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, France
{prenom.nom}@univ-avignon.fr
[2] Laboratoire Hubert Curien, France
christine.largeron@univ-st-etienne.fr

Mots-clés : *Graphes, classification, recherche de sous-graphes*

1 Introduction

En France, le secteur du marché public pèse plus de 110 milliards d'euros [1]. Dans un souci de transparence, tout contrat dépassant un seuil monétaire doit être publicisé, aussi bien pour ce qui est de son appel d'offres que de son attribution. Cette publication est réalisée dans un journal officiel : le Bulletin Officiel des Marchés Publics (BOAMP) pour les marchés dépassant le seuil français, ainsi que le Tenders Electronic Daily (TED) pour le seuil Européen.

Dans le TED et le BOAMP, chaque contrat est représenté par un ensemble d'attributs, dont une partie peut être utilisée pour déterminer une suspicion de fraude : par exemple, un appel d'offre qui ne reçoit qu'une seule offre est un signe de risque. Tous ces indicateurs de fraude sont nommés **red flags** [3]. Les travaux actuels sur la suspicion de fraude dans les marchés publics souffrent de deux limites. Premièrement, le manque de certaines données empêche de calculer les red flags pour tous les contrats. Ensuite, l'aspect relationnel des données n'est pas utilisé pour prédire la fraude : les études actuelles [1] utilisent des graphes, mais uniquement pour la visualisation.

Afin de pallier ces problèmes, nous proposons une méthode qui représente des ensembles de contrats sous forme de graphes, puis les caractérise d'après leurs sous-graphes fréquents. Elle peut alors prédire la suspicion de fraude, même en cas de données incomplètes.

2 Extraction de graphe et modélisation

Nous utilisons une version améliorée des données du TED concernant les contrats français [4]. Nous construisons chaque graphe à partir d'un sous-ensemble spécifique de contrats. Les sommets représentent les agents économiques (clients et fournisseurs). Les arêtes représentent les contrats conclus entre les agents. Chaque sommet et chaque arête possède un label décrivant certaines informations supplémentaires : budget d'un agent, temps de publicité, etc.

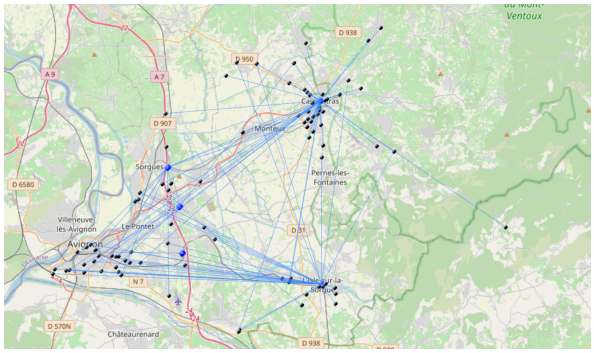
Nous obtenons alors une collection de graphes, dont un exemple est donné en Figure 1. Chaque graphe peut alors être classé dans 3 catégories : red flag, non red flag ou inconnu, selon la présence de red flags ou non dans les différents contrats considérés. En exploitant les sous-graphes, nous cherchons alors à classifier en suspect/non suspect les graphes de catégorie inconnue pour lesquels les redflags sont inconnus.

Pour résoudre cette tâche, nous extrayons les sous-graphes les plus fréquents dans la collection de graphes. Ce problème d'extraction de sous-graphes est NP-complet, car il nécessite de détecter des isomorphismes de sous-graphes. En effet, pour déterminer si un sous-graphe est potentiellement fréquent, il est nécessaire de calculer ses différents isomorphismes dans la collection de graphes. Plusieurs algorithmes tels que cgSpan [5] ou TKG [2] proposent des

1. https://www.economie.gouv.fr/

solutions pour extraire ces sous-graphes fréquents. Nous classons alors les sous-graphes en cherchant ceux qui sont les plus discriminants, c'est-à-dire ceux qui permettent d'affecter une classe spécifique à un graphe, et représentons nos graphes sous forme vectorielle, en utilisant les sous-graphes discriminants trouvés dans la partie précédente. Cette nouvelle représentation des graphes permet alors l'utilisation d'algorithmes de classification usuels.

3 Résultats



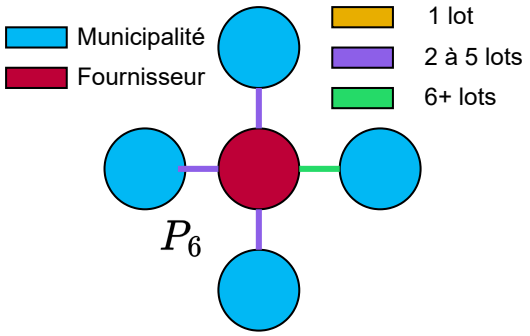


FIG. 1 – a) exemple d'un graphe b) Exemple d'un sous-graphe discriminant

Notre modèle permet, via le score discriminant et l'application d'un algorithme de classification, d'identifier des sous-graphes typiques de graphes avec red flags. La Figure 1 représente par exemple un fournisseur réalisant peu de lots (la couleur des arêtes indique le nombre de lots) avec plusieurs mairies et beaucoup de lots avec une unique mairie. Au niveau économique, ce sous-graphe peut alors être interprété comme une situation de favoritisme. Nous avons réalisé des premiers tests concluants sur un ensemble de 660 graphes, en extrayant plus de 13 000 sous-graphes.

4 Perspectives et Conclusion

Cette méthode permet d'utiliser l'information relationnelle pour déterminer la présence ou non de red flags dans des marchés publics quand l'information associée n'est pas disponible. De plus, cette information peut ensuite être associée à des comportements économiques via l'étude de ses sous-graphes.

Les perspectives futures portent sur l'utilisation de réseaux de neurones de graphes (GNN) et notamment de méthodes explicables parmi ces dernières. De plus, nous souhaitons également travailler sur l'algorithme d'extraction de motifs, pour directement identifier les sous-graphes discriminants au lieu de fréquents.

Références

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

[2] P. Fournier-Viger, C. Cheng, L. Chun-Wei J., U. Yun, and R. U. Kiran. TKG : Efficient mining of top-k frequent subgraphs. In *Big Data Analytics*, pages 209–226. Springer, 2019.

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

[4] L. Potin, V. Labatut, R. Figueiredo, C. Largeron, and P.-H. Morand. Foppa : a database of french open public procurement award notices. Technical report, Avignon Université, 2022.

[5] Z. Shaul and S. Naaz. cgspan : Closed graph-based substructure pattern mining. *CoRR*, abs/2112.09573, 2021.