Pattern Sampling in Distributed Databases

Lamine Diop, C. T. Diop, A. Giacometti, A. Soulet

Université Gaston Berger, LANI (Sénégal)

Université de Tours, LIFAT (France)

This work has been partly supported by CEA-MITIC

Outlines

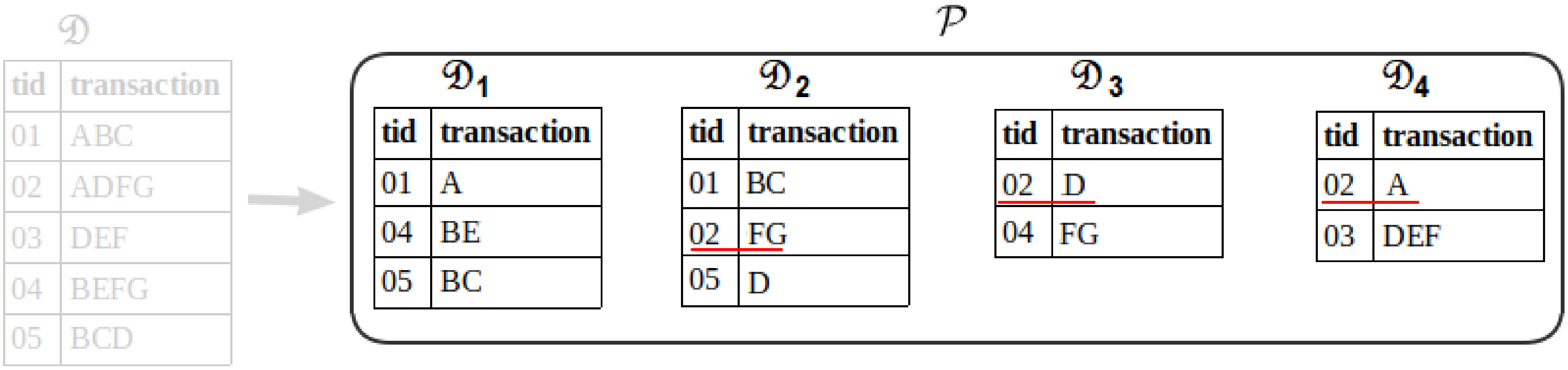
❑ Context and motivations

❑ Challenges for pattern sampling in distributed databases

❑ DDSampling (Distributed Database Sampling) : a generic two-step random procedure

