FORMULAIRE DE STAGE 2020 / 2020 INTERNSHIP FORM

Intitulé du stage / *Internship title* — Analyse de graphes signés pour détecter la corruption dans les marchés publics / *Analysis of signed graphs for the detection of corruption in public procurements*

Encadrants / *Advisors* — Rosa Figueiredo <rosa.figueiredo@univ-avignon.fr>
Vincent Labatut <vincent.labatut@univ-avignon.fr>

Lieu du stage / *Location* — Laboratoire Informatique d'Avignon (LIA), Avignon, France

Descriptif du stage / *Internship description*

Contexte / Context

L'ouverture massive des données publiques recouvre une importance particulière dans le cadre des marchés publics, et la récente réforme de ces marchés introduit explicitement l'open data dans la commande publique. Ce sont donc dès aujourd'hui une masse importante, et à court terme l'intégralité des données relatives aux marchés publics qui peuvent faire l'objet d'analyses qui jusqu'à présent se heurtaient à l'inexistence ou à la partialité de jeux de données. Que ce soit par les montants en jeu (la commande publique correspond à presque 15% du PIB, 200 milliards d'euros par an pour les seuls marchés publics), ou par les bénéfices économiques et sociétaux escomptés derrière cet impératif de transparence, le traitement et l'analyse de ces données constituent un enjeu tant académique que sociétal majeur.

Le travail motivant la présente offre de stage s'insère dans les deux axes transversaux du LIA : *Systèmes Complexes* et *Société Numérique*. À l'intersection de l'informatique, des sciences économiques et des sciences juridiques, il fait partie de **DéCoMaP** (Détection de la corruption dans les marchés publics) un projet à plus grande échelle financé par l'ANR et visant à collecter, traiter et analyser les données ouvertes relatives aux marchés publics français, afin d'élaborer un outil basé sur les méthodes de graphes signés pour la détection des risques de corruption qui peuvent exister entre acheteurs publics et entreprises. La prise en compte de la nature duale des liens unissant entreprises et donneurs d'ordres (relation contractuelle normale ou corruption) est fondamentale à la bonne modélisation du système étudié, et nécessite d'utiliser des réseaux signés. Il s'agit d'un type de réseau bien moins exploré que les réseaux non-signés classiques, et permettant d'inclure dans un même modèle des relations antagonistes.

L'élaboration d'outils de détection automatique des risques de corruption dans les marchés publics ("red flagging") n'est en soit pas totalement nouvelle, et se développe depuis quelques années déjà [1], notamment à l'échelle européenne [2]. Outre qu'aucun outil n'a à ce jour été adapté au cadre juridique et aux données des marchés publics français, les méthodes appliquées présentent une limite importante. Les approches existantes se focalisent sur des données individuelles, caractérisant de façon indépendante clients et fournisseurs, et ignorent les informations relationnelles, correspondant aux interdépendances et interactions entre ces différents acteurs. De ce fait, les outils produits passent à côté d'un certain nombre de propriétés émergentes, i.e. présentes à une granularité plus élevée que l'acteur isolé.

The advent of public open data is particularly important in the context of public procurements, as recent laws force public institutions to publish their public procurement-related data. Large amounts of data are already available, and soon the activity of this whole sector will be described and ready to be used by anyone. One can already perform analyses which were impossible not long ago, as it was difficult or even impossible to access these data. Processing and analyzing public procurements data is a major academic and societal challenge, due to the considerable amounts at stake (public procurements correspond to approximately 15% of the GDP, i.e. 200 billion Euros a year for France) as well as the economical and societal benefits expected from this imperative of transparency.

This internship is related to two themes developed at the LIA (Avignon Computer Science Lab): Complex Systems and Digital Society. It is connected to Computer Science, Economics, Law, and is a part of ***DéCoMaP*** *(Detection of corruption in public procurements), a larger project funded by the ANR (French NSF) and aiming at retrieving, processing and analyzing open data related to French public procurements, in order to design a tool able to assess corruption risks between public buyers and suppliers. In order to obtain reliable models, it is fundamental to take into account the dual nature of interactions between suppliers and buyers (i.e. sane vs. corrupted contractual relation). A good way of doing so is to use signed graphs as models. Unlike unsigned graphs, signed graphs allow representing antagonistic relationships. Moreover, they are much less studied than their unsigned counterparts, so there this domain is very open.*

Designing automatic tools for the assessment of risks of corruption in public procurements is a task called "red flagging". It is not completely new, as some teams have been working on it for a few years [1], especially at

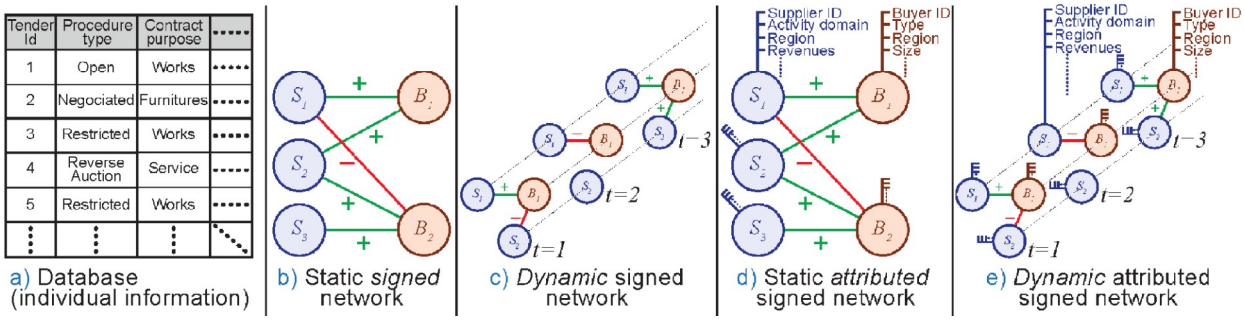
the European level [2]. However, none of them applies to French public procurements, as they do not handle its specifics (legal framework, nature and form of the available open data). Moreover, existing approaches focus on individual information, which characterize buyers and suppliers independently, and ignore relational information, which corresponds to interactions and interdependencies between these agents. Consequently, these approaches potentially miss a number of emerging features in the studied system, i.e. properties which exist only at a higher granular level than that of the isolated agent, and which can be essential to fully understand they system.

