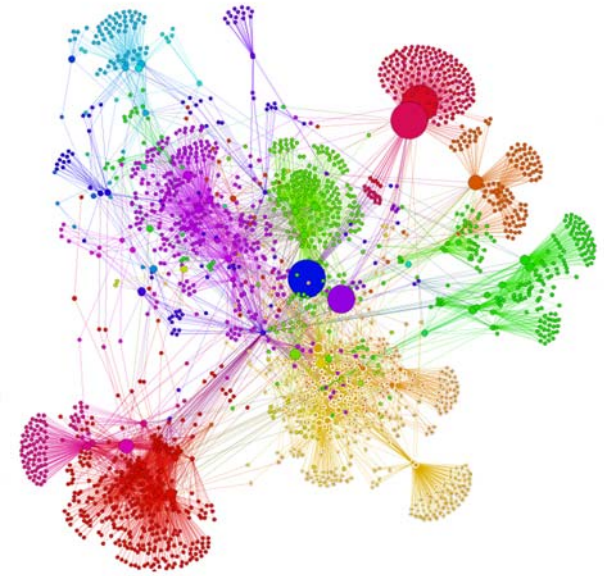
Approche comportementale pour la sécurisation des utilisateurs de réseaux sociaux numériques mobiles

Charles Pérez Perez

Thèse
de doctorat
de l'UTT

Charles PEREZ

Approche comportementale pour la sécurisation des utilisateurs de réseaux sociaux numériques mobiles



Spécialité :
Réseaux, Connaissances, Organisations

2014TROY0019 **Année 2014**

utt
université de technologie
Troyes

THESE

pour l'obtention du grade de

DOCTEUR de l'UNIVERSITE DE TECHNOLOGIE DE TROYES

Spécialité : RESEAUX, CONNAISSANCES, ORGANISATIONS

présentée et soutenue par

Charles PEREZ

le 21 mai 2014

Approche comportementale pour la sécurisation des utilisateurs de réseaux sociaux numériques mobiles

JURY

M. P. PARADINAS	PROFESSEUR CNAM PARIS	Président
M. P. BERTHOME	PROFESSEUR DES UNIVERSITES	Rapporteur
M. B. BIRREGAH	ENSEIGNANT CHERCHEUR UTT	Directeur de thèse
M. M. LEMERCIER	MAITRE DE CONFERENCES	Directeur de thèse
M. J.-F. MARCOTORCHINO	DIRECTEUR DE RECHERCHE THALES	Examinateur
Mme. E. MURISASCO	PROFESSEUR DES UNIVERSITES	Examinateur
M. J.-P. PINTE	MAITRE DE CONFERENCES	Examinateur
M. E. VIENNET	PROFESSEUR DES UNIVERSITES	Rapporteur

Personnalité invitée

M. D. BILLARD	PROFESSEUR

Ma thèse a été financée par le projet CPER Champagne-Ardenne CyNIC (Cybercriminalité, Nomadisme et Intelligence éConomique).

Ce projet CyNIC s'inscrit dans l'UMR STMR (Sciences et Technologies pour la Maîtrise des Risques) de l'Université de Technologie de Troyes. Il a débuté le 1^{er} avril 2011 et se terminera fin 2014. Le projet CyNIC propose d'apporter des solutions innovantes aux problèmes de sécurisation des appareils nomades et notamment des nouveaux téléphones évolués : les smartphones. La première tâche du projet est l'évaluation des équipements d'investigation numérique dédiés aux smartphones. La plupart des outils existants ne prennent en compte que les aspects concernant la téléphonie. Cependant, ces nouveaux appareils sont plus proches de micro-ordinateurs que de simples téléphones. La deuxième tâche est axée sur la problématique de prévention des fuites de données et de protection du patrimoine informationnel dans le cadre de la mobilité. L'émergence rapide des smartphones n'a pas été anticipée par la plupart des entreprises et leur RSSI. La troisième tâche s'intéresse aux solutions de sécurisation des logiciels pour équipements nomades en évaluant les outils existants et en proposant également le développement d'outils plus adaptés aux problématiques d'entreprise. La quatrième tâche est consacrée aux problématiques d'intelligence économique, de vie privée/professionnelle et aux aspects juridiques. Les recherches sont également axées sur des problématiques touchant les sciences sociales comme la perte d'anonymat, le renoncement à la vie privée et les différents recours juridiques. Les partenaires du projet sont l'ICD de l'UTT, l'IRCGN de la Gendarmerie nationale, les entreprises Devanlay et Eutech-SSI, l'Agence pour la diffusion de l'information technologique (ADIT) et le cabinet d'avocats Alain Bensoussan.

Mes travaux de thèse s'inscrivent essentiellement dans la quatrième tâche du projet qui traite de la sécurisation des données privées/professionnelles des utilisateurs de réseaux sociaux mobiles. L'objectif de cette thèse est de mettre en place des outils de protection des utilisateurs contre les acteurs malveillants qui exploitent les réseaux sociaux numériques comme vecteur d'attaque.

Résumé

Notre société doit faire face à de nombreux changements dans les modes de communication. L'émergence simultanée des terminaux nomades et des réseaux sociaux numériques permet désormais de partager des informations depuis presque n'importe quel lieu et potentiellement avec toutes les entités connectées.

Le développement de l'usage des smartphones dans un cadre professionnel ainsi que celui des réseaux sociaux numériques constitue une opportunité, mais également une source d'exposition à de nombreuses menaces telles que la fuites d'information sensible, le hameçonnage, l'accès non légitime à des données personnelles, etc.

Alors que nous observons une augmentation significative de la malveillance sur les plateformes sociales, aucune solution ne permet d'assurer un usage totalement maîtrisé des réseaux sociaux numériques. L'apport principal de ce travail est la mise en place de la méthodologie (SPOTLIGHT) qui décrit un outil d'analyse comportementale d'un utilisateur de smartphone et de ses contacts sur les différents médias sociaux. La principale hypothèse est que les smartphones, qui sont étroitement liés à leurs propriétaires, mémorisent les activités de l'utilisateur (interactions) et peuvent être utiles pour mieux le protéger sur le numérique.

Cette approche est implémentée dans un prototype d'application mobile appelé SPOTLIGHT 1.0 qui permet d'analyser les traces mémorisées dans le smartphone d'un utilisateur afin de l'aider à prendre les décisions adéquates dans le but de protéger ses données.

Mots-clés : criminalité informatique, analyse des données, réseaux sociaux (Internet), smartphones, identité numérique.

A Behaviour-based Approach to Protecting Mobile Social Network Users

Abstract

Our society is facing many changes in the way it communicates. The emergence of mobile terminals alongside digital social networks allows information to be shared from almost anywhere with the option of all parties being connected simultaneously.

The growing use of smartphones and digital social networks in a professional context presents an opportunity, but it also exposes businesses and users to many threats, such as leakage of sensitive information, spamming, illegal access to personal data, etc.

Although a significant increase in malicious activities on social platforms can be observed, currently there is no solution that ensures a completely controlled usage of digital social networks. This work aims to make a major contribution in this area through the implementation of a methodology (SPOTLIGHT) that not only uses the behaviour of profiles for evaluation purposes, but also to protect the user. This methodology relies on the assumption that smartphones, which are closely related to their owners, store and memorise traces of activity (interactions) that can be used to better protect the user online.

This approach is implemented in a mobile prototype called SPOTLIGHT 1.0, which analyses traces stored in users' smartphone to help them make the right decisions to protect their data.

Keywords: Cybercrime, Data analysis, Online social networks, Online identities, Smartphones.

Remerciements

Je tiens à remercier mes deux directeurs de thèse à savoir Marc Lemercier et Babiga Birregah qui ont encadré mon travail de recherche et ont réuni les conditions optimales pour ma progression scientifique.

Je remercie l'ensemble de mon jury de thèse pour les échanges riches ayant eu lieu pendant ma soutenance. J'exprime toute ma gratitude à mes rapporteurs Pascal Berthomé et Emmanuel Viennet, ainsi qu'à mes examinateurs Jean-Francois Marcotorchino, Elisabeth Murisasco, Pierre Paradinas et Jean-paul Pinte.

Je remercie le Conseil régional de Champagne Ardennes ainsi que le Conseil général de l'Aube pour la confiance qu'ils m'ont accordé et leur soutien.

J'adresse mes remerciements à Louis joseph Brossolet qui m'a accueilli chaleureusement à l'UTT dans le cadre de mon double diplôme ingénieur master.

J'exprime ma gratitude et mes pensées sincères à Gérard Sampité qui depuis ma formation à l'ESIEA a encouragé et soutenu ma démarche vers le monde de la recherche.

Je tiens à remercier Arlette Lofficier qui a su éveiller mon intérêt et ma curiosité envers l'enseignement et la pédagogie.

Je remercie Alain Corpel pour l'intérêt manifesté dans mes travaux et pour avoir partagé avec moi son expérience et ses compétences.

Je remercie Rony Germon pour son soutien et son accompagnement lors des différentes phases de réalisation de cette thèse.

J'ai une pensée particulière pour mes parents qui ont effectué d'innombrables sacrifices pour me permettre de réaliser mes objectifs.

Je remercie mon cercle familial et plus particulièrement mes soeurs, mes grands-mères, mes oncles, tantes et mes cousins qui ont toujours été présents pour moi.

Ces remerciements ne peuvent s'achever sans une pensée pour ma femme qui m'a sans cesse soutenu dans toutes mes entreprises.

Table des matières

Table des figures

Liste des tableaux

Introduction

Contexte de l'étude

Notre société doit faire face à de nombreux changements dans les modes de communication. L'émergence simultanée des terminaux nomades et des réseaux sociaux numériques permet désormais de partager des informations depuis presque n'importe quel lieu et, potentiellement, avec toutes les entités (p. ex. personnes, objets, robots) connectées. Cette opportunité nous permet de nous affranchir des infrastructures fixes et d'élargir nos champs de communication. En 2013, selon l'autorité de régulation des communications électroniques et de la poste, plus de la moitié des Français possèdent un smartphone et plus de 40 % s'en servent régulièrement pour accéder aux services en ligne tels que les réseaux sociaux numériques.

Tandis que certains acteurs de la sécurité informatique signalent une augmentation de la malveillance sur les plateformes sociales (p. ex. McAfee, ENISA), il n'existe actuellement aucune solution permettant d'assurer un usage contrôlé des réseaux sociaux numériques dans un cadre mobile. L'utilisateur est donc vulnérable non seulement aux attaques de phishing très présentes sur les plateformes de microblogging telle que Twitter mais également aux attaques par intrusion sur les réseaux sociaux de type Facebook.

Les smartphones et les réseaux sociaux numériques se sont développés et imposés au sein des entreprises. Ainsi, les smartphones sont désormais non seulement des appareils personnels mais aussi professionnels. Les utilisateurs peuvent interagir sur les plateformes sociales depuis leurs smartphones dans un but personnel ou professionnel. On observe ainsi une émergence de l'usage des smartphones ainsi que des réseaux sociaux numériques dans un cadre professionnel (pages professionnelles sur Facebook, réseaux professionnels du type LinkedIn).

Dans ce contexte d'entreprise, cette ubiquité des appareils nomades constitue une opportunité, mais également une plus forte exposition à des menaces. Parmi les menaces les plus redoutées, notons le vol de données à caractère personnel et le vol de données sensibles.

Pour répondre à ces menaces, de nombreuses entreprises ont intégré dans leur politique de sécurité une interdiction d'usage de certains services ou appareils. Ce blocage pur et simple ne constitue pas une solution durable car il conduit souvent à un contournement des règles de sécurité de la part des utilisateurs qui souhaitent quand même bénéficier des facilités d'usage des services.

Ce contexte émergent et ubiquitaire associé à l'usage des réseaux sociaux dans un cadre mobile et professionnel à fait apparaître de nouvelles formes de menaces qui sont à l'origine de la problématique présentée ci-dessous.

Problématique

Les nouvelles technologies de l'information et de la communication, en particulier les smartphones et réseaux sociaux numériques ont permis le développement d'une grande variété de canaux de communication (p. ex. appels, SMS, MMS, chat, blog, wiki). L'utilisateur évolue désormais dans un espace numérique relativement libre où il est possible de communiquer avec n'importe quelle autre entité quel que soit son emplacement géographique et quelle que soit sa nature (p. ex. humain, robot). Cette nouvelle relation avec l'espace et le temps a ouvert de nouvelles portes et a permis l'exploration de nouvelles interactions jusqu'alors impossibles.

Les terminaux nomades ont rendu l'utilisation des réseaux sociaux numériques plus intuitive et rapide. Il est donc possible de générer de l'information de tous types (p. ex. photos, messages, vidéos) sur plusieurs médias (p. ex. Facebook, Twitter) de manière simultanée et instantanée grâce aux applications mobiles.

Il est important de noter que les activités quotidiennes de l'utilisateur génèrent automatiquement des traces stockées à la fois sur les terminaux nomades et sur des serveurs distants. Celles-ci peuvent échapper au contrôle de l'internaute malgré toutes les actions qu'il peut mettre en œuvre pour protéger ses données privées. Les données privées étant à fortes valeurs ajoutés pour les acteurs malveillants, les utilisateurs doivent faire face à de nombreuses actions malveillantes.

La présence de la malveillance sur les réseaux sociaux numériques peut être motivée par de nombreuses raisons. Tout d'abord, la quantité importante d'utilisateurs présents sur les plateformes numériques sociales (plus d'un milliard sur Facebook et plus de 500 millions sur Twitter) peut engendrer des millions de victimes. Cette grande quantité d'utilisateurs est très certainement l'une des plus fortes motivations des cybercriminels. De plus, ces réseaux regroupent une quantité importante d'informations à caractère personnel et sensibles. Les réseaux sociaux numériques constituent une source de données de très forte valeur ajoutée pour les cybercriminels qui vont pouvoir tenter de les collecter et de les exploiter. L'agence européenne chargée de la sécurité des réseaux et de l'information (ENISA) pointe de manière plus précise deux principales origines de ces menaces. Elle indique que ces menaces proviennent du fait que les utilisateurs ont tendance à surévaluer la qualité de leur audience et à sous-estimer la taille de celle-ci. Les utilisateurs n'ont pas nécessairement conscience de la potentielle présence d'entités malveillantes et évaluent mal le nombre de personnes ayant accès à leurs données. Ils naviguent donc en permanence sur de nombreuses plateformes numériques où ils sont à la fois de potentielles victimes d'attaques malveillantes (p. ex. des attaques par phishing) et à la fois eux-mêmes vecteurs de divulgation d'informations sensibles à des contacts qui ne seraient pas légitimes.

Ces phénomènes peuvent être aggravés par certaines caractéristiques des réseaux sociaux numériques. Par exemple, la relation de confiance existant entre les amis d'un réseau social peut être utilisée par un attaquant pour légitimer son action sans éveiller les soupçons de sa victime. Il est possible de créer une quantité importante de profils de manière automatique, même si les plateformes essayent d'empêcher ce phénomène. Ainsi, l'attaquant peut disposer d'une armée virtuelle configurée pour accomplir des actions malveillantes. Ces profils ne divulgueront pas forcément l'identité de l'attaquant. Ceux-ci pourront éventuellement être détruits après utilisation et avant leur détection.

L'échange courant de contenus sur les plateformes sociales et la forte propension des utilisateurs à consommer de l'information rendent possibles les attaques par téléchargement de code source (p. ex. Koobface [1]). Celles-ci peuvent opérer par simple consultation d'une URL publiée dans les messages produits par des acteurs malveillants.

Enfin, certaines caractéristiques observées sur le graphe social (p. ex. le coefficient de clustering, la distance moyenne entre deux utilisateurs) des plateformes, engendrent des possibilités de diffusion et de propagation sans précédent ce qui rend le problème d'autant plus critique.

Dans ce contexte, il devient nécessaire de mettre en œuvre des approches méthodologiques de protection des données de l'utilisateur sur les réseaux sociaux numériques mobiles. Ces approches doivent intégrer une analyse des profils exploitant la plateforme sociale pour propager du contenu malveillant pouvant nuire à la sécurité de l'utilisateur et du système d'information auquel il est associé. Celles-ci doivent aussi permettre une mesure de légitimité des contacts (de l'audience) ayant accès à l'information publiées par l'utilisateur.

La problématique concerne donc la sécurisation des interactions et des données des utilisateurs de réseaux sociaux numériques mobiles. Il est nécessaire de mettre en place des modèles de représentation des médias de communication et des utilisateurs de ces médias. Un point crucial concerne l'intégration, la modélisation et l'exploitation des terminaux mobiles (qui sont au cœur des interactions) pour permettre de mieux sécuriser les données des utilisateurs, notamment en intégrant leur comportement dans les algorithmes proposés.

Les actions mises en place dans cette thèse ont pour objectif la sécurisation des utilisateurs des réseaux sociaux numériques en apportant des approches côté utilisateur. Bien que certaines approches présentées puissent être adaptées pour mener des actions de sécurisation côté opérateur, leurs mises en œuvre au niveau de l'utilisateur permet de tester l'approche plus rapidement et aussi de prendre en compte la spécificité de chaque individu dans la solution. Notons aussi qu'une solution centrée opérateur pourrait être trop dépendante de la plateforme analysée tandis que l'approche proposée dans cette thèse se veut plus généraliste tirant profit de l'ensemble des réseaux sur lequel l'utilisateur est présent est pouvant le sécuriser sur potentiellement l'ensemble de ces plateformes.

Contribution

Dans cette thèse, nous proposons une nouvelle méthodologie nommée SPOTLIGHT pour l'évaluation des profils sur les réseaux sociaux numériques. Cette évaluation a notamment pour but la mesure de la légitimité de l'appartenance de profils à une liste de contacts d'un utilisateur de smartphone. Cette évaluation est critique car la présence de contacts non légitimes dans la liste de contacts d'un utilisateur l'expose à de possibles fuites de données par l'accès illégitime de ses contacts aux données mais aussi par le potentiel accès à du contenu malveillant publié par de tels profils.

La méthodologie proposée se fonde sur deux principaux apports effectués dans le domaine de l'analyse des réseaux sociaux numériques.

Le premier apport est une analyse comportementale des profils de la plateforme Twitter dont l'objectif est de détecter des profils malveillants. Les profils malveillants sont ici des profils propageant des URL malveillantes sur la plateforme sociale. Cette méthodologie, nommée SPOT, est effectuée sur la base de multiples critères comportementaux issus des caractéristiques du profil, mais aussi des messages publiés par celui-ci et par l'analyse des URL contenues dans les messages. L'analyse des URL est cruciale car elle permet l'identification de contenu malveillant pouvant être propagé. Cette approche propose une évaluation de la virulence des profils suspects via une représentation tridimensionnelle intégrant le niveau d'activité, de visibilité et de danger de profils.

Notons que, en collaboration avec l'université de Ballarat, la mise en évidence de campagnes de profils malveillants a été proposée comme une évolution de SPOT. Cette détection se base sur l'identification de profils malveillants ayant un comportement similaire dans l'envoi et le contenu des messages. La similarité entre profils est évaluée par une approche d'attribution d'auteurs, combinée avec des indicateurs de synchronisation comportementaux. Parmi les apports de cette approche, nous avons pu identifier un ensemble de stratégies mises en œuvre par les acteurs malveillants pour augmenter l'efficacité des attaques réalisées sur les réseaux.

Le second apport présenté concerne la modélisation des données sociales d'un utilisateur de smartphone. Nous avons mis en œuvre une modélisation basée sur un graphe multicouche des données sociales issues d'un smartphone. Le graphe multicouche est constitué d'un ensemble de graphes modélisant chaque réseau social (Facebook, Twitter, etc.). Cet ensemble de graphe est interconnecté par la présence d'individus similaires sur différents réseaux sociaux. Ce modèle nous a permis la construction d'un indicateur du niveau de présence dans le smartphone des contacts appartenant au « cercle d'amis » de l'utilisateur. Le niveau de présence d'un individu sur de multiples réseaux sociaux est dénommé imbrication. Plus un individu est présent sur un ensemble de couches important, plus il possède un score d'imbrication élevé. Nous avons, dans ce cadre, proposé un nouvel indicateur de mesure de légitimité des contacts (dénommé SBRA et reposant sur cette imbrication), pouvant s'intégrer à la famille des indicateurs de prédiction de liens.

L'approche générale nommée SPOTLIGHT fournit une méthode d'évaluation des contacts d'un utilisateur de réseaux sociaux numériques mobiles. Celle-ci prend en compte non seulement le niveau de légitimité a priori d'un profil (défini par l'analyse du smartphone) dans la liste de contacts de l'utilisateur, mais aussi son propre comportement afin de mettre en évidence tout profil présentant une potentielle menace pour les données de l'utilisateur. Cette méthodologie s'applique tout autant aux réseaux sociaux centrés sur la notion d'identité tels que Facebook mais aussi sur les réseaux centrés sur le contenu tels que Twitter.

Organisation de la thèse

Ce document de thèse est organisé en trois parties.

La première partie présente le champ disciplinaire de l'analyse des réseaux sociaux numériques et, plus particulièrement, les travaux de l'état de l'art concernant la détection de la malveillance et la mesure de la confiance sur ces réseaux.

Cette première partie est composée de deux chapitres. Le premier chapitre présente une introduction aux réseaux sociaux numériques mobiles et à leur analyse. Le second chapitre présente l'état de l'art sur la sécurisation des données des utilisateurs de ces réseaux.

La seconde partie propose deux contributions apportées par cette thèse pour la sécurisation des utilisateurs de réseaux sociaux numériques. Celle-ci est composée de deux chapitres. Le premier chapitre est consacré à la présentation de SPOT, une méthodologie de détection des acteurs malveillants de la plateforme Twitter. Le second chapitre présente une méthodologie complémentaire fondée sur l'analyse des smartphones pour l'identification de contacts légitimes sur le réseau social Facebook.

Enfin, la dernière partie présente une approche unifiée des deux principales contributions et met en évidence la réalisation d'un prototype d'application mobile.

Nous terminerons par une conclusion générale. Celle-ci présente un bilan des propositions effectuées, ainsi que les limites et les perspectives de recherches associées à ce travail.

Partie I : Contexte de l'étude et état de l'art

« C'est le temps que tu as perdu pour ta rose qui fait ta rose si importante. »

(A. de Saint-Exupéry, 1943)

Chapitre 1

Réseaux sociaux numériques mobiles et sécurité des données des utilisateurs

Résumé du chapitre

Ce chapitre propose une introduction et un état de l'art des réseaux sociaux numériques mobiles. Cette présentation débute par un aperçu historique des réseaux sociaux tels que les concevait J. Barnes dans les années 1950. L'évolution de la discipline de l'analyse des réseaux sociaux est brièvement présentée puis mise en perspective avec les nouvelles opportunités apportées par le succès récent des plateformes du Web social, ou Web 2.0. Un aperçu du paysage des réseaux sociaux numériques est effectué et une présentation des deux plateformes les plus populaires est réalisée. Il se poursuit par la présentation des problèmes de sécurité associés aux interactions des utilisateurs de réseaux sociaux numériques. À ce sujet, dans un premier temps, nous présentons quelques failles traditionnelles de sécurité des systèmes d'information qui ont montré leur efficacité sur les réseaux sociaux numériques. Puis, dans un second temps, nous présentons les failles de sécurité propres à l'usage des réseaux sociaux numériques. Parmi les points clés, nous introduirons la notion de confiance entre deux contacts sur les plateformes sociales. Cette confiance qui est assumée mais pas toujours vérifiée est, en effet, souvent à l'origine de la vulnérabilité des utilisateurs des réseaux sociaux numériques.

1.1 Les réseaux sociaux

Les réseaux sociaux sont définis comme un ensemble d'entités sociales interagissant les unes avec les autres [2]. Ces entités peuvent être des hommes mais aussi des entreprises [3], des états, des animaux, etc. [4].

Parmi les exemples académiques les plus étudiés, on compte, par exemple, le réseau social constitué des relations d'amitié entre les trente-quatre membres d'un club de karaté d'une université américaine en 1970. Celui-ci a été étudié dans le but de comprendre les mécanismes à l'origine de la création de sous-groupes d'individus à l'intérieur du club. Un autre exemple est le réseau social créé à partir de la proximité mesurée entre un ensemble de dauphins dans leur espace naturel. L'objectif de ce réseau social était la mise en évidence de caractéristiques de communication, par exemple la préférence de communication entre animaux de même sexe.

L'analyse des réseaux sociaux correspond à l'étude des individus ou entités sociales et de leurs interactions afin de mettre en évidence certains mécanismes, de comprendre certaines situations et certaines dynamiques entre les acteurs [5].

La première étape nécessaire à une analyse des réseaux sociaux est la collecte de données issues des entités et des relations observées entre celles-ci. Avant l'existence des plateformes numériques, cette étape pouvait être conduite selon deux stratégies. La première était la réalisation d'un questionnaire et son envoi aux personnes concernées par l'étude. Un des plus célèbres exemples est le questionnaire de S. Milgram montrant l'existence en moyenne de seulement six degrés de séparation entre tout habitant des États-Unis [6]. Cette expérience est mondialement connue sous le nom des six degrés de séparation. La seconde méthode de collecte était la simple observation par un spécialiste, depuis l'intérieur d'une communauté, des interactions entre les individus.

Dans les années 1950, l'objectif de certains sociologues était d'effectuer des analyses locales dans l'optique de pouvoir arriver à une analyse plus générale qui permettrait une meilleure compréhension de notre monde. Dans cet esprit, J.-L. Moreno annonce en 1956 dans son ouvrage [7] que le travail d'analyse à une grande échelle ne pourrait se faire que si une plateforme commune d'échange était créée permettant la génération et la mémorisation de traces d'interactions exploitables pour reconstruire une partie du réseau social planétaire. Cette plateforme n'est aujourd'hui ni plus ni moins que les services proposés via l'Internet (pour la première fois mentionné en octobre 1972 par Robert E. Kahn) et prend vie sous la forme de réseaux sociaux numériques (RSN).

1.2 Les réseaux sociaux numériques

D'après la définition de [8], un réseau social numérique est un service s'appuyant sur une infrastructure informatique qui permet à un individu d'accomplir au minimum trois actions fondamentales.

La première est la possibilité de créer un profil en remplissant un formulaire plus ou moins exhaustif sur son identité. Un profil est généralement identifié de manière unique par une adresse de courriel ou un pseudonyme.

La seconde fonctionnalité est la possibilité de gérer une liste de contacts. Les contacts d'un utilisateur sont alors définis comme un ensemble de profils qui ont été identifiés par l'utilisateur. Cette identification de contacts permet d'activer des fonctionnalités spécifiques vis-à-vis de ces profils tels que les interactions via des messages textuels. Les

contacts peuvent être des connaissances acquises en dehors du site (p. ex. dans la vie réelle) ou bien acquises directement sur le réseau numérique. Notons que les réseaux sociaux numériques possèdent désormais, pour la plupart, des systèmes de recommandation qui permettent de proposer à l'utilisateur des profils susceptibles de les intéresser. Ce type de systèmes est le plus souvent élaboré sur la base de contacts existants et/ou des préférences indiquées par l'utilisateur.

Enfin, la dernière fonctionnalité nécessaire à un réseau social numérique est la possibilité de naviguer parmi les amis de nos amis. Cette fonctionnalité est à l'origine d'une certaine dynamique du réseau, car bien souvent les amis de nos amis nous ressemblent (c'est le principe d'homophilie) et peuvent ainsi être à l'origine de nouvelles relations.

Apparue en 1997, la première plateforme identifiée comme réseau social numérique est dénommée « Six Degrees », qui fait référence aux travaux de S. Milgram sur les six degrés de séparation. Une liste chronologique de l'apparition des réseaux sociaux numériques à partir de 1997 est présentée en figure 1.2.1.

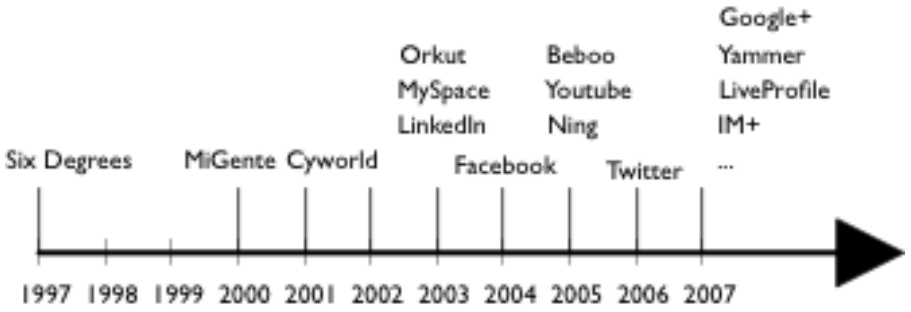


FIGURE 1.2.1 – Liste des réseaux sociaux numériques entre 1997 et 2007.

Les réseaux sociaux numériques possèdent naturellement de nombreuses différences. Afin de les distinguer plus finement, une classification en trois groupes a été proposée en 2009 par [9]. Cette classification propose de définir :

— les réseaux de *socialisation*, qui permettent à un individu de créer des relations numériques avec des personnes qu'il connaît déjà dans la vie réelle (p. ex. Facebook, Google+) ;

— les réseaux de *réseautage*, qui permettent à un individu d'acquérir de nouveaux contacts selon un ou plusieurs critères précis. Par exemple, LinkedIn et Viadeo permettent la création d'un réseau professionnel ;

— les réseaux de *navigation*, qui proposent à l'utilisateur d'accéder à du contenu pertinent pour lui. Ainsi, ce type de plateforme se focalise sur les messages circulant sur la plateforme (p. ex. Instagram, Twitter).

Nous présentons dans la suite les deux principaux réseaux sociaux numériques actuels que sont Facebook et Twitter. Ces deux plateformes serviront de cas d'étude durant cette thèse car elles illustrent deux grandes catégories de réseaux sociaux numériques que sont les réseaux de socialisation et les réseaux de navigation. De plus, elles regroupent un très grand nombre d'utilisateurs.

1.2.1 Le réseau de socialisation Facebook

Le leader des réseaux sociaux en ligne a été créé en 2004 par M. Zuckerberg. Facebook permet à un utilisateur de créer un compte à partir de quelques renseignements sur l'identité de son utilisateur. Cet utilisateur doit indiquer son nom, son prénom, une adresse de courriel, un mot de passe et une date de naissance (voir tableau 1.1). Il est désormais obligatoire, sur une grande partie des réseaux sociaux, de renseigner correctement les informations de son profil afin respecter les termes d'usage. Par exemple, Facebook requiert l'utilisation de son vrai nom. Un pseudonyme ou sobriquet ne pourra être utilisé qu'en complément. Sur Twitter, de telles restrictions ne sont pas mentionnées mais il est cependant interdit d'usurper l'identité de quelqu'un ou d'utiliser le nom d'une personne existante de manière offensante.

Après avoir rempli les informations principales, l'utilisateur se voit proposer une liste de contacts et doit accepter ou non la création d'un lien d'amitié avec ces derniers. Enfin, avant de pouvoir accéder à sa page personnelle, Facebook propose d'enrichir le profil de l'utilisateur en le complétant par son parcours scolaire, universitaire et professionnel (p. ex. lycée, université, employeur). Suite à l'ajout de ces dernières informations, un nouvel ensemble de contacts est proposé à l'utilisateur. Celui-ci peut de nouveau les ignorer ou les accepter. Le processus d'inscription se termine par l'ajout ou non d'une photo de profil.

Critères	**Facebook**	**LinkedIn**	**Twitter**	**Google+**
Nom, prénom	✓	✓	✓	✓
Courriel	✓	✓	✓	
Pseudonyme			✓	✓
Sexe	✓			✓
Date de naissance	✓			✓
Résidence		✓		
Code postal		✓		
Situation		✓		
Métier		✓		

Tableau 1.1 – Comparatif des informations nécessaires à la création de profils sur les quatre réseaux sociaux numériques Facebook, LinkedIn, Twitter et Google+.

Une fois inscrit, un utilisateur de Facebook peut encore optimiser son profil en l'enrichissant d'informations plus précises concernant son cursus mais aussi ses centres d'intérêt. Notons qu'il est même possible de renseigner ses préférences politiques, sa religion, son adresse et son numéro de téléphone. L'utilisateur a ensuite la liberté de rédiger des messages concernant ses activités du jour, mais aussi de publier des photos et vidéos pour enrichir sa page personnelle (également appelée « mur »). Il peut identifier un individu qu'il connaît sur une photo, il peut proposer un groupe sur une activité quelconque et ainsi rassembler des individus pour échanger de l'information sur un domaine particulier.

Finalement, il est possible d'interagir sur la plateforme de différentes manières. La première vocation de Facebook est de mettre un individu en relation avec des personnes qu'il connaît dans la vie réelle. Le slogan actuel de Facebook reflète ce positionnement : *« Facebook vous permet de rester en contact avec les personnes qui comptent dans votre vie »*. Un individu est généralement repérable par son nom et son prénom afin d'être reconnu des personnes qui le connaissent. Parmi les caractéristiques de Facebook, il est

important de noter qu'une relation ne peut être établie que sur accord mutuel des deux parties. Une fois la relation créée, deux amis peuvent communiquer en temps réel mais aussi commenter leurs activités et accéder à une partie ou à la totalité des informations l'un de l'autre.

1.2.2 Le réseau de navigation Twitter

Avec près de 500 millions d'utilisateurs, Twitter s'est imposé comme le leader des réseaux de navigation et plus particulièrement des plateformes de *microblog*. Le *microblog* est défini comme « un dérivé concis du blog, qui permet de publier un article plus court que dans les blogs classiques ».

À la création d'un compte Twitter, un utilisateur doit renseigner son nom complet (nom et prénom) mais aussi son pseudonyme, son adresse de courriel et son mot de passe (voir tableau 1.1). Son pseudonyme est l'identifiant unique d'un utilisateur et doit donc être inédit sur le réseau. Si ce n'est pas le cas, Twitter propose un ensemble de pseudonymes disponibles et similaires à celui recherché par l'utilisateur. Le processus d'inscription se poursuit comme pour Facebook à la recherche de contacts. Twitter étant centré sur l'information, les contacts proposés ne sont pas liés à notre identité réelle mais à nos centres d'intérêt. Dans ce but, Twitter propose lors de l'inscription de choisir des thèmes d'intérêt. Une liste des profils les plus reconnus sur les thèmes sélectionnés est proposée, libre à l'utilisateur de choisir alors ses contacts nommés *followees* (ou amis). Dans le langage commun, on dénomme l'utilisateur comme le *suiveur* de ses *followees.* L'ensemble des individus qui suivent un même et unique profil sont dénommés les *followers* de celui-ci.

Sur Twitter, les utilisateurs échangent de l'information par le biais de *tweets* qui sont de courts messages textuels d'une longueur maximale de 140 caractères. Ces messages semi-structurés peuvent contenir des *hashtags,* des *références* et des liens vers des sites web. Les *hashtags* sont des mots-clés précédés du caractère '#' qui permettent de caractériser un *tweet.* Par exemple, le hashtag *#sport* permet d'identifier le fait que le message traite de sport. Les *hashtags* permettent, entre autres, aux utilisateurs de rechercher les tweets en les filtrant. Les *références* permettent de citer un utilisateur dans un *tweet* sous la forme '@' suivi de son pseudonyme. Un utilisateur référencé recevra automatiquement celui-ci qu'il soit *follower* ou non du profil qui l'a émis.

La *timeline* d'un utilisateur est l'ensemble des tweets produits en temps réel par ses amis (ou *followees*). Cette *timeline* permet d'accéder à l'information produite par les contacts sélectionnés. Celle-ci est au cœur du service proposé par Twitter, car elle met à disposition les informations relatives à nos centres d'intérêt et *a priori* produites par des entités de confiance.

Il est important de noter que 90 % des profils de Twitter sont publics et donc visibles de tous. Au contraire de Facebook, n'importe qui peut accéder aux messages d'un individu public sans le consentement de celui-ci (la politique de confidentialité doit être acceptée). Au final, près de 99 % des données produites sur la plateforme Twitter sont publiques, ce qui en fait une des plateformes les plus accessibles mais aussi les plus vulnérables.

1.3 Les réseaux sociaux numériques mobiles

Avec l'apparition des smartphones, et compte tenu de la forte quantité d'utilisateurs de réseaux sociaux numériques, les plateformes sociales ont peu à peu été adaptées aux

smartphones et d'autres se sont uniquement portées sur les terminaux nomades afin d'utiliser le contexte de l'utilisateur (p. ex. l'heure, le lieu). Dans la suite, nous présentons ce type de réseau qui est défini dans cette thèse sous le terme de réseau social numérique mobile.

Les réseaux sociaux numériques ont connu une évolution notable avec l'arrivée des appareils nomades [10]. Durant ces dernières années, les téléphones mobiles ont proposé de nombreux services liés à la communication. La première génération, notée communément 1G et entièrement analogique, ne proposait que la possibilité de communiquer par la voix. La seconde génération (2G) devenue numérique a permis en plus d'échanger de l'information par le biais des SMS (messages textuels) et des MMS (p. ex. des images, des vidéos, des sons). Notons que cette seconde génération a été améliorée avec les versions 2.5G et même 2.75G afin d'augmenter le débit. Celle-ci a notamment permis l'apparition des premiers clients mails mobiles.

La troisième génération de mobile (3G) a permis l'augmentation significative des débits et le transfert de la voix et des données numériques en simultané. Cette évolution technologique a, par exemple, rendu possible l'intégration de la gestion des courriels et des réseaux sociaux numériques, qui ont permis d'élargir les moyens de communication à disposition d'un utilisateur nomade.

Depuis quelques années, grâce à la réussite d'Apple avec l'App Store (mais historiquement proposé par Nokia avec l'Ovi Store), le principe des magasins d'applications semble avoir fait l'unanimité. Ce principe de mise à disposition d'un lieu de partage pour la création et le téléchargement d'applications a été repris par Google avec Google Play, BlackBerry avec le BlackBerry App World. Cette évolution permet désormais à chacun de personnaliser son smartphone et de profiter de services adaptés aux usages.

Tandis que, depuis leur apparition en 1997, les réseaux sociaux numériques n'étaient accessibles que depuis un terminal fixe, à partir de 2006, un certain nombre ont été adaptés pour devenir accessibles depuis un terminal mobile, le plus souvent sous forme d'applications sociales ou via une interface Web adaptée pour les mobiles.

L. Monné annonce en 2009 que la majorité des réseaux sociaux numériques disponibles sur un smartphone ne sont en réalité que des *copies* des versions Web et que beaucoup n'ont pas apporté de réelles améliorations à leur système [11]. Il définit alors les réseaux sociaux numériques mobiles (RSM) comme des extensions de réseaux sociaux numériques capables d'intégrer le contexte spatio-temporel pour des services encore plus personnalisés et contextuels. Le contexte apporté par une utilisation mobile et nomade est un critère majeur d'évolution dans l'histoire des RSN. Ce phénomène de transfert et d'adaptations des tendances du Web 2.0 aux terminaux mobiles a été identifié sous le nom de Mobile 2.0 par [12].

Le tableau 1.2 présente un ensemble de plateformes sociales mobiles et pointe les différences entre les multiples services proposés. Les critères pris en compte pour la comparaison sont : le type de plateforme, noté *Type*, les caractéristiques propres à la plateforme, notées *Particularités*, la plateforme est-elle encore en activité, noté *Active*, la quantité d'utilisateurs, notée *Utilisateurs*. Concernant le type de plateforme, nous proposons d'utiliser le modèle de la société GoMo news [1] qui présente quatre principaux types de réseaux sociaux mobiles :

1. `http://www.gomonews.com/moso/`

— Les Réseaux de Partage de Contenu, notés RPC
Ce type de réseau propose l'envoi groupé de messages à une audience plus ou moins large. Ces messages sont généralement de courte taille (p. ex. microblogging) et ont pour objectif de diffuser de l'information liée au contexte spatio-temporel de l'utilisateur. Il peut s'agir aussi de réseaux sociaux numériques mobiles permettant le partage de fichiers sur des terminaux mobiles. Bien que la totalité des RSM permette d'échanger du contenu, ce type de réseau repose entièrement sur la notion de partage. Tout message sera publié de manière publique et les messages privés entre deux utilisateurs n'ont que peu de sens dans ce modèle.

— Les Réseaux Radar, notés RR
Ce type de réseau repose sur l'utilisation de la localisation géographique. Ils sont souvent référencés comme les services basés sur la localisation géographique, en anglais LBS, pour *Location Based Services*. L'objectif de tels réseaux est de permettre à un utilisateur donné de saisir des opportunités de rencontres avec d'autres membres du réseau (amis ou non de la vie réelle). Quelquefois, l'utilisateur peut aussi sélectionner des centres d'intérêt pour rencontrer des individus à proximité géographique et dont les activités peuvent être partagées.

— Les Réseaux de GéoTaggage, notés RGT
Ce type de réseaux est fondé sur le système de marquage géographique d'images ou de messages. Il s'agit de services basés sur la localisation géographique mais centrés sur le contenu et non sur les individus. Cela permet contextuellement d'obtenir des informations sur les lieux et sites visités. Par exemple, Dodgeball est un réseau social numérique qui permet d'associer un message textuel à un lieu géographique. Celui-ci sera reçu par les utilisateurs de la plateforme qui passeront dans la zone géographique concernée dans une période de temps donnée. Notons que ce réseau social a été racheté en 2009 par Google et a été remplacé ensuite par Google Latitude.

— Les Réseaux Sociaux Traditionnels, notés RST
Ces réseaux sociaux mobiles sont des réseaux sociaux traditionnels adaptés pour un usage mobile. Ils permettent aux utilisateurs déjà actifs de ces réseaux de rester connectés durant la journée pour publier ou consulter du contenu à n'importe quel moment et depuis n'importe où.

Nom	Type	Description	Actif	Utilisateurs
Dodgeball [10]	RR	Communication par SMS pour les personnes de proximité géographique		—
InfoRadar [13]	RR	Envoi de messages groupés et publics		—
CenceMe [14]	RR	Détection du comportement de l'utilisateur	✓	—
CitySence [15, 16]	RR	Détecte les points d'intérêt à San Francisco	✓	—
Twitter [17]	RPC	Microblog	✓	500 millions
Socialight [18]	RR	Localisation de contacts		—
WhozThat [19]	RPC	Partage d'informations entre deux individus qui se rencontrent		—
Google Latitude	RR	Localisation et partage de données avec des contacts		—
Path	RGT	Envoi de photos et localisation géographique	✓	3 millions
Yelp	RGT	Recherche de points d'intérêts	✓	17 millions
Instagram	RGT	Partage de photos	✓	80 millions
Foursquare	RGT	Localisation géographique à aspect ludique	✓	15 millions
Facebook	RST	Interaction avec des contacts	✓	1 milliard
Pinterest	RGT	Échange des liens vers des photos	✓	11 millions

Tableau 1.2 – Principaux réseaux sociaux numériques mobiles (chiffes de 2013).

Les réseaux sociaux numériques sont désormais, pour la plupart, accessibles depuis un terminal mobile. La frontière entre les réseaux sociaux numériques et les réseaux sociaux numériques mobiles est difficilement identifiable. La quasi-totalité des réseaux sociaux numériques peuvent désormais être identifiés comme réseaux mobiles traditionnels.

Nous avons représenté en figure 1.3.1, la quantité et le ratio d'utilisateurs nomades des principaux réseaux sociaux numériques actifs en 2013. Le ratio d'utilisateurs nomades est modélisé par le niveau d'imbrication entre chaque sphère illustrant un réseau social et la sphère représentant l'ensemble des utilisateurs de smartphones. Le diamètre de la sphère représentant chaque réseau social numérique est proportionnel à la quantité des utilisateurs de la plateforme. Le nombre d'utilisateur de chaque réseau social indiqué sur cette figure correspond aux valeurs fournies directement par les plateformes sur leur site officiel. Le nombre de smartphone est issu d'un article du spécialiste des marchés financiers Bloomberg LP [2].

2. `http://www.bloomberg.com/news/2012-10-17/smartphones-in-use-surpass-1-billion-will-double-by-20`
`html`

La figure 1.3.1 illustre l'aspect totalement nomade des réseaux Yelp, Foursquare, Instagram et Path. Ces derniers sont des réseaux sociaux numériques mobiles qui n'ont pas de sens lors d'une utilisation statique. La totalité de leurs participants y accèdent via un smartphone et l'accès Web ne permet que la consultation de données du réseau ou la création d'un profil utilisateur. D'autres plateformes ont un usage mixte avec un taux d'intégration d'utilisateurs nomades variable qui souvent résulte de leur notoriété et de leur capacité à intégrer des services supplémentaires.

Il est à noter que la plateforme Twitter comporte plus de la moitié d'utilisateurs nomades. Cette observation est liée à la nature de la plateforme, qui oblige les utilisateurs à être concis dans leurs messages (moins de 140 caractères) et donc particulièrement adaptée aux terminaux mobiles. Ce type de plateformes de *microblog* intègre de nombreuses fonctionnalités, dont la possibilité de référencer d'autres utilisateurs, l'insertion et la recherche par mots-clés, et la localisation géographique des messages.

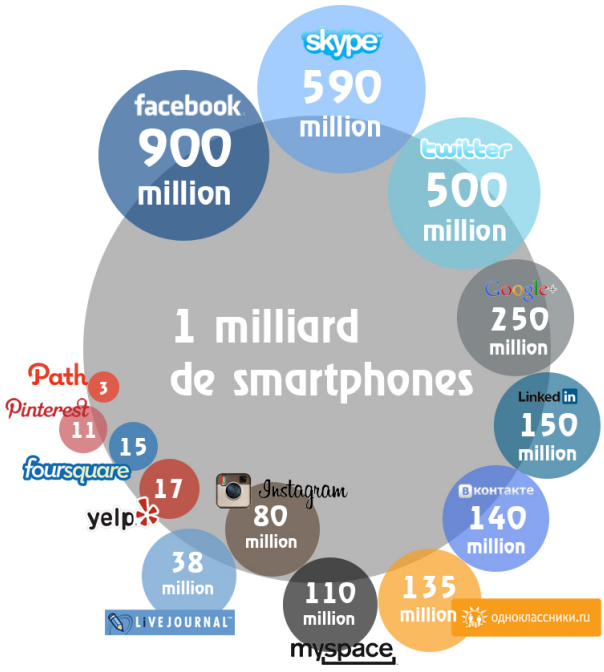


FIGURE 1.3.1 – Utilisateurs nomades des principaux réseaux sociaux numériques actifs en 2013.

1.4 Les failles de sécurité sur les réseaux sociaux numériques mobiles

1.4.1 Les failles traditionnelles des systèmes d'information

De nombreuses attaques ont été recensées depuis la création et avec le succès des réseaux sociaux numériques (figure 1.4.1). L'une des premières illustrations de la menace pesant

sur les RSN est le ver *Sammy* apparu en 2005. Il s'agit du premier qui était conçu pour se répandre sur le réseau social MySpace. Le ver avait pour objectif de modifier une partie du profil des utilisateurs sans *a priori* altérer, ni dérober, de données personnelles. Cette attaque a notamment été caractérisée par sa rapidité de propagation sur une large partie du réseau social.

En 2007, une faille de sécurité provenant du lecteur QuickTime a été exploitée sur le réseau social MySpace. Le téléchargement du virus s'effectuait lorsque l'utilisateur tentait de lire une vidéo depuis son profil.

Le célèbre ver *Koobface* est apparu en 2008-2009. Là encore, le téléchargement d'un code source s'effectuait lors de la tentative de lecture d'une vidéo. L'URL de la vidéo correspondait en fait à une copie du site YouTube. Une des originalités de ce ver est sa capacité de propagation sur plus de cinq réseaux sociaux numériques.

En 2009, un nouveau ver, *Mikeyy,* a fait son apparition sur le site de *microblogging* Twitter. Le ver avait pour objectif de modifier la page des utilisateurs avec des messages supplémentaires. Celui-ci publiait à l'insu de l'utilisateur des *tweets* contenant le mot *Mikeyy.*

Depuis 2009, une très forte quantité d'attaques ont été recensées et il est impossible d'établir une liste exhaustive de celles-ci. Cependant une récente étude Sophos a établi que 57 % des utilisateurs de RSN ont déjà été exposés au *spam* et 36 % ont déjà été exposés à des messages malveillants (conduisant au téléchargement de *malwares*).

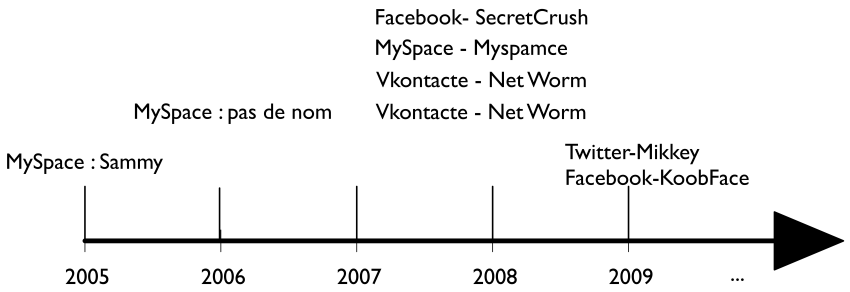


FIGURE 1.4.1 – Historique des attaques sur les réseaux sociaux numériques.

1.4.2 Les problèmes de sécurité propres aux réseaux sociaux numériques

Une grande quantité de données personnelles et professionnelles sont générées, stockées et manipulées tous les jours sur les réseaux sociaux numériques. Cela génère une quantité significative de risques, tels que les fuites de données, la perte de données, etc. [20, 21]. Dans cette section, et plus généralement dans cette thèse, nous nous concentrons sur le problème de la fuite de données, qui est considéré comme l'un des principaux risques [22]. La fuite de données est ici définie par l'accès non légitime à des données d'un utilisateur. Nous considérons deux types d'accès non légitimes : (1) l'utilisation de profils malveillants effectuant des attaques par phishing et (2) l'abus d'accès laissés à tord par l'utilisateur aux données par des profils non légitimés.

Ce risque peut être accru par l'utilisation des smartphones, car ceux-ci fournissent des fonctionnalités qui permettent une interaction facile et directe avec les réseaux sociaux. En effet, les utilisateurs de réseaux sociaux divulguent (volontairement ou non) une quantité

importante de données à caractère personnel. Ils ignorent souvent les conséquences qui peuvent découler de cette exposition de leurs données. De nombreux travaux ont abordé ces problèmes dans le but de protéger les données personnelles des utilisateurs et de rendre les réseaux sociaux plus sûrs [23, 24, 25, 26].

Le *courrier indésirable* (plus particulièrement le *spam* malveillant) est un problème de sécurité majeur pour les réseaux sociaux mobiles et en ligne. On définit comme *spams* les messages non sollicités reçus par un utilisateur sur un canal de communication donné.

Historiquement, le courriel a été le canal de communication privilégié par les *spammeurs*. Cependant, les réseaux sociaux le remplacent progressivement [22]. Ici, nous nous intéressons à un type particulier de messages non sollicités, les messages malveillants, c'est-à-dire ceux contenant une URL vers un site Web malveillant [27]. Les destinataires qui accèdent à l'URL sont dirigés vers une page Web malveillante qui exploitera des failles de sécurité traditionnelles pour dérober les données de l'utilisateur. Par exemple, le ver Koobface infectait les utilisateurs à partir d'un simple *tweet* ou d'un statut Facebook qui contenait un lien vers une vidéo YouTube. Avant la lecture de ce film, la fausse page Web YouTube demandait une mise à jour qui menait au téléchargement du ver et donc à l'infection du poste de travail de l'utilisateur. Le code téléchargé était en mesure de publier des messages à la place de l'utilisateur [28]. La confiance implicite entre deux amis d'un réseau social a joué un rôle important dans l'accélération et le succès de la propagation du ver.

Notons que les plateformes sociales contiennent des millions d'utilisateurs organisés sous forme de petit-monde (c.-à-d. n'importe quelle paire d'utilisateurs n'est séparée que par quelques connexions, et le réseau est fortement groupé), ce qui rend la diffusion des contenus potentiellement très efficace [29].

Le partage des données sensibles (personnelles et professionnelles) provient principalement du fait qu'un utilisateur n'est pas conscient de la qualité et de la quantité de son audience [22]. L'utilisateur qui publie des données stratégiques peut donc subir une variété de dommages. Lorsque des données sensibles sont distribuées, celles-ci peuvent être signalées et dupliquées si rapidement qu'il est impossible de les supprimer.

Une autre source de fuite de données est la présence de contacts malveillants dans la liste d'amis d'un utilisateur. Il a été observé que les demandes d'amis, même provenant d'une personne inconnue, ont un taux de réussite important sur Facebook [30, 31]. Notons que la politique de sécurité de Facebook permet à deux amis d'avoir accès aux renseignements personnels de chacun. Il est ainsi facile pour un utilisateur malveillant de créer un faux profil dans le but d'effectuer une intrusion dans la liste de contacts d'un utilisateur pour bénéficier des droits d'accès associés et accéder au contenu publié par celui-ci. Bien que les réseaux sociaux offrent des mécanismes pour empêcher la création automatique de tels profils, certains projets ont réussi à les surmonter [32]. Il a été aussi démontré en 2011 que la plupart des réseaux sociaux en ligne peuvent être infiltrés avec un taux de réussite de 80 % [30]. Il est donc tout à fait possible d'utiliser cette technique pour attaquer le système de protection de Facebook (protection contre la création de flottes de profils, etc.) [33] et ainsi accéder aux données personnelles des utilisateurs. L'existence d'intrusions a été confirmé par les travaux de [31] et [34], qui ont montré que tous les amis ne sont pas nécessairement dignes de confiance.

Le vol d'identité est l'une des conséquences possibles de la fuite de données. Il n'existe pas de solutions permettant de s'assurer qu'une identité numérique correspond à une personne physique et/ou une personne de confiance. Plusieurs travaux ont abordé le problème mais, jusqu'à présent, aucune solution n'a été intégrée dans les réseaux sociaux en ligne.