❑ Experimentations and Perspectives

Definition of a distributed transactional database



$\mathcal{D}$

tid	transaction
01	ABC
02	ADFG
03	DEF
04	BEFG
05	BCD

$\mathcal{P}$

$\mathcal{D}_1$

tid	transaction
01	A
04	BE
05	BC

$\mathcal{D}_2$

tid	transaction
01	BC
02	FG
05	D

$\mathcal{D}_3$

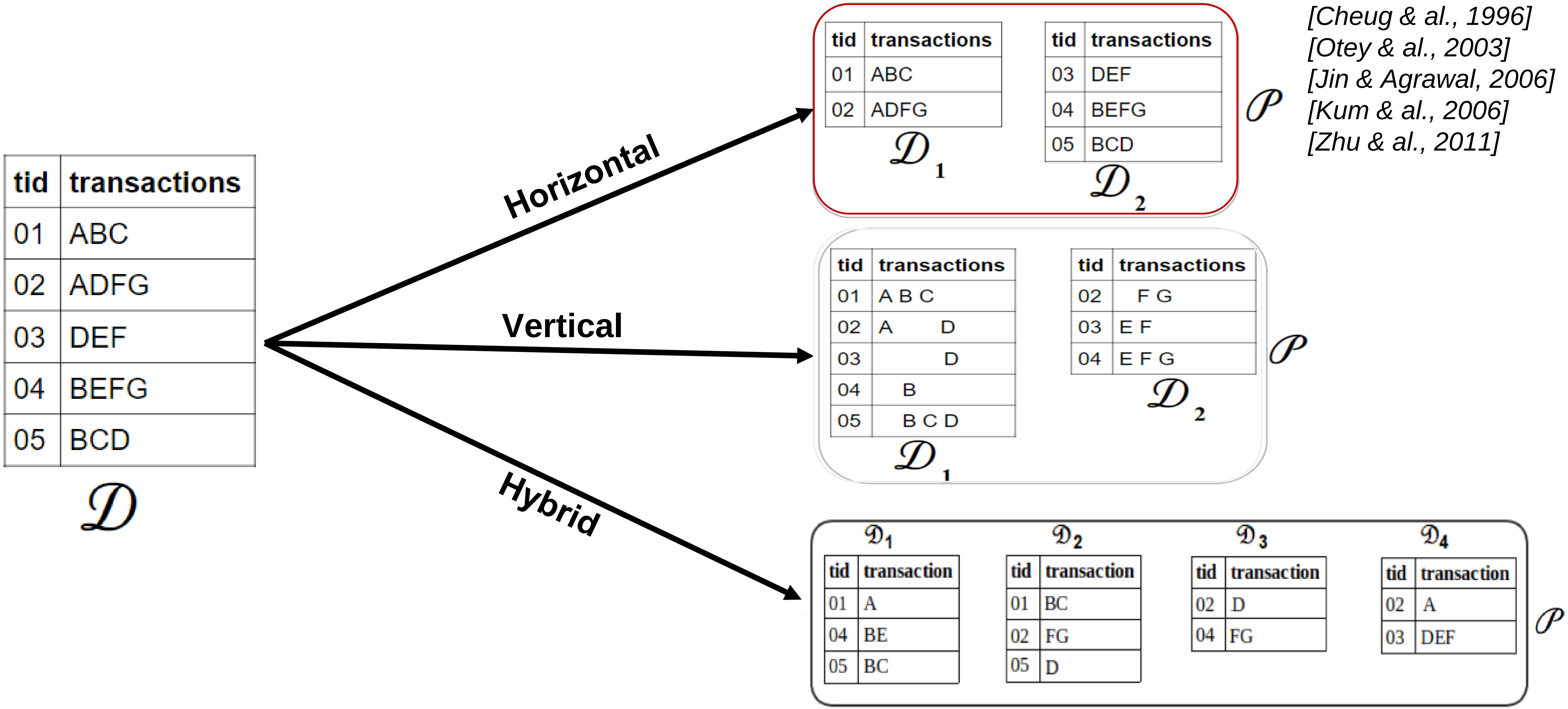
tid	transaction
02	D
04	FG

$\mathcal{D}_4$

tid	transaction
02	A
03	DEF

The transactions are distributed over several sites.
There is no repetition of items in the same transaction.

Types of fragmentation



tid	transactions
01	ABC
02	ADFG
03	DEF
04	BEFG
05	BCD

$\mathcal{D}$

Horizontal

tid	transactions
01	ABC
02	ADFG

$\mathcal{D}_1$

tid	transactions
03	DEF
04	BEFG
05	BCD

$\mathcal{D}_2$

$\mathcal{P}$

[Cheug & al., 1996]
[Otey & al., 2003]
[Jin & Agrawal, 2006]
[Kum & al., 2006]
[Zhu & al., 2011]

