FACEBOOK & NETWORK ANALYTICS

THE DARK SIDE OF FACEBOOK : LA CONNAISSANCE AU-DELA DES APPARENCES

VANDY BERTEN

Résumé Le présent document a pour but de sensibiliser le lecteur à la quantité

au-delà de ce que les utilisateurs indiquent de leur plein gré.

Des techniques de *network analytics* permettront de déduire beaucoup de choses à cible », telles que les différents groupes sociaux auxquels elle appartient ou le degré de proximité avec ses amis.

Des *méthodes d'inférence* peuvent se baser sur les informations publiées par ses amis pour identifier parmi les groupes sociaux ceux correspondant au travail, au

En utilisant des outils de *web crawling*, il sera possible de reconstituer une grande

Nous verrons par ailleurs que le même genre de technique permettra de comprendre les relations entre le
Facebook.

Samenvatting Dit document heeft als doel de lezer te sensibiliseren voor de aanzienlijke hoeveelheid informatie die men over hem kan terugvinden op de sociale netwerken, in het bijzonder op Facebook. We zullen zien dat men veel meer kan terugvinden dan wat de gebruikers zelf hebben aangegeven.

Met "network analytics"-technieken kan veel afgeleid worden over een "doelwit", zoals de verschillende sociale groepen waar hij deel van uitmaakt of hoe close hij is met zijn vrienden.

Inferentie-methodes kunnen gebaseerd zijn op de gegevens gepubliceerd door zijn vrienden om de sociale groepen te identificeren die overeenstemmen met werk, school, familie, hobby's, ...

Door "web crawling"-tools te gebruiken, zal het mogelijk zijn om een groot deel van het vriendennetwerk van een persoon opnieuw samen te stellen, zelf als hij ervoor gekozen heeft om zijn vriendenlijst niet te publiceren of om zijn foto's te verbergen.

We zullen overigens zien dat dezelfde soort techniek ervoor kan zorgen dat we de relaties tussen de verschillende groepsleden of de fans van een Facebookpagina beter kunnen begrijpen.

Table des matières

1. Introduction

Ce document est un complément de la Research Note 34 « Social Media & eGovernment ».

Dans le document en question
p

Avec la version Web, cependant, on peut se
quelconque et, sans en être ami, aller voir
paramètres de confidentialité choisis par cette personne, on pourra ne rien y voir, y voir toute la liste, ou encore uniquement les amis mutuels.

Dans le présent document, nous allons montrer dans un premier temps
peut obtenir comme information à propos de son propre réseau Facebook, en se

très similaire pourra être faite avec un réseau Twitter.

Ensuite, nous verrons dans la section 3

Nous verrons deux astuces pour y arriver. La première réside dans le fait que si A est ami avec B et que A cache sa liste d'amis mais pas B, on ne peut pas voir qu'ils sont amis sur le profil de A, mais on pourra le voir sur le profil de B (Section 3.1). La seconde se base sur les suggestions d'ajout que Facebook propose (Section 3.3).

4

se basant sur les informations diffusées par ses amis, grâce à des techniques

Dans la section 5, nous verrons comment des techniques similaires peuvent être utilisées pour mieux comprendre la communauté des utilisateurs tournant autour Facebook.

Nous insistons sur le fait que le but de ce document n'est pas de donner des outils à des gens mal-intentionnés (les techniques présentées ici sont largement connues de la communauté des hackers), mais d'attirer l'attention du citoyen « lambda » sur ce qui peut être fait à partir des informations qu'il dissémine sur les réseaux sociaux.

Par ailleurs, les techniques présentées ci-dessous se basent uniquement sur la

nous ne ferons usage de failles de séc
technique de *phishing.*

2. Explorer son propre réseau

« arcs »). À
l e société :

de

Dans le monde du Web, un
informatique au-
sont les utilisateurs et les liens sont -
orienté : si A est ami avec B, B est également ami avec A) ; sur Twitter, les arcs sont les relations de « suiveur-suivi » (on a ici un réseau orienté : A peut suivre B sans que B ne suive A en retour). O
on peut établir des relations entre un ensemble de pages Facebook et les utilisateurs qui les aiment (font un *like*), ou entre des *posts*[1] et les utilisateurs les ayant aimés ou commentés.

Dans cette section, nous allons voir comment on peut visualiser son propre réseau

connaissances et sur ses amis eux-mêmes. La pr
ses données depuis Facebook. Dans la suite, nous nous intéresserons au compte Facebook « `vandy.berten` »[2]. Une partie du texte qui suit sera rédigé à la première personne, de façon à alléger les phrases, permettant de dire « mon réseau » à la place de « le réseau de notre cible ».

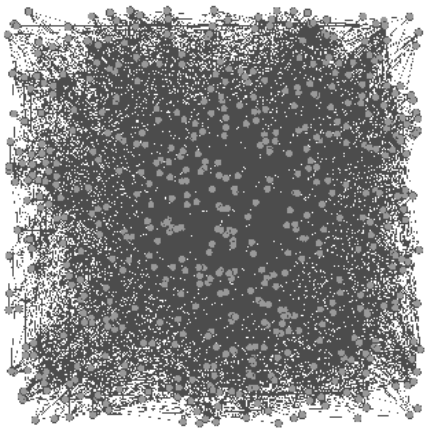


Figure 1- Graphe importé dans Gephi

Il existe plusieurs outils pour ce faire[3], tels que *Netvizz*[4] ou *NameGenWeb*[5] dernier que nous utilisons ici, avec lequel nous avons sélectionné tous les attributs possibles et choisi le résultat au format *GraphML* les

sont également amis entre eux. Remarquons que les amis de mes amis sent
pas dans le résultat. *NameGenWeb* propose également un outil de visualisation du réseau, mais nous avons

[1] Message écrit (publiquement) sur une page Facebook.

