finite **paths**, weighted by total cost.

Contribution(s) ([Francoisse-2017]) : A framework

- Probabilities of picking a path between nodes i and j ($\mathbf{Z}$).
- **Interpolates** between shortest path & CT distances.
- If T is high $\Rightarrow$ longer paths are favored.
- If T is small $\Rightarrow$ shorter paths are favored.

Motivation(s) :

- **Shortest path distance** is efficient but forget all network information outside the shortest path.
- **Random-walk-based distance** converges to useless or unrealistic values.



Picking a path $\wp_{ij}$ from the bag :
Probability distribution $\mathrm{P}(\wp) \propto T$ on $\mathcal{P}$

The bag-of-paths (BoP) distribution

$$\begin{array}{ll} \underset{\{\mathbf{P}(\wp)\}}{\text{minimize}} & \displaystyle\sum_{\wp \in \mathcal{P}} P(\wp)\, \tilde{c}(\wp) \\ \text{subject to} & \sum_{\wp \in \mathcal{P}} \mathrm{P}(\wp) \ln(\mathrm{P}(\wp)/\tilde{\pi}^{\mathrm{ref}}(\wp)) = J_0 \\ & \sum_{\wp \in \mathcal{P}} \mathrm{P}(\wp) = 1 \end{array}$$

with $\tilde{\mathrm{P}}^{\mathrm{ref}}(\wp) = \tilde{\pi}^{\mathrm{ref}}(\wp) / \sum_{\wp' \in \mathcal{P}} \tilde{\pi}^{\mathrm{ref}}(\wp')$

The bag-of-paths framework 3/5

The result of the minimization (see **[Francoisse-2017]** for details) is a **Boltzmann probability distribution** :

$$\mathrm{P}(\wp) = \frac{\tilde{\pi}^{\mathrm{ref}}(\wp)\exp\left[-\theta\tilde{c}(\wp)\right]}{\sum_{\wp' \in \mathcal{P}} \tilde{\pi}^{\mathrm{ref}}(\wp')\exp[-\theta\tilde{c}(\wp')]} \qquad (1)$$

Long (high cost) paths have low probability.
Short (low cost) paths have high probability.

The bag-of-paths probability

$$P(s = i, e = j) = \frac{\sum_{\wp \in \mathcal{P}_{ij}} \tilde{\pi}^{\mathrm{ref}}(\wp)\exp[-\theta\tilde{c}(\wp)]}{\sum_{\wp' \in \mathcal{P}} \tilde{\pi}^{\mathrm{ref}}(\wp')\exp[-\theta\tilde{c}(\wp')]} \qquad (2)$$

$\mathbf{Z}$ is a **counting machine** for paths from i to j :

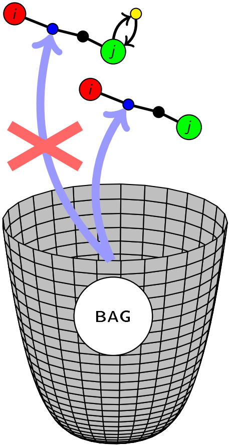
The bag-of-paths probability using Z

$$P(s = i, e = j) = \frac{z_{ij}}{\mathcal{Z}} \tag{3}$$

where $\mathbf{Z} = (\mathbf{I} - \mathbf{W})^{-1}$ and $\mathbf{W} = \mathbf{P}^{\mathrm{ref}} \circ \exp[-\theta \mathbf{C}]$

$\mathbf{Z}$ is the **fundamental matrix**.
$\mathcal{Z}$ is the **partition function** of the BoP system.

The bag-of-hitting-paths framework (absorbing path) :



Picking an hitting path $\wp^{\mathrm{h}}_{ij} \circ \wp_{jj} \in \mathcal{P}_{ij}$:
Probability distribution P_{h} on $\mathcal{P}_{\mathrm{h}}$

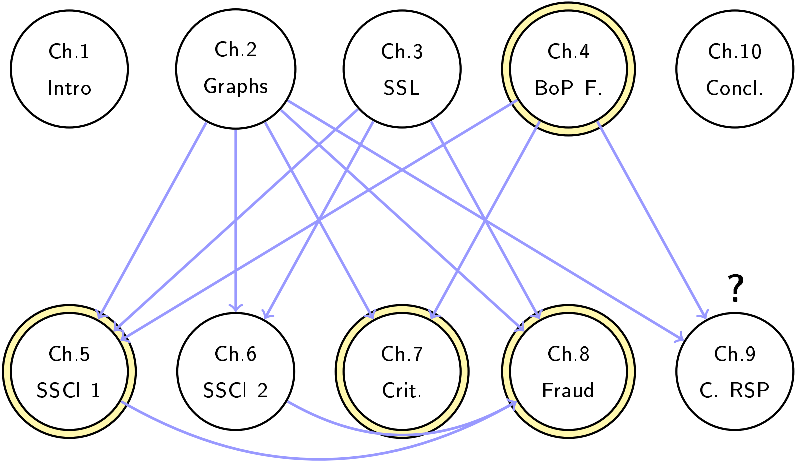
Which easily leads to

$$z^h_{ij} = z_{ij} / z_{jj}$$

and now $\tilde{\mathrm{P}}^{\mathrm{ref}}(\wp^{\mathrm{h}}) = \tilde{\pi}^{\mathrm{ref}}(\wp^{\mathrm{h}})$

The randomized-shortest-path
is just the BoHP for $i = 1$ and $j = n$.