Vertical

tid	transactions
01	A B C
02	A D
03	D
04	B
05	B C D

$\mathcal{D}_1$

tid	transactions
02	F G
03	E F
04	E F G

$\mathcal{D}_2$

$\mathcal{P}$

Hybrid

$\mathcal{D}_1$

tid	transaction
01	A
04	BE
05	BC

$\mathcal{D}_2$

tid	transaction
01	BC
02	FG
05	D

$\mathcal{D}_3$

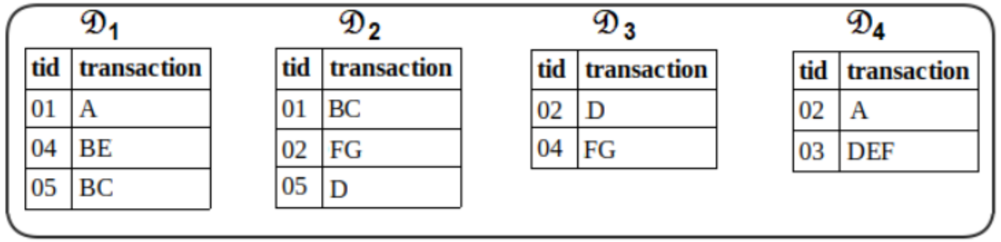
tid	transaction
02	D
04	FG

$\mathcal{D}_4$

tid	transaction
02	A
03	DEF

$\mathcal{P}$

Description of the communication model



$\mathcal{D}_1$

tid	transaction
01	A
04	BE
05	BC

$\mathcal{D}_2$

tid	transaction
01	BC
02	FG
05	D

$\mathcal{D}_3$

tid	transaction
02	D
04	FG

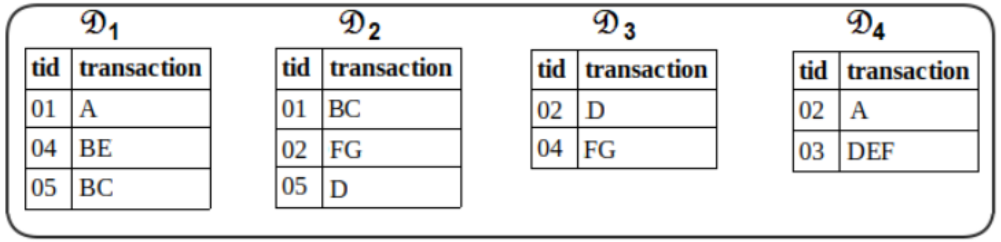
$\mathcal{D}_4$

tid	transaction
02	A
03	DEF

$lengthOf(j, \mathcal{D}_k)$: the number of items in the transaction j contained in the fragment $\mathcal{D}_k$

$itemAt(i, j, \mathcal{D}_k)$: the i^{th} item in the transaction j contained in the fragment $\mathcal{D}_k$

Challenges for pattern sampling in distributed databases



$\mathcal{D}_1$

tid	transaction
01	A
04	BE
05	BC

$\mathcal{D}_2$

tid	transaction
01	BC
02	FG
05	D

$\mathcal{D}_3$

tid	transaction
02	D
04	FG

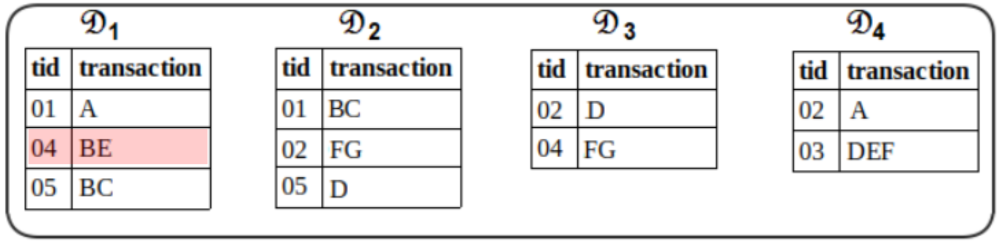
$\mathcal{D}_4$

tid	transaction
02	A
03	DEF

Challenge 1 : Weight the transactions in the transactional distributed database without centralizing any item

Challenge 2 : Decentralize pattern sampling

Challenge 1 : Weight the transactions in the distributed transactional database



$\mathcal{D}_1$

tid	transaction
01	A
04	BE
05	BC

$\mathcal{D}_2$

tid	transaction
01	BC
02	FG
05	D

$\mathcal{D}_3$

tid	transaction
02	D
04	FG

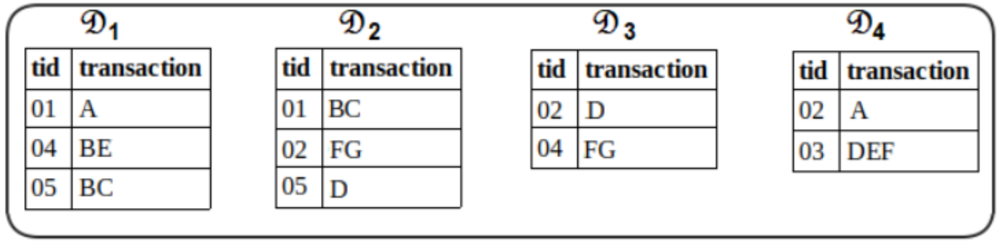
$\mathcal{D}_4$

tid	transaction
02	A
03	DEF

$lengthOf(j, \mathcal{D}_k)$: the number of items in the transaction j contained in the fragment $\mathcal{D}_k$

	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	
	$\mathbb{M}$				$\widetilde{\mathbb{M}}$
01	1	2	0	0	3
02	0	2	1	1	4
03	0	0	0	3	3
04	2	0	2	0	4
05	2	1	0	0	3

Challenge 1 : Weight the transactions in the transactional distributed database



$\mathcal{D}_1$

tid	transaction
01	A
04	BE
05	BC

$\mathcal{D}_2$

tid	transaction
01	BC
02	FG
05	D

$\mathcal{D}_3$

tid	transaction
02	D
04	FG

$\mathcal{D}_4$

tid	transaction
02	A
03	DEF

$lengthOf(j, \mathcal{D}_k)$: the number of items in the transaction j contained in the fragment $\mathcal{D}_k$

