Utilisation de graphes pour la détection de corruption dans les marchés publics

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

21/02/2023

LABORATOIRE INFORMATIQUE D'AVIGNON

ANR DeCoMaP

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

anr

AVIGNON UNIVERSITÉ

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. [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/

Détection de marchés frauduleux

Secteur à risque

Selon l'AFA [3], 16 % des responsables achats indiquent avoir fait l'objet d'une tentative de corruption au cours de leur carrière.

- Principale difficulté : pas de vérité terrain.
- Approche classique : passer par des **red flags** [MRM21]. Données uniquement attributaires [Nat16].
- Amélioration des liens entre les acteurs [Pot+22].
- Exploitation possible via l'utilisation des **graphes**.

3. https://www.agence-francaise-anticorruption.gouv.fr/files/files/Guide_maitrise_risque_corruption-Hyperlien.pdf

Détection de marchés frauduleux

Champ	Nombre d'offres	Prix d'attribution
% données manquantes	31%	32%

- Certaines données indispensables pour calculer les red flags sont manquantes.
- Peut-on utiliser l'information **relationnelle** pour identifier des marchés frauduleux en l'absence de red flags ?

Notre approche

Données

Soit une collection $\mathcal{G}$ de graphes attribués $G(V, E, \mathbf{X}, \mathbf{Y}, L)$ composés d'un ensemble de sommets V, un ensemble de liens E, une matrice d'attributs de sommets $\mathbf{X}$, une matrice d'attributs de 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**.

Extraction de motifs fréquents

Extraction de motifs fréquents

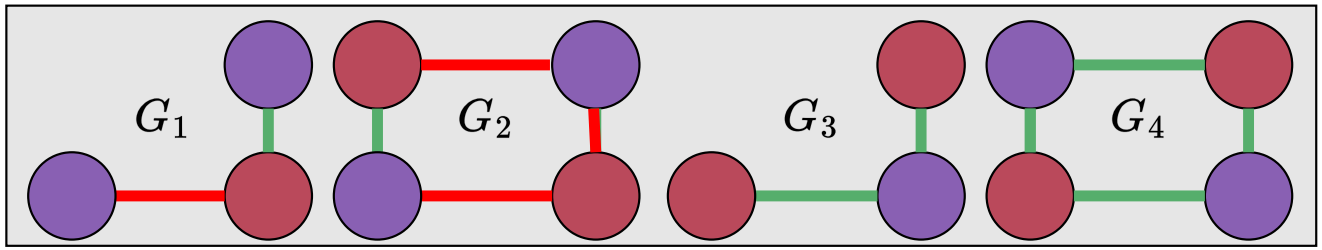
Soit une collection $\mathcal{G}$ de graphes attribués $G(V, E, \mathbf{X}, \mathbf{Y})$ et un seuil t. Le problème de l'extraction de motifs fréquents revient à trouver l'ensemble de sous-graphes apparaissant au moins dans t graphes.

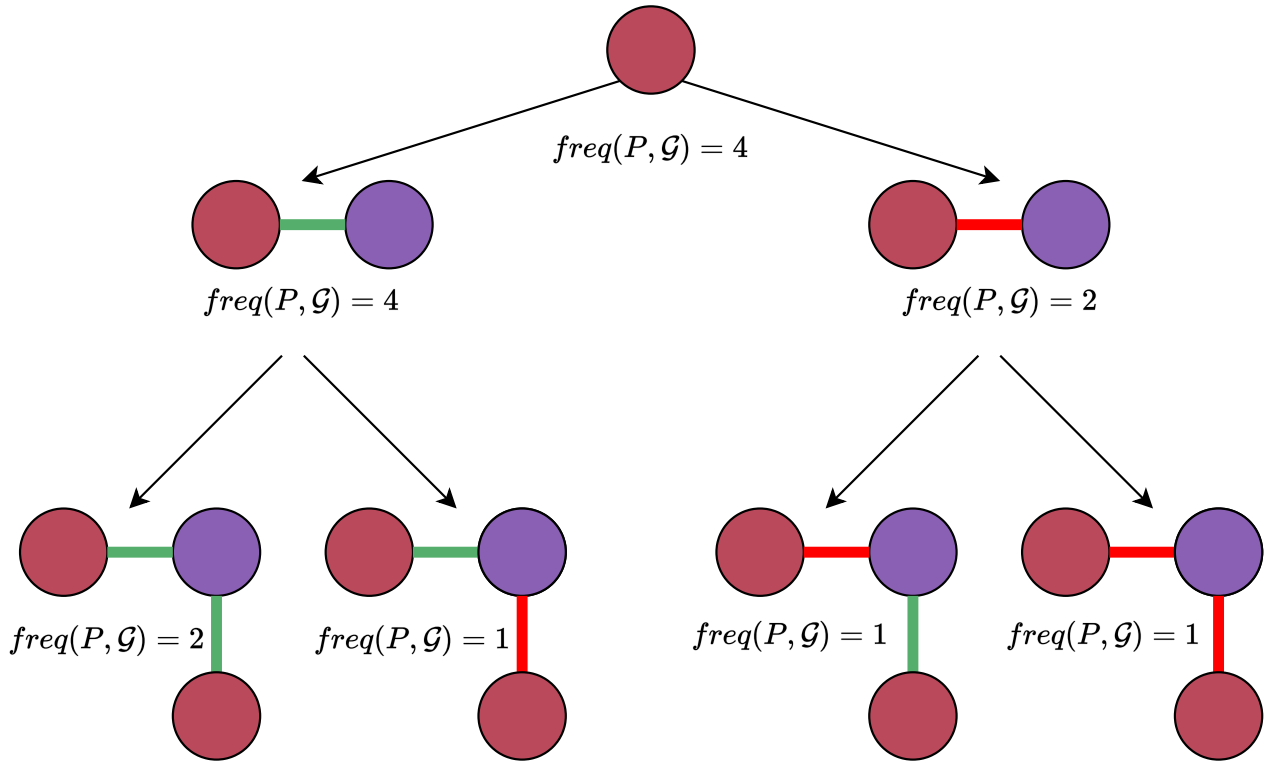
Definition (Support d'un motif)