Intuition : Can we predict if chapter 9 will be published ?



Motivation(s)

- Once we have informative networks, we are tempted to predict valuable information.
- (bio-)molecule network, text mining, web mining, social networks,...

Contribution(s)

- BoP $\Rightarrow$ BoP **betweeness**
- BoP betweeness $\Rightarrow$ BoP **group betweeness**
- BoP group betweeness $\Rightarrow$ semi-supervised **classifier**
- Compared to 7 algorithms on 13 datasets (on website).

The **bag-of-paths** betweenness :

$$\mathrm{bet}_j \triangleq \sum_{i=1}^{n} \sum_{k=1}^{n} \mathrm{P}(int = j | s = i, e = k; i \neq j \neq k \neq i) \tag{4}$$

The **bag-of-paths** group betweenness :

$$\mathrm{gbet}_j(\mathcal{C}_i, \mathcal{C}_k) \triangleq \mathrm{P}(int = j | s \in \mathcal{C}_i, e \in \mathcal{C}_k; s \neq int \neq e \neq s) \tag{5}$$

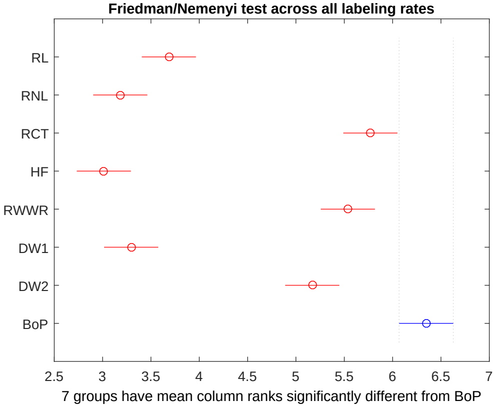
bet and **gbet**$(\mathcal{C}_i, \mathcal{C}_k)$ only $\propto$ **Z**.

The BoP classifier

$$\hat{\mathbf{y}} = \arg\max_{c \in \mathcal{L}} \{\mathbf{gbet}(\mathcal{C}_c, \mathcal{C}_c)\} \text{ with } \hat{\mathbf{y}} \propto \mathbf{Z}, \mathbf{y}^c \qquad (6)$$

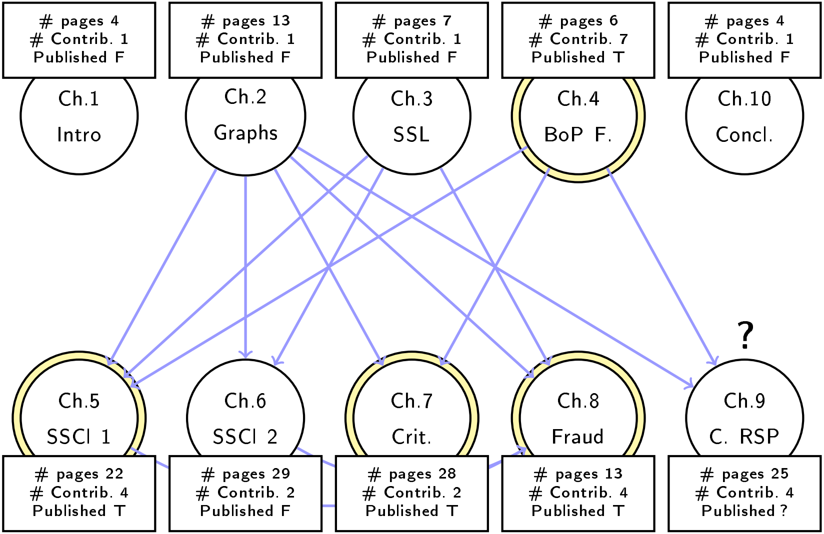
Experimental methodology :
7+1 transductive classifiers, 13 datasets, 5 runs
10 folds outer cross-validation ($l = \{10\%, 30\%, 50\%, 70\%, 90\%\}$)
10 folds inner cross-validation ($l = 90\%$)

Semi-supervised classification on graph 5/5



Special cases were also investigated :
unbalanced datasets, scarsely labeled datasets, runtimes.

Intuition : Can we predict if chapter 9 will be published ?



