Aide à l'Analyse Visuelle de Réseaux Sociaux pour la Détection de Comportements Suspects

Amyn Bennamane*, Hakim Hacid*, Arnaud Ansiaux*, Alain Cagnati**

*Alcatel-Lucent Bell Labs France
Centre de Villarceaux Route de Villejust, 91620 Nozay, France
{prénom.nom}@alcatel-lucent.com,
**Ministère de l'Intérieur, ST(SI)^2
alain.cagnati@interieur.gouv.fr

Résumé. Cet article traite de l'analyse visuelle de réseaux sociaux pour la détection de comportements suspects à partir de données de communications fournies à des enquêteurs suivant deux procédures : l'interception légale et la rétention de données. Nous proposons les contributions suivantes : (i) un modèle de données et un ensemble d'opérateurs pour interroger ces données dans le but d'extraire des comportements suspects et (ii) une représentation visuelle conviviale pour une navigation simplifiée dans les données de communication accompagnée avec une implémentation.

1 Introduction

Au delà de l'intérêt qu'apportent les réseaux sociaux en ligne et les moyens de communication à l'utilisateur et aux fournisseurs de services, ces moyens ne sont malheureusement pas toujours utilisés de manière licite et peuvent être utilisés, par exemple, dans la synchronisation d'opérations illégales. Une opération illégale peut aller d'une utilisation illégale simple, p.ex. fraude, des moyens de communication, à des formes plus complexes comme le terrorisme. Pour identifier des comportements suspects, les enquêtes sont généralement menées par des enquêteurs qui ne sont pas experts en informatique. En effet, les enquêteurs sont des «Hommes» de terrain et utilisent les outils informatiques en tant que source d'information additionnelle pour vérifier des hypothèses. Dans ce contexte, les tableurs comme Excel sont utilisés pour gérer des listes, faire des tris et des calculs dans l'espoir de construire des relations significatives entre les données, extrayant ainsi des connaissances utiles et des faits. Les tableurs sont très puissants pour analyser des données tabulaires (Lakshmanan et al. (1998)), mais ne sont pas adaptés pour la mise en évidence de patterns dans les données semi-structurées ou non structurées comportant de nombreuses relations, références croisées et métadonnées. À moins que l'utilisateur ne sache ce qu'il est en train de chercher, il est très difficile d'extraire directement de nouvelles connaissances depuis les tables.

Le problème majeur auquel nos utilisateurs sont confrontés est la capacité à extraire des informations utiles depuis des jeux de données de grande taille, hétérogènes, et éparpillés. Ce

travail complexe d'extraction est effectué dans le cadre du projet de recherche VIGIEs [1] dont le but est de fournir aux autorités françaises un outil pour capturer, stocker, et analyser efficacement toutes les informations interceptées de téléphonie fixe, VoIP [2], téléphonie mobile, etc. mais aussi d'Internet. Dans cet article, nous considérons le cas des données de télécommunications (fixes et mobiles) fournies par les fournisseurs de services, sous forme de ce qui est communément appelé *FADETs* [3], aux enquêteurs pour traiter des cas suspects. Nous proposons les contributions suivantes : (i) un modèle de données et un ensemble d'opérateurs pour interroger ces données. Ces opérateurs sont utilisés par des enquêteurs pour formuler des requêtes dans le but d'extraire des comportements suspects et (ii) une représentation visuelle conviviale et à la navigation simplifiée pour les données de communication et un prototype d'implémentation. Cet article est organisé comme suit : la Section 2 présente un bref état de l'art des différents domaines liés à notre travail. La Section 3 décrit notre approche pour l'analyse visuelle de réseaux sociaux. Enfin, la Section 4 conclut et donne des perspectives futures de ce travail.