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

- Plusieurs algorithmes existants : gSpan, FSM, TKG...
- Méthodes dites "Pattern Growth".

Méthodes Pattern Growth





Algorithme gSpan

- Problème d'isomorphisme de sous-graphes : très **coûteux**.
- Explosion du nombre de motifs.
- Algorithme **gSpan** [YH]
- 2 stratégies utilisées :
  - Représenter chaque sous-graphe par un code, nommé **code DFS**. Relation d'ordre entre les codes DFS.
  - Anti-monotonie du support.

Stratégies utilisées dans gSpan

Definition (Code DFS minimal)

Soit P et P' deux motifs. On a
$min(DFS(P)) = min(DFS(P')) \iff P \cong P'$

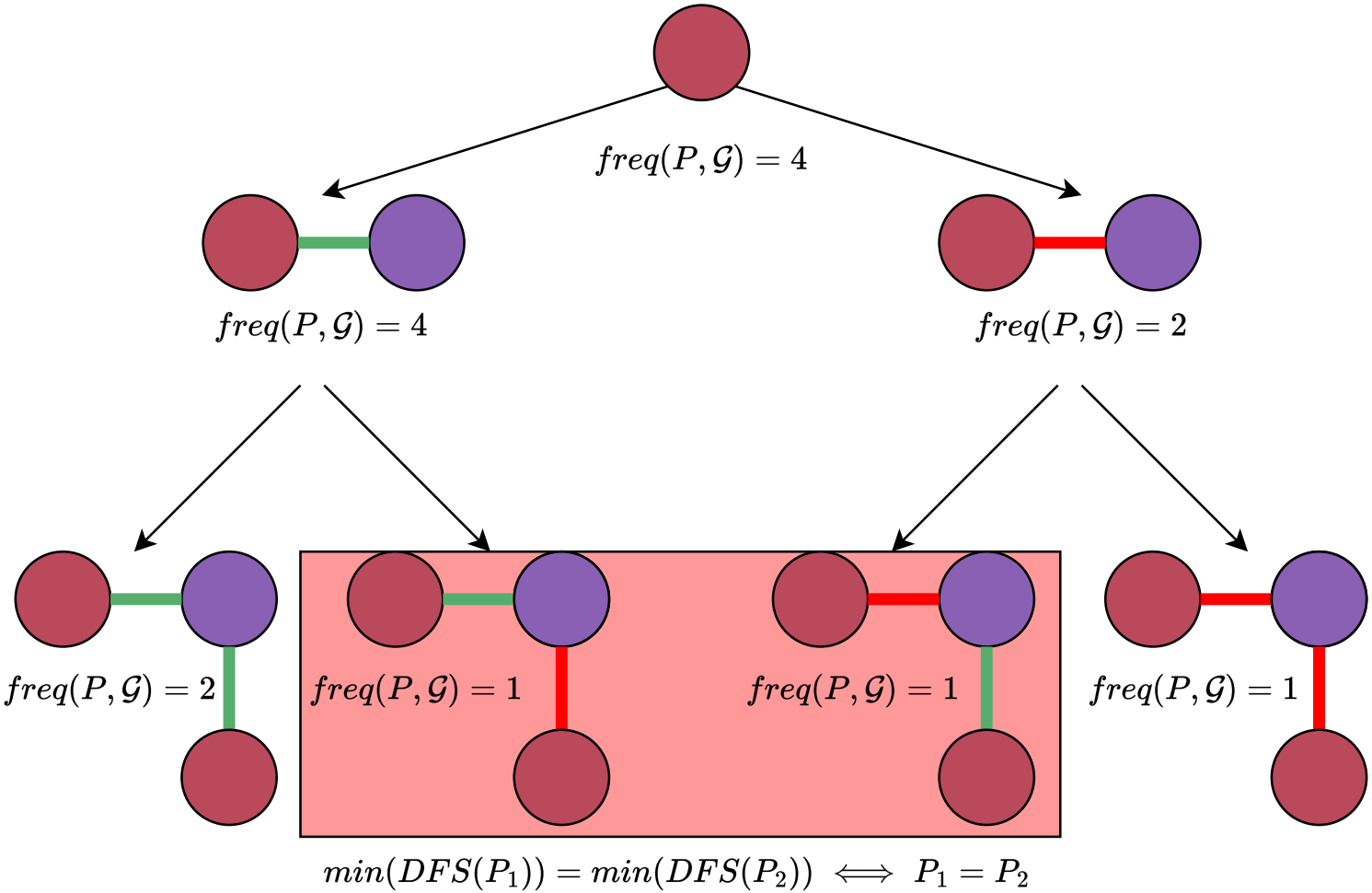
On peut ainsi déterminer si un motif a déjà été miné