Motivation(s)

- Basically the same than before, but with additional features
- Particular case of **Multi-view learning**
- Therefore which **data source** is the most useful ?
- **Spatial correlation** analysis (consistency)

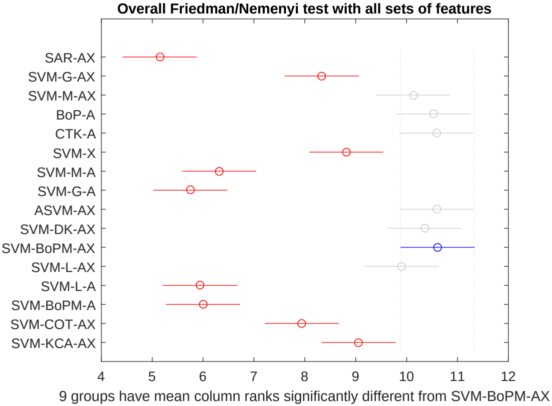
Contribution(s)

- Reviews different algorithms **[Fouss-2016]**.
- Compares 16 algorithms on 10 datasets (on website).
- Investigates **spatial correlation** analysis for classification.
- **General conclusions** to tackle this task

Experimental methodology (runtimes were also investigated) :
15 transductive classifiers + baseline SVM, 10 datasets, 5 runs
5 folds outer cross-validation ($l = 20\%$)
5 folds inner cross-validation ($l = 80\%$)
5 features sets ($n_f = \{100, 50, 25, 10, 5\}$)

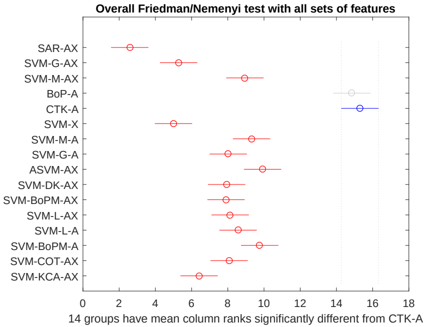
Classifier name	Used information	Familly
Bag of Path	Graph only	Graph-based
Regularized Commute Time Kernel	Graph only	Graph-based
SVM on Moran index only	Graph only	Graph embedding
SVM on Geary index only	Graph only	Graph embedding
SVM on LPCA only	Graph only	Graph embedding
SVM on BoP Modularity only	Graph only	Graph embedding [Devooght-2014]
SVM on Features only	Features only	Baseline
Spatial AutoRegressive model	Graph & Features	Extension of X-based classifier
SVM on Moran index and Features	Graph & Features	Graph embedding
SVM on Geary index and Features	Graph & Features	Graph embedding
SVM on LPCA and Features	Graph & Features	Graph embedding
SVM on BoP Modularity and Features	Graph & Features	Graph embedding [Devooght-2014]
SVM on Autocovariates and Features	Graph & Features	Extension of X-based classifier
SVM on Double Kernel	Graph & Features	Extension of X-based classifier
Co-training based on two SVMs	Graph & Features	Multi-view learning
SVM based on kernel canonical correl.	Graph & Features	Multi-view learning

considering all feature sets :

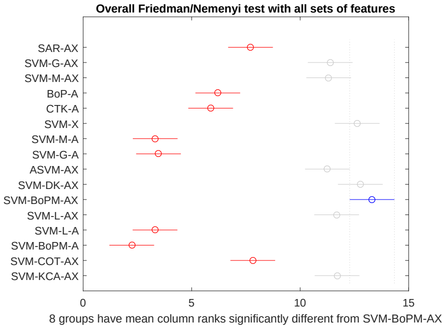


Other perspective were also investigated :
graph-based, dual sources, graph embedding methods only

considering dataset autocorrelation (Table 6.8) :

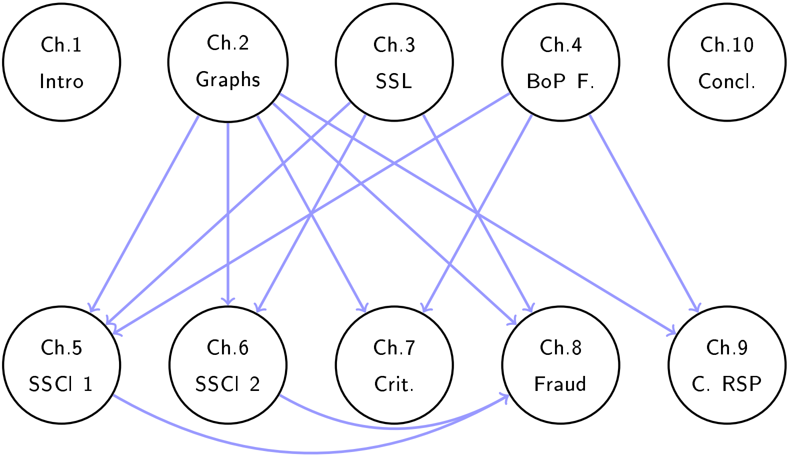


graph-driven datasets



features-driven datasets

Intuition : Which chapter is the most critical/important ?



A bag-of-paths node criticality measure 2/5

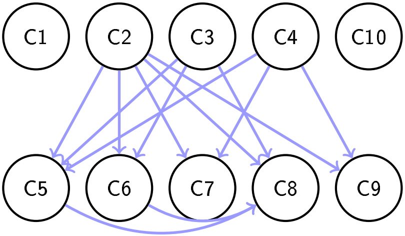
Motivation(s)

- Which node is **in-between**, **critical** for flow or **eccentric** ?
- Linked to the concept of betweenness
- Not application-dependent

Contribution(s)

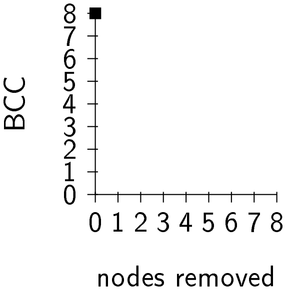
- A **new criticality measure** and a faster approximation
- 11 other criticality measures, we search for **correlations**.
- Compared using two disconnection strategies on random graphs and real-life social networks.

