Génération de bases de données de transactions synthétiques: Vers la prises en compte des bordures

Réalisé avec Didier Devaurs
laboratoire Liris, Lyon

Contexte: extraction des fréquents

- [AIS93] + [AS94]
- Une multitude de contributions depuis:
  - Représentations condensées: fermés, clés, maximaux
  - Structures de données: Prefix-Tree, FP-Tree, Patricia-Tree, tables de hachages, vertical bitmaps etc...
  - Techniques de parcours: par niveau, en profondeur, avec des sauts, aléatoires
  - Plusieurs implementations d'un même algo...
- FIMI '03 '04 : 21 programmes

FIMI: évaluer les algos

- Comparer, comprendre, prévoir, classer, progresser
- Nécessité reconnue : FIMI (Bayardo, Goethals, Zaki)
  - Comparaison impartiale
  - Constats positifs
    - impossible de déterminer « le meilleur » algo
    - Les meilleurs temps sont impressionnants
  - Constats négatifs
    - Manque de jeux de données
    - Impossible de tirer des conclusions interessantes

FIMI: Jeux d'essais

- 14 BD de transactions :

	#Articles	#Transact.	Taille Moyenne
Accident	468	340183	33,8
BMS1	497	59602	2,5
BMS2	3341	77512	5,6
BMSPOS	1658	515597	7,5
Chess	75	3196	37
Connect	129	67557	43
Kozarak	41270	99002	8,1
Mushroom	119	8124	23
Pumsb	2088	49046	50,5
Pumsb*	2113	4905	74
Retail	16469	88162	10,3
T10I5N1KP5KC0.25D200k	956	200000	10,3
T20I10N1KP5KC0.25D200k	979	200000	20,1
T30I15N1KP5KC0.25D200k	987	200000	29,7

Problèmes

- Petites transactions en moyenne
  - Les fréquents sont « bas » dans le treillis des parties
  - Beaucoup d'algos ne sont pas mis en difficulté
  - Mettre un petit support: d'où proviennent les échecs ?
- Les bases tiennent en mémoire
  - Diminuer la mémoire pénaliserai certains algorithmes
- Doit-on se focaliser sur des jeux réels ?
- Jeux synthétiques: Système Quest, d'IBM.
  - #transactions, #items, taille moyenne

Objectifs

- Générer des bases synthétiques, pour de réelles campagnes de tests.
- Qu'est ce qui est important ?
  - Nb de candidats générés, nb accès aux données, taille de la sortie, niveau d'apparition des fréquents...
  - Le Nb article est inintéressant
  - Nb Transactions: doit pouvoir varier de façon indépendante
  - Taille moyenne: trop grossier

Prise en compte des bordures (1)

- Bordure positive : les plus grands fréquents
- Bordure négative : les plus petits non-fréquents
- Equivalence entre :
  - Une bordure positive
  - Une bordure négative
  - Un ensemble de fréquents
- La plupart des algos parcourent au moins ces deux bordures

Prise en compte des bordures (2)

- Prendre en entrée une bordure (+ ou -) ?
  - Difficile de saisir une bordure à la main
  - Générer automatiquement: en fonction de quel critère ?
- La distribution d'une bordure:
  - Nb d'éléments dans la bordure à chaque niveau.
- Caractérise plus finement les jeux de données
  - Attention: ce n'est pas une caractérisation exacte

Quelques préliminaires (1/3)

- Lexicographique V.S. Colex
  - Ex.: Soient ABCDE les articles
    Lex: ABC ABD ABE ACD ACE ADE BCD BCE BDE CDE
    Colex: ABC ABD ACD BCD ABE ACE BCE ADE BDE CDE
  - Colex: un nouvel article est introduit lorsque toutes les combinaisons sont épuisées.
  - Conséquence : à un motif correspond un rang unique.
  - Ex: Si on a 6 articles ABCDEF
    Lex: ABC ABD ABE ABF ACD ACE ACF ADE ADF AEF BCD ...
    Colex: ABC ABD ACD BCD ABE ACE BCE ADE BDE CDE ABF ...

Quelques préliminaires (1/3)