Definition (Support anti-monotone)

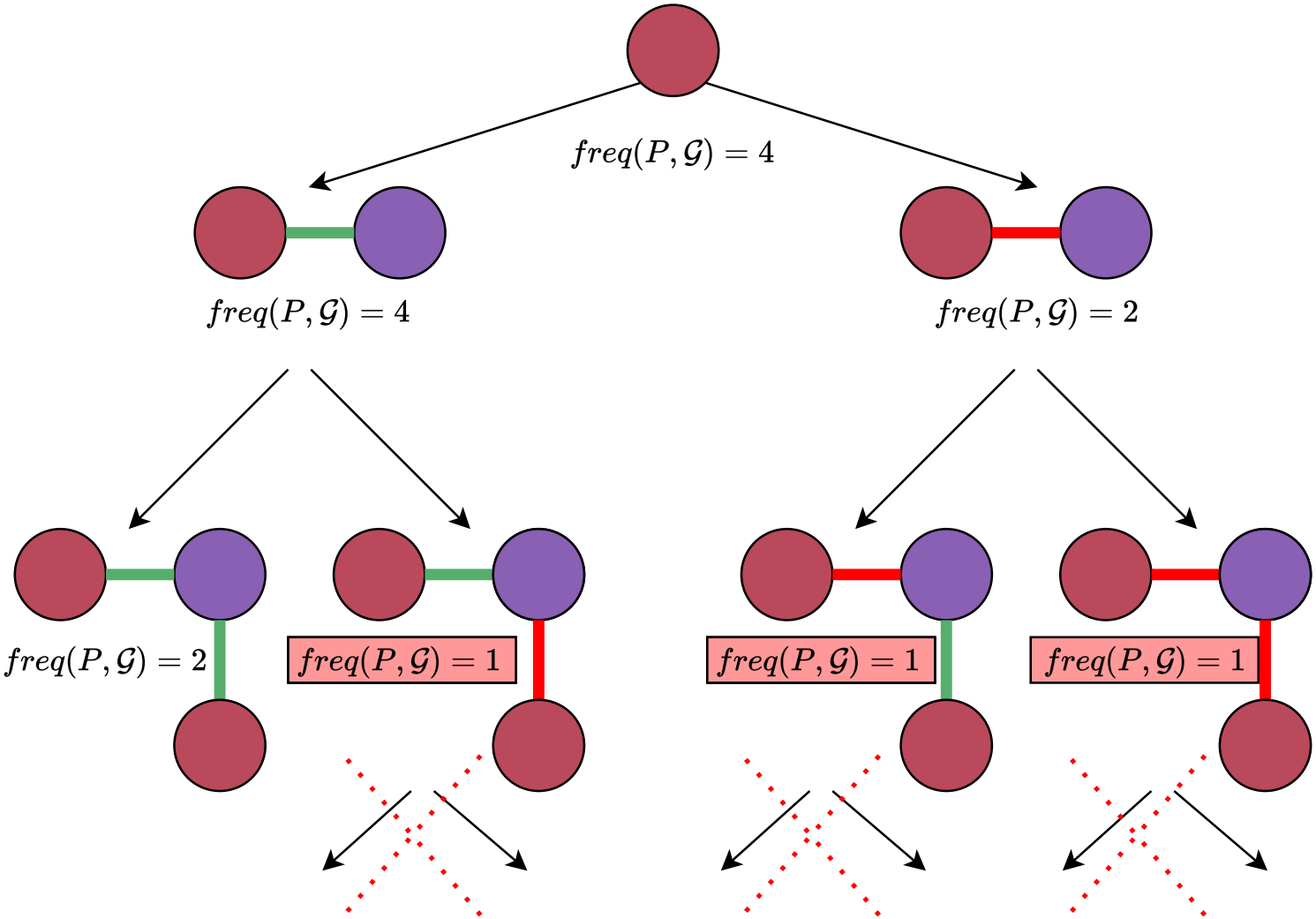
Soit P' un sur-motif de P et $\mathcal{G}$ un ensemble de graphes. On a alors $supp(P, \mathcal{G}) \geq supp(P', \mathcal{G})$