Experimental methodology (also updated ranking) :

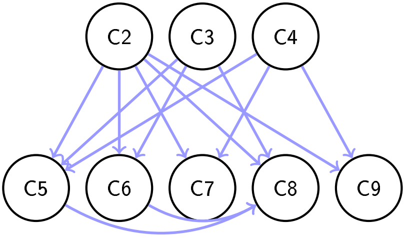


the ranking : [5 2 6 3 7 8 4 9]

The BCC : 8/8

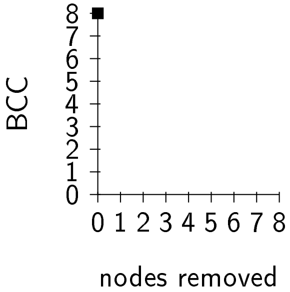


Experimental methodology (also updated ranking) :

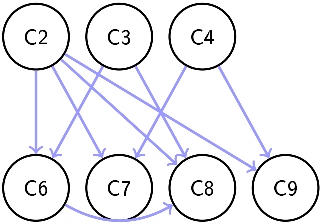


the ranking : [5 2 6 3 7 8 4 9]

The BCC : 8/8

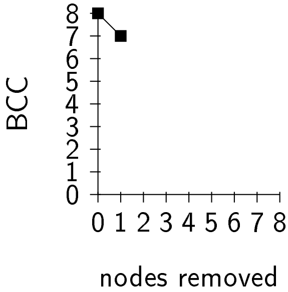


Experimental methodology (also updated ranking) :

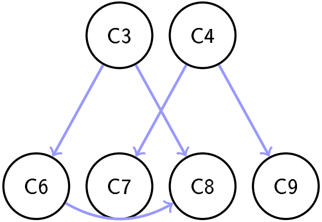


the ranking : [X 5 6 3 7 8 4 9]

The BCC : 8/8

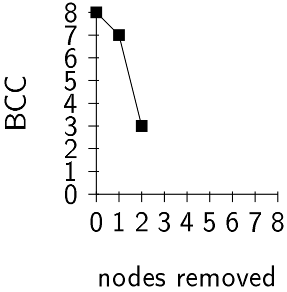


Experimental methodology (also updated ranking) :

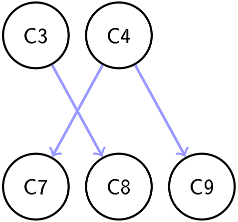


the ranking : [X X 6 3 7 8 4 9]

The BCC : 8/8

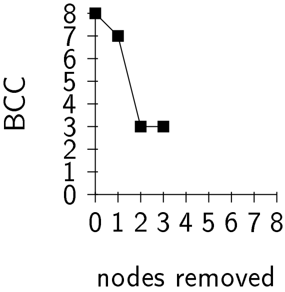


Experimental methodology (also updated ranking) :

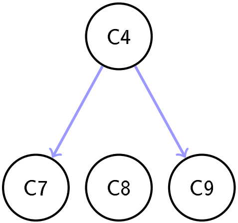


the ranking : [X X X 3 7 8 4 9]

The BCC : 8/8

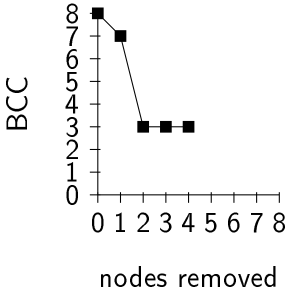


Experimental methodology (also updated ranking) :

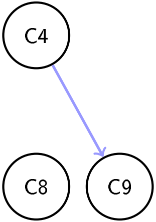


the ranking : [X X X X 7 8 4 9]

The BCC : 8/8

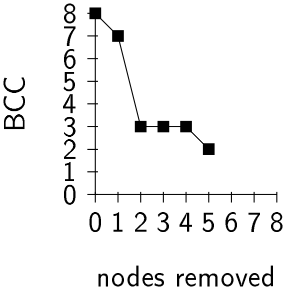


Experimental methodology (also updated ranking) :

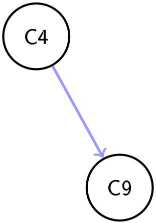


the ranking : [X X X X X 8 4 9]

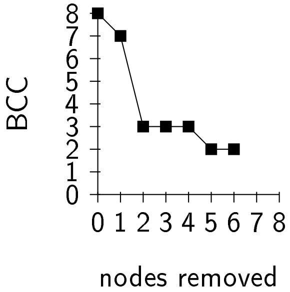
The BCC : 8/8



Experimental methodology (also updated ranking) :