Certains réseaux tels que Twitter et Facebook permettent cependant de signaler un vol d'identité. Il n'existe toujours aucun moyen pour empêcher une personne de créer un profil avec l'identité d'une autre personne. Cela nécessiterait l'authentification de l'utilisateur, mais ces technologies ne sont pas encore disponibles sur les réseaux sociaux en ligne.

Une autre conséquence de la fuite de données est le traçage des utilisateurs. Le traçage consiste à récupérer, sur les médias sociaux, les informations relatives au contexte spatio-temporel d'un utilisateur. Suivre un utilisateur en temps réel peut conduire au manque de respect de la vie privée. À titre d'exemple, le logiciel Raytheon [3] a prouvé qu'il est possible de traquer les utilisateurs à partir des données de localisation géographique présentes sur Twitter, Facebook et le réseau social Foursquare. L'outil démontre donc que la localisation d'un utilisateur est souvent divulguée. Des travaux ont pour but de prévenir ces problèmes en recommandant aux utilisateurs de prêter attention aux configurations de localisation de leurs appareils [35, 25]. Les réseaux sociaux mobiles qui sont fondés sur la localisation géographique des utilisateurs peuvent stocker et donner accès à la position d'un utilisateur par le biais d'une interface de programmation d'applications. Si ces interfaces sont accessibles, on peut être en mesure de déduire de celles-ci le lieu de travail et même l'adresse personnelle de l'utilisateur [36, 37, 38].

Conclusion

Les nouvelles technologies de l'information et de la communication, en particulier les smartphones et les réseaux sociaux numériques, offrent une quantité importante de médias de communication. L'utilisation nomade qui intègre les réseaux sociaux numériques en plus des technologies habituelles rend de plus en plus simple la publication et l'acquisition de contenus. Toutes ces nouvelles portes de communication sont naturellement aussi des vecteurs d'attaque pour les acteurs malveillants. Ceux-ci y voient une opportunité grandissante dans un monde où une information peut se révéler stratégique et monnayable.

Les utilisateurs n'ont pas toujours conscience ni de la distinction entre vie professionnelle et vie personnelle, ni de l'impact de certaines actions en ligne. Les utilisateurs naviguent sur les plateformes numériques où ils sont à la fois de potentielles victimes d'attaques malveillantes et à la fois eux-mêmes vecteurs de divulgation d'informations sensibles à leur public.

Un des éléments de base pour joindre un réseau social utilisateur est la notion de contacts (souvent appelés « amis »). Ces contacts sont les personnes avec lesquelles l'utilisateur souhaite échanger et partager de l'information. Tandis que les contacts de l'utilisateur sur un réseau social donné sont supposés être dignes de confiance, les travaux de [30] montrent que, sur la plupart des réseaux sociaux en ligne, les utilisateurs ne font pas suffisamment attention en sélectionnant ou en acceptant des amis. Cette vulnérabilité a été utilisée pour attaquer le système immunitaire Facebook et accéder aux données à caractère personnel des utilisateurs.

Bien que les travaux de recherche aient révélé l'existence de profils malicieux sur les réseaux sociaux numériques et plus particulièrement dans la liste de contacts des utilisateurs, pour l'instant, aucune solution claire à ce problème n'a été proposée pour l'instant. Par conséquent, alors qu'un utilisateur publie des informations personnelles sur des réseaux sociaux numériques, certains utilisateurs illégitimes peuvent obtenir l'accès à des

3. `http://www.raytheon.com/`

données potentiellement sensibles. De même, ces profils vont pouvoir publier du contenu malveillant qui sera visible des utilisateurs leur ayant octroyé leur confiance.

Comme indiqué dans les politiques de sécurité des plateformes sociales, l'utilisateur est responsable des personnes appartenant à son groupe d'amis, et ces personnes sont supposées être dignes de confiance. À titre d'exemple, Facebook indique clairement que l'on doit seulement envoyer des requêtes d'amis *« aux personnes avec lesquelles vous avez une connexion réelle, comme vos amis, famille, collègues ou camarades de classe »*. En outre, ils indiquent que, *« si vous êtes intéressé à recevoir des mises à jour des personnes que vous trouvez intéressantes, mais ne connaissez pas personnellement (p. ex. des journalistes, des célébrités, des personnalités politiques), essayez de les suivre au lieu de leur envoyer des demandes d'amis »*. Cette confiance est souvent un facteur favorable à la diffusion des vers et des attaques malveillantes car celles-ci se propagent de proche en proche grâce à la relation d'amitié et à l'implicite confiance qui existe entre l'utilisateur et ses contacts.

Dans ce contexte nouveau apporté par la convergence des technologies mobiles et sociales, il devient nécessaire de proposer des solutions personnalisées de sécurisation et entièrement intégrées dans le contexte nomade. De telles solutions doivent à la fois tirer parti des ressources offertes par les smartphones *pour proposer un outil d'accompagnement* de l'utilisateur et de sécurisation de celui-ci.

Les enjeux scientifiques émanant d'une telle approche sont multiples. Il est nécessaire de mettre en place des modèles de représentation des médias de communication et du comportement de l'utilisateur sur l'ensemble de ces médias. Il est aussi nécessaire de mettre en place des outils de détection de comportements malveillants à l'échelle égocentrique de l'utilisateur et aussi à grande échelle, et ce pour être en mesure d'intégrer des outils de prédiction de comportement malveillant dans un but d'anticipation.

Chapitre 2

État de l'art sur la modélisation des réseaux sociaux numériques mobiles et la sécurisation des interactions entre leurs utilisateurs

Résumé du chapitre

Ce chapitre propose un état de l'art des modèles existants pour la représentation d'un réseau social mais aussi de données sociales issues d'un smartphone. À cette occasion, les principaux indicateurs de l'analyse des réseaux sociaux seront présentés. Le modèle multicouche, qui constitue un modèle central de cette thèse, sera explicité tout comme les techniques associées permettant son élaboration. Nous présenterons ensuite les deux principales approches pour la sécurisation des interactions entre les utilisateurs : d'une part, les techniques de détection de comportements malveillants sur les plateformes sociales et, d'autre part, les approches de mesure de confiance entre deux utilisateurs. Le chapitre se termine par la mise en évidence de certains verrous auxquels nous répondons dans cette thèse.

2.1 La modélisation des réseaux sociaux numériques mobiles

2.1.1 Modéliser un réseau social

Il est possible de modéliser un réseau social et donc un réseau social numérique par un graphe noté $G = (N, A)$, dont l'ensemble des nœuds N représente les entités sociales et l'ensemble des arêtes A les relations ou les interactions entre ces individus. Il est à noter que, par construction, les arêtes sont des couples d'éléments de N. Proposé par J.-L. Moreno dans les années 1930, ce concept de graphe représentant les interactions sociales portait à l'origine le nom de « sociogramme » [7]. Ce mode de représentation d'un réseau social par un graphe est maintenant largement utilisé dans la littérature scientifique et permet de conduire des analyses à différentes échelles.

Nous proposons de représenter les différents niveaux d'analyse d'un graphe par la matrice M de la figure 2.1.1.

Cette matrice révèle trois niveaux observation pour l'analyse : le nœud (c.-à-d. l'utilisateur), le nœud et son voisinage (c.-à-d. le profil et ses contacts) et le réseau social global. Le profil utilisateur représenté par un nœud est l'unité la plus petite d'observation d'un réseau social numérique. Une échelle intermédiaire consiste à analyser un profil avec son ensemble de contacts. Enfin, il est possible de prendre en compte l'ensemble des utilisateurs dans une analyse à grande échelle.

Pour chaque échelle, il est possible de caractériser les entités observées sur trois plans particuliers : les entités individuelles, les entités liées entre elles et les flux générés à travers les liens.

Nous proposons dans la suite de ce chapitre un tour d'horizon des techniques de l'analyse des réseaux sociaux en fonction des données obtenues sur le graphe et de l'échelle d'analyse [39, 40]. Nous noterons par $M(x, y)$, une technique d'analyse appliquée à l'échelle x selon le plan d'analyse y.

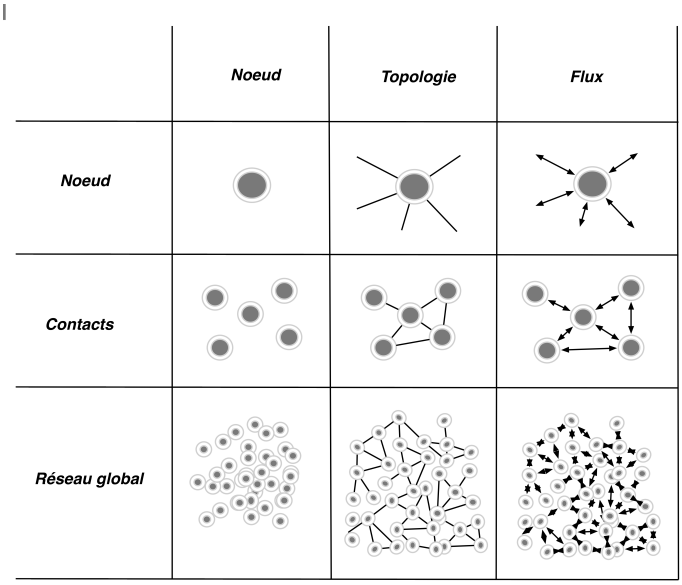


FIGURE 2.1.1 – Matrice M des échelles d'analyse d'un réseau social numérique.

Dans le cas des réseaux sociaux numériques, le nœud est associé à un individu, ou

profil utilisateur, c'est-à-dire : $M[N\text{œ}ud, N\text{œ}ud]$. Un individu est identifié alors par une adresse de courriel, un pseudonyme, un couple (nom, prénom) ou un identifiant numérique. Un nœud peut être caractérisé par un ensemble d'attributs issus de multiples sources. Des indicateurs issus des attributs d'un profil peuvent être analysés pour la recherche de tendances, d'anomalies ou de caractérisations plus précises d'un individu dans son contexte.

Sur Twitter, les attributs d'un profil sont multiples, et la figure 2.1.2 en illustre une partie. Le profil représenté est celui de la plateforme elle-même (c.-à-d. @twitter), mais tout profil public peut être extrait sous le même format. Parmi les attributs les plus pertinents, notons le nom de l'utilisateur, son pseudonyme, sa position géographique, sa description, l'adresse de son site Internet, le nombre de ses connexions entrantes et sortantes, le nombre de messages publiés. Notons que le profil présenté a été certifié par la plateforme Twitter pour garantir sa propre identité. Cela se traduit par l'affectation de la valeur *true* à l'attribut *verified* (voir ligne 12 de la figure 2.1.2).

```
1  <user>
2    <id>783214</id>
3    <name>Twitter</name>
4    <screen_name>twitter</screen_name>
5    <location>San Francisco, CA</location>
6    <description>Your official source for news, updates and tips</description>
7    <url>http://blog.twitter.com/</url>
8    <followers_count>17734324</followers_count>
9    <friends_count>120</friends_count>
10   <created_at>Tue Feb 20 14:35:54 +0000 2007</created_at>
11   <time_zone>Pacific Time (US & Canada)</time_zone>
12   <verified>true</verified>
13   <statuses_count>1571</statuses_count>
14   <lang>en</lang>
15 </user>
```

FIGURE 2.1.2 – Représentation XML du profil @twitter sur la plateforme Twitter.

Le nœud ainsi que ses voisins peuvent faire partie d'une analyse plus élargie d'un utilisateur (c.-à-d. $M[Contacts, N\text{œ}ud]$). À titre d'exemple, l'analyse en termes d'attributs d'un nœud et de ses voisins permet de mieux décrire le nœud analysé en partant du principe que son identité est aussi forgée par le choix et les caractéristiques de ses contacts. Ce concept d'identité sociale a été traité dans la littérature [41, 42].

L'ensemble des nœuds d'un graphe englobe l'ensemble des acteurs d'une plateforme sociale numérique avec leurs attributs (c.-à-d. $M[Global, N\text{œ}ud]$). L'analyse des données à cette échelle peut être réalisée à l'aide des techniques de fouille de données appliquées sur le profil de l'utilisateur (p. ex. sa description). Il est possible de pondérer l'importance d'un attribut d'un individu d'autant plus qu'il est observé par celui-ci, mais qu'il ne l'est pas chez les autres. Cette technique, semblable au calcul du score TF-IDF (Term Frequency-Inverse Document Frequency) pour l'analyse de corpus de documents, met en évidence l'intérêt d'une connaissance locale dans un cadre global [43].

La topologie d'un nœud (c.-à-d. $M[N\text{œ}ud, Topologie]$) concerne la quantité d'arcs entrants et/ou sortants de ce nœud. Certaines plateformes sociales numériques peuvent distinguer ces deux types de relations, ce qui permet la construction d'un indicateur de

degré entrant (noté $d_{entrant}(u)$) et d'un indicateur de degré sortant (noté $d_{sortant}(u)$). Sur Twitter, le degré entrant est appelé nombre de « suiveurs » et le degré sortant correspond au nombre d'amis (respectivement lignes 9 et 10 sur la figure 2.1.2). La somme des degrés entrant et sortant est identifiée comme le degré total. Dans le cas des plateformes numériques, ce degré est souvent interprété comme révélateur de l'importance d'un utilisateur. Un indicateur de réputation, noté $R(u)$, d'un individu $u \epsilon N$ a été proposé pour la plateforme Twitter. Celui-ci est calculé en fonction du degré entrant et sortant selon la définition 2.1 proposée par [44].

Définition 2.1. Réputation d'un profil Twitter

La réputation R d'un profil modélisé par un nœud u sur le réseau Twitter est définie par :

$$\forall u \in N, R(u) = \frac{d_{entrant}(u)}{d_{entrant}(u) + d_{sortant}(u)} \tag{2.1.1}$$

Avec :

$d_{entrant}(u)$ le degré entrant du nœud u
$d_{sortant}(u)$ le degré sortant du nœud u

Cet indicateur de réputation met en évidence l'attractivité du profil vis-à-vis des autres utilisateurs tout en prenant en compte sa propre propension à créer des liens avec les autres. Cet indicateur est calculé à partir du contenu des attributs *followers_count* et *friends_count*. La réputation du profil donné en exemple dans la figure 2.1.2 est quasiment maximale ($R(@twitter) \approx 1$).

Le réseau égocentrique d'un utilisateur (c.-à-d. $M[Contacts, Topologie]$) est constitué de l'ensemble de ses contacts et des relations entre ceux-ci. Il permet d'obtenir une vision locale sur les relations entre les contacts de l'utilisateur et de mettre en évidence certaines facettes de son identité numérique. [42] met en évidence la possibilité d'extraire du réseau égocentrique un ensemble de groupes de contacts possédant une forte connectivité et des centres d'intérêt communs. Cette analyse peut être améliorée, notamment pour mesurer l'importance de chacun des contacts en fonction de la quantité et du type des interactions observées. Celle-ci permet d'identifier un ensemble de facettes de la vie sociale d'un utilisateur [45].

La topologie globale d'un graphe représentant un réseau social numérique (c.-à-d. la cellule $M[Global, Topologie]$ de la matrice) peut permettre un très large nombre d'analyses. Nous illustrons la pertinence de traitements à cette échelle par le célèbre algorithme PageRank [46]. Cet algorithme évalue la réputation d'une page Web à partir de l'ensemble des liens existant entre les pages. L'algorithme est souvent référencé comme le modèle du surfeur aléatoire. Le score affecté à chaque page est égal à la quantité de temps passé par un individu (le surfeur) qui se déplacerait de page en page en suivant aléatoirement les liens. Pour éviter les boucles infinies, le surfeur se voit affecter à chaque itération une probabilité non nulle de se déplacer à n'importe quel endroit du Web. L'algorithme a été proposé par L. Page, le cofondateur de Google, il est aujourd'hui encore l'un des critères utilisés pour la remontée de pages Web par le célèbre moteur de recherche.

2.1.2 Modélisation de données sociales issues d'un mobile

Nous présentons un bref état de l'art concernant la modélisation et l'analyse des traces sociales d'un smartphone. Cet état de l'art se terminera par la présentation du modèle

multicouche (aussi noté ML) qui a été utilisé comme modèle de référence dans notre travail.

[47] a introduit le concept de *fouille de données mobile*, pour illustrer le potentiel d'une analyse des données sociales issues d'un smartphone. Le travail ainsi proposé, nommé *SocialMine,* permet de collecter sur les appareils nomades Android un ensemble de traces sociales. Parmi les données analysées, on peut recenser : les historiques d'appels et de SMS, les listes de contacts et les données issues de l'application mobile Facebook. L'objectif de l'étude est de montrer la faisabilité de la collecte de traces sociales. Pour cela, l'auteur a proposé un ensemble de statistiques illustrant l'utilisation des applications sociales des smartphones. Parmi les résultats les plus intéressants, l'application permet de générer le graphe pondéré des contacts de l'utilisateur du smartphone.

Récemment, les auteurs de [48] ont présenté un outil appelé *LogAnalysis* pour effectuer une analyse des appels téléphoniques. *LogAnalysis* permet de visualiser, de filtrer et de surveiller les journaux d'appels.

Dans le travail [45], les auteurs ont proposé une extension du simple graphe des contacts extraits depuis un smartphone. L'approche combine le graphe de contact extrait depuis un smartphone avec des données issues du Web. L'analyse des données du Web a pour objectif de déduire les relations entre les contacts de l'utilisateur de smartphone. Pour accomplir cette tâche, les auteurs proposent d'utiliser un moteur de recherche et de rechercher les occurrences communes des contacts. Cette analyse extrait les contacts et aussi les relations entre eux. Une fois le graphe égocentrique extrait de l'analyse, une technique de *clustering* est utilisée pour extraire du graphe les composantes importantes. Cela permet notamment de mettre en évidence certains aspects de la vie sociale (centres d'intérêt) de l'utilisateur de smartphone.

Les auteurs de [49] ont proposé une analyse qui tient compte de la mobilité, des relations sociales et des motifs de communication d'un utilisateur. L'objectif de cette démarche était d'analyser la corrélation entre des données issues de sources hétérogènes (c.-à-d. carnet d'adresses, SMS, Facebook et localisation géographique).

L'aspect nomade des smartphones a été analysé par les travaux de D. Quercia [50]. L'approche proposée étudie les relations spatio-temporelles entre les utilisateurs de terminaux mobiles. À cet effet, une application est intégrée dans le smartphone des utilisateurs à analyser. Un algorithme permet de détecter régulièrement les utilisateurs à proximité géographique de l'appareil et de les mémoriser dans une base de données. Ces relations sont ensuite traitées pour générer un graphe dynamique des utilisateurs. Le poids des connexions est alors proportionnel à la quantité et à la durée des rencontres physiques entre chaque paire d'utilisateurs.

Certains travaux ont utilisé le concept de graphes spatio-temporels et l'ont confronté avec le graphe de contacts [51]. Le résultat principal est la forte corrélation entre les deux graphes. Cela signifie qu'une partie importante des contacts présents sur un smartphone correspondent à des personnes rencontrées régulièrement dans la vie réelle.

2.1.3 Le modèle multicouche

2.1.3.1 Présentation du modèle

M. Magnani et L. Rossi [52] ont proposé un modèle multicouche (dénoté ML pour *Multi Layer*)), qui peut contribuer à unifier les multiples facettes d'un utilisateur de réseau social. Ce modèle a été appliqué et testé par les auteurs sur les réseaux FriendFeed et Twitter.

Le modèle multicouche peut être formalisé comme un ensemble L de K couches et un ensemble de matrices de correspondance entre ces couches. Chaque couche représente un réseau social numérique sur lequel l'utilisateur est présent. Une couche de réseau social L_i est représentée par un graphe $G_i = (N_i, A_i)$ où N_i représente l'ensemble des nœuds (c.-à-d. profils) et A_i l'ensemble des liens (c.-à-d. connexions). Les connexions entre les différentes couches sont modélisées par des matrices de correspondance. La figure 2.1.3 montre un exemple d'un réseau multicouche.

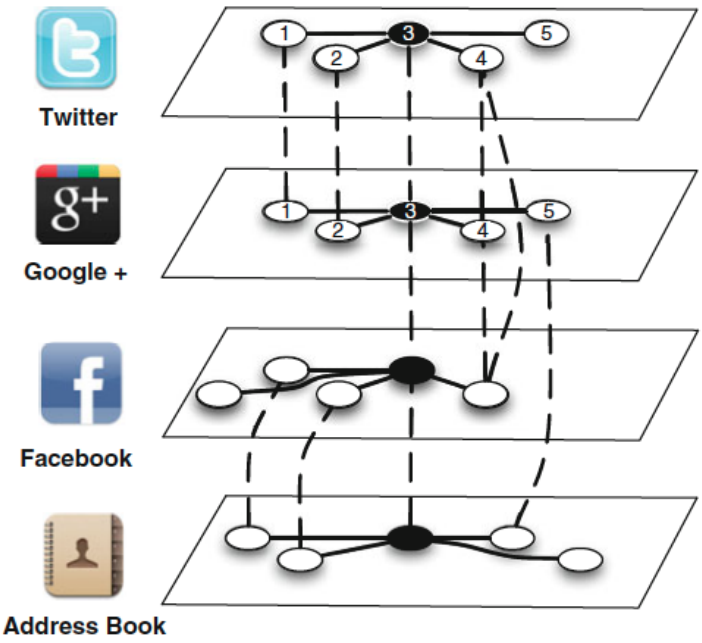


Figure 2.1.3 : Réseau social égocentrique multicouche d'un utilisateur de smartphone.

Il existe deux types de connexions dans les réseaux multicouche : (1) les connexions *intra-couche* (représentées par des traits continus), qui sont des connexions entre les utilisateurs d'un réseau social donné, et (2) les connexions *inter-couche* (représentées par des traits discontinus), qui représentent les relations entre les nœuds de réseaux sociaux distincts. Les connexions *intra-couche* sont identifiées par le réseau social considéré. Par exemple, si la couche considérée est Facebook, la nature de la relation est identifiée par l'appartenance à la liste d'amis. Les connexions *inter-couche* du modèle, représentées par des traits pointillés sur la figure, exigent la vérification d'une condition de correspondance. Un lien entre deux nœuds est créé si leurs profils associés sur deux couches différentes appartiennent à la même personne. Ces connexions étant nécessaires à la construction du modèle, nous proposons de faire un état des lieux des techniques pouvant aider à leur identification.

2.1.3.2 Établissement des liens inter-couche

Le domaine de la résolution d'entité à pour objectif de déterminer si deux profils de différents réseaux sociaux correspondent à la même entité [53, 54]. Ci-dessous, nous présentons un état de l'art de ce type d'approches par correspondance directe.

L'approche par correspondance directe est fondée sur les identifiants uniques d'un utilisateur sur plusieurs réseaux sociaux. Ces approches sont généralement liées au mode de

représentation d'un profil. Parmi les représentations de profils les plus communes (Resource Description Framework) est un modèle de métadonnées pour désigner tout type d'information comme une expression *sujet-prédicat-objet*. Ce modèle a été enrichi dans le domaine des relations sociales par deux ontologies, que sont FOAF (Friend of a Friend) [55] et SIOC (Semantically-Interlinked Online Communities) [56]. Ces ontologies contiennent des spécifications riches qui permettent d'identifier les relations entre les individus et leurs profils respectifs. Ces formats de description contiennent beaucoup d'identificateurs uniques qui peuvent être utilisés pour identifier directement une entité sur de multiples plateformes. Ces identifiants sont considérés comme *propriété fonctionnelle inverse* (IFP) et sont définis comme suit : *« Partout où vous voyez le sujet lié à un objet par ce prédicat particulier, alors que le sujet est le seul et unique sujet avec cet objet relié par le prédicat. Si jamais vous voyez un autre sujet lié à l'objet par le prédicat, vous saurez que l'autre sujet est réellement le même sujet ».*

Les travaux de [57] reposent sur l'utilisation de *foaf :mbox_sha1sum*, *foaf :homepage* et *foaf :name* pour faire correspondre deux profils. L'attribut *foaf :mbox_sha1sum* est le résultat de la fonction mathématique SHA1 appliquée à l'adresse e-mail. L'hypothèse sous-jacente est que les profils de la même personne sur plusieurs plateformes sociales contiennent la même adresse de courriel. La page d'accueil (*foaf :homepage*) est également utilisée comme identifiant pour effectuer une correspondance directe entre plusieurs entités. Un fort inconvénient de cette approche et qu'il n'existe aucune garantie que l'utilisateur ait indiqué sa page d'accueil sur ses multiples profils. Enfin, le nom de l'utilisateur (*foaf :name*) peut également être utilisé comme identifiant mais uniquement à condition que l'utilisateur ait indiqué son nom personnel sur tous ses profils numériques.

Le problème des correspondances directes en termes de nom d'utilisateur est la non-détection d'un lien si la moindre différence existe entre les deux noms comparés. Un ensemble de mesures a été proposé pour surmonter cette limitation. Les mesures proposées visent à capturer le degré de similitude non seulement entre deux chaînes de caractères en général [58], mais aussi, et plus précisément, entre deux noms de personnes [59]. Notons que même si sur certains réseaux les utilisateurs sont identifiés par un pseudonyme (p. ex. Twitter), le nom de l'utilisateur est toutefois requis à la création du profil et peut être exploité pour l'identification des liens inter-couche. Nous présentons ci-dessous quelques-unes de ces mesures de similarité.

La distance de Levenshtein entre deux chaînes de caractères est mesurée par le nombre de transformations nécessaires pour modifier la première chaîne afin d'en obtenir la seconde [60]. Les opérations de base autorisées sont l'insertion, la suppression et la substitution de caractères. La distance de Damerau-Levenshtein étend cette mesure en permettant en plus d'effectuer des transpositions [61].

La distance *q-gram* calcule le nombre de sous-séquences de caractères de longueur q (appelée *q-grams*) en commun entre les deux noms de personnes [62]. La normalisation est effectuée sur la base soit du nombre de *q-gram* dans la chaîne la plus courte, soit du nombre de *q-gram* dans la chaîne la plus longue, ou de la moyenne des deux.

La distance de Jaro calcule la similarité entre deux chaînes en tenant compte de l'ordre et du nombre de caractères communs [63]. La mesure de Wrinkler enrichit celle-ci en intégrant le fait que moins d'erreurs se produisent au début des noms de personnes lors de la saisie informatique de ceux-ci. Cet indicateur affecte ainsi plus d'importance aux premiers caractères des noms de personnes. La mesure de Winkler dite « triée » propose un tri par ordre alphabétique des mots qui constituent les chaînes (par exemple nom et prénom) avant de calculer la mesure de Winkler [64, 65].

Il est important de noter qu'un ensemble de mesures plus complexes existent pour analyser la similarité, par exemple les similarités qui sont fondés sur la phonétique des noms (p. ex. Soundex par [66]). Le lecteur peut se référer à [59] et [58] pour une liste exhaustive de ces mesures.

2.2 Détection de profils malveillants sur les réseaux sociaux numériques

Nous proposons de classer les travaux scientifiques de détection de profils malveillants sur les plateformes sociales selon sept dimensions. Rappelons que, dans cette thèse, nous identifions comme comportement malveillant tout profil qui peut nuire à l'utilisateur soit en accédant de manière illégitime à ses données soit en publiant lui-même du contenu illégitime. Un exemple de contenu non légitime est une URL pointant vers un site Web utilisant une faille de sécurité pour télécharger sur le matériel de la victime un virus ou un ver informatique. Les sept dimensions retenues pour la comparaison des approches de l'état de l'art sont :

— L'*objectif* principal du travail de recherche. Celui-ci peut être la mise en évidence, l'observation ou la détection de profils malveillants sur les réseaux sociaux numériques.

— L'*approche* utilisée, qui peut être basée sur une analyse manuelle, sur des techniques de classification, de l'analyse textuelle ou sur des techniques de groupement de données.

— L'*échelle* d'application de l'approche, qui se mesure en quantité de profils traités.

— La *dynamique,* qui reflète si l'approche de recherche peut être utilisée en temps réel ou si elle nécessite une phase de collecte suivie
— d'une phase d'analyse (les deux étant consécutives).

— Le *type de données* utilisées, qui peut varier de la simple analyse des messages à l'analyse des profils et même du comportement des utilisateurs.

— L'*analyse des sites Web* dont les URL sont présentes dans les messages envoyés. Cette analyse peut permettre d'avoir une idée de la criticité ou non des attaques.

— L'indicateur noté *« aide à la décision »* met en évidence l'apport ou non d'une solution graphique utilisable pour gérer le danger issu des profils malveillants et prendre ainsi les décisions adéquates

Les contributions les plus significatives sont présentées dans le tableau 2.1. Dans la suite de cette section, nous présenterons les différents travaux de recherche en fonction de l'approche qu'ils proposent.

	Objectif	Approche	Échelle	Dynamique	Données utilisées	Sites Web	Aide à la décision
[67]	Détection	Classification	500		Profils, comportement, graphe		
[27]	Détection	Géométrique	70 000		Profils, messages	✓	
[68]	Observation	Manuelle	800 000		Mots-clés	Partiellement	Ligne de temps
[69]	Mise en évidence	Classification	200 000		Profils, messages		
[70]	Détection	Groupement de données	3 500 000		Messages	Partiellement	Ligne de temps
[26, 44]	Détection	Classification	26 000		Graphe, messages		
[71]	Détection	Analyse sémantique latente	non connue		Messages		
[72]	Détection	Géométrique	25 000	✓	Sacs de mots, messages	Service-tiers	
[73]	Détection	Groupement de données	320		Like, URL, amis		
[74]	Détection	Classification	1 000 000	✓	Sac de mots, profils, messages		
[75]	Détection	Classification	54 000 000		Mots-clés, URL	✓	
[76]	Détection	Statistique	100 000		Horodatage		
[32]	Détection	Graduation	1 800 000		Graphe		

Tableau 2.1 – Comparaison des travaux de détection de profils malveillants sur les réseaux sociaux numériques.

2.2.1 Les approches fondées sur les techniques de classification

La détection de messages ou de profils malveillants par les techniques de classification a été abordée par de nombreux travaux scientifiques. Déjà utilisée pour la détection de *spams* dans les boîtes mail, cette approche a été adaptée aux besoins de la situation [77]. Les approches fondées sur la classification considèrent pour la majorité l'existence de deux classes : les profils malveillants et les profils inoffensifs. Il s'agit alors de constituer un ensemble d'apprentissage pour chacune des classes et d'appliquer les approches mathématiques concernées.

La qualité de la détection dépend bien souvent de la qualité de l'échantillon d'apprentissage, de la méthode utilisée et du type de variables prises en compte dans le modèle.

La constitution de l'échantillon de référence peut être réalisée de plusieurs manières. Tout d'abord, comme dans l'approche de [26, 44] réalisée au College of Information Sciences and Technology de l'université d'état de Pennsylvanie, la constitution d'un ensemble de profils de référence peut s'en remettre à l'analyse manuelle d'experts. Ainsi, un ensemble de taille variable (entre 100 et 500 profils) va être scruté sur la plateforme puis classifié avec l'un des deux labels selon la décision unique des experts.

Certains travaux [67, 69, 27] de l'université du Texas A&M ont proposé, dans le cadre du projet « Social Honeypot », la création de profils *pot de miel* visant à attirer les profils malveillants des plateformes sociales. Ceux-ci possèdent des caractéristiques particulières (p. ex. des photos de profil, une description) et leur gestion est optimisée dans le but d'obtenir l'attention des profils suspects de la plateforme. Des experts viennent ensuite valider la sélection opérée. Notons que, dans ce contexte, les performances de détection obtenues par les approches sont généralement de l'ordre de 90%. Il convient de préciser que les profils malveillants ainsi détectés ne sont pas forcément représentatifs de l'ensemble des profils malveillants de la plateforme, mais seulement de ceux ayant été attirés par des profils « pot de miel ». Aussi, par construction, ce type d'approches ne peut prétendre à une détection qu'à une petite échelle de profils malveillants.

Les auteurs de [75] de l'université fédérale de Minas Gerais proposent de constituer un ensemble de profils malveillants en partant des URL de sites répertoriés comme dangereux. La liste de sites dangereux est extraite depuis des bases de données constituées à cet effet. À titre d'exemple, sur la plateforme Phishtank, un utilisateur ayant un doute sur la qualité d'un site Web peut le signaler et le proposer pour analyse. Une équipe d'expert étudient le site en question et, si celui-ci est détecté malveillant, il est intégré dans la base de données.

L'approche va scruter sur le réseau les messages et par déduction les utilisateurs qui font référence à ces sites malveillants.

Afin de pouvoir classifier des messages ou des utilisateurs comme malveillants ou non, il est souvent nécessaire de pouvoir les représenter sous forme de vecteurs. Un vecteur contient un ensemble de caractéristiques décrivant l'entité représentée. En ce qui concerne les médias sociaux, ces caractéristiques sont les indicateurs mesurables de l'activité d'un profil.

Nous proposons une classification des approches existantes en fonction des composantes utilisées dans chaque modèle. Pour ce faire, nous considérons quatre principales catégories d'indicateurs : *profil*, *comportement*, *message* et *graphe*.

La dimension *profil* contient les indicateurs relatifs au profil numérique de l'entité sur la plateforme sociale. Elle contient généralement l'âge, le sexe et le lieu associé au profil.

Le *comportement* est défini par un ensemble d'indicateurs dynamiques, tels que le nombre de messages envoyés et leur fréquence.

	Profil				**Comportement**		
	Age	*Genre*	*Position*	*Autres*	*Nombre messages*	*Fréquence messages*	*Autres*
[27]					✓		
[78]					✓		
[26]							
[75]	✓				✓	✓	✓
[79, 32]							
[67, 69, 74]	✓	✓	✓	✓	✓	✓	

Tableau 2.2 – Indicateurs utilisés par les différentes approches selon les dimensions profil et comportement.

La dimension *message* concerne les attributs relatifs aux messages. Par exemple, la similarité entre les messages est souvent utilisée, car elle permet de détecter la redondance de messages. La quantité d'URL, le nombre de références et de citations observées dans les messages sont autant de variables que l'on peut aussi prendre en compte.

La dimension *graphe* reflète les indicateurs issus de la théorie des graphes, qui peuvent avoir de l'importance pour révéler la position d'un profil au sein du réseau social numérique. Ces indicateurs peuvent être locaux (p. ex. degré du nœud), de voisinage (p. ex. dépendant du degré des nœuds voisins) ou globaux (issus d'analyses globales du graphe). Cette dernière est très coûteuse dans l'analyse des plateformes numériques, car leur taille est très grande (plusieurs millions d'utilisateurs).

Les tableaux 2.2 et 2.3 présentent les indicateurs utilisés par les différents travaux de la littérature.

	Message					**Graphe**	
	Similarité	*Nombre URL*	*Nombre réfé-rences*	*Nombre mots-clés*	*Autres*	*Local*	*Global*
[27]	✓	✓				✓	
[72]		✓					
[26]	✓	✓				✓	
[75]		✓			✓	✓	
[79, 32]						✓	✓
[67, 69, 74]	✓	✓	✓	✓		✓	

Tableau 2.3 – Indicateurs utilisés par les différentes approches de la littérature selon les dimensions message et graphe

2.2.2 Les solutions fondées sur les techniques de groupement de données

Les techniques de groupement de données consistent à rassembler un ensemble d'objets en groupes. Chaque groupe contient des objets qui partagent une plus grande similitude entre eux qu'avec les objets des autres groupes.

Dans le cas de l'analyse des réseaux sociaux numériques, les objets considérés sont pour la plupart du temps des profils d'utilisateurs. Il s'agit alors de définir une notion de distance entre deux profils à partir de caractéristiques définies par le modèle proposé.

L'approche [70], issue de la collaboration entre l'université Northwestern et l'université de Santa Barbara, pour détecter les profils malveillants sur Facebook est principalement fondée sur les techniques de groupement de données. Les auteurs proposent de collecter les messages de la plateforme Facebook et de ne traiter que ceux contenant des liens vers des sites Web. L'objectif principal est de détecter des profils propageant des URL malveillantes. Les auteurs déduisent ensuite les connexions entres les messages et les profils à partir du contenu de ceux-ci. Les deux principales sources de liens sont les URL citées et les mots-clés référencés dans les messages. Les approches de groupement de données traditionnelles sont ensuite appliquées sur le graphe généré pour en extraire des groupes de profils malveillants. Les groupes de profils malveillants sont identifiés à l'aide de deux propriétés : la synchronisation des messages envoyés à l'intérieur du groupe et la quantité de profils ayant publié des messages.

Une approche de mise en évidence de profils malveillants propre aux services basés sur la localisation géographique a été proposée par les auteurs de [80] de l'université fédérale de Minas Gerais (Brésil). Cette approche prend en considération un indicateur important de la plateforme Foursquare : les recommandations. Les recommandations sont des messages associés à un lieu particulier, qui visent à indiquer aux autres utilisateurs un fait marquant nécessitant d'être relevé. Les autres utilisateurs peuvent ensuite ajouter et confirmer ou non la recommandation. Le travail proposé permet d'identifier les utilisateurs indiquant des recommandations qui ne correspondent pas avec le lieu indiqué. Ces recommandations contiennent la plupart du temps un lien vers un site Web qui n'a aucun rapport avec le lieu. Les auteurs ont groupé les utilisateurs en fonction des comportements mesurés sur la plateforme. Parmi les indicateurs pris en compte, le nombre de recommandations publiées, le nombre de messages reçus en retour de ces recommandations, le nombre de lieux indiqués comme visités, et le ratio du nombre de recommandations comportant des URL. Parmi les groupes d'utilisateurs mis en évidence, certains possèdent des caractéristiques correspondant à un comportement malveillant.

Les auteurs de [73] de l'université du Roi-Saoud ont proposé une approche par groupement de données basée sur la représentation des données sous forme de graphe. Les profils analysés (les nœuds) sont liés entre eux par un arc de poids calculé comme l'agrégat de trois indicateurs. Les trois indicateurs utilisés sont : la quantité d'amis, la quantité de pages et d'URL que les profils partagent. Les auteurs proposent ensuite d'appliquer une technique de groupement de données sur le graphe généré. Les résultats mettent en évidence l'existence de groupes de profils majoritairement composés de profils malveillants et de groupes majoritairement composés de profils inoffensifs. Cependant, certains groupes de profils sont mixtes, ce qui constitue une limite non négligeable à l'approche.

2.2.3 Les solutions basées sur l'analyse sémantique latente

L'analyse sémantique latente est un procédé de traitement de contenu de corpus de textes. Elle a pour objectif de déceler les relations entre les multiples documents et les termes d'un corpus. Cette approche repose généralement sur la construction d'une matrice de co-occurrence de termes dans les documents. Une décomposition matricielle est ensuite appliquée pour estimer l'usage des termes au travers de la totalité des documents. Chaque document et chaque terme sont alors représentés sous forme de vecteurs et la comparaison

des documents peut s'effectuer à l'aide de mesures de distance. Cette approche peut s'appliquer à des documents, des courriels, des *tweets*, etc. Généralement, une réduction en deux ou trois dimensions principales est appliquée afin d'obtenir une représentation visuelle de la proximité entre le document et/ou les termes dans un espace représentable graphiquement.

L'université nationale en informatique du Pakistan a proposée l'utilisation de cette approche pour la détection de profils malveillants sur les plateformes sociales numériques. Celle-ci repose sur la représentation des messages postés et de leurs termes parmi un corpus de messages de la plateforme [71]. Un corpus de messages (malveillants et bienveillants) est récolté puis représenté sous forme de matrice de co-occurrence. L'analyse sémantique latente est appliquée, puis les vecteurs propres résultants sont stockés dans une base de données de référence. Pour détecter un message malveillant, on recherche le vecteur de la base de données le plus proche du vecteur considéré. Dans ce but, l'approche la plus commune consiste à évaluer la similarité entre deux entités comme étant le cosinus de l'angle entre les deux vecteurs. Si le message le plus proches est un message malveillant, le message est identifié comme malveillant, et dans le cas contraire il est considéré comme non malveillant.

2.2.4 Les approches fondées sur la détection d'anomalies

Le domaine de la détection d'anomalies, souvent utilisé pour la prise de décision, a pour objectif de trouver des *pattern* dans les données qui ne sont pas conformes au comportement attendu [81]. Avec la croissance exponentielle des réseaux, la détection d'anomalies couvre un large éventail de domaines d'application tels que la détection de fraude (p. ex. les cartes de crédit), la détection d'intrusion, etc. [82].

La détection d'anomalies est une préoccupation centrale en cyber-sécurité au XXI^{ème} siècle [83, 84]. En effet, comme le stipule D.B. Skilicorn « même si tout comportement anormal n'est pas malveillant il est probable qu'un comportement malveillant soit anormal ».

Plusieurs méthodes sont utilisées pour détecter des anomalies dans un ensemble de données telles que les techniques statistiques, les K-plus proches voisins, les approches fondées sur le groupement de données et les approches spectrales [85]. Notons que les approches spectrales peuvent aussi être appliquées à des données qui sont représentées sous forme de graphes.

La plupart des algorithmes fournissent en sortie un score reflétant dans quelle mesure une entité donnée peut être considérée comme anormale. La figure 2.2.1 donne un aperçu des différentes approches existantes dans ce domaine.

Dans le reste de cette section, nous présentons la technique la plus communément utilisée : les K-plus proches voisins.

La technique des K-plus proches voisins (K-NN)

Cette méthode est classée comme une approche de détection de points anormaux basée sur la notion de densité ou de distance [86]. Ainsi, un élément donné sera considéré comme anormal s'il se trouve éloigné des autres. Dans ce cadre, tout nœud normal possède un voisinage compact ou dense.

La méthode $K - NN$ affecte à chaque nœud un score qui est égal à sa distance par rapport à son K^{eme} plus proche voisin. Les observations dont les scores sont plus élevés qu'un certain seuil sont classées comme anormales. On fait référence à $1 - NN$ ($K - NN$

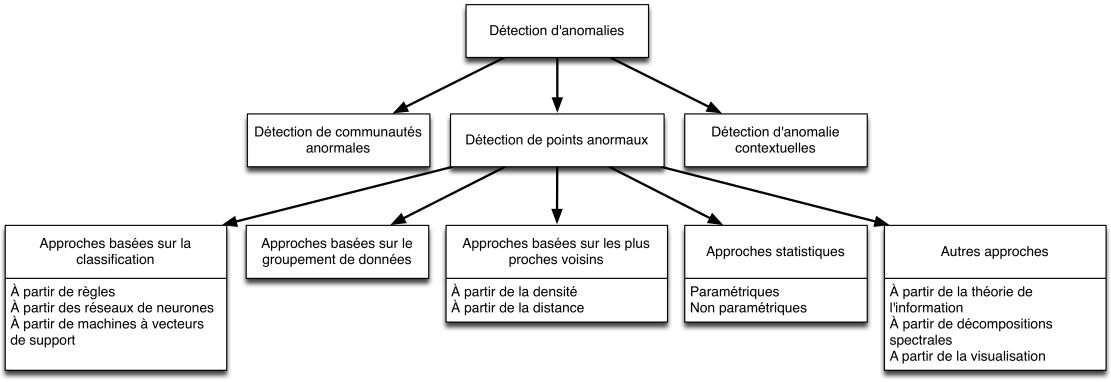


FIGURE 2.2.1 – Les champs de la détection d'anomalies.

pour $K = 1$) comme méthode du plus proche voisin. Dans ce cas, le score affecté à chaque nœud est simplement la distance à son voisin le plus proche.

Afin de déterminer les scores d'anomalies, il est nécessaire de définir la notion de distance. Cette distance est fréquemment fondée sur des critères spécifiques et les nombreuses approches de la littérature illustrent la diversité des distances applicables pour résoudre ce problème [87, 88]. Dans l'analyse des sociaux réseaux, cette distance peut dépendre des attributs et des liens entre les nœuds. Dans ce cas, l'approche $K - NN$ peut être mise en œuvre après une analyse préliminaire.

Notons que le score d'anomalies est sensible au paramètre K : une trop faible valeur de K peut conduire à un score de valeur aberrante et bruité alors qu'un score élevé peut conduire à trop lisser l'ensemble de scores. Dans [89], les auteurs suggèrent la méthode de validation croisée pour comparer et évaluer l'efficacité du paramètre K. Pour plus de détails sur la méthode $K - NN$ le lecteur peut se référer à [90].

2.2.5 Les solutions de détection de réseaux de profils malveillants

Nous présentons ci-dessous les approches visant à détecter les connexions entre les profils malveillants tels que les campagnes de *spams* sur les réseaux sociaux numériques.

Nous nous référons à des campagnes suspectes, comme un ensemble de profils suspects qui exécutent des actions similaires et agissent en coordination. Nous nous référons à des campagnes malveillantes comme un ensemble de profils qui exécutent des actions similaires et agissent en coordination pour propager des liens vers des sites Web malveillants.

[27] de l'université de Santa Barbara propose de récupérer les connexions entre les profils malveillants sur la base de leurs URL publiées. L'hypothèse est que deux profils qui publient des messages contenant la même URL malveillante appartiennent à la même campagne malveillante. Il est important de noter que l'utilisation des services de raccourcissement d'URL peut compliquer cette tâche.

Certaines approches reposent sur un algorithme de classification qui sert à distinguer les communautés de profils normaux des communautés de profils malveillants. Par exemple, [73] a proposé d'appliquer un algorithme de *clustering* sur un vaste ensemble de caractéristiques comportementales et basées sur le contenu des messages. Les résultats montrent que cette approche permet d'identifier les profils producteurs de contenus non sollicités ou malveillants avec une bonne précision. Cependant, cette approche ne permet pas l'identification des connexions étroites entre les profils de *spam* (p. ex. campagnes de *spams*).

[30] a introduit le concept de *Socialbot Networks* (SBN), un ensemble de profils qui

sont détenus et gérés par un contrôleur humain unique. Ces auteurs de l'université de British Columbia ont démontré l'efficacité de ces réseaux et ont également souligné que la connexion de profils gérés automatiquement peut artificiellement augmenter l'impact d'un profil malveillant donné.

[79] a enquêté sur les relations existantes entre les profils malveillants sur les réseaux sociaux en ligne. Ce travail, soutenu par le centre Max Planck Indo-German Centre for Computer Science (IMPECS), a permis d'analyser le problème du *« link farming »* qui correspond à la création de liaisons illégitimes afin d'accroître artificiellement la visibilité des profils. Les résultats de l'approche ont montré que les *spammers* utilisent fortement cette technique. Le travail a aussi mis en évidence le fait que de nombreux profils légitimes acceptent et génèrent des connexions avec des profils *spammers* pour augmenter eux aussi leur audience. Ce phénomène rend difficile la distinction de campagnes malveillantes en utilisant uniquement le graphe social d'une plateforme.

Le travail [91] réalisé à l'université d'État de Pennsylvanie a proposé une solution pour identifier la propagation des vers à l'intérieur d'un réseau social en ligne. Les auteurs proposent de créer un réseau leurre de profils et d'analyser le contenu reçu par ces profils. La détection repose sur l'utilisation de mesures de similarité entre les messages afin de déterminer s'il existe des preuves de propagation de contenus malveillants.

Les auteurs de [92] ont proposé une approche pour détecter les campagnes de spams sur Twitter. L'approche propose une mesure de similarité basée sur la quantité d'URL publiée conjointement entre deux comptes. De cette mesure, les auteurs de l'université de Floride construisent un graphe de similarité à partir duquel des sous-graphes sont identifiés comme des campagnes. Une classification des campagnes est réalisée en appliquant une classification (basée sur les machines à vecteurs de support) sur les caractéristiques des campagnes.

Une autre méthode de mise en évidence de la proximité entre deux profils peut être réalisée par la méthode de l'attribution d'auteurs. L'attribution d'auteurs vise à identifier les principales caractéristiques liées aux spécificités d'écriture d'un même individus sous des entités numériques distinctes. Une approche basée sur cette technique combinée avec des indicateurs comportementaux sera présentée dans cette thèse [93].

Les actions malveillantes connaissent un taux de réussite relativement important et une rapidité de diffusion extrêmement large. L'une des causes principales de ce succès est la notion de confiance qui existe entre les utilisateurs. Les amis sont très souvent considérés comme des entités de confiance et l'utilisateur n'émet pas de doute quant à la légitimité des messages provenant de ces amis. Cette observation est l'une des raisons principales expliquant la rapidité de diffusion du ver *Koobface*, qui a menacé les données personnelles de 200 millions de personnes. Il est désormais prouvé que la simple relation d'amitié existant entre utilisateurs de Facebook n'est pas une garantie suffisante de légitimité de celle-ci [34]. Pour répondre à cette problématique, de nombreux travaux se sont penchés sur la création d'outils et d'algorithmes d'évaluation de la confiance sur les plateformes numériques sociales. Nous en proposons un aperçu dans la suite de ce chapitre.

2.3 L'évaluation de la confiance sur les réseaux sociaux numériques

[94] a proposé une définition de la confiance pour les services basés sur le Web. Ce travail de l'université du Maryland établit que la confiance en une personne est *« l'engagement dans une action avec une personne basé sur la croyance que les futures actions de cette personne vont nous être bénéfiques »*.

Dans ce contexte, la notion de confiance est donc un facteur critique pour établir ou non la sécurité d'un individu ou de ses données. Un nombre important de travaux sont consacrés à la mise en place d'un indicateur de confiance sur les profils de plateformes sociales numériques. Les systèmes d'évaluation de confiance ont des applications diverses mais un objectif commun : aider l'utilisateur dans ses choix.

Ces indicateurs doivent être sans cesse améliorés et adaptés pour prendre en compte une diversité de comportements frauduleux et pour s'adapter à une majorité de réseaux sociaux numériques.

Nous pouvons distinguer deux grands types d'évaluations de la confiance sur les médias sociaux : tout d'abord les approches proposant un calcul d'évaluation directe de la confiance d'un profil, ensuite les approches fondées sur la notion de propagation de cette confiance sur les réseaux. Des discussions sur les avantages et inconvénients des deux approches sont disponibles dans le travail [95].

2.3.1 Approches topologiques

2.3.1.1 Par la mesure de similarité locale

Plusieurs indicateurs de similarité locale ont été proposés et une liste exhaustive peut être trouvée dans [96] et [97]. La mesure de similarité locale la plus simple entre un nœud x et un nœud y est l'indice des voisins communs (CN), qui compte simplement le nombre d'occurrences de voisins communs entre une paire de nœuds (x, y). Cet indicateur suppose qu'un lien entre deux nœuds est plus probable s'ils partagent un nombre important de voisins en commun. Le réseau social Facebook utilise cette hypothèse pour recommander des amis à un utilisateur [98]. D'autres mesures existent, telles que l'indice de Salton, Jaccard, Sørensen, valorisation des hubs (HPI), dévalorisation des hubs (HDI), Leicht-Holme-Newman (LHN), l'attachement préférentiel (PA), Adamic Adar (AA) et l'Allocation des ressources (RA). Le tableau 2.4 précise la mesure de similarité pour chaque indicateur existant. Dans cette table, nous noterons par $\Gamma(x)$ l'ensemble des voisins de x, et k_x le nombre de ses voisins (c.-à-d. $k_x = |\Gamma(x)|$).

L'indicateur de Jaccard (JA) compte le nombre d'amis communs mais assure la normalisation du score de similarité. Il prend en compte l'union des deux ensembles de voisins à cet effet.

L'indicateur de Salton (SA) est largement utilisé dans la littérature et est souvent désigné comme la similarité cosinus, ou coefficient d'Ochiai, en biologie. Cet indicateur est égal à la similarité cosinus entre les deux vecteurs de voisins.

L'indicateur de Sørensen donne un score de similarité qui est égal au ratio de deux fois la quantité de contacts partagés par le nombre total de contacts.

Les indicateurs HPI et HDI (*Hub promoted index* et *Hub deprecated index*) fournissent un moyen pour donner plus ou moins d'importance aux liens qui sont adjacents à des hubs [99]. Sur Facebook, l'utilisation de l'indicateur HDI pour prédire l'amitié permet de

Indicateur	Formule
Voisins communs	$s_{xy}^{CN} = \lvert\Gamma(x) \cap \Gamma(y)\rvert$
Jaccard	$s_{xy}^{Jaccard} = \frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert}$
Salton	$s_{xy}^{Salton} = \frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\sqrt{k_x * k_y}}$
Sørensen	$s_{xy}^{Sørensen} = \frac{2*\lvert\Gamma(x)\cap\Gamma(y)\rvert}{k_x+k_y}$
HPI	$s_{xy}^{HPI} = \frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{min(k_x,k_y)}$
HDI	$s_{xy}^{HDI} = \frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{max(k_x,k_y)}$
Leicht-Holme-Newman	$s_{xy}^{LHN} = \frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{k_x * k_y}$
Attachement préférentiel	$s_{xy}^{PA} = k_x * k_y$
Adamic Adar	$s_{xy}^{AA} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{log(k_z)}$
Allocation des ressources	$s_{xy}^{RA} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{k_z}$

Tableau 2.4 : Principaux indicateurs locaux de prédiction des liens

ne pas prendre en compte la similitude d'un profil avec celui d'une célébrité, qui a un très grand nombre de contacts.
L'indicateur d'attachement préférentiel (PA) suppose que la similarité entre deux nœuds est proportionnelle au produit de leur degré. Le modèle sous-jacent est utilisé pour générer des réseaux libres d'échelle. L'application d'un tel modèle aux réseaux sociaux implique que les profils célèbres sont plus susceptibles de créer des connexions que les profils non célèbres. Pour la même raison, ces connexions seront plus susceptibles d'impliquer d'autres profils célèbres.

La mesure de Leicht-Holme-Newman (LHN) donne un score élevé de similitude entre deux nœuds possédant un grand nombre de voisins communs en ce qui concerne le nombre attendu de ces voisins dans le modèle de configuration [100]. Le modèle de configuration défini dans [101] est une réalisation aléatoire d'un réseau particulier.