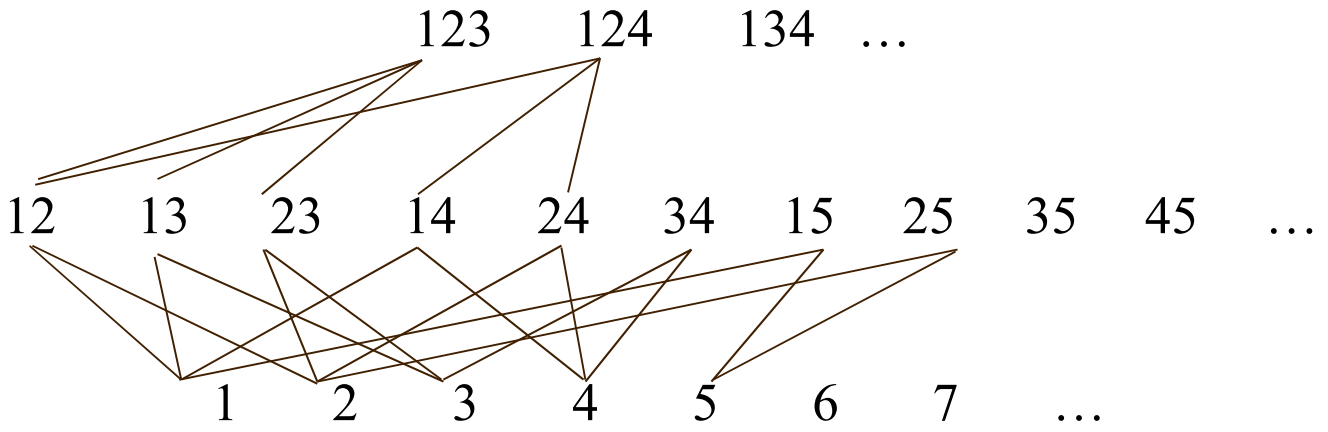
- Lexicographique V.S. Colex
  - Ex.: Soient ABCDE les articles
    Lex: ABC ABD ABE ACD ACE ADE BCD BCE BDE CDE
    Colex: ABC ABD ACD BCD ABE ACE BCE ADE BDE CDE
  - Colex: un nouvel article est introduit lorsque toutes les combinaisons sont épuisées.
  - Conséquence : à un motif correspond un rang unique.
  - Ex: Si on a 6 articles ABCDEF
    Lex: ABC ABD ABE ABF ACD ACE ACF ADE ADF AEF BCD ...
    Colex: ABC ABD ACD BCD ABE ACE BCE ADE BDE CDE ABF ...

Quelques préliminaires (1/3)

- Lexicographique V.S. Colex
  - Ex.: Soient ABCDE les articles
    Lex: ABC ABD ABE ACD ACE ADE BCD BCE BDE CDE
    Colex: ABC ABD ACD BCD ABE ACE BCE ADE BDE CDE
  - Colex: un nouvel article est introduit lorsque toutes les combinaisons sont épuisées.
  - Conséquence : à un motif correspond un rang unique.
  - Ex: Si on a 6 articles ABCDEF
    Lex: ABC ABD ABE ABF ACD ACE ACF ADE ADF AEF BCD ...
    Colex: ABC ABD ACD BCD ABE ACE BCE ADE BDE CDE ABF ...

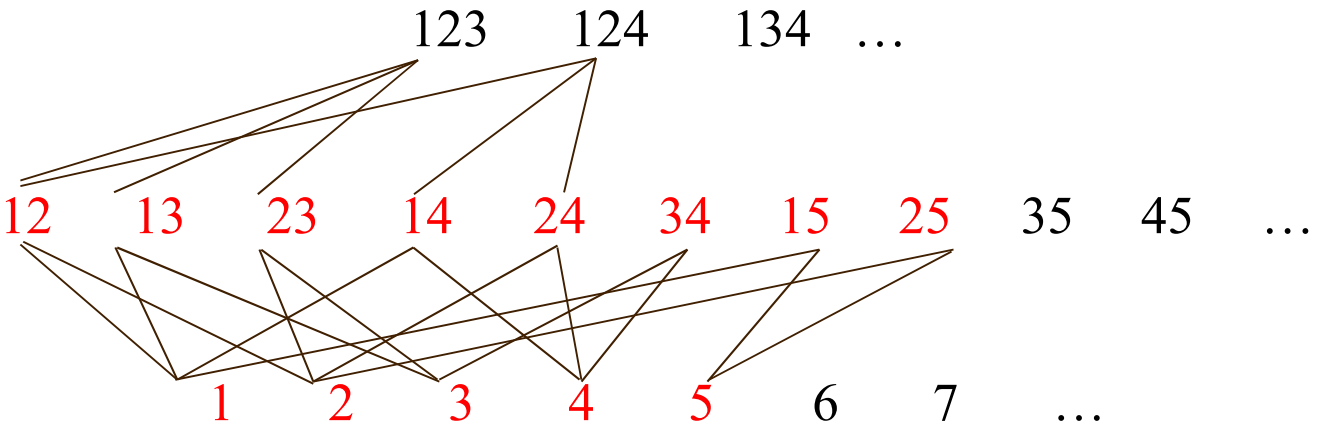
Quelques préliminaires (2/3)

- Calcul des motifs induits
  - Fréquents déduits d'autres fréquents par monotonie
  - Facilité par l'ordre Colex dans certains cas



