Extraction de motifs pour la détection d'anomalies dans des graphes : application à la fraude dans les marchés publics

Lucas Potin, Rosa Figueiredo, Vincent Labatut, Christine Largeron

Extraction de motifs pour la détection d'anomalies dans des graphes : application à la fraude dans les marchés publics

Lucas Potin*, Rosa Figueiredo*,
Vincent Labatut*, Christine Largeron**

* : {prénom.nom}@univ-avignon.fr,
** : christine.largeron@univ-st-etienne.fr

Résumé. Dans le cadre des marchés publics, il existe plusieurs indicateurs, appelés red flags, permettant d'estimer le risque de fraude. Ces red flags sont calculés en fonction des attributs spécifiques de chaque contrat et sont ainsi dépendants du bon remplissage des notices d'attributions. Dans cet article, nous proposons une méthode basée sur l'extraction de motifs pour la détection d'anomalies dans des graphes. Cette approche générique vise à identifier les sous-graphes associés à la présence de red flags, afin de construire un ensemble de nouveaux indicateurs. Ces motifs peuvent ensuite être utilisés dans les cas où l'information sur les red flags est manquante. Nos expériences confirment que la plupart des red flags peuvent être retrouvés en ne considérant qu'un petit pourcentage de ces motifs.

1 Introduction

Les marchés publics désignent l'achat de biens et de services par une autorité publique (le client), auprès d'une personne morale de droit public ou privé (le fournisseur). Lorsqu'un marché dépasse les seuils européens, l'appel d'offres ainsi que l'avis d'attribution de ce marché doivent être publiés au *Journal Officiel de l'Union Européenne* (JOUE). La version en ligne de ce journal, appelée *Tenders Electronic Daily*[1] (TED), publie plus de 650 000 avis de marché par an[2]. Par conséquent, le secteur des marchés publics fournit une énorme quantité de données accessibles au public.

Historiquement, les anomalies dans les marchés publics, qui font référence à des comportements suspicieux, sont liées à des caractéristiques spécifiques associées aux contrats. Dans la littérature, ces caractéristiques sont appelées *red flags*, et sont utilisées comme indicateurs de fraude potentielle (Fazekas et Tóth, 2014 ; Ferwerda et al., 2017) En pratique, un red flag correspond à une certaine valeur d'un ou plusieurs attributs. Par exemple, la modification du prix du contrat en cours de procédure, ou la réception

1. https://ted.europa.eu/
2. https://ted.europa.eu/TED/main/HomePage.do

d'une seule offre pour un appel d'offre donné, sont typiquement considérés comme des red flags (National Fraud Authority, 2016). Mais les informations nécessaires au calcul de ces red flags ne sont pas toujours disponibles. Dans les données françaises du TED, certains attributs sont largement absents (Potin et al., 2022), tels le nombre de réponses à un appel d'offres (vide pour 30% des lots) ou les informations sur la sous-traitance (52%), ce qui ne permet que de calculer des red flags *partiels*.

Dans le domaine de la détection de fraude dans les marchés publics, la plupart des études sont basées sur des données *tabulaires* (Carvalho et al., 2013 ; Carneiro et al., 2020), c'est-à-dire que chaque contrat est considéré séparément. Seuls quelques rares auteurs tentent de tirer parti des *relations* entre les contrats en adoptant une approche basée sur les graphes. Fazekas et Tóth (2016) proposent le CRI, un score composite de plusieurs red flags, mais n'utilisent les graphes que pour visualiser la distribution de ce score. Wachs et Kertész (2019) considèrent des graphes afin d'estimer la proportion de red flags chez les agents principaux, c'est-à-dire les clients et fournisseurs ayant les relations les plus fréquentes. Cependant, à notre connaissance, il n'existe pas de méthode dans la littérature, basée sur les graphes pour créer des modèles prédictifs. Nous considérons que cette modélisation offre l'avantage d'exploiter toutes les informations disponibles, notamment les relations entre les agents, en gérant également l'incomplétude des données.

Cette hypothèse nous amène à proposer une méthode basée sur les graphes pour détecter les anomalies dans les marchés publics. Elle s'inspire du modèle *sac-de-mots* issu de la recherche d'information, en utilisant des red flags connus pour construire une collection de graphes correspondant aux situations frauduleuses et non frauduleuses, puis extraire des motifs pour identifier les sous-graphes fréquents qui caractérisent ces situations. Ces sous-graphes constituent le vocabulaire utilisé pour décrire les graphes : chacun est représenté par un vecteur dont les valeurs indiquent la présence ou l'absence des sous-graphes. Enfin, nous entraînons un classifieur pour distinguer les deux classes, ce qui permet ensuite la classe des graphes de label inconnu, c'est-à-dire construits avec des contrats où les attributs des red flags sont manquants. Nous évaluons notre approche de manière expérimentale, sur des données des marchés publics français extraites du TED. Nos résultats montrent que notre méthode exploite avec succès les données incomplètes. De plus, elle permet d'expliquer les différences entre les graphes frauduleux et non frauduleux, en observant les motifs les plus discriminants. Enfin, son caractère générique permet aussi de l'appliquer pour identifier des anomalies dans d'autres contextes, et même dans un cadre de classification multi-classes. Notre contribution est triple. Premièrement, nous formulons le problème de détection des graphes anormaux comme un problème de classification standard. Deuxièmement, nous proposons la méthode *PANG* (Pattern-Based Anomaly Detection in Graphs), qui tire parti de l'exploration de motifs pour résoudre ce problème. Troisièmement, nous appliquons cette méthode au contexte de la détection de la fraude dans les marchés publics.

