Découverte de connaissances et apprentissage dans les données graphes : Application aux marchés publics français

Lucas Potin, Rosa Figueiredo, Vincent Labatut, Christine Largeron

Decade 2023 : Découverte de connaissances et apprentissage dans les données graphes

Application aux marchés publics français

Lucas Potin[1] Rosa Figueiredo[1] Vincent Labatut[1]
Christine Largeron[2]

[1]Laboratoire Informatique d'Avignon – LIA EA 4128
{prénom.nom}@univ-avignon.fr

[2]Laboratoire Hubert Curien – LabHC UMR 5516
christine.largeron@univ-st-etienne.fr

Atelier Decade – Strasbourg
6 juillet 2023

LABORATOIRE INFORMATIQUE D'AVIGNON

ANR DeCoMaP

LABORATOIRE HUBERT CURIEN
UMR • CNRS • 5516 • SAINT-ETIENNE

anr

AVIGNON UNIVERSITÉ

Sommaire

Contexte applicatif : les marchés publics

Marché public

Contrat conclu à titre onéreux par un ou plusieurs acheteurs publics avec un ou plusieurs opérateurs économiques. [1]

État des lieux en France

- 169 060 marchés en 2020, pour plus de **110 milliards** d'euros.
- Données disponibles au niveau Européen via le Tenders Electronic Daily (**TED**) [2].
- Notice : informations **relationnelles** et attributaires :
  - **Marché demandé par un ou plusieurs clients**
  - **Marché réalisé par un ou plusieurs fournisseurs**
  - Attributs sur le marché (prix, nombre d'offres, etc.)

1. https://www.economie.gouv.fr/entreprises/definition-marche-public
2. https://ted.europa.eu/

Contexte applicatif : les marchés publics

Marché public

Contrat conclu à titre onéreux par un ou plusieurs acheteurs publics avec un ou plusieurs opérateurs économiques. [3]

État des lieux en France

- 169 060 marchés en 2020, pour plus de **110 milliards** d'euros.
- Données disponibles au niveau Européen via le Tenders Electronic Daily (**TED**) [4].
- Notice : informations relationnelles et **attributaires** :
  - Marché demandé par un ou plusieurs clients
  - Marché réalisé par un ou plusieurs fournisseurs
  - **Attributs sur le marché (prix, nombre d'offres, etc.)**

3. https://www.economie.gouv.fr/entreprises/definition-marche-public
4. https://ted.europa.eu/

Détection de marchés frauduleux

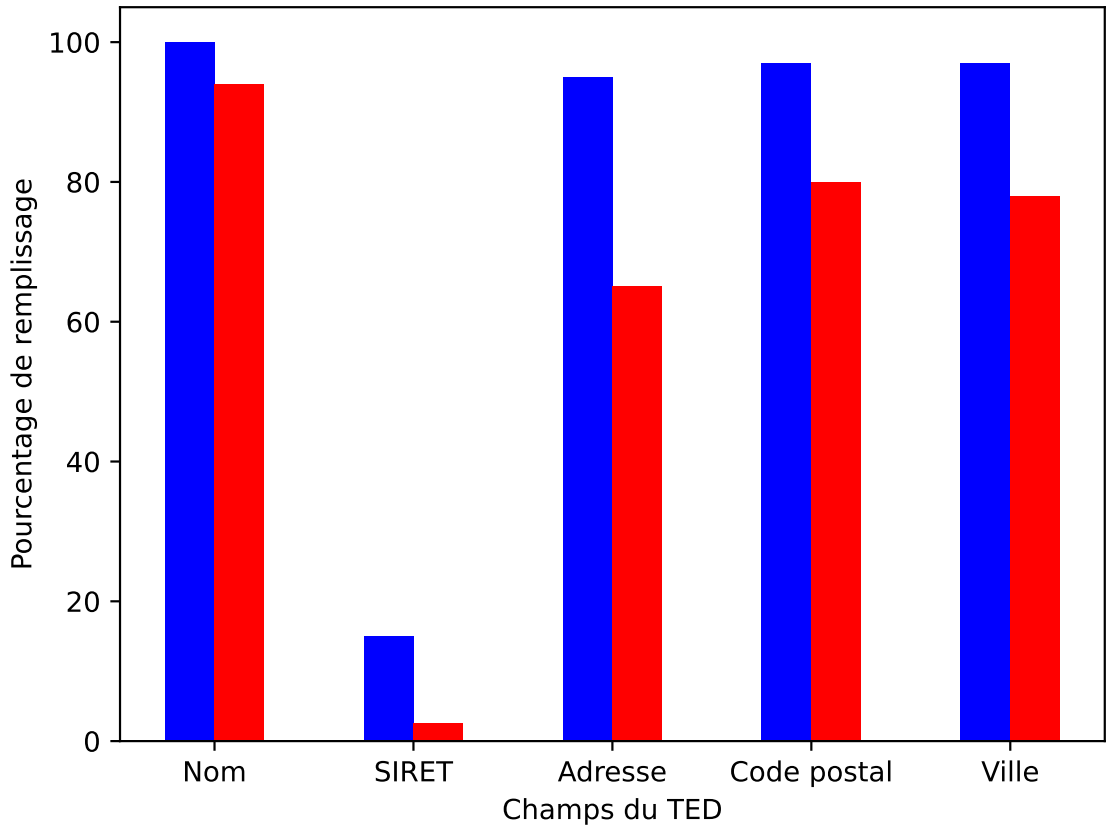
- Principale difficulté : pas de vérité terrain.
- Approche classique : passer par des **red flags** [2]. Données définies sur des données tabulaires [3].
- Remplissage des différents champs liés aux redflags