L'indicateur Adamic-Adar (AA) compte les voisins communs, en donnant aux nœuds les moins connectés plus de poids [102]. Cela peut être interprété comme l'insignifiance de partager un ami commun qui est une célébrité sur le réseau. Au contraire, l'obtention d'un ami commun avec une faible quantité de contacts aura un poids plus important. Cela permet d'illustrer le fait qu'une personne avec une faible quantité d'amis peut choisir plus rigoureusement ses contacts qu'une personne célèbre qui peut accepter l'amitié de beaucoup de gens.

L'indicateur d'allocation des ressources (RA) ajoute pour chaque voisin commun l'inverse du nombre de ses contacts. Cet indicateur, comme son nom l'indique, est étroitement lié au processus physique d'allocation de ressources [103]. Cet indicateur évalue la quan-

tité d'information qui peut être envoyé de x à y par leurs voisins communs. Chaque voisin commun est identifié comme étant un émetteur possédant une ressource qu'il va distribuer de manière équitable à tous ses voisins. Le montant total des ressources reçues par y provenant de x correspond au score de similitude.

2.3.1.2 Par l'analyse à grande échelle

Certains algorithmes de mesure de confiance sur les RSN se basent sur les algorithmes de centralité ou de prestige à une échelle globale. Dans un premier temps, nous présenterons les algorithmes incontournables de la mesure d'influence traditionnellement appliqués aux pages Web. Nous présenterons ensuite les adaptations de ces algorithmes dans le cas de la mesure de la confiance à échelle globale et sur les réseaux sociaux numériques.

Les algorithmes présentés ici se basent sur la modélisation de l'Internet par un graphe. On considère donc l'ensemble des pages Web comme des nœuds et l'ensemble des liens vers des pages comme des arcs orientés. Lorsqu'un arc existe entre un nœud $\mathcal{A}$ et un nœud $\mathcal{B}$ alors cela signifie que la page $\mathcal{A}$ possède un lien vers la page $\mathcal{B}$.

L'algorithme *Hyperlink-Induced Topic Search* (HITS) est un algorithme de référence de la mesure de la pertinence de pages Web [104]. Cet algorithme considère l'existence des deux types de pages Web : les *hubs* et les autorités. Les autorités sont des pages très connectées tandis que les hubs sont des pages qui contiennent des liens vers les autorités. Pour chaque requête, les deux scores sont calculés sur un sous-ensemble de pages correspondant aux critères de recherche.

L'algorithme PageRank est certainement le plus célèbre des algorithmes de mesure de la réputation des pages Web [46]. NodeRank est une adaptation de PageRank qui apporte deux évolutions [105]. La première énonce que le graphe peut être pondéré, ainsi le surfeur possède une probabilité non équiprobable de choisir une page. La probabilité du surfeur de choisir une page est alors proportionnelle au poids relatif de chaque lien à sa disposition. La deuxième concerne la probabilité de se déplacer à tout endroit du graphe à chaque itération. L'approche NodeRank propose de rendre cette probabilité proportionnelle à la quantité de pages qu'il est possible de choisir. Si la quantité de pages à choisir est trop restrictive, alors le surfeur décide de se déplacer.

Les travaux de [32] et de [79] ont permis d'adapter et de comparer chacun de ces algorithmes dans le cas des réseaux sociaux numériques pour construire un indicateur de confiance et ainsi détecter des individus malveillants. Dans le cas présent, le graphe sur lequel s'appliquent les algorithmes est le graphe constitué des utilisateurs d'une plateforme sociale numérique et les connexions entre ces nœuds peuvent être identifiées par la notion de contact.

De manière assez surprenante, les auteurs montrent qu'une partie des utilisateurs ayant un niveau de réputation important avec PageRank sont en fait, des profils malveillants (le *link farming* est une explication possible). Les auteurs concluent donc qu'il est nécessaire dans le futur d'intégrer d'autres paramètres dans ces algorithmes pour les rendre robustes à ce type de problèmes.

2.3.2 Les approches fondées sur les systèmes de réputation

Sur certaines plateformes, un utilisateur qui effectue une action engageant un autre utilisateur peut signaler aux autres utilisateurs son opinion sur celle-ci. Un exemple est WOT (Web of Trust), qui permet aux utilisateurs de donner leur opinion sur la visite d'un site Web et ainsi obtenir son niveau de sécurité. Le même principe existe pour de nombreux types d'entités différentes, et dans le cas des RSN, on souhaite souvent qualifier un message ou un utilisateur.

De manière générale, un système de mesure de confiance ou de réputation doit comporter au minimum trois actions fondamentales :

— Générer un retour d'expérience sur une action/un utilisateur.

— Diffuser ce retour d'expérience à large échelle.

— Agréger l'opinion pour donner un score de réputation.

La manière d'agréger l'opinion des utilisateurs est définie par le calcul de l'indicateur de confiance. La qualité et la quantité des informations transmises lors du retour d'expérience sont alors critiques pour l'évaluation de cet indicateur. Certaines techniques d'agrégation peuvent simplement se baser sur une valeur moyenne, une sommation, les densités de probabilités obtenues des différentes notes.

L'approche la plus triviale pour évaluer la confiance entre deux individus qui ne sont pas reliés (qui n'ont pas déjà interagi) repose sur le calcul du chemin le plus favorable en termes de confiance. On peut alors conserver comme score le maximum du produit de tous les scores de confiance qui apparaissent sur le plus court chemin. Cette approche comporte plusieurs biais, dont celui d'être trop optimiste pour être applicable. Pour résoudre cet aspect, il est possible d'évaluer la confiance comme la valeur moyenne des confiances obtenues sur tous les chemins entre les deux individus. Cependant, la grande connectivité entre les utilisateurs peut avoir tendance à rendre très faible cette valeur. De plus, la complexité de l'algorithme de recherche de chemins entre deux nœuds sur un graphe de grande taille (comportant des millions de nœuds) pose d'importantes limitations à ces solutions de bon sens.

Pour répondre aux limitations de ces approches, les auteurs de [106] ont proposé la notion de confiance omniprésente. Leur approche repose sur le calcul de la confiance entre $\mathcal{A}$lice et $\mathcal{B}$ob par rapport aux chemins qui existent entre ces deux utilisateurs en ajoutant une contrainte supplémentaire. En effet, un filtrage sur les chemins pris en compte est effectué pour ne traiter que les chemins passant par les utilisateurs ayant une forte confiance en $\mathcal{B}$ob.

Un ensemble d'algorithmes tels que EigenTrust, Trust-WebRank a été proposé pour permettre d'évaluer la confiance relative entre tout utilisateur en fonction de la confiance qu'il donne à ses contacts [107, 108]. Ces approches reposent sur le calcul du vecteur propre associé à la plus forte valeur propre. Ce vecteur est composé des scores relatifs de confiance entre chaque paire d'individus.

L'un des principaux problèmes associés aux systèmes de réputation est la nécessité de normaliser les scores locaux de confiance. La normalisation effectuée a naturellement tendance à dévaluer la confiance pour un utilisateur qui a effectué beaucoup d'évaluations par rapport aux autres. Pour autant, il n'y a pas de raison pouvant expliquer qu'un utilisateur affectant moins souvent un indicateur de confiance aux autres utilisateurs ait un impact plus important.

Contribution	Approche	Évaluation	Indicateurs
[109]	Contrôle d'accès	Diffusion	-
[110]	Mesure indirecte	Diffusion	-
[111]	Contrôle d'accès	Diffusion	-
[112]		Diffusion	-
[106]	Mesure indirecte	Diffusion	-
[113]	Contrôle d'accès	Diffusion	-
[79]	Classement	Diffusion	Topologie
[32, 114]	Classement	Diffusion	Topologie
[115]	Réputation	Directe	Types de relations
[116]	Réputation	Directe	Interactions
[117]	Graphe de confiance	Directe	Domaines
[44]	Topologique	Directe	Degrés

Tableau 2.5 – Approches d'évaluation de la confiance sur les réseaux sociaux numériques

Nous n'avons pas présenté en détail les algorithmes de filtrage collaboratif, qui font toutefois partie des algorithmes de diffusion de la confiance sur les réseaux sociaux numériques. Ceux-ci reposent sur un graphe bipartite composé d'utilisateurs mais aussi de produits (p. ex. items achetés).

Dans cette thèse, notre travail se focalise principalement sur les réseaux sociaux numériques, nous ne disposons donc pas d'une évaluation des entités analysées et le graphe d'analyse est n'est donc par nature pas bipartite.

2.3.3 Les limites de l'évaluation de la confiance sur les réseaux sociaux numériques

Les algorithmes d'évaluation de la confiance d'un utilisateur sur les réseaux sociaux numériques posent de nombreuses contraintes difficiles à résoudre. L'une des contraintes majeures est la nécessité pour les acteurs d'affecter un score aux utilisateurs de la plateforme. Ainsi, il est possible de biaiser l'indicateur par intégration de scores frauduleux ajoutés automatiquement par un unique profil ou un cartel de profils (p. ex. *shilling attack*). De plus, une majorité d'utilisateurs peuvent ne pas participer activement au système d'évaluation ce qui en limite les résultats. Notons que, sur la majorité des réseaux sociaux numériques, la mise en place d'évaluation n'est pas rendue possible (encore moins obligatoire) par le système. Notons aussi que l'évaluation d'un utilisateur ou d'une action par un utilisateur nécessite une connaissance de celui-ci. Tandis que sur la réception d'un produit il est facile d'évaluer la qualité de la transaction, lorsqu'il s'agit d'évaluer la potentielle dangerosité d'un profil vis-à-vis de la fuite de données, le problème devient beaucoup plus complexe. Un utilisateur pourrait affecter sa confiance en un ami qui, soit n'est pas réellement cette personne, soit effectue des actions malveillantes qui ne lui sont pas visibles. Ainsi, un score positif de confiance pourrait perdre tout son sens.

Conclusion

Nous avons présenté un ensemble d'outils pour l'analyse des réseaux sociaux numériques. Ceux-ci reposent en très grande partie sur les techniques issues de la théorie des graphes et aussi de fouille de données. En particulier, le modèle multicouche semble à ce jour, être le plus complet dans le sens ou il propose d'unifier des identités numériques diverses d'un même et unique utilisateur. Face aux nombreuses menaces malveillantes qui pèsent sur les réseaux sociaux numériques, nous avons présenté un état de l'art des techniques permettant (1) de détecter la malveillance, (2) de mesurer la confiance entre deux utilisateurs. L'évaluation de la confiance entre les utilisateurs est un problème critique pour la sécurité des interactions des utilisateurs de réseaux sociaux numériques. En effet, la simple consultation d'une URL contenue dans un message ou l'acceptation d'une demande d'amitié peut être, par exemple, à l'origine de fuite de données. Les approches présentées de l'état de l'art mettent en évidence le manque d'outils exploitables pour guider et assurer un niveau de sécurité de l'utilisateur dans son usage quotidien des réseaux sociaux numériques.

Partie II : Approches pour la sécurisation des utilisateurs de réseaux sociaux numériques

« Only in action can you fully realize the forces operative in social behavior. »

(S.Milgram, 1974)

Chapitre 3

Détection de comportements malveillants sur les réseaux sociaux numériques : le cas de Twitter

Résumé du chapitre

Ce chapitre présente le premier apport du travail, à savoir la détection comportementale, à large échelle, des acteurs malveillants d'un réseau social numérique. Tous les jours, environ cinquante millions de messages sont générés par plus de cinq cents millions de profils sur Twitter. Certains utilisateurs tentent d'exploiter le succès de cette plateforme de microblogging afin d'effectuer des actions malveillantes qui peuvent, par exemple, conduire à la fuite de données. Nous présentons, dans ce chapitre, une approche comportementale développée pour l'évaluation de profils sur Twitter (SPOT 1.0) grâce à un indicateur à trois dimensions qui implique le degré d'$\mathcal{A}$*ctivité, de* $\mathcal{V}$*isibilité et le niveau de* $\mathcal{D}$*anger. Une extension de ce travail, nommée REPLOT, a été réalisée et permet de détecter des campagnes de profils suspects. Cette méthodologie contient trois grandes phases : (1) les profils individuels sont analysés par SPOT 1.0 pour déterminer s'ils sont suspects ou non ; (2) les connexions entre les profils suspects sont identifiées en utilisant une approche de fusion de données, de modèles temporels et d'analyse de caractérisation d'auteurs ; (3) une technique de groupement de données est utilisée pour profiler les différentes campagnes malveillantes. Par cette méthode, les profils malveillants ne sont pas seulement découverts automatiquement, mais aussi profilés. Pour un analyste ayant besoin de découvrir les tendances et les modèles d'attaque des profils, ce niveau d'automatisation permet de réduire le nombre de profils devant être visités. Ainsi, plutôt que d'analyser un très grand nombre de profils individuels, un plus petit nombre de groupes peut être analysé.*

3.1 Introduction

La présence de sites Web malveillants sur l'Internet ne peut être contestée et de nombreux efforts sont entrepris pour les identifier (p. ex. la base PhishTank, le plugin WOT). À titre d'exemple, [118] présente une liste d'outils disponibles pour signaler et vérifier la qualité d'un site Web à partir de son URL.

Afin d'augmenter la visibilité d'un site malveillant, les auteurs ont généralement recours à des techniques de diffusion massive d'URL. L'un des principaux vecteurs de diffusion de ces sites Web malveillants a été jusqu'alors les courriels. Pour lutter contre cette menace, des approches de détection de courriels suspects ont été menées, celles-ci pouvant être fondées uniquement sur le contenu des courriels [119, 120], ou prendre en compte également les caractéristiques des sites référencés [77, 121].

Le succès des plateformes comme Twitter et la spontanéité des échanges ont fait des sites de *microbbloging* et des réseaux sociaux numériques un nouveau vecteur de propagation d'URL malveillantes. De récents travaux ont été proposés pour lutter contre cette nouvelle forme de menaces sur Twitter. [122, 26] proposent une classification des *tweets* pour détecter les profils malveillants à partir de leur comportement et du contenu des messages envoyés.

Nous proposons, dans ce chapitre, une méthodologie pour détecter et évaluer l'impact de profils malveillants sur Twitter. Notre étude est menée suivant trois axes : (1) le profil utilisateur, (2) les messages et (3) les liens que peuvent contenir ces messages. Cette méthodologie a donné lieu à la réalisation d'un outil qui présente trois avantages par rapport à l'état de l'art existant : 1) il travaille en temps réel, 2) il intègre un outil de visualisation permettant l'aide à la décision et 3) il mesure la présence réelle de malveillance.

L'outil proposé nommé SPOT (en anglais Scoring suspicious Profiles On Twitter) effectue une extraction des données à partir de la *timeline* de Twitter grâce à ses interfaces de programmation (API). L'analyse et l'évaluation dynamique de ces données permettent d'évaluer les profils.

Nous appellerons « profil suspect » tout profil présentant une ou plusieurs caractéristiques pouvant révéler un comportement anormal. Nous appellerons « profil malveillant » tout profil suspect ayant diffusé, au moins une fois, une URL malveillante sur le réseau.

3.2 Méthodologie

Notre méthodologie repose sur trois principaux axes d'investigation : (1) les profils utilisateurs, (2) les messages diffusés par ces profils et (3) les URL contenues dans les messages. Chacun de ces axes est analysé par le biais de multiples critères dans le but de détecter des profils malveillants. La figure 3.2.1 met en évidence les relations entre les trois axes et les principales caractéristiques qui leur sont associées. Dans la suite de cette section, nous présentons en détail les caractéristiques de chaque axe d'investigation.

3.2.1 La dimension profil

Un individu s'identifie sur une plateforme sociale numérique par un profil. Sur Twitter, les attributs d'un profil sont multiples, et la figure 3.2.2 en illustre une partie. Le profil représenté est celui de la plateforme elle-même (c'est-à-dire @twitter), mais tout profil public peut être extrait sous le même format. Parmi les attributs les plus pertinents,

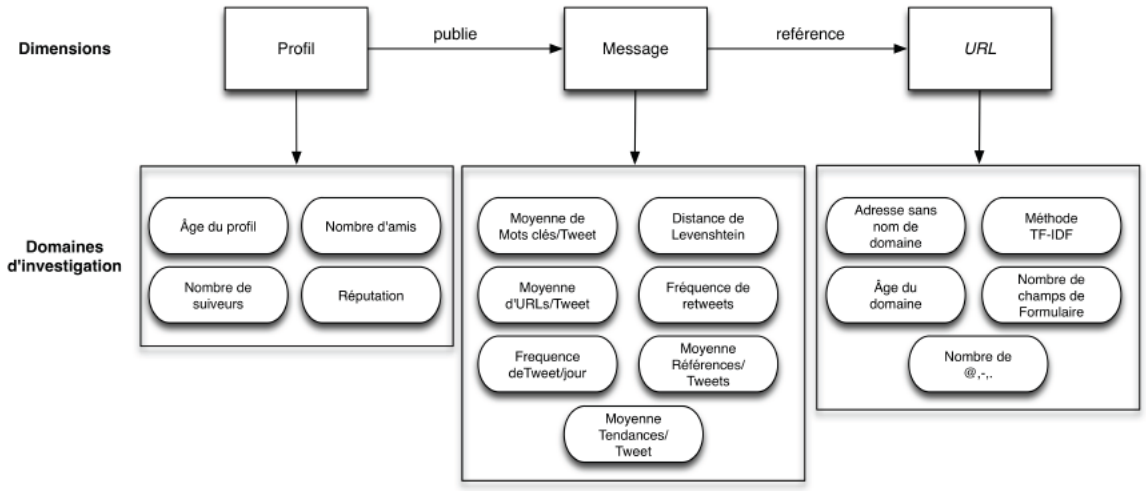


Figure 3.2.1 : Principales caractéristiques associées à chacun des trois axes d'investigation (Profil, Message, URL).

notons le nom de l'utilisateur, son pseudonyme, sa position géographique, sa description, l'adresse de son site Internet, le nombre de ses connexions entrantes et sortantes, et le nombre de messages publiés.

```
1  <user>
2    <id>783214</id>
3    <name>Twitter</name>
4    <screen_name>twitter</screen_name>
5    <location>San Francisco, CA</location>
6    <description>Your official source for news, updates and tips</description>
7    <url>http://blog.twitter.com/</url>
8    <followers_count>17734324</followers_count>
9    <friends_count>120</friends_count>
10   <created_at>Tue Feb 20 14:35:54 +0000 2007</created_at>
11   <time_zone>Pacific Time (US & Canada)</time_zone>
12   <verified>true</verified>
13   <statuses_count>1571</statuses_count>
14   <lang>en</lang>
15 </user>
```

Figure 3.2.2 – Représentation XML du profil @twitter sur la plateforme Twitter.

Généralement, un individu suspect s'abonne à un très grand nombre de contacts sans susciter l'intérêt des autres utilisateurs. Ce nombre très important de *followees* (*friends_count* en figure 3.2.2) est suspect car un individu sur Twitter ne peut gérer et tirer profit d'un nombre élevé d'amis. Par contre, les acteurs malveillants peuvent tirer profit d'un nombre important d'amis. Ils peuvent ainsi créer une base de données de contacts pour analyser le comportement d'un maximum de profils dans le but d'influencer une partie de ceux-ci. Pour détecter un fort déséquilibre entre le nombre d'amis et le nombre de *followers*, un indicateur de réputation peut être calculé [123].

3.2.2 La dimension message

Cette dimension intègre les caractéristiques associées aux messages. On pourrait croire qu'un message envoyé par un utilisateur est transmis uniquement à ses abonnés (c.-à-d. *followers*). Cependant, certaines caractéristiques contenues dans les messages peuvent permettre une plus grande diffusion vers des inconnus ou des cibles prédéterminées. Le message peut contenir directement des mentions à certains utilisateurs (p. ex. @charles___perez), qui recevront de ce fait le message. De plus, l'utilisation de mots-clés (*hashtags*) dans le message (p. ex. #football) peut avoir un impact sur l'audience du message. En effet, les mots-clés sont utilisés par l'outil de recherche de Twitter lorsqu'un utilisateur souhaite rechercher des *tweets* sur un domaine particulier. Ainsi, le fait d'utiliser les mots-clés les plus en vogue dits *« trends »* est une stratégie utilisée pour rendre le message visible à un plus grand nombre d'utilisateurs [122].

La distance de Levenshtein [60] (c.-à-d. nombre d'opérations élémentaires nécessaires pour passer d'un *tweet* à l'autre) entre chaque paire de messages est utilisée pour détecter la génération automatique de messages. Néanmoins, Twitter empêche le renvoi d'un message identique à un message préalablement envoyé. Dans certains cas, les profils automatisés transmettent des messages similaires dont seuls les mots-clés et les références sont modifiés, afin d'atteindre un public plus grand et s'intéressant à des domaines variés.

3.2.3 La dimension URL

La simple consultation d'un message n'a généralement pas de conséquence directe sur les données d'un utilisateur. Cependant, la consultation d'un site Internet dont l'URL est présente dans le message peut avoir un impact. La dimension URL telle que présentée ici est liée à cette notion d'impact. Nous proposons d'analyser et de classifier les URL présentes dans les messages afin de détecter les profils propagateurs d'URL malveillantes. Lorsque les URL sont réduites (par exemple via les services *tinyURL* ou *bitly*), nous récupérons dans un premier temps l'URL longue d'origine. L'âge du domaine est l'une des caractéristiques retenues pour les mêmes raisons que celui du profil. Le nombre de tirets, d'arobases et de points ('-', '@', '.') est aussi un paramètre régulièrement utilisé pour une telle classification. Notons que les paramètres sélectionnés pour une URL sont globalement ceux rencontrés dans la littérature [77, 124]. Enfin, la technique TF-IDF (Term Frequency-Inverse Document Frequency) utilisée pour la création de signatures lexicales d'un site est ici appliquée comme paramètre de notre classification [125, 126]. Cette technique sélectionne un ensemble de mots-clés en s'appuyant sur le score TF-IDF [121]. Lors de la recherche de ces mots sur un moteur de recherche, si le nom de domaine des sites trouvés ne correspond pas avec le nom de domaine du site évalué, alors il est possible que celui-ci soit une copie et qu'il ne soit pas légitime. En général, les sites les plus reconnus sont référencés le plus sur Internet et apparaissent en tête de liste de réponse par des algorithmes tels que PageRank.

3.3 Méthodologie SPOT

Cette section présente l'architecture de la méthodologie SPOT. Cette architecture, composée de six étapes, a pour but non seulement d'identifier les profils suspects ; mais aussi de les évaluer. Celle-ci est présentée en figure 3.3.1. Le premier module, directement relié à la *timeline* de Twitter, établit la connexion avec celui-ci et récupère les messages produits

par les utilisateurs pour les stocker dans une base de données. Depuis cette base de données, le second module génère un ensemble de caractéristiques contenant les indicateurs comportementaux des deux dimensions *profil* et *message*. Cette génération d'attributs est réalisée à chaque fois que la quantité de messages d'un individu dépasse un certain seuil (par défaut quinze messages). À partir de ces indicateurs, le module III classe les profils en deux catégories : suspects et non-suspects. Les profils identifiés comme suspects sont ensuite analysés selon la troisième dimension, les URL contenues dans les *tweets*. Le module IV analyse et stocke les URL contenues dans les tweets de profils suspects et génère un ensemble de caractéristiques qui sera utilisé par le module V pour une seconde classification. Cette seconde classification permet de connaître les URL malveillantes envoyées par les profils suspects. Enfin, un dernier module permet de créer un indicateur de virulence des profils suspects.

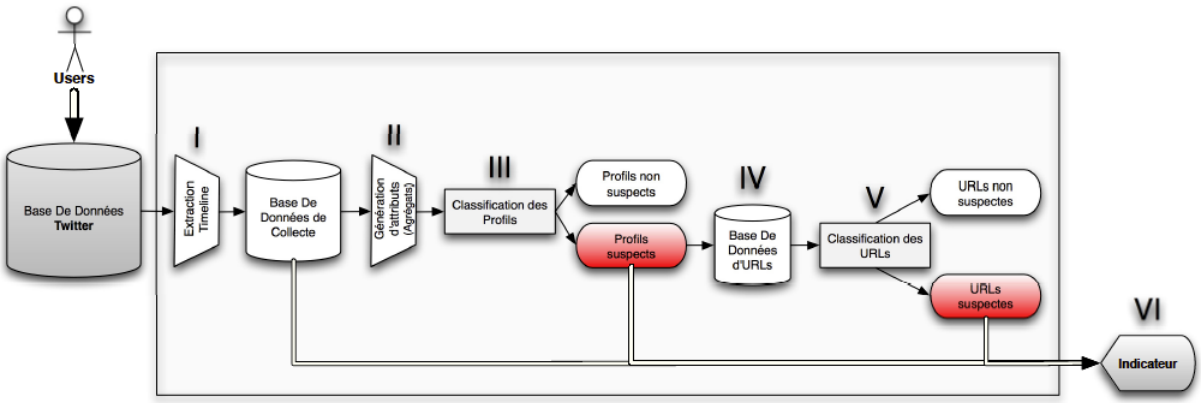


Figure 3.3.1 : Architecture de SPOT.

3.3.1 La collecte de données

Nous pouvons distinguer deux principales techniques permettant de collecter de l'information sur les réseaux sociaux numériques : les API et les *webcrawlers*.

Les « Applications Programming Interfaces » (API) proposent un ensemble de fonctions mises à disposition par un service tel que les réseaux sociaux numériques. Ces fonctions sont utilisables sous certaines réserves (p. ex. quantité d'appels) et permettent d'accéder à certaines données de ces plateformes. En fonction des plateformes, les données accessibles sont plus ou moins restreintes et les formats de données diffèrents (p. ex. json, xml, atom). Les APIs peuvent permettre la collecte de données selon des critères précisés sur les utilisateurs, la position géographique ou des mots-clés.

Les *webcrawlers* constituent une alternative intéressante pour collecter de l'information sur les réseaux sociaux numériques. Les *webcrawlers* sont des outils automatisés qui parcourent un ensemble de pages web pour lesquelles ils ont été configurés. Ils peuvent récupérer le contenu de ces pages pour les traiter et les analyser. Les données accessibles par les *webcrawlers* sont identiques à celles pouvant être visualisées par un utilisateur de la plateforme sociale. Ils sont donc soumis aux limitations dues au niveau de confidentialité de chaque profil.

Nous présentons, dans la suite, un aperçu des interfaces de programmation de Twitter et la solution de collecte qui a été retenue.

3.3.1.1 Les interfaces de programmation de Twitter

Twitter propose trois principales les API *Search*, *Social Graph* et *Streaming* (voir figure 3.3.2). L'API *Search* permet de questionner la base de données de *tweets* de Twitter afin d'en récolter une partie correspondant à une requête. L'API *Social Graph* comporte un ensemble de fonctions plus larges permettant de traiter des demandes relatives aux profils et aux relations entre ces profils. Enfin, l'API de *streaming* permet d'ouvrir une connexion pour pouvoir recevoir des informations de Twitter en temps réel. Ces données sont une partie de la *timeline* générale de Twitter. Cette *timeline* peut être réduite ou non à un ensemble d'individus, de mots-clés ou à un emplacement géographique.

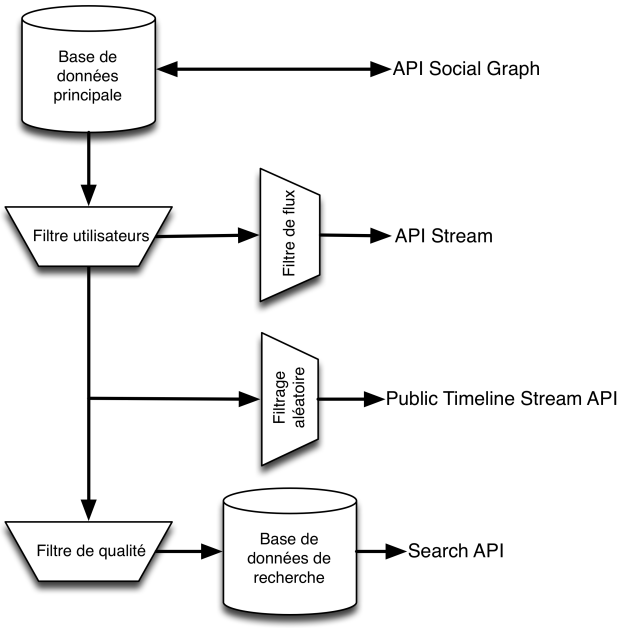


FIGURE 3.3.2 – Taxonomie des interfaces de programmation de Twitter.

L'API *Social Graph* permet de questionner de manière directe la base de données principale de Twitter. Cette API permet d'effectuer des requêtes sur un profil à partir de son pseudonyme ou de son identifiant. Elle permet ainsi d'accéder à une vingtaine de caractéristiques du profil, dont le nombre de *followers*, le nombre de *followees*, la date de création du profil (ligne 17) et le nombre de tweets envoyés depuis sa création (figure 3.2.2).Cette API permet aussi de questionner le réseau sur les relations entre les profils. Bien que cette API soit très riche, car permettant de reconstruire le graphe social de Twitter, elle est soumise à des restrictions et donc peu exploitable pour des collectes de données de grande taille. À ce jour, un utilisateur non authentifié de cette API peut effectuer 150 requêtes par heure, tandis qu'un utilisateur authentifié pourra en effectuer 350.

L'API *Search* de recherche permet comme tout moteur de recherche de trouver une réponse à une recherche précise. Cette API n'est pas reliée directement à la base de données principale de Twitter. En effet, par souci de qualité, le service se connecte à une base de données secondaire constituée de *tweets* répondant à un degré de qualité imposé par Twitter. Le filtre appliqué en entrée de la base de données de recherche permet, entre autres, d'éliminer les *tweets* redondants de la plateforme. L'API permet de sélectionner les *tweets* provenant d'un individu particulier ou contenant un ensemble de mots-clés particulier, mais aussi de récolter des *tweets* provenant d'une certaine zone géographique

à un instant précis. L'API de recherche possède les mêmes contraintes de nombre de requêtes que l'API précédente.

L'API *Stream* a un fonctionnement différent des deux autres. Celle-ci permet d'ouvrir un flux et ainsi de travailler en temps réel sur Twitter. Cette API nécessite obligatoirement l'authentification de son utilisateur et permet d'accéder aux *tweets* envoyés par les utilisateurs de la plateforme. L'API *Stream* fournit environ dix messages par seconde, soit actuellement moins de un pour cent de l'ensemble des messages émis sur sa plateforme. Le processus de sélection des messages retenus parmi l'ensemble des messages de la plateforme est aléatoire.

La solution proposée pour collecter les données du réseau Twitter est la collecte via la *timeline*. La collecte n'est donc pas centrée sur le graphe du réseau mais sur les flux de messages. Cette collecte n'est pas rétroactive, donc seuls les messages envoyés à partir du lancement du programme peuvent être récupérés. Notre outil utilise l'API *Stream* de Twitter. Cette API permet de récupérer les messages échangés en temps réel sur la plateforme. Celle-ci est relativement complète puisqu'à chaque message récupéré sont associées les principales informations de son auteur. L'analyse en temps réel a nécessité la parallélisation des tâches. En moyenne, plus de dix messages par seconde peuvent être reçus en simultané. Nous avons donc mis en place un système de parallélisation de tâches. Ainsi, à chaque message reçu une tâche spécifique est créée et uniquement consacrée à l'extraction et à la génération des attributs associés à celui-ci.

3.3.1.2 Le modèle de la base de données

Le stockage des données est crucial, car il conditionne la facilité ou non d'accès à l'information lors de l'analyse. Le stockage des données se fait à l'aide d'une base de données de type MySQL. Ce choix a été effectué car, pour un grand amas de données, les bases de données de type *MySQL* ont un temps de réponse réduit (c'est le choix effectué par la plateforme Twitter). De plus, en cas de besoin, il est possible d'exporter la base ou certaines de ses tables sous de multiples formats.

L'entité fondamentale de notre analyse est l'individu. Comme nous l'avons déjà mentionné, sur Twitter, l'individu est identifié sur la plateforme par un pseudonyme et un identifiant numérique. L'identifiant considéré dans la base de données est l'identifiant numérique. Les principales informations conservées sur un profil sont : la localisation de l'individu, le nombre d'amis, le nombre de suiveurs, le nombre de messages publiés, sa langue, sa description, etc. Ces informations permettent la création d'attributs démarquant les différents individus en fonction de leur comportement.

Le modèle relationnel de la base de données est représenté sur la figure 3.3.3. L'entité principale est la personne (table *Person*) composée de dix-sept champs, dont la clé est l'identifiant noté Id. Cet identifiant est l'identifiant sur dix digits du profil d'un utilisateur sur le réseau Twitter. Le graphe social de Twitter est stocké grâce aux tables *Follows* et *Friends*, celles-ci représentent la relation de suiveur ou d'ami entre deux personnes d'identifiants *PersonId1* et *PersonId2*. Ces relations sont marquées de la date d'extraction des relations afin de pouvoir conserver la dynamique du réseau durant la collecte. Enfin, la table *Status* permet de stocker les *tweets*, ceux-ci sont identifiés de manière unique par un identifiant donné par la plateforme. De ces *tweets* sont extraits les références à des personnes, les mots-clés et les URL. Par souci de normalisation de la base, les tables *Status_Hastag*, *URL_Status* et *Status_Person* servent de liaison, et les tables *Person*, *Hashtag* et *URL* permettent de stocker les entités sans redondance.

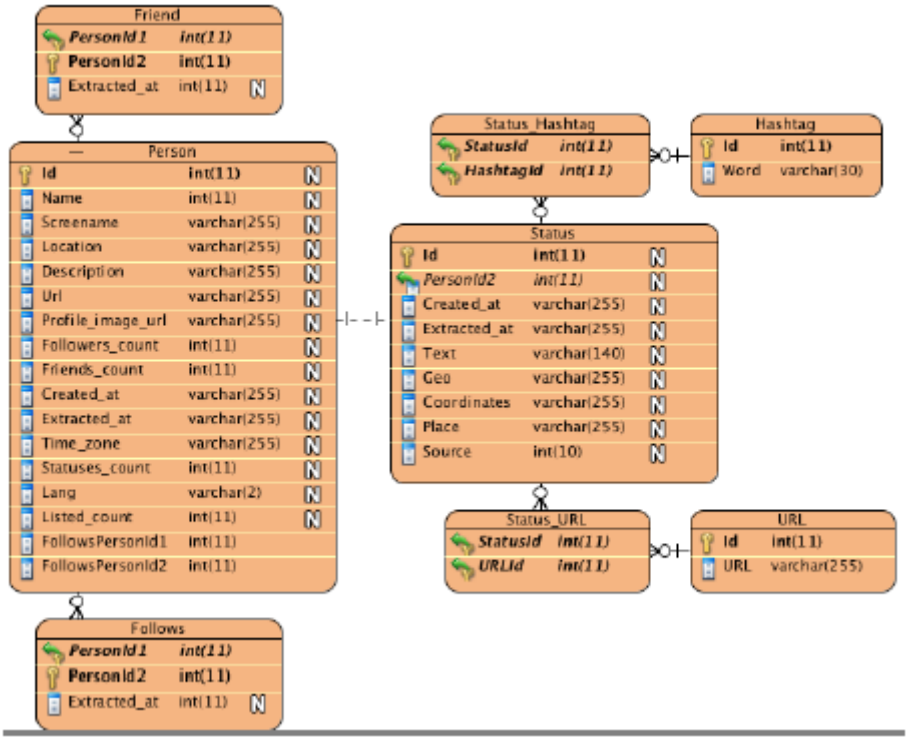


FIGURE 3.3.3 – Modèle relationnel de la base de données relative aux utilisateurs de Twitter.

3.3.2 La génération des attributs issus du profil et des messages

La génération d'attributs réalisée par le module II de la plateforme intègre les dimensions *profil* et *message*. Les attributs des valeurs ou agrégats permettent de modéliser un profil comme un vecteur nécessaire à la classification en phase III. Cette section présente l'approche nécessaire à la génération de différentes composantes du vecteur. Le tableau 3.1 illustre une partie des composantes du vecteur représentant trois utilisateurs de la plateforme Twitter.

Identifiant	**Âge**	**#Amis**	**#Suiveurs**	**Moyenne d'URL**	**...**
xxxxxx	630	4017	6077	0.98	...
yyyyyy	7	112	28	0.35	...
zzzzzzz	6	0	6	0.54	...

Tableau 3.1 – Exemple de caractéristiques générées pour quelques utilisateurs de la plateforme Twitter.

La quantité d'amis et de suiveurs
Les deux composantes les plus simples à générer sont les nombres de *followers* et *followees* qui sont directement récupérés sans modification depuis la base de données principale.

Les moyennes de références, hashtags et URL
Les moyennes d'URL, de référence et de mots-clés par *tweets* sont extraites à l'aide d'une requête avec jointure et agrégation sur la base de données principale.

L'âge du compte
L'âge du compte d'un profil est établi à partir de la date de création du profil et de la date de calcul de cet âge. Cet âge est calculé en jours et constitue la sixième composante du vecteur.

La fréquence de tweets
La fréquence de *tweets* par jour est calculée comme le quotient du nombre total de *tweets* produits par un individu, avec l'âge du compte calculé auparavant.

La fréquence de retweets
Un *retweet* est un message adressé en réponse à un message, cet attribut est notable car assez significatif d'un comportement humain. En effet, les profils automatisés ne répondent généralement pas aux messages qui leurs sont attribués. Un *retweet* est identifiable facilement puisqu'il doit être marqué par les caractères « RT : » en début de *tweets*. Afin de détecter les *retweets* parmi l'ensemble des *tweets* de la base de données, une expression régulière est utilisée comme filtre dans la requête SQL de sélection.

La distance de Levenshtein entre les tweets
La distance de Levenshtein entre deux chaînes de caractères est définie en fonction du nombre d'opérations basiques nécessaires pour passer d'une chaîne de caractères à une autre. La moyenne des distances est ensuite calculée pour former l'attribut du profil. Cette distance affecte une valeur comprise entre 0 et 1. Cette distance est nulle pour des chaînes identiques et s'approche de 1 pour des chaînes complètement différentes. La distance est calculée sur les *tweets* dont sont retirées les entités (c.-à-d. URLs, mots-clés, références), afin de déceler les robots dont seuls les références, les mots-clés et les URL changent au cours du temps (p. ex. figure 3.3.4).

3.3.3 La détection des profils suspects

La classification du module III consiste à isoler les profils suspects des profils non suspects. Afin de réaliser cette tâche, nous proposons l'utilisation d'une machine de classification par apprentissage. Ces machines ont pour but d'évaluer une ou plusieurs règles de classification à partir de cas déjà classés.

Afin de pouvoir réaliser la phase d'apprentissage, certains travaux proposent une classification manuelle de profils [123]. Dans ce travail, afin de gagner du temps, l'identification de profils suspects a été réalisée de façon pseudo-automatique. En effet, Twitter propose aux utilisateurs de marquer un individu comme *spammer* (voir figure 3.3.5). Un utilisateur peut envoyer un *tweet* comportant la référence @spam ou @spammer suivi des pseudonymes des utilisateurs repérés suspects (p. ex. @suspect).

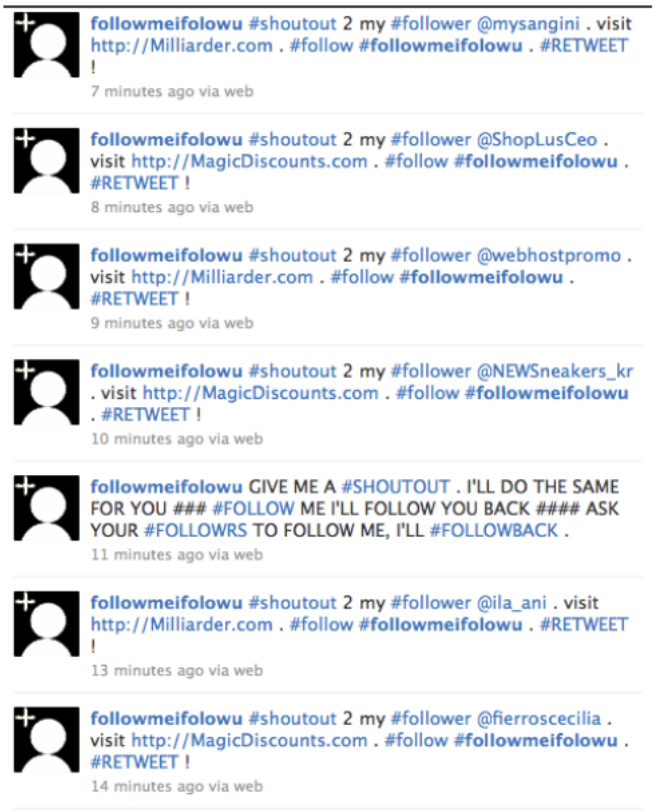


FIGURE 3.3.4 – Exemple de messages produits par un profil suspect.

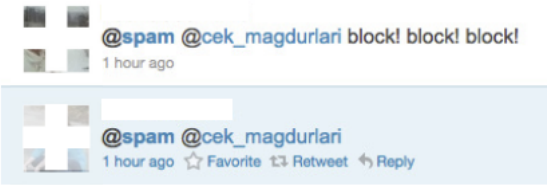


FIGURE 3.3.5 – Signalement d'un profil perturbateur par deux utilisateurs de Twitter.

Les profils @spam, @spammer sont gérés par Twitter pour lutter contre la malveillance. Nous avons utilisé l'outil de recherche de *tweets* via l'API *Search* pour récupérer les messages contenant la mention @spam ou @spammers ; de ces *tweets,* nous avons extrait les mentions aux profils suspects.

En ce qui concerne l'identification de profils non suspects nécessaire à notre ensemble d'apprentissage, la piste investiguée est la *vérification* de profils. En effet, sous la demande d'un utilisateur, Twitter peut vérifier si un profil correspond à une personne réelle. Ce processus de vérification proposé par Twitter assure aux utilisateurs qu'un profil est non malveillant. Un utilisateur vérifié est repérable à l'aide d'un label apparaissant proche de son pseudonyme sur sa page Twitter. Nous pouvons aussi, grâce au champ *is_verified* de son profil XML, savoir si l'utilisateur est certifié ou non. Pour cette étape, nous avons parcouru notre base de données de collecte à la recherche de profil possédant le label *is_verified*.

Au total, nous avons identifié 250 profils suspects et 250 profils certifiés (vérifiés manuellement par un expert). Une fois les profils suspects et non suspects repérés et classifiés, une phase d'écoute de leur activité a été mise en place. Un flux sur les identifiants a été ouvert, nous permettant de récolter les messages en temps réel et de générer les attributs nécessaires à l'apprentissage.

3.3.4 La génération des attributs des URL

Pour l'analyse des URL et la génération des attributs de ces URL, nous avons mis en place une seconde base de données spécifique. Son modèle relationnel est décrit dans la figure 3.3.6. Celle-ci comporte cinq tables, la table principale étant la table URL contenant les caractéristiques d'une URL, deux tables de jointure *URL_mining* et *URL_domaine* qui permettent d'éviter les doublons dans les tables *Domain* et *Token*. La table *Domain* contient les noms de domaines et les dates de création, de mise à jour et d'expirations associées. La table *Token* contient des lemmes extraits du site Web.

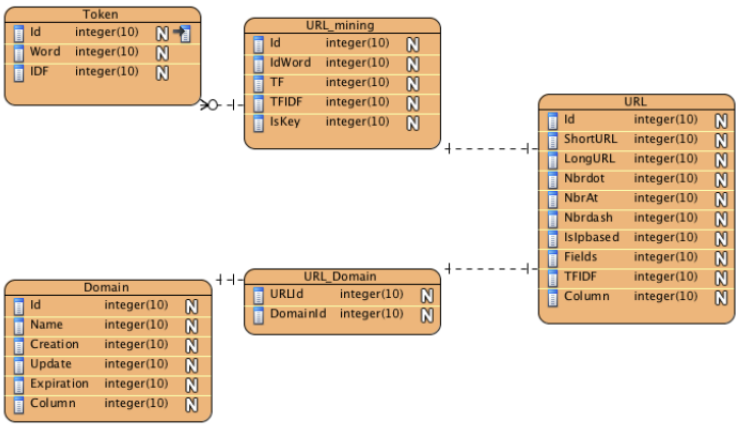


FIGURE 3.3.6 – Modèle relationnel de la base de données de stockage des composantes des URL.

3.3.4.1 L'âge du domaine

L'âge d'un domaine est récupéré automatiquement grâce à un appel au service proposé par Internic [1]. Celui-ci est capable de nous fournir l'âge des noms de domaines de nombreux premiers niveaux tels que : .com, .edu, .net, .org, etc. Ainsi, pour ne pas effectuer de requêtes inutiles, les URL sont vérifiées à l'aide d'une expression régulière et seules les URL comportant les noms de domaines des premiers niveaux précédents sont transmises au service. Ce service permet d'obtenir les dates de création, de mise à jour et d'expiration, les trois sont stockées dans notre base, mais pour l'instant, seule la date de création est utilisée par SPOT 1.0.

3.3.4.2 Le score TF-IDF

La technique TF-IDF est une méthode de fouille de données basée sur l'occurrence de lemmes dans un document par rapport à un corpus de textes. Ici, le texte analysé est le contenu de la page Web, et le corpus est l'ensemble des documents du Web. Pour obtenir le contenu de la page Web, nous récupérons la page HTML et en extrayons le contenu textuel à l'aide de la librairie Jsoup [2]. Le texte est ensuite nettoyé de ses articles et les occurrences de chacun des mots restants sont calculées pour en extraire la fréquence. Ce calcul statistique TF (Text Frequency) [127] est défini comme suit :

$$TF_{i,j} = \frac{N_{i,j}}{\sum_k N_{k,j}} \tag{3.3.1}$$

Avec :

$TF_{i,j}$ la fréquence d'occurrence du lemme i dans la page j
$N_{i,j}$ le nombre d'occurrences du lemme i dans la page j
$\sum_k N_{k,j}$ le nombre total d'occurrences des lemmes dans la page j

Ce calcul est pondéré d'un coefficient IDF (Inverse Document Frequency), qui représente l'occurrence d'un lemme dans un corpus de texte. Ici, notre corpus est le Web entier, et pour mesurer l'occurrence d'un lemme sur le Web, l'API du moteur de recherche Yahoo [3] a été utilisé et le nombre de résultats à la recherche du lemme a été enregistré. La taille du Web (en termes de nombre de documents) a été approximée au nombre de documents indexés par Yahoo.

La formule pour obtenir le score IDF est la suivante :

$$IDF_i = log(\frac{T}{|\{l_i \in t,\, t \in T\}|}) \tag{3.3.2}$$

Avec :

T le nombre de documents du Web ($\approx 10\, Milliards$)
$|\{l_i \in t,\, t \in T\}|$ le nombre de documents contenant le lemme i

Le score calculé est significatif de l'occurrence du lemme dans le texte (équation 3.3.1) mais aussi de son occurrence dans le reste des documents (équation 3.3.2). L'indicateur

1. `http://www.internic.net/whois.html`
2. `http://jsoup.org/`
3. `http://developer.yahoo.com/search/`

TF-IDF d'un lemme i dans un page Web j est défini comme le produit des deux valeurs précédentes :

$$TF - IDF_{i,j} = TF_{i,j} * IDF_i \tag{3.3.3}$$

Un lemme présent avec une forte occurrence dans un document mais une faible occurrence dans le corpus aura un fort score TF-IDF. Ce lemme est donc représentatif du document dans le corpus ; pour cette raison, il est généralement sélectionné pour construire une signature lexicale du document. Dans notre cas, il s'agit de construire une signature lexicale d'une page Web. Le travail [125] a mis en évidence qu'une signature convenable pour un document du Web est constituée de cinq lemmes, et celui-ci s'en est servi pour proposer le concept de liens robustes. Lorsque la signature lexicale d'un site a été trouvée, celle-ci est utilisée pour vérifier sa légitimité. La signature lexicale de l'URL est envoyée comme paramètre de requête au moteur de recherche Yahoo. Ensuite, la légitimité du site est évaluée à partir du nombre de réponses dans le TOP 20 comportant le même nom de domaine que l'URL analysée. Nous créons donc un indicateur basé sur la méthode TF-IDF.

3.3.5 La classification des URL issues des profils suspects

Le module V consiste à classifier les URL comme suspectes ou non suspectes. Cette classification se base sur une phase d'apprentissage à partir d'exemples de référence. Les exemples d'URL suspectes ont été récoltés à partir de la base de données de sites malveillants Phishtank [4], et les 250 résultats les plus récents ont été extraits et analysés pour générer les caractéristiques nécessaires. En ce qui concerne les URL non suspectes, c'est à partir de la liste des sites Web les plus visités du monde qu'ils ont été sélectionnés. Le site Web *mostpopularWebsites* [5] a été exploré (via un *Webcrawler*) sur plusieurs pages afin d'extraire cette liste. Au total, 250 URL des sites les plus visités ont été analysées. L'algorithme de classification s'est entraîné à partir des 500 URL ainsi classifiées.

Nous proposons de mettre en évidence les résultats de notre analyse comportementale des profils en affectant un niveau de virulence à chacun. Pour cela nous présentons dans la suite de ce chapitre un indicateur tridimensionnel qui prend en compte le niveau d'activité, de visibilité et de danger associé à chaque profil évalué. Cet indicateur permet d'obtenir une vision globale de la malveillance des profils sur la plateforme Twitter et peut aussi faire office d'outil d'aide à la décision.

3.3.6 L'indicateur tridimensionnel : *Activité, Visibilité, Danger*

Afin de caractériser le niveau de virulence associé à un profil, et potentiellement à une attaque effectuée par ce profil, nous proposons trois dimensions. Les trois dimensions sont : (1) le caractère malveillant des URL, (2) la visibilité plus ou moins grande des *tweets* générés, (3) le comportement plus ou moins actif sur la plateforme. Lorsque l'on considère une attaque par propagation de contenu malveillant, ces dimensions permettent de mettre en évidence l'audience potentielle de l'attaque et le taux de répétition de l'attaque. Nous détaillons dans la suite la manière d'évaluer ces trois dimensions.

4. `http://www.phishtank.com/`
5. `http://mostpopularWebsites.net/`

3.3.6.1 L'évaluation de la menace par les URL

La dimension danger est calculée à l'aide du nombre d'URL dangereuses contenues dans les messages de l'individu suspect (voir définition 3.1). Plus le rapport entre le nombre d'URL envoyées et le nombre d'URL détectées suspectes est grand, et plus le danger est important pour un individu accédant aux *tweets* de l'utilisateur.

Définition 3.1. Mesure de danger
Nous proposons de définir l'indicateur de danger pour un profil dénoté p comme suit :

$$\mathcal{D}p = \frac{|U_m^p|}{|Up|} \tag{3.3.4}$$

Avec :

U_m l'ensemble des URL malveillantes diffusées par un profil dénoté p pendant la durée de collecte

U l'ensemble des URL diffusées par p pendant la durée de collecte

Ce ratio peut être perçu comme la probabilité, pour un utilisateur de la plateforme, de se retrouver sur un site malveillant lors de la consultation d'une URL envoyée par le profil analysé. Si la valeur $\mathcal{D}p$ est maximale, alors un utilisateur consultant une URL issue des messages du profil p est certain de tomber sur un site malveillant. Si $\mathcal{D}p$ est nul, il n'y a aucun risque lié à la consultation du lien. Par construction, l'ensemble des valeurs possibles de cet indicateur sont contenues dans l'intervalle $[0, 1]$.

3.3.6.2 La mesure de l'activité d'un profil

La mesure du comportement en termes d'activité se fait selon deux variables, (1) le nombre de *tweets* envoyés par heure (f_{tweet}) et (2) le nombre de *followees* ajoutés par heure ($f_{followees}$) tel qu'indiqué dans la définition 3.2. Un utilisateur malveillant va naturellement augmenter les chances de réussir son action en répétant l'attaque plusieurs fois. Ainsi, la fréquence de *tweets* révèle le nombre d'attaques effectuées par heure. L'autre paramètre important est sa tendance à vouloir suivre des individus sur le réseau. En suivant un maximum d'individus, l'attaquant augmente le nombre de victimes directes (en augmentant le nombre de personnes qui le suivront en retour), mais il peut également se constituer aussi une liste de contacts qu'il pourra définir comme cibles. Twitter limitant ce type d'opérations automatiques à 150 par heure, un attaquant exploitant au maximum le réseau répartira l'envoi des *tweets* et l'ajout de contacts de sorte à atteindre ce maximum.

Définition 3.2. Mesure d'activité
Nous proposons de définir l'activité d'un utilisateur p sur Twitter comme suit :

$$\mathcal{A}p = \frac{N_{tweets} + N_{followees}}{150} \tag{3.3.5}$$

Avec :

N_{tweets} le nombre de *tweets* envoyés par heure

$N_{followees}$ le nombre de *followees* ajoutés par heure

L'ensemble des valeurs possibles de cet indicateur sont contenues dans l'intervalle $[0, 1]$, une valeur proche de l'unité est significative d'un profil très actif, tandis qu'une valeur nulle est significative d'un profil inactif.

3.3.6.3 L'indicateur de visibilité

La mesure de la visibilité des *tweets* d'un attaquant est l'un des paramètres significatifs reflétant le nombre de potentielles victimes de son attaque. Les trois aspects retenus dans la définition 3.3 sont les mots-clés, les URL et les références (voir la Figure 3.3.6). Nous partons du principe qu'une attaque optimale portée par les *tweets* doit utiliser au mieux les 140 caractères pour rendre visible au maximum d'utilisateurs une ou plusieurs URL souhaitées (URL). Pour ce faire, il doit contenir des URL, des mots-clés et des références. L'indicateur de visibilité indique à quel point les trois dimensions ont été utilisées dans les messages de l'attaquant.

Faire n références à des utilisateurs coûte en moyenne $n*10,6$ signes, ajouter n *hashtags* coûte en moyenne $n*10,3$ signes, et ajouter n URL coûte en moyenne $n*15$ signes [6].