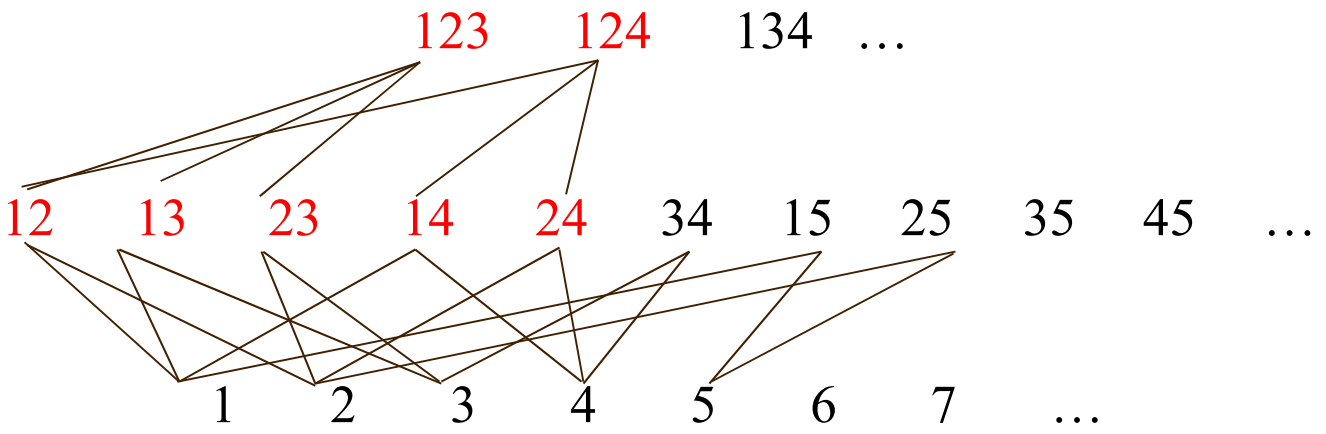
Quelques préliminaires (2/3)

- Calcul des motifs induits
  - Fréquents déduits d'autres fréquents par monotonie
  - Facilité par l'ordre Colex dans certains cas



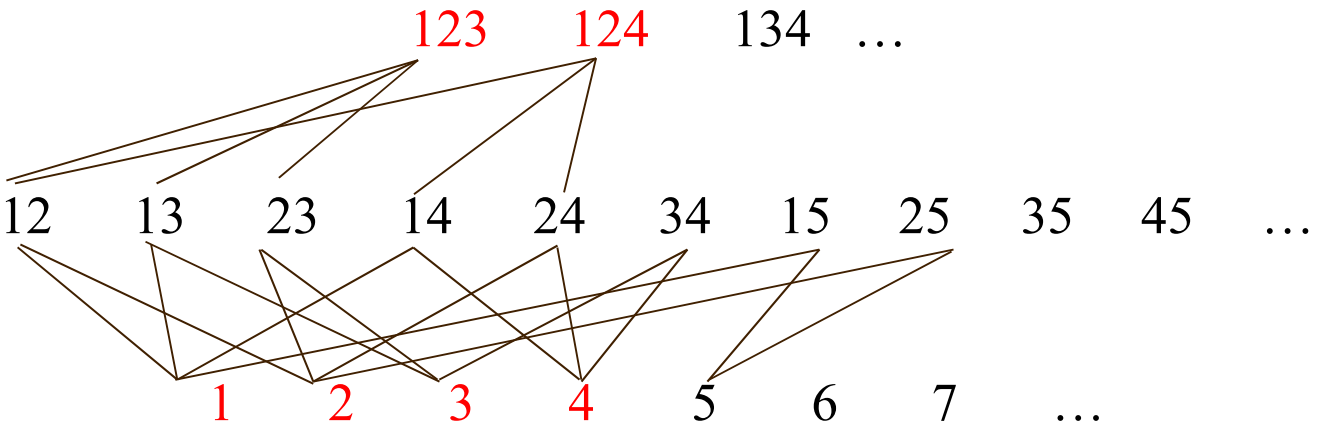
Quelques préliminaires (2/3)

- Calcul des motifs induits
  - Fréquents déduits d'autres fréquents par monotonie
  - Facilité par l'ordre Colex dans certains cas



Quelques préliminaires (2/3)

- Calcul des motifs induits
  - Fréquents déduits d'autres fréquents par monotonie
  - Facilité par l'ordre Colex dans certains cas



Quelques préliminaires (3/3)

Soit F_k les n premiers fréquents de taille k.

- Le nombre de fréquents induits au niveau i<k est calculable efficacement.
- Soit $(F_j)_{j=1..K}$ les n_j premiers fréquents de taille j. Alors les fréquents induits au niveau i<j sont calculables efficacements.

Une méthode de génération (1)

- Proposée par [Maniatty, Ramesh, Zaki] PODS'03
- Entrée : une distribution S de bordure positive
- Sortie : une BD de transactions *d*, et un support *minsup*
  - La bordure positive des motifs de support > *minsup* dans *d* possède la distribution S.
- En fait, la méthode est plus générale: prend en entrée une **séquence de séquences**.

Une méthode de génération (2)

- Deux étapes:
  - génération de BD^+
  - Génération de la base à partir de BD^+
- ex.: $S = < 2, 3, 2 >$

123 124

Une méthode de génération (2)

- Deux étapes:
  - génération de BD^+
  - Génération de la base à partir de BD^+
- ex.: $S = < 2, 3, 2 >$

123 124

12 13 23 14 24

Une méthode de génération (2)

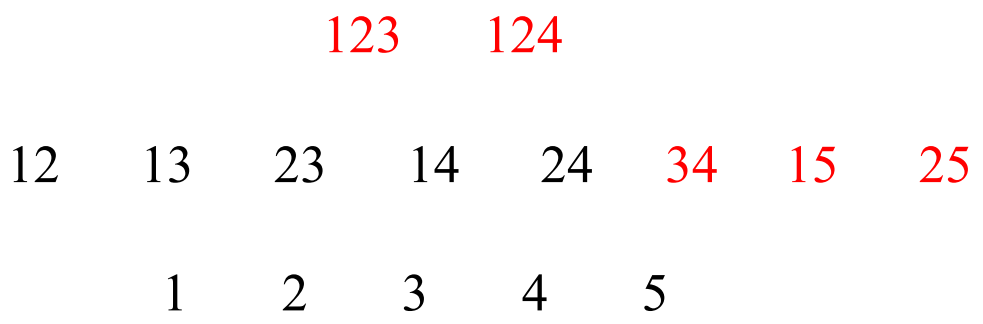
- Deux étapes:
  - génération de BD^+
  - Génération de la base à partir de BD^+
- ex.: $S = < 2, 3, 2 >$