Nombre d'offres	Présence de critères	Contrat sous-traité
$69,3\%$	$76,5\%$	$50,7\%$

- **Notre objectif** : proposer une méthode utilisant les relations entre les acteurs
- Peut-on utiliser l'information **relationnelle** pour identifier des marchés frauduleux en l'absence de red flags ?

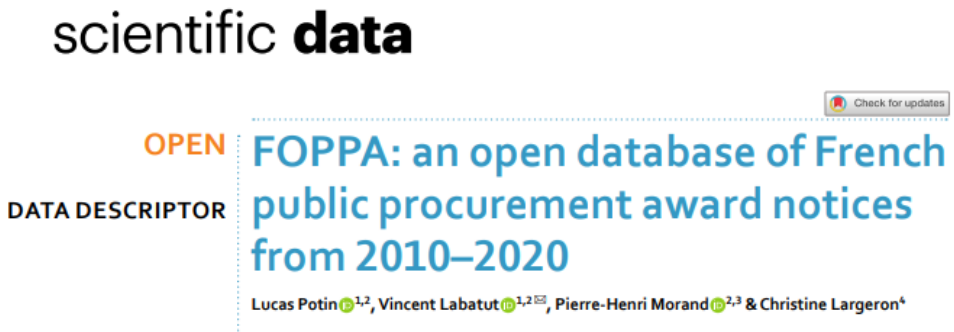
Données des marchés publics

- Nécessite d'avoir une identification précise des agents.
- **Problème** : SIRET pas renseigné pour les clients et fournisseurs.



La base FOPPA

- Création de la base FOPPA [6]
- TED français entre 2010 et 2020 $\longrightarrow$ 1,3 M de lots
- Consolidation des données
- Sirétisation correcte de 80% de la base



- Github : `https://github.com/compnet/foppainit`
- Zenodo : `https://zenodo.org/record/7808664`

Extraction des graphes

Construction de graphes :

- Plus d'1 Million de contrats $\longrightarrow$ Graphe de très grande taille.
- Comportements différents selon le secteur d'activité [1]