Le reste de l'article est structuré comme suit. La Section 2 introduit la terminologie utilisée dans cet article, ainsi que la formulation de notre problème. La Section 3 décrit notre algorithme PANG. La Section 4 présente le jeu de données et le protocole expérimental utilisé pour évaluer les performances de l'algorithme. La Section 5 décrit et discute nos résultats. Enfin, nous commentons les aspects les plus notables de notre

travail dans la Section 6, et identifions ses principales perspectives.

2 Formulation du problème

Nous adoptons une approche inspirée de la recherche d'information. De la même manière qu'un document peut être modélisé comme un sac-de-mots, nous proposons de représenter un graphe comme un sac de l'ensemble de ses sous-graphes, appelés *motifs*. Étant donné un ensemble de graphes, nous construisons un dictionnaire, composé des motifs apparaissant dans ces graphes. Sur la base de ce dictionnaire, chaque graphe peut être représenté comme un vecteur de longueur fixe, utilisable ensuite en entrée de n'importe quel algorithme classique d'apprentissage automatique. Les motifs ont déjà été employés pour représenter un graphe, notamment dans le domaine de la classification d'images (Acosta-Mendoza et al., 2016).

Dans cette partie, nous décrivons d'abord comment nous définissons cette représentation vectorielle, afin de formuler notre problème comme un problème de classification. Nous présentons notre approche de manière générique, car elle est applicable à différents contextes, puis la mettons en œuvre dans le cas spécifique de la détection des fraudes dans les marchés publics dans la Section 4.

Définition 1 (Graphe attribué) *Un graphe attribué est défini comme un tuple* $G = (V, E, \mathbf{X}, \mathbf{Y})$ *dans lequel* V *est l'ensemble des* n *sommets,* E *l'ensemble des* m *arêtes de* G, $\mathbf{X}$ *une matrice de taille* $n \times d_v$ *dont la ligne* $\mathbf{x}_i$ *est le vecteur de taille* d_v *associé aux attributs du sommet* $v_i \in V$, *et* $\mathbf{Y}$ *une matrice de taille* $m \times d_e$ *dont la ligne* $\mathbf{y}_i$ *est le vecteur de taille* d_e *associé aux attributs de l'arête* $e_i \in E$.

Nous supposons que nous disposons d'une collection de tels graphes, comme illustré dans la Figure 1. Dans cet exemple, chaque sommet possède un attribut (bordeaux ou violet) de même que chaque arête (vert ou rouge).

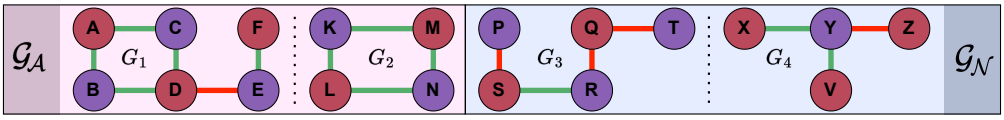

FIG. 1 – Exemple d'une collection de graphes $\mathcal{G}$, incluant les ensembles des graphes anormaux ($\mathcal{G}_A$) et normaux ($\mathcal{G}_N$).

Nous supposons que chaque graphe G possède un label l_G choisi dans $\mathcal{L} = \{A, N\}$, indiquant respectivement un graphe anormal ou normal. Ce label n'est pas connu pour tous les graphes à notre disposition. Soit $\mathcal{G}$ l'ensemble des graphes dont le label est connu. Cet ensemble peut être divisé en deux sous-ensembles disjoints : $\mathcal{G} = \mathcal{G}_A \cup \mathcal{G}_N$ ($\mathcal{G}_A \cap \mathcal{G}_N = \emptyset$). Le sous-ensemble $\mathcal{G}_A$ contient les graphes anormaux tandis que $\mathcal{G}_N$ contient les graphes normaux.

Notre objectif est d'entraîner un classifieur en utilisant les labels connus, afin de prédire les labels inconnus pour les graphes sans label. Pour ce faire, nous utilisons l'extraction de motifs pour construire une représentation vectorielle des graphes, qui

peut être exploitée par les méthodes classiques de classification supervisée. Nous introduisons maintenant les notions relatives aux motifs qui sont requises pour définir notre représentation.