Inutile de continuer si un motif est peu fréquent.

gSpan : sous-graphes identiques

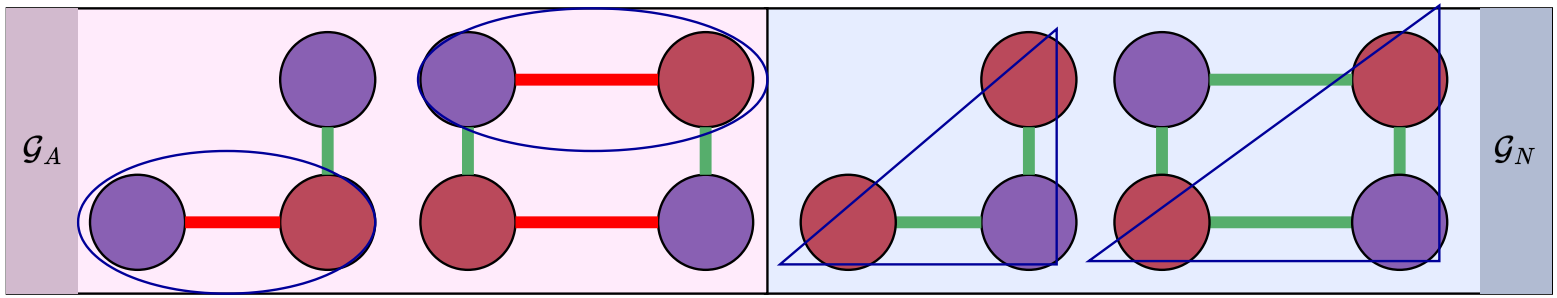


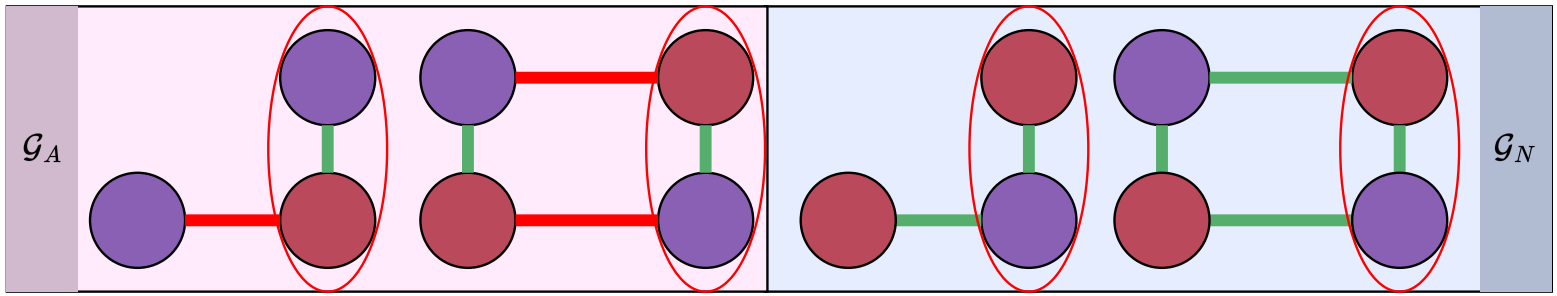
gSpan : sous-graphes peu fréquents



Motif discriminant

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



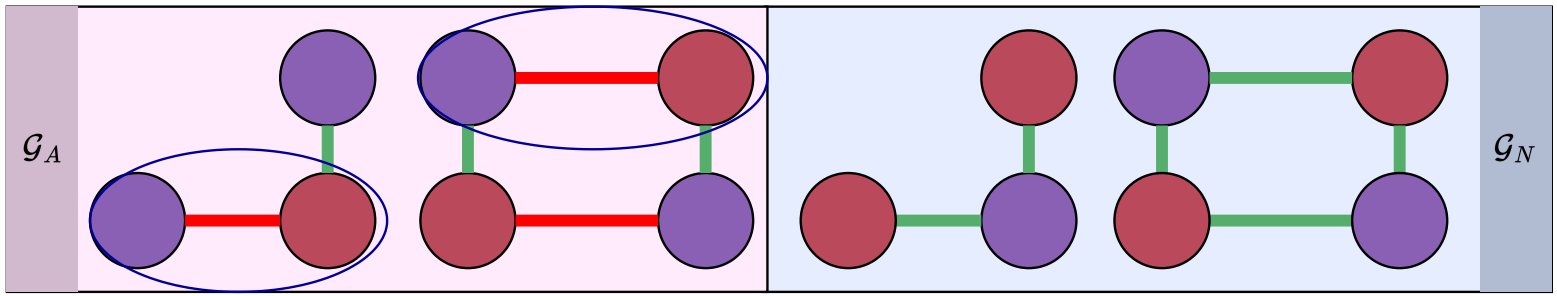


Extraire les motifs fréquents ne suffit pas !

Motif discriminant

Definition (Score discriminant)

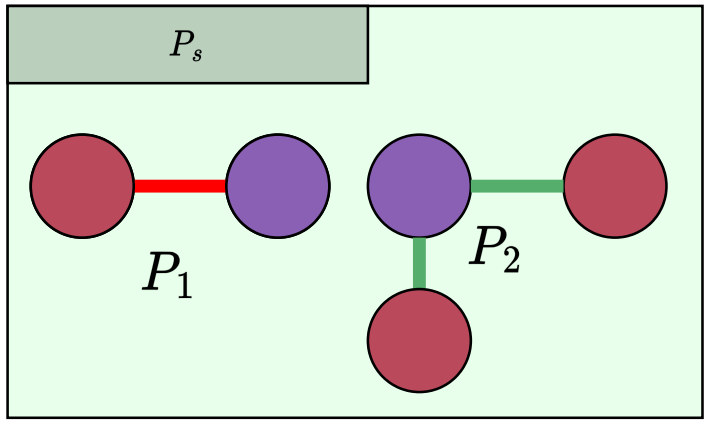
Étant donné un motif P de $\mathcal{G}$, le score discriminant de P est défini par $dis(P) = |supp(P, \mathcal{G}_A) - supp(P, \mathcal{G}_N)|$.



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

Méthode PANG

- Identification d'un ensemble de motifs discriminants $\mathcal{P}_s$
- Représentation vectorielle
- Utilisation de méthodes usuelles supervisées pour classer les vecteurs obtenus.