2 État de l'art

La taille immense des réseaux sociaux en ligne actuels impose des contraintes liées aux mesures que l'on peut appliquer sur ce genre de réseaux pour comprendre les structures sous-jacentes (Du et al. (2010); Faloutsos (2010)). Les mesures existantes n'ont ainsi plus de sens dans un tel contexte (Wasserman et Faust (1994)), et les structures cachées ne peuvent plus être extraites en utilisant les techniques existantes. Cela suppose de travailler davantage dans ces graphes sociaux pour proposer de meilleures techniques et stratégies. Une grosse partie de l'analyse s'effectue actuellement à l'aide de méthodes statistiques et/ou d'apprentissage automatique, ce qui sous-entend une certaine expertise des utilisateurs finaux. Le processus de SNA peut offrir une valeur ajoutée pour plusieurs recommandations de contenus, de services, pour la recherche d'information, les systèmes d'optimisation, la gestion de la vie privée, etc. (Amer-Yahia et al. (2008)). Notre approche est certainement complémentaire à celle-ci dans le sens où nous visons à ne pas obliger l'utilisateur à comprendre les techniques d'analyse mais au contraire il devient de la responsabilité du chercheur (i.e., de l'outil offert par le chercheur) de lui offrir un moyen simple et efficace pour la vérification de ses hypothèses, généralement acquises et accumulées suite à une expérience de plusieurs années.

Une partie du travail exposé dans cet article est lié à la visualisation de données. La visualisation de données est un vaste domaine (de Oliveira et Levkowitz (2003)), et les techniques impliquées vont du simple positionnement de données sur un plan à des représentations plus complexes telles que le voisinage, les communautés, etc. Dans ce travail, nous nous sommes intéressés principalement à un domaine particulier dans la visualisation : les réseaux criminels. La visualisation de réseaux criminels a trouvé un essor notable dans la littérature scientifique après les événements du 11 Septembre 2001. Xu et Chen (2005) présentent un état de l'art des caractéristiques structurelles des réseaux criminels, ainsi qu'une classification des outils de visualisation et d'analyse de réseaux criminels en trois catégories, qui reposent toutes

1. VIGIEs : Visualisation, Interprétation et Gestion des Interceptions Electroniques. Projet de Recherche de l'Agence Nationale de la Recherche (ANR) - programme CSGOSG 2008.
2. VoIP : *Voice over IP*, Voix sur IP
3. FADETs : Factures Détaillées

sur la représentation des données sous forme de graphes, aisément assimilables par l'humain. Notre approche pour l'analyse de réseaux sociaux et la détection de réseaux criminels est une approche semi-automatique dans le sens où nous considérons la génération automatique de graphes sociaux tout en faisant intervenir l'utilisateur dans la mise en place de la chaine d'analyse nécessaire pour la vérification de ses hypothèses.

3 Aide à l'analyse visuelle de réseaux sociaux

3.1 Modèle de données

Les discussions menées avec différents acteurs concernés par l'analyse des interactions suspectes nous a permis d'en apprendre plus sur leur façon de procéder dans leur travail de tous les jours. Les enquêteurs qui cherchent à extraire des informations opèrent souvent de manière séparée suivant le type d'interaction dont ils disposent. Pour être capables de fournir un outil simple avec de hautes capacités d'analyse pour l'utilisateur, nous avons commencé par observer la façon dont travaillent les enquêteurs lorsqu'ils analysent les données de communications. Au niveau logique, les utilisateurs suivent généralement les activités d'une certaine entité, p.ex. une personne. Ces activités peuvent avoir lieu sur différents canaux, p.ex. téléphone, email, transferts bancaires, etc. Une fois que chaque canal est analysé séparément, les conclusions de chaque canal sont agrégées pour avoir une perspective de plus haut niveau sur l'affaire. Intuitivement, alors que les considérations relatives aux interactions basiques servent à tirer des conclusions précises, l'agrégation est utile au niveau de l'interprétation. Ainsi, notre modèle vise à traduire cette observation. Nous commençons par définir une structure spécifique appelée *s-Graph* pour traduire les entités suivies :

Définition 1 ***(Super graphe (s-Graph))*** *Un graphe dirigé, valué et étiqueté, agrégeant plusieurs sous-graphes appelés graphes de propriétés (p-Graph) suivant une relation contenant/ contenu guidée par plusieurs stratégies d'interaction.*