Définition 2 (Motif d'un graphe) *Soit G un graphe attribué. Un graphe P est un motif de G s'il est isomorphe à un sous-graphe P' de G, i.e. $\exists P' \subset G : P \cong P'$*

Comme nous ne considérons que des graphes attribués, nous utilisons la définition de l'isomorphisme de graphe proposée par Hsieh et al. (2006), c'est-à-dire qu'un isomorphisme doit préserver non seulement les arêtes, mais aussi les attributs des sommets et des arêtes. Nous considérons que P est un motif d'un ensemble de graphes $\mathcal{G}$ lorsque P est un motif d'au moins un de ses graphes. La Figure 2 représente trois exemples de motifs de l'ensemble de graphes de la Figure 1.

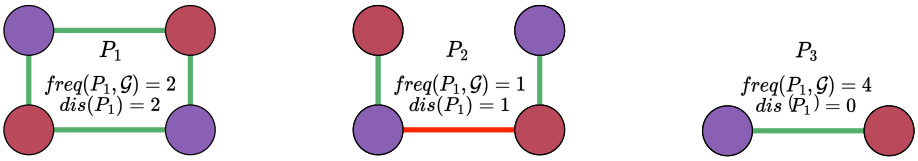

FIG. 2 – Exemple de motifs présents dans le graphe G_1 de la Figure 1.

Nous définissons $\mathcal{P}_A$ et $\mathcal{P}_N$ comme les ensembles de motifs de $\mathcal{G}_A$ et $\mathcal{G}_N$. De plus, nous notons $\mathcal{P} = \mathcal{P}_A \cup \mathcal{P}_N$ l'ensemble de tous les motifs de $\mathcal{G}$. Tous les motifs ne sont pas équivalents. Ainsi, parmi ceux de la Figure 2, P_3 est beaucoup plus présent que P_1 et P_2, dans l'ensemble des graphes de la Figure 1. La principale approche utilisée dans la littérature dans ce cas de figure est d'identifier les motifs *émergents* (García-Vico et al., 2017), c'est-à-dire caractéristiques d'une classe par rapport au reste des données. Nous adoptons une approche similaire, également basée sur un score mesurant le pouvoir discriminant des motifs, mais exploitant les motifs propres à chacune des deux classes.

Définition 3 (Fréquence d'un motif) *Soit $\mathcal{G}$ un ensemble arbitraire de graphes attribués. La fréquence $freq(P, \mathcal{G})$ d'un motif P dans $\mathcal{G}$ est le nombre de graphes dans $\mathcal{G}$ ayant P comme motif : $freq(P, \mathcal{G}) = |\{G \in \mathcal{G} : \exists P' \subset G \text{ t.q. } P \cong P'\}|$.*

Dans cette définition, la fréquence est basée sur le nombre de graphes contenant le motif, et non sur le nombre d'occurrences par graphe. En considérant l'ensemble des graphes de la Figure 1, les motifs présentés dans la Figure 2 ont une fréquence dans $\mathcal{G}$ respectivement de 2, 1 et 4, malgré les multiples occurrences de P_3 dans G_1. Nous exploitons ces fréquences pour calculer le score discriminant d'un motif.

Définition 4 (Score discriminant) *Étant donné un motif P de $\mathcal{G}$, le score discriminant de P est défini par $dis(P) = |freq(P, \mathcal{G}_A) - freq(P, \mathcal{G}_N)|$.*

Un score proche de 0 indique un motif aussi fréquent dans $\mathcal{G}_A$ que dans $\mathcal{G}_N$. Au contraire, un score élevé signifie que le motif est plus fréquent dans l'un des deux sous-ensembles. Par exemple, dans la Figure 2, P_1 est plus discriminant que P_3, puisque $dis(P_1) = 2$ alors que $dis(P_3) = 0$.

Nous utilisons ce score pour classer les motifs de $\mathcal{P}$, et sélectionner les s ($1 \leq s \leq |\mathcal{P}|$) plus discriminants afin de construire l'ensemble des motifs discriminants, noté $\mathcal{P}_s$. Le paramètre s permet de contrôler la taille de la représentation vectorielle des graphes. Comme dans l'exploration de texte, nous supposons que la réduction de cette taille, obtenue en ne considérant que les sous-graphes les plus discriminants, peut accroître la performance de classification. Sur la base de $\mathcal{P}_s$, nous construisons une représentation vectorielle $\mathbf{h}_i \in \mathbb{R}^s$ de chaque graphe $G_i \in \mathcal{G}$. Chaque valeur de $\mathbf{h}_i$ mesure l'importance du motif correspondant dans ce graphe spécifique. Nous discutons du calcul de ces valeurs dans la Section 3. Nous obtenons ainsi une matrice $\mathbf{H} \in \mathbb{R}^{|\mathcal{G}| \times s}$ dont la ligne i représente $\mathbf{h}_i^T$. Par conséquent, H_{ij} correspond à la valeur associée au motif P_j dans le graphe G_i.