[2] C

[3]
r ainsi que les connexions entre ceux-ci. La future version, qui sera obligatoire en avril 2015, ne semble plus le permettre. Il ne sera

[4] https://apps.facebook.com/netvizz/

[5] https://apps.facebook.com/namegenweb/

Gephi[6], logiciel open source spécialisé dans la visualisation de graphes, dans lequel nous importons le résultat fourni par *NameGenWeb* (ce document

auprès des puristes, nous ne ferons pas de différence entre « graphe » et « réseau » dans la suite de cet article.

Comme on peut le voir sur la Figure 1 n dans *Gephi* ne semble pas très

un algorithme de *layout* *Gephi* nous en propose une quinzaine. Nous sélectionnons « *Force Atlas* », qui convient très bien pour ce type de graphe.

2.1. Modularisation

Le résultat est un petit peu meilleur, mais peu explicite. Nous allons maintenant lancer un algorithme de modularisation, *clusters*, -à-

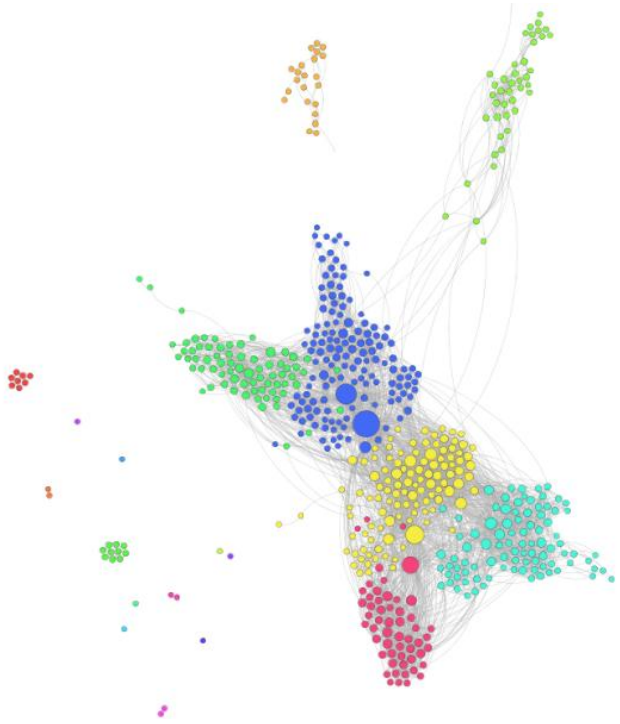


Figure 2 - Résultat du réseau de vandy.berten dans Gephi. Les couleurs font apparaitre les partitions distinctes, la taille des nœuds donne une indication de leur centralité.

beaucoup de connexions entre eux. Ceci se fera en cliquant sur le bouton « *Run* » à côté de « *Modularity* » (onglet « *Statistics* »), puis en coloriant le graphe dans *Partition*, sur la base de « *Modularity Class* ». Pour mieux comprendre ce

faire apparaître les labels (icône avec un T ; nous suggérons par ailleurs de choisir « *Hide non-selected* »

Notons -à-dire la personne dont

volontairement pas présent

cas, il y aurait un arc entre ce n

[6] « Quick Review 65 :» consacrée à *Gephi* : http://www.smalsresearch.be/publications/document/?docid=118

2.2. Premières constatations

-à-dire un ensemble de personnes se connaissant mutuellement), certaines connectées

Facebook ne connaissant (en tout cas sur Facebook) aucune autre personne du réseau. robablement de gens rencontrés à une occasion très ponctuelle,

ant de voir à quel point les couleurs utilisées représentent en effet des groupes sociaux de la vie réelle :

. Des gens partageant donc quelque -être des convictions politiques ou religieuses

Nous pouvons également voir que la plus grande composante connexe[7] du réseau reprend 95 « *Connected Components* »). Ce qui veut dire que chaque personne parmi ces 95 % est amie avec un de mes amis, qui est

composante. Le plus long chemin a une longueur de 12 sauts (mesure « *Eccentricity* » « *Avg. Path Length* » 11 de mes amis pour connecter les deux amis les plus éloignés de mon réseau. Notons que ça ne remet en rien en cause la théorie des six degrés de séparation[8], qui veut que chaque personne sur terre est connectée à chaque autre par une

pas un autre chemin plus court reliant mes deux amis éloignés, mais composé de personnes avec lesquelles je ne suis pas ami .

2.3. Identifier les amis proches ?

Dans un grand nombre de cas, il est proches » (bien que ça soit difficile à définir formellement). Dans le domaine de la théorie des graphes, on parle de « centralité ». Il existe plusieurs façons de mesurer cette centralité : une de ces mesures est la centralité de degré, qui mesure le nombre

Faceboo
(sur Facebook).

On peut visualiser ses amis proches en se servant du *ranking* (sélectionner le petit diamant, puis « *Degree* »), qui va faire apparaître en grand une personne avec laqu (voir Figure 2)

c
partie de ma famille et des participants à ladite activité).

[7] Une composante connexe est un sous-

connexe.

[8] http://fr.wikipedia.org/wiki/Six_degr%C3%A9s_de_s%C3%A9paration

2.4. Des amis communs inattendus ?

Gephi permet

vacances. On va pour

betweeness centrality). Pour effectuer cette mesure, *Gephi* ayant

iarité sera sur un grand nombre de plus courts istincts.