123 124

12 13 23 14 24 34 15 25

Une méthode de génération (2)

- Deux étapes:
  - génération de BD^+
  - Génération de la base à partir de BD^+
- ex.: $S = < 2, 3, 2 >$



Une méthode de génération (2)

- Deux étapes:
  - génération de BD^+
  - Génération de la base à partir de BD^+
- ex.: $S = < 2, 3, 2 >$

123 124

12 13 23 14 24 34 15 25

1 2 3 4 5 6 7

Une méthode de génération (3)

- La génération de la base est ensuite triviale

6
7
3 4
1 5
2 5
1 2 3
1 2 4

Support : $\pi = \frac{1}{7}$

Une méthode de génération (4)

- Deux résultats importants:
  - La méthode fonctionne pour toute séquence en entrée
  - Le nombe d'articles utilisés est minimal
- Expérimentations:
  - Calculer la bordure positive des bases connues
  - Reproduire ces bases de façon synthétiques.
- Mais les BD sont-elles vraiment « similaires » en terme de complexité ?

La bordure négative synthétique (1)

- Quelle est la bordure négative des bases générées par [Maniatty, Ramesh, Zaki] ?
- Méthode proposée:
  - Entrée: une séquence de BD+
  - Sortie: la séquence de BD- correspondante dans la méthode de [Maniatty, Ramesh, Zaki]
- Idée: soient F les n premiers motifs de taille k
  - On peut calculer les ensembles qui les induisent au niveau k+1

La bordure négative synthétique (2)

- Ex: $< Bd^{+}(F) > = < 2, 3, 2 >$

123 124

12 13 23 14 24 34 15 25

1 2 3 4 5 6 7

La bordure négative synthétique (2)

- Ex: $< Bd^{+}(F) > = < 2, 3, 2 >$

123 124

12 13 23 14 24 34 15 25 35 45 16 26 36 46 56 17 27 37 47 57 67

1 2 3 4 5 6 7

La bordure négative synthétique (2)