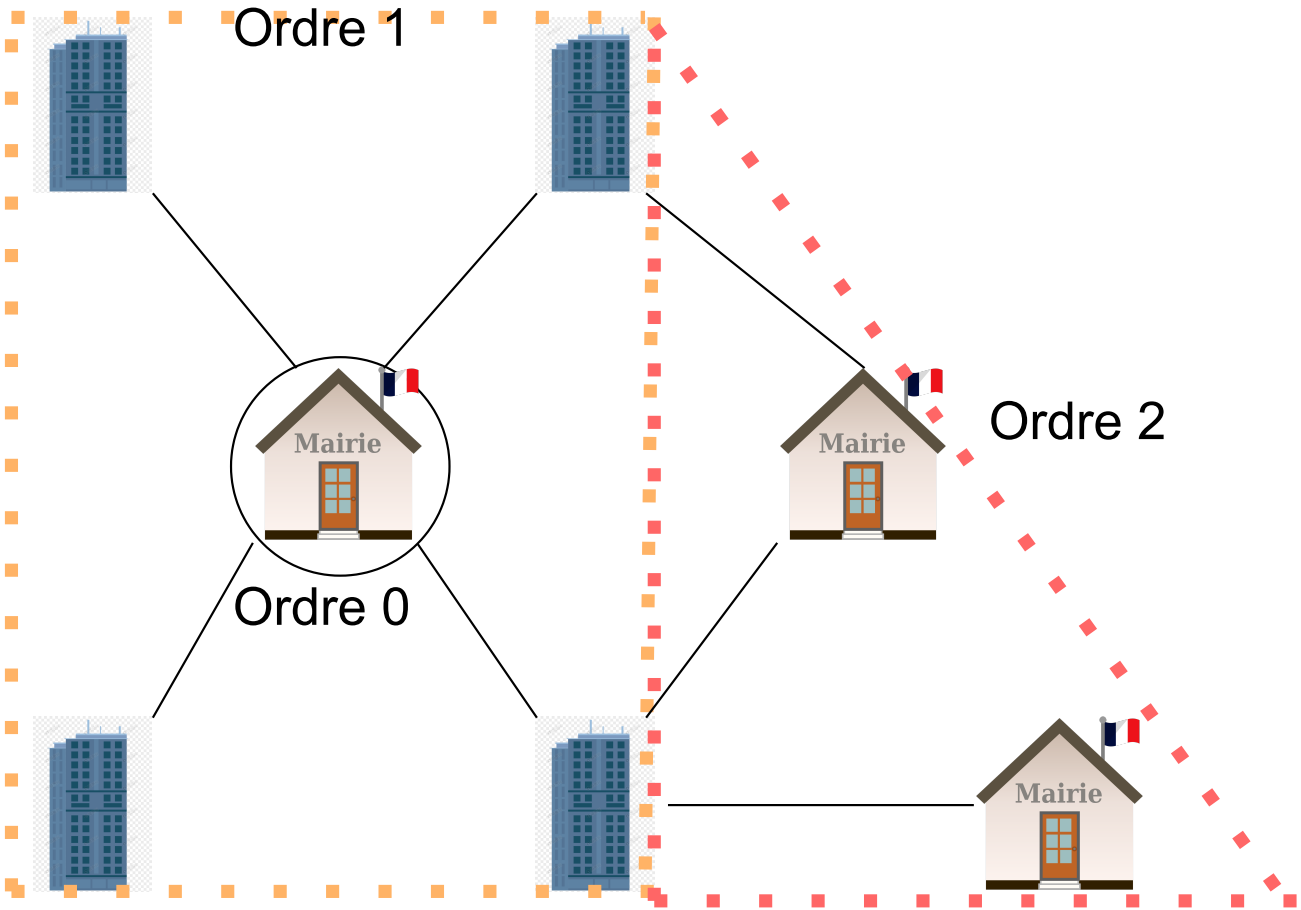
$\rightarrow$ Choix de **restreindre** la sélection des marchés

Construction d'ensemble de contrats cohérents :

- Temporel : plage d'un an.
- Géographique : restriction au département.
- Secteur d'activité : marché de travaux (CPV 45).

Création des graphes

Construction d'égo réseaux d'ordre 2 à partir de municipalités.



Framework PANG

Données

Soit une collection $\mathcal{G}$ de graphes attribués $\mathcal{G} = G(V, E, \mathbf{X}, \mathbf{Y}, L)$ composés d'un ensemble de sommets V, un ensemble de liens E, une matrice d'attributs des sommets $\mathbf{X}$, une matrice d'attributs des liens $\mathbf{Y}$ et un label du graphe L qui peut être `Anormal` ou `Normal` en fonction des red flags.

Objectif

Prédire les labels inconnus pour les graphes sans label.

Idée principale

Représenter les graphes en fonction de leurs sous-graphes : **motifs**.