Dans *Gephi*, on obtient « *Avg. Path Length* ». Ensuite, on applique le *ranking* correspondant. En affichant la liste, on e

degré, mais très vite apparaissent des noms surprenants, qui ne sont pas spécialement proches. En regardant de plus près, on voit

2.5. D'autres notions de centralité

Comme évoqué ci-

téressons principalement au réseau Facebook

2.5.1. Degree centrality

La définition la plus simple mesure le nombre de voisins (*likes* ou *friends* sur Facebook, *followers* On

mesure globalement (nombre

followers Twitter) ou localement (parmi les amis de la cible).

Un message diffusé par un *degree centrality* sera vu par beaucoup de monde.

Un message diffusé par un

followers , mais seulement quelques-uns dans le réseau analysé aura moins

de *followers*, donc la plupart dans le réseau cible.

Dans un réseau personnel

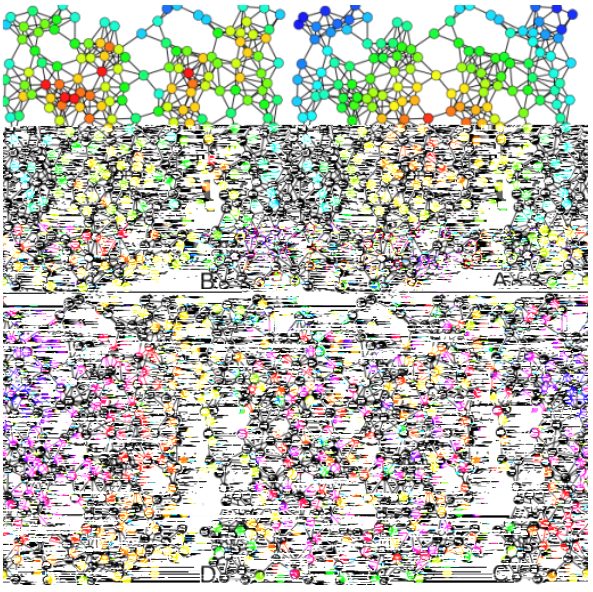


Figure 3 - A: Degree centrality ; B : Closeness centrality ; C : Betweeness centrality ; D: Eigenvector centrality
Image : http://en.wikipedia.org/wiki/Centrality

(dans lequel la cible est exc

2.5.2. Closeness centrality

closeness centrality »

2.5.3. Betweenness centrality

La *betweeness centrality* intermédiarité, compte le nombre de plus
e
Elle m ou connecteur. Un
betweenness centrality est lié en général à plusieurs *clusters*.
Sur Facebook, ce sera typiquement un ami ayant des amis communs inattendus.

2.5.4. Eigenvector centrality

Mesure basée sur les valeurs propres de la *matrice* décrivant le graphe. Elle se base sur le f diffuse » vers ses voisins. Par exemple,

son *eigenvector centrality* . Une
technique proche de cette notion est utilisée pour le « PageRank » de Google.

2.5.5. Social neighbors

Proposé par WolframA [9], cette

également des amis de A. Si cette mesure est faible, cela veut dire que A connait la grande majorité des amis de B. Cela peut donner une idée du niveau de proximité ou de dépendance de B par rapport à A.

2.5.6. Excentricité

esure de centralité, mais permet
de (la
plus courte)
rtante, plus certains
Une excentricité faible caractérise un réseau
fortement connecté.
général du réseau.

2.5.7. Autres mesures

On peut aussi étudier les messages envoyés p
personne : nombre de fois que les tweets sont retweetés, commentés, mis en favoris

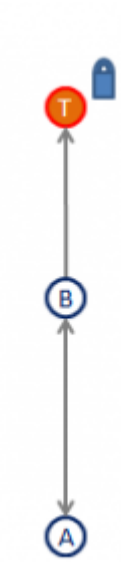
[9] http://www.wolframalpha.com/facebook/

3. Reconstituer le réseau d'un compte Facebook

extrayait elle-même les informations nécessaires. Nous allons voir maintenant que, reconstituer une grande partie de son réseau.

3.1. Via les amis mutuels

En mai dernier, le chercheur en sécurité Shay Priel[10] faisait remarquer ce qui lui semblait être une vulnérabilité de Facebook (et que Facebook considère comme une fonctionnalité) permettant de reconstituer la liste d'amis de n'importe quel profil. Supposons que l'on s'intéresse à une cible T, qui cache sa liste d'amis. Si on se rend sur https://www.facebook.com/T/friends (remplacer « T » par un nom d'utilisateur Facebook pour obtenir un véritable exemple), une liste vide s'affichera.



Supposons maintenant que soit parvenu à déterminer qu'un certain A est ami avec T et que ce A ne cache pas sa liste d'amis. Si on se rend sur sa liste d'amis (https://www.facebook.com/A/friends), on pourra y trouver T, ce qui indiquera que T est un ami de A. Étant donné que la relation d'amitié sur Facebook est symétrique, on saura également que A est un ami de T.

Relâchons maintenant cette contrainte et supposons que A ne soit pas forcément ami avec T, mais qu'ils aient uniquement des amis en commun. En rajoutant à l'adresse ci-dessus « ?and=T » (soit, donc https://www.facebook.com/A/friends?and=T), et si A accepte toujours de publier sa liste d'amis, on obtiendra une liste de profils qui sont des amis mutuels de T et A (B dans le schéma ci-contre), et donc, par définition, de T également. Notons que seuls apparaîtront dans cette liste ceux qui ont gardé publique leur liste d'amis.

