Les tables de nomenclatures

Définition et Identification

Julien Delplanque, Olivier Auverlot, Anne Etien et Nicolas Anquetil

{prénom}.{nom}@inria.fr, olivier.auverlot@univ-lille1.fr

Université de Lille

Inria
inventeurs du monde numérique

RMod
https://rmod.inria.fr/web/

Plan

1. Propriétés et motivation
2. Cas d’étude : AppSI
    1. Analyse des données structurelles
    2. Analyse des données d’évolution
    3. Construction d’un modèle de classification des tables de nomenclature
3. Conclusion et travaux futurs

Intuition

Tables ayant pour finalité de rassembler et d’apporter des informations complémentaires aux lignes des tables constituant le coeur de la base.

id	nom	prénom	civilité_id
1	Delplanque	Julien	1
2	Auverlot	Olivier	1
3	Etien	Anne	2
4	Anquetil	Nicolas	1

personnes

id	nom	abréviation
1	Monsieur	Mr
2	Madame	Mme

civilités

Propriétés

(proposition)

1. Chaque ligne est **identifiable de façon unique**.

2. Chaque ligne est **unique**.

3. Souvent **référencée par d'autres tables** et **référence rarement** des tables via des contraintes de clés étrangères.

4. **Évolue peu** et majoritairement d'ajouts de lignes, rarement de modifications et encore plus rarement de suppressions.

5. Certaines **contraintes** d'intégrité référentielle sont **incompatible**.

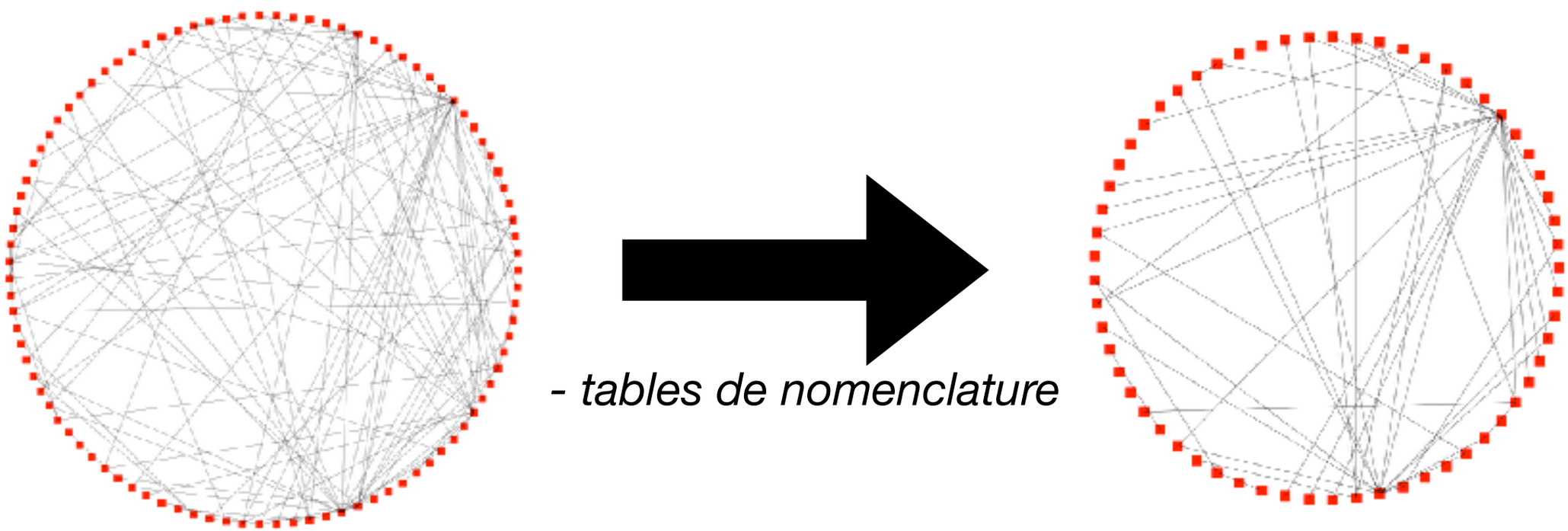
Compatibilité des contraintes d'intégrité référentielles