Sur la base de cette représentation de nos données, notre problème de détection d'anomalies revient à classer des graphes avec des labels inconnus comme anormaux ou normaux. Plus formellement, étant donné, pour chaque graphe $G \in \mathcal{G}$, le label l_G et la représentation vectorielle $\mathbf{h}$, il s'agit d'apprendre une fonction $f : \mathbb{R}^s \to \{A, N\}$, qui associe un label anormal ou normal à la représentation vectorielle du graphe.

3 Algorithme PANG

Pour résoudre ce problème de classification, nous proposons l'algorithme PANG (Pattern-Based Anomaly Detection in Graphs)[3]. Il est constitué de quatre étapes, précédées d'une étape préliminaire consistant à extraire les graphes. Cette partie du processus étant dépendante des données, nous reportons sa description à la Section 4. Les quatre étapes, représentées dans la Figure 3, sont les suivantes :

1. Identification de tous les motifs de $\mathcal{G}$ et construction de $\mathcal{P}$.
2. Sélection des motifs les plus discriminants parmi eux.
3. Construction de la représentation vectorielle de chaque graphe.
4. Utilisation de ces représentations pour entraîner un classifieur.

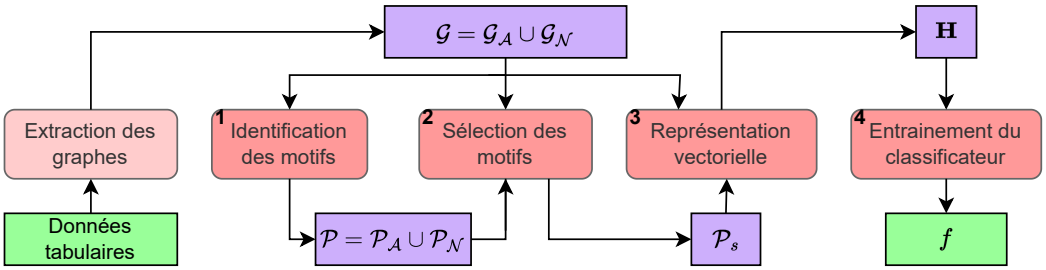

FIG. 3 – Étapes de l'algorithme PANG proposé.

3. https://github.com/CompNet/Pang

Identification des motifs Afin de créer $\mathcal{P}$, nous utilisons un outil existant d'extraction de motifs de graphes. Plusieurs outils de ce type sont disponibles, tels que gSpan (Yan et Han, 2002), FFSM (Huan et al., 2003) ou TKG (Fournier-Viger et al., 2019). Nous choisissons TKG (Fournier-Viger et al., 2019), car selon notre évaluation expérimentale, nous obtenons des résultats identiques aux autres algorithmes, pour un temps de calcul plus court. TKG est implémenté en Java, et est inclus dans le logiciel SPMF (Fournier-Viger et Lin, 2016) [4].

TKG permet de trouver les K motifs les plus *fréquents* dans un ensemble de graphes. Cependant, nous voulons *tous* les motifs, donc nous fixons K de manière à exploiter chaque motif dans notre ensemble de données. TKG s'appuie sur un algorithme itératif, qui part d'un motif fréquent et cherche exhaustivement à l'étendre en ajoutant une arête. Les nouveaux motifs ne sont alors stockés que s'ils sont fréquents.

Sélection des motifs Après avoir obtenu tous les motifs de $\mathcal{G}$ à l'étape précédente, nous calculons leurs scores discriminants comme expliqué dans la Section 2. Nous conservons les s motifs les plus discriminants afin de construire $\mathcal{P}_s$. Ce paramètre a pour but de limiter la taille de l'espace de représentation.

Représentation vectorielle Après la création de $\mathcal{P}_s$, nous construisons la représentation vectorielle de chaque graphe dans $\mathcal{G}$. Ici, plusieurs approches sont possibles. Dans ce travail, nous choisissons de construire un vecteur binaire indiquant la présence ou l'absence de chaque motif dans le graphe considéré. Nous définissons la matrice $\mathbf{H}$ comme suit : pour chaque graphe $G_i \in \mathcal{G}$ et chaque motif $P_j \in \mathcal{P}$, nous attribuons 1 à H_{ij} si ce motif P_j est présent dans G et 0 s'il est absent. Par conséquent, notre représentation actuelle ne permet pas de distinguer les cas où un motif apparaît une ou plusieurs fois dans un graphe. Cependant, elle peut être étendue pour intégrer d'autres pondérations telles que TF-IDF ou BM25 (Amini et Gaussier, 2013). Ainsi, nous pourrions envisager d'utiliser le nombre d'occurrences des motifs dans chaque graphe plutôt que de se contenter de leur présence/absence, comme nous le faisons. Cependant, TKG et les autres outils d'exploration de motifs considérés ne permettent que d'identifier la présence de motifs dans les graphes, sans compter leurs occurrences.