- Ex: $< Bd^{+}(F) > = < 2, 3, 2 >$

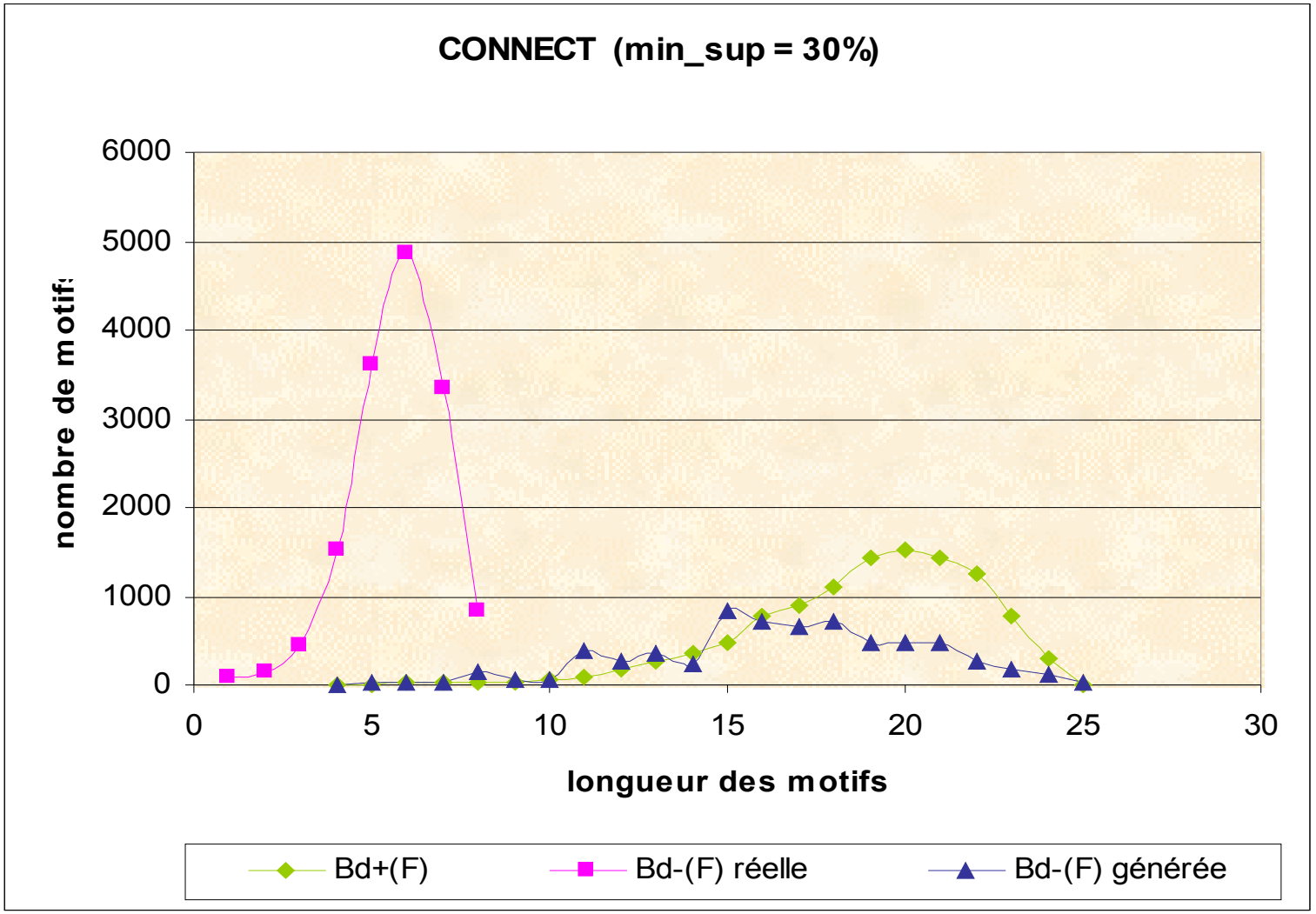
123 124 134 234 125

12 13 23 14 24 34 15 25 35 45 16 26 36 46 56 17 27 37 47 57 67

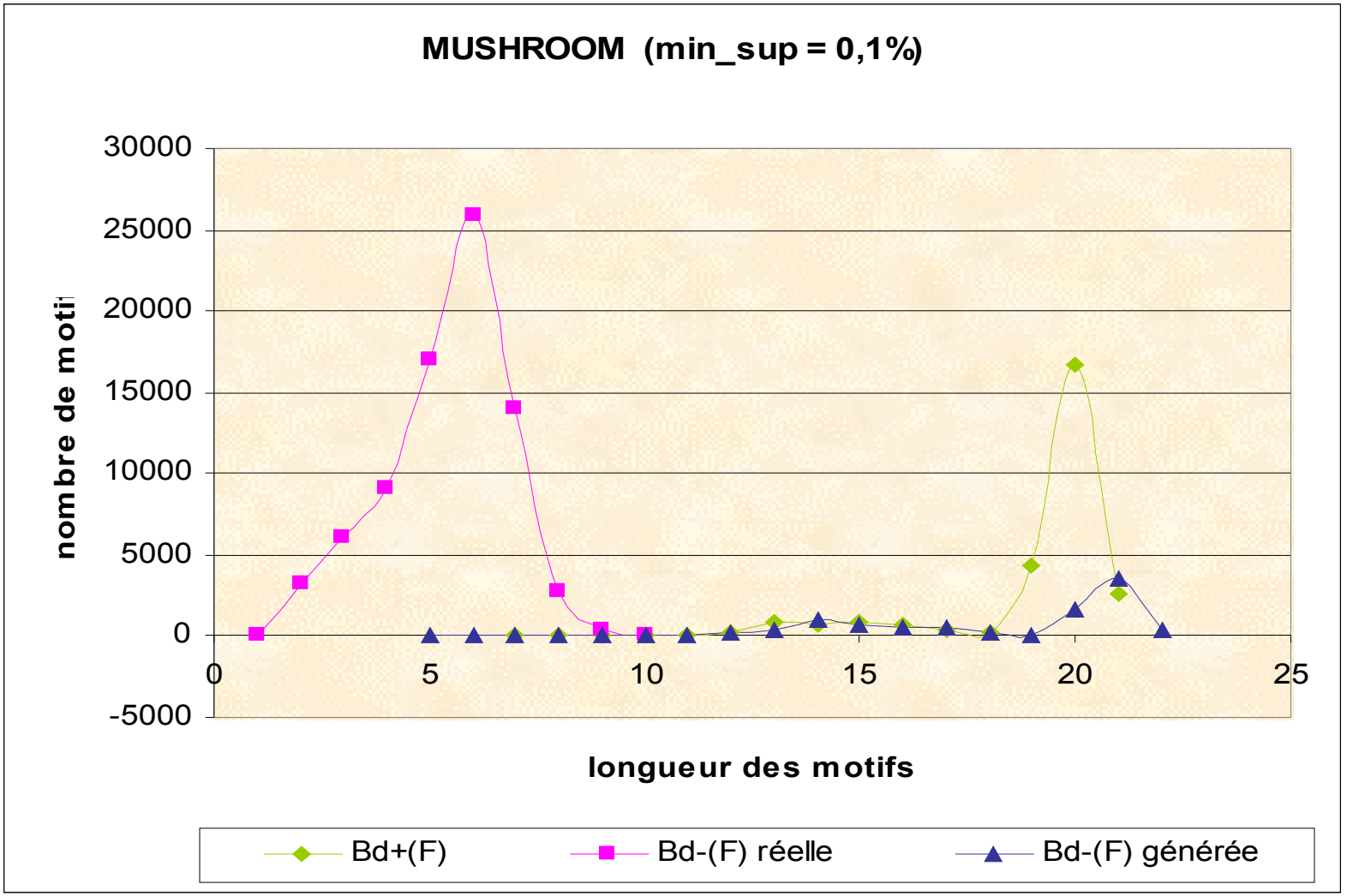
1 2 3 4 5 6 7

- Donc: $< Bd^{-}(F) > = < 0, 13, 3 >$
- Au plus, la bordure négative dépasse d'un niveau la bordure positive