Si A a bien été choisi, on vient donc d'obtenir quelques amis de T. Il suffit maintenant, pour chaque ami B de A, de refaire la même chose que pour A (https://www.facebook.com/T/friends?and=B). On obtiendra probablement en partie les mêmes amis, mais avec un petit peu de chance, également l'un ou l'autre nouveau (en plus de A que obtiendra nécessairement). Pour autant que puisse trouver quelques candidats de départ, on pourra ainsi, de proche en proche, construire une partie du réseau de la « cible ». Nous verrons plus bas que nous avons, de cette façon, retrouvé 80 % des amis ayant leur liste d'amis publique du profil de l'auteur de ce document.

Notons qu'avec le processus ci-dessus, on trouve no graphe, mais également les connexions : on peut établir une connexion entre B et tous les profils repris sur la page https://www.facebook.com/T/friends?and=B.

Le processus décrit ci-dessus peut être très laborieux. Nous attirons cependant l'attention sur le fait que l'utilisation de script *crawler*[11], permettant une automatisation de cette construction, comme celui proposé par Shay Priel, est

[10] http://blog.cyberint.com/2014/05/facebook-hidden-friends-vulnerability.html

[11] Web, en se déplaçant de liens en lien.

contraire aux conditions d'utilisation de Facebook (point 3.2)[12]. L'auteur s'expose donc à une suspension de son compte Facebook sans préavis : « *You will not collect users' content or information, or otherwise access Facebook, using automated means (such as harvesting bots, robots, spiders, or scrapers) without our prior permission*. ».

3.2. Initialisation du processus

Pour initier ce processus, il faut trouver l'un ou l'autre profil ayant des amis communs avec T (mais pas forcément ami direct avec T). Une façon simple est de se servir de l'outil « *Graph Search* »[13]. Une requête telle que « *People tagged in photos of T* » liste toutes les personnes taguées dans des photos publiques sur lesquelles T est également tagué. Il y a donc des chances qu'il y ait l'un ou l'autre ami (d'amis) de T parmi celles-ci.

On peut également se baser sur l'appartenance à des groupes et choisir comme requête « *People of groups T joined* » ou « *People who like [une page que T pourrait aimer]* ». Il est également possible de se baser sur le nom de famille (« *Find all people named [Nom de famille de la cible]* »), dans l'espoir de tomber sur l'un ou

s'agit d'un nom de famille répandu). De multiples possibilités sont disponibles, y compris celles d'avoir déjà suffisamment d'information sur la cible pour en connaître quelques amis.

Notons qu'avec cette technique, dont l'efficacité est illustrée plus bas, rien ne permet à la cible de se rendre compte que quelqu'un tente de reconstituer son réseau.

Nous avons appliqué la technique présentée ci-dessus (en partant des personnes taguées) sur le profil de l'auteur de ce document, dont la liste d'amis est cachée. La -ci, 298 seulement ont rendu leur liste d'amis publique. Nous avons retrouvé grâce à cette méthode 249 amis, soit 44 % du total (561), mais 83 % des amis avec une liste publique (298), qui sont les seuls qu'il est possible de retrouver par cette méthode.

3.3. Via les suggestions de Facebook

Une seconde méthode, présentée par Irene Abezgauz[14] se repose sur les suggestions d'amis que Facebook fait en permanence pour aider les utilisateurs à étendre leur réseau. Elle consiste à créer un compte Facebook « bidon » et à envoyer à partir de ce compte une demande d'amitié à la cible. Même si la cible ignore la requête ou la refuse, étant donné que le compte « attaquant » est vierge, le seul élément dont Facebook dispose pour suggérer des nouveaux amis est la liste d'amis de la cible. Et c'est bien ce qu'on observe... Il suffit de se rendre maintenant sur https://www.facebook.com/friends/requests/ pour voir apparaître une série de suggestions qui ne sont autre que des amis de la cible. Ils n'apparaissent cependant pas tous : toujours sur le compte `vandy.berten`, 89 suggestions ont été faites, toutes correctes. La même requête faite à d'autres moments donne exactement le même résultat, mais un essai avec un autre compte « bidon » donne d'autres résultats : 99 suggestions nous ont été faites avec cet autre compte, dont près de la moitié ne faisait pas partie des suggestions du premier compte.

[12] https://www.facebook.com/legal/terms

[13] http://www.smalsresearch.be/la-vie-privee-selon-facebook/

[14] http://www.quotium.com/resources/facebook-vulnerability-discloses-friends-lists-defined-as-private/

Par ailleurs, parmi ces quelque 150 amis listés, 35 n'avaient pas été trouvés par la méthode présentée dans la section précédente. Et en réappliquant l'algorithme de recherche d'amis mutuels sur ces nouveaux amis, nous avons obtenu au total 296 amis, soit 99 % des 298 amis publiant leur liste d'amis ! Nous aurions probablement pu encore augmenter ce score en créant de nouveaux comptes « bidon ».

Notons que cette méthode est moins transparente que la précédente, puisque la cible reçoit une demande d'amitié, qui pourrait sembler suspecte.

3.4. Approximation de la structure

En appliquant l'analyse de modularité (voir Section 2.1) au graphe ainsi créé (voir Figure 4), nous avons pu constater que, en dehors des profils tout à fait isolés, 91 % des profils ont été classés dans la même classe que lorsque nous avions la totalité du graphe. Par ailleurs, près de la moitié des différences provient de deux de mes cercles sociaux ayant beaucoup de connexions, avec un certain nombre de personnes faisant partie des deux.