Framework PANG

Données

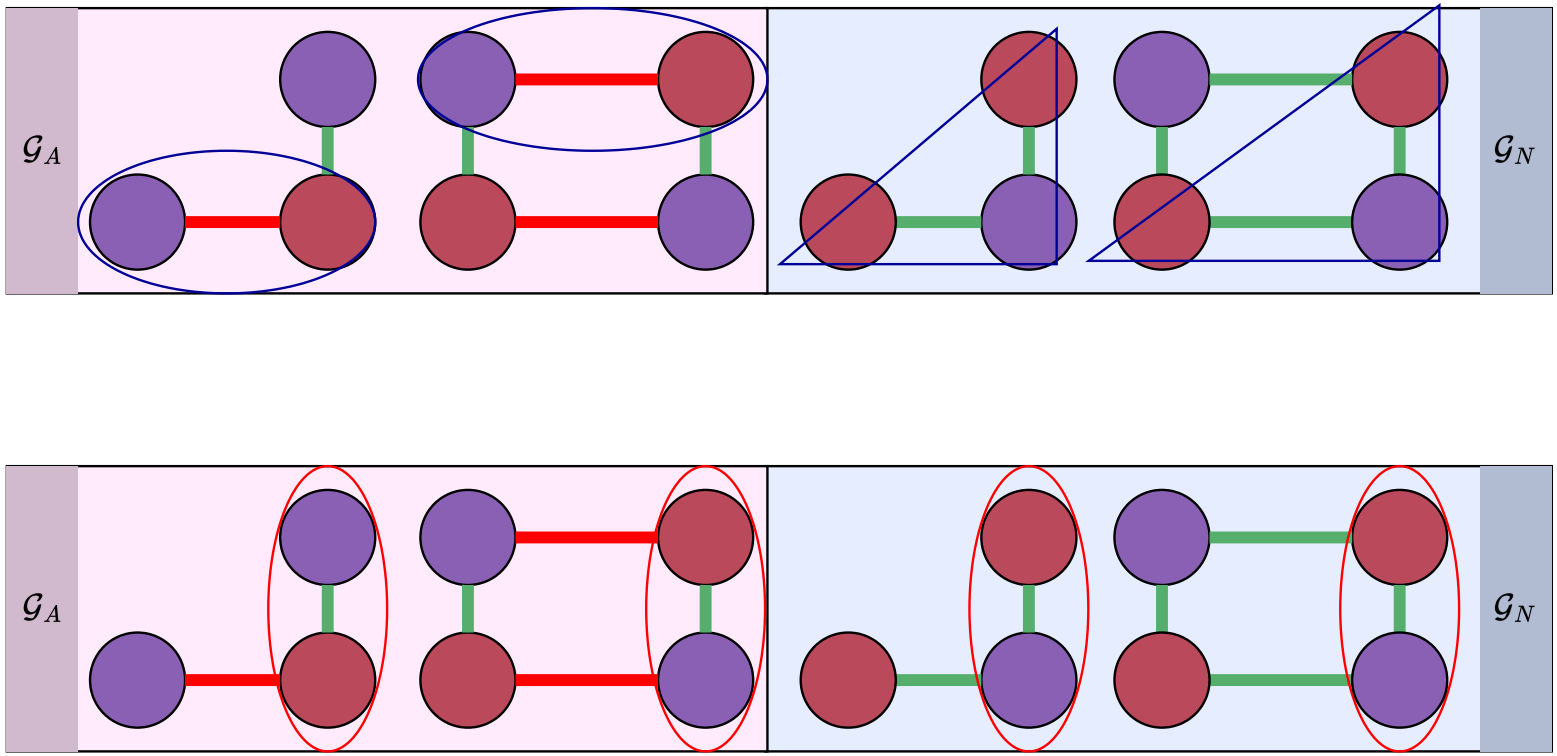
Soit une collection $\mathcal{G}$ de graphes attribués $\mathcal{G} = G(V, E, \mathbf{X}, \mathbf{Y}, L)$ composés d'un ensemble de sommets V, un ensemble de liens E, une matrice d'attributs des sommets $\mathbf{X}$, une matrice d'attributs des liens $\mathbf{Y}$ et un label du graphe L qui peut être `Anormal` ou `Normal` en fonction des red flags.

Dans notre contexte applicatif :

- V : les différents agents.
- E : les différentes relations (1 lot ou +) entre agents.
- $\mathbf{X}$: le type de l'agent (client ou fournisseur).
- $\mathbf{Y}$: le nombre de lots.
- L : selon le nombre de red flags dans le graphe.

Motif discriminant

Certains motifs figurent principalement dans une des deux classes de graphes, tandis que d'autres sont génériques.



Motif discriminant

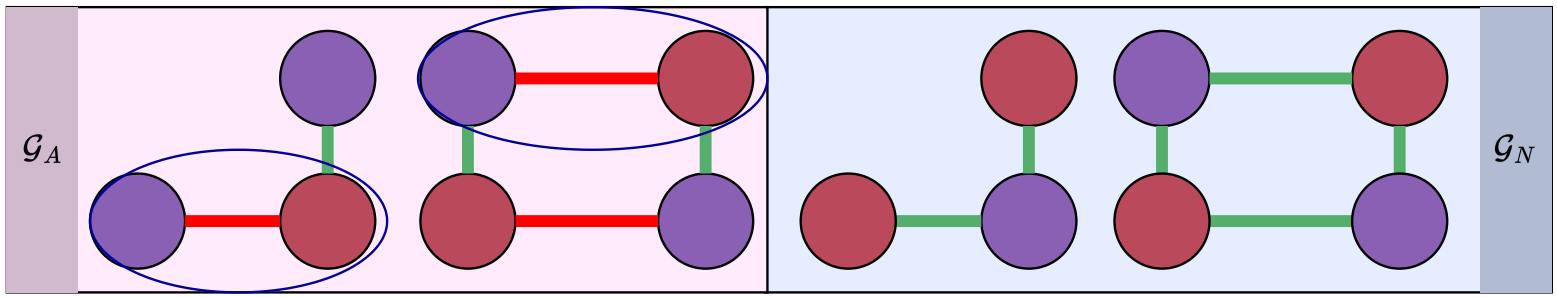
Définition (Fréquence d'un motif)