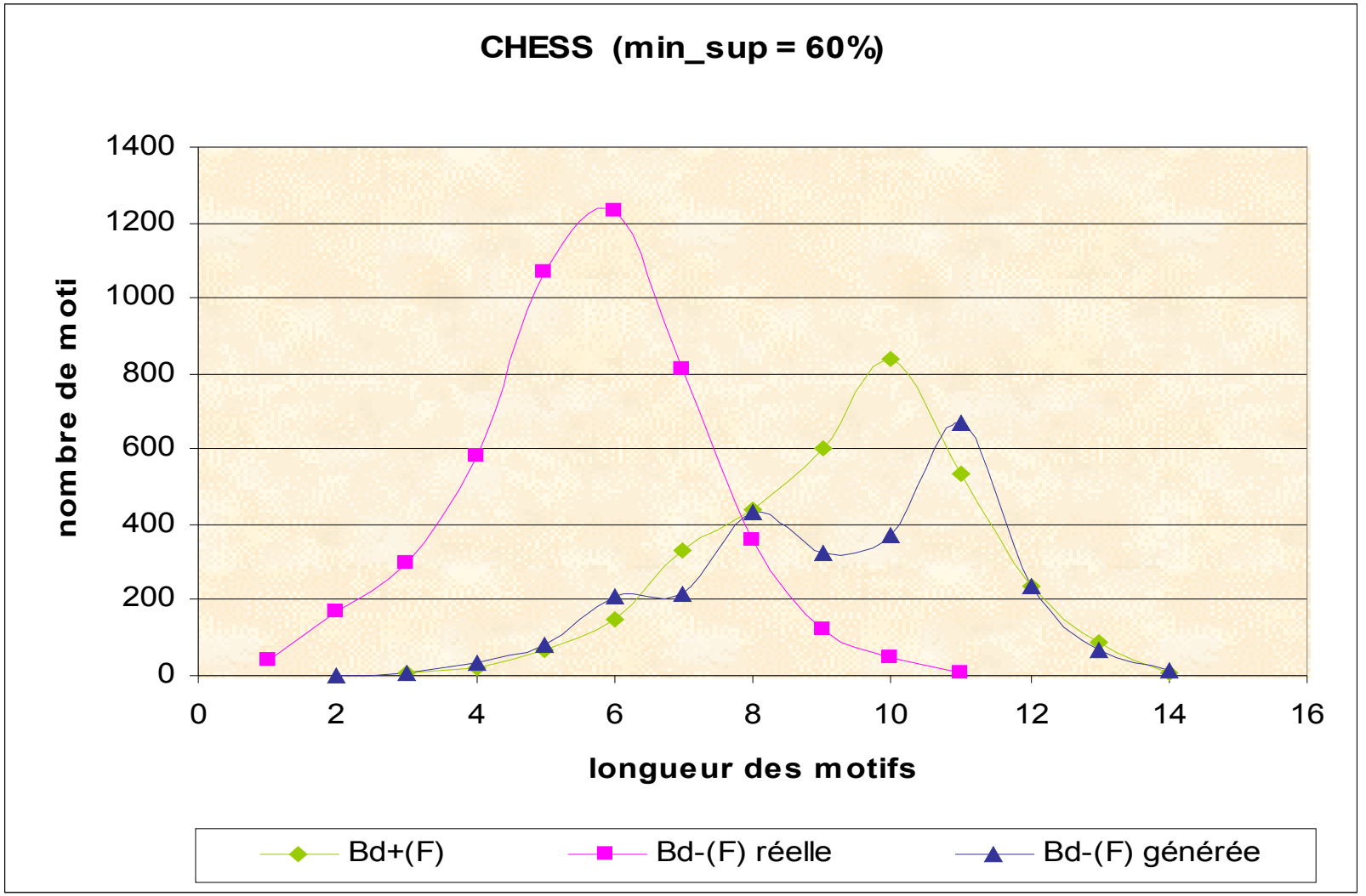
Comparaison avec les BD existantes



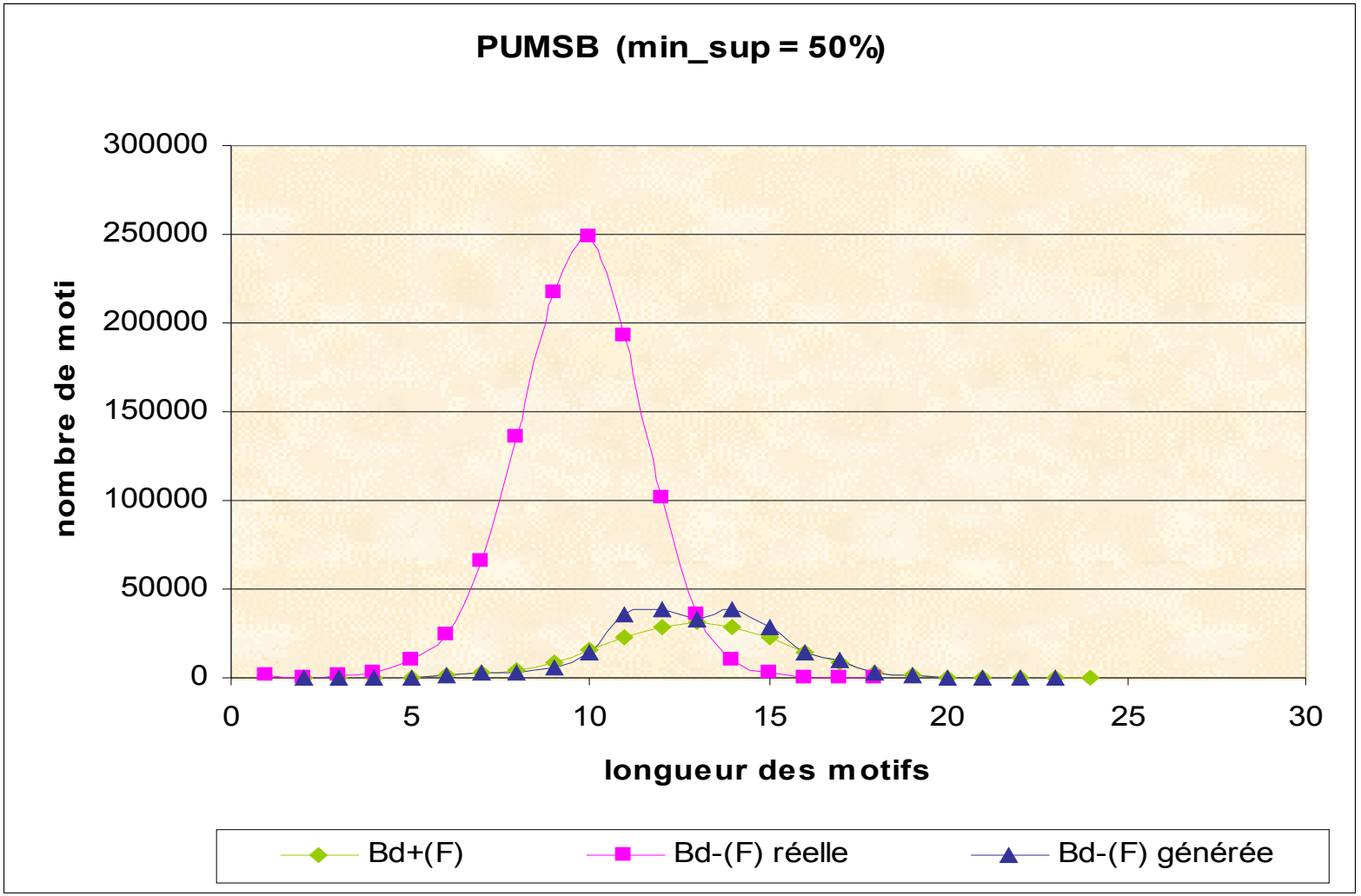
Comparaison avec les BD existantes



Comparaison avec les BD existantes



Comparaison avec les BD existantes



Prendre BD^- en entrée (1)

- *Entrée*: une séquence pour BD^-
- *Sortie*: Une BD et un support représentatif
- A chaque niveau, on utilise la même disposition des motifs que dans [Maniatty, Ramesh, Zaki, 2002]
  - De gauche à droite, à chaque niveau, on trouve:
    - Les fréquents non maximaux
    - Les motifs de BD+
    - Les motifs de BD-
    - Les non fréquents non minimaux
- Permet d'utiliser les propriétés intéressantes

Prendre BD^- en entrée: exemple

- Ex.: <BD-> = <0,2,3,1>
  - Etape 1: recherche du nombre minimal d'articles

1234

Prendre BD^- en entrée: exemple

- Ex.: <BD-> = <0,2,3,1>
  - Etape 1: recherche du nombre minimal d'articles

1234

123 124 134 234

Prendre BD^- en entrée: exemple