Méthode PANG

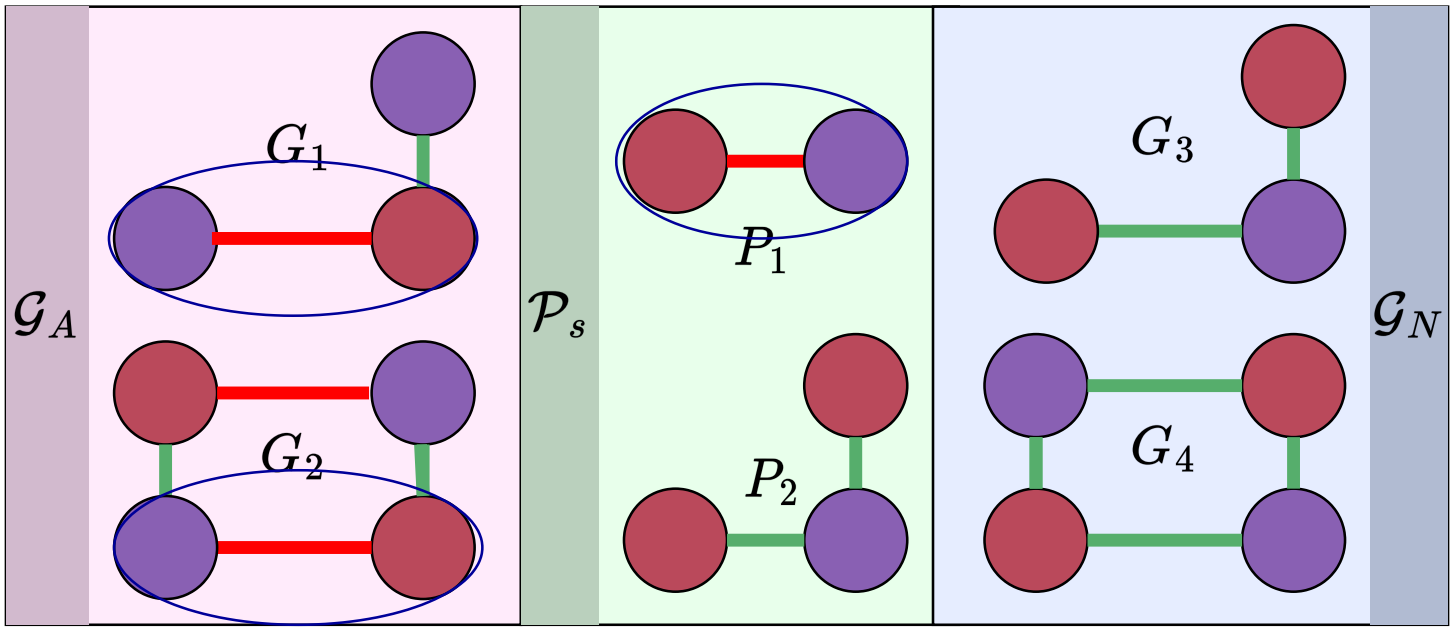
- Identification d'un ensemble de motifs discriminants $\mathcal{P}_s$
- Représentation vectorielle
- Utilisation de méthodes usuelles supervisées pour classer les vecteurs obtenus.

Représentation vectorielle