Soit Ω dénotant un s-Graph d'objets, défini comme $\Omega(\overline{V}, \overline{A}, \overline{L}(\overline{V}), \overline{W}(\overline{A}))$, où $\overline{V}$ représente un ensemble de n nœuds du graphe (correspondant à un ensemble de personnes dans ce cas). $\overline{A}$ représente un ensemble d'arcs liant l'ensemble de nœuds du graphe. Il doit être noté que $\overline{A}$ est un ensemble virtuel d'arcs (qui est construit, comme nous le verrons dans le paragraphe suivant, par agrégation de plusieurs arcs venant des sous-graphes). Du fait que Ω est un graphe dirigé et valué, nous pouvons définir une fonction $\omega : \overline{V} \times \overline{V} \rightarrow R^+$ telle que : $\forall \overline{v}_i, \overline{v}_j \in \overline{V}^2, (\overline{v}_i, \overline{v}_j) \in \overline{A}\ iff\ \omega(\overline{v}_i, \overline{v}_j) \in R^+$. Ainsi, ω associe un poids à chaque couple de nœuds qui ont des interactions communes. Ω étant un graphe dirigé, la propriété suivante s'applique pour chaque arc : $\forall \overline{v}_i, \overline{v}_j \in \overline{V}^2, (\overline{v}_i, \overline{v}_j) \neq (\overline{v}_j, \overline{v}_i)$. $\overline{L}(\overline{V})$ représente un ensemble d'étiquettes associées à chaque nœud du super graphe. Intuitivement, comme un s-Graph est une agrégation de graphes, tous ses composants devraient aussi être une agrégation de sous-composants constituant ces sous-graphes participant à l'agrégation. Pour rendre cela compréhensible dans le modèle sous-jacent, introduisons la notion de *propriété communicante* définie comme suit :

Définition 2 ***(Propriété communicante)*** *Une propriété communicante est un attribut qui peut identifier et capturer l'existence d'une interaction entre les nœuds d'un s-Graph.*

Une propriété communicante est similaire à la définition d'un attribut dans le modèle Entité Association, excepté que dans notre cas nous ne considérons comme propriétés que les attributs décrivant des nœuds du s-Graph et qui ont une capacité de connectivité. Cela signifie que les propriétés sont les attributs qui aident à matérialiser les liens entre des individus, comme les numéros de téléphones, des adresses e-mail, des comptes bancaires, etc. Comme une propriété communicante est une partie d'un nœud d'un s-Graph qui permet la liaison à d'autres nœuds, chaque propriété peut être considérée comme une partie séparée du système pour, par exemple, une analyse plus approfondie. En faisant ainsi, nous pouvons construire plusieurs graphes d'après le type de chaque propriété. Nous introduisons alors une structure de graphe que nous appelons *graphe de propriétés* (ou *p-Graph* pour faire court).

Soit $P = \{p_1, ..., p_k\}$ l'ensemble des propriétés communicantes. Considérons T traduisant un ensemble de types comme suit : $T = \{t_i | 1 \leq i \leq s, \forall i, j : t_i \neq t_j\}$. Nous définissons une fonction $\tau : P \rightarrow T$ telle que : $\forall p_i \in P, \tau(p_i) = t_i, t_i \in T$. Comme une propriété communicante est une partie d'un nœud d'un s-Graph qui permet la liaison à d'autres nœuds, chaque propriété peut être considérée comme une partie séparée du système pour, par exemple, une analyse plus approfondie. En faisant ainsi, nous pouvons construire plusieurs graphes d'après le type de chaque propriété. Nous introduisons alors une structure de graphe que nous appelons *graphe de propriétés* (ou *p-Graph* pour faire court).

Définition 3 ***(Graphe de propriétés (p-Graph))*** *Un graphe dirigé, valué et étiqueté qui : (i) lie les nœuds représentant des propriétés communicantes ayant obligatoirement le même type et (ii) matérialise une interaction.*

Formellement, un p-Graph dénoté par G peut être dénoté par $G(V, A, L(V), W(A))$ où V correspond à l'ensemble des m nœuds (comprendre propriétés communicantes d'un type particulier). A est un ensemble d'arcs résultant des connexions entre les propriétés communicantes. Les arcs sont contraints par le type de propriétés comme expliqué dans le paragraphe précédent. Nous revisitons alors la définition des arcs de l'Équation **??** en ajoutant cette contrainte spécifique : $\forall v_i, v_j \in V^2, (v_i, v_j) \in A \; ssi \; \omega(v_i, v_j) \in R^+ \wedge t(v_i) = t(v_j)$.