Définition 3.3. Mesure de visibilité
La visibilité d'un profil p est définie par :

$$\mathcal{V}p = \frac{Avg(@)*10.6 + Avg(\#)*10.3 + Avg(URLs)*15.0}{140} \tag{3.3.6}$$

L'ensemble des valeurs possibles de cet indicateur sont contenues dans l'intervalle $[0, 1]$, une valeur proche de l'unité est significative de *tweets* optimisés pour la diffusion, tandis qu'une valeur proche de 0 est significative de *tweets* peu visibles.

3.4 Prototype SPOT 1.0

SPOT 1.0 est une application développée en langage Java à l'aide de l'IDE Netbeans. Celle-ci permet de lancer la collecte, le traitement et l'analyse des profils de la plateforme Twitter en temps réel. La représentation des profils dans le repère tridimensionnel est l'élément clé de l'interface, qui permet de mettre en évidence les profils les plus virulents. Pour permettre la prise de décision, un ensemble de caractéristiques comportementales de chaque profil souhaité peut être affiché et la consultation du profil sur la plateforme est aussi rendue possible par l'application. Nous présentons dans la suite les deux principaux écrans du logiciel : l'interface d'activation et de désactivation des modules et l'interface représentant les profils dans le repère tridimensionnel.

3.4.1 L'écran d'accueil

La figure 3.4.1 présente la fenêtre d'accueil du logiciel SPOT. Cette fenêtre permet de visualiser trois actions fondamentales de SPOT : (1) la collecte de *tweets* et des profils associés, (2) la génération de caractéristiques de profils les plus actifs, (3) la détection des profils suspects.

6. Ces valeurs ont été évaluées sur la base des 2 Millions de tweets collectés

— Le premier module affiche en temps réel la quantité de *tweets* collectés, ce module peut être activé et désactivé par le biais du bouton *« track »*. Un ensemble de données concernant la collecte sont affichées pour permettre au gestionnaire de suivre l'évolution de la collecte. Les données affichées sont : le nombre de *tweets* collectés, le nombre de profils collectés, le nombre de mots-clés collectés, de références collectées et le nombre d'URL collectées.

— Le second module permet d'activer et de désactiver la génération des caractéristiques des profils collectés. Celui-ci affiche en temps réel le nombre de profils analysés ainsi que leurs caractéristiques, telles que la moyenne de *tweets* envoyés, l'âge moyen et la distance moyenne entre les *tweets*.

— Le dernier module active la détection de profils suspects et affiche les caractéristiques de ceux-ci ainsi que leur quantité détectée sur la plateforme.

Depuis cette fenêtre principale, il est possible de visualiser les profils dans notre repère tridimensionnel permettant l'aide à la décision.

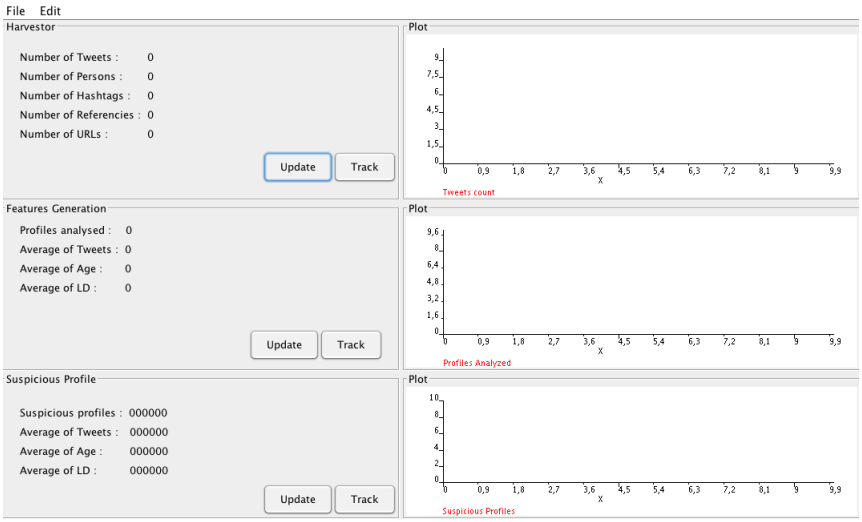


FIGURE 3.4.1 – Fenêtre principale de l'outil SPOT.

3.4.2 La représentation tridimensionnelle

La figure 3.4.2 présente les profils dans le repère tridimensionnel $\mathcal{A}ctivité$, $\mathcal{V}isibilité$, $\mathcal{D}anger$. Il est possible d'interagir avec la représentation pour faire pivoter les axes et sélectionner un ensemble de profils à investiguer. Sur la figure, cet ensemble est représenté par un rectangle de sélection que l'utilisateur peut créer à volonté. Lorsque la sélection est effectuée, l'onglet de droite est mis à jour pour représenter les caractéristiques des profils choisis.

La figure 3.4.3 représente un exemples de sélection de profils. Le tableau présente pour chaque profil l'identifiant de celui-ci sur la plateforme, son pseudonyme, l'URL de son profil, la date de création du profil et ses coordonnées dans le repère tridimensionnel. Sur

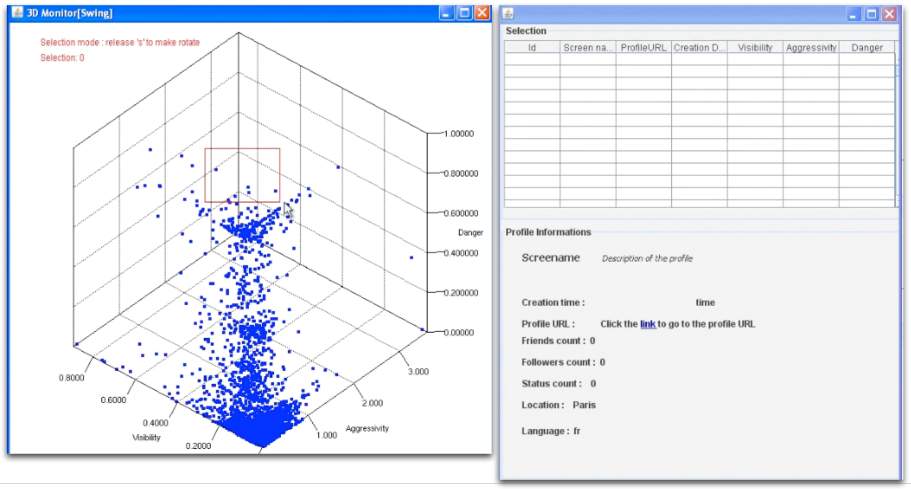


FIGURE 3.4.2 – Repère tridimensionnel avec un rectangle de sélection des profils représentés en rouge.

l'exemple de la figure, seuls des profils de danger maximal ont été sélectionnés. Chaque profil peut être individuellement analysé en sélectionnant sa position dans le tableau. L'investigation peut se terminer par la consultation de son profil directement depuis la plateforme.

3.5 Résultats

La figure 3.5.1 représente les profils analysés selon les trois dimensions de l'analyse. Un profil correspond à un point de coordonnée $p(\mathcal{V}, \mathcal{A}, \mathcal{D})$ défini par l'indicateur. La couleur d'un profil indique s'il a été détecté suspect ou non après la classification de premier niveau de SPOT 1.0. Un profil suspect est en rouge (gris clair en N&B) tandis qu'un profil non suspect est en bleu (gris foncé en N&B). Les profils suspects (en rouge) sont les seuls dont l'analyse est étendue sur une dimension supplémentaire (c.-à-d. les URL). Ainsi, les profils non suspects sont situés dans le plan ($\mathcal{V}isibilité$, $\mathcal{A}ctivité$). La représentation en trois dimensions des profils permet de mettre en évidence les profils les plus dangereux. Ceux-ci se situent près du point $I(1, 1, 1)$ tandis que les profils non dangereux se situent près de l'origine $O(0, 0, 0)$. La représentation de la figure a été réalisée après seulement deux jours d'analyse. Celle-ci met en évidence le comportement actif des profils suspects ainsi que leur visibilité au-dessus de la moyenne. On observe aussi que le danger est bien présent et n'est pas négligeable sur le réseau. Environ 100 des 6 000 profils collectés présentent un risque pour les utilisateurs. Certains profils non suspects se démarquent des autres profils et ont une position relativement stratégique en cas d'attaque.

3.5.1 Évaluation de la classification

La qualité de la méthode de classification se mesure à l'aide de quatre variables de base. Le nombre de cas positifs correctement classés (a), le nombre de cas positifs mal classés (b), le nombre de cas négatifs mal classés (c) et le nombre de cas négatifs correctement classés (d). Ces variables sont présentées dans la matrice de confusion du tableau 3.2.

Les trois indicateurs choisis pour évaluer l'efficacité de la méthode de classification sont la précision, l'exactitude et le rappel. La précision qui reflète le pourcentage de profils

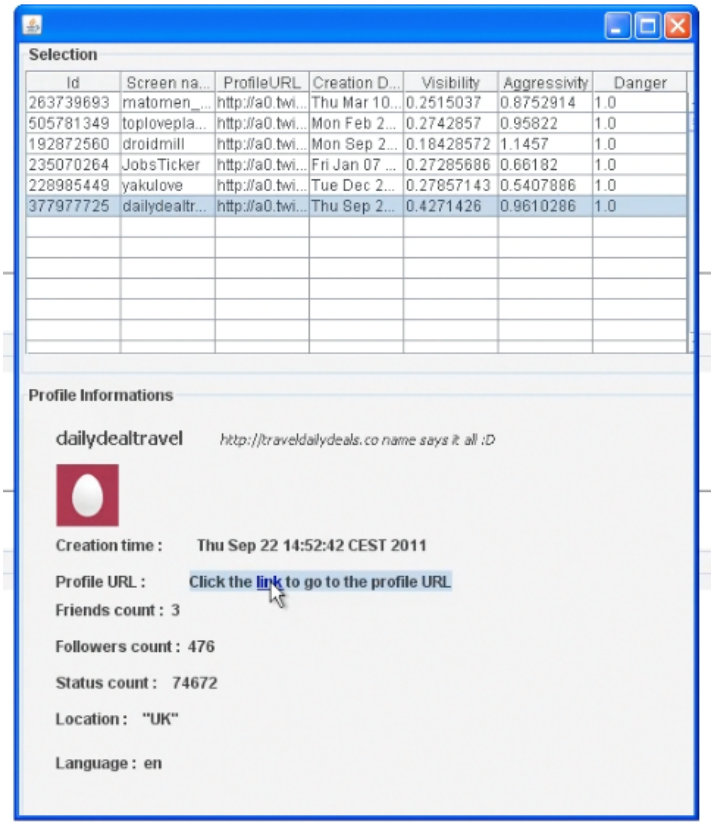


FIGURE 3.4.3 – Affichage des informations relatives à un profil.

	Prédiction positive	**Prédiction négative**
Classe positive	a	b
Classe négative	c	d

Tableau 3.2 – Matrice de confusion d'une technique d'apprentissage supervisé.

correctement classifiés est mathématiquement définie comme suit :

$$Précision = \frac{a+d}{a+b+c+d} \tag{3.5.1}$$

L'exactitude, qui reflète le pourcentage de profils identifiés correctement comme suspects, est définie comme suit :

$$Exactitude = \frac{a}{a+c} \tag{3.5.2}$$

Enfin, le rappel, qui est le pourcentage de profils suspects correctement identifiés, est défini comme suit :

$$Rappel = \frac{a}{a+b} \tag{3.5.3}$$

Notons que les indicateurs de performance basés sur la précision ne sont pas nécessairement les seules options, surtout quand la notion de sécurité entre en jeu. Il est possible d'utiliser d'autres indicateurs ou simplement des poids pour chaque classe afin d'identifier qu'un profil malveillant non détecté est plus critique qu'un profil normal détecté comme

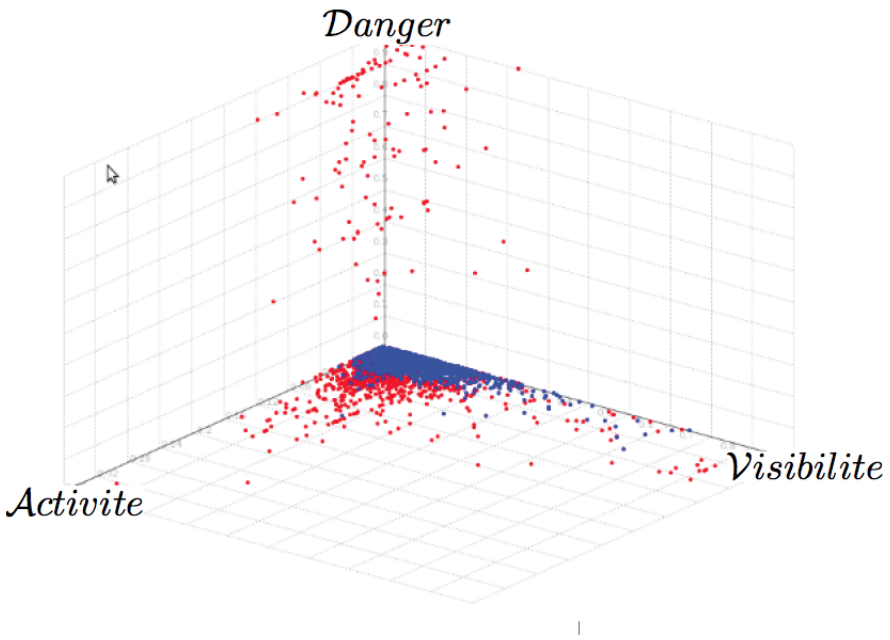


Figure 3.5.1 : Représentation des profils dans le repère tridimensionnel.

malveillant. Dans ce travail, nous avons simplement utilisé la manière la plus traditionnelle d'évaluer les performances, mais cette adaptation peut être aussi prise en compte.

3.5.1.1 La classification des profils

La classification du module III a été testée sur un ensemble de 500 profils suspects et non suspects. Le processus de validation croisée est constitué de dix groupes et les résultats sont les suivants. La *précision* globale est de 79 %, cela signifie qu'en moyenne quatre profils sur cinq sont correctement classifiés. L'*exactitude* ou ratio de profils identifiés suspects à juste titre est de 83 %. Cela signifie que cinq profils sur six sont identifiés suspects à juste titre et seulement un sur six est un faux positif. Le *rappel* est de 73 %, cela signifie que plus de sept profils suspects sur dix sont détectés.

3.5.1.2 La classification des URL

La classification du module V est celle des URL. Cette classification a été réalisée à partir de 500 URL suspectes et non suspectes. Là encore, dix groupes ont été générés aléatoirement pour effectuer une validation croisée. Le ratio d'URL correctement classifiées est de 76 %. L'*exactitude* est de 70 %, ce qui correspond au nombre d'URL classées suspectes à juste titre. Enfin, 90 % des URL suspectes sont détectées, ce qui signifie que seulement une URL suspecte sur dix n'est pas détectée.

3.5.1.3 Quelques remarques sur les performances

Les performances indiquées ci-dessous bien qu'acceptables ne sont pas idéales car certains profils suspects peuvent de ne pas être détectés par notre outil.

3.5.2 La répartition des classes sur les paramètres

La classification effectuée par SPOT 1.0 permet *a posteriori* d'évaluer l'importance de chacune des variables utilisées lors de la classification. Ainsi, nous avons représenté les graphiques correspondant à chacune des variables en marquant en bleu la répartition des profils non suspects et en rouge la répartition des profils suspects. Les résultats sont présentés dans la suite et permettent de mettre en évidence la pertinence de chacune des variables et leur rôle plus ou moins discriminant. Les graphiques ont été réalisés sur la base de 12 000 profils classifiés par SPOT 1.0.

3.5.2.1 Le nombre de suiveurs

La figure 3.5.2 met en évidence une forte répartition de profils suspects avec peu de *followers*. Au contraire, elle met en avant une forte répartition de profils non suspects autour de la valeur 100. Ainsi, nous observons qu'un profil suspect attire beaucoup moins de *followers* que les profils non suspects. Malgré leur volonté de créer des relations sur le réseau, la majorité des utilisateurs ne suivent pas de profils suspects. Le reste de la figure ne permet pas d'identifier de réelles distinctions supplémentaires entre les deux classes de profil.

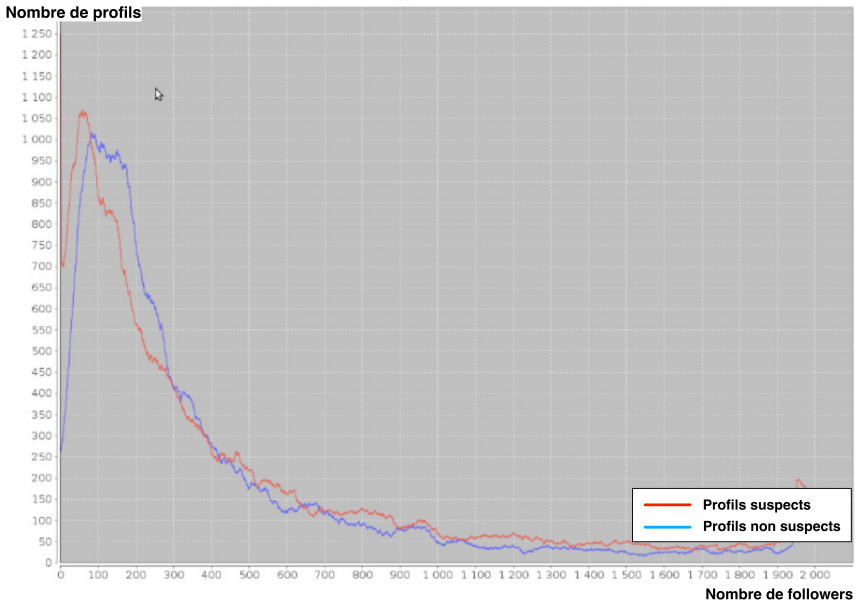


Figure 3.5.2 : Répartition des profils en fonction du nombre de *followers*.

3.5.2.2 Le nombre d'amis

La figure 3.5.3 permet de distinguer un nombre de *followees* plus faible pour une majorité de profils suspects que pour les profils non suspects. Cependant, au-delà de 700 suiveurs, les profils suspects sont majoritaires, et ce, jusqu'à la valeur caractéristique 2 000. Autour de la valeur 2 000 un nombre important de profils suspects est observé. Cela signifie que la majorité des profils de Twitter dont le nombre de *followers* n'est pas égal à 2 000 mais dont le nombre de *followees* est égal a 2 000 sont suspects. Cela met en évidence l'action menée par Twitter pour obliger un contact à obtenir 2 000 *followers* avant de pouvoir suivre plus de 2 000 profils.

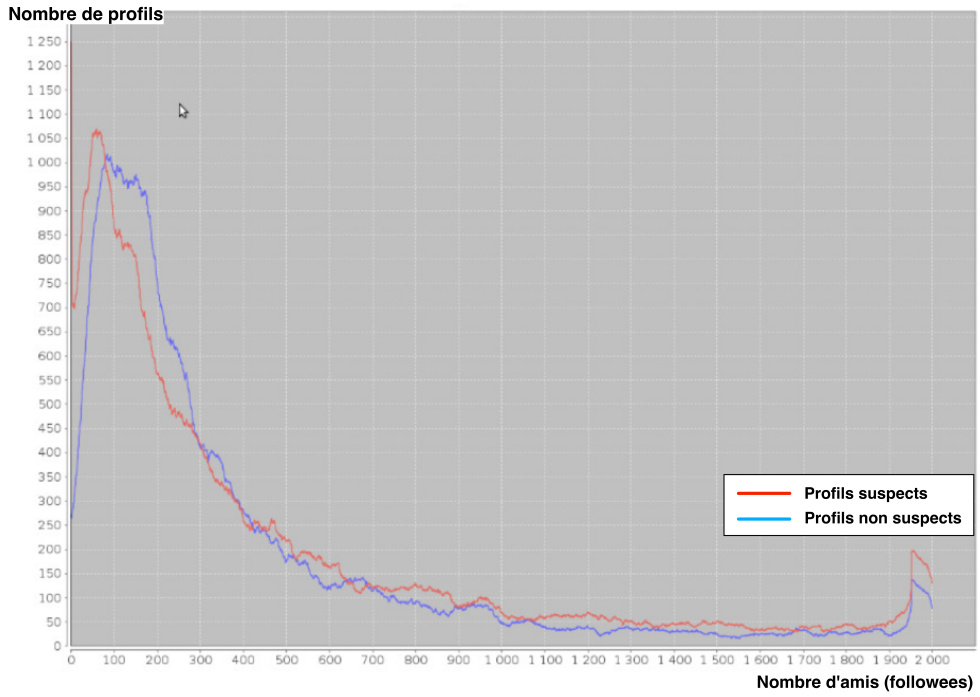


Figure 3.5.3 : Répartition des profils en fonction du nombre de *followees*.

3.5.2.3 La fréquence de tweets par jour

L'observation de la répartition des profils par rapport à la fréquence de *tweets* confirme notre intuition de départ. La figure 3.5.4 met en évidence une faible fréquence quotidienne de *tweets* des profils non suspects, majoritairement inférieure à 75, tandis que la répartition des profils suspects est plus étalée, jusqu'à des valeurs proches de 250.

3.5.2.4 La distance de Levenshtein entre les *tweets*

La représentation en figure 3.5.5 met en évidence le fait qu'une partie relativement importante des profils suspects génère des messages quasi similaires (distance de Levenshtein proche de 1). Les profils non suspects quant à eux produisent des messages plus variés et très peu de messages de contenu similaire.

3.5.2.5 La réputation

La répartition de la réputation des profils est assez proche pour les profils suspects et les profils non suspects. La figure 3.5.6 révèle cependant que les profils avec une forte réputation sont majoritairement classés comme suspects.

3.5.2.6 L'âge des profils

La figure 3.5.7 représente la répartition de l'âge des profils suspects et des profils non suspects. La figure permet de mettre en évidence l'âge relativement faible des profils suspects et l'âge relativement plus âgé de la majorité des autres profils. Cette observation confirme le fait que Twitter détecte une grande partie de profils suspects relativement rapidement. Cependant, certains profils suspects persistent dans le temps jusqu'à presque

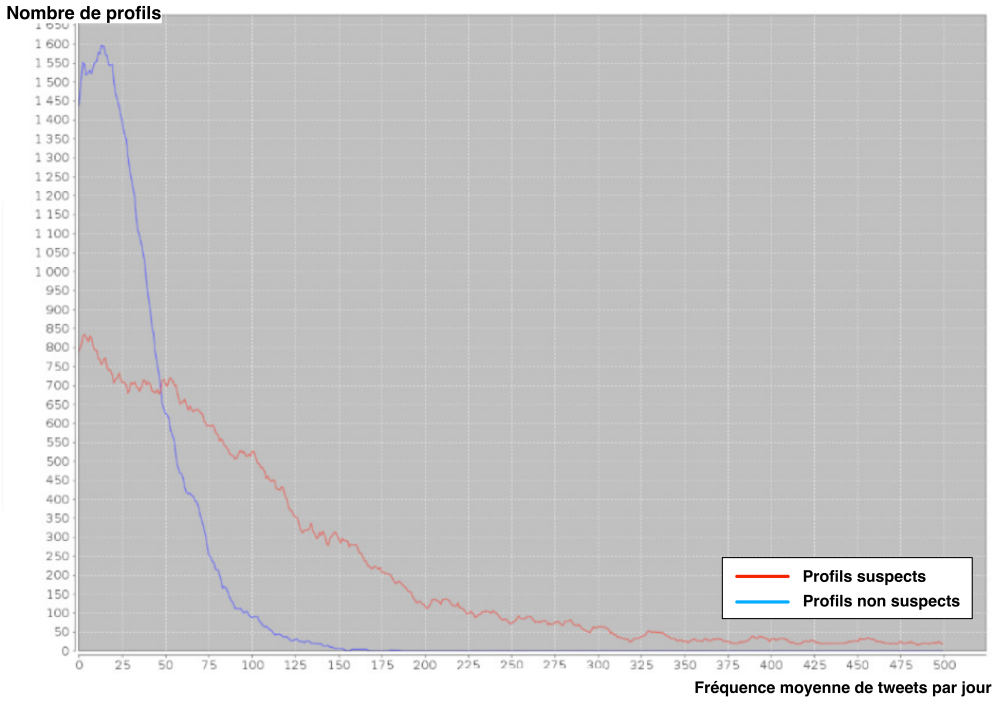


Figure 3.5.4 : Répartition des profils en fonction de la réputation.

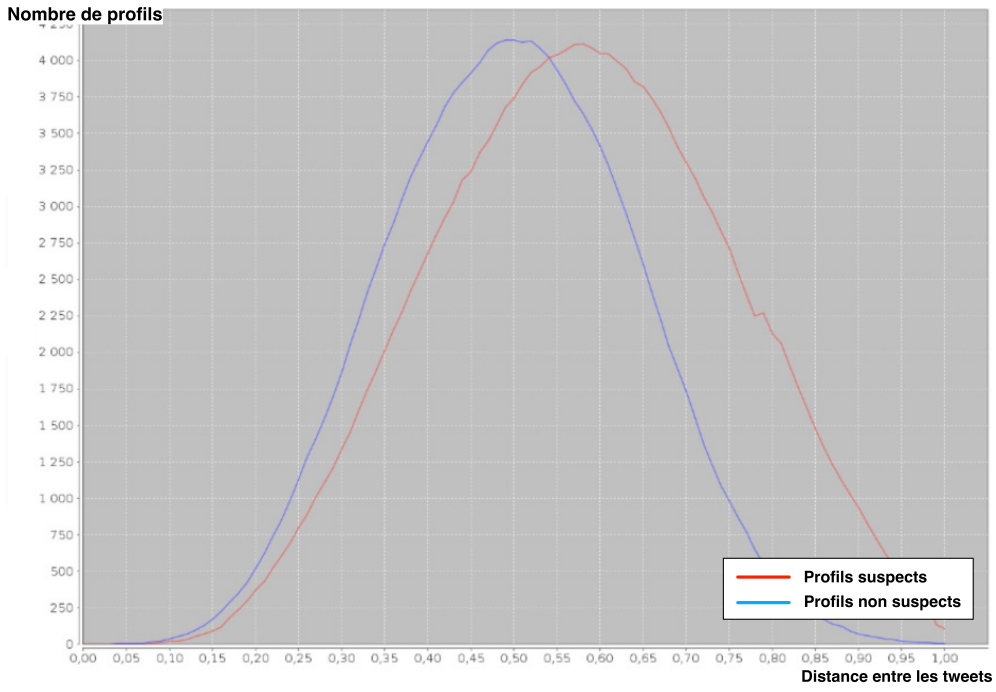


Figure 3.5.5 : Répartition des profils en fonction de la distance entre les *tweets*.

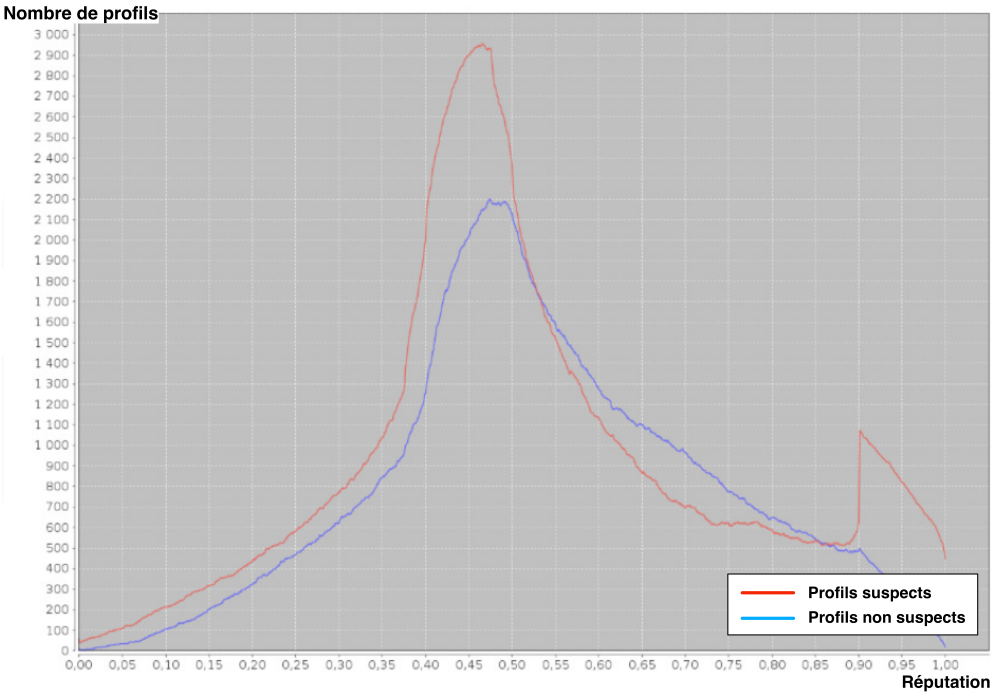


Figure 3.5.6 : Répartition des profils en fonction de la réputation.

deux ans. La représentation indique donc que l'âge est un des paramètres discriminants nécessaires à prendre en compte lors de la classification.

3.5.2.7 Résumé des observations et discussion

Le tableau 3.3 représente les valeurs moyennes des dix indicateurs choisis pour la classification des profils sur 12 000 profils issus de la plateforme Twitter. Les profils suspects ont une moyenne d'âge beaucoup plus faible que les profils non suspects de la plateforme. Les profils suspects ont en moyenne plus de *followers* et moins de *followees* que les profils non suspects. Cette observation peut venir du fait qu'un profil contrôle le nombre de *followees* mais non le nombre de *followers*. Il est simple de se créer un nombre important de *followees* et il est beaucoup plus compliqué de forcer les profils à nous suivre. Les profils suspects publient en moyenne davantage d'URL que les autres profils. Ceci peut être expliqué par la volonté des acteurs malveillants à emmener leurs victimes vers un site malveillant. Il s'agit de l'un des principaux vecteurs d'action possibles. Le nombre de références à des personnes dans un *tweet* ne paraît pas être un facteur discriminant. Les mots-clés par contre sont davantage utilisés par les profils suspects, notamment pour augmenter la légitimité d'un message et augmenter sa visibilité. Les profils suspects ne répondent quasiment pas aux messages leur étant destinés, au contraire des individus non suspects qui répondent à 20 % de ceux-ci. La fréquence de *tweets* envoyés par les profils suspects est en moyenne cinq fois plus grande que celle des profils non suspects. Les profils suspects produisent des messages plus similaires que les autres utilisateurs. La réputation ne semble pas être un facteur discriminant de la classification.

Nous proposons de calculer l'écart-type de chacune des variables observées pour les deux types de profils. Cette mesure met en évidence l'écart mesuré entre chaque observation individuelle et la moyenne de ces observations.

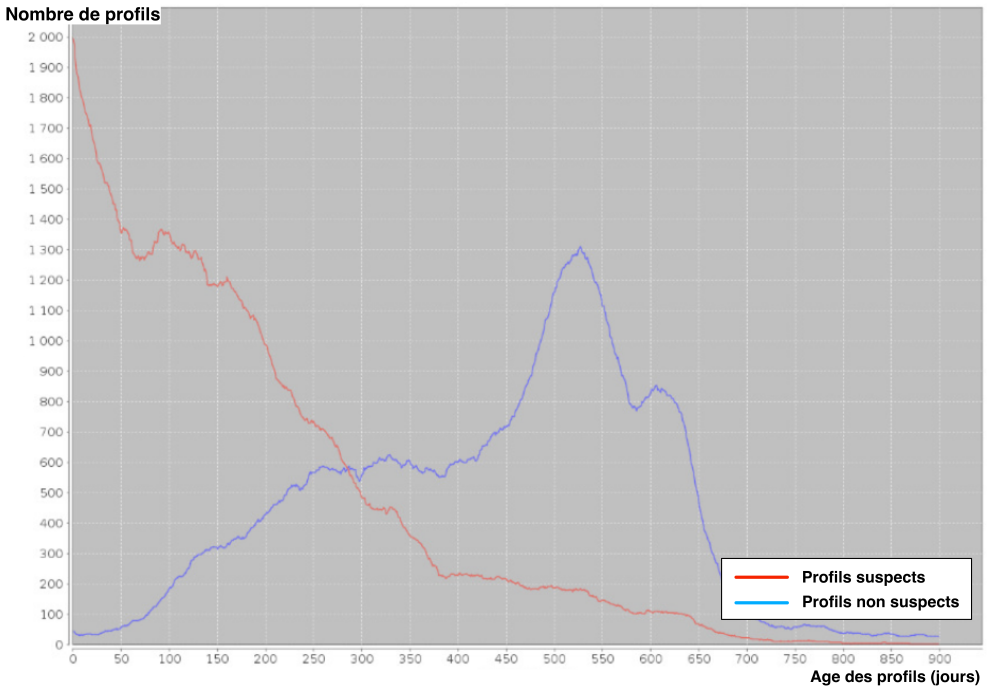


Figure 3.5.7 : Répartition des profils en fonction de l'âge des profils.

Indicateur	**Suspect**	**Normal**
Âge du profil	190	**465**
Nombre d'amis	**941**	753
Nombre de *followers*	1 600	**1 724**
Moyenne d'URL par *tweet*	**0.22**	0.07
Moyenne de références par *tweet*	0,80	**0.84**
Moyenne de *hashtags* par *tweet*	**0.25**	0,14
Ratio de *retweets*	0.09	**0.2**
Fréquence de *tweets* par jour	**131**	35
Distance entre les *tweets*	**0.68**	0.59
Réputation	0.58	**0.59**

Tableau 3.3 – Comparaison des valeurs moyennes des indicateurs par classe. Les valeurs maximales sont représentées en gras.

Le tableau 3.4 met en évidence le fait qu'un comportement normal est plus facilement identifiable qu'un comportement suspect. Ainsi, les écarts-types sont plus faibles pour les profils non suspects, et les profils suspects sont plus hétérogènes en moyenne. Cela illustre la diversité des techniques d'attaque sur la plateforme. Seuls deux indicateurs - l'âge et la fréquence de *retweets* - ont un écart-type faible. Ainsi, le jeune âge et la très faible quantité de *retweets* sont les indicateurs les mieux partagés par les profils suspects.

Indicateur	**Suspect**	**Normal**
Âge du profil	164	**175**
Nombre d'amis	**3 642**	2 679
Nombre de *followers*	**9 562**	9 000
Moyenne d'URL par *tweet*	**0.38**	0.16
Moyenne de références par *tweet*	**0.75**	0.58
Moyenne de *hashtags* par *tweet*	**0.76**	0.27
Ratio de *retweets*	0.19	**0.25**
Fréquence de *tweets* par jour	**151**	26
Distance entre les *tweets*	**0.18**	0.16
Réputation	**0.21**	0.18

Tableau 3.4 – Comparaison des écart-types des indicateurs par classe. Les valeurs maximales sont représentées en gras.

3.6 REPLOT : Détection de campagnes de profils malveillants

De la simple utilisation d'un profil unique pour exécuter des attaques, les acteurs malveillants sont maintenant passés à une façon plus collective et synchronisée pour entreprendre des actions malveillantes. Cela leur permet d'accroître l'impact de leur attaque en augmentant artificiellement leur réputation et de la même manière le nombre total de victimes (profils ciblés). Par conséquent, la détection de profils malveillants est souvent insuffisante pour éliminer les campagnes malveillantes et une caractérisation d'entre eux est nécessaire. Pour cela nous proposons une extension de l'approche SPOT pour intégrer cette nouvelle mouvance de la malveillance.

La méthodologie appelée REPLOT contient trois grandes phases : (1) les profils individuels sont analysés pour déterminer s'ils sont suspects ou non ; (2) les connexions entre les profils suspects sont identifiées en utilisant une approche de fusion de données, de modèles temporels et d'analyse de caractérisation d'auteurs ; (3) une technique de groupement de données est utilisée pour profiler les différentes campagnes malveillantes. L'identification des campagnes est réalisée via la création d'un graphe fondé sur les indicateurs analysés en deuxième phase en vue de trouver des sections du graphe avec une similitude interne élevée.

Par cette méthode, les profils malveillants ne sont pas seulement découverts automatiquement, mais également profilés. Pour un analyste ayant besoin de découvrir les tendances et les modèles d'attaque des profils, ce niveau d'automatisation permet de réduire le nombre de profils devant être consultés.

3.6.1 Méthodologie générale

La méthodologie proposée repose sur une combinaison d'analyses comportementale et de contenu des profils Twitter. L'analyse comportementale repose sur l'outil SPOT et l'analyse concernant le contenu est assurée par la méthode NUANCE [128, 129]. La figure 3.6.1 présente un aperçu global de l'approche proposée.

La première étape vise à identifier les utilisateurs suspects via la méthodologie SPOT 1.0. La deuxième étape vise à identifier les relations importantes entre les profils suspects. Ce travail est effectué en utilisant une mesure de similarité basée sur le contenu. Cette similitude combine une méthode d'attribution d'auteur et une mesure temporelle axée sur les entités. Enfin, un algorithme de groupement de données est effectué pour identifier les campagnes. Les campagnes détectées sont caractérisées pour évaluer leurs particularités (p. ex. la taille, la dynamique, la stratégie) et leur niveau de danger. Cette section fournit des détails sur les deux étapes additionnelles à SPOT.

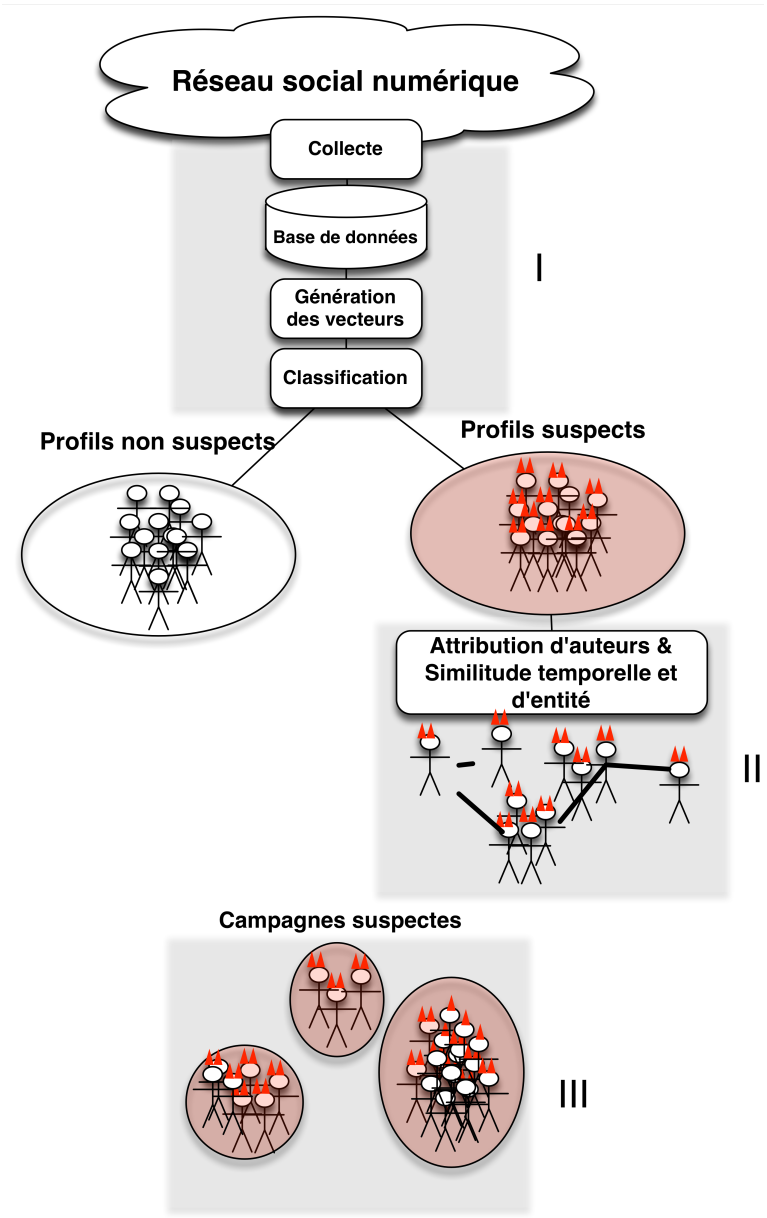


FIGURE 3.6.1 – Méthodologie de détection de campagnes malveillantes.

Une fois les profils suspects identifiés par SPOT, la deuxième étape de l'approche proposée concerne l'enquête sur les profils obtenus. Une analyse basée sur le contenu identifie les connexions entre les profils. Tout d'abord, un algorithme d'attribution d'auteurs est appliqué sur les tweets recueillis des profils suspects [130, 131]. Ensuite, une similitude temporelle est mesurée à partir des entités contenues dans les messages. Les deux analyses contribuent à la mesure de la similitude entre les profils.

L'analyse d'attribution d'auteur a été effectuée en utilisant une version allégée de l'algorithme de NUANCE [129], qui a été utilisé à l'origine dans [131]. Dans l'algorithme NUANCE normal, un ensemble de données est utilisé pour regrouper des documents par auteur. Alors qu'il a été montré efficace, l'algorithme a une complexité élevée, ce qui rend impossible son utilisation dans sa forme actuelle pour résoudre ce problème. Au lieu de cela, nous avons utilisé une seule instance de l'algorithme recentré de profils locaux [132] afin de comparer les profils de chaque compte.

La distance entre toutes les paires de comptes a été calculée et les valeurs de probabilité ont été mesurées en comparant les distances précalculées à partir d'un ensemble de données d'apprentissage contenant des correspondances de compte connus (voir [131] pour plus de détails). La distribution des distances précalculées est utilisée pour calculer ces probabilités empiriques, la probabilité d'appartenance de deux comptes au même auteur est évaluée comme le pourcentage de couples de comptes dans l'ensemble de données d'entraînement avec une distance plus élevée.

Les comptes correspondant à une probabilité supérieure à un seuil donné sont considérés comme du même auteur. Bien que les seuils « normaux», tels que 0.9 ou 0.95, pourraient être utilisés, nous avons calculé le seuil optimal en utilisant un ensemble de données qui maximise la F-mesure (c'est-à-dire la moyenne harmonique de la précision et du rappel) et utilisé cette valeur, qui était de 0.989. L'utilisation de seuils plus élevés abouti à des valeurs de rappel plus faibles, tandis que les seuils inférieurs ont donné des valeurs de précision inférieures.

En ce qui concerne la similitude temporelle axée sur les entités, la durée de l'expérience est divisée en trames temporelles T_i de longueur égales à δt. L'algorithme 3.1 présente la méthode d'évaluation de la mesure de similitude de deux profils u et v pour une période de temps T_i donnée. Cette similitude a pour but d'identifier les profils qui agissent en coordination. L'hypothèse sous-jacente est que les campagnes sociales qui ont un but commun doivent, à un instant donné, présenter un minimum de similitude. Les trois types d'entités de Twitter sont considérés : les références (p. ex. @nom_d'utilisateur), les mots-clés (p. ex. *#hashtags*) et les URL. Un couple de profil partageant les mêmes hashtags peut révéler une approche commune pour attirer l'attention des utilisateurs. Enfin, l'utilisation d'une même URL malveillante par un même couple de profils peut révéler l'utilisation de la même stratégie (vulnérabilité).

L'entrée de l'algorithme est un couple de profils suspects notée (u, v) et le résultat de celui-ci est un vecteur $B_T(u, v) = (B_{T_1}(u, v), B_{T_2}(u, v), ...)$ où $B_{T_i}(u, v)$ est la similitude temporelle au pas de temps T_i. Le vecteur de similitude temporelle est initialisé à la ligne 1. Une boucle est effectuée sur chaque entité e qui a été publiée par les deux profils u et v (ligne 2). Le moment de la publication de ces entités est analysé pour identifier le pas de temps correspondant (lignes 3 et 4). Lorsque l'entité est identifiée comme appartenant à la trame de temps en cours, le score de similarité est augmenté tel qu'indiqué ligne 5. Le premier volet du score est calculé sur la base de la fonction fréquence inverse de document (IDF) [133]. Dans notre cas, un document se réfère à un message et un terme correspond à une entité e. m_j désigne l'ensemble des tweets qui contiennent l'entité et M est le nombre

Algorithme 3.1 Algorithme de mesure de la similitude entre deux profils

Entrée: (u, v) // Un couple de profils Twitter
Entrée: $E(u)$ // L'ensemble d'entitées publiées par un profil u
Entrée: t_e^u // Temps de publication de l'entité e par le profil u
Sortie: $B_T(u, v)$ // Le vecteurs contenant les scores de similitude $B_{T_i}(u, v)$ pour chaque pas de temps i

$B_T(u, v) \leftarrow \emptyset$
pour tout $e \in E(u) \cap E(v)$ **faire**
  pour tout $T_i \in [0, T]$ **faire**
    si $t_e^u \in \left[T_i - \frac{\delta t}{2}, T_i + \frac{\delta t}{2}\right]$ **ou** $t_e^v \in \left[T_i - \frac{\delta t}{2}, T_i + \frac{\delta t}{2}\right]$ **alors**
      $B_{T_i}(u, v) = B_{T_i}(u, v) + log(\frac{|M|}{|\{m_j : e \in m_j\}|}) * e^{-\frac{|t_e^u - t_e^v|}{\tau}}$
    fin si
  fin pour
fin pour
renvoyer $B_T(u, v)$

de tweets concernés dans l'expérience.

Le second volet du score est une mesure de la synchronisation observée entre les entités communes qui sont publiées. La mesure de la synchronisation repose sur une décroissance exponentielle. La durée de vie moyenne τ, permet d'ajuster le score de similarité en fonction du niveau de synchronisation souhaité. Le réglage de τ à une valeur importante permettra la détection de messages fortement synchronisés. Inversement, une petite valeur permettra la détection des profils similaires, même s'ils ne sont pas strictement simultanés.

Ce score temporel de similarité est combiné avec le résultat de l'attribution d'auteur pour fournir une similitude entre un couple de profils. Cette similarité globale est identifiée comme la force d'un lien entre les profils. Nous proposons de définir la force $w_{T_i}(u, v)$ entre un profil u et un profil v à l'instant T_i tel qu'indiqué dans la définition suivante.

Définition 3.4. Similarité entre deux profils suspects u et v La similarité entre deux profils u et v est définie comme suit :

$$w_{T_i}(u, v) = \alpha A(u, v) + \beta B_{T_i}(u, v) \tag{3.6.1}$$

Où :
$A(u, v)$ se réfère au résultat de l'attribution d'auteur entre u et v
$B_{T_i}(u, v)$ se réfère à la similarité temporelle entre u et v

Notons que nous pratiquons pas l'attribution d'auteurs de manière dynamique. Une des raisons est que le nombre de messages nécessaires pour effectuer une telle analyse et obtenir de bonnes performances est relativement important (environ 200 *tweets*). Dans notre cas, cela est rendu possible par la collecte des 200 derniers tweets de chaque profil suspect avec l'API de recherche Twitter. Cependant, nous ne pouvons garantir qu'une quantité suffisante de messages soit publiée à chaque pas de temps pour appliquer l'attribution d'auteur dynamiquement. La complexité de l'algorithme est une autre raison de ce choix. Les coefficients alpha et bêta doivent être adaptés aux contraintes de l'expérience (p. ex. le temps de réponse, le nombre de messages, le nombre de profils).

Dans la troisième étape, nous cherchons à révéler des profils suspects qui partagent des similitudes importantes. Pour ce faire, nous proposons de construire un graphe suspect

noté $G(N, A)$ qui est composé d'un ensemble de nœuds N (c.-à-d. profils suspects) et d'un ensemble d'arêtes A (c.-à-d. relations). Le poids sur une arête entre deux profils u et v est mesuré comme la somme, sur la durée de l'expérience, des poids $w_{T_i}(u, v)$.

Nous appliquons un seuil pour éliminer les liens non significatifs sur le graphe, puis nous effectuons un algorithme de clustering basé sur la modularité sur ce graphe pour détecter les campagnes suspectes [134]. La modularité d'un graphe est une mesure de la force d'une partition spécifique d'un réseau en communautés [135]. La modularité est grande quand une partition obtient de nombreux liens au sein des communautés et peu de liens entre celles-ci. L'algorithme utilise une optimisation afin de trouver les partitions des graphiques qui ont une grande modularité. Il commence par l'examen de chaque nœud en tant que communauté et fusionne les communautés qui génèrent la plus forte augmentation dans le score de modularité.

Les campagnes sont caractérisées à la fois par des attributs basés sur le comportement et basés sur le contenu afin d'identifier clairement leur but. La caractérisation des campagnes a un double objectif : d'abord, elle permet de mieux comprendre les stratégies utilisées par les acteurs malveillants et, ensuite, elle peut permettre l'identification de profils détectés potentiels qui appartiennent à ces groupes en appliquant des mesures telles que l'affinité.

3.6.2 Résultats

Nous avons testé notre méthode sur un ensemble de profils actifs de la plateforme Twitter pendant la durée de l'expérience (à partir du 01/03/2012 et jusqu'au 10/04/2012). Pour cette expérience, le paramètre de seuil K a été fixé à 20 messages (un tel seuil permet d'identifier uniquement les profils très actifs). Un ensemble de 1 000 profils suspects identifiés par SPOT a été sélectionné pour enquête plus approfondie. La figure 3.6.2 représente le graphe suspect $G(N, A)$ qui contient les liens (représentant la similitude) entre les profils (les paramètres alpha et bêta sont de poids égal). Les graphes sont représentés avec le logiciel Gephi [136].

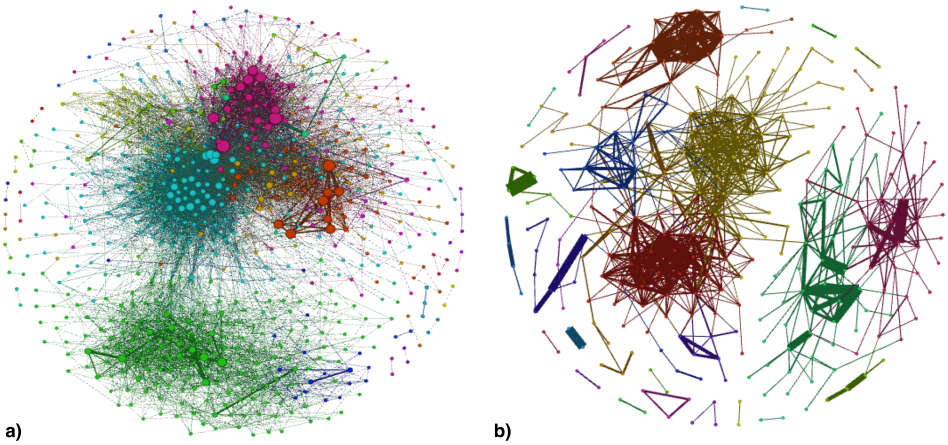


FIGURE 3.6.2 – Graphe suspect contenant les campagnes détectées.

Dans la sous-figure a) le graphe contient tous les nœuds de degrés non nuls identifiés au cours de l'expérience. Le nombre de nœuds est égal à 680, soit environ 70 % des individus suspects analysés. Cette première observation indique que moins de 30 % des profils suspects de notre ensemble de données semble se comporter de manière totalement indépendante de tout type de campagne. Le coefficient de clustering mesuré est égal à

0.505, la modularité est égale à 0.681 et la longueur du chemin moyen est égal à 2.87. Dans la sous-figure b), nous avons mis en évidence les principaux groupes en supprimant les connexions non-significatives (liens avec un faible poids < 5 % de la valeur maximale).

Nous présentons ci-dessous les résultats de la caractérisation du comportement pour les trois plus grandes campagnes identifiées.

La figure 3.6.3 illustre le nombre de profils de chaque campagne au cours de la durée de l'expérience.

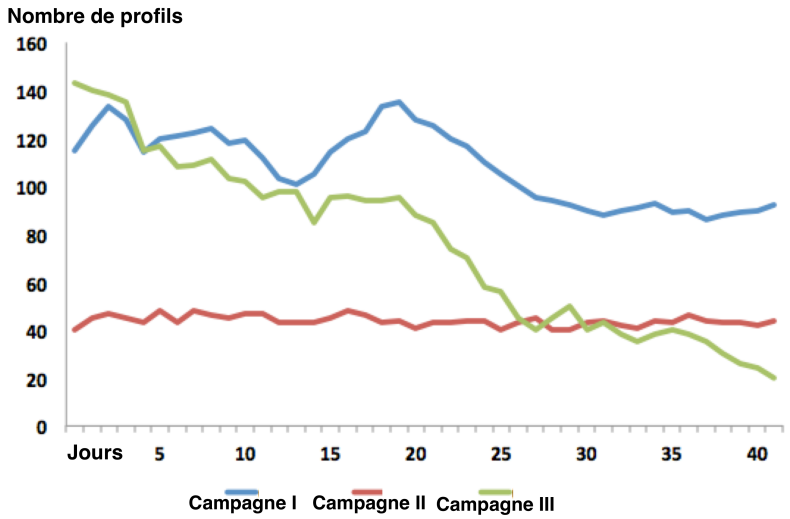


FIGURE 3.6.3 – Évolution du nombre de profils appartenant aux trois campagnes.

La première campagne compte environ 100 participants avec un nombre maximum de 140 à 20 jours. La seconde campagne maintient un nombre relativement stable de participants de 40. La troisième campagne enregistre une diminution régulière du nombre de participants : 140 au premier jour et seulement 25 au vingtième jours. Les premières observations montrent que les réseaux suspects sont dynamiques et évoluent au fil du temps. Une diminution de la taille d'un réseau peut s'expliquer par divers facteurs tels que la suppression des profils malveillants par Twitter. Il est également plausible que le gestionnaire d'un réseau malveillant décide d'arrêter la campagne, temporairement ou définitivement. Inversement, une augmentation du nombre de participants à un réseau suspect peut s'expliquer par la création de nouveaux profils, l'augmentation du nombre de profils suspects agissant en coordination à un moment donné.