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

Méthode PANG

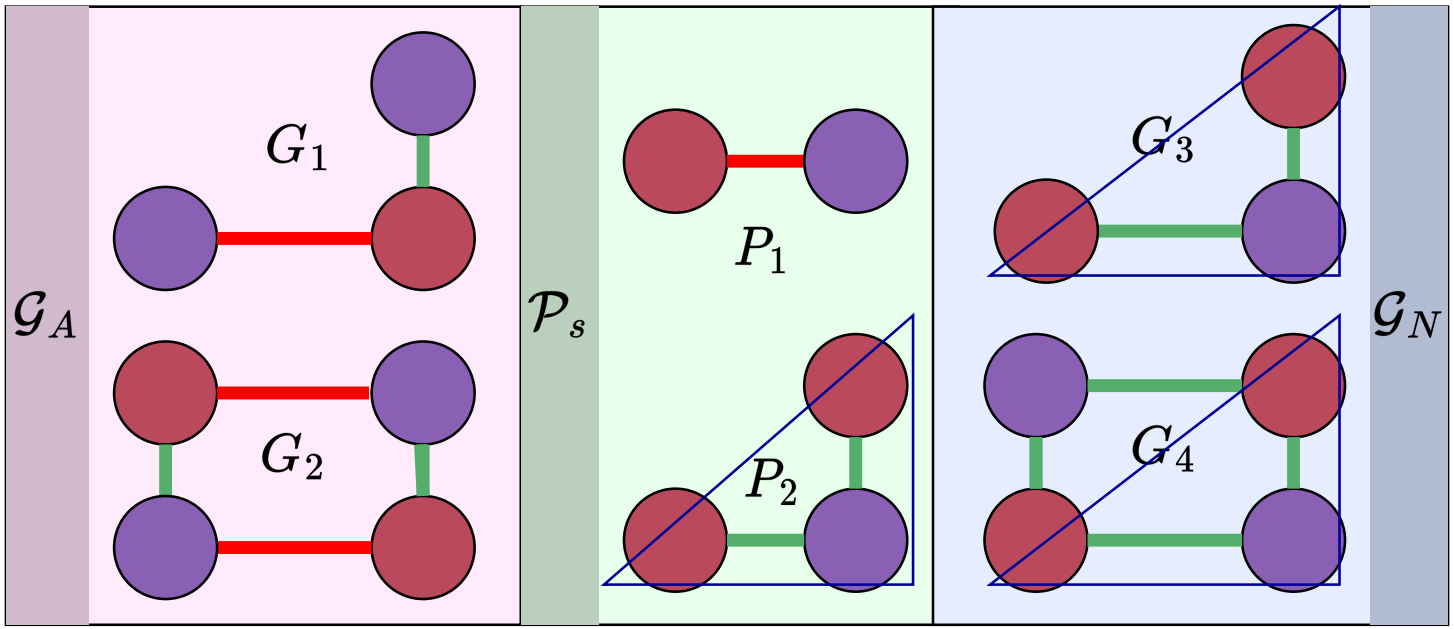
- Identification d'un ensemble de motifs discriminants $\mathcal{P}_s$
- Représentation vectorielle
- Utilisation de méthodes usuelles supervisées pour classer les vecteurs obtenus.