Enfin, compte tenu du modèle et du contexte, les propriétés suivantes s'appliquent : (i) Un nœud d'un s-Graph peut avoir plusieurs propriétés de chaque type. Par exemple, il n'est pas exclu d'avoir une personne avec de nombreux numéros de téléphones et d'adresses e-mail. (ii) Les nœuds du s-Graph n'ont pas nécessairement le même nombre de propriétés. Ceci peut être considéré comme une conséquence de l'observation précédente. Cette propriété est utile pour traduire la diversité des informations qu'un analyste pourrait avoir sur les objets analysés. Les différents concepts de ce modèle sont illustrés dans la Figure 1.

3.2 Opérateurs pour l'analyse visuelle de réseaux sociaux

Le choix de la stratégie suivante, orientée opérateurs plutôt que fonctionnalités, est motivé par (i) la lourdeur du processus (en termes de ressources et de temps) nécessaire pour implémenter chaque besoin utilisateur comme une fonctionnalité dans l'outil, et (ii) le besoin grandissant de puissance expressive de l'utilisateur qui rejoint et complique (i). Ces opérateurs visent à permettre aux utilisateurs d'exprimer leurs besoins basiques puis, en les combinant, d'exprimer des opérations de plus en plus complexes. Enfin, deux types d'opérateurs sont proposés : (i) des opérateurs de définition de données, dédiés à satisfaire le besoin de création de

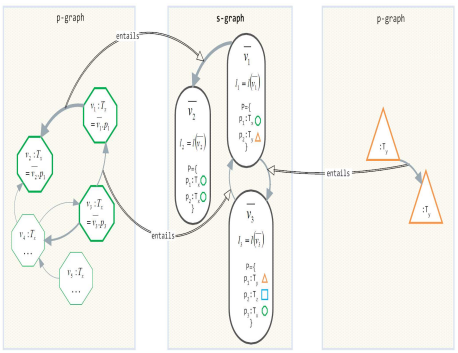

FIG. 1: Illustration des différents concepts du modèle

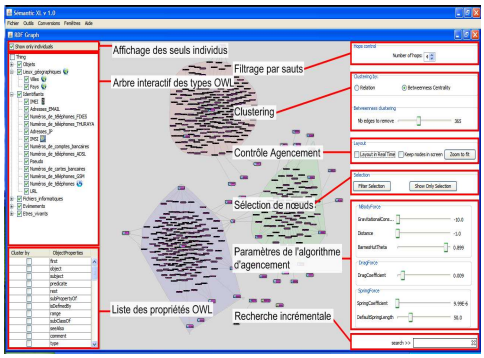

FIG. 2: SemanticXL : Illustration du prototype actuel

composants du modèle de données, p.ex des nœuds, liens, etc. Ce type d'opérateurs est nécessaire, p.ex., pour permettre aux enquêteurs de grouper des interactions en un objet physique. Nous définissons deux opérateurs : *CREATE* et *ASSOCIATE*. (ii) Des opérateurs de manipulation de données décrits ci-après :

Origine et destination d'un arc ($\overrightarrow{\eta}$, $\overleftarrow{\eta}$) : ces deux opérateurs capturent deux des opérations les plus fréquemment effectuées par les enquêteurs lorsqu'ils analysent un réseau de communication : l'origine et la destination d'une communication. Intuitivement, cet opérateur travaille sur les arcs associés à un individu qui sont généralement utile pour comprendre ou capturer des phénomènes sociaux comme les chaînes et les flux d'information.

Union ($\bigcup$) : L'union opère sur plus d'un nœud et traduit le besoin de récupérer des nœuds qui ont participé à différents interactions avec les nœuds sélectionnés ; chose utile pour réduire l'ensemble des nœuds à analyser en détail.

