Laboratoire LIP6

PLDAC - 2021

Modélisation et Interrogation de Réseaux d'Interaction Économiques et Financières

Contacts : Bernd AMANN bernd.amann@lip6.fr

Contexte

L'affaire « Panama Papers » au centre des scandales d'évasion fiscale, de blanchiment d'argent, de corruption ou encore de contournement des sanctions, a démontré qu'il y a un réel besoin de connaître l'écosystème des entreprises citées et ne pas s'arrêter aux entreprises elles-mêmes. C'est dans ce contexte que la 5e directive européenne anti- blanchiment (Directive UE 2018/843) a été mise en place, ayant pour objectif de renforcer les procédures de vérification d'identité et de connaissance du client, et de faciliter la coopération et les échanges d'informations entre les Etats membres mais aussi au sein des groupes d'assurances et de banques.

Le secteur bancaire montre néanmoins encore de graves défaillances dans les systèmes et les processus de lutte contre le blanchiment de capitaux et le financement du terrorisme qu'il est censé mettre en place. Ces défaillances reflètent une inefficacité des outils de conformité employés à l'heure actuelle, notamment le KYC (Know Your Client) où l'examen d'une société se focalise exclusivement sur l'identification et l'analyse de celle-ci à l'exclusion de ses relations avec son écosystème[1].

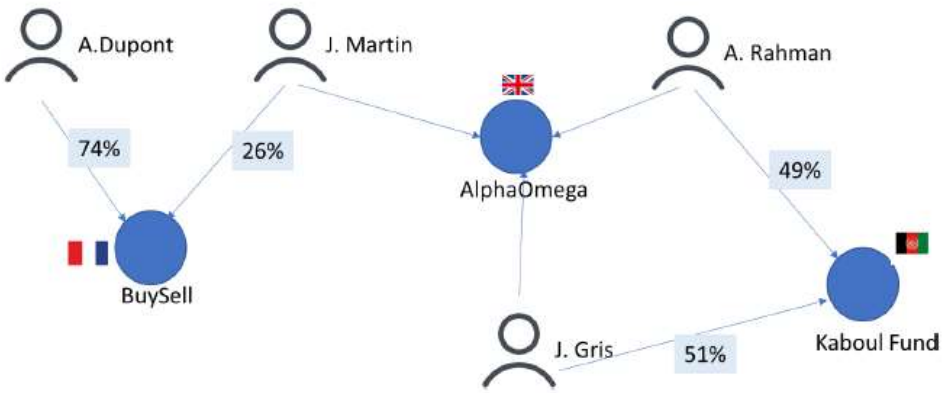
Objectifs généraux

La représentation des données sur les acteurs (sociétés, personnes) et leurs relations (achat, vente, actions, financement) sous forme de Réseaux d'Interaction Économiques et Financières offre de nombreuses possibilités d'analyse complexes pour identifier des schémas frauduleux comme le blanchiment d'argent, le financement illégal ou la fraude fiscal. Ces analyses correspondent en général à des workflows de collecte et de transformation de documents textuels et de données semi-structurées pour extraire des informations sémantiques et temporelles qui sont ensuite analysées par les experts. La définition de ces workflows nécessite une bonne maîtrise de différents modèles et outils informatiques (NLP, bases de données, analyse de données, machine learning).

Le schéma suivant montre que l'entreprise BuySell a un actionnaire, M. Martin, qui s'avère être un bénéficiaire effectif (+25% de parts), et qui détient par ailleurs des actions dans une société AlphaOmega aux côtés de deux autres actionnaires, M. Rahman et M. Gris. A première vue, les logiciels de diligence anti-blanchiment ne verraient rien d'anormal. Un examen plus approfondi néanmoins, montrerait que M.Rahman et M. Gris détiennent à eux seuls une entreprise en Afghanistan, pays classé à risque sur la liste du Gafi et de l'UE.

Le graphe d'écosystème d'une entreprise ne s'arrête pas aux relations actionnariales. Les relations entre sociétés peuvent être de toute nature qui ferait que l'entreprise en relation mérite d'être diligentée dans le contexte de l'entreprise cliente. Les relations peuvent être commerciales (fournisseur, client, partenaire), financières (créancier), ou toute co-occurrence significative (subvention d'une même association, soutien à une même cause etc.). Le graphe d'écosystème devra prendre en compte ces relations, de manière à intégrer la notion du temps pour démontrer leur évolution ou au contraire leur obsolescence.

[1] http://www.revue-banque.fr/risques-reglementations/article/kyc-know-your-client-au-kyn-know-your-%20network



Pour répondre à ces enjeux, le projet PLDAC consiste à étudier des approches de modélisation et l'exploration de réseaux d'interactions économiques et financières. Le modèle de données doit être capable de représenter la sémantique, l'évolution et la qualité des informations (acteurs, relations) et de permettre la formulation de requêtes complexes sur les différentes dimensions (sémantique, évolution, qualité) pour l'exploration de ces informations et identifier des schémas frauduleux.

Travail à réaliser :

- Définition d'un schéma sémantique (ontologie) KYC formalisant les types d'entités (sociétés, personnes, …) et leur relations économiques et financières (acquisition, vente, actionnaire, propriétaire, …).
- *Etat de l'art sur l'extraction d'entités* nommés (NER)
- Développement d'un premier prototype pour
    - Instancier l'ontologie KYC à partir des sources identifiés (csv, XML, JSON) sous forme d'un graphe de connaissance
    - *Interroger et analyser les données extraites*.

Prérequis : bases de données, notebook, python.

Ce projet pourra éventuellement être prolongé sous forme de stage M1.

Références bibliographiques:

[2] ContentCheck: Models, Algorithms and Tools for Data Journalism and Journalistic Fact-Checking https://team.inria.fr/cedar/contentcheck/

[3] Open source tools used: https://www.searchtechnologies.com/blog/panama-papers-open-source-projects