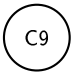


The BCC : 8/8



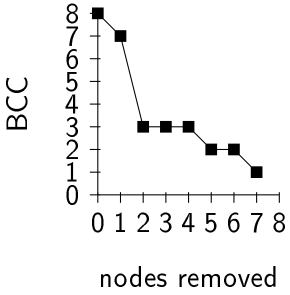
the ranking : [X X X X X X 4 9]

Experimental methodology (also updated ranking) :



the ranking : [X X X X X X X 9]

The BCC : 8/8



A bag-of-paths node criticality measure 3/5

The bag-of-paths criticality

$$\mathrm{cr}_j = \sum_{i,k=1(\;!)}^{n} P_{ik}^{(-j)}(\mathbf{A}) \log\left(\frac{P_{ik}^{(-j)}(\mathbf{A})}{P_{ik}(\mathbf{A}^{(-j)})}\right) \qquad (7)$$

KL divergence on accessibility, before and after node removal, between (+ fast approximation) :

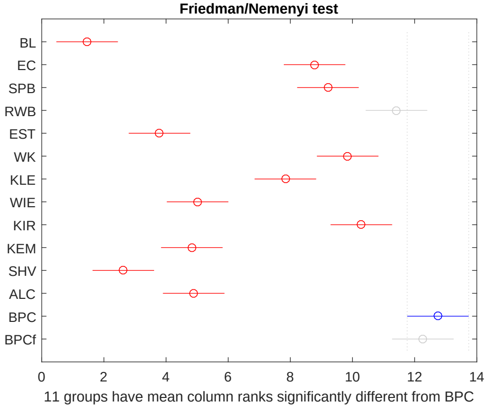
$$\mathrm{P}_{ik}(\mathbf{A}^{(-j)}) = \frac{z_{ik}(\mathbf{A}^{(-j)})}{\sum\limits_{i',k'=1(\;!)}^{n} z_{i'k'}(\mathbf{A}^{(-j)})}$$

j being removed from the graph.

$$\mathrm{P}_{ik}^{(-j)}(\mathbf{A}) = \frac{z_{ik}^{(-j)}(\mathbf{A})}{\sum\limits_{i',k'=1(\;!)}^{n} z_{i'k'}^{(-j)}(\mathbf{A})}$$

j is ignored in $\mathbf{Z}$.

The smaller AUC, the better the measure (also updated ranking) :



100 A-B graphs, 1 ranking. **Other results** were also investigated : E-R graphs, small social networks, clustering on ranking correlations

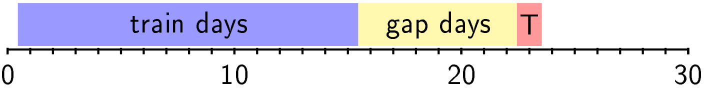
Graph-based fraud detection 1/5

Intuition :

From the Brufence project : it is to prevent error or **frauds**.

Worldline uses $\approx$ 200 rules to prevent, some being **data driven**.

Most of investigation by humans $\Rightarrow$ Propose $\approx$ 100 frauds/day.



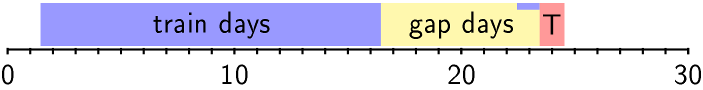
Trx/Card Precision@100 (T.P./100 most probable frauds)

Intuition :

From the Brufence project : it is to prevent error or **frauds**.

Worldline uses $\approx$ 200 rules to prevent, some being **data driven**.

Most of investigation by humans $\Rightarrow$ Propose $\approx$ 100 frauds/day.



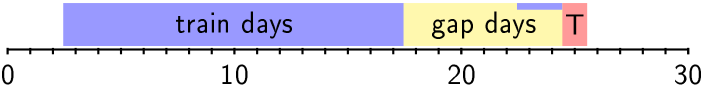
Trx/Card Precision@100 (T.P./100 most probable frauds)

Intuition :

From the Brufence project : it is to prevent error or **frauds**.

Worldline uses $\approx$ 200 rules to prevent, some being **data driven**.

Most of investigation by humans $\Rightarrow$ Propose $\approx$ 100 frauds/day.



Trx/Card Precision@100 (T.P./100 most probable frauds)

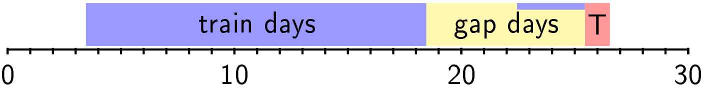
Graph-based fraud detection 1/5

Intuition :

From the Brufence project : it is to prevent error or **frauds**.

Worldline uses $\approx$ 200 rules to prevent, some being **data driven**.

Most of investigation by humans $\Rightarrow$ Propose $\approx$ 100 frauds/day.



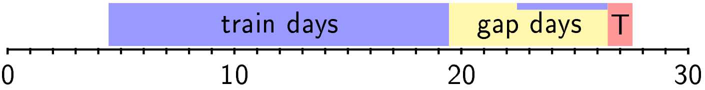
Trx/Card Precision@100 (T.P./100 most probable frauds)

Intuition :

From the Brufence project : it is to prevent error or **frauds**.