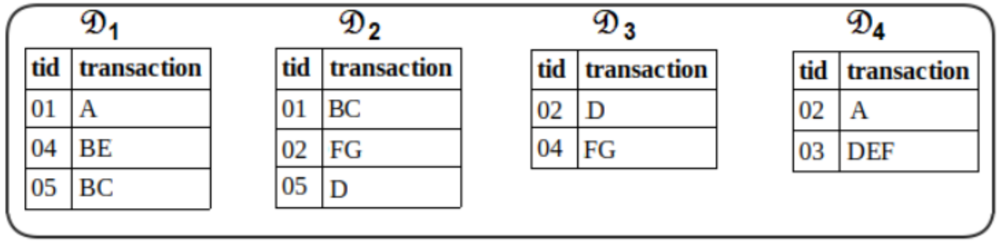
	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$		$u_{\leq 2}$
	$\mathbb{M}$				$\widetilde{\mathbb{M}}$	$\omega_{\leq 2}(j)$
01	1	2	0	0	3	7
02	0	2	1	1	4	11
03	0	0	0	3	3	7
04	2	0	2	0	4	11
05	2	1	0	0	3	7

[Diop et al., 2019]

Norm-based utility : $u(\varphi) = f_u(||\varphi||)$ with $f_u : \mathbb{N} \longrightarrow \mathbb{R}^+$

Class of norm-based utility measures $\mathcal{M}$: $m_u(\varphi, \mathcal{D}) = freq(\varphi, \mathcal{D}) \times f_u(||\varphi||)$

Challenge 2 : Decentralize pattern sampling



$\mathcal{D}_1$

tid	transaction
01	A
04	BE
05	BC

$\mathcal{D}_2$

tid	transaction
01	BC
02	FG
05	D

$\mathcal{D}_3$

tid	transaction
02	D
04	FG

$\mathcal{D}_4$

tid	transaction
02	A
03	DEF

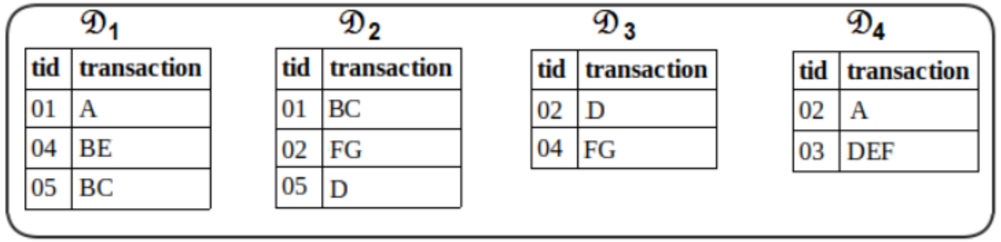
$lengthOf(\mathrm{j}, \mathcal{D}_k)$ **: the number of items in the transaction j contained in the fragment** $\mathcal{D}_k$

	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$		
	$\mathbb{M}$				$\widetilde{\mathbb{M}}$	$\omega_{\leq 2}(j)$
01	1	2	0	0	3	7
02	0	2	1	1	4	11
03	0	0	0	3	3	7
04	2	0	2	0	4	11
05	2	1	0	0	3	7

Step 1 : Draw a transaction identifier j proportionally to $\omega_{\leq 2}(j)$

Tid	$\ell = 0$	$\ell = 1$	$\ell = 2$	$\ell = 3$
03	$\emptyset$	i_1, i_2, i_3	$i_1 i_2, i_1 i_3, i_2 i_3$	$i_1 i_2 i_3$

Challenge 2 : Decentralize pattern sampling



$\mathcal{D}_1$

tid	transaction
01	A
04	BE
05	BC

$\mathcal{D}_2$

tid	transaction
01	BC
02	FG
05	D

$\mathcal{D}_3$

tid	transaction
02	D
04	FG

$\mathcal{D}_4$

tid	transaction
02	A
03	DEF

$lengthOf(\mathbf{j}, \mathcal{D}_k)$: the number of items in the transaction j contained in the fragment $\mathcal{D}_k$

	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$		
	$\mathbb{M}$				$\widetilde{\mathbb{M}}$	$\omega_{\leq 2}(j)$
01	1	2	0	0	3	7
02	0	2	1	1	4	11
03	0	0	0	3	3	7
04	2	0	2	0	4	11
05	2	1	0	0	3	7

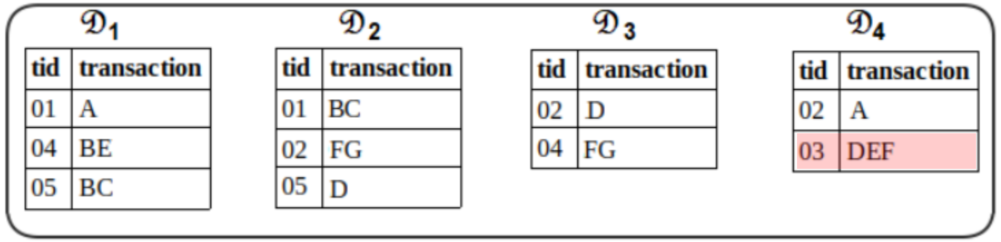
Step 1 : Draw a transaction identifier **j** proportionally to $\omega_{\leq 2}(j)$