Intersection ($\bigcap$) : cet opérateur récupère les nœuds qui ont été contactés par tous les nœuds spécifiés. C'est une opération importante car elle permet aux enquêteurs de distinguer des nœuds importants et de comprendre les flux d'information, les séquences, et les principaux liens de communications entre individus.

Pour pouvoir tirer profit entièrement des capacités du modèle de données, nous avons étendu la définition de certains opérateurs de base. Les opérateurs qui sont considérés par cette extension sont : le voisinage, l'union, et l'intersection. L'idée basique derrière cette extension est la suivante : pour toute opération entre nœuds de p-Graphs différents (c.-à-d. canaux de communication différents), la réponse à la requête passe forcément par les nœuds correspondants dans le s-Graph. De plus, un jeu de métadonnées est généralement attaché à chaque interaction, p.ex. le temps, la localisation, etc. Ces métadonnées sont actuellement utilisées à travers l'association de ces informations aux liens. En raison du manque de place, nous ne détaillons pas ces opérateurs.

La Figure 2 illustre la version actuelle du prototype SemanticXL. comme signalé précédament, ce prototype dans sa version basique a été implémenté au Ministère de l'Intérieur français comme initiative pour aider les enquêteurs à mieux utiliser les données de communications. Ce travail commun a permis d'améliorer largement le prototype en ajoutant d'autres fonctionnalités et en abstrayant le problème de l'analyse des données de communications. Bien

que les opérateurs ne soient pas complètement implémentés, certains d'entre eux le sont sous forme de fonctionnalités dans cette version et illustrent très bien l'intérêt d'une telle approche. Dans ce qui suit, nous décrivons les différentes fonctionnalités disponibles dans l'outil.

4 Conclusion et perspectives

Nous avons décrit, dans cet article, un modèle et un outil pour l'analyse visuelle de données de communications (c.à.d réseaux sociaux). Nous avons proposé un modèle générique pour l'analyse de communication multicanaux, une contribution très importante dans ce domaine. Le modèle est principalement basé sur l'expérience métier des utilisateurs finaux, c.-à-d. des enquêteurs judiciaires. Ce travail entend aussi apporter une aide aux utilisateurs qui sont plutôt des utilisateurs novices en informatique. Nous avons présenté un prototype implémentant différentes propositions du papier. Comme travail futur immédiat, nous prévoyons de continuer dans l'amélioration de l'outil en offrant une représentation visuelle des opérateurs dans le but d'offrir encore une plus grande flexibilité aux utilisateurs finaux.

Références

Amer-Yahia, S., V. Markl, A. Y. Halevy, A. Doan, G. Alonso, D. Kossmann, et G. Weikum (2008). Databases and web 2.0 panel at vldb 2007. *SIGMOD Record 37*(1), 49–52.

de Oliveira, M. C. F. et H. Levkowitz (2003). From visual data exploration to visual data mining : A survey. *IEEE Transactions on Visualization and Computer Graphics 9*, 378–394.

Du, N., H. Wang, et C. Faloutsos (2010). Analysis of large multi-modal social networks : Patterns and a generator. In *ECML/PKDD (1)*, pp. 393–408.

Faloutsos, C. (2010). Mining billion-node graphs : Patterns, generators and tools. In *ECML/PKDD (1)*, pp. 1.

Lakshmanan, L. V. S., S. N. Subramanian, N. Goyal, et R. Krishnamurthy (1998). On query spreadsheets. In *ICDE*, pp. 134–141.

Wasserman, S. et K. Faust (1994). *Social Network Analysis : Methods and Applications (Structural Analysis in the Social Sciences)* (1 ed.). Cambridge University Press.

Xu, J. et H. Chen (2005). Criminal network analysis and visualization. *Commun. ACM 48*(6), 100–107.

Summary

This paper deals with visual social networks analysis for suspicious behavior detection from large communications data provided by communication services providers for criminal investigators following two procedures: lawful interception and data retention. We propose the following contributions: (i) a data model and a set of operators for querying this data in order to extract suspicious behavior and (ii) a user friendly and easy-to-navigate visual representation for communication data with a prototype implementation.