- Ex.: <BD-> = <0,2,3,1>
  - Etape 1: recherche du nombre minimal d'articles

1234

123 124 134 234 125 135 235

Prendre BD^- en entrée: exemple

- Ex.: <BD-> = <0,2,3,1>
  - Etape 1: recherche du nombre minimal d'articles

1234

123 124 134 234 125 135 235

12 13 23 14 24 34 15 25 35

Prendre BD^- en entrée: exemple

- Ex.: <BD-> = <0,2,3,1>
  - Etape 1: recherche du nombre minimal d'articles

1234

123 124 134 234 125 135 235

12 13 23 14 24 34 15 25 35 45 16

Prendre BD^- en entrée: exemple

- Ex.: <BD-> = <0,2,3,1>
  - Etape 1: recherche du nombre minimal d'articles

1234

123 124 134 234 125 135 235

12 13 23 14 24 34 15 25 35 45 16

1 2 3 4 5 6

Prendre BD^- en entrée: exemple

- Ex.: <BD-> = <0,2,3,1>
  - Etape 1: recherche du nombre minimal d'articles

1234

123 124 134 234 125 135 235

12 13 23 14 24 34 15 25 35 45 16

1 2 3 4 5 6

- 6 articles seront nécessaires
- On ne peut pas s'arrêter ici !

