



HAL
open science

Détection automatique de fraude dans les marchés publics : Application aux cas français

Lucas Potin, Rosa Figueiredo, Vincent Labatut, Christine Largeron,
Pierre-Henri Morand

► To cite this version:

Lucas Potin, Rosa Figueiredo, Vincent Labatut, Christine Largeron, Pierre-Henri Morand. Détection automatique de fraude dans les marchés publics : Application aux cas français. Meetup LIAvignon, Nov 2022, Avignon, France. 2022. hal-03833237

HAL Id: hal-03833237

<https://hal.science/hal-03833237>

Submitted on 28 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Détection automatique de fraude dans les marchés publics : Application aux cas français

Lucas Potin¹, Rosa Figueiredo¹, Vincent Labatut¹, Christine Largeron², Pierre-Henri Morand³

1: LIA – Laboratoire Informatique d'Avignon / 2: LabHC – Laboratoire Hubert Curien / 3: LBNC – Laboratoire Biens, Normes, Contrats

Objectifs et méthodes

- Projet **DeCoMaP** : Détecter la Corruption dans les Marchés Publics.
- Collection, traitement et analyse** des données relatives aux marchés publics français, afin d'élaborer des outils de **détection automatique** des risques de **fraude**.
- 3 méthodes, qui sont :
 - Extraction et enrichissement** d'une base de données de marchés publics français en fonction des différentes sources disponibles.
 - Représentation des marchés publics sous la forme de **graphes**.
 - Utilisation de l'aspect relationnel des graphes pour **prédire** des comportements frauduleux.

Les marchés publics

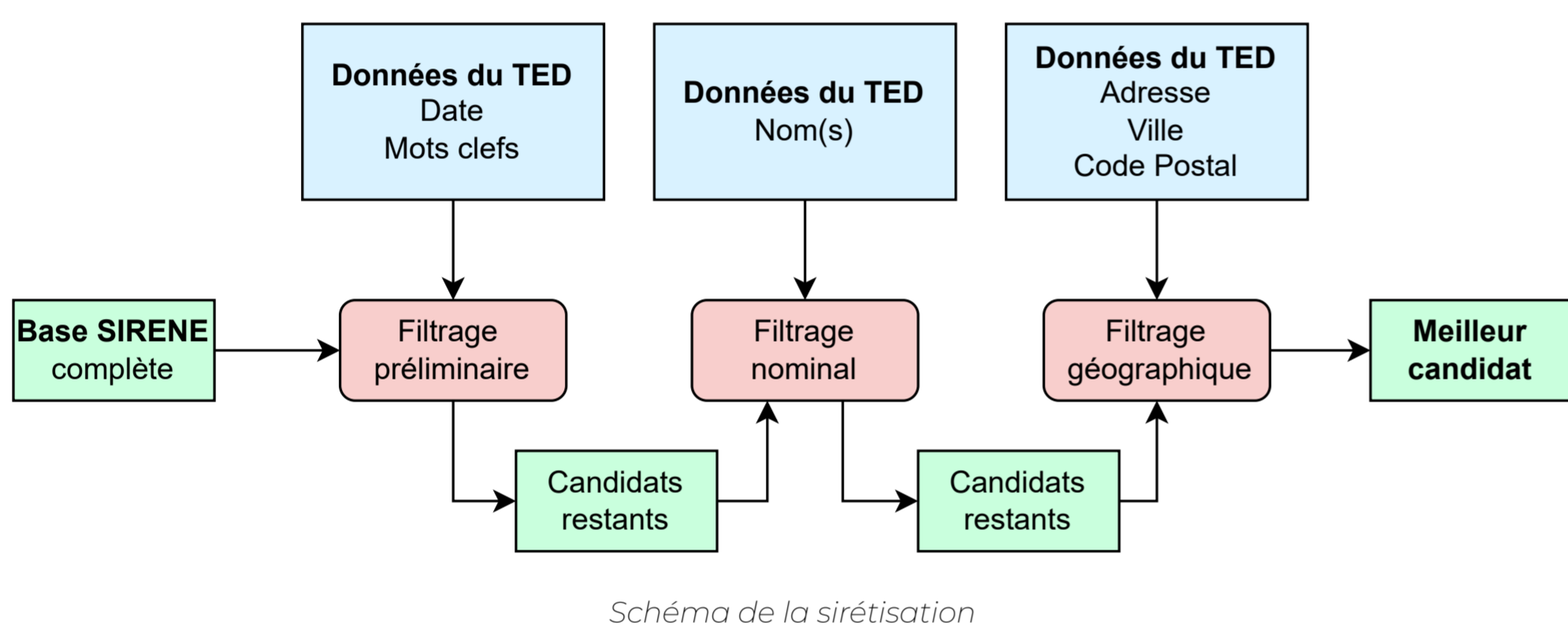
169 060 marchés publics en 2020 en France, pour un montant de plus de **110 milliards** d'euros, entre des clients et des fournisseurs, nommés **agents**.