Step 2 : 2-1. Draw a norm ℓ
2-2. Draw uniformly a subset of ℓ=2 indexes

$\omega^{0}_{\leq 2} = 1$ $\omega^{1}_{\leq 2} = 3$ $\omega^{2}_{\leq 2} = 3$ $\omega^{3}_{\leq 2} = 0$

Tid	$\ell = \mathbf{0}$	$\ell = \mathbf{1}$	$\ell = \mathbf{2}$	$\ell = \mathbf{3}$
03	$\emptyset$	i_1, i_2, i_3	$\boldsymbol{i_1 i_2}, i_1 i_3, i_2 i_3$	$i_1 i_2 i_3$

Challenge 2 : Decentralize pattern sampling



$\mathcal{D}_1$

tid	transaction
01	A
04	BE
05	BC

$\mathcal{D}_2$

tid	transaction
01	BC
02	FG
05	D

$\mathcal{D}_3$

tid	transaction
02	D
04	FG

$\mathcal{D}_4$

tid	transaction
02	A
03	DEF

$\boldsymbol{lengthOf}$(j, $\mathcal{D}_k$) : the number of items in the transaction j contained in the fragment $\mathcal{D}_k$

$\boldsymbol{itemAt}$(i, j, $\mathcal{D}_k$) : the i^{th} item in the transaction j contained in the fragment $\mathcal{D}_k$

$\boldsymbol{itemAt}(i_1, 3, \mathcal{D}_4)$=D
$\boldsymbol{itemAt}(i_2, 3, \mathcal{D}_4)$=E

	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$		
	$\mathbb{M}$				$\tilde{\mathbb{M}}$	$\omega_{\leq 2}(j)$
01	1	2	0	0	3	7
02	0	2	1	1	4	11
03	0	0	0	3	3	7
04	2	0	2	0	4	11
05	2	1	0	0	3	7

Step 1 : Draw a transaction identifier j proportionally to $\omega_{\leq 2}(j)$

Step 2 : 2-1. Draw a norm ℓ
2-2. Draw uniformly a subset of ℓ=2 indexes

	$\omega_{\leq 2}^{0} = 1$	$\omega_{\leq 2}^{1} = 3$	$\omega_{\leq 2}^{2} = 3$	$\omega_{\leq 2}^{3} = 0$
Tid	$\ell = 0$	$\ell = 1$	$\ell = 2$	$\ell = 3$
03	$\emptyset$	i_1, i_2, i_3	$\boldsymbol{i_1 i_2}, i_1 i_3, i_2 i_3$	$i_1 i_2 i_3$

