# Analyse d'activité et exposition de la vie privée sur les médias sociaux

Younes Abid*, Abdessamad Imine*, Amedeo Napoli*,
Chedy Raïssi*, Marc Rigolot**, Michaël Rusinowitch *

*INRIA-Nancy, 54600 Villers-lès-Nancy
prenom.nom@inria.fr
** Fondation MAIF, 50 avenue Salvador Allende , 79000 Niort
prenom.nom@fondation.maif.fr

Dans ce travail [1] nous avons réalisé une enquête sur l'usage des médias sociaux pour déterminer les sujets sensibles et détecter des vulnérabilités de vie privée. Nous avons collecté 232 réponses complètes et valides d'utilisateurs de médias sociaux. La corrélation par rapport à la variable "âge" entre notre échantillon et la population des internautes français [2] est 0,8 et s'élève à 0,95 pour les internautes de plus de 18 ans. Nous avons analysé le comportement des internautes sur les médias sociaux suivant quatre critères et défini les sujets sensibles comme étant ceux qui appartiennent à *au moins deux* ensembles parmi les suivants : L'ensemble $E_{discussion}$ des sujets dont la fréquence de discussion globale est inférieure à la fréquence moyenne moins l'écart type. Dans notre étude $E_{discussion}$ est {"*Argent*", "*Achats*", "*Religion*", "*Rencontre*"}. L'ensemble $E_{activite}$ des forums et sites internet dont le taux d'activité globale est inférieur au taux moyen moins l'écart type est {"*Sortie, Rencontre, Chat*", " *Philosophie, Religion, Libre pensée*"}. L'ensemble $E_{anonyme}$ des sites et forums dont le taux de publication anonyme (sans identification ou avec des profils anonymes) dépasse la moyenne de 8.7 % est {"*Économie, Politique, Actualité, Infos*", "*Philosophie, Religion, Libre pensée*", "*Jeux, Musique, Film, Humour, Art, Livre*", "*Santé, Courses, Cuisine, Maison, Astuce*"}. Nous avons simulé des entretiens individuels directifs pour identifier les sujets évités sur les réseaux sociaux. Les participants avaient la possibilité de développer une réponse libre dans sa forme et dans sa longueur. Nous avons ensuite analysé les réponses et défini l'ensemble $E_{evite}$ des sujets évités en se basant sur les sujets mentionnés et les mots répétés fréquemment par les répondants. $E_{evite}$ = {"*Politique*", "*Religion*", "*Vie personnelle et familiale*" , "*Vie sentimentale et sexuelle*", "*Vie financière*", "*Actualité*", "*Vie professionnelle*", "*Santé*", "*Art*" "*Vacances et Voyages*" }

Nous avons normalisé les séries de pourcentages calculées dans notre enquête pour les rendre comparables. La transformation des données a consisté à diviser chaque valeur par la moyenne de la série. Étant donné un sujet : moins il est discuté sur des médias sociaux plus il est sensible. Aussi, nous définissons le *coefficient de sensibilité* $C$ par opposition au taux de discussion sur les médias sociaux. Le tableau 1 classe les sujets et les données personnelles inférées du plus sensible vers le moins sensible sur les médias sociaux.

1. réalisé dans le cadre d'un projet financé par la Fondation MAIF.

2. http://vingthuitzerotrois.fr/marketing/attachment/repartition-age-reseaux-sociaux/

| Sujet $x$ | données personnelles inférées | Coefficient de sensibilité $C(x)$ |
|---|---|---|
| Religion | les opinions philosophiques ou religieuses | 2.25 |
| Argent | la situation financière | 2.18 |
| Politique | les appartenances politiques | 2.08 |
| Rencontre | la vie sentimentale et les rencontres | 2.00 |
| Achats | les achats et les dépenses | 1.85 |
| Santé | la santé | 1.63 |

TAB. 1 – *Ordre décroissant des données sensibles.*

Un attaquant peut réaliser les étapes suivantes pour inférer des informations personnelles sensibles et divulguer l'identité de sa cible.

1. **Associer des profils sur plusieurs médias sociaux.** Les résultats du sondage montrent que 65.75 % des répondants utilisent les mêmes pseudonymes ou des pseudonymes qui se ressemblent sur plusieurs médias sociaux. 52.05 % des internautes sondés utilisent les mêmes adresses email pour créer plusieurs profils (l'adresse email est un exemple d'attribut).
2. **Construire un profil complet.** 72.16 % des répondants ont créé des profils qu'ils n'utilisent plus et dont ils ignorent l'existence.
3. **Construire le réseau d'amitié de la cible à travers les médias sociaux.** 90 % des utilisateurs des réseaux sociaux sondés ont des amis en communs entre plusieurs réseaux.
4. **Collecter des données sensibles à travers les profils associés et les liens d'amitié directe.** 84.04 % des utilisateurs des réseaux sociaux sondés publient des photos d'autres personnes sans demander leur accord.
5. **Identifier la cible à travers ses attributs, interactions et amis directs.** 35.34 % des répondants utilisent leur vrai nom comme pseudonyme et 11.64 % utilisent leur vraie photo.
6. **Attaquer itérativement les amis indirects.** 61.03 % des répondants trient leurs amis et partagent les mêmes points de vue qu'eux formant ainsi des communautés.
7. **Identifier la communauté à laquelle appartient la cible.** 67.14 % des répondants affirment que leurs amis sont amis entre eux.
8. **Identifier les membres de la communauté pour identifier la cible.** 56.34 % des internautes sondés sont amis avec leurs collègues, camarades, voisins et membres de famille sur le même profil.

# Summary

Anonymous use of Social network do not prevent users from privacy risks resulting from infering and cross-checking information published by themselves or their relationhips. With this in mind we have conducted a survey in order to measure sensitiveness of personal data published on social media and to analyze the users behaviors. We have shown that 76 % of internet users that have answered the survey are vulnerable to identity or sensitive data disclosure. Our study is completed by the description of an automatic procedure that shows how easily these vulnerabilities can be exploited and motivates the need for more advanced protection mechanisms.