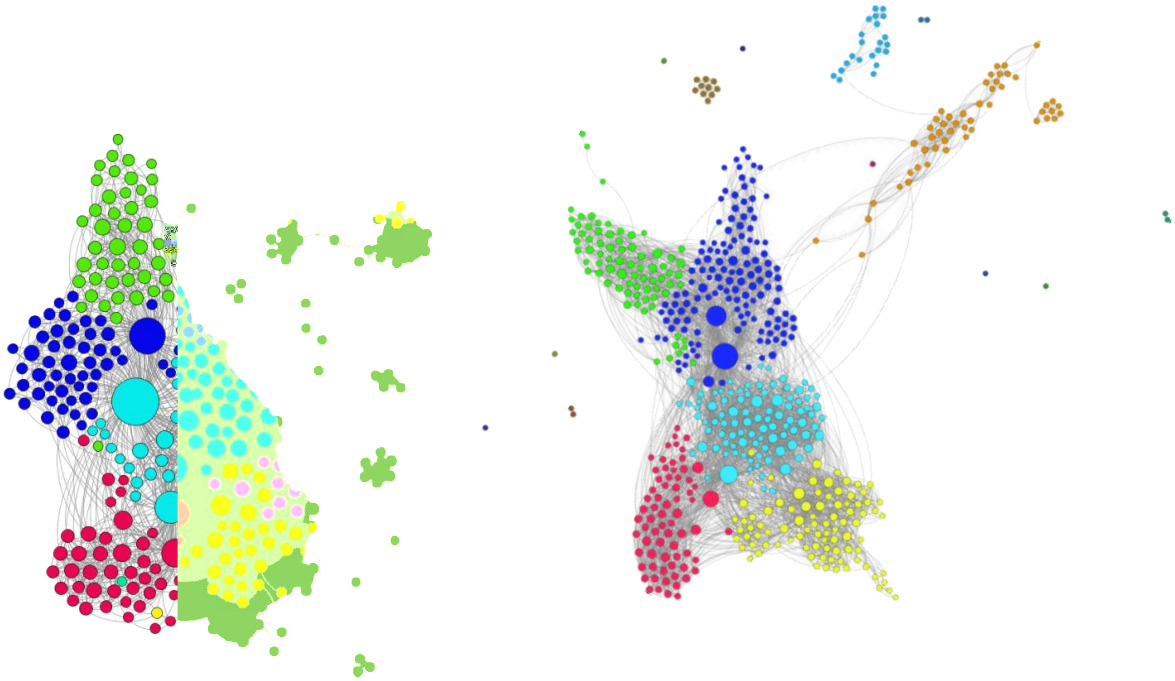


Figure 4 - Gauche : réseau approximé ; Droite - réseau original

Le tableau suivant reprend plus en détail l'exactitude des résultats. Les deux
rt au total ayant été
classés dans le groupe de la ligne. La troisième colonne indique la proportion de

groupe qu'avec le graphe complet.

	Graphe complet	Graphe extrait	% égaux
Partition 1	25 %	26 %	91 %
Partition 2	16 %	14 %	88 %
Partition 3	13 %	15 %	91 %
Partition 4	13 %	14 %	100 %
Partition 5	12 %	13 %	85 %
Partition 6	12 %	13 %	100 %

Tableau 1 - Degré d'approximation du graphe extrait

3.5. Deviner des amis

Nous avons essayé d'aller plus loin, en cherchant à savoir s'il était possible de trouver des amis qui rendent leur liste privée, à partir du moment où l'on acceptait un certain degré d'erreur dans le résultat (c'est-à-dire potentiellement considérer comme ami quelqu'un qui ne l'est pas). L'idée est la suivante : si un grand nombre d'amis de T disent être amis avec A, il est probable que A et T soient en réalité amis, même s'ils cachent tous les deux leur liste d'amis. Pour ce faire, nous avons listé la totalité des amis de chaque ami de T publiant sa liste d'amis et avons étudié les recouvrements.

En fixant un seuil à 15 amis communs, nous avons obtenu 72 amis potentiels de `vandy.berten`. Parmi ces 72, 61 faisaient réellement partie de mes amis sur Facebook, mais ; j'en connaissais personnellement 8 autres et j'estime qu'ils me connaissent également (sans que nous ne soyons amis sur Facebook). Seuls 3 m'étaient très vaguement ou pas du tout connus. En considérant que quelqu'un s'intéresse à mon réseau social réel (c'est-à-dire les personnes que je connais, peu importe que nous soyons amis sur Facebook ou non), on aurait pu retrouver près de 370 de mes amis, avec seulement 3 erreurs.

En fixant par contre le seuil à 10, 173 auraient été trouvés, dont 120 effectifs sur Facebook (avec liste privée), 32 connus en vrai mais pas amis sur Facebook et 21 inconnus ou très vaguement connus. Le taux d'erreur devient alors plus élevé, et le problème est que seule la cible (moi, en l'occurrence) est capable de dire si l'approximation est correcte ou non, ce qui complique la tâche de l'attaquant. Une recherche plus approfondie, à partir d'un certain nombre de cibles collaborantes, prêtes à faire la même expérience, permettrait sans doute de déterminer un seuil raisonnable ou optimal, en fonction de paramètres à déterminer. Mais le travail, qui doit largement être mené à la main, est très laborieux.

Nous n'avons pas étudié de méthodes plus avancées, mais il serait probablement possible d'améliorer cette technique, par exemple en pondérant les amis communs en fonction de leur centralité : une personne ayant 10 amis communs avec une haute centralité de degré moyen (ou une autre définition de centralité) avec la cible a sans doute plus de chance d'être amie avec la cible qu'une personne ayant 10 amis en commun ayant tous une faible centralité de degré.