Experimental protocol for DDSampling

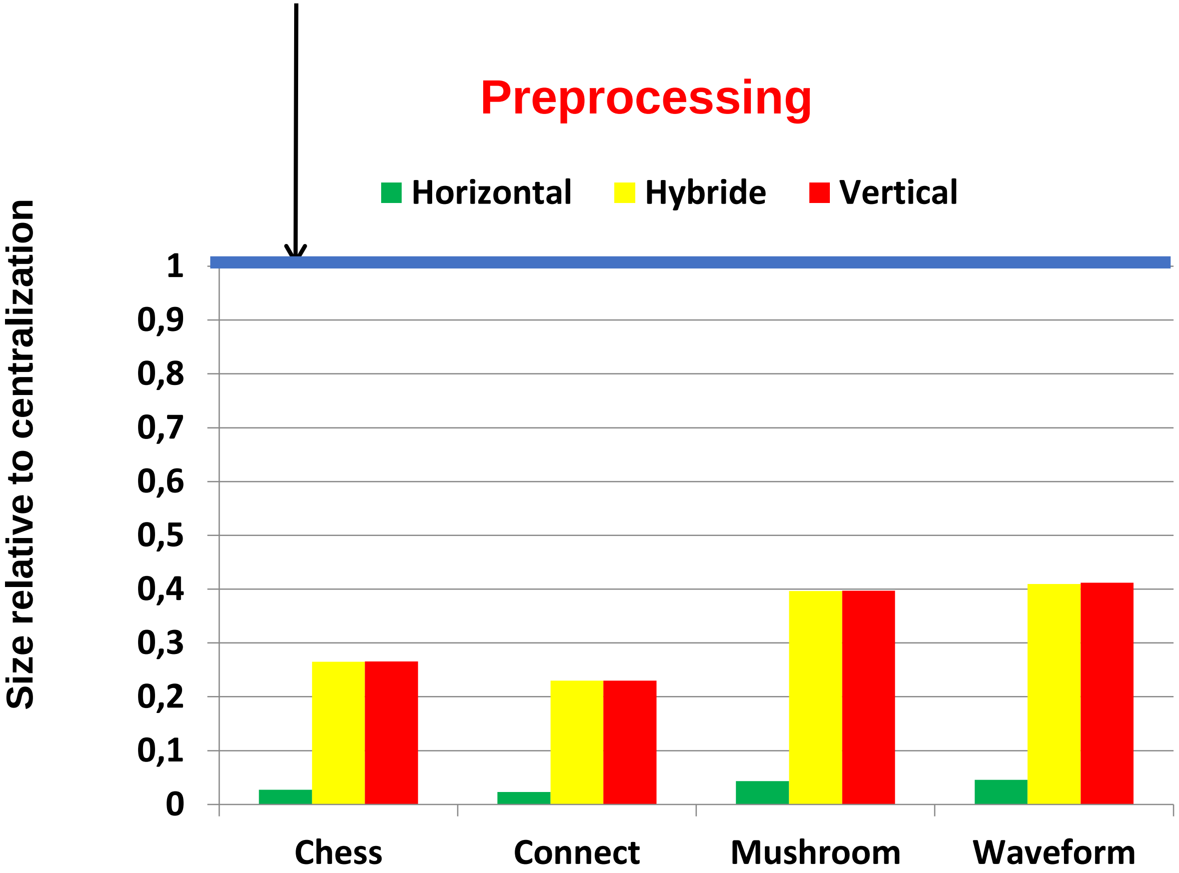
- Artificial distributed databases : *Uniform fragmentation in K = 10 fragments*

- **Horizontal partitioning :** every transaction is described in only one fragment
- **Vertical partitioning :** every item is present in only one fragment
- **Hybrid partitioning :** neither horizontal nor vertical

- Utility : $\boldsymbol{u}_{[\boldsymbol{\mu}..\boldsymbol{M}]} = u_{\geq\mu} \times u_{\leq M} \;\; with\; \mu = 1\; and\; M \in [1..5]$