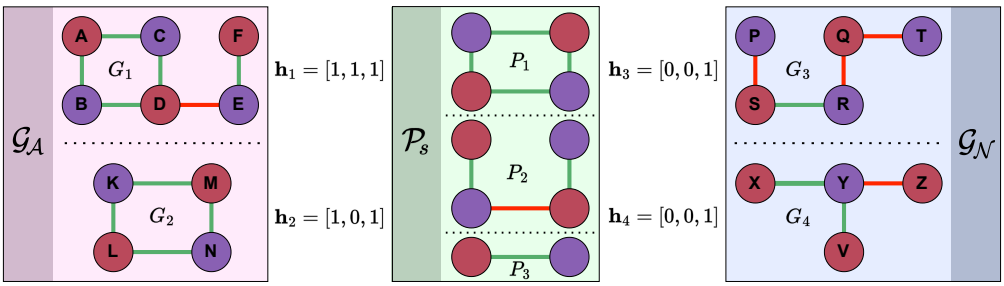

FIG. 4 – Représentation vectorielle des graphes de la Figure 1, pour $s = 3$.

4. https://www.philippe-fournier-viger.com/spmf/

La Figure 4 montre la représentation vectorielle des graphes de la Figure 1. Notons qu'avec notre méthode, deux graphes différents, G_3 et G_4, peuvent avoir la même représentation vectorielle.

Apprentissage du classifieur Après l'étape précédente, nous représentons chaque graphe par un vecteur de taille fixe, quel que soit son nombre de sommets ou d'arêtes. Nous utilisons cette représentation pour entraîner un classifieur qui prédit les labels des graphes. Dans notre méthode, tout classifieur peut être utilisé. Ici, nous nous concentrons sur trois classifieurs standards : Random forest (Ho, 1995), SVM (Cortes et Vapnik, 1995) et k-Nearest Neighbors (Altman, 1992). Nous les appliquons tous et comparons leurs performances lors de notre évaluation expérimentale (Section 5).

En plus d'une bonne performance, nous cherchons également à construire un classifieur explicable, afin d'identifier les motifs qui sont les plus utiles pour faire la prédiction. Ces derniers pourraient être interprétés par un expert du domaine, et fournir des informations sur ce qui distingue les relations frauduleuses des relations normales.

4 Application aux marchés publics

Nous appliquons PANG sur un ensemble de données portant sur les marchés publics, afin de détecter de potentielles fraudes. Dans cette partie, nous présentons d'abord les données brutes, puis le processus utilisé pour filtrer et extraire les graphes attribués.

Données brutes Nous utilisons le site *Tenders Electronic Daily*[5], qui se consacre à la publication des avis d'attributions relatifs aux marchés publics en Europe depuis 2005. Dans le TED, chaque contrat implique *au moins* deux *agents* : un client et un fournisseur, et il peut être constitué de plusieurs lots, qui sont décrits par un ensemble d'attributs tels que le prix ou le nombre d'offres. Cet ensemble de données complet contient des millions d'entrées, mais nous ne nous intéressons qu'au sous ensemble des contrats publiés en France au cours de la période 2015–2019, contenant 417 809 lots. Ce sous-ensemble présente des problèmes au niveau de la qualité des données, notamment le manque de nombreux identifiants d'agents. Nous résolvons ce problème crucial en exploitant une version améliorée du TED (Potin et al., 2022).

Filtrage des contrats Afin d'obtenir un ensemble de graphes, et non pas seulement un seul énorme graphe, nous décidons de ne travailler qu'avec une partie des données disponibles, en les filtrant selon cinq aspects : catégorie d'agent, secteur d'activité, période temporelle, région géographique et taille. En ce qui concerne les agents, nous nous concentrons sur les municipalités, car leur identification est de meilleure qualité que pour les autres agents publics. Pour chaque municipalité présente dans l'ensemble de données, nous définissons un sous-ensemble de contrats contenant non seulement les contrats propres à la municipalité, mais aussi ceux impliquant ses fournisseurs et d'autres municipalités. Les quatre autres filtres nous permettent de contrôler la taille de ces sous-ensembles de contrats, tout en conservant une certaine homogénéité dans

5. https://ted.europa.eu/

les pratiques liées aux contrats : nous ne gardons que des sous-ensembles de contrats de *travaux*, couvrant des périodes d'un an et ne concernant que des municipalités appartenant au même département.