Travail demandé / Tasks

Le travail effectué avant ce stage a permis de collecter et structurer l'ensemble des données issues des marchés publics français, telles que proposées par le *Bulletin Officiel des Annonces des Marchés Publics* (BOAMP) (annonces et avis d'attribution). Il s'agit maintenant d'identifier les données *a priori* pertinentes, et de les enrichir à partir à la fois de l'état de l'art académique (analyse juridique, approche économique théorique et empirique) et de l'expertise d'un Professeur en économie, Pierre-Henri Morand (LBNC – AU), qui porte le projet DéCoMaP. En fonction du déroulement du projet, il est envisageable de compléter la BD existante en exploitant d'autres sources de données telles que TED (l'équivalent européen du BOAMP).

Le stagiaire utilisera **les graphes signés afin de modéliser, visualiser et analyser les réseaux** complexes formés par les entreprises (fournisseurs) et les collectivités (donneurs d'ordres). Il s'agit de graphes dont les liens sont caractérisés par un signe, pouvant être positif ou négatif, afin de représenter des relations antagonistes. Ces graphes seront construits à partir des données obtenues à la première étape. Comme illustré dans la Figure ci-dessous, plusieurs méthodes sont possibles pour effectuer cette opération, que le stagiaire devra mettre en œuvre et comparer.

Une fois les graphes obtenus, ils seront principalement utilisés pour deux tâches. La première est leur **analyse via des outils descriptifs** classiques dans le champ des réseaux complexes (taille, densité, transitivité, centralités diverses…). La seconde vise à **résoudre des problèmes de partitionnement définis sur les graphes signés** en utilisant les méthodes de résolution existantes plus adaptées [3,4] aux réseaux obtenus. Avec les solutions obtenues émergeront des groupes d'acteurs (acheteurs publics et entreprises fournisseurs) susceptibles d'être liés entre eux par des pratiques délictueuses.



The work conducted before this internship allowed us to retrieve and store in a local database all the data regarding French public procurements as described by the French official journal (Bulletin Officiel des Annonces des Marchés Publics - BOAMP). We now want to identify the relevant information, and complete them using both state of the art results (legal analysis, theoretical and empirical economical approaches) and human expertise through Pierre-Henri Morand (LBNC – AU), the economics professor leading the DeCoMaP project. Depending on how the project goes, it can be necessary to enhance the database by leveraging other sources such as TED (European version of the BOAMP).

*The intern will use **signed graphs to model, visualize and analyze the complex networks** representing the interactions between companies (suppliers) and public administrations (buyers). These are graphs whose edges are characterized by a sign, which can be either positive or negative, allowing to represent antagonistic relationships. These graphs will be built based on our database. As illustrated by the above Figure, this task can be conducted through several different methods. The intern will have to implement, assess and compare them.*

*The extracted graphs will be used for two distinct purposes. The first is to perform a **descriptive analysis through standard tools** coming from complex networks analysis (size, density, transitivity, centrality, etc.). The second is to solve various **signed graph partitioning problems** through the application of existing resolution methods [3,4], as well as versions of these methods that the intern will adapt to the characteristics of our own networks. The obtained partitions will be used to identify groups of agents (buyers and suppliers) likely to be connected by unlawful practices.*

Perspectives

Sous réserve qu'à la fois le stagiaire et les encadrants soient satisfaits du déroulement du stage, celui-ci est susceptible de déboucher sur un doctorat prévu dans le cadre du projet ANR DéCoMaP déjà mentionné. Ce doctorat sera co-encadré par Rosa Figueiredo et Vincent Labatut (LIA - AU), comme le stage, et en plus par Christine Largeron (Laboratoire Hubert Curien - Université Jean Monnet).

If both the intern and adivsors are satisfied with the way the internship went, the intern will have the opportunity to start a PhD funded by the previously mentioned DéCoMaP ANR project. This PhD will be co-advised by Rosa Figueiredo and Vincent Labatut (LIA - AU), as for this internship, and additionally by Pr Christine Largeron (Laboratoire Hubert Curien - Université Jean Monnet).

Références / References

[1] Fazekas, Mihály, et István János Tóth. New ways to measure institutionalised grand corruption in public procurement, U4 Brief, n°9 (2014).

[2] Ferwerda, Joras, et Ioana Deleanu. Identifying and Reducing Corruption in Public Procurement in the EU. European Commission, OLAF, 2013.

[3] Figueiredo, Rosa et Yuri Frota. The maximum balanced subgraph of a signed graph: Applications and solution approaches. European Journal of Operational Research, 236(2) : 473-487, 2014.

[4] Labatut, Vincent. Generalized Measures for the Evaluation of Community Detection Methods, International Journal of Social Network Mining, n°2(2015):44–63.

Thématique(s) associée(s) au stage / *Keywords*

Analyse de graphes signés / *Signed graphs analysis*
Visualisation / *Visualization*
Marchés publics / *Public procurements*
Données ouvertes / *Open data*
Partitionnement de graphes / *Graph partitioning*