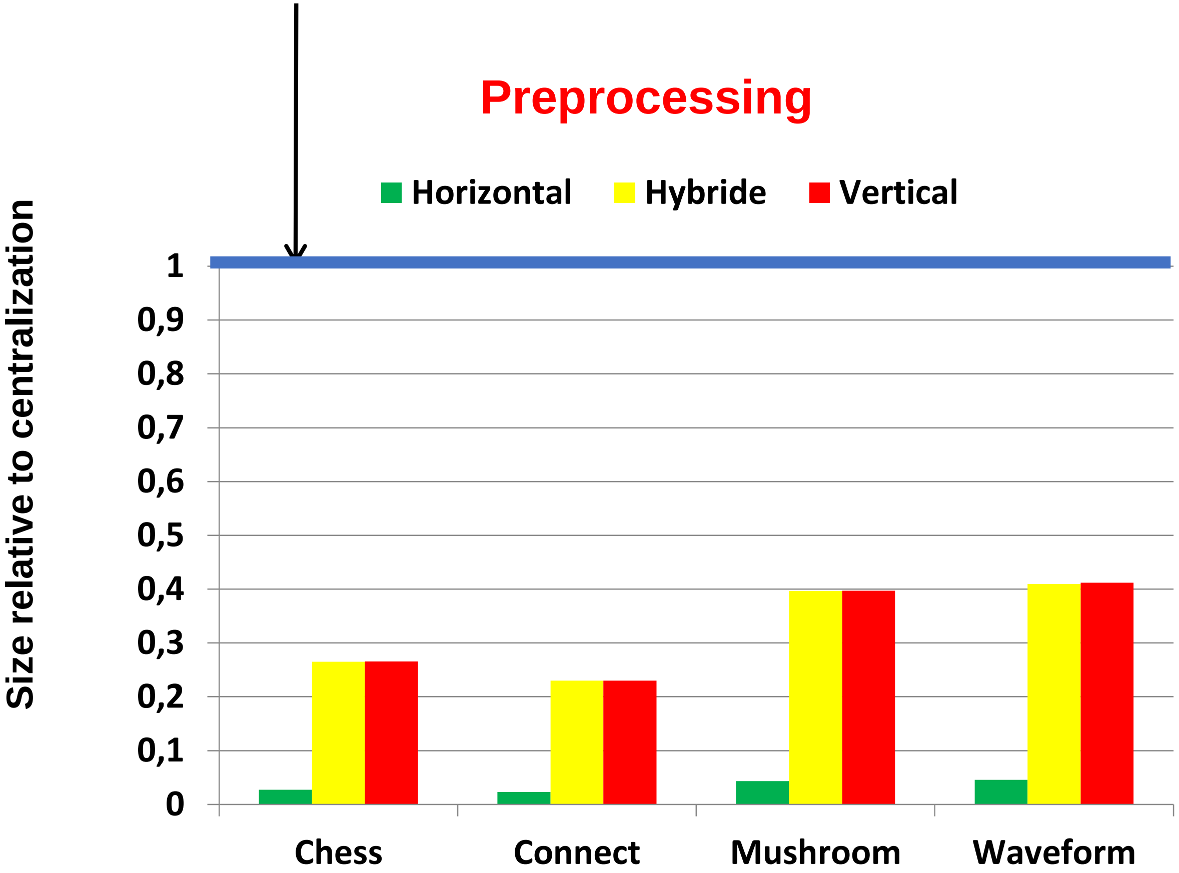
DDSampling communication cost

Baseline : Centralisation cost.



DDSampling communication cost

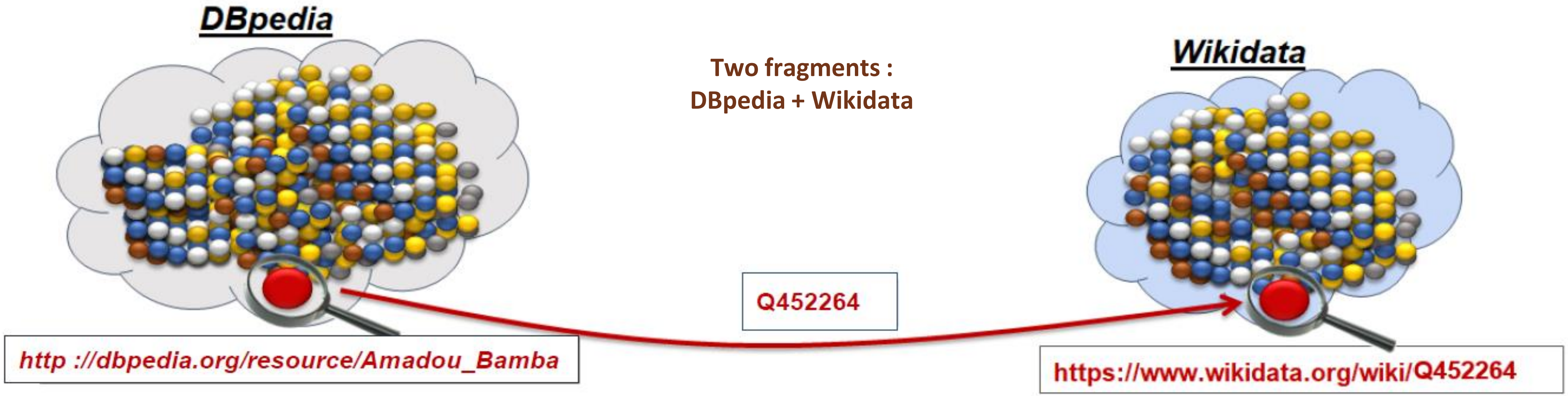
Baseline : **Centralisation cost.**



	Vertical : Nb patterns$(M = 5)$ before reaching the cost of centralization	Cost of centralization in nb $itemAt$
Chess	**17 976**	**118 252**
Connect	**460 165**	**2 904 951**
Mushroom	**23 973**	**186 852**
Waveform	**13 820**	**110 000**