Nous exposons dans les tableaux 3.6 à 3.8 les caractéristiques comportementales des classes obtenues. Pour des fins de comparaison, nous avons affiché dans le tableau 3.5 les valeurs moyennes pour les mêmes caractéristiques sur un grand nombre de profils normaux. Nous observons que parmi les paramètres les plus pertinents, le nombre moyen de *hashtags* utilisé dans les tweets est beaucoup plus élevé pour les profils suspects que pour les profils normaux. Cette observation confirme le fait qu'un profil suspect vise à gagner en visibilité en incluant les *hashtags* dans ses messages (les *hashtags* permettent à une communauté d'intérêts de recevoir des tweets).

D'autre part, les profils normaux utilisent plus de références ($Moyenne@ = 0.68$) que les profils suspects ($Moyenne@ = 0.30$). Rappelons-nous que les *tweets* contenant des références (par exemple @nom_d'utilisateur) sont automatiquement reçus par la personne référencée. Les scores révèlent que les campagnes malveillantes ne visent pas des individus particuliers, mais plutôt des groupes identifiés par les *hashtags*. Le nombre

Tableau 3.5 – Moyennes des indicateurs pour les profils normaux

	Moyennes
Âge du compte	513
Fréquence de tweet	94.0
Moyenne #	0.18
Moyenne @	0.68
Moyenne URL	0.20
Fréquence de RT	0.16
#Amis	1 049
#Suiveurs	2 836
#Tweets	110

Tableau 3.6 – Moyenne des caractéristiques pour la campagne I

	Moy	Min	Max	Med
Âge du compte	668	312	1 323	560
Fréquence de tweet	106	0.11	429	91.9
Moyenne #	2.00	0.10	5.50	1.7
Moyenne @	0.60	0.00	3.9	0.00
Moyenne URL	0.69	0.00	1.23	0.94
Fréquence de RT	0.20	0.00	3.30	0.00
#Amis	5 086	0.00	22 127	3234
#Suiveurs	6473	3.00	22 680	4328
#Tweets	321	229	485	297

moyen d'URL par tweet est également beaucoup plus élevé pour les profils appartenant à des communautés suspectes ($MoyenneURL = 0.70$) que pour les profils normaux ($MoyenneURL = 0.20$). Il en ressort que ces profils agissent sur la plateforme principalement pour augmenter la visibilité d'un ensemble donné de sites Internet.

Les caractéristiques comportementales des groupes suspects indiquent qu'il n'existe pas de distinction significative entre les trois campagnes suspectes. Afin de mieux les caractériser, nous proposons d'étudier la qualité des URL qu'ils diffusent.

Les figures 3.6.4 à 3.6.6 présentent les boîtes de Tukey des scores obtenus pour l'ensemble des URL produites par chaque campagne sur un ensemble de quatre indicateurs. Sur ces figures, les points représentent les valeurs anormales de l'échantillon de données. Le minimum et le maximum (hors anomalies) sont représentés par les lignes horizontales inférieures et supérieures. Les limites du rectangle principal sont définies par les valeurs des quartiles inférieur et supérieur. La médiane est représentée par la ligne en gras.

Cette évaluation a été effectuée sur la base de l'API Web Of Trust (WOT) qui utilise une approche de crowdsourcing pour évaluer la qualité d'un site web. Les quatre critères

Tableau 3.7 – Moyenne des caractéristiques pour la campagne II

	Moy	Min	Max	Med
Âge du compte	462	308	1 363	365
Fréquence de tweet	88.8	0.07	402	40.5
Moyenne #	2.03	0.33	6.53	1.61
Moyenne @	0.01	0.00	0.09	0.00
Moyenne URL	0.81	0.01	1.07	0.96
Fréquence de RT	0.00	0.00	0.00	0.00
#Amis	1 349.1	0.00	9 819	191
#Suiveurs	1 297	0.00	8 949	460
#Tweets	268	232	314	268

Tableau 3.8 – Moyenne des caractéristiques pour la campagne III

	Moy	Min	Max	Med
Âge du compte	615	322	989	575
Fréquence de tweet	139	11.0	370	87.2
Moyenne #	2.01	1.01	3.48	1.78
Moyenne @	0.04	0.00	0.15	0.01
Moyenne URL	0.81	0.43	1.00	0.9
Fréquence de RT	0.02	0.00	0.08	0.01
#Amis	999	6.00	2 210	890
#Suiveurs	1 404	155	2 882	1 289
#Tweets	293	237	381	278

sont : 1) la confiance, 2) la fiabilité, 3) le respect de la vie privée et 4) la sécurité des enfants. La confiance fait référence à la confiance globale que l'on donne au site (p. ex. service, sécurité), la fiabilité se réfère à la confiance que l'on donne sur le site pour effectuer des transactions d'affaires (p. ex. par exemple, achat, vente). Le respect de la vie privée se réfère à la confiance en ce qui concerne la manipulation des renseignements personnels. Enfin, la sécurité des enfants se réfère à l'existence de contenus inappropriés à certains âges sur le site. À chaque critère est attribué une note comprise entre 0 et 100, qui reflète la valeur du site et qui doit être interprétée comme suit : $100 \geqslant note > 80$ Excellent ; $80 \geqslant note > 60$ Bien ; $60 \geqslant note > 40$ Non satisfaisant ; $40 \geqslant note > 20$ Mauvais ; $20 \geqslant note \geqslant 0$ Très mauvais.

Les résultats mettent en évidence que les campagnes I, II et III semblent citer majoritairement des sites dignes de confiance. Cependant, quelques anomalies pour les campagnes II et III révèlent le fait que certains des sites ont un très mauvais score en ce qui concerne la sécurité des enfants. La campagne II révèle un ensemble important de sites anormaux avec des caractéristiques très mauvaises. Cela souligne clairement le fait qu'un ensemble

de sites est malveillant et que la campagne a des intentions malveillantes (p. ex. phishing).

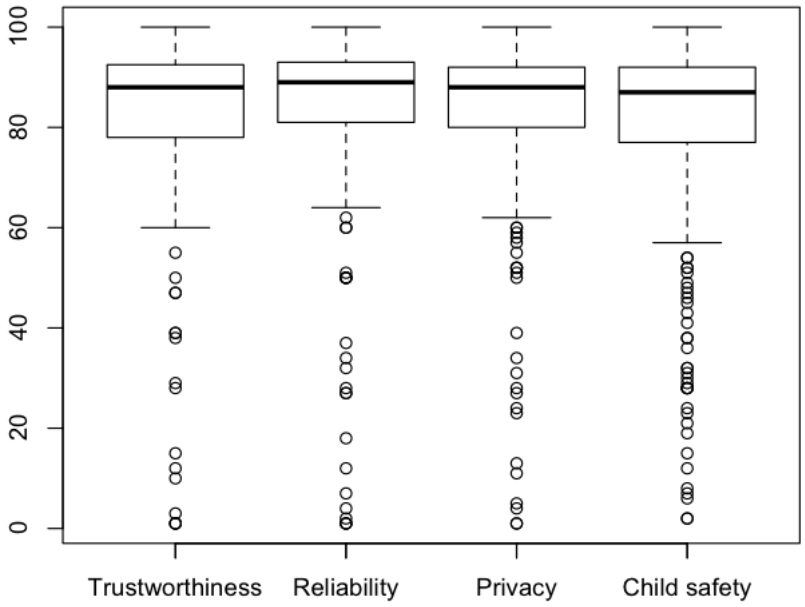


FIGURE 3.6.4 – Caractérisation des URL produites par la campagne I.

Nous mesurons la performance de notre approche avec les indicateurs de pureté, pureté inverse, précision et la mesure F. Ces mesures visent à évaluer les différences entre l'échantillon de données de test (c.-à-d. des profils avec des étiquettes vérifiées) notées L avec l'échantillon de données obtenues (c.-à-d. des profils avec des étiquettes issues de l'algorithme) noté C.

Étant donné un ensemble de campagnes identifiées malveillantes $L = L_1, ..., L_m$ et un ensemble de campagnes détectés $C = C_1, ..., C_k$, la pureté d'une approche de classification est définie comme suit :

$$Pureté = \sum_{i=1}^{k} \frac{|C_i|}{k} \max_{j \in 1,...,m} Précision(C_i, L_j) \tag{3.6.2}$$

La précision d'une campagne C_i détectée, pour une catégorie L_j est définie comme suit :

$$Précision(C_i, L_j) = \sum_{i=1}^{k} \frac{|C_i \cap L_j|}{|C_i|} \tag{3.6.3}$$

La relation fondamentale entre la précision et le rappel est obtenue par la formule suivante :

$$Rappel(C_i, L_j) = Précision(L_i, C_j) \tag{3.6.4}$$

La pureté inverse est définie par :

$$PuretéInverse = \sum_{i=1}^{m} \frac{|L_i|}{n} \max_{j \in 1,...,k} Rappel(C_j, L_i) \tag{3.6.5}$$

À noter que pour chaque campagne détectée, la précision est évaluée en identifiant la campagne qui possède le plus grand nombre de profils communs. De même, le rappel d'une

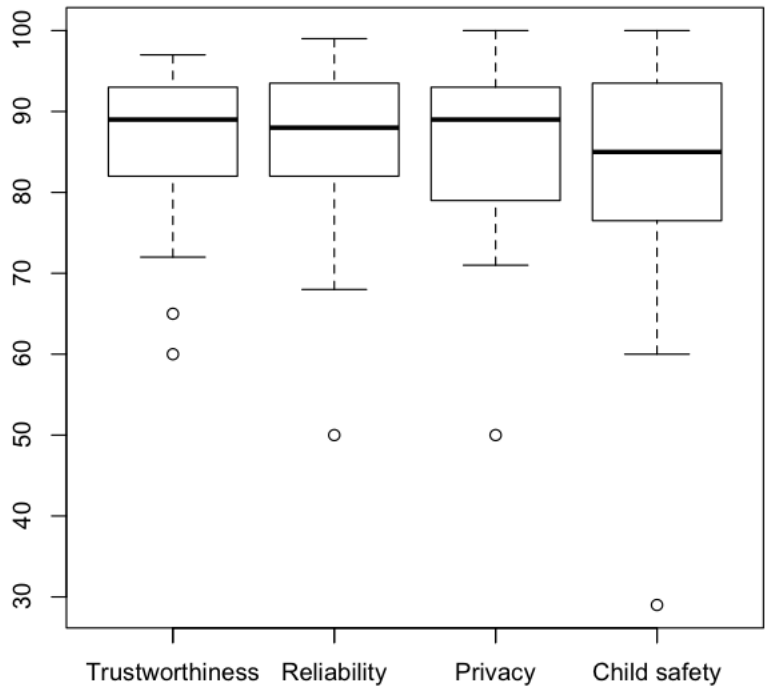


FIGURE 3.6.5 – Caractérisation des URL produites par la campagne II.

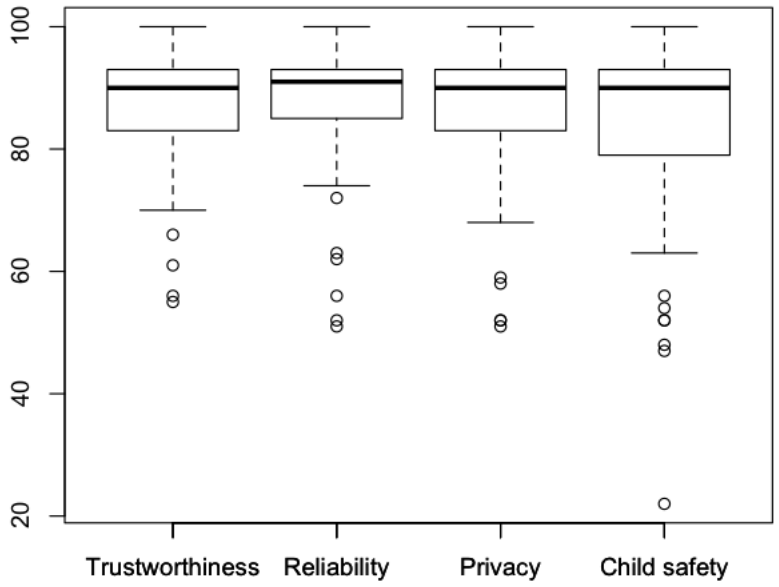


FIGURE 3.6.6 – Caractérisation des URL produites par la campagne III.

campagne identifiée est calculée sur la base de la campagne détectée qui contient le plus grand nombre de profils communs.

Enfin, les performances sont évaluées avec la F-mesure, qui est définie dans l'équation 3.6.6.

$$F = \sum_{i=1}^{m} \frac{|L_i|}{n} max_{j \in 1,\ldots,k} F(L_i, C_j) \tag{3.6.6}$$

Avec :

$$F(L_i, C_j) = \frac{2 * Rappel(L_i, C_j) * Précision(L_i, C_j)}{Rappel(L_i, C_j) + Précision(L_i, C_j)} \tag{3.6.7}$$

La figure 3.6.7 présente la précision et le rappel obtenus pour les trois principales campagnes détectées dans notre échantillon de données. Nous avons obtenu une valeur de pureté égale à 0,81 et une pureté inverse égale à 0,74 inverse. La mesure F obtenue est égale à 0,76.

	Campagne I	Campagne II	Campagne III
Précision	0.79	0.74	0.81
Rappel	0.98	0.67	0.74

FIGURE 3.6.7 – Valeurs de précision et de rappel pour les trois principales campagnes détectées dans l'échantillon de données.

On notera que ces scores sont calculés sur la base du temps complet de l'expérience. L'aspect dynamique de la campagne peut générer des variations des résultats qui dépendent du délai sélectionné.

Campagnes Découvertes

Nous avons évalué la réputation liée à un profil normal moyen comme celle d'un profil appartenant à des campagnes détectées. La réputation d'un profil Twitter notée u et représentée par un nœud dans le graphe social est définie dans l'équation 3.6.8 proposée par [123].

$$R(u) = \frac{d_{in}(u)}{d_{in}(u) + d_{out}(u)} \tag{3.6.8}$$

Sur la base de nos résultats, un profil normal a une réputation globale de 0.73 qui révèle qu'il attire un nombre de suiveurs qui équivalent à peu près au double du nombre de ses amis. Rappelons-nous que les valeurs de nombre moyen d'amis et de suiveurs sont relativement élevées en raison du processus de sélection de SPOT, qui analyse seulement les profils les plus actifs de Twitter.

Les trois campagnes possèdent une réputation moyenne de 0.56, 0.49 et 0.58 respectivement pour les campagnes I, II et III. Bien que ces scores soient plus faibles que pour un profil normal ces dernières valeurs sont néanmoins élevées. Ces scores soulignent le fait que les campagnes malveillants réussissent étonnamment à attirer une quantité d'adeptes qui est supérieure ou au moins aussi importante que leur nombre d'amis. Cette observation peut s'expliquer par les trois raisons suivantes :

— l'organisation de profils dans les campagnes permet d'augmenter artificiellement la réputation en créant des connexions artificielles (c'est ce que l'on appelle *« link farming »*) [79] ;

— de nombreux profils peuvent trouver dans les profils malveillants l'occasion d' accroître leur audience et leur réputation (ils suivent généralement les personnes qui les suivent) ;

— ils exploitent pleinement les *hashtags* ou des références dans le but d'augmenter leur réputation. Il est important de noter que, comme la réputation de campagne augmente les profils malveillants, le nombre de profils vulnérables augmente également.

Des trois campagnes, la campagne II présente le plus faible score de réputation. Une raison possible est la baisse du nombre de participants (environ 40 profils). Cela rend l'augmentation artificielle de la réputation moins efficace que pour les campagnes I et II qui possèdent un plus grand nombre de profils (respectivement 100 et 140). La campagne II possède un nombre moyen de tweets, retweets et de références faibles. Ces trois facteurs peuvent également être pris en compte pour expliquer ce score de réputation plus faible.

Nous avons étudié manuellement les trois campagnes identifiées et avons mis en évidence un ensemble de techniques qu'ils appliquent afin d'atteindre leur objectif. Nous présentons ci-dessous les quatre principales techniques identifiées et présentons dans le tableau 3.9 le niveau d'utilisation de ces techniques par les trois campagnes principales.

— L'utilisation de *has*htags qui encouragent les utilisateurs à suivre les profils d'une campagne (noté Suiveurs). Parmi les hashtags les plus couramment utilisés, nous avons identifié : #500aday, #1000ADAY, #f4f, #Follow4Follow, #followme, # InstantFollow, # MustFollow, #OpenFollow, #OpenFollowPro, #TeamAutoFollow, #TeamFollowBack. Notons que les hashtags de ce type, bien qu'ils puissent paraitre suspects, sont efficaces car de nombreux utilisateurs de Twitter agissent dans l'optique de gagner de l'audience. De même, suivre un profil qui vous suit devient pour certains utilisateurs comme un contrat implicite.

— L'utilisation des mots et des *hashtags* qui encouragent les lecteurs à télécharger, afficher un média ou visiter une page Web (désigné Téléchargement). Parmi les mots et les *hashtags* les plus courants, on peut citer : #download, #freedownload, #freedownloads.

— L'utilisation de tweets pseudo-automatiques (noté Automatisation). Ces *tweets* ont une racine commune et seulement quelques mots, chiffres ou signes de ponctuation sont modifiés.

Tableau 3.9 – Techniques utilisées par les différentes campagnes.

	Suiveurs	Téléchargement	Automatisation	Média
Campagne I	+	++	+	
Campagne II				++
Campagne III	++			

— L'utilisation des médias ou des outils externes pour créer les tweets (noté Média). Certains profils envoient chaque période de temps un tweet qui est extrait à partir d'un flux RSS, mais aussi de quelques phrases pouvant être extraites de livres. Notons qu'à cet effet des outils de gestion des médias sociaux tels que TweetAdder, Twitterfeed, HootSuite peuvent être utilisés.

Nous avons évalué chaque technique pour chaque campagne sur une échelle à trois niveaux : (« ++ ») pour une utilisation très forte, («+») pour un utilisation moyenne et (« ») pour une non utilisation.

Les résultats fournis dans le tableau 3.9 indiquent que la campagne III (qui possède la réputation la plus forte) a un usage intensif de *hashtags* pour encourager les suiveurs. Nous avons constaté que la plupart de leurs tweets ne contiennent que des hashtags qui se réfèrent à des processus de suiveurs / amis. Une représentation des mots et des *hashtags* les plus courants est proposée figure 3.6.8.

Figure 3.6.8 – Nuage de mots produits par la campagne III

La campagne I, comme déjà indiqué dans la section précédente, contribue à la propagation de nombreuses URL malveillantes. Cette observation est confirmée par les résultats fournis dans le tableau 3.9 et la figure 3.6.9 qui mettent en évidence le fait que les tweets publiés par cette campagne encouragent fortement les utilisateurs à télécharger ou à consulter une page Web donnée. Notons que cette campagne utilise également des messages pseudo-automatiques et des *hashtags* qui encouragent les suiveurs. Ces observations peuvent expliquer le fait que, malgré la propagation malveillante d'URL, cette campagne a un niveau assez élevé de réputation.

Enfin, nous avons observé que la campagne II repose uniquement sur l'utilisation de médias pour la création de tweets. Cela signifie qu'elle se base sur des sections de livres ou de sites Web existants pour publier des tweets. Cette campagne n'encourage pas les suiveurs ni le téléchargement mais propage néanmoins de temps en temps des URL qui représentent une menace pour les utilisateurs.

FIGURE 3.6.9 – Nuage de mots produits par la campagne I

Conclusion

Ce chapitre a présenté une méthodologie nommée SPOT pour la détection de profils malveillants sur Twitter. En moyenne, le prototype SPOT 1.0 reçoit 100 000 messages par jour qu'il analyse en temps réel pour générer les caractéristiques nécessaires à l'analyse du comportement des individus de la plateforme. Une moyenne de 2 000 profils sont détectés chaque jour et sont analysés pour évaluer le danger qu'ils représentent. Cette analyse du danger passe par l'analyse quotidienne de 2 500 URL. La plateforme proposée qui évalue le danger par rapport aux URL peut être adaptée pour évaluer toutes sortes de dangers. L'objectif de la plateforme a été de mettre en évidence la présence d'acteurs malveillants sur les nouvelles plateformes sociales mais aussi de mieux comprendre le comportement d'un attaquant sur une telle plateforme. La méthodologie REPLOT a permis de faire évoluer SPOT pour une détection des campagnes de profils malveillants. Cette détection permet de focaliser l'étude sur des groupes de profils et donc d'alléger le travail d'un expert.

Il est important de préciser que cette solution ne permet d'identifier qu'une partie des profils malveillants : la performance et l'échelle d'analyse constituent des verrous. Par conséquent, certains utilisateurs peuvent être exposés, et ce malgré la mise en place de l'outil. Afin d'apporter une solution à ces contraintes, et dans l'objectif de proposer une approche plus proche de chaque utilisateur, une solution mobile et intégrant le comportement de l'utilisateur est proposée dans les prochains chapitres.

Chapitre 4

Détection de contacts légitimes et non légitimes par l'analyse de smartphone

Résumé du chapitre

Ce chapitre propose de répondre au problème de la détection de contacts légitimes (resp. non légitimes) dans la liste de contacts d'un utilisateur. Cette notion de légitimité, implicite mais pas toujours vérifiée, est à l'origine de nombreuses failles de sécurité pour les données des utilisateurs mais aussi pour les systèmes d'information auxquels ils sont associés. Tandis que la majeure partie des travaux de la sécurité des systèmes d'information considèrent le smartphone comme une source de vulnérabilité supplémentaire, nous proposons d'analyser les traces d'activité qu'il comporte pour sécuriser l'utilisateur. Cette analyse locale vise à rechercher et à identifier les individus redondants sur le smartphone. La redondance d'un individu est mesurée à l'aide d'un indicateur d'imbrication qui révèle la présence d'un contact sur de multiples réseaux sociaux de l'utilisateur (p. ex. courriel, SMS, Twitter, Facebook). L'imbrication des contacts est ensuite intégrée dans une mesure de similarité nommée « allocation des ressources pondérée par un smartphone ». Nous testons cette mesure, en comparaison avec les indicateurs concurrents, sur les deux principaux réseaux sociaux numériques que sont Twitter et Facebook. Le chapitre se termine par la présentation de notre prototype d'application mobile « SOCIALYSER » fonctionnel sur les plateformes iOS et Android.

4.1 Application du modèle multicouche dans un cadre mobile

4.1.1 Illustration du modèle dans un cadre nomade

Le modèle multicouche peut être formalisé comme un ensemble L de K couches et un ensemble de matrices de correspondance entre ces couches. Chaque couche représente un réseau social numérique sur lequel l'utilisateur est présent. Une couche de réseau social L_i est représentée par un graphe $G_i = (N_i, A_i)$ où N_i représente l'ensemble des nœuds (c.-à-d. profils) et A_i l'ensemble des liens (c.-à-d. connexions). Les connexions entre les différentes couches sont modélisées par des matrices de correspondance. La figure 4.1.1 montre un exemple d'un réseau multicouche qui peut être extrait à partir du smartphone de l'utilisateur (identifié par les nœuds de couleur noirs).

Dans le cas du réseau multicouche d'un utilisateur de smartphone chacune des couches est un réseau en étoile. En d'autres termes, la seule information disponible sur le smartphone est la liste de contacts de l'utilisateur sur les différents réseaux sociaux.

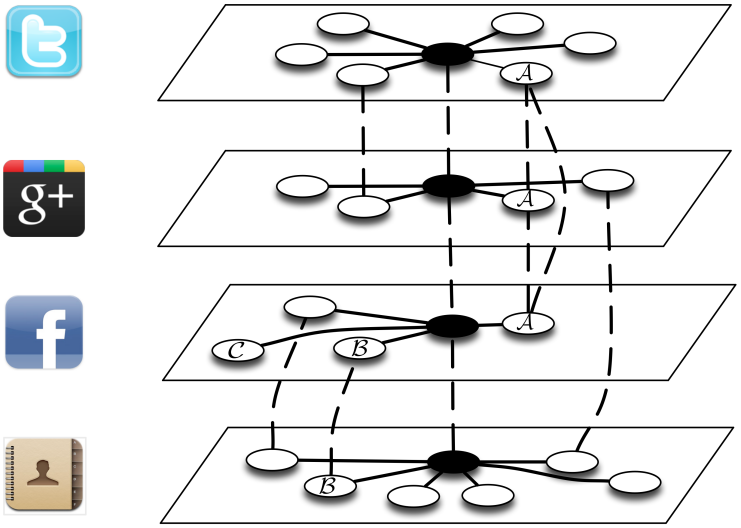


Figure 4.1.1 : Le réseau social égocentrique multicouche d'un utilisateur de smartphone.

Tel qu'indiqué dans le chapitre d'état de l'art, il existe deux types de connexions dans les réseaux multicouche : (1) les liaisons *intra-couche*, qui sont des connexions entre les utilisateurs d'un réseau social, et (2) les connexions *inter-couche*, qui représentent la relation entre les nœuds de réseaux sociaux distincts. Les connexions *intra-couche* sont identifiées par le réseau social considéré. Par exemple, si la couche considérée est Facebook, la nature de la relation est identifiée par l'amitié. Les connexions *inter-couche* du modèle, représentées en traits pointillés sur la figure, exigent la vérification d'une condition de correspondance. Dans ce travail, nous proposons de créer un lien entre deux nœuds si leurs profils associés sur deux couches différentes appartiennent à la même personne.

4.1.2 La recherche des connexions inter-couche sur le smartphone

Les smartphones stockent une grande quantité d'informations sur les multiples interactions d'un utilisateur. Les données d'un smartphone sont généralement gérées avec le

système intégré SQLite, qui est un modèle de gestion de bases de données relationnelles. Celui-ci permet à toute application de stocker des données localement sur un smartphone. Le système SQLite est maintenant disponible sur les principaux systèmes d'exploitation pour smartphone (e.g. iOS, Android) [137, 134].

Il faut noter que les profils FOAF des utilisateurs ne sont pas directement accessibles sur les appareils mobiles. Cela rend les approches telles que [138] non applicables dans ce contexte. Toutefois, une approche possible pour réaliser l'identification d'entité peut consister en l'analyse des bases de données locales. À titre d'exemple, certaines applications permettent aux contacts Facebook et Twitter d'être inclus et mis en correspondance avec les contacts du carnet d'adresses. Sur un iPhone 3GS, nous avons observé que cette opération conduit au stockage de données de correspondance. La table de correspondance est nommée *abentries* et contient les paires de correspondances entre les membres du carnet d'adresses et les utilisateurs de Facebook.

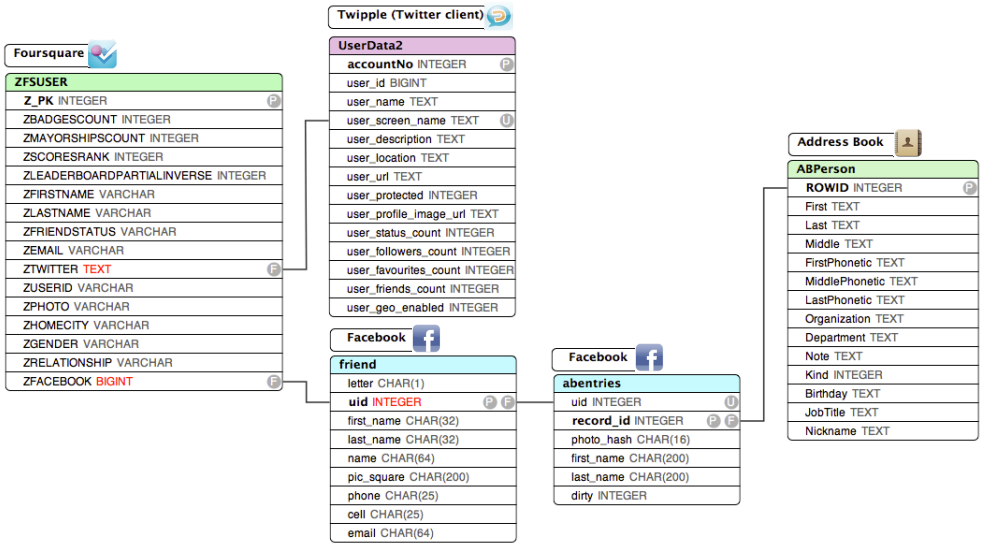


Figure 4.1.2 : Bases de données SQLite d'applications sociales récupérées sur un iPhone 3GS.

La figure 4.1.2 représente les liens qui peuvent être créés entre le carnet d'adresses, Facebook, Foursquare et Twipple (une application cliente de Twitter). Cette figure ne contient que les tables pertinentes pour l'analyse, mais de nombreuses autres tables existent pour chaque application sociale. La figure met en évidence les liens que l'on peut identifier directement entre les contacts trouvés sur les multiples plateformes. Notons que, en fonction des terminaux mobiles et de la version, les correspondances entre les profils peuvent différées de la figure 4.1.2. Dans le cas où aucune donnée de correspondance spécifique n'est déjà présente sur le smartphone, il sera nécessaire d'appliquer des approches classiques telles que la mesure de similarité de Winkler.

4.2 Construction des indicateurs d'imbrication du modèle multicouche

Nous présentons dans cette section un des principaux apports de ce travail à savoir un ensemble d'indicateurs de mesure d'imbrication associées au modèle multicouche. Le réseau social égocentrique multicouche de la figure 4.1.1 servira d'exemple tout au long de la présentation des définitions suivantes.

4.2.1 L'imbrication de profil à profil

Nous proposons de formaliser la fonction qui permet d'identifier les connexions intercouche du modèle tel qu'indiqué dans la définition 4.1. Pour un utilisateur u (c.-à-d. un profil), rencontré au moins une fois dans une des couches du modèle, nous dénoterons par u_k le nœud représentant cet utilisateur sur la couche L_k. Deux nœuds u_k et v_l coïncident si nous détectons par les approches présentées précédemment qu'ils appartiennent au même utilisateur. Dans le cas inverse, les nœuds seront considérés non coïncidents.

Définition 4.1. Fonction de correspondance entre deux nœuds
Soient deux nœuds $u_k \in L_k$ et $v_l \in L_l$ et étant donné un paramètre $\theta \in \mathbb{N}$, la fonction de correspondance d'ordre θ est définie comme :
$\forall k \neq l \in \{1, ..., K\}$,

$$I^{(\theta)}(u_k, v_l) = \begin{cases} 1 & si\ u_k\ et\ v_l\ coincident\ par\ rapport\ à\ \theta \\ 0 & sinon \end{cases} \tag{4.2.1}$$

Le facteur θ représente le nombre maximum d'intermédiaires qui permettent d'identifier par transitivité que deux nœuds correspondent. Par exemple, si un profil Twitter (noté $u_{Twitter}$) correspond à un profil Facebook (noté $v_{Facebook}$) et que ce dernier correspond au profil Google+ (noté $w_{Google+}$), mais qu'aucune correspondance n'est directement identifiée entre les profils Twitter et Google+, alors : $I^{(0)}(u_{Twitter}, v_{Facebook}) = 1$, $I^{(0)}(v_{Facebook}, w_{Google+}) = 1$, $I^{(0)}(u_{Twitter}, w_{Google+}) = 0$ et $I^{(1)}(u_{Twitter}, w_{Google+}) = 1$. Ce paramètre peut permettre l'identification indirecte de correspondances entre profils notamment lorsque différentes techniques sont appliquées. Dans la suite de ce travail on considérera uniquement le cas ou $\theta = 0$ et pour alléger l'écriture nous noterons $I(u_k, v_l)$ à la place de $I^{(0)}(u_k, v_l)$.

Il est important de noter que la fonction de correspondance telle qu'indiquée dans la définition 4.1 est binaire (c.-à-d. une correspondance existe ou pas). Toutefois, en pratique, rarement deux entités sont identifiées comme correspondantes avec certitude. Au lieu de cela, une mesure de similarité est effectuée et des paires d'entités qui possèdent une similarité supérieure à un seuil fixé sont identifiées comme correspondantes. Ainsi, la performance de l'approche repose en partie sur le choix d'un seuil adéquat. À cet effet, il est courant d'utiliser un ensemble de données de référence avec des correspondances et des non-correspondances connues, et de calculer à partir de ces données le seuil qui permet d'obtenir les résultats les meilleurs. Ce seuil est ensuite utilisé pour prédire si une correspondance existe entre deux profils sur de nouveaux échantillons.

La fonction de correspondance entre deux profils nous permet de définir les matrices de correspondance, qui sont les matrices d'adjacence des graphes qui relient chaque paire de réseaux.

Définition 4.2. Matrices de correspondance
Les matrices de correspondance entre deux couches L_k et L_l sont définies par :

$$\forall k \neq l \in \{1, ..., K\}, M(L_k, L_l)_{u_k, v_l} = I(u_k, v_l) \tag{4.2.2}$$

Ces matrices révèlent la façon dont un utilisateur de smartphone organise et communique sur les RSN. Certains utilisateurs peuvent gérer leur présence sur chaque réseau social en fonction de la nature des plates-formes [9]. Par exemple, une matrice composée de valeurs nulles signifie qu'aucune correspondance n'existe entre les deux listes de contacts de l'utilisateur. Ce cas illustre le fait qu'une séparation nette existe entre ces couches. Au contraire, une matrice qui est surtout composée de valeurs 1 indique qu'une relation étroite existe entre la paire considérée de réseaux sociaux. En d'autres termes, l'audience de l'utilisateur dans les deux réseaux sociaux est très similaire.

À titre d'exemple, pour la figure 2.1.3, la matrice de correspondance entre les couches Twitter et Google+ est donnée ci-dessous. Notons que l'ordre des nœuds est indiqué par leur numéro d'identification tel que précisé sur la figure.

$$M(Twitter, Google+) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Il est important de noter que, puisque nous étudions un smartphone, la vision du réseau (plus précisément des contacts) que nous obtenons est la vision de l'utilisateur et non une image globale du réseau. Cependant, cette vue située du réseau peut être riche en information et peut aider à comprendre certains aspects de la vie de l'utilisateur sur les réseaux sociaux numériques.

4.2.2 L'imbrication de profil à réseau

Nous introduisons le concept d'imbrication dans le réseau multicouche extrait d'un smartphone. Nous prenons en compte deux alternatives possibles du modèle : le mode pilier et le mode général. En mode pilier, le modèle est caractérisé par le fait qu'une personne ne peut être représentée qu'une fois sur chaque réseau social (c.-à-d. sur chaque couche du modèle). Cela signifie qu'un nœud d'une couche ne peut pas correspondre à plus d'un nœud d'une autre couche. Dans le cadre du mode général, aucune contrainte n'est imposée quant au nombre de profils d'une personne dans un réseau. On peut avoir plusieurs comptes sur le même réseau social, et donc un nœud d'une couche peut correspondre à plusieurs nœuds d'une autre couche. Concernant le mode général, l'imbrication notée I_G d'un nœuds dans une couche est définie comme étant le rapport entre les nœuds de la couche qui correspondent à ce nœud par la quantité totale de nœuds de la couche.

Définition 4.3. Imbrication d'un nœud dans une couche (mode général)

L'imbrication d'un nœud u_k dans une couche L_l est définie par :

$$\forall k \neq l \in \{1, ..., K\}, I_G(u_k, L_l) = \frac{\sum_{v_l \epsilon N_l} I(u_k, v_l)}{|N_l|} \tag{4.2.3}$$

Dans le cas du mode pilier, l'imbrication dénotée I_P est donnée par le nombre de nœud(s) de la couche qui correspondent à ce nœud. Cette variante tient compte du nombre maximum de connexions inter-couche qui existent entre un nœud et une couche.

Définition 4.4. Imbrication d'un nœud dans une couche (mode pilier)

L'imbrication d'un nœud u_k dans une couche L_l, est définie par :

$$\forall k \neq l \in \{1, ..., K\}, I_P(u_k, L_l) = \sum_{v_l \epsilon N_l} I(u_k, v_l) \tag{4.2.4}$$

Nous illustrons l'imbrication entre un nœud et une couche en utilisant le nœud u de la figure 4.1.1. Ce nœud, appartenant au réseau social Twitter, est identifié à la fois sur Google+ et Facebook. En utilisant le mode pilier, on peut donc observer les résultats suivants :

$$I_P(\mathcal{A}_{Twitter}, Google+) = 1$$

$$I_P(\mathcal{A}_{Twitter}, Facebook) = 1$$

$$I_P(\mathcal{A}_{Twitter}, Carnet\ d'adresses) = 0$$

Le premier constat est que le nœud u a une certaine présence dans la vie numérique du propriétaire de smartphone.

4.2.3 L'imbrication de profil à réseaux

La définition de l'imbrication d'un nœud dans un ensemble de couches est directement déduite de l'imbrication entre un nœud et une couche. Ces indicateurs permettent d'établir les contacts qui sont les plus redondants dans un ensemble de réseaux sociaux.

Définition 4.5. Imbrication d'un nœuds dans un ensemble de couches (mode général)

L'imbrication d'un nœud u_k dans un ensemble de couches S est définie par :

$$\forall L_k \notin S, u_k \in L, I_G(u_k, S) = \frac{\sum_{l \in S} I_G(u_k, l)}{\sum_{l \in S} |N_l|} \tag{4.2.5}$$

Pour le modèle de pilier, la définition est adaptée comme suit :

Définition 4.6. Imbrication d'un nœud dans un ensemble de couches (mode pilier)_

L'imbrication d'un nœud u_k dans un ensemble de couches S est définie par :

$$\forall L_k \notin S, u_k \in L_k, I_P(u_k, S) = \frac{\sum_{l \in S} I_P(u_k, l)}{|S|} \tag{4.2.6}$$

Dans l'exemple donné précédemment, en ce qui concerne le modèle pilier, le nœud u est observé dans trois des quatre couches analysées, et donc $I_P(\mathcal{A}_{Twitter}, L) = 0.75$.

4.2.4 L'imbrication de réseau à réseaux

Nous proposons l'extension de la notion d'imbrication à une paire de couches. Pour les deux variantes du modèle, cette imbrication est égale au nombre de connexions définies entre ces deux couches, divisé par le nombre de connexions maximum autorisé par le modèle.

Définition 4.7. Imbrication entre deux couches (mode général)

L'imbrication d'une couche L_k dans une couche L_l est définie comme suit :
$\forall k \neq l \in \{1, ..., K\}$

$$I_G(L_k, L_l) = \frac{\sum_{u_k \in N_k} I_P(u_k, L_l)}{|N_k|} \tag{4.2.7}$$

Pour le mode pilier, le nombre de connexions autorisées est égal au nombre minimal de nœuds entre les deux couches.

Définition 4.8. Imbrication entre deux couches (mode pilier)

L'imbrication de la couche L_k dans la couche L_l est définie comme suit :
$\forall k \neq l \in \{1, ..., K\}$

$$I_P(L_k, L_l) = \frac{\sum_{u_k \in N_k} I_P(u_k, L_l)}{min(|N_k|, |N_l|)} \tag{4.2.8}$$

Dans l'exemple de la figure 4.1.1, en mode pilier, on peut observer que pour Twitter et Google+ le nombre maximal de connexions est égal à cinq. Parmi ces cinq liens possibles, trois connexions sont observées. Cela signifie que : $I_P(Twitter, Google+)= 0.6$. Les scores

	Carnet d'adresses	Facebook	Google +	Twitter
Carnet d'adresses	1.0	0.4	0.2	0.0
Facebook	0.4	1.0	0.4	0.2
Google +	0.2	0.4	1.0	0.6
Twitter	0.0	0.2	0.6	1.0

Tableau 4.1 – Imbrication en mode pilier entre chaque paire de couches

d'imbrication entre chacune des autres paires de couches sont indiqués dans le tableau 4.1.

Ces observations permettent de mettre en évidence les usages des réseaux sociaux d'un utilisateur. D'une part, les couples de réseaux qui se chevauchent fortement peuvent révéler les possibles utilisations similaires de deux réseaux sociaux. D'autre part, les couches qui ne se chevauchent absolument pas peuvent révéler une séparation volontaire créée par l'utilisateur entre deux de ses facettes numériques. Dans l'exemple, Twitter et le carnet d'adresses sont entièrement disjoints, tandis que Facebook et le carnet d'adresses se chevauchent.

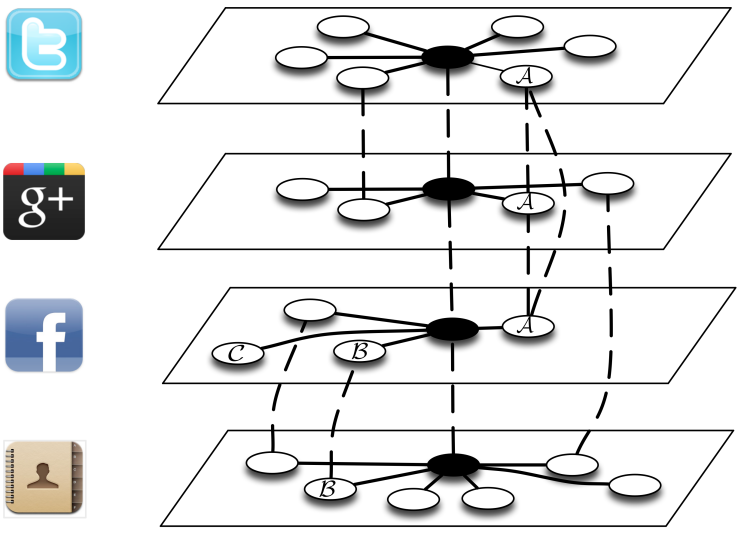


FIGURE 4.2.1 – Exemple de réseau multicouche d'un utilisateur pour la mesure d'imbrication de contacts Facebook $\mathcal{A}$lice, $\mathcal{B}$ob et $\mathcal{C}$arole.

Considérons le réseau multicouche de la figure 4.2.1. Nous souhaitons évaluer l'imbrication des trois contacts Facebook : $\mathcal{A}$lice, $\mathcal{B}$ob et $\mathcal{C}$arole notés $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ dans l'ensemble des réseaux S constitué de Twitter, Facebook, Google+ et du carnet d'adresses. Pour cela nous évaluons $I(\mathcal{A}_{Facebook}, S)$, $I(\mathcal{B}_{Facebook}, S)$ et $I(\mathcal{C}_{Facebook}, S)$.

$\mathcal{A}$lice est identifiée sur les strates Facebook, Google+ et Twitter, par conséquent, et d'après la formule de mesure de chevauchement, la présence d'$\mathcal{A}$lice sur Twitter apporte 1/6, celle sur Facebook apporte 1/4, celle sur Google+ apporte 1/4, le tout étant normalisé par la quantité totale de contacts des quatre couches à savoir 20. Nous obtenons donc une

imbrication d'$\mathcal{A}$lice approximativement égale à 0.033. Les tableaux 4.2 et 4.3 présentent les éléments de détail du calcul.

En ce qui concerne $\mathcal{B}$ob, sa présence sur les réseaux Facebook mais aussi dans le carnet d'adresses lui permet d'obtenir une imbrication de 0.021.

	$I(\_, Twitter)$	$I(\_, Google+)$	$I(\_, Facebook)$	$I(\_, Carnet\,d'adresses)$
$\mathcal{A}$	1/6	1/4	1/4	0/6
$\mathcal{B}$	0/6	0/6	1/4	1/6
$\mathcal{C}$	0/6	0/6	1/4	0/6

Tableau 4.2 – Imbrication de $\mathcal{A}$, $\mathcal{B}$ et $\mathcal{C}$ dans chaque strate.

	Imbrication Totale	**Approximation**
$\mathcal{A}$	$(1/6+1/4+1/4+0/6)/(6+4+4+6)$	0.033
$\mathcal{B}$	$(0/6+0/4+1/4+1/6)/(6+4+4+6)$	0.021
$\mathcal{C}$	$(0/6+1/4+0/4+0/6)/(6+4+4+6)$	0.012

Tableau 4.3 – Calcul de l'imbrication de $\mathcal{A}$, $\mathcal{B}$ et $\mathcal{C}$ dans les quatre réseaux.

$\mathcal{C}$arole n'étant identifiée que sur Facebook ; sont imbrication est la plus faible et approximativement égale à 0.012.

Cet exemple met en évidence, l'importance en termes d'imbrication accordée à un contact qui est présent sur de multiple réseaux sociaux ($\mathcal{A}$lice) par rapport à des contacts qui sont présents sur une quantité plus faible de réseaux ($\mathcal{B}$ob et $\mathcal{C}$arole). Aussi, la présence du contact sur chaque réseau est pondérée par la quantité de contacts de l'utilisateur sur ce réseau. La présence d'un profil sur un réseau social dont l'utilisateur possède beaucoup de contacts est par construction plus faible que celle d'un profil sur un réseau sur lequel l'utilisateur a moins de contacts. Nous pouvons interpréter cette observation de la manière suivante : l'importance d'un contact dans un réseau social peut être identifiée comme étant la probabilité pour l'utilisateur d'interagir avec ce contact s'il répartit son temps de communication de manière égalitaire entre chacun. Notons que, même si cette hypothèse ne soit que rarement vérifiée, la mesure d'imbrication affecte de l'importance à la présence d'un individu dans plusieurs réseaux plutôt qu'à des interactions à l'échelle d'un unique réseau.

4.2.5 Le concept de graphe d'identité

Le graphe d'identité (voir définition suivante) permet de visualiser plus facilement la mesure d'imbrications d'un contact dans un ensemble de réseaux sociaux numériques. Celui-ci est défini comme le graphe contenant un nœud, tous les nœuds qui correspondent avec lui, ainsi que leurs connexions inter-couches correspondantes. Si un contact est représenté sur une seule couche, le graphe d'identité correspondant est un nœud unique.

Définition 4.9. Graphe d'identité
Le graphe d'identité $H^S_{u_k}$ d'un nœud u_k appartenant à une couche k dans un ensemble de couches $S \subset L$, est noté :

$$H_{u_k}^{S}(N^{'}, A^{'}) \tag{4.2.9}$$

Avec :

$$N^{'}(H_{u_k}^{S}) = \{v_l \epsilon N, l \in S |\, I(u_k, v_l) = 1\}$$

Et :

$$A^{'}(H_{u_k}^{S}) = \{(u, v) \epsilon N^{'} \times N^{'} | I(u, v) = 1\}$$

Une manière simple de construire le graphe d'identité d'une personne est de traverser ses connexions inter-couches en commençant par un de ses profils identifiés. Le graphe résultant d'une telle opération est noté $H_{u_k}^{S}(N^{'}, A^{'})$, où u est le nœud de départ et S l'ensemble des couches prises en compte. Sur la figure 4.2.2, en ce qui concerne le mode pilier, nous avons représenté les deux graphes d'identité $H_{\mathcal{A}_{Twitter}}^{S}(N^{'}, A^{'})$ et $H_{\mathcal{B}_{Facebook}}^{S}(N^{'}, A^{'})$. On peut facilement vérifier que les indicateurs proposés sont directement liés aux caractéristiques du graphe d'identité (par exemple, étant donné un nœud u appartenant à une couche k et un ensemble de couches $S \subset L$, $I_P(u_k, S) = \frac{\left|N'(H_{u_k}^{S})\right|}{|S|}$. L'imbrication de la personne correspondante est donc proportionnelle au nombre de nœuds d'un tel graphe.

FIGURE 4.2.2 – Graphe d'identité de deux profils u et v appartenant respectivement aux réseaux sociaux Twitter et Google+.

4.3 L'imbrication comme indicateur de légitimité

Dans cette section, nous présentons un exemple d'application des indicateurs d'imbrication pour l'évaluation de la fiabilité des contacts d'un utilisateur de smartphone. Le système d'évaluation proposé repose sur le domaine de la prédiction de liens. La figure 4.3.1 illustre la situation. Nous voulons affecter un score de légitimité pour les contacts d'un utilisateur de smartphone sur un réseau social, comme Facebook ou Twitter. Le nœud x représente un profil Facebook qui est évalué et le nœud y représente le profil d'utilisateur de smartphone. Notre approche vise à identifier la similitude entre x et y, à partir uniquement des données représentées sur cette figure (c.-à-d. les voisins de x sur le réseau analysé et le réseau multicouche de l'utilisateur y).

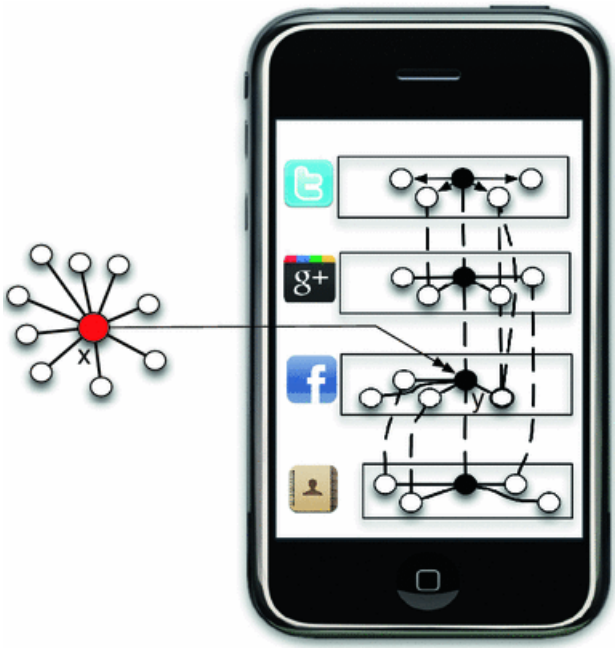


FIGURE 4.3.1 – Réseau multicouche d'un utilisateur de smartphone.

4.3.1 Similarité entre deux profils dans le contexte local

Le problème de prédiction de liens peut être formalisé comme suit : étant donné un aperçu d'un réseau social au temps t, quelles sont les connexions (entre profils) qui sont les plus susceptibles d'apparaître dans le futur ? Les techniques courantes de prédiction de liens attribuent une note de similarité pour chaque paire de nœuds et prédisent des connexions pour les meilleurs scores. Les mesures de similarité peuvent être organisées en trois groupes principaux : les indices globaux, les indices semi-locaux et les indices locaux [96]. Le contexte mobile de notre analyse diffère du problème de prédiction de liens habituel. Les principales différences sont les suivantes :

(1) nous n'avons pas une vision globale du réseau social ;
(2) nous sommes seulement intéressés par les liens qui apparaissent entre un profil et l'utilisateur de smartphone ;
(3) nous disposons du réseau multicouche de la personne demandée.

La situation est illustrée par la figure 4.3.1. Nous proposons de transposer le problème de prédiction de liens d'un point de vue utilisateur comme suit : étant donnée l'information locale d'un réseau social en ligne, quels contacts (c.-à-d. amis futurs ou existants) apparaissent comme les plus légitimes ?

Plusieurs indicateurs de similarité locale ont été proposés et une liste exhaustive peut être trouvée dans l'état de l'art de cette thèse ainsi que dans les travaux [96] et [97].

4.3.2 Proposition d'un nouvel indicateur basé sur les données du smartphone

Les réseaux sociaux en ligne n'effectuent pas généralement de distinction entre les différents types de relations que l'on peut construire (famille, collègues de travail). Ainsi, et par défaut, chaque contact est traité de manière égale en termes d'autorisations et

d'importance. De plus, le système de recommandation d'amis de Facebook est fondé principalement sur le nombre de voisins communs [98]. Ce modèle n'est clairement pas un moyen suffisant pour assurer le niveau de fiabilité d'un certain nombre de contacts dans le réseau.

Pour répondre à cette limitation, une solution serait d'intégrer la quantité de contacts de ces voisins communs (c'est ce que propose l'indicateur d'allocation de ressource). Bien que cela semble être une bonne solution, il est clair que le degré d'un contact n'est pas suffisant pour déterminer son importance. Par exemple, sur Facebook, de nombreux contacts peuvent être ignorés par une personne en raison de leur présence passive, accidentelle ou non significative dans leur ensemble d'amis. Au contraire, les membres de la famille de l'utilisateur devraient être identifiés comme plus importants et donc avoir une incidence sur le score de similarité plus forte. Dans cet objectif, nous proposons d'enrichir l'indicateur d'allocation des ressources en établissant l'imbrication comme paramètre de répartition des informations. L'hypothèse sous-jacente est que les émetteurs importants sont plus susceptibles d'être des profils dont le chevauchement dans le réseau social multicouche est important. L'indicateur proposé d'allocation des ressources pondérée par le smartphone (SBRA pour *Smartphone Based Ressources Allocation*) est présenté dans la définition 4.10.

Définition 4.10. Allocation des ressources pondérée par le smartphone (SBRA)

$$s_{xy}^{SBRA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{I_P(z, S)}{k_z}. \tag{4.3.1}$$

Avec :

x l'utilisateur du smartphone ;
y l'utilisateur à analyser ;
S l'ensemble des couches du modèle ;
k_z le degré du nœud z ;
$I(z, S)$ l'imbrication de z dans S.

Cet indicateur affecte un score de similarité entre deux profils x et y d'autant plus fort que : (1) ils partagent une quantité importante de voisins communs, (2) leurs voisins communs ont une quantité de contacts qui est faible, (3) leur voisins communs ont une forte imbrication dans le réseau multicouche extrait du smartphone de l'utilisateur y. Notons que, de cette manière, le score de similarité s_{xy}^{SBRA} n'est pas une relation symétrique car les scores d'imbrication des voisins communs à x et y extraits du réseau multicouche de x, ne sont pas nécessairement identiques aux scores d'imbrication des voisins communs à x et y extrait du réseau multicouche de y. De plus, l'ensemble des interactions mémorisées sur le smartphone sont exploitées et permettent ainsi d'intégrer le comportement de l'utilisateur pour établir la mesure.