Figure 5 - Friendship page

On pourrait aussi imaginer explorer la *Friendship page* (https://www.facebook.com/USER1?and=USER2) qui reprend les histoires en commun entre deux comptes, même
On y trouve les photos publiques où les deux ont été tagués, mais également les

3.6. Aspects temporels

Sur la *Friendship page* présentée ci-dessus, on peut voir apparaitre, dans le cas où

devenus amis. En collectant cette information partout où cela est possible, on obtient
moment donné, mais on sait
comment il a évolué. La dynamique des réseaux est souvent très instructive. Par

partition sont apparues en même temps, ce qui permettrait de penser q
passé quelque chose dans la vie de la « cible » à ce moment-là.

voir quand sont apparues des amitiés, on ne
peut par contre pas voir celles qui ont disparu.

beaucoup de potentiel.

3.7. Tests de reconstitution

Le 24/9/2014, nous avons appliqué cette méthode au réseau du compte `vandy.berten`. Nous avons pour ce faire développé un script en Python, utilisant la librairie « Selenium », qui

tagged in Vandy Berten photos », obtenu % - 83 % des cs) et 1569 connexions (31 %) ;

- En rajoutant la méthode « People you may know », nous montons à 296 amis, soit 99 , et 1638 connexions (32 %) ;
- Si nous changeons les paramètres du compte de façon à rendre publique la de `vandy.berten`, nous obtenons 561 t 4111 connexions (81 %).

Avec le script que nous avons développé (mais que nous ne publierons pas, pour éviter toute utilisation inappropriée), voici la séquence de commandes que nous utilisons :

1. Recherche de candidats en se basant sur les personnes apparaissant sur les photos où la cible est « tagguée » :

   ```
   python fb-crawler.py -username ***** -password ***** -query "People tagged in Vandy Berten's photos" -output infosession-vandy-query-photos.txt
   ```

   20 résultats, 36 secondes

2. À partir des candidats, rechercher par amis mutuels les amis cachés de la cible :

   ```
   python fb-crawler.py -username ***** -password ***** -action hidden -target vandy.berten -profilesfile infosession-vandy-query-photos.txt -output infosession-vandy-hidden.txt
   ```

   247 résultats, 12 minutes

3. Recherches grâce à « People you may know » (avec deux comptes différents) :

   ```
   python fb-crawler.py -username ***** -password ***** -action pymk -output infosession-vandy-pymk.txt
   ```

   89 résultats, 30 secondes

4. Fusion des recherches par les photos + « people you may know »

   ```
   python merge.py infosession-vandy-pymk* infosession-vandy-hidden.txt > infosession-vandy-all.txt
   ```

   284 résultats

5. Re-crawling :

   ```
   python fb-crawler.py -username **** -password ***** -action hidden -target vandy.berten -profilesfile infosession-vandy-all.txt -output infosession-vandy-all-r2.txt
   ```

   296 résultats, 12 minutes

6. Recherche du réseau :

   ```
   python fb-crawler.py -username ***** -password ***** -action network -target vandy.berten -profilesfile infosession-vandy-all-r2.txt -output infosession-vandy.gdf
   ```

   , 26 minutes

devons faire une *friend request* ; cela pourrait également être automatisé, mais nous

4. Inférence

Une fois les partitions identifiées, on peut dans certains cas comprendre à quoi elles correspondent dans la réalité. L'idée est la suivante : supposons que je puisse déterminer que parmi les 200 amis d'un certain Albert, 20 pratiquent le fitness (par exemple parce qu'ils sont publiquement membres d'un groupe autour du fitness, ou plus simplement l'affichent explicitement sur leur profil). Tel quel, ça ne me donne pas beaucoup d'information sur Albert. Il se pourrait que 10 % des gens en général pratiquent le fitness (cette affirmation purement arbitraire, sans aucun fondement statistique, n'est qu'un exemple illustratif !). Il n'est donc pas étonnant de retrouver cette proportion dans les amis d'Albert. Par contre, si grâce aux techniques mentionnées ci-dessus, j'identifie une « partition » de 30 personnes, particulièrement bien connectées et dont Albert fait partie, et que parmi ces gens-là, j'en identifie 20 affichant leur intérêt pour ce sport, il y a beaucoup de chances qu'il s'agisse d'un groupe d'adeptes du fitness, qui se connaissent probablement grâce à leur pratique. Je pourrais donc en déduire qu'il est « probable » (nous n entrerons pas dans des calculs de statistiques et probabilités) qu'Albert pratique également ce sport.

En analysant ainsi les différentes partitions du réseau d'Albert, ou pourra sans doute dans certains cas identifier, par exemple, une école fréquentée par la majorité des gens d'une partition, une université... et donc déduire pas mal de choses sur la « victime ».

) à partir de quelques-uns

de cette partition.

4.1. Via les noms de famille

Sur le réseau extrait du profil `vandy.berten`, nous avons dans un premier temps voulu voir si l'on pouvait identifier la partition correspondant à sa famille. L'idée étant que, dans cette partition, il est fort probable de trouver un plus grand nombre de personnes portant le même nom de famille. Et ça marche ! La proportion de personnes portant un des 3 noms de famille les plus fréquents de la partition est de très loin plus importante que dans toutes les autres partitions. Le même constat a pu être fait sur d'autres réseaux.

4.2. Via la section « à propos »

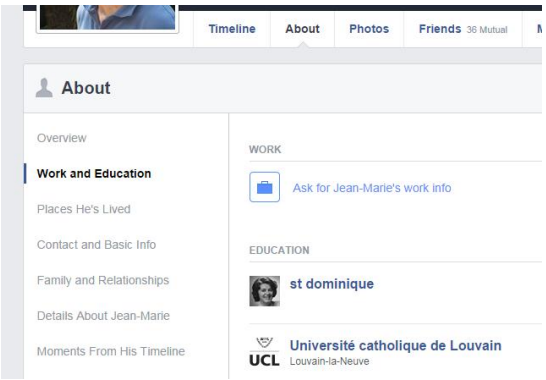


Figure 6 - Section "About" > "Work and Education"

Nous avons aussi examiné la section « à propos » (*About*) de chacun des profils de mon réseau, en ne considérant que ceux accessibles publiquement. Nous nous sommes intéressés de plus près à quatre partitions : une correspond à l'école où effectué toute ma scolarité, l'autre mes études supérieures et deux autres à des activités qui ne sont pas référencées dans la section « à propos ».