- **DDSampling is very parsimonious in communication cost**

Outlier entities detection in semantic Web databases



DBpedia

Two fragments :
DBpedia + Wikidata

Wikidata

Q452264

http ://dbpedia.org/resource/Amadou_Bamba

https://www.wikidata.org/wiki/Q452264

Class Person contains 772,432 entities described by DBpedia and Wikidata

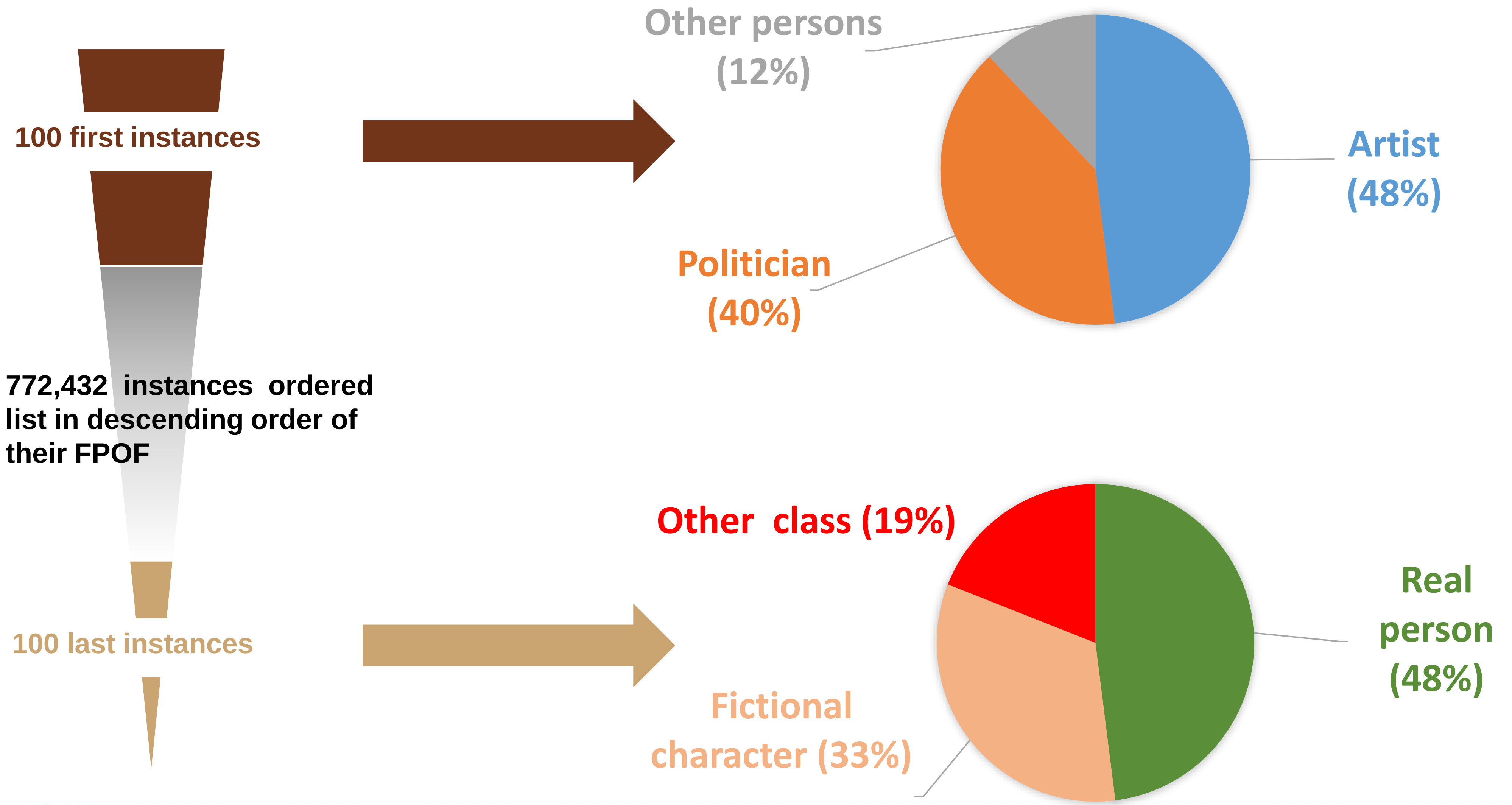
Degree of aberration of each instance

- **FPOF** *[He et al., 2005]* : Frequent Pattern Outlier Factor

- **FPOF** with utility : $u_{\geq 1} \times u_{\leq 3}$

- Approximate the FPOF with a sample of 10,000 patterns *[Giacometti & Soulet, 2016]* under norm constraints

Qualitative evaluation of outlier detection on Person



100 first instances

772,432 instances ordered list in descending order of their FPOF

100 last instances

The sampled patterns are effective in detecting outliers in real life databases, even if they have the long tail curse.

Conclusion

- A Two-Step random procedure on distributed transactional databases
  - First sampling method on distributed transactional databases
  - Generic algorithm ({horizontal, **vertical**, **hybrid**} × {**area**, **frequency**, **norm constraints**})
  - Very accurate to detect outliers in transactional databases
- Perspectives
  - Weight correction mechanism to counterbalance the breakdown of a site
  - Replace the exact drawing of transactions with a stochastic method to minimize the cost

Try it https://github.com/DDSamplingRDF/ddsampling !

université de TOURS

UNIVERSITE GASTON BERGER

L'excellence au service du développement

Thank you for your attention!