Considérons l'exemple de la figure 4.3.2 pour illustrer le calcul de la mesure de similarité entre l'utilisateur de smartphone x et trois profils Facebook notés y_1 et y_2 et y_3. Le résultat de l'imbrication des nœuds voisins de x dans l'ensemble des réseaux disponibles est indiqué à proximité de leur position. On peut facilement observer, par la typologie du réseau multicouche, que le score d'imbrication du nœud $\mathcal{D}$ est égal à celui du nœud $\mathcal{B}$. À partir

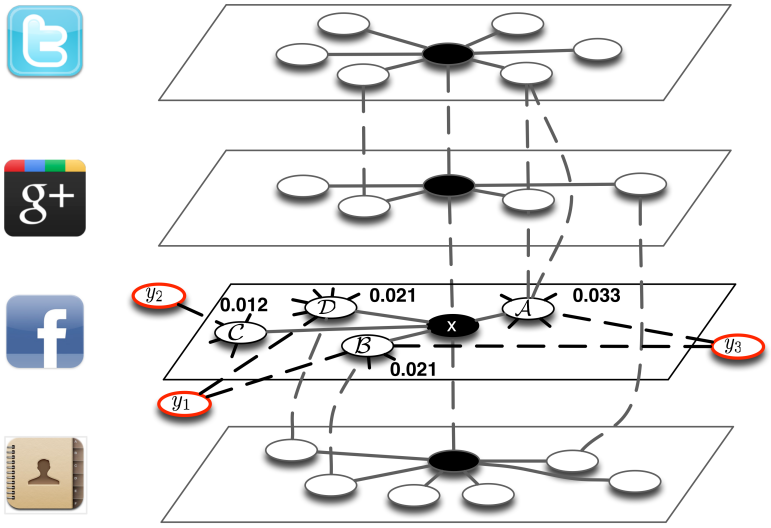


FIGURE 4.3.2 – Mesure de similarité entre l'utilisateur x et trois utilisateurs de Facebook notés y_1 et y_2 et y_3.

de la définition 4.10, nous avons indiqué dans le tableau 4.4 les différentes opérations de calcul. Nous pouvons observer que, conformément à l'intuition, la valeur de similarité entre x et y_3 est la plus forte. Cela s'explique à la fois par le fait de partager sur le réseau Facebook un ensemble de connexions communes et par l'importance de ces connexions sur le réseau multicouche de x. Même si le nœud y_1 possède autant de connexions en commun que le nœud y_3, la forte imbrication d'$\mathcal{A}$lice dans le réseau multicouche permet à y_3 d'obtenir un score de similarité plus fort. Il ne s'agit donc pas seulement d'évaluer la quantité de voisins communs mais aussi, par le biais de l'imbrication de pondérer leur importance. Il est logique, à quantité de voisins communs identique, d'accorder plus de similarité à y_3, à partir du moment où sa présence sur de nombreux réseaux de l'utilisateur est détectée. Sans cette analyse multi-niveau, la vision de similarité est beaucoup moins fine. Enfin, le nœud y_2 ne partageant qu'une connexion en commun et celui-ci étant très faiblement imbriqué, la mesure de similarité associée est la plus faible.

	$\Gamma(x) \cap \Gamma(\_)$	$\sum_{z \epsilon \Gamma(x) \cap \Gamma(\_)} \frac{I(z,L)}{\lvert\Gamma_z\rvert}$	$s^{SBRA}_{x\_}(\mathcal{I}dentité)$
y_1	$\{\mathcal{B}, \mathcal{D}\}$	$0.021/5 + 0.012/5$	0.0066
y_2	$\{\mathcal{C}\}$	$0.012/5$	0.0024
y_3	$\{\mathcal{A}, \mathcal{B}\}$	$0.033/6 + 0.021/5$	0.0097

Tableau 4.4 – Calcul de la similarité entre x et y_1, y_2, y_3 sur Facebook.

4.3.3 L'évaluation des performances de l'indicateur proposé pour évaluer la légitimité

Nous avons testé notre approche sur un ensemble de réseaux sociaux (carnet d'adresses, Twitter, Google+ et Facebook) extraits de smartphones et sur un ensemble de contacts (amis et non-amis) extraits de Facebook et de Twitter. Les graphes d'identité extraits du réseau multicouche d'un utilisateur donné sont représentés sur la figure 4.3.3. Chaque composante connexe du graphe représente le graphe d'identité d'un contact Facebook

parmi les quatre autres réseaux. Cette représentation du smartphone d'un utilisateur permet de mettre en évidence la présence d'une majorité de contacts sur une seule et unique strate. Certains contacts sont cependant présents sur deux, trois ou même l'ensemble des réseaux analysés. Ces individus tiennent une place importante pour l'utilisateur car il communiquent et échange de l'information sur de nombreux et divers canaux de communications. Notons que l'identification de tels types d'individus est importante dans l'identification de contacts légitimes et intégrée dans l'approche via la mesure de similarité présentée dans la définition 4.10.

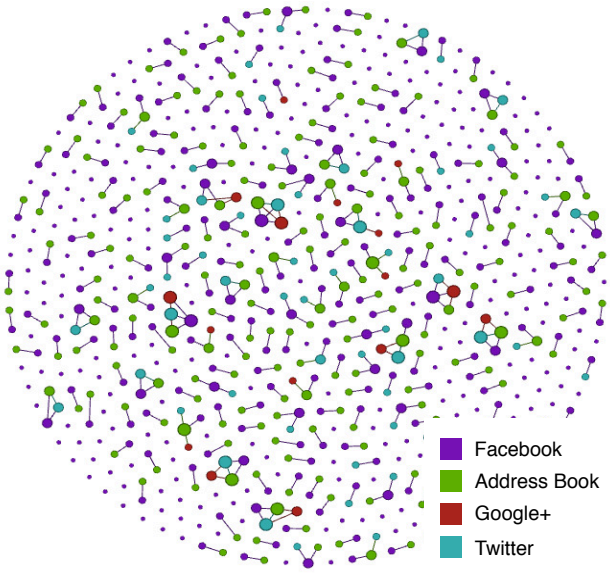


FIGURE 4.3.3 – Graphes d'identité illustrant l'imbrication des contacts Facebook parmi les trois autres réseaux considérés.

Nous avons testé notre approche sur deux échantillons de graphes de Facebook et Twitter. Dans le cas de Facebook, le graphe contient 25 230 nœuds reliés par 29 521 arêtes. Le coefficient de *clustering* mesuré est égal à 0,636 et la longueur du chemin moyen entre deux nœuds est égale à 4,48. Dans le cas de Twitter, le graphe contient 15 088 nœuds et 17 236 arêtes. Le coefficient de *clustering* est 0,247 et la longueur du chemin moyen est de 3,2. Ces graphes sont obtenus pour chaque utilisateur de smartphone de la manière suivante. À partir du smartphone, on récupère les identifiants Twitter et Facebook de l'utilisateur. Ensuite, un *webcrawler* est configuré pour effectuer sur chaque réseau un parcours en profondeur d'ordre 2 en partant du profil de l'utilisateur.

Pour évaluer l'approche, il est nécessaire de supposer qu'un sous-ensemble des liens existants n'est pas connu. Ainsi, l'efficacité de chaque algorithme repose sur la capacité à prédire de telles connexions *non observées* mais existantes. Fondamentalement, on s'attend à ce qu'une bonne classification donne plus de poids aux liens *non observés* qu'aux liens *non existants*. Dans notre cas, les liens sont des connexions qui illustrent une relation légitime entre l'utilisateur de smartphone et ses contacts. Cette relation est identifiée comme une amitié qui est certifiée comme légitime par la personne concernée et validée par un expert. Cette étape a été réalisée en interrogeant directement le propriétaire de smartphone pour étiqueter les liens comme légitimes ou non. La détection d'activités malveillantes potentielles des profils (p. ex. la propagation d'URL malveillantes) est également effectuée pour vérifier ces profils (cette vérification repose sur SPOT [128]). Si un profil est détecté malveillant, celui-ci est supprimé de la liste de contacts et considéré

Indices	Facebook		Twitter	
	échantillon 1	*échantillon 2*	*échantillon 3*	*échantillon 4*
CN	0.908	0.892	0.782	0.755
Salton	0.910	0.889	0.640	0.610
Jaccard	0.907	0.897	0.547	0.540
Sørensen	0.907	0.892	0.465	0.525
HPI	0.905	0.891	**0.935**	**0.905**
HDI	0.918	0.886	0.670	0.675
LHN	0.913	0.878	0.552	0.575
PA	0.535	0.559	0.500	0.535
AA	0.901	0.833	0.732	0.740
RA	0.886	0.843	0.715	0.745
SBRA	**0.983**	**0.921**	0.695	0.750

Tableau 4.5 – Valeurs d'AUC pour les ensembles de données (les meilleurs résultats sont affichés en gras et les moins bons sont soulignés).

comme appartenant à la liste de liens *non existants*. Un tel pré-traitement garantit que la performance de chaque approche est fondée sur sa capacité à détecter des relations légitimes. Ici, nous évaluons l'efficacité de notre indicateur SBRA à l'égard des autres mesures présentées en utilisant la valeur de l'aire sous la courbe ROC (fonction d'efficacité du récepteur) [139]. Le score AUC (Area Under the ROC Curve) peut être calculé sur la base de n comparaisons indépendantes entre un lien *non observé* et un lien *non existant*. Le nombre de fois que l'algorithme affecte à un lien *non observé* un score supérieur à un lien *non existant* est égal à n' (et à n'' s'ils ont le même score). Compte tenu de ces paramètres, l'AUC est calculé comme indiqué dans l'équation 4.3.2.

$$AUC = \frac{n' + 0.5 * n''}{n} \tag{4.3.2}$$

La qualité de la prédiction est aussi élevée que le nombre AUC est proche de la valeur 1. Un algorithme aléatoire est supposé avoir une valeur d'AUC qui est proche de 0.5. Les résultats obtenus pour les indices sont présentés dans le tableau 4.5 pour les ensembles de données Twitter et Facebook.

En ce qui concerne Facebook, nous observons que notre indice SBRA obtient les meilleurs résultats pour les deux ensembles de données. L'indicateur d'attachement préférentiel est moins efficace dans la prédiction de la relation de deux contacts sur le réseau Facebook. Cela signifie que la relation n'est pas bien décrite si uniquement le nombre de contacts des profils est pris en compte. Nous pouvons également constater que les indicateurs basés sur les voisins communs obtiennent de bonnes performances. D'après les résultats, la fonction d'imbrication permet de discriminer une paire de nœuds qui ne pourrait pas être identifiée comme légitime sur le simple nombre de voisins communs.

En ce qui concerne les ensembles de données de Twitter, nous pouvons constater que notre approche n'obtient pas les meilleurs résultats. Cela nous permet de supposer que la nature des connexions n'est pas fondée principalement sur l'identité des personnes. En effet, comme déjà indiqué, les relations sur Twitter reposent d'avantage sur l'information que sur les personnes. Nous notons que l'indice HPI est le seul indicateur qui obtient de bonnes performances. Cela signifie que les utilisateurs se connectent à des personnes qui possèdent déjà de nombreuses connexions sur le réseau. Ce constat s'explique par le

fait que, sur Twitter, le degré d'un nœuds est souvent considéré comme un niveau de prestige et de légitimité. Ce phénomène est encore soutenu par le fait que Twitter fonde son système de recommandation d'amis sur ce principe. Le travail de [140] a prouvé que les profils malveillants sur Twitter obtiennent un score élevé pour les algorithmes de centralité les plus courants (p. ex. PageRank, HITS, NodeRank). De plus, [141] a identifié que cela est principalement dû au fait que les utilisateurs légitimes suivent une grande quantité de profils malveillants. Ces observations révèlent un problème de sécurité très important. Sur la base des faits présentés, la création d'une relation sur Twitter est principalement basée sur une fonctionnalité qui n'est pas capable d'assurer la fiabilité.

Nous constatons que, pour les réseaux sociaux de navigation tels que Twitter, l'approche présentée n'obtient pas de bonnes performances. Ceci est principalement dû au fait que notre approche repose sur la modélisation d'un dispositif personnel basé sur des réseaux sociaux de socialisation (p. ex. carnet d'adresses, courriels). Une solution possible à l'identification d'une relation légitime sur Twitter pourrait être l'adaptation du calcul d'imbrication afin qu'il puisse capter le niveau d'importance d'un contenu (et non d'un utilisateur) dans le réseau ML de l'utilisateur de smartphone. Cette amélioration, et adaptation de l'approche, sera présentée dans le chapitre suivant.

Pour les réseaux sociaux axés sur l'utilisateur, l'indicateur d'imbrication semble être un outil complémentaire aux approches existantes afin de mieux distinguer les contacts légitimes et les contacts non-légitimes. Cela confirme le fait que les connexions légitimes entre les utilisateurs de Facebook sont davantage susceptibles d'apparaître s'ils partagent des amis qui ont un score d'imbrication important. En d'autres termes, les résultats illustrent le fait que les données disponibles sur le smartphone d'un utilisateur peuvent être utilisées pour discriminer les contacts légitimes et non légitimes sur les réseaux sociaux centrés sur l'utilisateur.

Nous avons appliqué le modèle multicouche sur les réseaux sociaux d'un utilisateur de smartphone. Au-delà de l'enrichissement de la vision de ses multiples aspects de la vie numérique, le travail réalisé peut contribuer à un outil d'aide à la décision pour la prévention de fuites de données.

4.4 Le prototype *SOCIALYSER 1.0*

Nous avons développé une application iOS pour illustrer la faisabilité de notre approche dans un contexte mobile. Cette application, appelée *SOCIALYSER*, fonctionne sur iPhone et iPad. Ce prototype analyse les cinq réseaux suivants : Facebook, Twitter, Google+, courriel et carnet d'adresses (dénommé « téléphone » dans le reste de cette section). L'intégration de Facebook et Twitter est simplifiée grâce à l'intégration native de ces réseaux sociaux dans iOS6. Notre application dispose de trois écrans principaux. Les multiples contacts de l'utilisateur sur les différents médias sont présentés sur le premier écran, figurant sur le côté gauche de la figure 4.4.1. L'utilisateur peut changer de réseau facilement avec l'élément de menu en haut à gauche, appelé *Switch*.

Le second écran, sur le côté droit de la figure 4.4.1, montre le résultat de l'analyse du calcul de l'imbrication (non normalisé et en mode pilier) des contacts en regard des couches que l'on souhaite prendre en compte. Par défaut, l'application affiche l'imbrication concernant toutes les couches qui contiennent une liste non nulle de contacts. L'utilisateur peut modifier ce paramètre et activer ou désactiver l'une des couches. Le troisième écran, sur le côté gauche de la figure 4.4.2, affiche le score légitimité accordé à chaque contact Facebook. Il intègre la possibilité d'afficher les amis communs et leurs scores d'imbrication

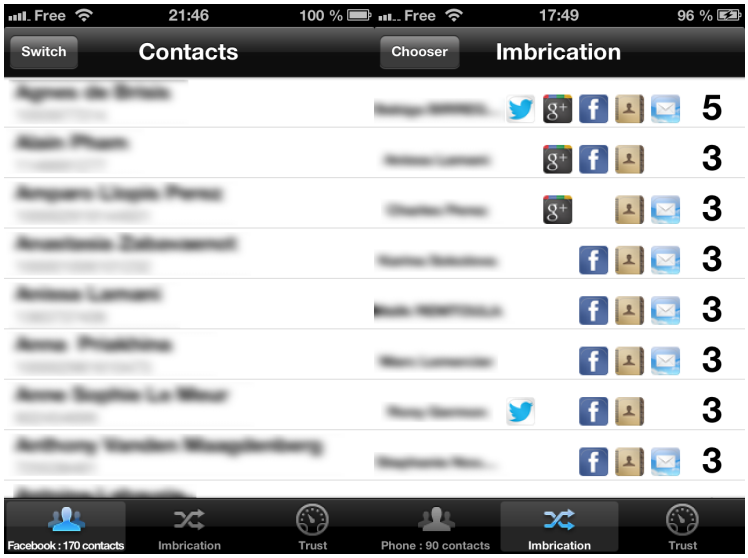


FIGURE 4.4.1 – Liste de contacts et imbrication.

FIGURE 4.4.2 – Imbrication des amis communs et mesure de légitimité.

pour un contact sélectionné. La valeur d'imbrication des amis communs est illustrée sur le côté droit de la figure 4.4.2. Enfin, pour effectuer une analyse à grande échelle, l'application permet d'exporter et d'envoyer par courriel les résultats de l'extraction.

4.5 Résultats

Nous présentons quelques résultats associés aux scores de légitimité recueillis à partir de l'application *SOCIALYSER*. Même s'ils ne sont pas nécessaires pour la validation de notre approche, ces résultats peuvent contribuer à donner un nouvel éclairage sur les utilisations des réseaux sociaux et le niveau relatif de la sécurité des données de ceux-ci [142]. Un nombre total de trente personnes ont participé à l'expérience. Tous ces utilisateurs possèdent un iPhone ou un iPad et utilisent le réseau social Facebook. Bien que non exhaustifs, ces résultats donnent un aperçu des perspectives possibles de ce travail et illustrent la faisabilité de notre approche.

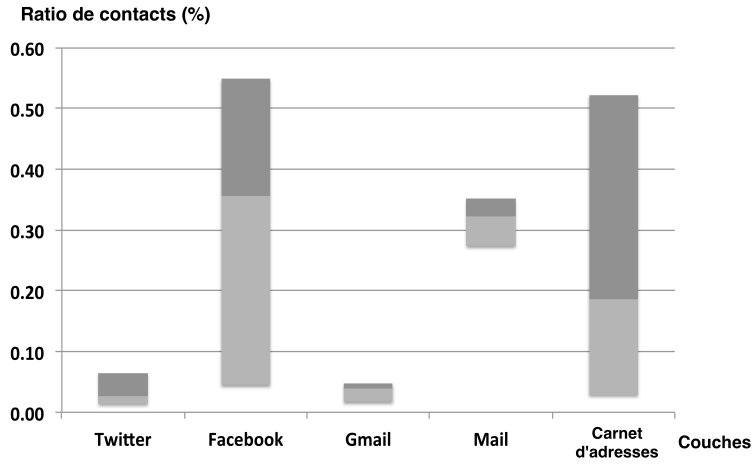


FIGURE 4.5.1 – Ratio de contacts pour chaque réseau.

La figure 4.5.1 montre la distribution moyenne des contacts sur les réseaux sociaux des utilisateurs de smartphone. Sur cette figure, on peut observer le minimum, le maximum et les scores médians pour chaque réseau. Twitter et Google+ sont les réseaux les moins représentés. Toutefois, le courrier électronique représente en moyenne 40 % des contacts. Facebook et les listes de contacts du téléphone sont généralement bien représentés, mais leur ratio est très clairsemé. Sur la base de ces observations, nous pouvons affirmer que, dans la plupart des cas, au moins trois couches (Facebook, courriel et téléphone) sont disponibles sur le smartphone. *A contrario*, les deux autres couches ne sont pas significativement représentées.

La figure 4.5.2 illustre la distribution des scores d'imbrication (non normalisé) sur les cinq réseaux. Nous avons constaté que plus de 10 % des contacts ont une imbrication supérieure à un.

Nous présentons dans le tableau 4.6, l'imbrication moyenne, entre chaque paire de réseaux.

On peut voir que Twitter obtient une imbrication faible avec tout autre réseau pris en considération. La raison principale est que Twitter n'est pas un réseau axé sur la personne, mais plutôt un système orienté vers le contenu. Ainsi, la nature du réseau est fondamentalement différente des autres strates. On peut observer que la strate téléphone

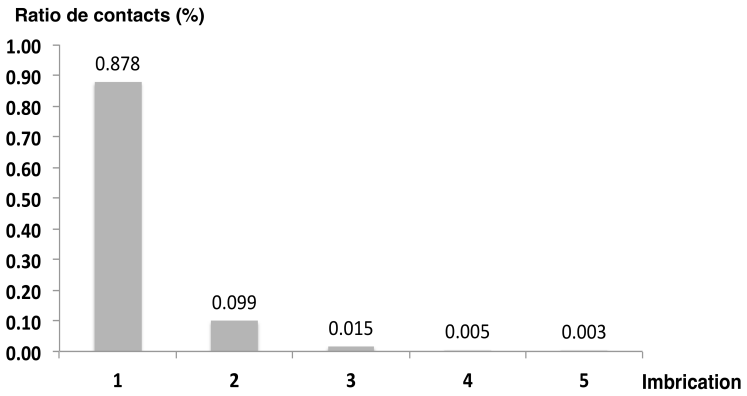


FIGURE 4.5.2 – Imbrication moyenne des profils dans les cinq couches.

	Twitter	Facebook	Gmail	Mail	Carnet d'adresses
Twitter	0.00	0.02	0.00	0.01	0.01
Facebook	0.02	0.00	0.02	0.04	0.20
Gmail	0.00	0.02	0.00	0.02	0.04
Mail	0.01	0.04	0.02	0.00	0.14
Carnet d'adresses	0.01	0.20	0.04	0.14	0.00

Tableau 4.6 – Imbrication moyenne en mode pilier entre chaque paire de réseaux

est fortement imbriquée dans les réseaux Facebook, Google+ et Mail. L'imbrication importante de Facebook dans les autres réseaux est un signe de la réussite de notre approche.

La répartition de la légitimité par rapport à la valeur maximale est illustrée en figure 4.5.3. Une telle représentation permet d'évaluer les contacts d'un utilisateur Facebook. Nous pouvons voir que près de 30 % des contacts d'une personne peuvent être considérés comme légitimes, car ils possèdent entre 70 % et 100 % de la valeur maximale de légitimité. Environ 35 % des contacts possèdent un score qui varie entre 30 % et 70 % du maximum. Ceux-ci partagent une quantité importante de contacts communs considérés comme légitimes. Environ 25 % des contacts sont situés dans la plage de 10 % à 20 %. Ce résultat prouve que l'on peut posséder un ensemble de contacts qui n'ont que peu de contacts en commun. Enfin, environ 15 % des contacts ont un très faible score de légitimité. Ces personnes figurent parmi les membres du réseau social d'une personne mais sans légitimité *a priori*. Ces contacts sont complètement déconnectés de l'interaction observée de l'utilisateur dans les quatre autres réseaux. Ces personnes peuvent accéder aux données privées et personnelles d'une personne sans réelle légitimité.

Conclusion

Dans ce chapitre, nous avons proposé une nouvelle approche pour aborder la question de l'évaluation de la fiabilité des contacts d'un utilisateur de réseau social. La méthode repose sur le graphe multicouche, qui modélise les multiples interactions d'un utilisateur de smartphone avec ses multiples médias sociaux. Un tel modèle, enrichi par les indicateurs d'imbrication, permet d'identifier les contacts importants et d'intégrer de tels contacts dans l'évaluation de la légitimité. Notre application de ce modèle à Facebook a prouvé l'efficacité de notre approche, et le déploiement de l'application *SOCIALYSER* a montré

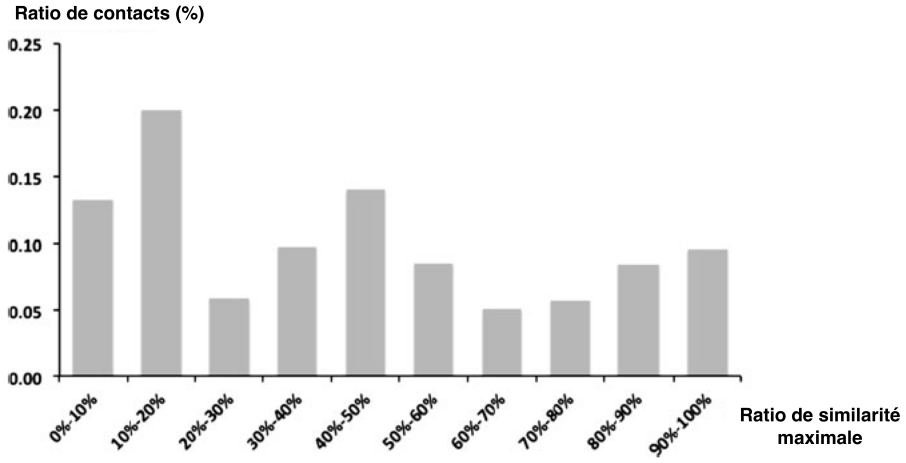


FIGURE 4.5.3 – Répartition de la légitimité des contacts.

sa faisabilité. À notre connaissance, cette approche est la première tentative d'analyse de légitimité entre utilisateurs à partir de la modélisation et de l'analyse d'un smartphone. L'intégration directe d'une telle approche dans les smartphones présente de nombreux avantages. L'évaluation du score de légitimité ne nécessite aucune participation de l'utilisateur et elle s'appuie uniquement sur les données disponibles au niveau local. L'analyse est effectuée sur le smartphone de l'utilisateur, aucune règle de confidentialité n'est violée et aucun accès privilégié à l'information n'est nécessaire.

Partie III : Vers une sécurisation personnalisée des utilisateurs de réseaux sociaux numériques

« Strong ties are the people you really trust, people whose social circle tightly overlap with your own. »

(M. Granovetter, 1973)

Chapitre 5

Approche combinée pour la protection de données sur les RSN

Résumé du chapitre

Ce chapitre présente une approche combinée des travaux de SPOT décrits lors du chapitre 3 et de la mesure de légitimité définie dans le chapitre 4. La méthodologie ainsi détaillée, nommée SPOTLIGHT, permet sur différents types de réseaux sociaux numériques (socialisation et navigation) non seulement d'évaluer le niveau de proximité pouvant être associé à un conftact d'un utilisateur, mais aussi de mesurer sa propension à se comporter de manière anormale. Nous montrons ainsi que l'apport de cette double analyse permet l'identification de contacts de confiance de manière plus performante. Cette confiance est identifiée en regard des deux principales menaces liées aux RSN. D'une part, la non-propagation de contenu malveillant par le contact analysé et, d'autre part, sa présence légitime dans la liste de contacts de l'utilisateur. L'approche est testée sur deux ensembles de données issus de Facebook et Twitter.

5.1 Cadre général

5.1.1 Contexte de l'approche

La figure 5.1.1 fournit un exemple de graphe représentant les relations sociales dans un réseau social numérique. Sur ce graphe, l'utilisateur x possède un ensemble de contacts (identifiés par des nœuds avec doubles contours) qui possèdent également un ensemble de contacts. Sur un réseau social numérique, certains utilisateurs représentent une menace (par exemple en publiant des URL malveillantes, en obtenant l'accès illégitime aux données personnelles), et sont colorés en gris dans la figure. Dans ce contexte, notre approche combinée vise à détecter et à évaluer les contacts légitimes de l'utilisateur x parmi les autres profils. En d'autres termes, deux classes de profils peuvent être prises en compte : (1) les contacts légitimes de l'utilisateur x et (2) les autres profils. Les résultats de cette approche consistent en deux indicateurs qui ont pour but d'aider l'utilisateur de smartphone (nœud central de la figure) dans l'évaluation de sa/son ensemble d'amis et de prendre des décisions appropriées en ce qui concerne les profils (qui appartiennent ou souhaitant appartenir à son cercle d'amis) qui ne sont pas légitimes ou dont le comportement est malveillant.

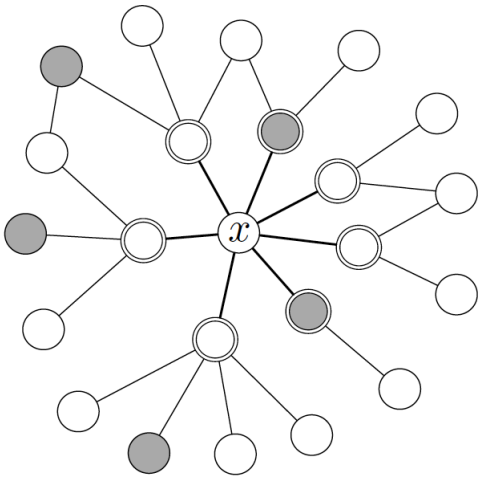


FIGURE 5.1.1 – Réseau social d'un utilisateur noté x composé de son ensemble de contacts (nœuds avec doubles contours) et de leurs contacts respectifs (nœuds avec un contour simple). Les nœuds légitimes sont colorés en blanc et les autres sont colorés en gris.

Un aperçu de l'approche SPOTLIGHT est présenté en figure 5.1.2. Nous proposons de modéliser les traces d'interactions sociales d'un smartphone sous forme de graphe multicouche où chaque couche est un média social de communication accessible à partir de l'appareil. Dans l'exemple, les médias considérés sont le carnet d'adresses, Facebook, Twitter et Google+. Ce modèle fournit des indices importants sur le comportement de l'utilisateur et permet d'évaluer les personnes et les intérêts principaux de l'utilisateur de smartphone. Ceux-ci sont ensuite intégrés dans une mesure de distance / similarité afin d'évaluer la légitimité (sur un réseau social) d'un profil y appartenant à la liste d'amis du propriétaire du smartphone (identifié par x). Une analyse comportementale complémentaire est effectuée sur la base de mesures de l'*activité* et de la *visibilité* des contacts dans le but de détecter un comportement anormal. Un classement est finalement réalisé à partir de ces deux caractéristiques principales pour identifier les profils légitimes.

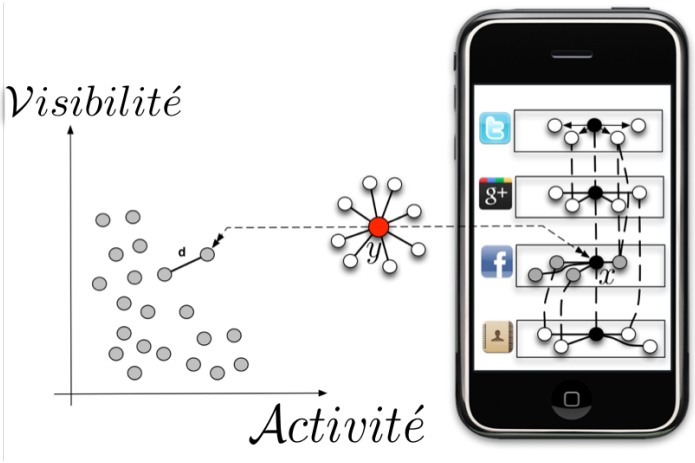


FIGURE 5.1.2 – Vue d'ensemble de la démarche SPOTLIGHT pour évaluer la fiabilité d'un profil dénoté y appartenant à l'ensemble de contacts de l'utilisateur de smartphone noté x.

5.1.2 Proposition de classification des applications sociales mobiles

La figure 5.1.3 illustre les applications sociales que l'on peut trouver sur un smartphone. Nous identifions deux principaux types d'applications sociales : (1) les médias sociaux traditionnels de communication (p. ex. appels téléphoniques et les SMS), (2) les réseaux sociaux numériques tels que Facebook, Tumblr, Google+ ou Twitter. Comme le montre la figure, nous proposons de classer les réseaux sociaux numériques en deux principaux groupes étiquetés comme $\mathcal{C}ontenu$ et $\mathcal{I}dentité$. Les réseaux de type $\mathcal{C}ontenu$ correspondent aux réseaux de navigation (ce qui importe est le contenu) et les réseaux de type $\mathcal{I}dentité$ correspondent aux réseaux de socialisation (ce qui compte sont les personnes).

Notons que, même si la frontière entre ces deux types de réseaux n'est pas toujours évidente, il est souvent possible de dégager une tendance principale. À titre d'exemple, les réseaux Pinterest et Instagram regroupent les utilisateurs autours du contenu (photos). Ceux-ci identifient des utilisateurs par rapport à l'intérêt qu'ils portent au contenu qu'ils produisent et non à qui ils sont. Pour ces raisons, ils seront classifiés comme des réseaux de type $\mathcal{C}ontenu$. Le réseau Whatsapp se rapproche dans son mode de fonctionnement aux échanges de type SMS. Ces SMS sont, dans la plupart des cas, envoyés à des personnes que l'utilisateur connait personnellement. Pour cette raison, Whatsapp sera classé dans les réseaux de type $\mathcal{I}dentité$.

Il est important de préciser que dans les perspectives de notre approche, il est envisageable de considérer des réseaux mixtes donc les coefficient plus ou moins forts seraient affectés à chaque catégorie.

Dans la suite, nous proposons une méthodologie générale pour répondre au problème de l'identification de contacts légitimes sur un réseau social numérique. Cette méthodologie utilise tous les moyens de communication disponibles sur le smartphone. De plus, elle est adaptée à la fois pour les réseaux de type $\mathcal{C}ontenu$ et $\mathcal{I}dentité$ et permet donc de répondre aux limitations observées par l'approche présentée lors du chapitre précédant.

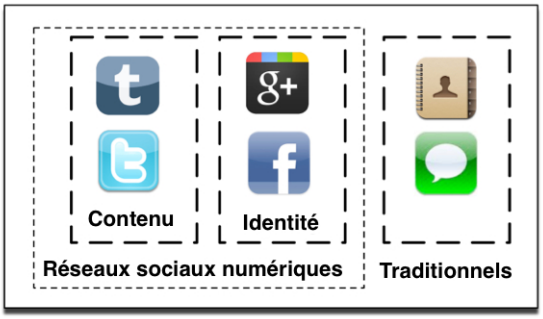


FIGURE 5.1.3 – Proposition de classification des applications sociales mobiles.

5.2 Mesure de similarité locale au smartphone

5.2.1 Quelques définitions préliminaires

Tel qu'indiqué dans l'état de l'art, le modèle multicouche (ML) peut être représenté comme un ensemble L de K couches. Chaque couche L_i est représentée comme un graphe $G_i = (N_i, A_i)$ où N_i représente l'ensemble des nœuds et A_i représente l'ensemble des arêtes. Sur la figure 5.1.2, le propriétaire du smartphone est représenté par un nœud noir et, par hypothèse, il appartient à chaque strate d'analyse.

Comme l'illustre la figure 5.2.1, nous proposons deux variantes du modèle ML lorsqu'il est appliqué à un smartphone. Alors que la première variante se concentre sur les relations entre les profils dans le but de détecter des contacts « pertinents », la seconde se concentre sur des sujets (*termes*) dans le but de détecter les centres d'intérêt « pertinents ». Les deux variantes ont pour but de fournir une mesure de similarité/distance qui exprime la proximité entre un utilisateur de smartphone x et un utilisateur y par rapport à des sujets pertinents pour les réseaux de contenu notés $\mathcal{C}ontenu$ ou des individus importants pour les réseaux de type $\mathcal{I}dentité$. Dans la suite de ce chapitre, un nœud pourra illustrer un profil utilisateur pour les réseaux de type $\mathcal{I}dentité$ mais aussi un terme pour les réseaux de type $\mathcal{C}ontenu$. Cette dualité est illustrée en figure 5.2.1.

En effet, l'ensemble des couches contient deux catégories d'applications sociales : les réseaux sociaux numériques (notés $L_{SNS} = \{Facebook, Twitter, Google+\}$) et les moyens de communication traditionnels (notés $L_{Traditionnel} = \{CarnetAdresses\}$). La méthodologie proposée repose sur les deux types de média ($L = L_{Traditionnel} \cup L_{SNS}$), mais a seulement pour but de fournir un moyen efficace pour caractériser la légitimité d'un profil dans la liste des contacts de l'utilisateur de smartphone sur un réseaux social numérique (représenté en gris sur la figure).

La détection de profils/termes clés est effectuée par la recherche de ses occurrences sur les différentes couches de l'utilisateur du smartphone. Cela se traduit par une fonction d'appariement entre deux nœuds u et v appartenant aux couches k et l notée $I(u_k, v_l)$. La définition de cette fonction d'appariement est similaire à la définition 4.1 du chapitre précédent. De même, l'identification de l'imbrication d'un nœud dans une ou plusieurs couches du modèle sont identiques aux définitions 4.3 à 4.5. Nous présentons dans la suite un exemple illustrant ces métriques pour les réseaux sociaux de type $\mathcal{C}ontenu$. La mesure permet d'identifier les termes qui sont souvent utilisés dans les messages du propriétaire de smartphone sur un réseau social, mais également sur un ensemble de réseaux.

Nous présentons un nouvel exemple en figure 5.2.2. Cette fois-ci, le réseau multicouche

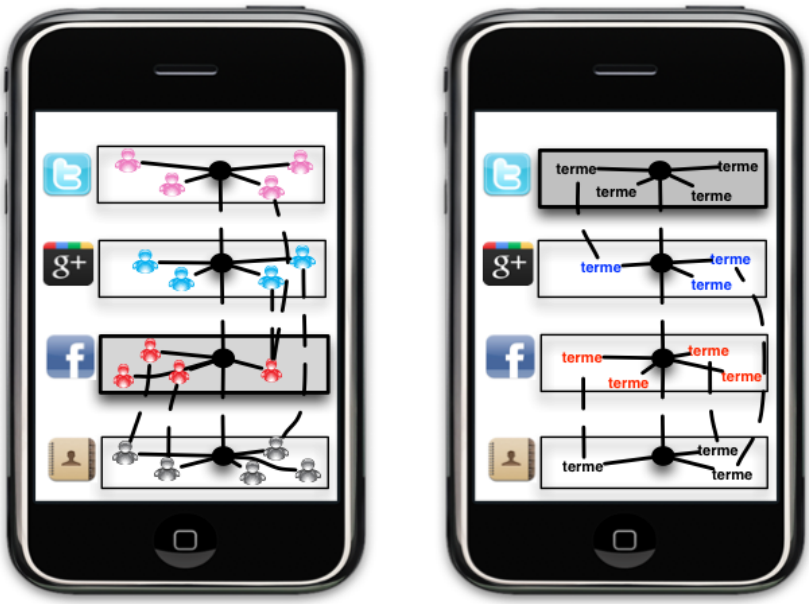


FIGURE 5.2.1 – Modèles ML de représentation du smartphone pour les réseaux de type *Identité* (gauche) et *Contenu* (droite).

a été construit non pas à partir des contacts de l'utilisateur sur chaque strate, mais partir des termes contenus dans les messages envoyés sur les différents réseaux. Nous avons identifié trois termes : $\mathcal{A}$rbre $\mathcal{B}$allon et $\mathcal{C}$hat notés respectivement $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ et ayant été publiés par l'utilisateur sur le réseau social Twitter. Nous souhaitons évaluer leur niveau d'imbrication dans le réseau multicouche de l'utilisateur.

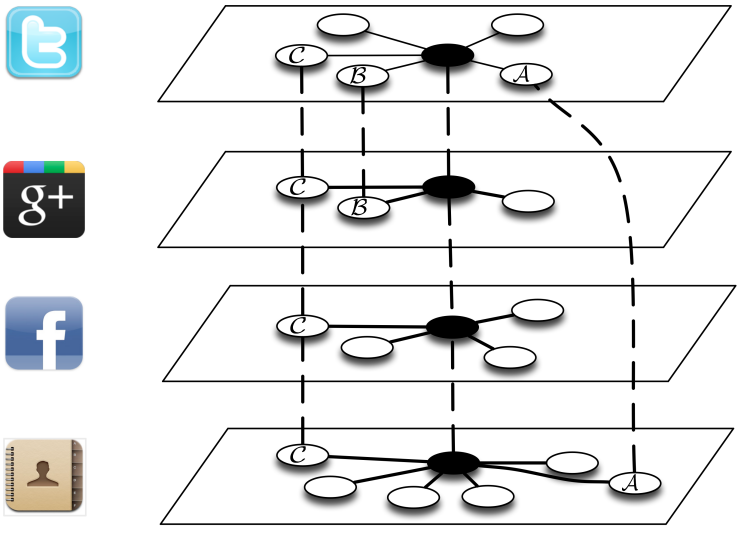


FIGURE 5.2.2 – Exemple de réseau multicouche d'un utilisateur pour la mesure d'imbrication des termes $\mathcal{A}$rbre, $\mathcal{B}$allon et $\mathcal{C}$hat.

À cet effet, nous évaluons $I(\mathcal{A}_{Twitter}, S)$, $I(\mathcal{B}_{Twitter}, S)$ et $I(\mathcal{C}_{Twitter}, S)$ avec l'ensemble des réseaux S constitué de Twitter, Facebook, Google+ et du carnet d'adresses. Cette analyse nous permet d'identifier à quel point l'utilisateur communique ou non avec les mêmes termes sur les différents canaux de communication lui étant accessibles. Nous présentons

dans les tableaux 5.1 et 5.2 les détails des opérations pour la mesure d'imbrication. Nous observons que le terme $\mathcal{C}$hat apparaît sur tous les réseaux analysés tandis que les deux autres termes n'apparaissent que dans deux réseaux. Assez logiquement ce terme obtient une valeur d'imbrication plus forte que les deux autres termes ($I(\mathcal{C}_{Twitter}, S) \approx 0.053$). D'après le calcul de l'imbrication, le terme $\mathcal{A}$rbre obtient un niveau d'imbrication plus faible ($I(\mathcal{A}_{Twitter}, S) \approx 0.030$) que le terme $\mathcal{B}$allon ($I(\mathcal{B}_{Twitter}, S) \approx 0.030$). Ceci est principalement dû au fait que ce dernier apparaît dans le réseau Google+ sur lequel l'utilisateur s'exprime sur moins de sujets. Cette importance du terme est associée à l'hypothèse suivante : si l'utilisateur s'exprime sur de nombreux sujets il est difficile de connaitre réellement ces centres-d'intérêt ; par contre s'il s'exprime sur une quantité plus faible de sujets, il est vraisemblable que ceux-ci aient plus d'importance.

	$I(\_, Twitter)$	$I(\_, Google+)$	$I(\_, Facebook)$	$I(\_, Carnet\ d'adresses)$
$\mathcal{A}$	1/5	0/3	0/4	1/6
$\mathcal{B}$	1/5	1/3	0/4	0/6
$\mathcal{C}$	1/5	1/3	1/4	1/6

Tableau 5.1 – Imbrication de $\mathcal{A}$, $\mathcal{B}$ et $\mathcal{C}$ dans chaque strate.

	Imbrication Totale	**Approximation**
$\mathcal{A}$	$(1/5+0/3+0/4+1/6)/(5+3+4+6)$	0.020
$\mathcal{B}$	$(1/5+1/3+0/4+0/6)/(5+3+4+6)$	0.030
$\mathcal{C}$	$(1/5+1/3+1/4+1/6)/(5+3+4+6)$	0.053

Tableau 5.2 – Calcul de l'imbrication de $\mathcal{A}$, $\mathcal{B}$ et $\mathcal{C}$.

Nous proposons de définir la fonction de voisinage $\Gamma_l^{\mathcal{C}}(x)$ qui renvoie l'ensemble des voisins d'un nœud x pour une couche L, sous un contexte $\mathcal{C}$ tel qu'indiqué ci-dessous.

Définition 5.1. Fonction de voisinage

La fonction de voisinage $\Gamma_l^{\mathcal{C}}(x)$ qui définit l'ensemble des voisins d'un nœud x pour une couche l, sous un contexte $\mathcal{C}$ est définie comme suit :
$\forall \mathcal{C} \in \{\mathcal{I}dentité, \mathcal{C}ontenu\}$;

$$\Gamma_l^{\mathcal{C}}(x) = \{w \in N_l | E(x, w) \in E\}$$

Tel qu'illustré dans la figure 5.2.1, w peut être soit un profil ($\mathcal{C} = \mathcal{I}dentité$) soit un terme pertinent à l'utilisateur ($\mathcal{C} = \mathcal{C}ontenu$).

Nous noterons dans la suite par $l_{\mathcal{C}}$, la projection d'un réseau social l dans le contexte associé $\mathcal{C}$. Notons que, par nature un réseau social de type contenu tel que Twitter est considéré de contexte $Contenu$ et un réseau de type Facebook est considéré de contexte $\mathcal{I}dentité$. Pour mesurer la similarité de deux utilisateurs Twitter, nous utiliserons la variante $Contenu$ et pour mesurer la similarité entre deux utilisateurs Facebook nous utiliserons la variante $\mathcal{I}dentité$. Pour un réseau social, il convient d'avoir une bonne connaissance de celui-ci afin d'utiliser la variante de notre modèle la mieux adaptée.

Bien qu'il soit tout à fait envisageable de considérer un réseau comme étant à la fois un réseau de contextes $Contenu$ et $\mathcal{I}dentité$, nous considérons dans ce travail qu'un réseau social n'est associé qu'à un unique contexte.

Nous proposons de mesurer le degré de « similitude » entre deux entités comme suit :

Définition 5.2. Mesure de similarité entre les entités
La similarité entre un nœud x et un nœud y sur un réseau social l de contexte $\mathcal{C}$ est définie par :

$$s_{xy}^{SBRA}(l_{\mathcal{C}}) = \sum_{z \epsilon \Gamma_l^{\mathcal{C}}(x) \cap \Gamma_l^{\mathcal{C}}(y)} \frac{I(z,S)}{popularité_{\mathcal{C}}(z)} \tag{5.2.1}$$

Avec :
l est le réseau social considéré,
S est un ensemble de réseaux sociaux,
$I(z,S)$ est le chevauchement de z dans un ensemble de réseaux S,
y l'utilisateur évalué,
x le propriétaire du smartphone,
$\mathcal{C}$ est le contexte,
$popularité_{\mathcal{C}}(z)$ est la popularité d'une entité z sous un contexte $\mathcal{C}$,
$\Gamma_l^{\mathcal{C}}(x)$ est le voisinage de l'utilisateur x pour le contexte $\mathcal{C}$.

La formule 5.2.1 est une variante de l'indice de similarité d'allocation des ressources [103] qui est souvent utilisé pour la prédiction de liens dans les réseaux sociaux. La fonction proposée ajoute pour chaque entité commune z, le score de chevauchement de cette entité dans le smartphone divisé par son score de popularité qui dépend de la variante du modèle. L'objectif est d'affecter plus d'importance à des entités communes qui sont significatives (elles ne sont pas si communes sur la plateforme).

5.2.2 La mesure de similarité et les algorithmes associés

Nous proposons de définir la mesure de similarité entre un utilisateur de smartphone dénoté x et un de ses contacts dénoté y sur un réseau social l de contexte $\mathcal{C}$ tel qu'indiqué dans les définitions 5.3 et 5.4.

Définition 5.3. Similarité entre les entités pour les réseaux de type identité

Dans le cadre des réseaux de type $\mathcal{I}dentité$, la similarité entre un nœud x et un nœud y sur un réseau social l de contexte $\mathcal{C}$ est définie comme suit :

$$s_{xy}^{SBRA}(l_{\mathcal{I}dentité}) = \sum_{z \epsilon \Gamma_l^{\mathcal{I}dentité}(x) \cap \Gamma_l^{\mathcal{I}dentité}(y)} \frac{I(z,S)}{|\Gamma_z|} \tag{5.2.2}$$

Avec :
S est un ensemble de réseaux sociaux,
$I(z,S)$ est le chevauchement de z dans un ensemble de réseaux S,
y est l'utilisateur évalué,
x est le propriétaire du smartphone,
$\Gamma_l^{\mathcal{I}dentité}(x)$ le voisinage de l'utilisateur x pour le contexte $\mathcal{I}dentité$.

Dans ce cas, le score de popularité dépend du nombre et de la qualité des amis communs. La mesure affecte plus d'importance aux voisins communs qui ont un fort chevauchement dans le smartphone et qui sélectionnent soigneusement leurs amis (faible nombre d'amis) qu'à des profils qui ont un faible score de chevauchement et qui peuvent se créer une quantité importante d'amis sur le réseau social.

Définition 5.4. Similarité entre les entités pour les réseaux de type contenu

Dans le cas du contexte $\mathcal{C}ontenu$, la similarité entre un nœud x et un nœud y sur un réseau social l de contexte $\mathcal{C}$ est définie comme suit :

$$s_{xy}^{SBRA}(l_{\mathcal{C}ontenu}) = \sum_{z \in \Gamma_l^{\mathcal{C}ontenu}(x) \cap \Gamma_l^{\mathcal{C}ontenu}(y)} \frac{I(z, L)}{1 + |\{m : z \in m\}|} \tag{5.2.3}$$

Avec :

L est un ensemble de réseaux sociaux,
$I(z, S)$ est le chevauchement de z dans un ensemble de couches S,
y est l'utilisateur évalué,
x est le propriétaire du smartphone,
$\Gamma_l^{\mathcal{C}}(x)$ le voisin de l'utilisateur x pour le contexte $\mathcal{C}ontenu$,
m est un ensemble de messages de la plateforme.

Le score de popularité d'un terme z est mesuré par le nombre de messages qui contiennent ce sujet sur un échantillon de la plateforme sociale. La similitude entre deux profils augmente quand ils partagent un grand nombre de sujets communs, que ces sujets sont pertinents pour l'utilisateur (chevauchement élevés) et que ces sujets sont rarement utilisés sur la plateforme.

Prenons l'exemple de la figure 5.2.3 pour illustrer le calcul de la mesure de similarité entre l'utilisateur de smartphone x et trois profils Twitter notés y_1, y_2 et y_3.

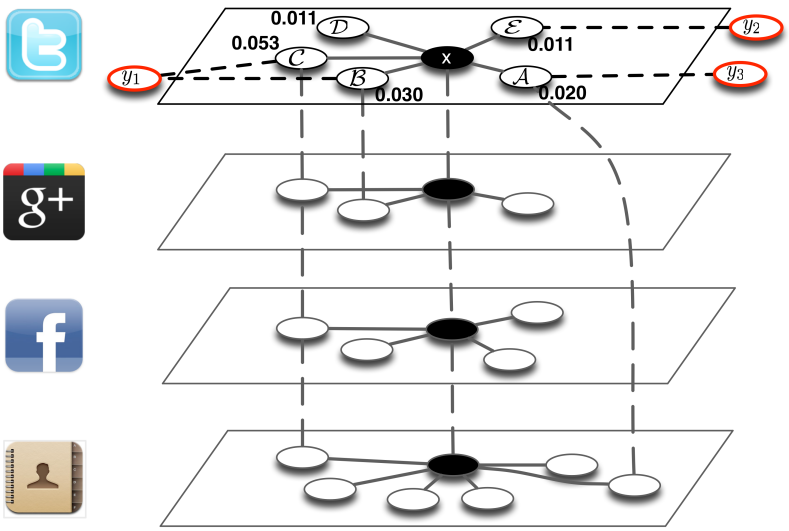


FIGURE 5.2.3 – Mesure de similarité entre l'utilisateur x et les utilisateurs de Twitter notés y_1,y_2, y_3.

Twitter étant un réseau de type $\mathcal{C}ontenu$, nous utilisons donc le contexte associé de même type. Sur la figure cela se traduit par le fait que seuls les nœuds x, y_1, y_2, y_3

représentent des utilisateurs. Les autres nœuds représentent des termes extraits à partir des interactions. Le résultat de l'imbrication des nœuds voisins de x dans l'ensemble des réseaux disponibles est indiqué à proximité de leur position.

Concernant les mesures popularité d'un terme, des valeurs fictives sont indiquées dans le tableau 5.3 et seront utilisées comme référence pour notre exemple.

Valeurs de z	**Popularité** $(1 + \lvert\{m : z \in m\}\rvert)$
$\mathcal{A}$	15
$\mathcal{B}$	2
$\mathcal{C}$	4
$\mathcal{E}$	5

Tableau 5.3 – Nombre de messages contenant les termes $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{E}$ sur un ensemble de messages de la plateforme.

La similarité entre l'utilisateur et le nœud y_1 est assez logiquement la plus forte. Ce qui est notamment dû au fait que ces deux utilisateurs partagent en commun dans leurs interactions les termes $\mathcal{B}$ et $\mathcal{C}$. De plus, ces deux termes ont une imbrication forte dans le réseau multicouche de l'utilisateur ce qui révèle l'importance de ceux-ci. Enfin, la faible popularité de ces deux termes sur la plateforme Twitter contribue aussi à ce score de similarité. En ce qui concerne les nœuds y_2 et y_3, tous les deux comportent uniquement un terme en commun avec l'utilisateur. Une simple analyse locale permet d'identifier que le terme $\mathcal{A}$ est plus imbriqué que le terme $\mathcal{E}$ dans le réseau local de l'utilisateur. Cependant la forte popularité de $\mathcal{A}$ sur la plateforme indique l'aspect non spécifique de celui-ci. D'après la formule 5.2.3, le nœud y_2 obtient une similarité plus forte que le nœud y_3.

	y_1	y_2	y_3
$\Gamma^{\mathcal{C}ontenu}(x) \cap \Gamma^{\mathcal{C}ontenu}(\_)$	$\{\mathcal{B}, \mathcal{C}\}$	$\{\mathcal{E}\}$	$\{\mathcal{A}\}$
$\frac{I(z,L)}{1+\lvert\{m:z\in m\}\rvert}$	$0.030/2 + 0.053/4$	$0.011/5$	$0.020/15$
$s_{x\_}^{SBRA}(\mathcal{C}ontenu)$	0.0282	0.0022	0.0013

Tableau 5.4 – Calcul de la similarité entre x et y_1, y_2 et y_3 sur Facebook.

L'algorithme de la figure 5.1 détaille le calcul cette mesure de similarité pour un utilisateur de smartphone.

Tout d'abord, les voisins des utilisateurs sont collectés pour chaque média disponible sur le smartphone (lignes 1 à 3). Ensuite, chaque voisin de l'utilisateur est examiné pour détecter leur présence potentielle dans l'ensemble des couches. Chaque fois qu'un profil est détecté dans une couche le score de chevauchement est mis à jour (ligne 7). Enfin, comme dernière étape, la mesure de similarité est calculée sur la base des voisins communs et sur leurs chevauchements respectifs (ligne 12). L'algorithme renvoie l'ensemble de contacts triés en fonction de leur similarité avec l'utilisateur de smartphone (ligne 14).