Soit $\mathcal{G}$ un ensemble de graphes attribués. La fréquence $freq(P, \mathcal{G})$ d'un motif P dans $\mathcal{G}$ est le nombre de graphes dans $\mathcal{G}$ contenant P : $freq(P, \mathcal{G}) = |\{G \in \mathcal{G} : \exists P' \subset G \text{ t.q. } P \cong P'\}|$.

Définition (Score discriminant)

Étant donné un motif P de $\mathcal{G}$, le score discriminant de P est défini par $dis(P) = |freq(P, \mathcal{G}_A) - freq(P, \mathcal{G}_N)|$.

Motif discriminant

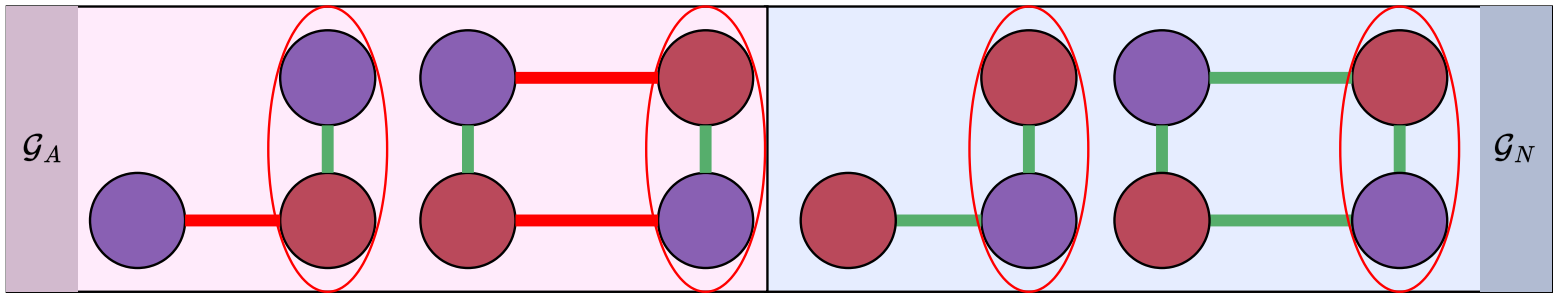


Définition (Score discriminant)

Étant donné un motif P de $\mathcal{G}$, le score discriminant de P est défini par $dis(P) = |freq(P, \mathcal{G}_A) - freq(P, \mathcal{G}_N)|$.

$$dis(P) = |2 - 0| = 2$$

Motif discriminant



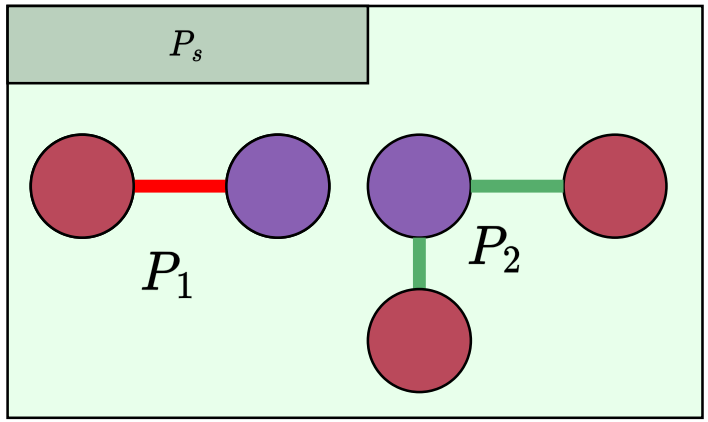
Définition (Score discriminant)

Étant donné un motif P de $\mathcal{G}$, le score discriminant de P est défini par $dis(P) = |freq(P, \mathcal{G}_A) - freq(P, \mathcal{G}_N)|$.

$$dis(P) = |2 - 2| = 0$$

Framework PANG

- **Identification d'un ensemble de motifs discriminants $\mathcal{P}_s$ parmi tous les motifs P de $\mathcal{G}$**
- Représentation vectorielle
- Apprentissage d'un modèle permettant de classer un nouveau graphe sans label à partir de sa représentation vectorielle