Données disponibles, au niveau français, via le **Bulletin Officiel des Annonces des Marchés Publics (BOAMP)** et européen via le **Tenders Electronic Daily (TED)**.

Utilisation d'indicateurs standards sur les propriétés des marchés pour repérer des mauvaises pratiques, nommés **redflags**.

Beaucoup de données manquantes, notamment les numéros de SIRET clients (**80%**) et fournisseurs (**95%**).

Première étape : **sirétisation** du TED.



- Sirétisation de plus de **85%** de la base TED, avec un taux de réussite de **70%** pour les clients et **65%** pour les fournisseurs.
- Ce taux monte à **80%** et **75%** si on considère le SIREN.
- Étape supplémentaire de clustering : matching d'agents via **fuzzy-matching**.

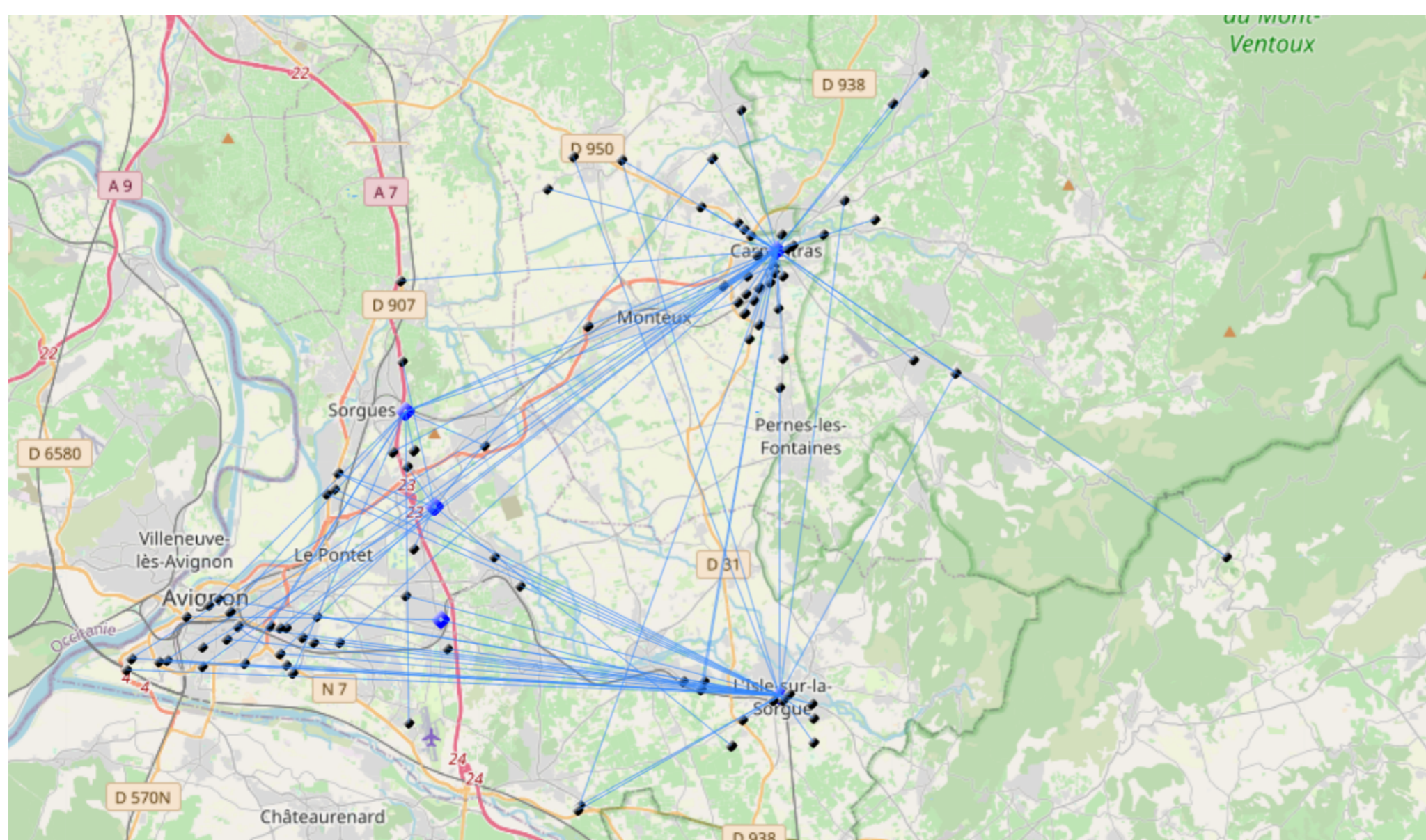
Au départ	Après la sirétisation	Après le fuzzy-matching
889 692	273 525	252 910