Worldline uses $\approx$ 200 rules to prevent, some being **data driven**.

Most of investigation by humans $\Rightarrow$ Propose $\approx$ 100 frauds/day.



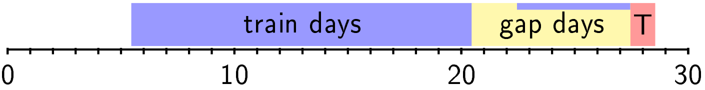
Trx/Card Precision@100 (T.P./100 most probable frauds)

Intuition :

From the Brufence project : it is to prevent error or **frauds**.

Worldline uses $\approx$ 200 rules to prevent, some being **data driven**.

Most of investigation by humans $\Rightarrow$ Propose $\approx$ 100 frauds/day.



Trx/Card Precision@100 (T.P./100 most probable frauds)

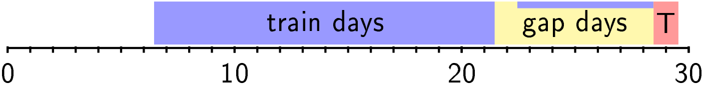
Graph-based fraud detection 1/5

Intuition :

From the Brufence project : it is to prevent error or **frauds**.

Worldline uses $\approx$ 200 rules to prevent, some being **data driven**.

Most of investigation by humans $\Rightarrow$ Propose $\approx$ 100 frauds/day.



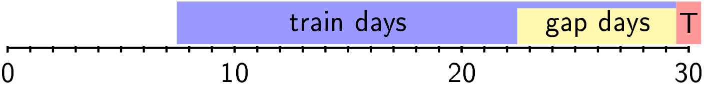
Trx/Card Precision@100 (T.P./100 most probable frauds)

Intuition :

From the Brufence project : it is to prevent error or **frauds**.

Worldline uses $\approx$ 200 rules to prevent, some being **data driven**.

Most of investigation by humans $\Rightarrow$ Propose $\approx$ 100 frauds/day.



Trx/Card Precision@100 (T.P./100 most probable frauds)

Motivation(s)

- If we improve detection of 1%, we can save 350X my house...
- Graphs have been poorly investigated for Fraud detection.
- Concept drift, Fast, Big data (3V) and Unbalanced data

Contribution(s)

- Improve an FDS named APATE **[Van vlasselear-2015]**.
  - Realistic scenario (previous slide)
  - Uses human feedback (previous slide).
  - Damp hubs (next slides).
- Prove that graph analysis is useful for applied fraud detection.

Risk scores are obtained by iterating on a **tripartite graph** (APATE) :

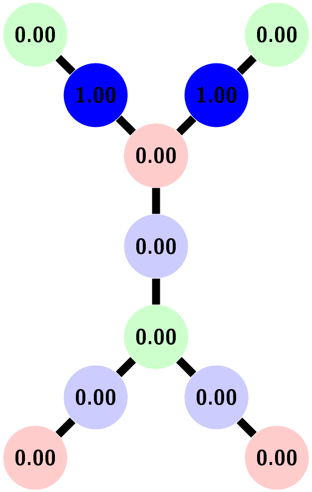
The random-walk with restart

$$\vec{r}_k = \alpha \cdot \mathbf{P}^T \vec{r}_{k-1} + (1-\alpha) \cdot \vec{r}_0$$

$\vec{r}_k$ can be divided by **d** (RCTK) to **damp hubs**.

3x4 features are created :
Trx - Merch - CH and 4 time decay.

Tractable (**one update per day**) but issues with new Trx/March/CH.



Risk scores are obtained by iterating on a **tripartite graph** (APATE) :

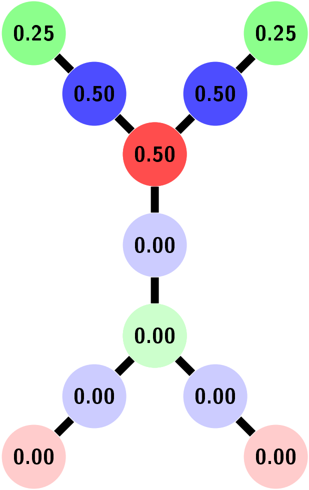
The random-walk with restart

$$\vec{r}_k = \alpha \cdot \mathbf{P}^T \vec{r}_{k-1} + (1-\alpha) \cdot \vec{r}_0$$

$\vec{r}_k$ can be divided by **d** (RCTK) to **damp hubs**.

3x4 features are created :
Trx - Merch - CH and 4 time decay.

Tractable (**one update per day**) but issues with new Trx/March/CH.



Risk scores are obtained by iterating on a **tripartite graph** (APATE) :

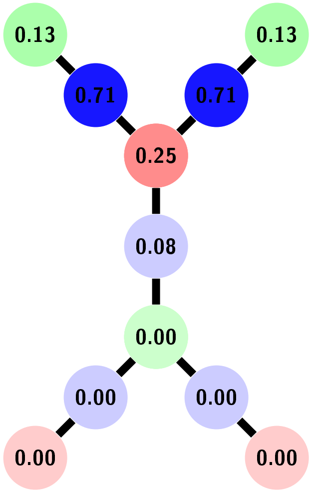
The random-walk with restart

$$\vec{r}_k = \alpha \cdot \mathbf{P}^T \vec{r}_{k-1} + (1-\alpha) \cdot \vec{r}_0$$

$\vec{r}_k$ can be divided by **d** (RCTK) to **damp hubs**.