Après ce filtrage, nous obtenons plusieurs sous-ensembles de contrats. Pour chacun de ces contrats, nous calculons un red flag reflétant le risque qu'il soit frauduleux, en utilisant un indicateur standard de la littérature. Nous attribuons un red flag à un contrat si son nombre d'offres reçues est strictement égal à 1, ce qui révèle un manque de concurrence (National Fraud Authority, 2016). Nous comptons également le nombre de relations sans information, c'est-à-dire où le nombre d'offres est manquant pour chaque contrat. Si ce nombre est supérieur à 2, nous considérons que le sous-ensemble ne peut pas avoir de label.

Extraction de graphes Pour chaque sous-ensemble de contrats obtenu après le filtrage, nous extrayons un graphe G. Dans le contexte des marchés publics, en raison de la complexité des données, on peut extraire différents types de graphes (Fazekas et Tóth, 2016), en fonction de ce que représentent les sommets, les arêtes et leurs attributs. Dans notre approche, les sommets représentent les agents, avec un attribut pour distinguer un client d'un fournisseur, et les arêtes représentent la présence de contrats entre des agents, avec un attribut lié au nombre de lots associés à une relation. Nous considérons trois valeurs possibles pour ce dernier attribut : exactement un lot, entre 2 et 5 lots, et 6 lots ou plus. Cette représentation nous permet d'identifier les cas où un client a de nombreux contrats avec un seul fournisseur, un comportement généralement associé à des red flags dans la littérature (Falcón-Cortés et al., 2022).

Afin d'attribuer un label à chaque graphe, nous utilisons les red flags calculés pour chaque contrat. Nous considérons qu'une arête est anormale si elle contient *au moins* un contrat marqué d'un red flag. Le label d'un graphe dépend de son nombre d'arêtes anormales : si ce nombre est inférieur à 2, nous considérons le graphe comme normal (label N), et comme anormal (label A) dans le cas contraire.

Propriétés de l'ensemble de données Notre méthode d'extraction produit 389 graphes normaux et 330 anormaux. Afin d'obtenir des classes équilibrées pour calculer la fréquence des motifs sans biais, nous sous-échantillonnons pour ne garder qu'une partie des graphes normaux, égale au nombre de graphes anormaux.

Label du graphe	**Nombre de graphes**	**Nombre moyen de sommets (e-t)**	**Nombre moyen d'arêtes (e-t)**
Anormal	330	15.76 (5.56)	17.09 (7.86)
Normal	330	12.54 (5.41)	12.59 (6.90)

TAB. 1 – Caractéristiques du dataset.

La Table 1 indique le nombre de graphes pour chaque label, le nombre moyen de sommets et d'arêtes, ainsi que les écarts-types associés pour notre ensemble de données. Les graphes anormaux sont légèrement plus grands que les graphes normaux, en moyenne, car les graphes plus grands sont plus susceptibles de contenir certains

contrats avec red flags. Étant donné la petite taille de l'ensemble de données, nous adoptons une validation croisée à 5 blocs pour évaluer la performance du classifieur.

5 Résultats

Dans cette partie, nous évaluons les performances de PANG sur le jeu de données décrit dans la Section 4. Tout d'abord, nous décrivons les motifs discriminants obtenus avec notre score. Ensuite, nous comparons les performances obtenues avec les classifieurs considérés, ainsi que l'effet du nombre de motifs sélectionnés. Enfin, nous discutons des motifs considérés comme discriminants par les classifieurs.

Motifs discriminants Lorsqu'il est appliqué à notre jeu de données, TKG renvoie un nombre total de 15 793 motifs distincts. La Figure 5.a indique la distribution des scores discriminants. Nous observons que la plupart des motifs (85%) ont un score compris dans $[0; 20]$, et peuvent donc être considérés comme non discriminants. La Figure 5.b présente deux exemples de motifs discriminants P_4 et P_5, de scores respectifs 91 et 64. ils représentent plusieurs relations avec un nombre moyen de lots, qui sont plutôt présentes dans les graphes de grande taille, plus souvent associés au label A.

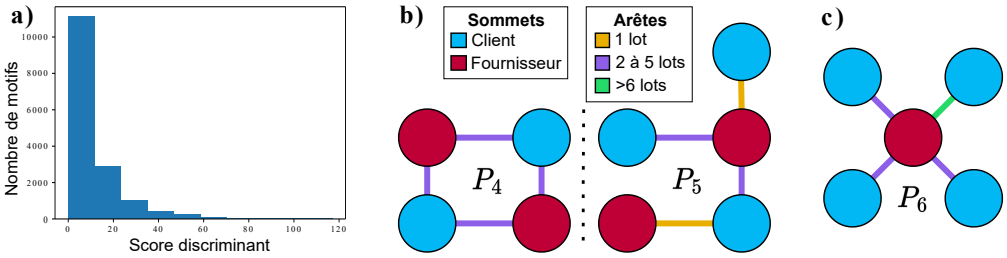

FIG. 5 – (a) Distribution des scores discriminants. (b) Deux exemples de motifs discriminants d'après notre score. (c) Motif lié à du favoritisme.