Algorithme 5.1 Algorithme de calcul de la similarité entre l'utilisateur d'un téléphone intelligent noté x et un ensemble de profils testés $\mathcal{P}$ sur un réseau social l

Entrée: x // Utilisateur de smartphone
Entrée: L // L'ensemble de strates
Entrée: $l \in L_{SNS}$ // Un réseau social numérique de contexte $\mathcal{C}$
Entrée: $\mathcal{P}$ // Un ensemble de profils évalués tels que
$\mathcal{P} = \{p : p \in \Gamma_l^{\mathcal{C}ontext}(x)\}$
Sortie: $\mathcal{P}$ // Liste des utilisateurs évalués ordonnés par valeur de similarité

pour tout réseau $l^{'} \in L$ **faire**
  Construire le graphe $G_{l'}$ en collectant $\Gamma_{l'}^{\mathcal{C}}(x)$
fin pour
pour tout profil $p \in \Gamma_l^{\mathcal{C}ontexte}(x)$ **faire**
  pour tout $u_k \in l_k$ **faire**
    $I(p, L) = I(p, L) + I(p, u_k)$
  fin pour
fin pour
pour tout profil $p \in \mathcal{P}$ **faire**
  $s_{xp}^{SBRA}(\mathcal{C}) = \sum_{z \in \Gamma_l^{\mathcal{C}}(x) \cap \Gamma_l^{\mathcal{C}}(p)} \frac{I(z,L)}{popularité(z)}$
fin pour
renvoyer $\mathcal{P}$ // Liste de profils ordonnée en fonction des scores s_{xp}^{SBRA}

5.3 La modélisation du comportement de l'utilisateur

5.3.1 La mesure du score d'*activité* et de *visibilité*

Nous proposons de modéliser le comportement d'un utilisateur de réseaux sociaux numériques par deux caractéristiques nommées *activité* et *visibilité*. Afin de maintenir la possibilité d'intégrer notre approche dans un smartphone, nous évaluons uniquement ces deux indicateurs basés sur les messages disponibles localement sur l'appareil. Pour les couches qui correspondent à des médias traditionnels de communication (c.-à-d. $L_{Traditionnel}$), les messages sont envoyés et reçus à partir du dispositif mobile et sont donc disponibles localement. En ce qui concerne les sites de réseaux sociaux (c.-à-d. L_{SNS}), les messages sont également disponibles car ils sont au cœur de l'utilisation. Cette représentation du comportement des utilisateurs est détaillée ici pour une application sur les deux principaux types de réseaux sociaux numériques $\mathcal{I}dentité$ et $\mathcal{C}ontenu$.

L'*activité* d'un profil est mesurée par la quantité d'actions qu'il effectue sur le réseau social en ligne au cours d'une période de temps. Des exemples d'actions que l'on peut identifier à partir de messages sont : l'envoi d'un message, l'ajout d'un contact, *tagger* quelqu'un dans une photo, commenter un statut, etc. Il est important de noter que nous ne considérons pas les actions impliquant l'utilisateur en tant qu'*activité* si il/elle n'est pas à l'origine de celle-ci (p. ex. *être taggé dans une photo*, *recevoir un message*).

Définition 5.5. Activité d'un profil sur les réseaux de contexte $\mathcal{C}ontenu$ et $\mathcal{I}dentité$

L'*activité* d'un profil p au cours d'une période notée T_i est définie comme suit :

$$A_{T_i}^{p}(\mathcal{I}dentité) = |Messages_{T_i}(p)| + |Connexions_{T_i}(p)| \tag{5.3.1}$$

et :

$$A^p_{T_i}(\mathcal{C}ontenu) = A^p_{T_i}(\mathcal{I}dentité)$$

où :

$Messages_{T_i}(p)$ est l'ensemble des messages envoyés par p pendant une période de temps T_i.

$Connexions_{T_i}(p)$ est l'ensemble des connexions créées par un profil p pendant une période de temps T_i.

Notons que selon le réseau social, le nombre de connexions créées peut ou non être accessibles localement sous forme de messages. Par exemple, sur Facebook, lorsqu'un utilisateur crée une ou plusieurs nouvelles connexions, un message apparaît automatiquement dans sa *timeline*. Ce message est prototypé de la façon suivante : *"Profil_1 est maintenant connecté à profil_2 et N autres personnes"*.

L'analyse de ces messages peut permettre d'aborder la mesure de l'activité d'un profil pour identifier que $N + 1$ connexions ont été créées par *Profil_1*.

La *visibilité* d'un profil désigne le niveau de visibilité que le profil possède (lorsque mesurable) ou cherche à produire (dans le cas contraire) sur un réseau social numérique. Sur les réseaux de type *Identité*, les messages du mur d'un individu permettent une mesure directe de la visibilité car toute action impliquant l'utilisateur apparaît dans sa *timeline* (p. ex. a été identifiée sur la photo, sur le message posté par). Une analyse du contenu des messages permet de distinguer les messages qui sont envoyés par l'utilisateur (*activité*) de ceux qui impliquent l'utilisateur (*visibilité*).

Définition 5.6. Visibilité d'un profil (réseaux basés sur l'identité)

La visibilité d'un profil p pendant une période de temps T_i est définie sur les réseaux de type *Identité* comme suit :

$$V^p_{T_i}(\mathcal{I}dentité) = |Mur_{T_i}(p)| - |Messages_{T_i}(p)| \tag{5.3.2}$$

où :

$Messages_{T_i}(p)$ désigne l'ensemble des messages envoyés par p pendant la période de temps T_i.

$Mur_{T_i}(p)$ désigne l'ensemble de tous les messages qui apparaissent sur le mur de p pendant la période de temps T_i.

Sur les réseaux de type *Contenu*, une mesure directe de la visibilité n'est souvent pas disponible à partir des messages. Cependant, il est possible de mesurer la quantité d'action que l'utilisateur effectue pour rendre ses messages visibles aux autres. La visibilité d'un message est alors liée à la quantité de références et de *hashtags* qu'il contient. Sur la plupart des réseaux de type *Contenu* en ajoutant une référence à un message, celui-ci est automatiquement reçu par le profil indiqué. De même, l'ajout d'un *hashtag* dans un message le rend visible par une communauté qui adhère à ce sujet ou par les recherches associées à ce tag sur le réseau.

Définition 5.7. Visibilité d'un profil (réseaux basés sur le contenu)

La visibilité d'un profil p pendant une période de temps T_i est définie sur les réseaux de type $\mathcal{C}ontenu$ comme suit :

$$V^p_{T_i}(\mathcal{C}ontenu) = |Hashtags_{T_i}| + |Références_{T_i}| \tag{5.3.3}$$

où :

$Hashtags_{T_i}(p)$ désigne l'ensemble des *hashtags* envoyés par p pendant la période de temps T_i.

$Références_{T_i}(p)$ désigne l'ensemble des références envoyées par p pendant la période de temps T_i.

Notons que ces définitions diffèrent légèrement par rapport aux indicateurs d'activité et de visibilité présentés lors du chapitre 4. D'une part, les scores étant dynamiques ceux-ci ne sont pas normalisés mais identifiés pour une période de temps. D'autre part, les contraintes associées par l'application de ces mesures aux deux types de réseaux et dans un cadre mobile nous ont conduits à opérer quelques modifications.

5.3.2 La mesure du score d'anomalie

Notre objectif est de fournir un score d'anomalie, à un moment donné T_i, qui tienne compte de l'*activité* et de la *visibilité* d'un profil pour cet intervalle de temps, mais aussi de son comportement plus ancien. La valeur d'anomalie d'un profil pour chaque période de temps est définit par la distance Euclidienne de ce profil à son K-plus proche voisin dans l'espace à deux dimensions (activité et visibilité). Notons que la position de chaque profil dans le repère *activité* et de la *visibilité* est définit par la valeur de la moyenne mobile selon les deux indicateurs. La formule de la moyenne mobile d'un l'indicateur $I \in \{A, V\}$ est définie comme suit :

$$\tilde{I}^p_{T_i} = \frac{1}{H} \sum_{k=0}^{H-1} I^p_{T_{i-k}}$$

où :

H est la longueur de l'historique qui est prise en compte

L'algorithme 5.2 décrit le mécanisme requis pour évaluer ce score. L'ensemble des profils $\mathcal{P}$ qui sont connectés à l'utilisateur de smartphone x sur un réseau social dénoté l, sont pris en considération pour l'analyse. Le paramètre H se réfère à la longueur (en jours) de l'historique qui est considéré pour effectuer la moyenne mobile des caractéristiques comportementales. Le paramètre T_i indique la valeur actuelle du pas de temps. La valeur retournée par l'algorithme est la liste des utilisateurs évalués selon le classement par score d'anomalie. L'algorithme procède comme suit, à chaque pas de temps, les messages des profils testés sont récoltées et les valeurs d'*activité* et de *visibilité* sont mesurées (lignes 3-4). Celles-ci sont utilisées pour évaluer la moyenne mobile de l'*activité* et la *visibilité* (ligne 6), qui sont nécessaires pour calculer le score d'anomalie (ligne 7). Ce score est mis à jour à chaque pas de temps et une liste triée est restituée.

La figure 5.3.1 présente un exemple de représentation des profils de la plateforme Twitter dans le repère à deux dimensions. À chaque période de temps (ici chaque jour), les mesures d'*activité* et de *visibilité* de chaque profil sont mesurées puis utilisées pour évaluer la nouvelle valeur de la moyenne mobile du profil selon ces deux dimensions. Chaque profil,

Algorithme 5.2 Algorithme de calcul du score d'anomalie de l'ensemble des profils $\mathcal{P}$ testés sur un réseau social

Entrée: x // Utilisateur de smartphone
Entrée: $l \in L_{SNS}$ //Un réseau social numérique de contexte $\mathcal{C} \in \{\mathcal{C}ontenu, \mathcal{I}dentité\}$
Entrée: H // Longueur de l'historique pris en compte dans l'analyse
Entrée: $\mathcal{P}$ // Un ensemble de profils évalués tels que
$\mathcal{P} = \{p : p \in \Gamma_l^{\mathcal{C}}(x)\}$
Entrée: i // Pas de temps numéro i dénoté T_i
Sortie: $\mathcal{P}$ // Liste des profils évalués et triés par score d'anomalie

```
pour i = 0 to t faire
    pour tout profil p ∈ Γ_l^C(x) faire
        Collecter Messages_{T_i}(p) et Mur_{T_i}(p)
        Calculer A^p_{T_i}(C) and V^p_{T_i}(C))
    fin pour
    Calculer Ã^p_{T_t}(C) et Ṽ^p_{T_t}(C)
    Calculer la distance Euclidienne au K^ème plus proche voisin
fin pour
renvoyer P // Liste des profils évalués triés par valeur d'anomalie
```

représenté par un point bleu sur la figure, se déplace donc au cours du temps. Nous avons illustré par des flèches les distances de deux profils p_1 et p_2 à leur plus proche (1-NN) et second plus proche voisin (2-NN). Le profil p_2 apparaissant dans une zone dense tandis que le profil p_1 apparaissant dans une zone peu dense, la valeur d'anomalie propre à p_1 sera plus grande que la valeur affectée à p_2.

5.4 Mise en œuvre et résultats

Nous avons testé notre méthodologie sur un ensemble d'utilisateurs de smartphone. L'objectif étant, à partir du graphe multicouche (intégrant les couches Mail, carnet d'adresses, Google+, Twitter et Facebook), d'évaluer les contacts appartenant aux réseaux sociaux numériques Facebook et Twitter. Un contact fiable est identifié comme un profil qui n'effectue pas d'action malveillante (p. ex. les attaques de *phishing*) et qui est identifié par l'utilisateur comme légitime dans sa liste de contacts. Tout profil qui soit : (1) ne fait pas partie des contacts de l'utilisateur soit (2) est détecté comme auteur de *phishing*, soit (3) est identifié comme non légitime par l'utilisateur lui-même sera considéré comme non fiable. Le point (2) est vérifié par le travail d'experts et par l'algorithme SPOT [128] (cf. chapitre 4). La performance de notre approche repose alors sur sa capacité à identifier les contacts légitimes des autres (c.-à-d. ceux satisfaisant au moins un critère identifié précédemment).

Pour chaque utilisateur, les données nécessaires à l'analyse sont : (1) le réseau multicouche de l'utilisateur ; (2) un échantillon de son graphe égocentrique Twitter ; (3) un échantillon de son graphe égocentrique Facebook.

Tandis que les points (2) et (3) peuvent être obtenus via les plateformes considérés en utilisant les APIs ou un webcrawler, le point (1) nécessite une analyse locale du smartphone de l'utilisateur. Pour cela, nous avons utilisé trois méthodes. La première est l'extraction avec l'accord de l'utilisateur, du contenu de son smartphone et ce via des outils d'analyse forensique. Un inconvénient est la nécessité d'avoir un accès direct à l'appareil.

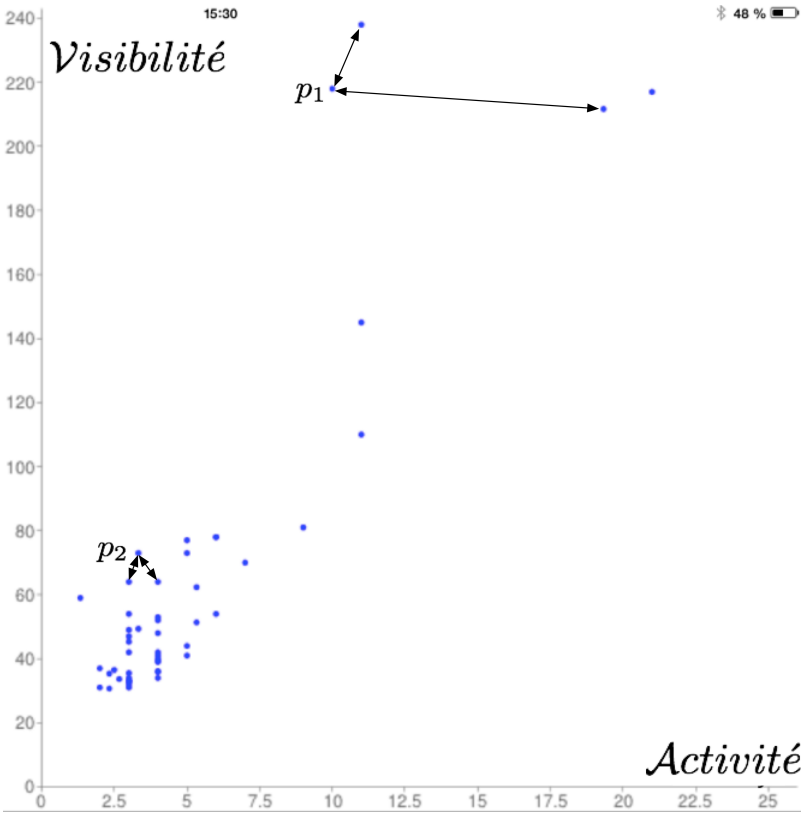


FIGURE 5.3.1 – Représentation du score d'anomalie des profils dans le repère *activité* et la *visibilité*

Il est aussi possible de récupérer et d'analyser les fichiers de synchronisation de l'appareil avec l'ordinateur de l'utilisateur. Celui-ci doit donc identifier les fichiers de synchronisation et les transmettre pour analyse. Enfin, nous avons réalisé une application mobile SPOTLIGHT qui sert non seulement de prototype à notre méthodologie mais qui permet d'exporter le graphe multicouche et son envoi par courriel.

5.4.1 Construction du graphe multicouche

La construction du modèle multicouche, repose principalement sur l'identification des liens inter-couches. Comme indiqué dans l'état de l'art, il existe un grand nombre d'algorithmes pour l'identification de ces liens. À titre d'exemple, nous avons présenté en figure 5.4.1, les performances obtenues par cinq algorithmes pour l'identification des connexions entre des contacts issues du carnet d'adresses et des profils Facebook. Dans notre cas, ce sont les couples (noms, prénoms) de profils qui servent à l'identification. Pour évaluer les performances, nous avons manuellement constitué des ensembles de paires de noms et prénoms qui correspondent aux mêmes utilisateurs et d'autres qui ne correspondent pas. Étant donné l'importance accordée par notre méthodologie aux connexions inter-couche qui sont utilisées pour évaluer la légitimité ; nous ne pouvons pas tolérer un taux de faux positifs élevé car ceci engendrerait à tort une augmentation de la valeur d'imbrication de certains contacts. Nous avons donc fixé un seuil maximal à ne pas dépasser pour le taux de faux positif. Ce seuil a été fixé à 0.0001 pour assurer qu'il n'existe pas plus d'une erreur de correspondance identifiée à tort lors de la comparaison de 100 contacts d'une strate avec 100 contacts d'une autre strate. Étant donné cette contrainte, les taux maximum de

détection de correspondances par les différents algorithmes sont présentés en figure 5.4.1. Des algorithmes testés, seul SOUNDEX, l'algorithme phonétique de correspondance n'a pas satisfait les contraintes imposées. En effet, celui-ci obtient un taux de faux positifs égal à 0.002, soit 20 fois supérieur à celui que nous avons fixé. Notons cependant que malgré ce critère d'irrecevabilité, le taux de détection de cet algorithme était supérieur à 90 %.

L'utilisation d'une stricte égalité comme identification de correspondances permet d'obtenir un taux de vrais positifs de 73 %, et par construction un taux de faux positifs nul (Il n'y avait pas d'homonymes dans l'ensemble de données qui nous a servi d'échantillon).

L'ensemble des autres algorithmes analysés obtiennent une valeur maximale de vrais positifs (indiqué sur la figure) lorsque le critère de faux positifs est tout juste atteint. Conformément aux expériences passées l'algorithme de Jaro-Winkler obtient le meilleur taux de détection (88 %). Suivent ensuite, la similarité basée sur les 3-grammes (85 %), la distance de Levenstein (84 %) et enfin l'algorithme basé sur les 2-grammes (84 %).

Notons qu'aucun algorithme ne permet d'obtenir une identification quasi parfaite des correspondances. Cette observation est due, en partie, à la présence de couples de nom et prénom incomplets dans le carnet d'adresses (p. ex. « mon conseillé » au lieu du nom et du prénom) mais aussi à la possible utilisation de comptes Facebook avec un nom incorrect (p. ex. ludovic_de_france au lieu de nom, prénom). La correspondance est connue de l'utilisateur mais ne peut être identifiée par les algorithmes. Dans dans de tels cas, il peut être possible pour l'utilisateur de renseigner manuellement une correspondance notamment via les applications mobiles de gestion de contacts (lorsque celles-ci sont associées aux réseaux sociaux).

Nous avons finalement retenu l'algorithme de Jaro-Winkler pour tester notre méthodologie car celui-ci obtient de meilleures performances que les algorithmes concurrents sur les différents ensembles des données que nous avons testés. De plus, celui-ci permet de respecter la contrainte fixée pour les faux positifs.

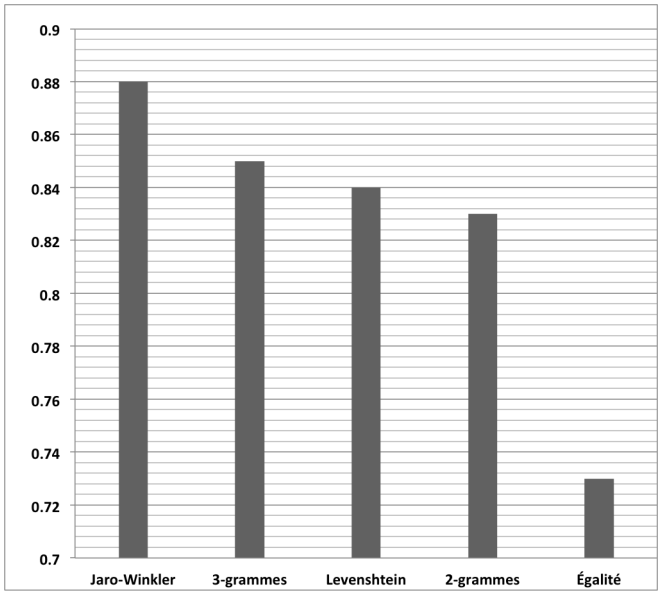


FIGURE 5.4.1 – Performances des algorithmes d'identification des connexions inter-couches sur un ensemble de contacts de Facebook et du carnet d'adresses.

5.4.2 Aperçu de quelques échantillons tests de Twitter et Facebook

Nous avons testé notre approche sur les deux réseaux sociaux que sont Facebook et Twitter et qui appartiennent respectivement aux réseaux de type *Identité* et *Contenu*. Les caractéristiques des graphes sont détaillées dans le tableau 5.5.

Nous observons que le degré moyen est plus élevé pour Twitter que pour Facebook. Cela signifie qu'un profil sur Twitter possède en moyenne un nombre moyen d'amis qui est supérieur à un profil Facebook. Nous constatons que le coefficient de *clustering* sur Twitter est très élevé par rapport à un graphe aléatoire d'Erdős–Rényi équivalent ($0.247 \gg 0.0001$) et que sa distance moyenne entre deux nœuds est relativement faible (elle ne dépasse pas le logarithme du nombre de nœuds ($3.951 < log(15\,088)$). Le graphe social de Twitter présentent les caractéristiques d'un petit monde. Les mêmes observations peuvent être faites pour l'ensemble de données de Facebook qui respecte également les conditions pour être un petit monde.

Mesure	**Twitter**	**Facebook**
Nombre de nœuds	15 088	35 269
Nombre de liens	17 187	39 798
Degré moyen	2.278	1.128
Distance moyenne	3.951	5.249
Coefficient de *clustering*	0.247	0.025

Tableau 5.5 : Caractéristiques des graphes sociaux de Twitter et Facebook.

La figure 5.4.2 illustre les boîtes de Tukey de l'*activité* et de la *visibilité* pour un ensemble d'utilisateurs de Twitter et de Facebook sur la période d'une journée. Ces résultats ne représentent que les états actifs et montrent que dans l'ensemble, le niveau d'activité d'un profil est plus élevé sur Twitter que sur Facebook. En d'autres termes, en un jour donné, un profil Twitter génère plus de messages qu'un profil Facebook. Cette observation peut révéler le fait qu'un *tweet* est un message court qui peut être envoyé en quelques secondes tandis que les messages de Facebook peuvent être plus personnels et contenir plus de caractères. Il est également important de noter que seuls les messages publics sont comptés dans l'analyse, ce qui signifie qu'une conversation directe avec un contact n'est pas comptée dans le calcul de l'*activité*. Nous notons que la visibilité d'un profil Twitter est relativement étalée. Cette observation révèle le fait qu'un *tweet* unique peut contenir plusieurs entités qui permettent au message d'accroître en visibilité. En outre, comme indiqué dans [128], de nombreux profils Twitter malveillants utilisent les *hashtags* pour promouvoir leurs *tweets*. Nous observons que, dans l'ensemble, l'*activité* et la *visibilité* d'un profil Facebook se situent entre les bornes un et trois. Ceci exprime le fait qu'un profil Facebook envoie rarement plus de trois messages par jour, et que ces messages provoquent rarement la création de plus de trois messages.

Les occurrences des couples *activité/ visibilité* pour un jour donné sont représentées en figures 5.4.3 et 5.4.4. On peut observer que, sur les deux réseaux sociaux, les couples les plus fréquents se produisent pour de faibles valeurs de l'*activité* et de la *visibilité*. Cela révèle le fait que la plupart des profils ne sont pas actifs tous les jours sur les réseaux sociaux et que quand ils sont actifs ils produisent seulement quelques messages. Nous notons quelques valeurs très élevées de *visibilité* sur Facebook ce qui peut exprimer un message controversé ou un événement particulier (p. ex. les anniversaires). Sur Twitter,

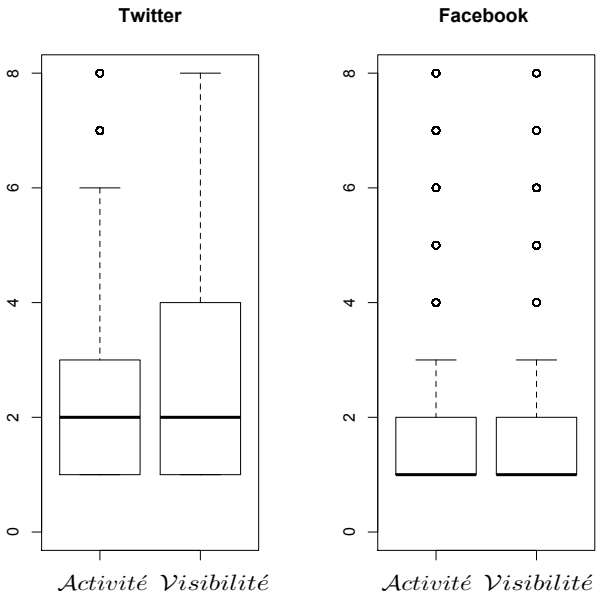


Figure 5.4.2 : Boîtes de Tukey de l'*activité* et de la *visibilité* des utilisateurs sur Twitter et Facebook sur la période d'une journée.

des valeurs élevées de *visibilité* expriment l'utilisation d'un grand nombre de *tweets* qui chacun contiennent un grand nombre d'entités.

La figure 5.4.5 illustre pour un utilisateur de smartphone, les valeurs d'anomalies et de similarité pour chacun des contacts. Chaque carré représente un profil qui appartient au réseau social Facebook d'un utilisateur de smartphone. On peut observer que la plupart des profils ont un faible score d'anomalie et de similarité. Ces profils sont généralement non malveillants mais également non étroitement reliés à l'utilisateur. Quelques profils ont un score élevé de similarité (p. ex. A, B, C) et sont *a priori* des membres de la famille, des amis proches, etc. À l'opposé, un ensemble de profils qui ont une faible similarité ont aussi un comportement anormal (p. ex. D, E), ces profils présentent des caractéristiques de profils malveillants. Nous n'observons pas de contacts qui ont à la fois des valeurs très élevées d'anomalies et de similarité. Ceci est principalement dû au fait que si un contact anormal possède une relation étroite avec l'utilisateur, il est probable que l'utilisateur va détecter rapidement ce comportement anormal et prendre la décision appropriée. Nous notons que quelques contacts (p. ex. F, G) ont cependant des scores relativement élevés de similarité et d'anomalies. Pour ces profils, l'identification peut être compliquée. Ces profils partagent un lien fort avec l'utilisateur, mais présentent aussi un comportement anormal.

5.5 L'évaluation de l'approche

5.5.1 Évaluation sur Twitter et Facebook

L'évaluation de l'approche est effectuée sur les deux ensembles de données présentés dans la section précédente. Nous fournissons les performances obtenues par la mesure de similarité pour identifier les contacts légitimes et les performances obtenues lorsque le

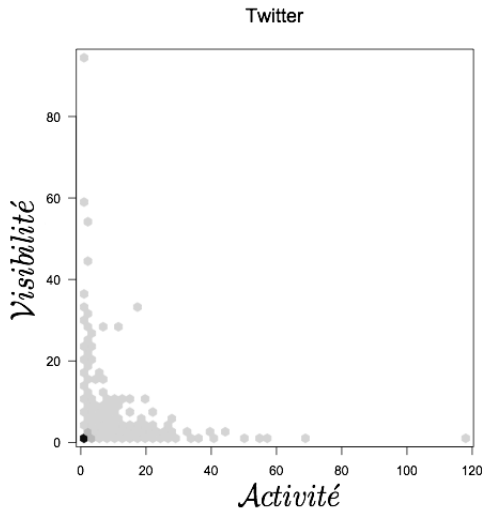


Figure 5.4.3 : Occurrences des couples activité, visibilité sur Twitter.

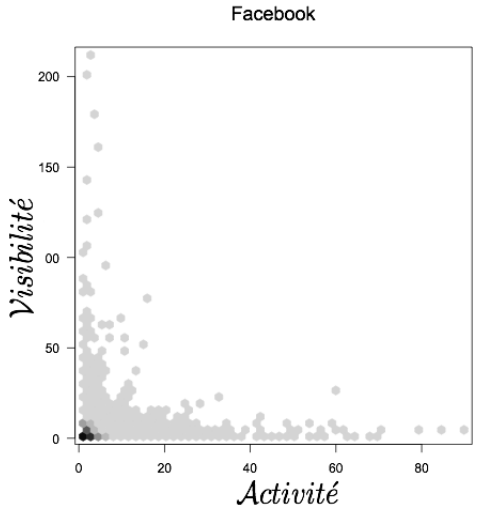


Figure 5.4.4 : Occurrences des couples activité, visibilité sur Facebook.

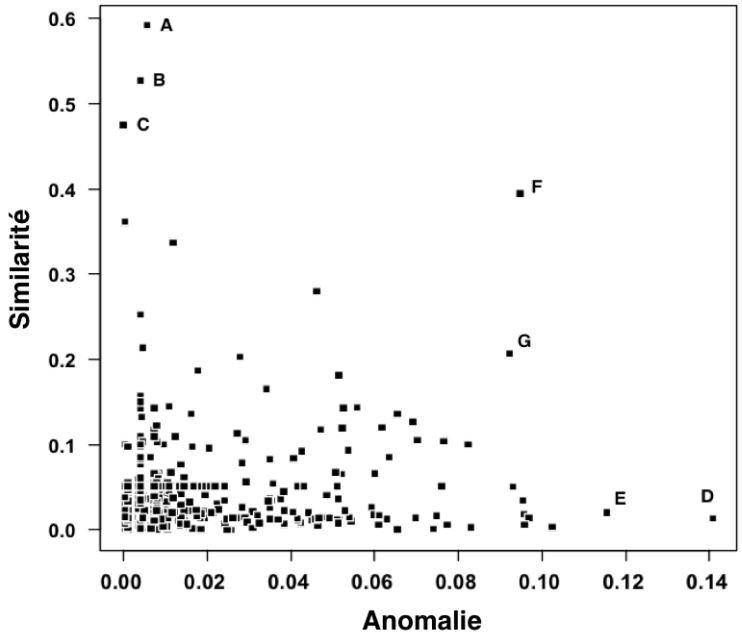


FIGURE 5.4.5 – Représentation des contacts Facebook d'un utilisateur de smartphone dans le repère similarité et anomalie.

comportement (score d'anomalie) des profils testés est en plus pris en compte.

La performance de la mesure $SBRA$ pour prédire les contacts légitimes a été mesurée en comparaison avec les mesures de similarité existantes par la valeur de l'aire sous la courbe ROC (aussi nommée fonction d'efficacité du récepteur).

Nous fournissons dans le tableau 5.6, une liste complète des mesures testées et leurs performances à la fois sur Twitter et sur Facebook.

Il est important de noter que les résultats présentés dans ce tableau correspondent aux moyennes des meilleures valeurs d'AUC obtenues pour chaque individu sur un ensemble d'utilisateurs. Le fait de considérer chaque utilisateur indépendamment autorise l'affectation d'un seuil distinct de similarité pour chacun. Ceci peut expliquer les faibles différences obtenues entre les différents indicateurs. En complément nous présenterons, dans la sous-section suivante, les performances obtenues lorsque un seuil seulement est appliqué pour l'ensemble des utilisateurs considérés. Cette méthode de validation permet de mieux tester la robustesse des approches.

Les meilleurs résultats sont obtenus pour les deux réseaux sociaux lorsque l'analyse du smartphone est intégrée dans la mesure. Cela signifie que le comportement local de l'utilisateur dans les médias sociaux disponibles sur un smartphone donne une indication importante sur qui ou ce qui est le plus pertinent pour lui.

En ce qui concerne Twitter, les deux types de contexte ont été testés. Lorsque le contexte est défini comme $\mathcal{C}ontenu$, la mesure se base sur la pertinence des centres d'intérêt communs. Lorsqu'il est défini comme $\mathcal{I}dentité$, la mesure se base sur la pertinence des voisins communs. Nous observons que lorsque le contenu est pris en compte l'analyse du smartphone permet d'augmenter les performances obtenues par rapport aux autres mesures. Les termes imbriqués (obtenus par son utilisation du smartphone) semblent être un critère important lors de la création de contacts fiables sur Twitter ($SBRA = 0.72$). En outre, la diversité ou non des sujets d'intérêt des profils Twitter est un bon indicateur de la création d'une relation légitime ($PA = 0.70$). Les connexions fiables peuvent se produire lorsque les spécialistes créent des liens avec des spécialistes et les non spécialistes avec des non spécialistes. Toutefois, comme indiqué dans [34] et dans le chapitre précédent lorsque le contexte $\mathcal{I}dentité$ est considéré alors l'analyse du smartphone ne fournit pas de résultats intéressants. Les contacts fiables sur Twitter sont identifiés comme étant pour la plupart des hubs ($HPI = 0.93$).

Indicateurs	**Twitter** ($\mathcal{C}ontenu$)	**Twitter** ($\mathcal{I}dentité$)	**Facebook** ($\mathcal{I}dentité$)
CN	0.51	0.78	0.89
Salton	0.59	0.64	0.88
Jaccard	0.61	0.54	0.89
Sørensen	0.66	0.46	0.89
HPI	0.56	**0.93**	0.89
HDI	0.68	0.67	0.88
LHN	0.66	0.55	0.87
PA	0.70	0.50	0.56
AA	0.51	0.73	0.83
RA	0.50	0.71	0.84
SBRA	**0.72**	0.69	**0.90**

Tableau 5.6 : Scores AUC des principaux indicateurs de prévision des liens pour la détection de contact légitimes sur Twitter et Facebook (les meilleurs résultats sont affichés en gras et les moins bons sont soulignés).

En ce qui concerne Facebook, seule la similarité basée sur l'identité a été testée (les messages Facebook ne sont pas publics). La typologie du réseau permet d'identifier que l'indice d'attachement préférentiel n'obtient pas de bonnes performances. Cela indique que les célébrités ne se connectent pas à des célébrités (resp. pour les non célébrités). Globalement, les indicateurs basés sur les voisins communs ont de bonnes performances. Nous observons que l'indice d'allocation des ressources ne fonctionne pas aussi bien que l'indice de voisin commun lorsque le chevauchement des contacts n'est pas introduit dans la mesure.

5.5.2 Approfondissement de l'évaluation sur Facebook

La figure 5.5.1 fournit les courbes ROC obtenues pour prédire les contacts fiables sur Facebook avec l'indicateur SBRA. Les courbes illustrent les effets de chaque réseau social dans la performance globale de l'indicateur. Il semble que Twitter soit la couche la moins importante pour la prédiction. Cela peut révéler le fait que les contacts Facebook d'un utilisateur ne sont pas représentés parmi les contacts Twitter ou que leur présence sur Twitter n'est pas pertinente pour la prédiction des contacts légitimes. Notons aussi que la couche Twitter ne se focalise que sur le contenu et non sur les personnes, alors que les autres couches sont davantage liées à la notion d'identité. Nous observons que les couches courriels, Google+ et carnet d'adresses sont nécessaires pour obtenir de bonnes performances. Si l'une de ces couches n'est pas disponible, cela peut réduire de manière significative les performances de l'approche.

Les résultats ci-dessus nous indiquent que chaque strate n'a pas la même importance dans la prédiction de contacts légitimes. Nous avons donc poursuivi notre évaluation de l'approche en affectant pour chaque strate un coefficient. Ensuite, grâce à une optimisation de l'erreur quadratique moyenne appliquée à la performance de l'approche, nous avons pu mettre en évidence le poids optimal de chaque strate pour la prédiction de contacts légitimes sur Facebook dans le cadre de notre mesure de similarité basée sur la smartphone. La figure 5.5.2, présente le poids de chaque réseau permettant une performance optimale

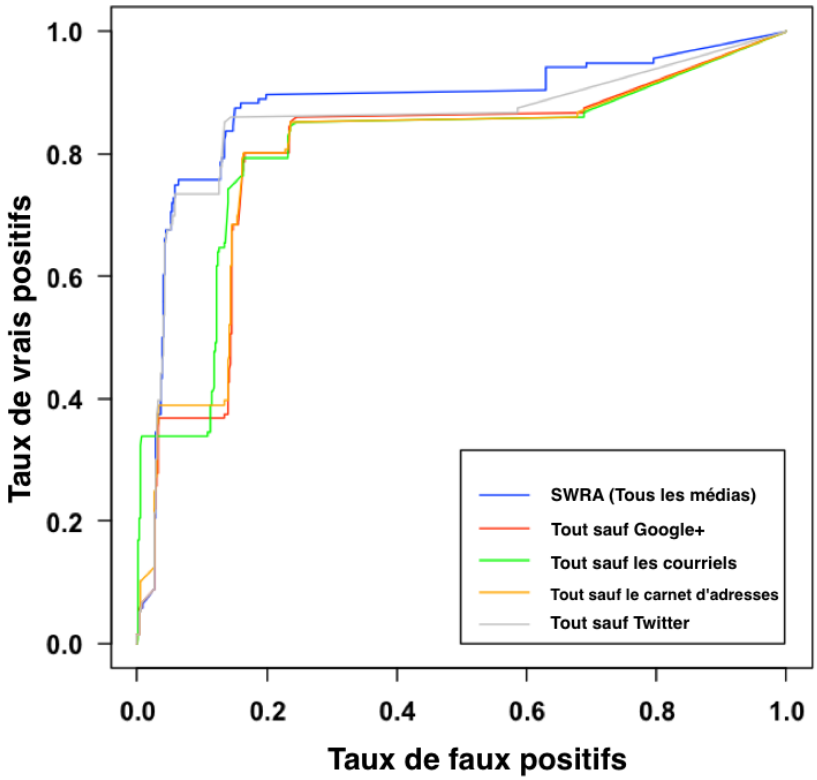


Figure 5.5.1 : Performances de la mesure de similarité *SBRA* sur Facebook en fonction des médias pris en compte.

dans la détection de contacts légitimes. Nous observons que la strate carnet d'adresses du téléphone est la plus importante devant la strate Mail, Google+, Facebook et enfin Twitter. Notons que, par construction, la présence d'un contact sur Facebook est inhérente à notre indicateur de similarité. Ceci peu expliquer, le faible score d'importance de cette strate qui finalement n'est pas révélatrice d'une imbrication. *A contrario*, nous observons que notre modèle est optimal lorsque la présence d'un contact dans le carnet d'adresses est mise en évidence. Ceci indique que la légitimité d'un contact est obtenue par le partage d'amis communs qui se retrouvent dans le carnet d'adresses de l'utilisateur. Aussi, la présence dans les contacts mail (0.34) et Google+ (0.15) est aussi significative. Enfin, la strate Twitter n'apporte aucun élément décisif pour la mesure de la légitimité d'un contact Facebook. Notons que ces résultats corroborent les observations de la figure 5.5.1 qui présente les performances de l'approche en fonction des réseaux pris en considération.

Nous présentons dans le tableau 5.7, les valeurs de performance obtenues sur un ensemble de données de Facebook. Contrairement au tableau 5.6, les performances ainsi présentées ont été obtenues par l'application d'un unique seuil pour l'ensemble des données analysées. Ces valeurs représentent de meilleure manière la robustesse des approches. L'hétérogénéité de la typologie du graphe est donc prise en compte et peut affecter les performances de manière plus significative.

Nous observons que les valeurs obtenues pour l'ensemble des indicateurs sont plus faibles. Cela confirme le fait que l'application d'un même seuil de détection pour des utilisateurs dont les comportements peuvent varier est une contrainte importante. Cependant, la performance de notre approche SBRA est moins affectée et sa robustesse apparaît plus importante. Cette observation, peut notamment provenir du caractère personnalisé de notre indicateur qui permet d'intégrer les fluctuations ponctuelles dans le

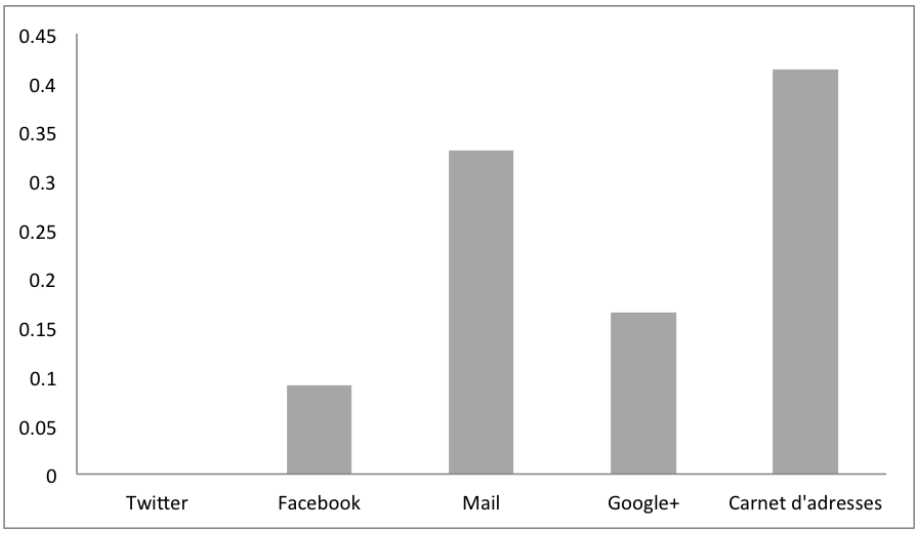


FIGURE 5.5.2 – Importance des strates dans la mesure de similarité basée sur le smartphone pour l'identification des contacts Facebook légitimes.

comportement des utilisateurs. Nous avons aussi indiqué la performance obtenue lorsque les pondérations optimales des strates sont appliquées (noté SBRA pondéré). Celle-ci nous a permis d'augmenter de 0.04 le score d'AUC obtenu par rapport au modèle dont chaque strate à la même importance.

Indicateurs	**Facebook** (*Identité*)
CN	0.66
Salton	0.64
Jaccard	0.61
Sørensen	0.64
HPI	0.62
HDI	0.62
LHN	0.61
PA	0.51
AA	0.64
RA	0.61
SBRA	0.75
SBRA(pondéré)	**0.79**

Tableau 5.7 – Scores AUC pour les principaux indicateurs de prévision des liens.

La figure 5.5.3 représente les performances obtenues par le score d'anomalie pour la détection des contacts suspects (potentiels auteurs de *phishing* d'après SPOT 1.0) sur Twitter et Facebook (score AUC). L'échelle de temps représente la durée d'observation des profils en termes de jours d'analyse. Tandis que sur Twitter la meilleure performance est obtenue après seulement 10 à 15 jours, sur Facebook la performance après un tel délai est relativement faible (environ 65 %). La performance sur Facebook continue d'augmenter avec le temps, la performance sur Twitter diminue après une période de 20 jours et se stabilise après une période d'observation de 90 jours.

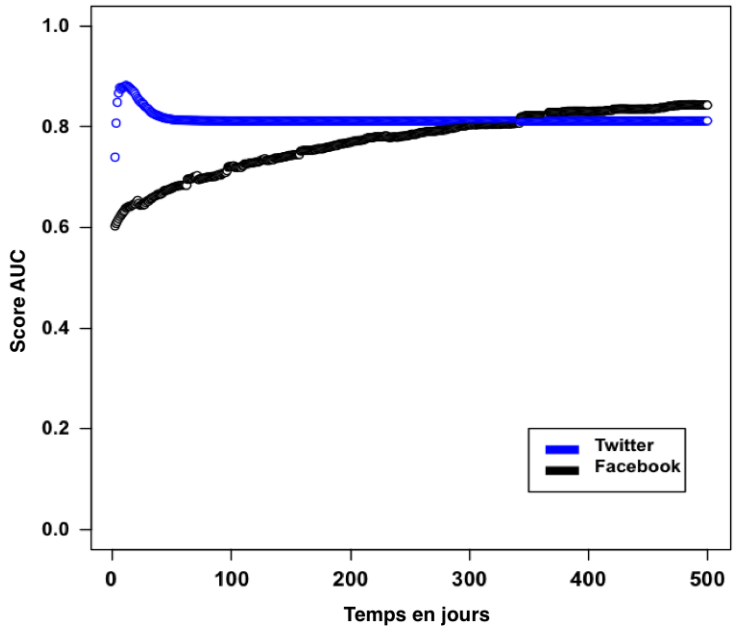


Figure 5.5.3 : Performances obtenues par le score d'anomalie pour détecter les profils suspects sur Twitter et Facebook.

Nous avons enfin testé le gain apporté par le score d'anomalie dans la prédiction des contacts légitimes (en complément de la mesure de similarité). Nous observons que l'intégration du score d'anomalie permet d'augmenter la performance en moyenne de 9 % sur Twitter et de 4 % sur Facebook (après 10 jours d'observation).

5.6 Prototype SPOTLIGHT 1.0

Nous présentons dans cette section, l'interface graphique de l'application SPOTLIGHT qui fait office de prototype. Pour une documentation technique plus complète, les lecteurs sont conviés à consulter le document disponible à l'URL suivante : `bit.ly/1lFYssc`

Bien que l'application soit disponible à la fois sur Android et iOS pour iPhone et iPad, nous présentons ci-dessous la version iOS pour iPad qui est à ce jour la plus aboutie.

Tout d'abord, la visualisation des contacts de l'utilisateur est présentée sur la figure 5.6.1. Sur cet affichage, la liste de réseaux gérés par l'application est disponible sur la partie gauche de l'écran. Dans la version actuelle, le carnet d'adresses, Facebook, Twitter, Gmail et Mail sont pris en compte. La quantité de contacts de l'utilisateur est indiquée sous le nom de chaque réseau. Une mise à jour est réalisable grâce au bouton destiné à cet effet. Notons aussi que sur la gauche, une aide est présentée pour faciliter la compréhension de l'utilisateur.

Nous observons, sur la droite de l'écran, la liste des contacts Facebook de l'utilisateur. Pour chacun d'entre eux les niveaux d'activité, de visibilité et d'imbrication sont présentés.

Le deuxième écran présente les messages publiés par les contacts de l'utilisateur sur les deux réseaux Twitter et Facebook. Ceux-ci sont organisés par jour et heure de publication. Il est possible d'activer et de désactiver l'affichage des messages de Facebook et/ou de Twitter à n'importe quel moment. Nous planifions d'intégrer un filtre permettant d'accéder aux messages associés à un mot-clé ou à un expéditeur. Notons qu'il est possible de publier des messages depuis cet écran. L'envoi de message est configurable en fonction du

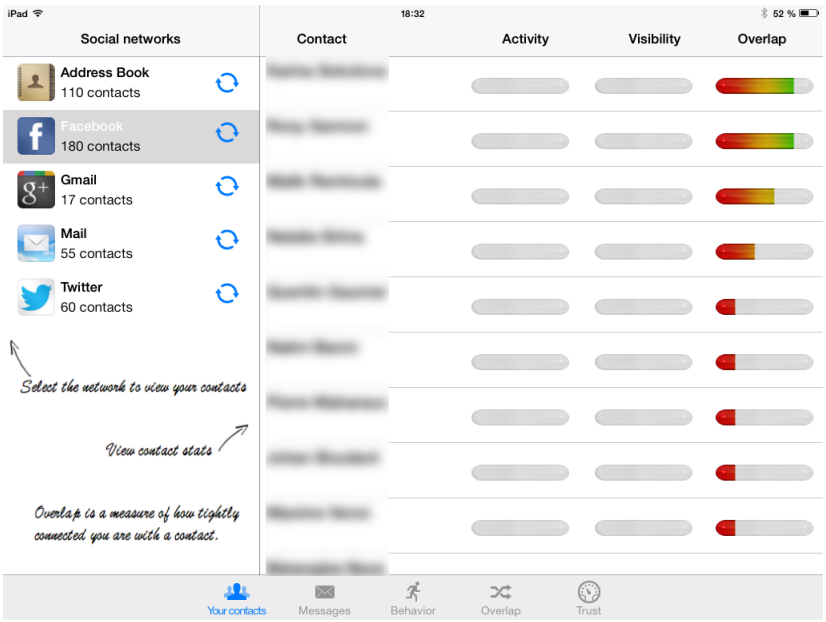


FIGURE 5.6.1 – Onglet de visualisation des contacts de l'utilisateur.

réseau pour spécifier l'audience du message et l'ajout ou non de données de localisation géographique.

Les messages présentés sur cet écran sont par défaut tous les messages de moins d'un jour. Lorsqu'une mise à jour des messages est effectuée, l'ensemble des messages de plus d'un jour est supprimé et les indicateurs comportementaux sont synthétisés et intégrés dans la base de données locale. Nous prévoyons de rendre paramétrable cette fonctionnalité dans une version future.

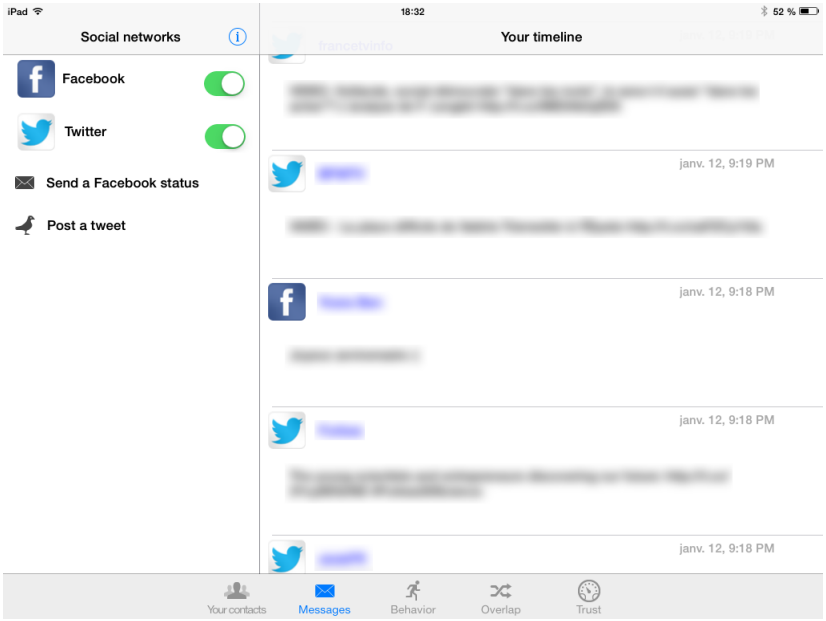


FIGURE 5.6.2 – Onglet de gestion des messages Twitter et Facebook.

La figure 5.6.3 met en évidence le comportement des contacts sur les réseaux Facebook et Twitter. La représentation bidimensionnelle des contacts a pour objectif la mise en

évidence de comportements suspects de certains contacts. Cette représentation illustre les valeurs d'activité et de visibilité des contacts à partir de leur comportement mesuré. Tel qu'indiqué dans cette thèse, il est possible d'identifier automatiquement avec de bonnes performances les contacts n'étant pas légitimes à partir de ces deux dimensions. Pour ce faire, sur la partie gauche de l'écran sont indiqués par ordre décroissant d'activité et de visibilité les profils du réseau sélectionné. L'icône d'œil permet de visualiser le profil sur le réseau concerné pour vérifier la pertinence de ce contact et, si nécessaire, de le supprimer de la liste d'amis. Le symbole danger indique la nécessité pour l'utilisateur de vérifier le comportement du contact qui paraît suspect.

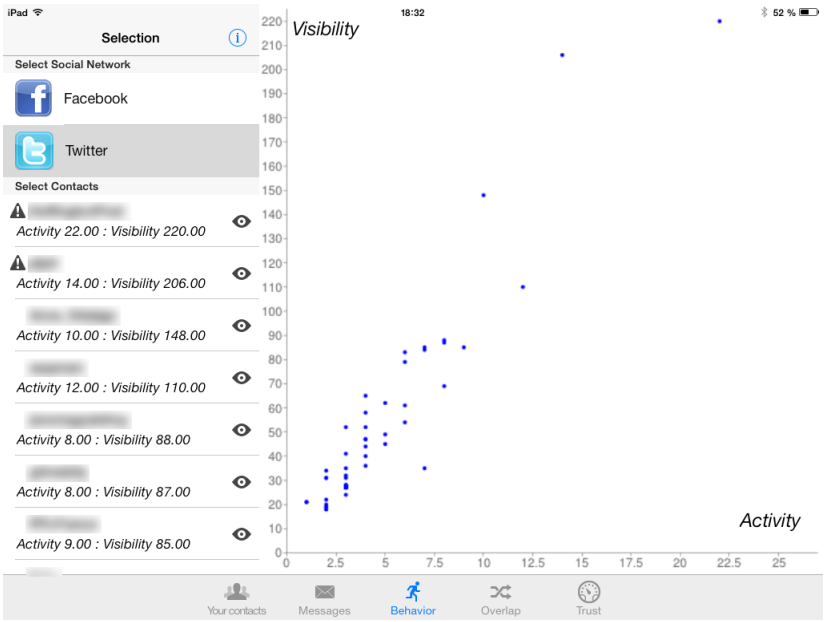


FIGURE 5.6.3 – Onglet de visualisation du comportement des utilisateurs sur Twitter et Facebook.

La figure 5.6.4 illustre la possibilité de sélectionner un profil pour le visualiser dans la représentation. Un profil sélectionné s'affiche en rouge et un message contextuel peut apparaître à cette occasion. Dans l'exemple, le profil sélectionné possède une valeur élevée de visibilité ce qui a conduit à l'affiche d'une alerte.

La figure 5.6.5 présente le quatrième onglet de l'application. Cet onglet présente le niveau d'imbrication des contacts dans les différents réseaux considérés. Il est possible, via l'écran de gauche, d'activer ou désactiver des réseaux. Les contacts sont automatiquement triés en fonction de leur mesure d'imbrication par valeur décroissante.

L'écran de mesure de la légitimité d'un contact sur Facebook est présenté sur la figure 5.6.6. Cette légitimité ne s'applique, dans cette version, que pour le réseau Facebook. Les contacts Facebook sont classés en fonction de cette mesure. Sur simple sélection d'un contact par l'utilisateur, l'ensemble des amis communs et leurs scores respectifs d'imbrication est affiché. À la première ouverture de cet écran, les mesures de légitimité ne sont pas calculées. L'utilisateur peut alors sélectionner un contact pour lancer l'exécution du calcul et afficher le score obtenu. La figure 5.6.7 montre qu'il est possible d'analyser tous les contacts pour obtenir en une seule fois la totalité des scores de légitimité. Pour cela, l'utilisateur accède au bouton d'option et sélectionne l'option « calculer pour tous les contacts ».