Dans chacun de ces quatre réseaux, entre 40 et 50 % des profils ont une section « à propos » publique. Le Tableau 2 reprend la proportion des partitions reprenant soit le nom de mon école secondaire, soit celui de mon université.

Sur l'ensemble de mes amis, les deux noms (école et université) apparaissent pour chacun dans 12 % de leur section « à propos ». Cette proportion est cependant nettement plus basse pour une grande partie des partitions (les 1 et 4 du tableau ci-dessous, mais toutes les autres avaient le même genre de valeur). Par contre, on trouve une partition dans laquelle 53 % de personnes publiant leur parcours mentionnent le nom de mon école secondaire, et une partition avec 57 % de références à mon université, soit des proportions largement supérieures à toutes les autres partitions. Il n'y aurait donc aucune difficulté à déduire mon parcours d'enseignement rien qu'en observant mon réseau.

	[École secondaire]	[Université]	Taille partition	% publique
Total	12 %	12 %	562	38 %
Partition 1	4 %	3 %	142	53 %
Partition 2	53 %	25 %	83	46 %
Partition 3	0 %	57 %	42	50 %
Partition 4	8 %	5 %	95	37 %

Tableau 2 – Inférence : deux partitions (Partition pour l'école secondaire et 3 pour l'université) ressortent très nettement, ce qui permet de caractériser ces partitions.

4.3. Via les groupes

La même chose a pu être faite en se basant sur les groupes : la requête de Graph Search « *XXX's groups* » liste tous les groupes Facebook auxquels une personne appartient. Cette information est publique pour chaque utilisateur, sauf pour les groupes « secrets » (la majorité des groupes étant soit ouverts tout le monde peut en voir les membres ainsi que le contenu des discussions ou fermés tout le monde peut en voir les membres, mais pas le contenu des discussions). On a retrouvé le nom d'une organisation dont je fais partie dans au moins un nom de groupe de 13 % de mes amis. Mais, alors que ce nombre descend à 0 ou 1 % pour la plupart des partitions, il monte à 29, 36 et 50 % pour trois d'entre elles, qui correspondent effectivement à mes contacts au sein de cette organisation.

L'accès aux *likes* étant également largement public, le même genre de travail peut être fait à partir des *likes* de chaqu .

4.4. Mauvaises utilisations possibles

On pourrait se dire que ça n'est en rien problématique : il ne s'agit pas là d'informations très sensibles. Mais on pourrait aller plus loin : certaines personnes, certes peu nombreuses, affichent explicitement leur orientation religieuse. D'autres ne le font pas explicitement, mais une analyse rapide de leur profil permet de le deviner : pages liées à l'église ou à un Imam influent *likées*, commentaires ou photos au contenu à connotation explicitement religieuse... une analyse, certes très laborieuse, des différentes partitions du réseau d'une cible pourrait permettre de deviner sa religion, même si elle ne donne aucune information à ce propos sur son

profil. Et il en va de même pour les préférences politiques ou l'orientation sexuelle. Il est bien évident que dans une démocratie digne de ce nom, il n'y a pas grand-chose à en craindre. Mais c'est sans doute moins le cas dans des pays où la liberté d'expression est plus que restreinte. Sans parler d'un malfrat qui voudrait mieux connaître la famille de sa cible avant de commettre un méfait. Il n'y a pas beaucoup de doute sur le fait que certaines polices ou autres services de renseignement se servent de ce genre de méthodes pour mieux cibler un suspect.

Par ailleurs, beaucoup de tentatives de piratage de type *phishing* ou hameçonnage à l'heure actuelle sont anonymes. Mais le jour où un courriel incitant quelqu'un à installer un logiciel ou à se connecter sur une copie de Facebook commencera par « Bonjour XXX [votre nom] ; ce week-end j'ai discuté avec YYY [un ami avec une haute centralité de degré] et ZZZ [un autre ami ayant beaucoup d'amis en commun avec YYY] à la fête de AAA [grâce aux photos ou évènements publiées par YYY ou ZZZ], qui m'ont dit que tu serais intéressé par ... », il y a de fortes chances pour que les gens se méfient beaucoup moins. De façon générale, les techniques de piratage par ingénierie sociale[15] ont beaucoup de nouvelles perspectives à explorer.

Un document de la police fédérale consacré au vol d'identité[16] écrit que l'intérêt pour des criminels « d'avoir accès aux réseaux sociaux est double : c'est d'abord d'avoir une connaissance de la hiérarchie ou des structures de réseaux d'amis des personnes concernées, et ensuite de dérober des renseignements classiques personnels via la technique du phishing ».

[15] http://www.smalsresearch.be/social-engineering-watch-out-because-there-is-no-patch-for-human-stupidity/

[16] « », Info nouvelles 2057, Police fédérale, juillet 2011 (http://polsupport.be/FILE/DGS/DSI/2057F.pdf)

5. Identifier les communautés autour d'une page Facebook

Avec des techniques similaires à celles présentées ci-dessus, il est possible de

quelconque de profils, par exemple

ne sont

« officiellement

une ensemble de personnes, ce qui peut être très intéressant pour un service de police tentant de mieux comprendre un

organisation rivale.