Prendre BD^- en entrée: exemple

- Ex.: <BD-> = <0,2,3,1>
  - Etape 2: on détermine les éléments de BD-

1 2 3 4 5 6

Prendre BD^- en entrée: exemple

- Ex.: <BD-> = <0,2,3,1>
  - Etape 2: on détermine les éléments de BD-

12 13 23 14 24 34 15 25 35 45 16 26 36 46 56

1 2 3 4 5 6

Prendre BD^- en entrée: exemple

- Ex.: <BD-> = <0,2,3,1>
  - Etape 2: on détermine les éléments de BD-

12 13 23 14 24 34 15 25 35 45 16 26 36 46 56

1 2 3 4 5 6

Prendre BD^- en entrée: exemple

- Ex.: <BD-> = <0,2,3,1>
  - Etape 2: on détermine les éléments de BD-

123 124 134 234 125 135 235 245 345 126 136 236

12 13 23 14 24 34 15 25 35 45 16 26 36 46 56

1 2 3 4 5 6

Prendre BD^- en entrée: exemple

- Ex.: <BD-> = <0,2,3,1>
  - Etape 2: on détermine les éléments de BD-

123 124 134 234 125 135 235 245 345 126 136 236

12 13 23 14 24 34 15 25 35 45 16 26 36 46 56

1 2 3 4 5 6

Prendre BD$^-$ en entrée: exemple

- Ex.: <BD-> = <0,2,3,1>
  - Etape 2: on détermine les éléments de BD-

1234 1235 1245 1345 2345

123 124 134 234 125 135 235 245 345 126 136 236

12 13 23 14 24 34 15 25 35 45 16 26 36 46 56

1 2 3 4 5 6

Prendre BD$^-$ en entrée: exemple

- Ex.: <BD-> = <0,2,3,1>
  - Etape 2: on détermine les éléments de BD-

1234 1235 1245 1345 2345

123 124 134 234 125 135 235 245 345 126 136 236

12 13 23 14 24 34 15 25 35 45 16 26 36 46 56

1 2 3 4 5 6

Prendre BD^- en entrée: exemple

- Ex.: <BD-> = <0,2,3,1>
  - Etape 3: on détermine les éléments de BD+

1234 1235 1245 1345 2345

123 124 134 234 125 135 235 245 345 126 136 236

12 13 23 14 24 34 15 25 35 45 16 26 36 46 56

1 2 3 4 5 6

Prendre BD$^-$ en entrée: exemple

- Ex.: <BD-> = <0,2,3,1>
  - Etape 3: on détermine les éléments de BD+

1234 1235 1245 1345 2345

123 124 134 234 125 135 235 245 345 126 136 236

12 13 23 14 24 34 15 25 35 45 16 26 36 46 56

1 2 3 4 5 6

Prendre BD^- en entrée: exemple

- Ex.: <BD-> = <0,2,3,1>
  - Etape 3: on détermine les éléments de BD+

1234 1235 1245 1345 2345

123 124 134 234 125 135 235 245 345 126 136 236

12 13 23 14 24 34 15 25 35 45 16 26 36 46 56

1 2 3 4 5 6

Prendre BD^- en entrée: exemple

- Ex.: <BD-> = <0,2,3,1>
  - Etape 3: on détermine les éléments de BD+

1234 1235 1245 1345 2345

123 124 134 234 125 135 235 245 345 126 136 236

12 13 23 14 24 34 15 25 35 45 16 26 36 46 56

1 2 3 4 5 6

- BD+ = {1234, 1235, 1245, 1345, 16, 26, 36}

Prendre BD^- en entrée: exemple

1 2 3 4
1 2 3 5
1 2 4 5
1 3 4 5
1 6
2 6
3 6

Support : $\pi = \frac{1}{7}$

Conjectures

- Pas encore de preuve formelle
- Résultats à venir:
  - La méthode est juste :-)
  - Toute séquence en entrée est acceptable
  - Le nombre d'articles est minimal

Perspectives (1)

- Prendre en compte les deux bordures en entrée:
  - Permet de mieux diversifier les espaces de recherche
  - Toutes les distributions sont-elles réalisables ?
    - Oui, mais avec une organisation différente
    - On perd quelques bonnes propriétés

Perspectives (2)

- Considérer d'autres caractéristiques décisives des jeux de données:
  - La distribution des bordures des clés (générateurs)
  - La distribution des fermés
  - La « taille » des classes d'équivalences
- ... ou d'autres contraintes que la fréquence

Perspectives (3)

- Prendre en entrée un ensemble de règles d'associations, ou des caractéristiques des règles
  - Notion de BD de transaction d' « Armstrong »
  - Difficulté: prendre en compte une mesure d'intérêt des règles, comme la confiance
  - Il faut certainement retrouver les fermés fréquents avant de générer.

Perspectives (4)

- Effectuer des tests...
  - Génération de batteries de jeux d'essais pertinents
  - Fournir des benchmarks à la communauté
  - Trouver les points faibles et points forts des algos
  - Déterminer des classifications des BD
- Elaborer des stratégies adaptatives en fonction des données
  - Détecter rapidement les configurations