Framework PANG

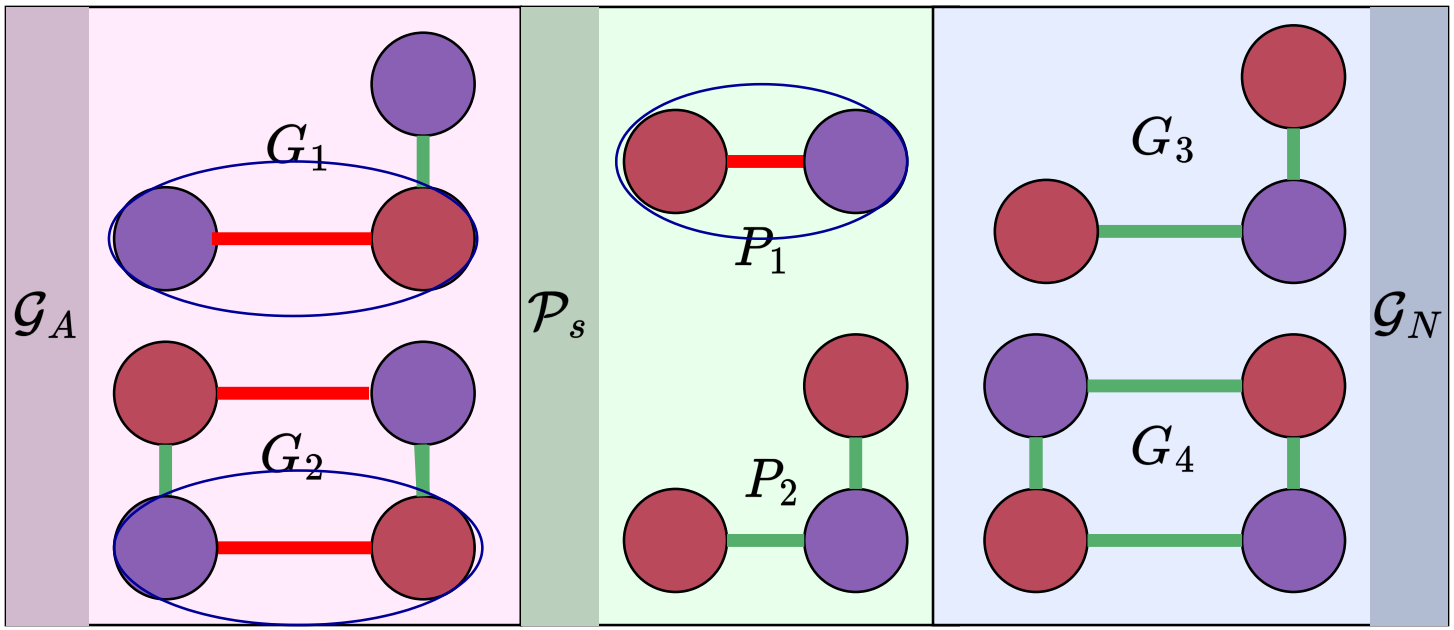
- Identification d'un ensemble de motifs discriminants $\mathcal{P}_s$ parmi tous les motifs P de $\mathcal{G}$
- **Représentation vectorielle**
- Apprentissage d'un modèle permettant de classer un nouveau graphe sans label à partir de sa représentation vectorielle

Représentation vectorielle

Soit G un graphe attribué et $\mathcal{P}_s = \{P_i\}$ un ensemble de s motifs discriminants choisis.
On note $\mathbf{h} = \{h_i\}$ la représentation vectorielle de G avec $h_i = 1$ si P_i est présent dans G, 0 sinon.