	UPDATE	DELETE
NO ACTION/RESTRICT	X	X
CASCADE	X	
SET NULL		

« X » indique que la contrainte est compatible

Motivations

- Mise en avant des tables constituant le **noyau** de la base de données



Motivations

- Mise en avant des tables constituant le **noyau** de la base de données
- Selection d’une méthode d’**indexation optimisée**
- **Localisation des littéraux** dans les requêtes SQL
- Création d’une **nouvelle instance** de base de données

Cas d’étude: AppSI

- Base de données PostgreSQL
- Conçue et utilisée par le Pôle Informatique et Technique du Laboratoire CRIStAL (Université de Lille)
- **97** tables, **63** vues, **64** fonctions PL/pgSQL et **20** triggers
- **21** extractions contenant le schéma SQL et les données anonymisées (10 novembre 2017 - 13 mars 2018)

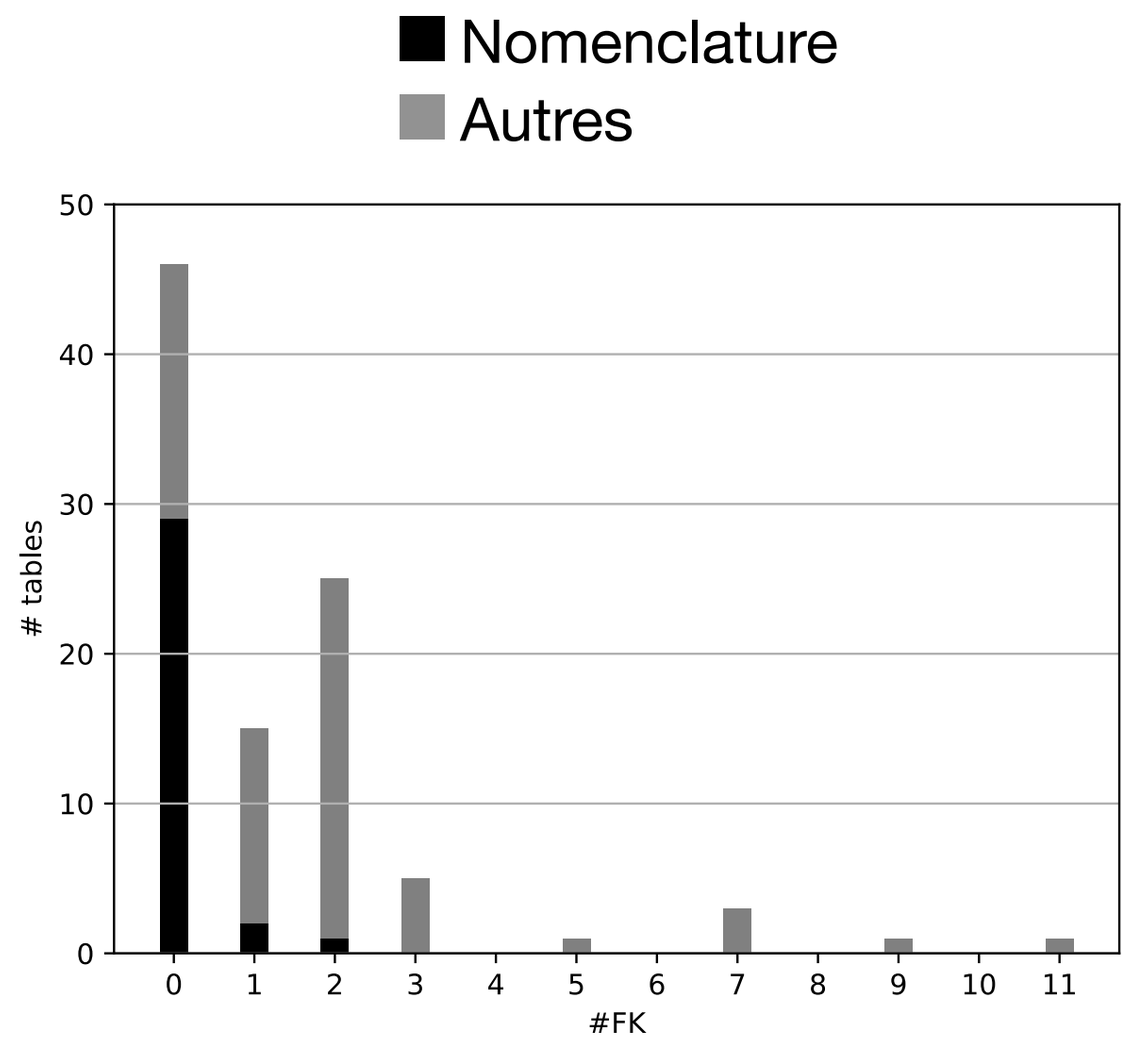
Métriques

- d : Nombre de lignes dupliquées dans une table
- #PK : nombre de clés primaires dans une table
- #FK : nombre de clés étrangères dans une table
- T+, T-, T~ : métriques d'évolution (voir plus loin)
- Respect ou non de la propriété 5.
- estNomenclature? : avis de l'architecte de la BDD

Données structurelles

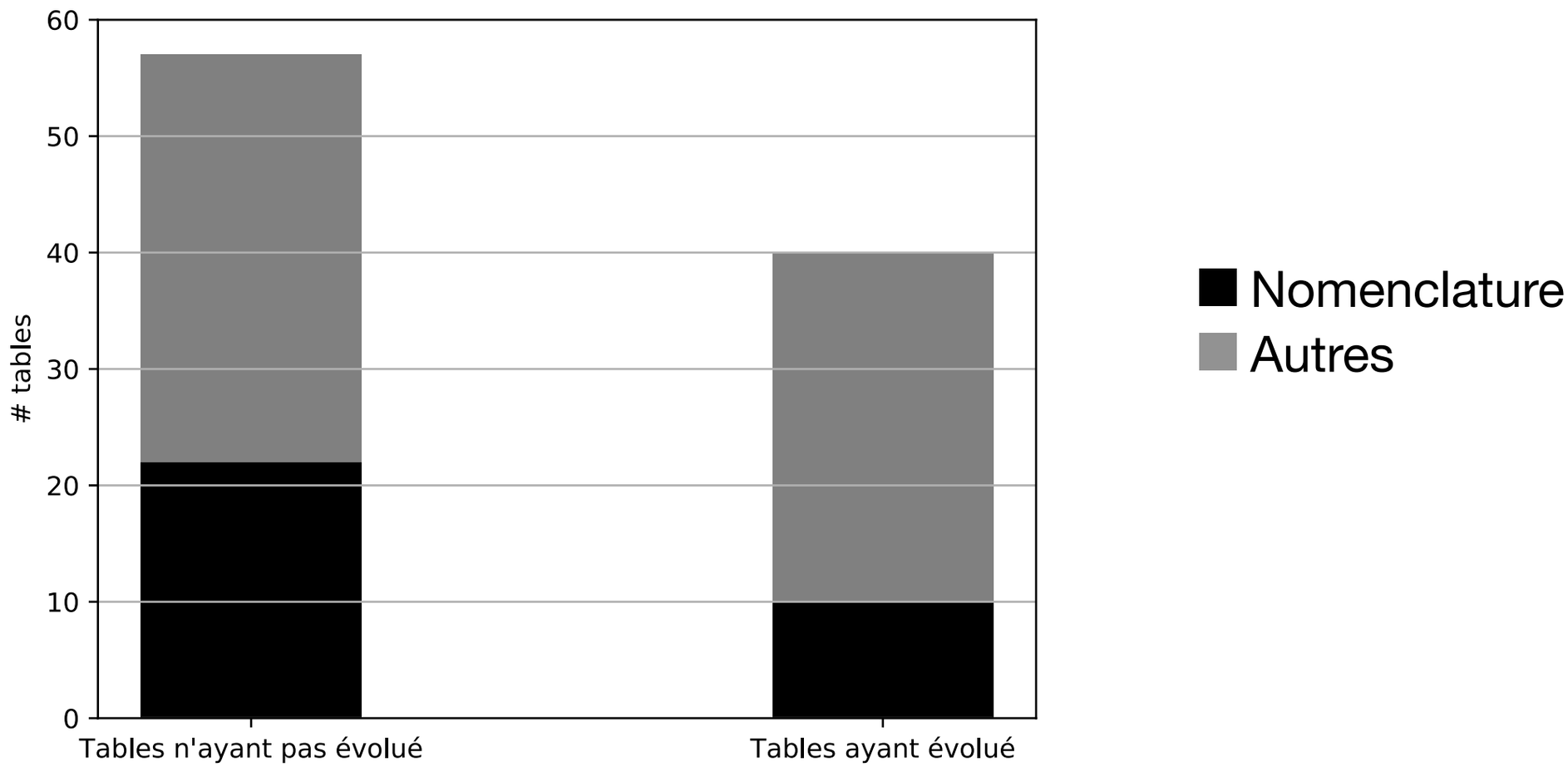
(dernière version de la BDD)