$h_1 = (1, 0)$ $h_2 = (1, 0)$ $h_3 = (0, 1)$ $h_4 = (0, 1)$

Méthode PANG

- Identification d'un ensemble de motifs discriminants $\mathcal{P}_s$
- Représentation vectorielle
- Utilisation de méthodes usuelles supervisées pour classer les vecteurs obtenus.



$h_1 = (1, 0)$ $h_2 = (1, 0)$ $h_3 = (0, 1)$ $h_4 = (0, 1)$

Méthode PANG

- Identification d'un ensemble de motifs discriminants $\mathcal{P}_s$
- Représentation vectorielle
- Utilisation de méthodes usuelles supervisées pour classer les vecteurs obtenus.

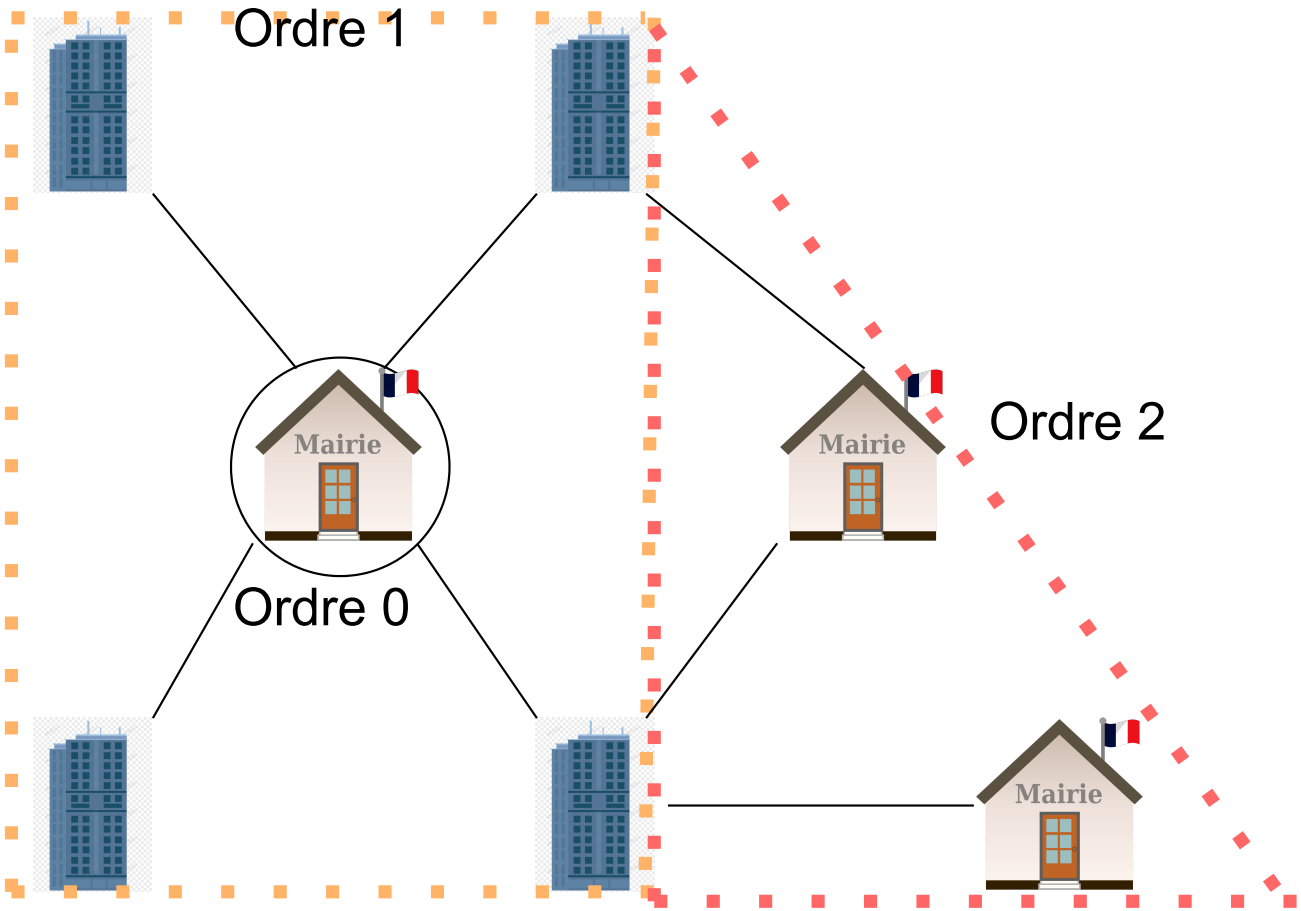
	h	**L**
G_1	1 0 1 1	A
G_2	1 0 0 1	A
G_3	0 1 0 0	N
G_4	0 1 0 0	N
G_5	0 1 1 0	?