5.1. Reconstitution sur tous les utilisateurs liés à une page

sur une page (commentaires, *likes*, partages

connexions entre les utilisateurs de cette « communauté ». Une méthode similaire avec celle présentée plus ha

: avec un *web crawler*, on extrait la liste complète des amis de chaque utilisateur faisant partie de cette

a

Ainsi, on trouvera de façon certaine chaque connexion entre deux utilisateurs publiant leur list

publiant leur liste, on peut malgré tout évaluer le taux de complétude de

on obtiendrait, pour chaque connexion (le fait que A et B soient amis) deux arcs (A est da

Nombre …	Formule	AFSCA	Smals
… de nœuds	N	342	208
… de nœuds publics	P	136 (39 %)	101 (48 %)
… d'arc	A	97	587
… de connexions	C	81 (19.7 % sym)	453 (29 % sym)
… moyen d'arcs / nœuds	$\frac{A}{P}$	97/136 = 0.713	587/101 = 5.812
… approximé d'arcs	$\frac{N \times A}{P}$	342x0.713 = 243	208 x 5.812= 1208

… approximé de connexions	$\frac{N \times A}{2P}$	243/2 = 121.5	1208/2 = 604
Taux de complétude	$C \Big/ \frac{N \times A}{2P}$	81/121.5 = 66.7 %	453/604 = 75 %

Tableau 3 Calcul du taux d'approximation

it

publics, chacun de ceux-ci possède cinq arcs au sein de la communauté. On pourrait donc extrapoler et estimer que cette moyenne est également respectée pour les on

500 connexions (la moitié de 1000 arcs) et nous en avons déjà trouvé 400. Nous aurions donc un taux de complétude (approximé) de 80 %.

Le Tableau 3 donne deux exemples de reconstitution, le premier avec la page de

de Smals.

5.2. Reconstitution des fans (likers) d'une page

Une méthode similaire permet de reconstit
principe uniquement accessible aux administrateurs.

Il existe (au moins) deux méthodes : la première se sert de Graph Search, la

page web.

Pour la première méthode, on commencera par une simple requête « *People who like PageXXX* ». Malheureusement, cette page ne liste pas la totalité des fans, mais uniquement un sous-ensemble (vraisemblablement aléatoire). Effectuer la même requête plusieurs fois de suite ne donnera quasiment aucune différence. Par contre, on peut légèrement varier la requête, avec par exemple « *People who are men and like PageXXX* » suivi de « *People who are female and* » ou « *are single* », « *are married* », « *live in Belgium*
différent, avec un large recouvrement, mais également quelques différences.

Par ailleurs, la même requête depuis des comptes différents donnera parfois des réponses légèrement différentes.

Le Tableau 4 présente les résultats de cette démarche avec la page Facebook de
.

Requête : People who …	**Nombre de profils**	**Nouveau par rapport au 1er**
like AgenceAlimentaire	420	/
are men and like…	181	23
are women and like …	293	37
live in Belgium and like …	82	0

speak French and like …	90	11
speak Dutch and like …	25	2
are single and like …	25	1
are married and like …	51	7

Tableau 4 - Reconstitution des fans de la page de l'AFSCA

Remarque : les nouveaux profils fan sont indiqués par rapport aux 420 profils trouvés dans la première requête, pas par rapport à la fusion des résultats précédents.

Au total, nous obtenons 480 profils distincts, soit 51 % des 932 fans de la page.

liste de fans :

https://www.facebook.com/plugins/fan.php?connections=100&id=AgenceAlimentaire

Pour extraire une liste de fans, il suffit maintenant de rechercher les adresses de profil dans le code HTML généré. Le plugin génère 100 profils ; une seconde requête en génère une nouvelle, en général quasiment identique mais pas dans le même ordre.

La fusion des deux méthodes nous donne 496 fans, soit 53 %.

En refaisant les mêmes requêtes le lendemain, nous avons obtenu un nouveau résultat très légèrement différent : un nombre similaire de profils fans, avec au total 9 nouveaux par rapport à ce que nous avions obtenu au préalable.

Le surlendemain, on passait à 517 fans, mais le nombre réel est lui passé de 932 à 958. Nous passons donc à 54 %.

6. Conclusions

Comme nous le montrons dans ce document, les réseaux sociaux, et Facebook en particulier, regorgent d'informations sur ceux qui y sont inscrits. Il n'est même pas nécessaire d'en être un utilisateur actif : a qu'une utilisation très passive de Facebook et ne poste quasiment jamais rien, certainement pas publiquement.

L'utilisateur n'a que très peu de contrôle sur ses informations, car elle sont souvent implicites (émanant de la structure des connexions ou de la fréquence d'apparition d'information) ou dans les mains d'autres utilisateurs (il est difficile de demander à chacun de ses amis de masquer sa liste d'amis, de cacher son appartenance à tel ou tel groupe ou de supprimer toutes ses photos).

Ceci n'est bien évidemment pas un appel au boycott : à chacun de faire la balance entre ce que les réseaux sociaux lui apportent en termes humains, de loisir ou autre, et l'information très précieuse que l'on donne à son propos, qui dépasse très largement celle qu'on fournit explicitement. Nous espérons que ce document permettra au lecteur de le faire de façon plus éclairée !

Par ailleurs, on pourrait facilement imaginer que des services de police ou de renseignement se servent de ce genre de techniques pour obtenir des informations

techniques soient déjà utilisées. Elles paraissent évidemment très intrusives, mais ne le sont probablement pas plus que des

La section Recherche de Smals produit régulièrement des publications couvrant de nombreux domaines du marché IT actuel. Vous pouvez obtenir ces publications via le site web de la section Recherche :

http://www.smalsresearch.be