Comparaison des classifieurs La Table 2 indique la précision (Pre), le rappel (Rec) et le F-score (FS) obtenus avec les trois classifieurs sélectionnés dans la Section 3, pour les deux classes. Il apparaît que Random Forest fournit les meilleurs résultats pour notre jeu de données, et ce, pour les trois métriques.

Type de classifieur	Graphes anormaux			Graphes normaux		
	Pre	Rec	FS	Pre	Rec	FS
RF	**0.94**	**0.90**	**0.92**	**0.89**	**0.93**	**0.91**
SVM	0.85	0.89	0.87	0.86	0.81	0.84
K-Means	0.76	0.69	0.72	0.67	0.74	0.70

TAB. 2 – Résultats pour $\mathcal{P}_s = \mathcal{P}$, i.e. utilisant tous les motifs extraits.

Nombre de motifs Nous nous intéressons maintenant à la performance de Random Forest en fonction de s, i.e. la taille de la représentation vectorielle. Pour comparaison, nous appliquons également ce classifieur à une représentation à base de plongement de graphe entier, apprise automatiquement avec Graph2Vec (Annamalai et al., 2017) [6]

La Table 3 indique les résultats pour chaque métrique selon différentes valeurs de s. La ligne *Tous* signifie que tous les motifs disponibles sont considérés comme discriminants ($s = |\mathcal{P}|$). Nous observons que la construction d'une représentation vectorielle avec seulement 100 motifs, c'est-à-dire moins de 1% du nombre total de motifs, qui est égal à 15 793, est suffisante pour obtenir un résultat supérieur à 80%, avec les trois métriques, et pour les deux classes. Cela représente plus de 90% du F-Score maximum obtenu avec tous les motifs. L'augmentation de s à 150 nous amène à 95% de cette performance. Nous obtenons alors des performances comparables à Graph2Vec, mais notre méthode a l'avantage d'être plus explicable et interprétable, en permettant l'analyse des motifs jugés discriminants par le classifieur.

Nombre	**Graphes anormaux**				**Graphes normaux**			
de motifs	**Pre**	**Rec**	**FS**	**%Max**	**Pre**	**Rec**	**FS**	**%Max**
10	0.69	0.77	0.72	79	0.68	0.59	0.63	69
100	0.84	0.84	0.84	91	0.81	0.81	0.81	88
150	0.89	0.85	0.87	**95**	0.88	0.87	0.87	**95**
Tous	**0.94**	**0.90**	**0.92**	100	**0.89**	**0.93**	**0.91**	100
Graph2Vec	0,88	0.89	0.88	96	0.88	0.86	0.87	95

TAB. 3 – Résultats du classifieur selon la taille de $\mathcal{P}_s$.

Analyses des motifs Notre méthode permet d'identifier directement les motifs les plus discriminants, et donc de tirer parti de l'expertise humaine pour comprendre la signification de ces motifs, du point de vue économique. Dans la Figure 5, P_4 et P_6 sont deux exemples de motifs discriminants d'après Random Forest. P_4 est également discriminant selon notre propre score. Ce motif représente une relation entre deux clients et deux fournisseurs, avec quelques contrats (pas seulement un) entre chacun d'eux. Nous supposons que ces motifs se produisent plus fréquemment dans les graphes avec plus de contrats, ce qui est le cas en moyenne pour nos graphes anormaux. Ce motif est alors considéré comme significatif pour cette classe. Dans le motif P_6, nous observons la présence d'une seule arête verte pour un fournisseur parmi plusieurs clients. Nous pouvons alors interpréter ce phénomène comme un potentiel favoritisme : un fournisseur travaille beaucoup plus avec une mairie qu'avec les autres : la mairie est alors plus susceptible de réaliser des appels d'offres sur mesure pour ce fournisseur, ce qui réduit le nombre d'offres pour un contrat donné.

6 Conclusion

Dans cet article, nous proposons PANG, une méthode utilisant l'extraction de motifs pour représenter des graphes sous forme de vecteurs, et pour les classifier. Nous

6. Implémentation de la bibliothèque Karateclub (Rozemberczki et al., 2020).

l'utilisons ensuite pour détecter la fraude dans les marchés publics, en l'appliquant sur des graphes extraits de données provenant du TED. Nos expériences montrent que PANG est capable d'identifier un ensemble de motifs qui peuvent être utilisés pour représenter chaque graphe sous forme de vecteur. Cette représentation vise à classer les graphes sans information sur les red flags. Elle permet ensuite d'associer des motifs à des comportements humains dans le domaine des marchés publics, comme du favoritisme. Mais, cette méthode générique peut également être employée dans d'autres domaines et pour un plus grand nombre de classes.