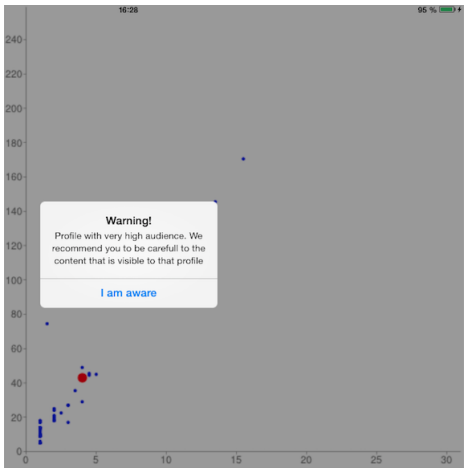


FIGURE 5.6.4 – Affichage d'une alerte pour un contact avec un fort score de visibilité.

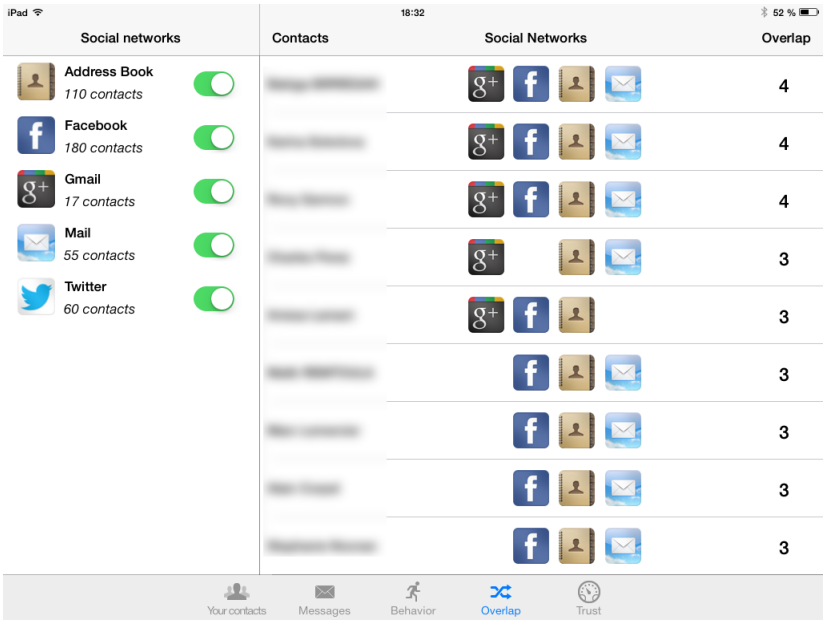


FIGURE 5.6.5 – Affichage des contacts de l'utilisateur triés par score d'imbrication.

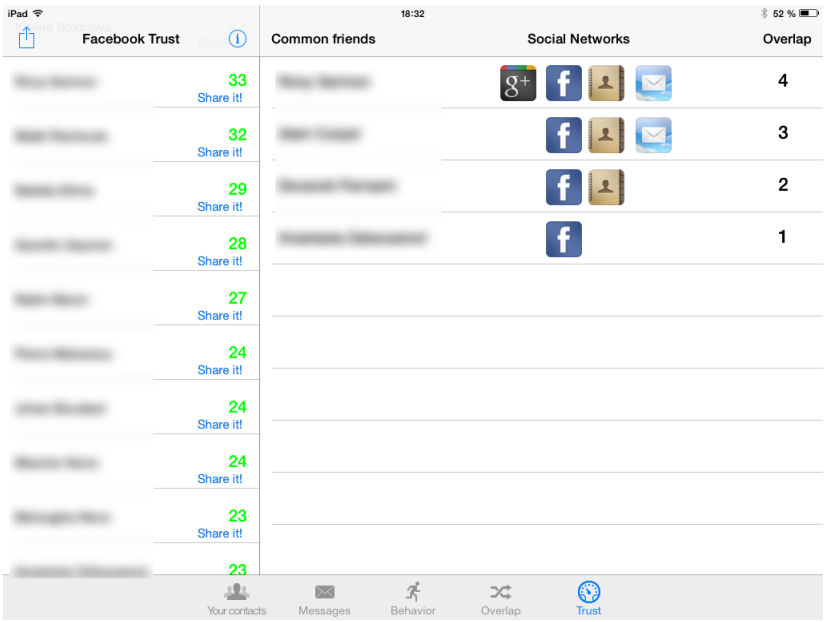


FIGURE 5.6.6 – Affichage du score de légitimité des utilisateurs de Facebook. Pour chaque profil, la liste des voisins communs est affichée avec le score d'imbrication associé.

L'autre option possible est l'envoi de certaines données anonymisées au développeur de l'application dans un but de recherche. Cette option telle que présentée en figure 5.6.8 génère un mail avec une pièce jointe contenant un fichier texte. Ce fichier résume les valeurs de légitimité et d'imbrication des contacts de l'utilisateur. Ce fichier de données anonymes peut être utilisé pour générer des statistiques et tester de nouveaux modèles sur les interactions entre les utilisateurs de réseaux sociaux numériques. Il serait aussi possible d'y inscrire le jugement de l'utilisateur sur les contacts détectés pour connaître l'impact de l'application sur la prise de décision.

Conclusion

Nous avons présenté une approche pour la détection de contacts légitimes (resp. contacts non légitimes) sur les réseaux basés sur la notion d'identité, mais aussi ceux basés sur le contenu. À notre connaissance, cette approche est la première tentative basée à la fois sur le comportement de l'utilisateur de smartphone et le comportement du profil testé pour détecter les contacts légitimes. L'approche est conçue pour être intégrée sur le smartphone d'un utilisateur et repose sur deux analyses complémentaires. Tout d'abord, nous analysons les données sociales disponibles sur un smartphone pour identifier les personnes et/ou les termes clés. Ces entités sont utilisées pour mesurer un degré de similitude entre un profil et l'utilisateur de smartphone. Ensuite, une mesure de comportement anormal d'un profil est calculée en fonction de son niveau d'a*ctivité* et de *visibilité*. L'application a montré que le niveau de similarité, lorsqu'il est adapté au bon type de réseau social, obtient de bonnes performances pour identifier les contacts légitimes. Nous avons aussi montré que l'intégration du score d'anomalie permet d'améliorer la qualité de la prédiction sur les deux types de réseaux.

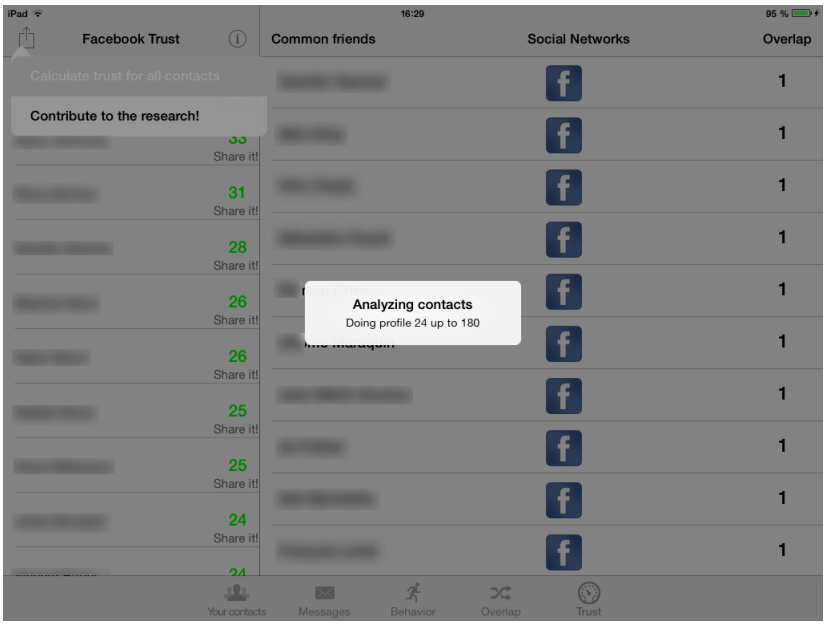


FIGURE 5.6.7 – Message de calcul en cours de la légitimité.

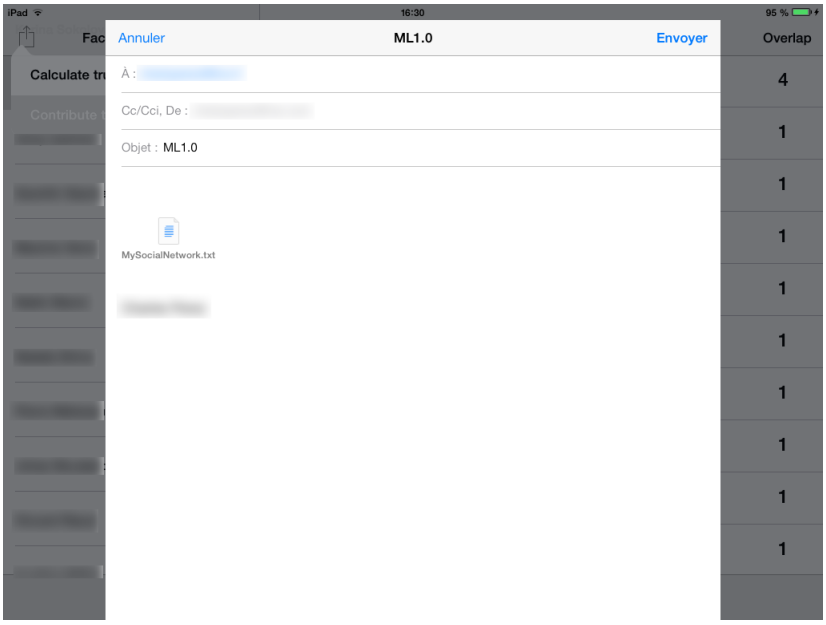


FIGURE 5.6.8 – Courriel permettant l'exportation des données de l'application.

Conclusion et perspectives

Rappel des objectifs

Les nouvelles technologies de l'information et de la communication, en particulier les smartphones et réseaux sociaux numériques, ont permis le développement d'une grande variété de canaux de communication. L'utilisateur évolue désormais dans un espace numérique relativement libre où il est possible de communiquer avec n'importe quelle autre entité, quel que soit son emplacement géographique et quelle que soit sa nature (p. ex. humain, robot). Cette nouvelle relation avec l'espace et le temps a ouvert de nouvelles portes et a permis l'exploration de nouvelles interactions jusqu'alors impossibles.

Les activités quotidiennes de l'utilisateur génèrent automatiquement des traces stockées à la fois sur les terminaux nomades et sur des serveurs distants. Celles-ci peuvent échapper au contrôle de l'internaute malgré les actions qu'il peut mettre en œuvre pour protéger ses données privées. L'utilisateur doit donc faire face à de nombreuses actions malveillantes.

Les utilisateurs n'ont pas nécessairement conscience de la potentielle présence d'entités malveillantes et évaluent mal le nombre de personnes ayant accès à leurs données. Ils naviguent donc en permanence sur de nombreuses plateformes numériques où ils sont à la fois de potentielles victimes d'attaques malveillantes (p. ex. attaques par *phishing*) et eux-mêmes vecteurs de divulgation d'informations sensibles à des contacts qui ne seraient pas légitimes.

Dans ce contexte, l'objectif de cette thèse était la mise en œuvre d'approches méthodologiques de protection des données de l'utilisateur sur les réseaux sociaux numériques mobiles. Ces approches doivent intégrer une analyse des profils exploitant la plateforme sociale pour propager du contenu malveillant pouvant nuire à la sécurité de l'utilisateur et du système d'information auquel il est associé. Celles-ci doivent aussi permettre une mesure de légitimité des contacts (de l'audience) ayant accès à l'information publiée par l'utilisateur.

Bilan des travaux effectués

Le travail de ces trois années de thèse a permis la réalisation de deux approches complémentaires et intégrées dans la méthodologie SPOTLIGHT.

À l'échelle d'Internet comme vecteur d'attaque massif (ou à large spectre), les travaux ont permis de valider le rôle des différents indicateurs (visibilité, agressivité et danger) qui ont été proposés pour la classification des profils sur les plateformes de socialisation. Grâce à une collaboration avec l'Université de Ballarat nous avons pu développer une approche commune qui intègre ces indicateurs avec des approches de fouille de données centrées sur l'identification des auteurs (authorship attribution). L'analyse des résultats obtenus a permis de démontrer la pertinence de nos indicateurs dans le cadre de la détection d'attaques à grande échelle tel que les « campagnes » de propagation d'URL malveillantes,

de contenus illicites, etc. Au-delà, une analyse plus fine a permis d'identifier non seulement le mode opératoire mais aussi des fragments de l'identité sociale de l'attaquant qui se cache derrière chaque campagne.

À l'échelle du smartphone comme point de vulnérabilité exploitable pour des attaques plus ciblées : l'approche « multicouche » a été proposée et repose sur l'exploration des traces des différentes applications sociales du téléphone (téléphonie, messagerie, réseaux sociaux, etc.) pour la construction d'un indicateur de légitimité. Cet indicateur est calculé sur la base de la fonction de redondance des contacts dans les différentes couches.

Enfin, l'approche générale SPOTLIGHT permet de fournir à l'utilisateur du téléphone exploré une vision du niveau de confiance qu'il peut accorder aux contacts présents dans ses réseaux. L'indicateur proposé apparaît comme un critère de légitimation qui permet de discriminer les contacts du smartphone.

Les perspectives et les limites de la méthodologie SPOTLIGHT

Nous proposons un inventaire des pistes de recherches et d'amélioration qui n'ont pas pues être explorées dans le cadre de cette thèse.

La pondération de strates

Le modèle ML adapté pour la mesure d'imbrication a prouvé son efficacité. Cependant, chaque utilisateur déploie un usage qui lui est propre, plus ou moins important de chaque réseau social numérique. Ainsi, la force de l'interaction avec un de ces contacts est d'autant plus forte qu'il communique avec celui-ci sur de multiples medias de communication, mais également d'autant plus forte qu'il utilise fortement les medias sur lesquels il est présent. Dans ce travail, ce problème de pondération des medias sociaux n'a pas été entièrement pris en compte. Notons que certains tests ont été effectués dans ce sens (pour Facebook) et présentent des résultats encourageants pour la suite. Il est important de garder cette piste ouverte pour travailler sur un indicateur de confiance encore plus personnalisé. Cette piste d'évolution doit aussi faire face à des contraintes techniques notamment de mesure de l'usage d'un réseau social numérique qui n'est actuellement pas réalisable sur tout type d'appareil.

L'intégration des contraintes spatio-temporelles

Parmi les pistes les plus sérieuses n'ayant pu être traitées, l'intégration d'une strate spatio-temporelle dans le modèle multicouche. Il est assez cohérent de penser que certaines relations sont tissées non pas grâce à des connaissances communes, ni par des centres d'intérêt communs mais simplement par des rencontres fortuites. Ce type de rencontres est certainement à l'origine de la performance non excellente de notre approche. Il paraît donc tout à fait intéressant d'analyser les rencontres entre les utilisateurs (par le biais du bluetooth par exemple) pour mesurer une distance spatio-temporelle entre les utilisateurs de smartphone. La présence dans le même contexte et régulière de deux personnes physiques donne tout naturellement un critère de légitimité qui peut être pris en compte dans un modèle futur. Notons tout de même, que l'analyse des rencontres spatio-temporelles entre deux utilisateurs de média sociaux a été traitée mais non présentée dans cette thèse car non intégrée au modèle final [143]. Ce travail a notamment permis de mettre en évidence la faisabilité de la mesure de distance spatio-temporelle entre deux utilisateurs de Twitter sous certaines conditions. Les trois conditions mises en

évidence sont : (1) les données utilisateur doivent être accessibles publiquement, (2) un service de localisation géographique doit être intégré à la plateforme et (3) les utilisateurs doivent être actifs sur la plateforme. En ce qui concerne la première exigence, nous ne pouvons analyser que les plateformes qui fournissent une quantité importante de données publiques. La deuxième condition est d'avoir accès à des données spatio-temporelles et cela est désormais possible avec les réseaux sociaux en ligne tels que Twitter et Facebook et l'utilisation de logiciels de réseaux sociaux mobiles comme Foursquare. La dernière condition est obligatoire pour agrandir la portée et l'intérêt de l'expérience. L'analyse et la comparaison de ces trois conditions sur les principales plateformes nous a conduit à définir la plateforme de microblogging Twitter comme source de données pour la mesure de similarité spatio-temporelle.

L'intégration du retour d'expérience de l'utilisateur

Le développement et le déploiement à large échelle de l'application SPOTLIGHT peut faire émerger une piste importante d'évolution de l'outil. Il est en effet important de pouvoir analyser l'impact de l'application sur les décisions prises par l'utilisateur. En d'autres termes, la mise en évidence de certains comportements déviants et de certains risques est-elle suffisante comme méthode de sécurisation de l'utilisateur ? Cela pose la question de l'ergonomie de l'application mais aussi de la pertinence des indicateurs proposés vis-à-vis d'un utilisateur lambda. D'un point de vue technique et scientifique, cela nécessite l'enregistrement des actions de l'utilisateur et cela pose naturellement le problème de respect de la vie privée.

Vers un modèle personnalisé, contextuel et dynamique

La perspective la plus sérieuse d'évolution des travaux de thèse est l'intégration de l'aspect dynamique dans un outil plus complet d'aide à la décision (cf. figure 5.6.9). L'idée première est de conserver les deux dimensions testées dans cette thèse, à savoir le niveau de similarité d'un contact et la mesure du comportement plus ou moins anormal de celui-ci. Dans un cadre plus complet nous avons mis en évidence quatre principales zones théoriques issues de ce repère.

La zone I correspond à un profil dont la légitimité (en termes de proximité) est forte avec l'utilisateur et dont le comportement est tout à fait normal. Ce type de profil à un instant donné ne présente pas de risque pour l'utilisateur.

La zone II contient les profils dont la similarité avec l'utilisateur est relativement faible mais dont le comportement est a priori normal. Ce type de profil peut représenter un risque sur les plateformes de socialisation ou des données personnelles sont partagées avec les contacts. Ainsi, un risque de fuite de données est réel et des mesures sur la suppression ou non de ce contact peuvent être mises en place.

La zone III présente la liste des profils les plus critiques en termes de sécurité. Ces profils n'ont pas de légitimité en termes de similarité avec l'utilisateur et présentent un comportement atypique. Il convient, pour obtenir un niveau de sécurité maximal, d'être particulièrement prudent quant à la lecture de contenus produits par ce type de profil. Aussi, la publication de contenus personnels peut être assez dangereuse dans ce contexte. Il convient pour l'utilisateur de se poser la question de la réelle valeur ajoutée et des risques présentés par ces contacts dans son réseau.

La zone IV contient la dernière catégorie de profils. Ceux-ci ont une forte similarité avec l'utilisateur mais présentent tout de même un niveau anormal de comportement.

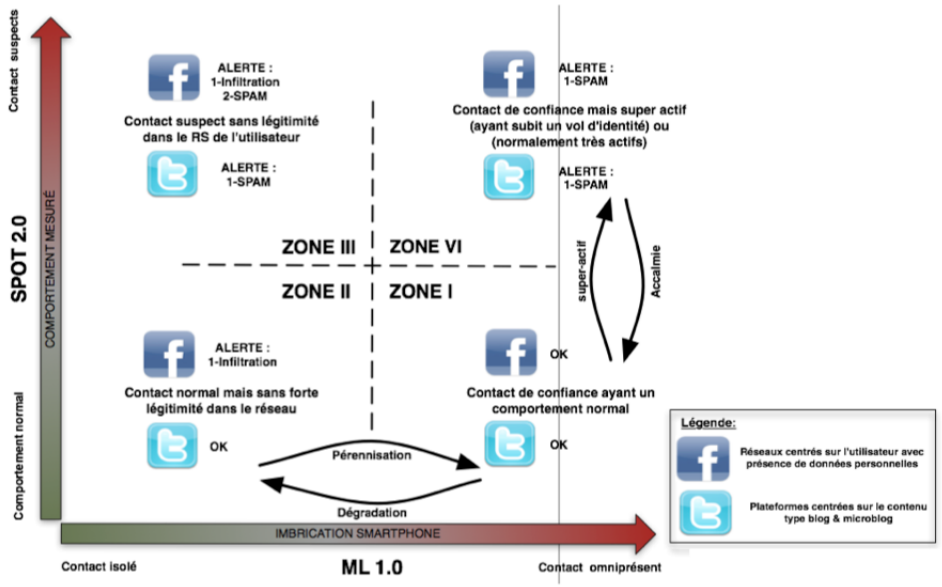


FIGURE 5.6.9 – Modèle théorique d'évaluation des contacts d'un utilisateur de smartphone.

Cette observation peut révéler une automatisation de ces profils et potentiellement la propagation de contenus malveillants.

L'aspect dynamique du modèle peut permettre d'éclaircir la situation d'un tel profil. Quels sont les états habituels de ce type de profils ? Quelle évolution est mesurée quant à la position de ces profils dans le repère théorique ? Nous identifions deux types de transitions : (1) d'une part, une évolution significative du comportement du profil évalué et, (2) d'autre part, une évolution de sa similarité avec l'utilisateur de smartphone.

Le premier type peut concerner le passage d'un état normal à un état anormal ; le plus souvent d'un état peu actif à un état très actif. Si nous prenons comme échelle de temps une journée, ce type de transition doit être suivi avec attention et nous devons pouvoir identifier si cet état est transitoire ou s'il persiste dans le temps. Il est vraisemblable que si cet état persiste trop longtemps, des mesures de sécurisation vis-à-vis de ce profil doivent être prises. Dans le cas contraire, il peut s'agir d'une perturbation ponctuelle qui n'affecte pas la sécurité de l'utilisateur. Ainsi, une accalmie sera observée dans les heures ou jours qui suivent cette transition.

Le second type de transition concerne le passage d'un niveau de proximité avec l'utilisateur à un autre. La pérennisation d'une relation peut se conclure par une augmentation significative de la mesure de similarité entre le profil analysé et l'utilisateur de smartphone. Cette évolution signifie a priori qu'une relation de confiance est en train de prendre forme. A contrario, une dégradation de la relation peut être observée. Dans ce cas il est vraisemblable que l'évolution du réseau de l'utilisateur mène à positionner ce profil d'un état désirable vers un état non désirable.

La dynamique générée par l'action de l'utilisateur de smartphone mais aussi par le comportement de ses contacts doit être prise en compte pour proposer une sécurisation personnalisée mais aussi adaptative et contextuelle. Le cadre théorique tel que présenté ici a été testé sur seulement quelques aspects dans cette thèse. Nous avons mesuré l'apport de la prise en compte du comportement des profils par rapport à une approche basée uniquement sur la similarité. Les premiers résultats sont encourageants, mais il apparaît

évident que toute les possibilités offertes par ce modèle n'ont pas été testées.

Publications

Articles dans des revues répertoriées dans les bases de données internationales

Charles Perez, Babiga Birregah et Marc Lemercier, A smartphone-based online social network trust evaluation system. Social Network Analysis and Mining, 18 p. Springer, 2014.

Chapitre d'ouvrage

Charles Perez, Marc Lemercier, Babiga Birregah et Cyril Debard, « Nouvelle approche de l'investigation numérique à l'aide de synergies entre smartphones et réseaux sociaux », Chapitre 4 du numéro sur l'identité numérique, Collection les cahiers du numérique sous la responsabilité de Jean Paul Pinte, 2014.

Communications avec actes répertoriés dans les bases de données internationales

Charles Perez, Babiga Birregah, Robert Layton, Marc Lemercier et Paul Watters, REPLOT : REtrieving Profile Links On Twitter for suspicious networks detection. Proceedings of the 2013 International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 8 p. IEEE, 2013.

Robert Layton, Charles Perez, Babiga Birregah, Paul Watters et Marc Lemercier, Indirect information linkage for OSINT through authorship analysis of aliases. The International Workshop on Data Mining Applications in Industry & Government (DMApps 2013), 12 p. 2013.

Charles Perez, Babiga Birregah et Marc Lemercier, Familiar Strangers detection in online social networks. Proceedings of the 2013 International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 8 p. 2013.

Charles Perez, Babiga Birregah et Marc Lemercier, The Multi-layer Imbrication for Data Leakage Prevention from Mobile Devices. Proceedings of the 11th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 813–819. IEEE, 2012.

Babiga Birregah, Tony Top, Charles Perez, Eric Châtelet, Nada Matta, Marc Lemercier et Hichem Snoussi, Multi-layer Crisis Mapping : A Social Media-Based Approach. Proceedings of the 2012 IEEE 21st International Workshop on Enabling Technologies : Infrastructure for Collaborative Enterprises, 379–384. IEEE, 2012.

Charles Perez, Marc Lemercier, Babiga Birregah et Alain Corpel, SPOT 1.0 : Scoring Suspicious Profiles on Twitter. Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 377–381. IEEE, 2011.

Communications avec actes non répertoriés dans les bases de données internationales

Vincent Lemoine, Charles Perez, Marc Lemercier, Pierre Vitard, Virginie Bensoussan, Alain Corpel, Rida Khatoun et Babiga Birregah, Protection of IT infrastructures against cyber crime. Interdisciplinary Workshop on Global Security (WISG 2013), 22-23 January 2013, Troyes. 2013.

Bibliographie

[1] The Real Face of KOOBFACE : The Largest Web 2.0 Botnet Explained. pages 1–18, juillet 2009.

[2] John Arundel Barnes : Class and committees in a norwegian island parish. *Human Relations*, 7(1):39–58, 1954.

[3] Walter W. Powell, Kenneth W. Koput et Laurel S. Doerr : Interorganizational Collaboration and the Locus of Innovation : Networks of Learning in Biotechnology. *Administrative Science Quarterly*, 41(1):116–145, mars 1996.

[4] Ronald L Breiger : Handbook of Data Analysis. pages 505–526. SAGE, 2004.

[5] Jens Krause, David Lusseau et Richard James : Animal social networks : an introduction. *Behavioral Ecology and Sociobiology*, 63:967–973(7), 2009.

[6] Stanley Milgram : The Familiar Stranger : An aspect of the urban anonymity. *Newsletter*, Division 8, 1972.

[7] Jacob Moreno : Who Shall Survive ? Foundations of Sociometry, Group Psychotherapy and Sociodrama. 1953.

[8] Danah Boyd et Nicole B. Ellison : Social Network Sites : Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13(1-2), novembre 2007.

[9] Aurélie GIRARD et Bernard FALLERY : Digital Social Networks : literature review and research perspectives. *In Association Information and Management*, juin 2009.

[10] Nina D. Ziv et Bala Mulloth : An Exploration on Mobile Social Networking : Dodgeball as a Case in Point. *In Proceedings of the International Conference on Mobile Business*, Washington, DC, USA, 2006. IEEE Computer Society.

[11] Laura Marcia Villalba Monné : A survey of mobile social networking. Rapport technique, Helsinki University of Technology, 2009.

[12] Fredrik Johansson : Extending mobile social software with contextual information. Rapport technique, 2008.

[13] Matti Rantanen, Antti Oulasvirta, Jan Blom, Sauli Tiitta et Martti Mäntylä : InfoRadar : Group and Public Messaging in the Mobile Context. Rapport technique, 2004.

[14] Emiliano Miluzzo, Nicholas D. Lane, Kristóf Fodor, Ronald Peterson, Hong Lu, Mirco Musolesi, Shane B. Eisenman, Xiao Zheng et Andrew T." Campbell : Proceedings of the 6th ACM conference on Embedded network sensor systems - SenSys '08. *In the 6th ACM conference*, pages 337–350, New York, New York, USA, 2008. ACM Press.

[15] Markus Loecher, Markus Loecher et Tony" Jebara : CitySense TM : multiscale space time clustering of GPS points and trajectories.

[16] Rohan Murty, Abhimanyu Gosain, Matthew Tierney, Andrew Brody, Amal Fahad, Josh Bers et Welsh Matt : CitySense : A vision for an urban-scale wireless networking testbed. 2008.

[17] Mario Cataldi, Luigi Di Caro et Claudio Schifanella : Emerging topic detection on Twitter based on temporal and social terms evaluation. *In Proceedings of the Tenth International Workshop on Multimedia Data Mining*, pages 4 :1–4 :10, New York, NY, USA, 2010. ACM.

[18] Melinger Daniel, Bonna Karen, Sharon Michael et SantRam Mohit : Socialight : A Mobile Social Networking System. Rapport technique, 2004.

[19] Aaron Beach, Mike Gartrell, Sirisha Akkala, Jack Elston, John Kelley, Keisuke Nishimoto, Baishakhi Ray, Sergei Razgulin, Karthik Sundaresan, Bonnie Surendar, Michael Terada et Richard Han : WhozThat ? evolving an ecosystem for context-aware mobile social networks. *IEEE Network*, 22(4):50–55.

[20] Wajeb Gharibi et Maha Shaabi : Cyber threats in social networking websites. *arXiv.org*, cs.CR, février 2012.

[21] Harvey Jones et José Soltren : Facebook : Threats to Privacy. *Project MAC : MIT Project on Mathematics and Computing*, 2005.

[22] Giles Hogben et Marnix Dekker : Smartphones : Information security risks, opportunities and recommendations for users. ENISA, 2010.

[23] Goran Delac, Marin Silic et Jakov Krolo : Emerging security threats for mobile platforms. *In MIPRO*, pages 1468–1473. IEEE, 2011.

[24] Marc Fossi : Symantec Global Internet Security Threat Report : Trends for 2010. 2011.

[25] Philip Wolny : Foursquare and Other Location-Based Services : Checking In, Staying Safe & Being Savvy. Digital and Information Literacy, 2011.

[26] Alex Hai Wang : Detecting spam bots in online social networking sites : a machine learning approach. *In DBSec'10 : Proceedings of the 24th annual IFIP WG 11.3 working conference on Data and applications security and privacy*, pages 335–342, Berlin, Heidelberg, juin 2010. Springer-Verlag.

[27] Gianluca Stringhini, Christopher Kruegel et Giovanni Vigna : Detecting spammers on social networks. *In the 26th Annual Computer Security Applications Conference*, page 1, New York, New York, USA, 2010. ACM Press.

[28] Kurt Thomas et David M. Nicol : The Koobface botnet and the rise of social malware. *Malicious and Unwanted Software (. . .*, 2010.

[29] Jon Kleinberg : The small-world phenomenon : an algorithm perspective. *In Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170, New York, NY, USA, 2000. ACM.

[30] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov et Matei Ripeanu : The socialbot network : When Bots Socialize for Fame and Money . *In the 27th Annual Computer Security Applications Conference*, page 93, New York, New York, USA, 2011. ACM Press.

[31] Frank Nagle et Lisa Singh : Can Friends Be Trusted ? Exploring Privacy in Online Social Networks. *In Social Network Analysis and Mining, 2009. ASONAM '09. International Conference on Advances in*, pages 312–315, 2009.

[32] Daniel Gayo-Avello et David J. Brenes : Overcoming spammers in twitter-a tale of five algorithms. *In 1st Spanish Conference on Information Retrieval, Madrid, Spain*, 2010.

[33] Tao Stein, Erdong Chen et Karan Mangla : Facebook immune system. *In SNS '11 : Proceedings of the 4th Workshop on Social Network Systems*. ACM Request Permissions, avril 2011.

[34] Charles Perez, Babiga Birregah et Marc Lemercier : A smartphone-based online social network trust evaluation system. *Social Netw. Analys. Mining*, 3(4): 1293–1310, 2013.

[35] Lin Yao, Chi Lin, Xiangwei Kong, Feng Xia et Guowei Wu : A Clustering-based Location Privacy Protection Scheme for Pervasive Computing. *arXiv.org*, cs.CR, novembre 2010.

[36] Nan Li et Guanling Chen : Sharing location in online social networks. *IEEE Network*, 24(5):20–25, 2010.

[37] Marcello Paolo Scipioni, Marcello Paolo Scipioni et Marc Langheinrich : I'm Here ! Privacy Challenges in Mobile Location Sharing. *Second International Workshop on Security and Privacy in Spontaneous Interaction and Mobile Phone Use (IWSSI/SPMU 2010)*, 2010.

[38] Marco Gruteser et Dirk Grunwald : Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking. *In Proceedings of the 1st international conference on Mobile systems, applications and services*, pages 31–42, New York, NY, USA, 2003. ACM.

[39] Phillip Bonacich : Power and Centrality : A Family of Measures. *American Journal of Sociology*, 92(5):1170–1182, 1987.

[40] Satu Elisa Schaeffer : Graph clustering. *Computer Science Review*, 1(1):27–64, août 2007.

[41] Nitin Agarwal, Huan Liu, Sudheendra Murthy, Arunabha Sen et Xufei Wang : A social identity approach to identify familiar strangers in a social network. *In* Eytan Adar, Matthew Hurst, Tim Finin, Natalie S. Glance, Nicolas Nicolov et Belle L. Tseng, éditeurs : *ICWSM*. The AAAI Press, 2009.

[42] Dieudonné Tchuente, Nadine Jessel, André Péninou, Marie-Françoise Canut et Florence Sèdes : A community based algorithm for deriving users' profiles from egocentrics networks : experiment on Facebook and DBLP. *Social Network Analysis and Mining*, 3(3):667–683, septembre 2013.

[43] Karen Spärck Jones : A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[44] Alex Wang : Don't follow me : Spam detection in twitter. *In Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, 2010.

[45] Fabio Dellutri, Luigi Laura, Vittorio Ottaviani et Giuseppe F Italiano : Extracting social networks from seized smartphones and web data. *Information Forensics and Security, 2009. (WIFS)*, pages 101–105, 2009.

[46] Lawrence Page, Sergey Brin, Rajeev Motwani et Terry Winograd : The pagerank citation ranking : Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[47] Ajita Gupta : Social Data mining on Smartphones. Rapport technique, juillet 2011.

[48] Salvatore Catanese, Emilio Ferrara et Giacomo Fiumara : Forensic analysis of phone call networks. *Social Network Analysis and Mining, 3(1) :15-33*, 2013.

[49] Theus Hossmann, Franck Legendre, George Nomikos et Thrasyvoulos Spyropoulos : Stumbl : Using Facebook to Collect Rich Datasets for Opportunistic Networking Research. *Information Forensics and Security, (WIFS),2009.*

[50] Daniele Quercia, Jonathan Ellis et Licia Capra : Using Mobile Phones to Nurture Social Networks. *IEEE Pervasive Computing*, 9(3):12–20, 2010.

[51] Nathan Eagle, Alex Pentland et David Lazer : From the cover : Inferring friendship network structure by using mobile phone data. *Proceedings of The National Academy of Sciences*, 106:15274–15278, 2009.

[52] Matteo Magnani et Luca Rossi : The ML-Model for Multi-layer Social Networks. *In Advances in Social Networks Analysis and Mining (ASONAM)*, pages 5–12. IEEE Computer Society, 2011.

[53] Elie Raad, Richard Chbeir et Albert Dipanda : User Profile Matching in Social Networks. *In 13th International Conference on Network-Based Information Systems (NBiS)*, pages 297–304. IEEE, 2010.

[54] Elie Raad, Richard Chbeir et Albert Dipanda : Discovering relationship types between users using profiles and shared photos in a social network. *Multimedia Tools Appl.*, 64(1):141–170, 2013.

[55] Dan Brickley et Ramanathan V Guha : RDF Vocabulary Description Language 1.0 : RDF Schema. Rapport technique, février 2004.

[56] Uldis Bojars, Alexandre Passant, Richard Cyganiak et John Breslin : Weaving SIOC into the Web of Linked Data. *In Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW), Beijing, China*, 2008.

[57] Li Ding, Lina Zhou, Timothy W Finin et Anupam Joshi : How the Semantic Web is Being Used : An Analysis of FOAF Documents. *In Hawaii International Conference on System Sciences (HICSS)*, 2005.

[58] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis et Vassilios S. Verykios : Duplicate record detection : A survey. *EEE Transactions on Knowledge and Data Engineering*, 19(1):1–16, janvier 2007.

[59] Peter Christen : A comparison of personal name matching : Techniques and practical issues. *In Workshop on Mining Complex Data (MCD), held at IEEE ICDM'06, Hong Kong*, pages 290–294, 2006.

[60] Vladimir Levenshtein : Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[61] Fred J. Damerau : A technique for computer detection and correction of spelling errors. *Communication of the ACM*, 7(3):171–176, mars 1964.

[62] Karen Kukich : Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24(4):377–439, décembre 1992.

[63] Matthew A Jaro : Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 1989.

[64] Edward H. Porter, William E. Winkler, Bureau Of The Census et Bureau Of The Census : Approximate string comparison and its effect on an advanced record linkage system. *In Advanced Record Linkage System. U.S. Bureau of the Census, Research Report*, pages 190–199, 1997.

[65] William E. Yancey : Evaluating string comparator performance for record linkage. 2005.

[66] Justin Zobel et Philip Dart : Phonetic string matching : lessons from information retrieval. *In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR)*, pages 166–172, New York, NY, USA, 1996. ACM.

[67] Kyumin Lee, James Caverlee et Steve Webb : The social honeypot project : protecting online communities from spammers. *In WWW '10 : Proceedings of the 19th international conference on World wide web*. ACM, avril 2010.

[68] Sarita Yardi, Daniel M Romero, Grant Schoenebeck et Danah M Boyd : Detecting spam in a Twitter network. *First Monday*, 15(1):1–13, 2010.

[69] Kyumin Lee, James Caverlee et Steve Webb : Uncovering social spammers : social honeypots + machine learning. *In SIGIR '10 : Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM Request Permissions, juillet 2010.

[70] Hongyu Gao, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen et Ben Y Zhao : Detecting and characterizing social spam campaigns. *In IMC '10 : Proceedings of the 10th annual conference on Internet measurement*. ACM Request Permissions, novembre 2010.

[71] Zahid Halim, Mian Maqsood Gul, Najam ul Hassan, Rauf Baig, Shafiq Ur Rehman et Farhat Naz : Malicious users' circle detection in social network based on spatio-temporal co-occurrence. *In Computer Networks and Information Technology (ICCNIT), 2011 International Conference on*, pages 35–39, July 2011.

[72] Saeed Abu-Nimeh, Thomas M. Chen et Omar Alzubi : Malicious and spam posts in online social networks. *IEEE Computer*, 44(9):23–28, 2011.

[73] Faraz Ahmed et Muhammad Abulaish : An MCL-Based Approach for Spam Profile Detection in Online Social Networks. *IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications*, pages 1–7, juin 2012.

[74] Kyumin Lee, James Caverlee, Krishna Y Kamath et Zhiyuan Cheng : Detecting collective attention spam. *In WebQuality '12 : Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*. ACM Request Permissions, avril 2012.

[75] Fabricio Benevenuto, Tiago Rodrigues, Virgilio Almeida, Jussara Almeida et Marcos Gonçalves : Detecting spammers and content promoters in online video social networks. *In SIGIR '09 : Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM Request Permissions, juillet 2009.

[76] Chao Michael Zhang et Vern Paxson : Detecting and Analyzing Automated Activity on Twitter. *Passive and Active Measurement*, 6579:102, 2011.

[77] Ian Fette, Norman Sadeh et Anthony Tomasic : Learning to Detect Phishing Emails. Rapport technique, 2006.

[78] Saeed Abu-Nimeh et Thomas Chen : Proliferation and Detection of Blog Spam. *IEEE Security & Privacy Magazine*, 8(5):42–47.

[79] Saptarshi Ghosh, Bimal Viswanath, Farshad Kooti, Naveen Sharma, Gautam Korlam, Fabricio Benevenuto, Niloy Ganguly et Krishna P Gummadi : Understanding and combating link farming in the twitter social network. *WWW '12 : Proceedings of the 21st international conference on World Wide Web*, 2012.

[80] Marisa Affonso Vasconcelos, Saulo Ricci, Jussara Almeida, Fabricio Benevenuto et Virgilio Almeida : Tips, dones and todos : uncovering user profiles in foursquare. *In WSDM '12 : Proceedings of the fifth ACM international conference on Web search and data mining*. ACM Request Permissions, février 2012.

[81] Varun Chandola, Arindam Banerjee et Vipin Kumar : Anomaly detection. *ACM Comput. Surv.*, 41(3):1–58, juillet 2009.

[82] Steven A Hofmeyr, Stephanie Forrest et Anil Somayaji : Intrusion Detection using Sequences of System Calls. *Journal of Computer Security*, 6:151–180, 1998.

[83] Haipeng Shen et Jianhua Z Huang : Analysis of call centre arrival data using singular value decomposition : Research Articles. *Appl. Stoch. Model. Bus. Ind.*, 21(3):251–263, 2005.

[84] Jeff Patti, Nadya Belov, Patrick Craven et Timothy Thayer : Applying Adaptive Anomaly Detection to Human Networks. *Human Behavior-Computational Modeling Intelligence Modeling Workshop*, 2009.

[85] Vinod Jakkula et D. Justin Cook : Anomaly detection using temporal data mining in a smart home environment. *Methods of information in medicine*, 47(1):70–75, 2008.

[86] Konstantinos Manikas : *Outlier Detection in Online Gambling*. Thèse de doctorat, IT University of Goteborg, 2008.

[87] Jigang Wang, Predrag Neskovic et Leon N Cooper : Improving nearest neighbor rule with a simple adaptive distance measure. *Pattern Recogn. Lett.*, 28(2):207–213, 2007.

[88] Varun Chandola, Arindam Banerjee et Vipin Kumar : Anomaly detection : A survey. *ACM Comput. Surv.*, 41:15 :1–15 :58, juillet 2009.

[89] Matthew Mullin et Rahul Sukthankar : Complete Cross-Validation for Nearest Neighbor Classifiers. *In 17th International Conference on Machine Learning (ICML*, 2000.

[90] Ke Zhang, Marcus Hutter et Huidong Jin : A New Local Distance-Based Outlier Detection Approach for Scattered Real-World Data. pages 813–822. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[91] Wei Xu, Fangfang Zhang et Sencun Zhu : Toward worm detection in online social networks. *In ACSAC '10 : Proceedings of the 26th Annual Computer Security Applications Conference*. ACM Request Permissions, décembre 2010.

[92] Chi Zhang, Jinyuan Sun, Xiaoyan Zhu et Yuguang Fang : Privacy and security for online social networks : challenges and opportunities. *Network, IEEE*, 24(4):13–18, 2010.

[93] Charles Perez, Babiga Birregah, Robert Layton, Marc Lemercier et Paul Watters : REPLOT : REtrieving Profile Links On Twitter for suspicious networks detection. *In 2013 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 1–8, septembre 2013.

[94] Jennifer Ann Golbeck : *Computing and Applying Trust in Web-based Social Networks*. Thèse de doctorat, College Park, MD, USA, 2005. AAI3178583.

[95] Paolo Massa et Paolo Avesani : Trust Metrics on Controversial Users. *International Journal on Semantic Web and Information Systems*, 3(1):39–64, 2007.

[96] Linyuan Lü et Tao Zhou : Link prediction in complex networks : A survey. *Physica A : Statistical Mechanics and its Applications*, 390(6):1150–1170, mars 2011.

[97] David Liben-Nowell et Jon Kleinberg : The link prediction problem for social networks. *CIKM '03 : Proceedings of the twelfth international conference on Information and knowledge management*, 2003.

[98] Facebook : People you may know, sep 2013.

[99] Erzsébet Ravasz et Albert-László Barabási : Hierarchical organization in complex networks. *Physical Review E*, 67(2):026112, février 2003.

[100] E Leicht, Petter Holme et M Newman : Vertex similarity in networks. *Phys. Rev. E*, 73(2):026120, février 2006.

[101] Michael Molloy et Bruce Reed : A critical point for random graphs with a given degree sequence. *Random Structures Algorithms*, 6(2-3):161–180, mars 1995.

[102] Lada A. Adamic et Eytan Adar : Friends and neighbors on the web. *Social Networks*, 25:211–230, 2001.

[103] Tao Zhou, Linyuan Lü et Yi-Cheng Zhang : Predicting missing links via local information. *The European Physical Journal B*, 71(4):623–630, octobre 2009.

[104] Jon M. Kleinberg : Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, septembre 1999.

[105] Josep M Pujol, Ramon Sangüesa et Jordi Delgado : Extracting reputation in multi agent systems by means of social network topology. *In AAMAS '02 : Proceedings of the first international joint conference on Autonomous agents and multiagent systems : part 1*. ACM Request Permissions, juillet 2002.

[106] Guanfeng Liu, Yan Wang et Mehmet A. Orgun : Trust transitivity in complex social networks. *In AAAI*, 2011.

[107] Sepandar D. Kamvar, Mario T. Schlosser et Hector Garcia-Molina : The eigentrust algorithm for reputation management in p2p networks. *In Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 640–651, New York, NY, USA, 2003. ACM.

[108] Frank Edward Walter, Stefano Battiston et Frank Schweitzer : A model of a trust-based recommendation system on a social network. *Autonomous Agents and Multi-Agent Systems*, 16(1), février 2008.

[109] Bader Ali, Wilfred Villegas et Muthucumaru Maheswaran : Proceedings of the 2007 conference of the center for advanced studies on Collaborative research - CASCON '07. *In the 2007 conference of the center for advanced studies*, page 288, New York, New York, USA, 2007. ACM Press.

[110] Ugur Kuter et Jennifer Golbeck : SUNNY : a new algorithm for trust inference in social networks using probabilistic confidence models. *In AAAI'07 : Proceedings of the 22nd national conference on Artificial intelligence*. AAAI Press, juillet 2007.

[111] Barbara Carminati et Elena Ferrari : Enforcing relationships privacy through collaborative access control in web-based Social Networks. *In 5th International ICST Conference on Collaborative Computing : Networking, Applications, Worksharing*. IEEE.

[112] Jennifer Golbeck : Trust and nuanced profile similarity in online social networks. *ACM Transactions on the Web*, 3(4):1–33, septembre 2009.

[113] Talel Abdessalem et Imen Ben Dhia : A reachability-based access control model for online social networks. *DBSocial '11 : Databases and Social Networks*, juin 2011.

[114] Daniel Gayo-Avello : All liaisons are dangerous when all your friends are known to us. *arXiv.org*, cs.SI, décembre 2010.

[115] Amel Bouzeghoub Sadok Ben Yahia Sana Hamdi, Alda Lopes Gancarsky : IRIS : A Novel Method of Direct Trust Computation for Generating Trusted Social Networks. *In Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on*, pages 616–623. IEEE Computer Society, 2012.

[116] Surya Nepal, Wanita Sherchan et Cecile Paris : STrust : A Trust Model for Social Networks. *In Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on*, pages 841–846, 2011.

[117] Wenjun Jiang et Guojun Wang : SWTrust : Generating Trusted Graph for Trust Evaluation in Online Social Networks. *In Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on*, pages 320–327. IEEE Computer Society, 2011.

[118] Eric Freyssinet : Reflexions pour un plan d'action contre les botnets. *In Symposium sur la Securite des Technologies de l'Information et de la Communication (SSTIC)*, 2010.

[119] R Rajaram S Appavu alias Balamurugan : Data mining techniques for suspicious email detection : a comparative study. *In European Conference on Data mining*, Portugal, 2007.

[120] Minoru Sasaki et Hiroyuki Shinnou : Spam detection using text clustering. *In 2005 International Conference on Cyberworlds (CW'05)*, pages 4 pp.–319. IEEE.

[121] Yue Zhang, Jason I Hong et Lorrie F Cranor : CANTINA : a content-based approach to detecting phishing web sites. *In WWW '07 : Proceedings of the 16th international conference on World Wide Web*, pages 639–648, New York, NY, USA, 2007. ACM Press.

[122] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues et Virgilio Almeida : Detecting spammers on Twitter. *In Proceedings of the 7th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010.

[123] Alex Hai Wang : Don't follow me : Spam detection in twitter. *In Int'l Conference on Security and Cryptography (SECRYPT)*, 2010.

[124] Justin Ma, Lawrence K Saul, Stefan Savage et Geoffrey M Voelker : Identifying suspicious URLs : an application of large-scale online learning. *In Proceedings of the 26th Annual International Conference on Machine Learning*, pages 681–688, New York, NY, USA, 2009. ACM.

[125] Thomas A Phelps et Robert Wilensky : Robust Hyperlinks : Cheap, Everywhere, Now. *In Digital Documents and Electronic Publishing*, Munich, Germany, septembre 2000.

[126] Seung-Taek Park, David M Pennock, C Lee Giles et Robert Krovetz : Analysis of lexical signatures for improving information persistence on the World Wide Web. *ACM transaction on information systems*, 22:2004, 2004.

[127] Gerard Salton et Christopher Buckley : Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[128] Charles Perez, Marc Lemercier, Babiga Birregah et Alain Corpel : SPOT 1.0 : Scoring Suspicious Profiles on Twitter. *In 2011 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 377–381. IEEE, 2011.

[129] Robert Layton, Paul Watters et Richard Dazeley : Automated unsupervised authorship analysis using evidence accumulation clustering. *Natural Language Engineering*, 19:95–120, 1 2013.

[130] Robert Layton, Paul Watters et Richard Dazeley : Authorship Attribution for Twitter in 140 Characters or Less. *In 2010 Second Cybercrime and Trustworthy Computing Workshop*, pages 1–8. IEEE, juillet 2010.

[131] Robert Layton, Charles Perez, Babiga Birregah, Paul Watters et Marc Lemercier : Indirect information linkage for OSINT through authorship analysis of aliases. *The International Workshop on Data Mining Applications in Industry & Government (DMApps 2013)*, pages 1–12, janvier 2013.

[132] Robert Layton, Paul Watters et Richard Dazeley : Recentred local profiles for authorship attribution. *Journal of Natural Language Engineering*, 2011.

[133] Karen Sparck Jones : A Statistical Interpretation of Term Specificity and its Application in Retrieval. 1972.

[134] Mark E. J. Newman : Fast algorithm for detecting community structure in networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 69(6 Pt 2):066133–066133, mai 2004.

[135] Mark E. J. Newman et Michelle Girvan : Finding and evaluating community structure in networks. *Phys. Rev. E*, août 2003.

[136] Mathieu Bastian, Sebastien Heymann et Mathieu Jacomy : Gephi : An Open Source Software for Exploring and Manipulating Networks. *In International AAAI Conference on Weblogs and Social Media (AAAI)*, 2009.

[137] Jay A. Kreibich : *Using SQLite*. O'Reilly Media, 1st édition, 2010.

[138] E Raad, R Chbeir et A Dipanda : Discovering relationship types between users using profiles and shared photos in a social network. *Multimedia Tools and Applications*, 2011.

[139] James A Hanley et Barbara J McNeil : The Meaning and Use of the Area under a Receiver Operating (ROC) Curvel Characteristic. *Radiology*, 143(1):29–36, 1982.

[140] Daniel Gayo Avello : All liaisons are dangerous when all your friends are known to us. *In Proceedings of the 22nd ACM conference on Hypertext and hypermedia (HT)*, pages 171–180, New York, NY, USA, 2011. ACM.

[141] Kyumin Lee, James Caverlee et Steve Webb : The social honeypot project. *In Proceedings of the 19th international conference on World wide web (WWW)*, page 1139, New York, New York, USA, 2010. ACM Press.

[142] Mihaela Vorvoreanu : Managing identity across social networks. *Poster session at the 2010 Conference on Computer Supported Cooperative Work*, 2010.

[143] Charles Perez, Babiga Birregah et Marc Lemercier : Familiar strangers detection in online social networks. *In International Conference on Advances in Social Networks Analysis and Mining- ASONAM 2013*, page 8. IEEE/ACM.

Charles PEREZ

Doctorat : Réseaux, Connaissances, Organisations

Année 2014

Approche comportementale pour la sécurisation des utilisateurs de réseaux sociaux numériques mobiles

Notre société doit faire face à de nombreux changements dans les modes de communication.
L'émergence simultanée des terminaux nomades et des réseaux sociaux numériques permet désormais de partager des informations depuis presque n'importe quel lieu et potentiellement avec toutes les entités connectées.
Le développement de l'usage des smartphones dans un cadre professionnel ainsi que celui des réseaux sociaux numériques constitue une opportunité, mais également une source d'exposition à de nombreuses menaces telles que la fuites d'information sensible, le hameçonnage, l'accès non légitime à des données personnelles, etc.
Alors que nous observons une augmentation significative de la malveillance sur les plateformes sociales, aucune solution ne permet d'assurer un usage totalement maîtrisé des réseaux sociaux numériques. L'apport principal de ce travail est la mise en place de la méthodologie (SPOTLIGHT) qui décrit un outil d'analyse comportementale d'un utilisateur de smartphone et de ses contacts sur les différents médias sociaux. La principale hypothèse est que les smartphones, qui sont étroitement liés à leurs propriétaires, mémorisent les activités de l'utilisateur (interactions) et peuvent être utiles pour mieux le protéger sur le numérique.
Cette approche est implémentée dans un prototype d'application mobile appelé SPOTLIGHT 1.0 qui permet d'analyser les traces mémorisées dans le smartphone d'un utilisateur afin de l'aider à prendre les décisions adéquates dans le but de protéger ses données.

Mots clés : criminalité informatique - analyse des données - réseaux sociaux (internet) – smartphones - identité numérique.

A Behaviour-based Approach to Protecting Mobile Social Network Users

Our society is facing many changes in the way it communicates. The emergence of mobile terminals alongside digital social networks allows information to be shared from almost anywhere with the option of all parties being connected simultaneously.
The growing use of smartphones and digital social networks in a professional context presents an opportunity, but it also exposes businesses and users to many threats, such as leakage of sensitive information, spamming, illegal access to personal data, etc.
Although a significant increase in malicious activities on social platforms can be observed, currently there is no solution that ensures a completely controlled usage of digital social networks. This work aims to make a major contribution in this area through the implementation of a methodology (SPOTLIGHT) that not only uses the behaviour of profiles for evaluation purposes, but also to protect the user. This methodology relies on the assumption that smartphones, which are closely related to their owners, store and memorise traces of activity (interactions) that can be used to better protect the user online.
This approach is implemented in a mobile prototype called SPOTLIGHT 1.0, which analyses traces stored in users' smartphone to help them make the right decisions to protect their data.

Keywords: cybercrime - data analysis - online social networks – smartphones - online identities.

Thèse réalisée en partenariat entre :

Aube Conseil Général

utt université de technologie Troyes

Ecole Doctorale "Sciences et Technologies"