Résultats de la sirétisation

- Division du nombre d'agents par **4**.
- Pour les entités sirétisées : ajout d'informations (budget pour les collectivités, **géolocalisation**, etc.).

Construction de graphes

- Construction du graphe global.
- Collection de petits graphes en filtrant sur le domaine d'activité, la zone géographique.
- Agents en sommets, contrats en arêtes.



Exemple de graphe, avec projection sur une carte géographique.

Représentation vectorielle de graphes et classification

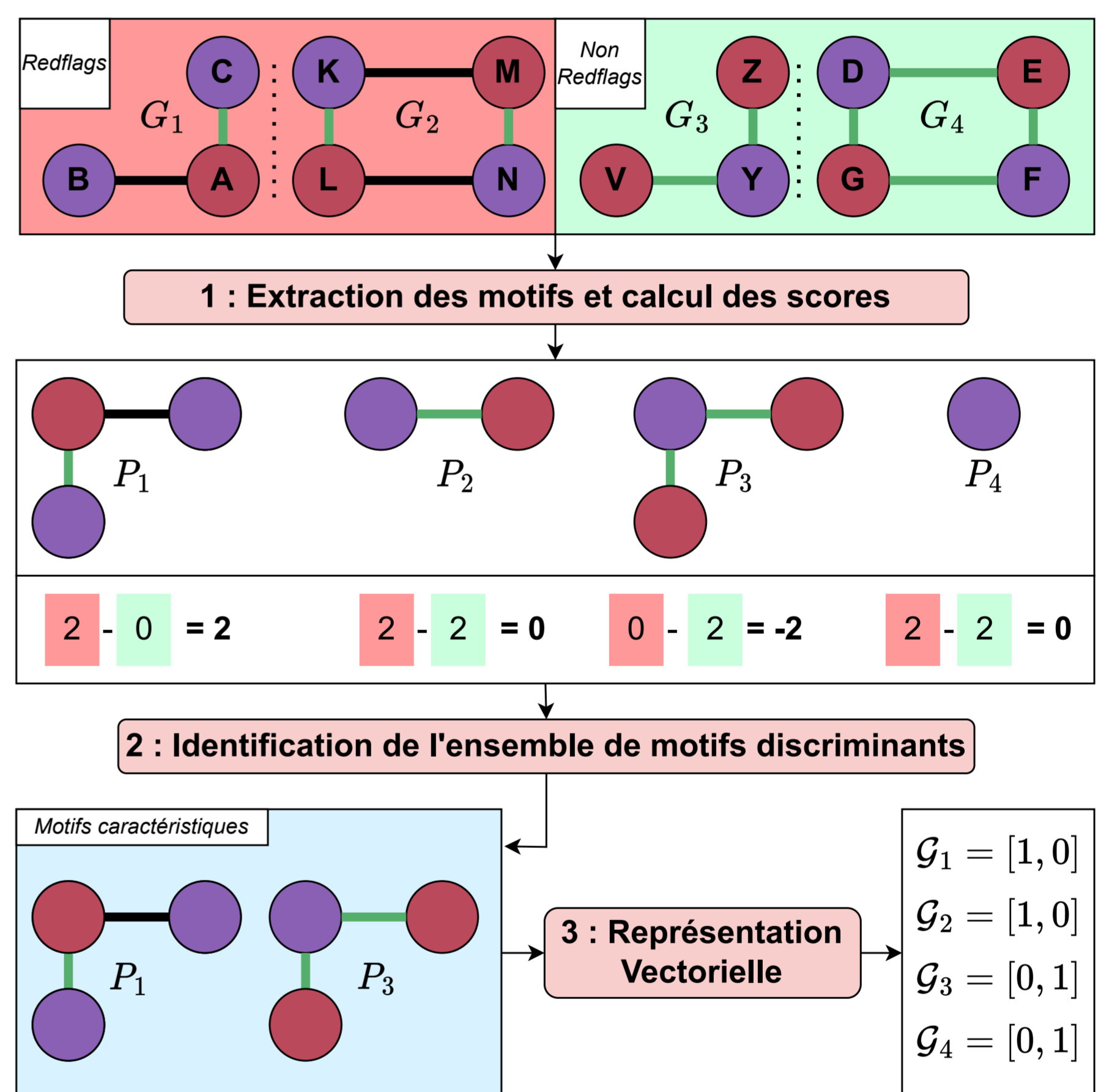
Classification d'une collection de graphes en deux catégories : **redflags** et **non-redflags**.

Méthode analogue au modèle **sac-de-mots** utilisé en recherche d'information : description de graphes en termes de sac-de-sous-graphes, appelés **motifs**.

Calcul d'un score **discriminant** pour les motifs indiquant la présence dans une classe par rapport à l'autre, identification des motifs les plus discriminants

Représentation de chaque graphe comme un vecteur **booléen**, indiquant la **présence** ou l'absence de chaque motif discriminant.

Classification des vecteurs par une méthode **classique** : SVM, RF, K-Neighbours.



Représentation vectorielle des graphes

- Meilleure performance de classification obtenue avec **Random Forest**.
- Résultats sur des **égo-réseaux** de municipalités.

Méthode	Graphes redflags		Graphes non-redflags	
	Pre	Rec	Pre	Rec
Graph2Vec	0,88	0,89	0,88	0,86
150-motifs	0,89	0,85	0,88	0,87
Tous-motifs	0,94	0,90	0,89	0,93

Résultats de la classification

- Performance **équivalente** à une méthode d'embedding classique (Graph2Vec).
- Méthode **interprétable**, via l'analyse des motifs considérés comme discriminants pour identifier des phénomènes **économiques** (favoritisme, collusion).

Perspectives

- Utilisation d'autres sources de données (BANATIC, BODACC, INPI).
- Extraction de motifs **émergents**.



ANR-19-CE38-0004



CV HAL



LinkedIn



Lucas Potin

Doctorant en 2ème année informatique
lucas.potin@univ-avignon.fr