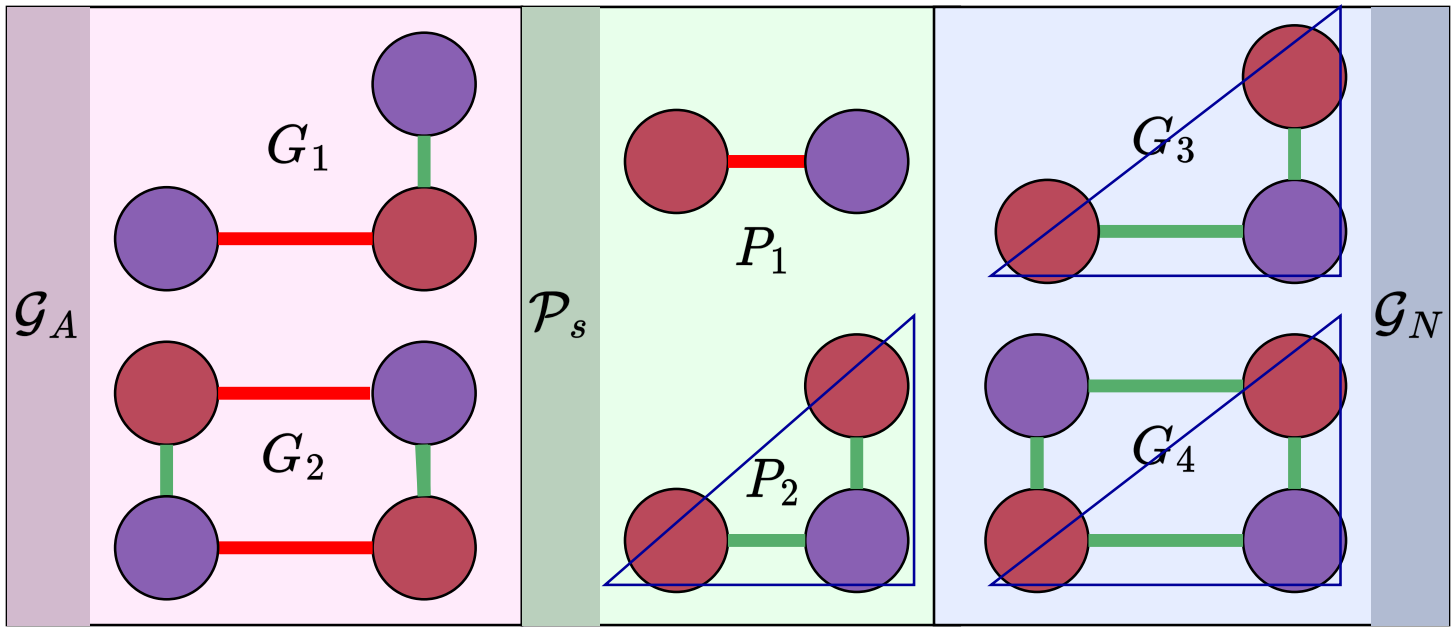
Framework PANG

- Identification d'un ensemble de motifs discriminants $\mathcal{P}_s$ parmi tous les motifs P de $\mathcal{G}$
- **Représentation vectorielle**
- Apprentissage d'un modèle permettant de classer un nouveau graphe sans label à partir de sa représentation vectorielle



Framework PANG

- Identification d'un ensemble de motifs discriminants $\mathcal{P}_s$ parmi tous les motifs P de $\mathcal{G}$
- **Représentation vectorielle**
- Apprentissage d'un modèle permettant de classer un nouveau graphe sans label à partir de sa représentation vectorielle



Framework PANG

- Identification d'un ensemble de motifs discriminants $\mathcal{P}_s$ parmi tous les motifs P de $\mathcal{G}$
- Représentation vectorielle
- **Apprentissage d'un modèle permettant de classer un nouveau graphe sans label à partir de sa représentation vectorielle**

	h	L
G_1	1 0 1 1	A
G_2	1 0 0 1	A
G_3	0 1 0 0	N
G_4	0 1 0 0	N
G_5	0 1 1 0	?

Expérimentation sur le jeu des marchés publics

- Ensembles de contrats liés à des municipalités.

Label du graphe	Nombre de graphes	Nombre moyen de sommets (st)	Nombre moyen d'arêtes (st)
Anormal	330	15,76 (5,56)	17,09 (7,86)
Normal	330	12,54 (5,41)	12,59 (6,90)

- Tests de différents classifieurs (RF, SVM).
- Utilisation de la validation croisée (5-Folds).

Résultats en termes de précision (P), rappel (R) et F-score (F)

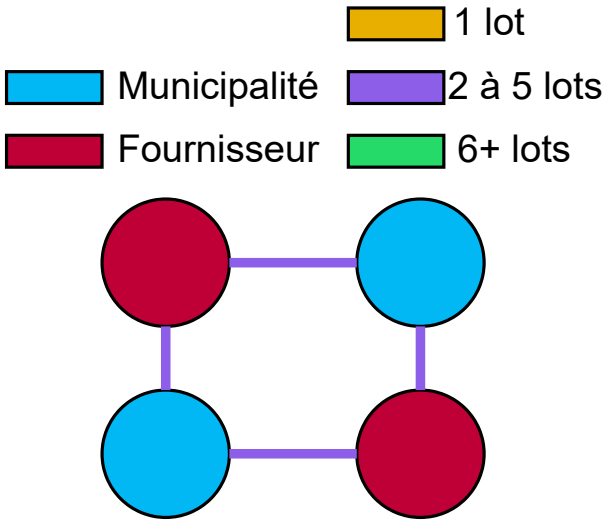
Résultats avec Random Forest selon la valeur de s.