- d = 0 pour toutes les tables
- #PK ≠ 1 pour **2** tables
- Utiliser la propriété 5 pour classer les tables fourni **81%** de précision et **71%** de rappel



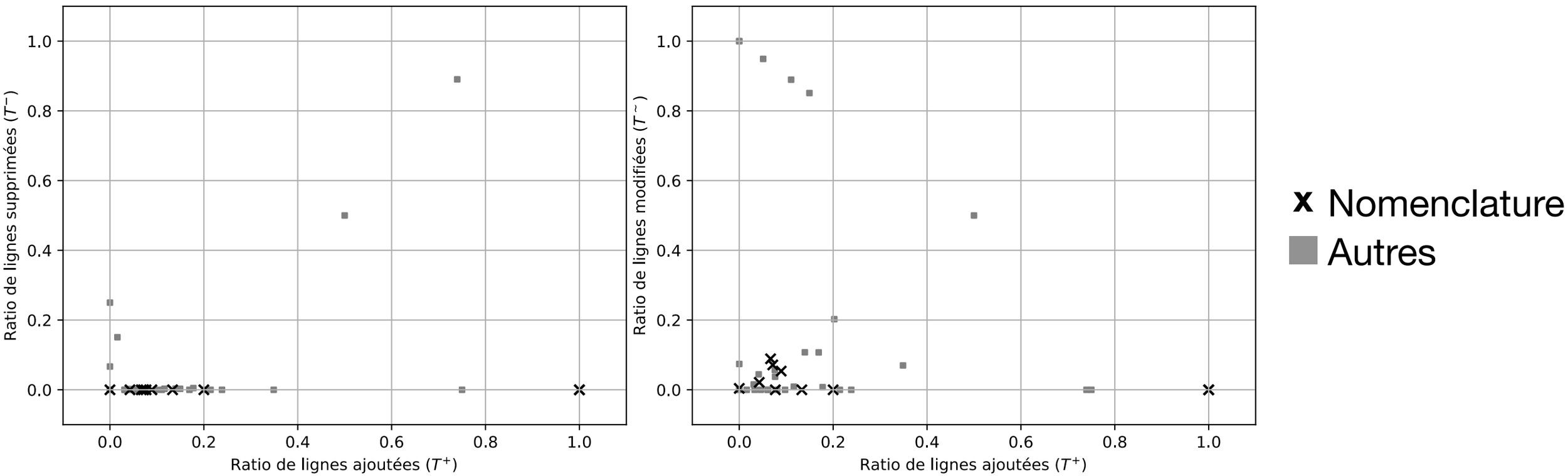
Nombre de tables ayant un certain nombre de clés étrangères