Dans le futur, nous prévoyons d'étendre cette approche de plusieurs manières. Tout d'abord, nous voulons améliorer la représentation vectorielle en exploitant le nombre d'occurrences des différents motifs dans le graphe considéré, au lieu de simplement représenter leur présence/absence, similairement à $TF\text{-}IDF$ ou BM25. Deuxièmement, cette représentation pourrait également inclure directement des informations tabulaires dans le vecteur, comme le prix moyen du contrat dans le graphe. Enfin, nous prévoyons de prendre en compte d'autres red flags identifiés dans la littérature, afin d'affiner l'interprétation du modèle discriminant construit.

Références

Acosta-Mendoza, N., A. Gago-Alonso, J. A. Carrasco-Ochoa, J. Francisco Martínez-Trinidad, et J. Eladio Medina-Pagola (2016). Improving graph-based image classification by using emerging patterns as attributes. *Eng Appl Artif Intell 50*, 215–225.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician 46*(3), 175.

Amini, M. R. et E. Gaussier (2013). *Recherche d'Information - applications, modèles et algorithmes*. Algorithmes. Eyrolles.

Annamalai, N., C. Mahinthan, V. Rajasekar, C. Lihui, L. Yang, et J. Shantanu (2017). graph2vec : Learning distributed representations of graphs.

Carneiro, D., P. Veloso, et A. Ventura (2020). Network analysis for fraud detection in portuguese public procurement. In *IDEAL*, pp. 390–401. Springer.

Carvalho, R. N., S. Matsumoto, K. B. Laskey, P. C. G. Costa, M. Ladeira, et L. L. Santos (2013). Probabilistic ontology and knowledge fusion for procurement fraud detection in brazil. In *URSW II*, pp. 19–40. Springer.

Cortes, C. et V. Vapnik (1995). Support-vector networks. *Mach. Learn. 20*(3), 273–297.

Falcón-Cortés, A., A. Aldana, et H. Larralde (2022). Practices of public procurement and the risk of corrupt behavior before and after the government transition in méxico. *EPJ Data Science 11*(1), 1–26.

Fazekas, M. et I. J. Tóth (2014). New ways to measure institutionalised grand corruption in public procurement. Technical report, U4 Anti-Corruption Resource Centre.

Fazekas, M. et I. J. Tóth (2016). From corruption to state capture : A new analytical framework with empirical applications from hungary. *PRQ 69*(2), 320–334.

Ferwerda, J., I. Deleanu, et B. Unger (2017). Corruption in public procurement : finding the right indicators. *Eur. J. Crim. Policy Res. 23*(2), 245–267.

Fournier-Viger, P., C. Cheng, L. Chun-Wei J., U. Yun, et R. U. Kiran (2019). TKG : Efficient mining of top-k frequent subgraphs. In *Big Data Analytics*, pp. 209–226.

Fournier-Viger, P. et J. C.-W. Lin (2016). The SPMF open-source data mining library version 2. In *Machine Learning and Knowledge Discovery in Databases*, pp. 36–40.

García-Vico, A. M., C. J. Carmona, D. Martín, M. García-Borroto, et C. J. del Jesus (2017). An overview of emerging pattern mining in supervised descriptive rule discovery : taxonomy, empirical study, trends, and prospects. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery 8*, e1231.

Ho, T. K. (1995). Random decision forests. In *3rd International Conference on Document Analysis and Recognition*, pp. 278–282.

Hsieh, S.-M., C.-C. Hsu, et L.-F. Hsu (2006). Efficient method to perform isomorphism testing of labeled graphs. In *ICCSA*, pp. 422–431. Springer.

Huan, J., W. Wang, et J. Prins (2003). Efficient mining of frequent subgraphs in the presence of isomorphism. In *3rd ICDM*. IEEE Comput. Soc.

National Fraud Authority (2016). Red flags for integrity : Giving the green light to open data solutions. Technical report, Open Contracting Partnership.

Potin, L., V. Labatut, R. Figueiredo, C. Largeron, et P.-H. Morand (2022). FOPPA : a database of french open public procurement award notices.

Rozemberczki, B., O. Kiss, et R. Sarkar (2020). Karate Club : An API Oriented Open-source Python Framework for Unsupervised Learning on Graphs. In *29th ACM CIKM*, pp. 3125–3132.

Wachs, J. et J. Kertész (2019). A network approach to cartel detection in public auction markets. *Scientific Reports 9*, 10818.

Yan, X. et J. Han (2002). gspan : graph-based substructure pattern mining. In *2002 IEEE International Conference on Data Mining*, pp. 721–724.

Summary

In the context of public procurement, several indicators, called red flags, are used to estimate fraud risk. These red flags are calculated according to certain contract attributes and are therefore dependant on the proper filling of the award notices. In this paper, we propose a general framework based on pattern extraction to detect anomalous graphs. It aims to identify subgraph patterns associated with the presence of red flags, in order to construct a set of new red flag indicators. These patterns can then be used in cases where red flags information is missing. Our experiments show that most of the red flags can be retrieved with a small percentage of patterns.