3x4 features are created :
Trx - Merch - CH and 4 time decay.

Tractable (**one update per day**) but issues with new Trx/March/CH.



Risk scores are obtained by iterating on a **tripartite graph** (APATE) :

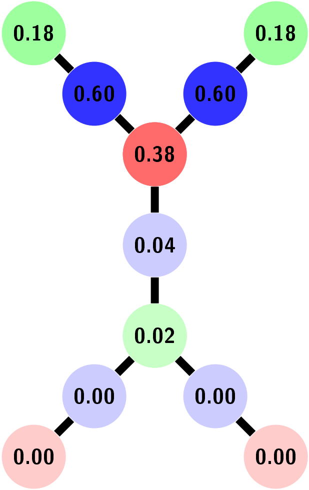
The random-walk with restart

$$\vec{r}_k = \alpha \cdot \mathbf{P}^T \vec{r}_{k-1} + (1-\alpha) \cdot \vec{r}_0$$

$\vec{r}_k$ can be divided by **d** (RCTK) to **damp hubs**.

3x4 features are created :
Trx - Merch - CH and 4 time decay.

Tractable (**one update per day**) but issues with new Trx/March/CH.



Risk scores are obtained by iterating on a **tripartite graph** (APATE) :

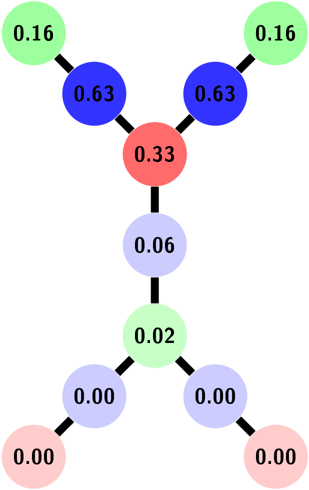
The random-walk with restart

$$\vec{r}_k = \alpha \cdot \mathbf{P}^T \vec{r}_{k-1} + (1-\alpha) \cdot \vec{r}_0$$

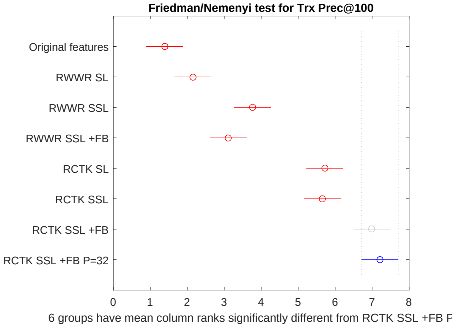
$\vec{r}_k$ can be divided by **d** (RCTK) to **damp hubs**.

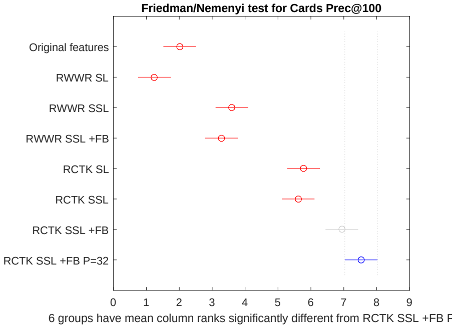
3x4 features are created :
Trx - Merch - CH and 4 time decay.

Tractable (**one update per day**) but issues with new Trx/March/CH.



Best methods combine **SSL**, **feedback** and **hubs damping**.

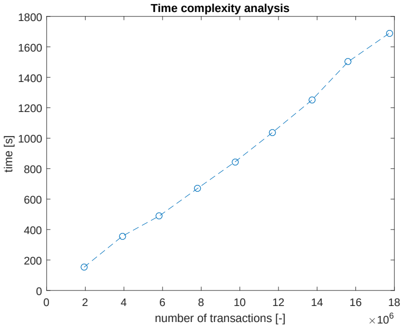




in terms of transactions

in terms of cards

Scalability was one of the main goal of this research



Time and space complexity **is O(n)**
In Matlab, R and Python !

Regular laptop can produce 3 risk scores in **a few minutes** for 20M transactions per day.

Other **minor improvements** were considered.

(notwithstanding classification)

Intuition : a **stochastic process** (decisions and random rewards)

MDP :
- a set of states
- a set of actions
- transition function, **independent of previous states**
- costs (or rewards) function

421 dice game :
- $3 * 6^3$ states (or 3*56)
- 2^3 actions
- 6 **ramdom transitions** per dice

reward			
1st D	**2nd D**	**3rd D**	**Score**
4	2	1	8
1	1	1	7
n	1	1	n
t	t	t	3
s	s-1	s-2	2
other			1

Interactions with other players are not taken into account...