Expérimentation sur le jeu des marchés publics

- Utilisation des données de FOPPA [Pot+22].
- Restrictions pour obtenir des graphes de taille réduites :
  - Temporel : plage d'un an.
  - Géographique : restriction au département.
  - Secteur d'activité : marché de travaux (CPV 45).
  - Pas plus de 8 relations pour une municipalité (restriction du degré du nœud.

Expérimentation sur le jeu des marchés publics

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



Expérimentation sur le jeu des marchés publics

Définition des graphes :

- 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.

Expérimentation sur le jeu des marchés publics

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, K-Means).
- Utilisation de la validation croisée (5-Folds).

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

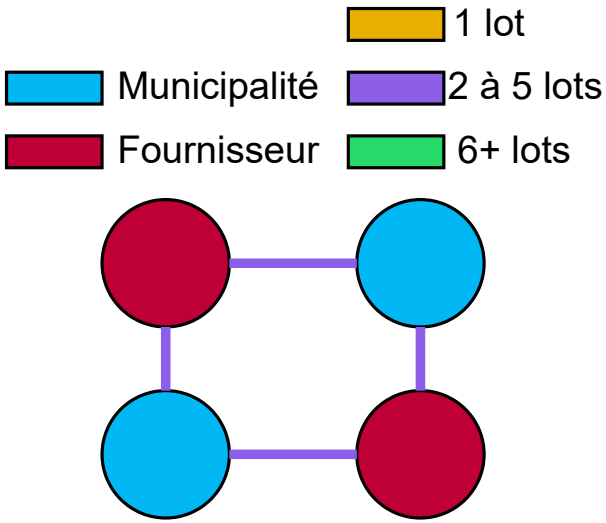
Résultats avec Random Forest selon la valeur de s.

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

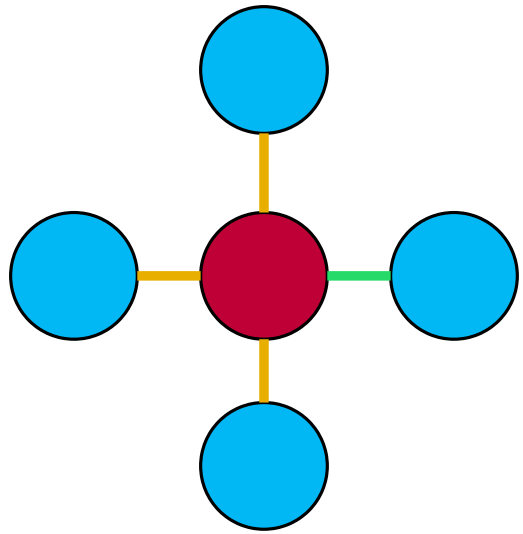
En utilisant 1% des motifs, les résultats sont équivalents à 95% de la performance totale.

Résultats

Objectif de notre approche : relier les motifs à des comportements économiques.



Motif lié à des secteurs sans grosse concurrence.



Motif pouvant être lié a du favoritisme.

Perspectives

- Peut-on choisir d'autres types de motifs ?
  - Motifs fermés
  - Motifs induits
  - Nombre d'occurrences d'un motif
- Miner directement les motifs discriminants.
- Définition des graphes.

Merci pour votre attention, avez vous des questions ?

AVIGNON
UNIVERSITÉ

Bibliographie

[MRM21] 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.

[Nat16] 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.

[Pot+22] L. Potin, V. Labatut, R. Figueiredo, C. Largeron et P.-H. Morand. *FOPPA : a database of French Open Public Procurement Award notices.* Rapp. tech. Avignon Université, 2022. url : https://hal.archives-ouvertes.fr/hal-03796734 (visité le 12/10/2020).

[YH] X. Yan et J. Han. « gSpan : graph-based substructure pattern mining ». In : *2002 IEEE International Conference on Data Mining*. IEEE Comput. Soc. doi : 10.1109/icdm.2002.1184038.