Évolution des données des tables



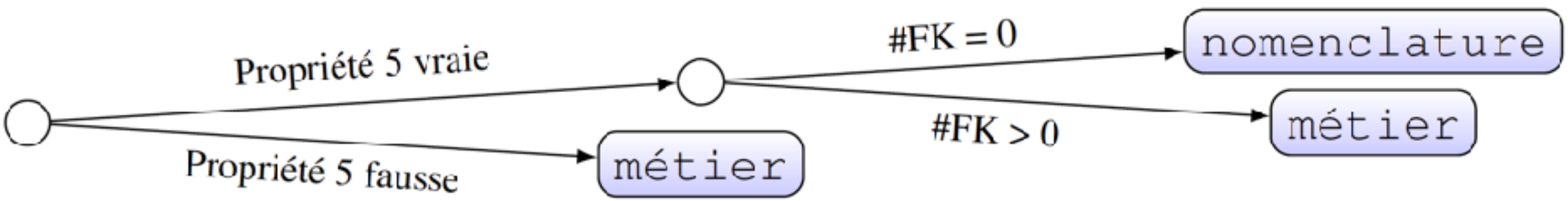
Proportion de tables ayant évolué et n’ayant pas évolué pour les données historique d’AppSI

Évolution des données des tables



- T+ : ratio de lignes **ajoutées** dans une table sur toutes les versions du schéma par rapport au nombre total de lignes ayant existé dans cette table
- T- : ratio de lignes **supprimées** dans une table sur toutes les versions du schéma par rapport au nombre total de lignes ayant existé dans cette table
- T~ : ratio de lignes **modifiées** entre la première et la dernière version du schéma divisé par le nombre de lignes dans la dernière version du schéma

Arbre de décision généré par C4.5



Rappel : La propriété 5 concerne la compatibilité ou non des contraintes d'intégrité référentielles

Précision : **88,6%**

Rappel : **88,7%**

Conclusion

- Définition du concept de table de nomenclature
- Intérêts en terme d’étude, maintenance et évolution d’une base de données
- 5 critères proposés pour discriminer ces tables
- Évaluation empirique de ces critères plutôt positive

Travaux futurs

- Observer les propriétés proposées sur les tables d’autres bases de données
- Répéter l’analyse sur un jeu de données couvrant une plus grande période

https://juliendelplanque.be/phd.html

Utilisation de littéraux

```sql
SELECT personnes.name
FROM
  personnes
WHERE
  personnes.civilite_id = 0;
```

Métriques

- $t^{+}_{i,j}$ est le nombre de lignes ajoutées,
- $t^{-}_{i,j}$ est le nombre de lignes supprimées,
- $t^{\sim}_{i,j}$ est le nombre de lignes modifiées,
- t_i et t_j sont les nombres de lignes respectivement dans les version i et j.
- $N = t_0 + \sum_{i=0}^{n-2} t^{+}_{i,i+1}$ est le nombre total de lignes ayant existé dans une table,
- $T^{+} = \frac{\sum_{i=0}^{n-2} t^{+}_{i,i+1}}{N}$ est le ratio de lignes ajoutées dans une table sur toutes les versions du schéma par rapport au nombre total de lignes ayant existé dans cette table,
- $T^{-} = \frac{\sum_{i=0}^{n-2} t^{-}_{i,i+1}}{N}$ est le ratio de lignes supprimées dans une table sur toutes les versions du schéma par rapport au nombre total de lignes ayant existé dans cette table,
- $T^{\sim} = \frac{t^{\sim}_{0,n-1}}{t_{n-1}}$ est le ratio de lignes modifiées entre la première et la dernière version du schéma divisé par le nombre de lignes dans la dernière version du schéma.