Motivation(s)

- Deterministic policy leads to a predictable behavior.
- If environment is changing over time, good to **randomize**.
- Integrating the concept of connectivity

Contribution(s)

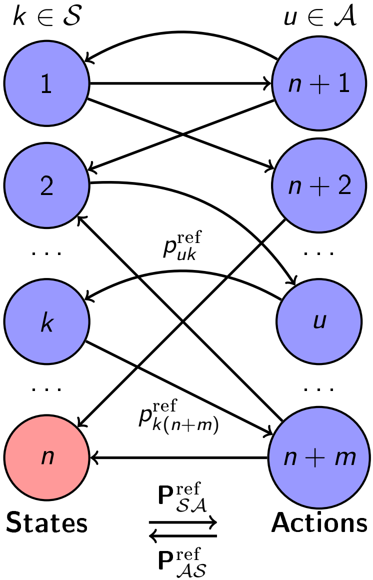
- Extends the RSP framework to tackle Markov decision process.
- We propose 2 algorithms, with mixed strategies as output.

How the MDP is modeled (RSP on bipartite graph) :

$\langle\tilde{c}\rangle$ is minimum and only $\propto$ **Z**.

$\mathcal{A} \rightarrow \mathcal{S}$ transitions are **constrained**
$\mathcal{S} \rightarrow \mathcal{A}$ transitions are free.
$\mathcal{S} \rightarrow \mathcal{A}$ transitions are the **mixed policy**.

Can be done with **modified edges costs**



Constrained RSP is identical to **a soft value iteration algorithm**.

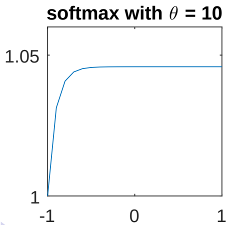
The simple value-iteration uses the Bellman-Ford algorithm between nodes 1 and n. It reads :

Bellman-Ford algorithm

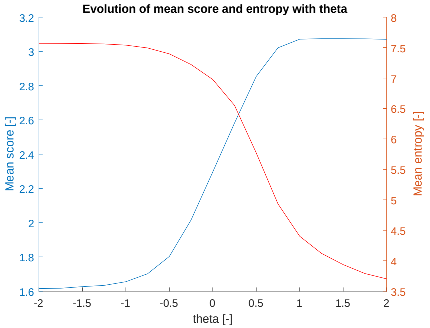
$$v_{kn} = \begin{cases} \min\limits_{u \in \mathcal{U}(k)} \left\{ c_{ku} + \sum\limits_{l \in \mathcal{S}ucc(u)} p^{\mathrm{ref}}_{ul} v_{ln} \right\} & \text{if } k \neq n \\ 0 & \text{if } k = n \end{cases}$$

min is replaced by a **softmin** :

$$\mathrm{softmin}(\mathbf{x}) = -\frac{1}{\theta} \log \left(\sum_{j=1}^{n} q_j \exp[-\theta x_j] \right)$$



Soft VI and constrained RSP : same mixed strategy interpolating between VI and random behavior



Non-soft VI is optimal.
Mean reward & entropy as expected.

DDD	Score	SVI reroll 1	SVI reroll 2
111	7	0 0 0	0 0 0
211	2	1 0 0	1 0 0
221	1	0 1 0	0 1 0
222	3	0 0 0	0 0 0
311	3	1 0 0	1 0 0
321	2	1 0 0	1 0 0
322	1	1 1 1	1 1 1
331	1	1 1 0	1 1 0
332	1	1 1 1	1 1 1
333	3	0 0 0	0 0 0
411	4	1 0 0	1 0 0
421	8	0 0 0	0 0 0
422	1	0 0 1	0 0 1
431	1	0 1 0	0 1 0
432	2	0 1 0	0 1 0
433	1	0 1 1	0 1 1
441	1	0 1 0	0 1 0
442	1	0 1 0	0 1 0
443	1	1 0 1	1 0 1
444	3	0 0 0	0 0 0
511	5	1 0 0	0 0 0
521	1	1 0 0	1 0 0
522	1	1 1 1	1 1 1
531	1	1 1 0	1 1 0
532	1	1 1 1	1 1 1
533	1	1 1 1	1 1 1
541	1	1 0 0	1 0 0
542	1	1 0 0	1 0 0
...	...	...	...

Limitations :

In all cases, introducing the BoP, and its underlying interpolation, improves the performance.

This interpolation comes with an increasing computational cost.

Chap.	BoP	Main limitation	Main further work
5	Yes	full $n \times n$ inversion	more efficient implementation
7	Yes	full $n \times n$ inversion	more efficient implementation
9	Yes	solve syst. of n equation	more efficient implementation
6	Yes	Lot of parameters	large graph analysis
8	No	(Field constrains)	compare with feature engineering **[Dal Pozzolo-2015]**

Conclusion 2/2

No nodes were injured during this thesis.

According to Chapter 5, Chapter 9 will be published.
According to Chapter 6, Chapter 9 will be published.
According to Chapter 7, Chapter 2 is the most critical (cfr title).
According to Chapter 8, paying twice an amount is not a fraud.
According to Chapter 9, we can play 421 after the defense and I will (probably) win.

Now it is question time...