Nombre de motifs	Graphes anormaux P	R	F	Graphes normaux P	R	F
10	0,69	0,77	0,72	0,68	0,59	0,63
100	0,84	0,84	0,84	0,81	0,81	0,81
150	0,89	0,85	0,87	0,88	0,87	0,87
Tous	**0,94**	**0,90**	**0,92**	**0,89**	**0,93**	**0,91**
Graph2Vec	0,88	0,89	0,88	0,88	0,86	0,87

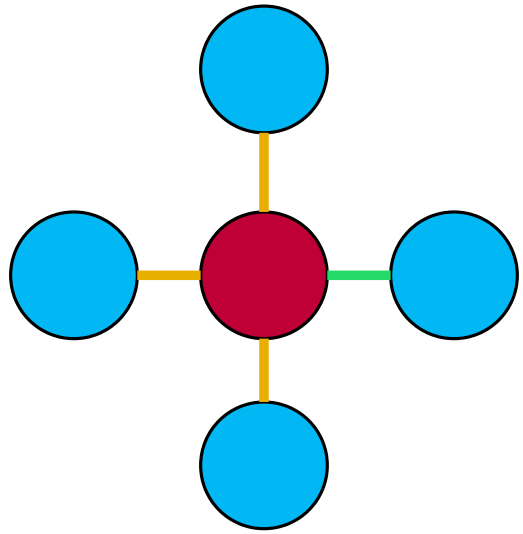
En utilisant 1% des motifs, les résultats sont équivalents à 95% de la performance totale.

Résultats

Avantage de notre approche : interprétabilité des résultats



Motif lié à des secteurs sans grosse concurrence.



Motif pouvant être lié a du favoritisme.

Pour développer

Deux articles :

- L. Potin, R. Figueiredo, V. Labatut et C. Largeron. « Extraction de motifs pour la détection d'anomalies dans des graphes : application à la fraude dans les marchés publics ». In : *23ème Conférence Extraction et Gestion des Connaissances*. Lyon, FR : Éditions RNTI, 2023, p. 289–296. url : https://editions-rnti.fr/index.php?inprocid=1002829
- L. Potin, R. Figueiredo, V. Labatut et C. Largeron. « Pattern Mining for Anomaly Detection in Graphs : Application to Fraud in Public Procurement ». In : *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2023*. 2023. doi : 10.48550/ARXIV.2306.10857

Conclusion

Contributions

- PANG : Méthode générique de détection d'anomalies dans des graphes, multi-classe et applicable à d'autres domaines.
- Construction d'un jeu de données à partir de marchés publics.
- Résultats expérimentaux confirmant une bonne identification des fraudes en l'absence des redflags.

Perspectives

- Paramétrage de PANG (type de motifs, de représentation...)
- Ajout des personnes physiques

Merci pour votre attention, avez vous des questions ?

AVIGNON
UNIVERSITÉ

Bibliographie I

[1] F. Maréchal et P.-H. Morand. « Are social and environmental clauses a tool for favoritism ? Analysis of French public procurement markets ». In : *3e Conférence annuelle de l'Association Française d'Economie du Droit* october (2018). url : https://mre.edu.umontpellier.fr/le-seminaire-de-mre/.

[2] N. Modrušan, K. Rabuzin et L. Mršic. « Review of Public Procurement Fraud Detection Techniques Powered by Emerging Technologies ». In : *International Journal of Advanced Computer Science and Applications* 12.2 (2021). doi : 10.14569/ijacsa.2021.0120272.

[3] National Fraud Authority. *RED FLAGS for integrity : Giving the green light to open data solutions*. Rapp. tech. Open Contracting Partnership, 2016. url : https://www.open-contracting.org/wp-content/uploads/2016/11/OCP2016-Red-flags-for-integrityshared-1.pdf.

[4] L. Potin, R. Figueiredo, V. Labatut et C. Largeron. « Extraction de motifs pour la détection d'anomalies dans des graphes : application à la fraude dans les marchés publics ». In : *23ème Conférence Extraction et Gestion des Connaissances*. Lyon, FR : Éditions RNTI, 2023, p. 289-296. url : https://editions-rnti.fr/index.php?inprocid=1002829.

Bibliographie II

[5] L. Potin, R. Figueiredo, V. Labatut et C. Largeron. « Pattern Mining for Anomaly Detection in Graphs : Application to Fraud in Public Procurement ». In : *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2023*. 2023. doi : 10.48550/ARXIV.2306.10857.

[6] L. Potin, V. Labatut, C. Largeron et P. H. Morand. « FOPPA : an open database of French public procurement award notices from 2010–2020 ». In : *Scientific Data* 10 (2023), p. 303. doi : 10.1038/s41597-023-02213-z.