Karimi Belhal
Master IREN

Exercie de mémoire :

La divulgation des informations personnelles sur les réseaux sociaux

Serge Pajak
Université Paris-Sud

Table des matières

Remerciements

Je tiens particulièrement à remercier Serge Pajak, professeur au master IREN et chercheur en économie à l'université Paris-Sud pour son encadrement tout au long de cette étude et sa sélection d'articles très intéressants sur le sujet.

Je remercie également Yannick Perez pour son tutorat à SUPELEC concernant la rédaction d'un exercice de mémoire.

I – Introduction

Après avoir fait une étude bibliographique de la divulgation des données personnelles sur les réseaux sociaux, on a pu remarquer que les individus sont parfois irrationnels dans leur manière d'interagir sur Facebook, par exemple.

La question de divulgation des données personnelles sur les réseaux sociaux pose inévitablement la question des motivations pour révéler de telles données sur internet. On peut alors se demander pourquoi les gens acceptent de publier aussi souvent leur vie privée et leurs données personnelles. Kirman et Acquisti, deux professeurs d'économie, ont mis en exergue ces questions en y mêlant la rationalité et de manière plus générale l'économie comportementale.

Dans cette deuxième partie de l'étude, j'ai décidé de me concentrer sur les réglages de vie privée sur Facebook. En effet, mon objectif est de comparer le comportement de classes de personnes, que l'on définira en fonction de leur âge, profession… afin de savoir les différences de réglages de vie privée et donc de divulgation selon la classe de la personne.

Pour cela, je me suis fixé des groupes bien définis et ai élaboré un script en langage de programmation Python afin de scraper les profils en question et d'en retirer le pourcentage de personnes dévoilant leurs informations privées.

Nous savons maintenant les différentes pratiques des utilisateurs de Facebook, il nous reste à comprendre ce qu'ils divulguent exactement et donc ce qui est consultable à partir d'un profil public. Même si certaines informations sont non renseignées, que l'on arrivera donc pas à extraire, il peut y avoir une différence entre les personnes qui décident de ne pas les publier et celles qui les renseignent sur leur profils mais les bloquent à la vue de tous.

Cette étude sera d'un intérêt double. Premièrement on pourra identifier les différences de comportement selon les appartenances aux groupes, qu'il s'agisse d'étudiants, d'adultes ou de sexe différent. Deuxièmement, comprendre de manière plus générale comment les utilisateurs du réseau social utilisent Facebook. Nous scraperons alors un nombre important de profils pour avoir des résultats significatifs et après avoir fait des hypothèses sur les comportements de divulgations, on essaiera de valider ces dernières à l'aide de nos statistiques.

II – La divulgation des informations

a) Les pratiques

Dans « Student Awareness of the Privacy Implications When Using Facebook », Tabreez Govani et Harriet Pashley élaborent un panel de pratiques des étudiants concernant les réseaux sociaux. Ils se basent sur de multiples sondages et expérimentations pour donner une idée précise de la manière avec laquelle les jeunes étudiants ou non utilisent les réseaux sociaux.

Parmi les raisons d'utiliser facebook, deux sont évidentes chez les jeunes étudiants. D'abord l'effet de mode : Facebook a su s'intégrer dans la tendance générale du web et des liens sociaux virtuels tout en innovant son image d'une pointe de jeunesse et de liberté. Et ensuite le fait que Facebook soit devenu très vite populaire si bien qu'il soit devenu un standard en termes de réseau social, tout le monde l'utilise. Dans leur article, Govani et Pashley décrivent comment les étudiants interagissent via facebook et par quel procédé les informations personnelles s'échangent, parfois à leur insu. La question qu'ils vont se poser entre autres est de savoir si ces jeunes font réellement attention aux réglages de vie privée ou bien ils sont tout simplement non informés.
Tout d'abord, Facebook est arrivé dans les universités américaines alors déjà équipées d'intranet. L'intranet est utilisé par tous les étudiants pour s'échanger des informations, des fichiers dans un but purement scolaire. On y trouve également les adresses mail de tous les élèves accompagnées d'une photo d'identité. On aurait pu imaginer que Facebook serait passé invisible auprès des étudiants déjà satisfaits de leur intranet, cependant Facebook arrive avec une particularité qui a su faire la différence : la profondeur de ses informations, « Reasons for Facebook's popularity as a campus networking tool over other campus networking tools include the depth of information that is encouraged by the site to be shared ».
C'est avec cet état d'esprit que Govani et Pashley nous dévoile leur étude sur les utilisations de Facebook par les étudiants à travers de nombreux sondages sur le campus de la Carnegie Mellon University. Les études, faites sur près de 50 élèves de l'université utilisant facebook révèlent plusieurs résultats importants et une classification des utilisations des réseaux sociaux. Tout d'abord, ces élèves ont dû répondre à un questionnaire sur ce qu'ils étaient prêts à montrer aux élèves de leur école. Les résultats montrent que plus de 80% des élèves montrent leurs centres d'intérêts, nom, école, date de naissance et leur ville de résidence. De plus, plus de la moitié laisse publique leurs jobs et opinions politiques. Cette même étude met en avant le fait que les élèves ne dévoilent par leur mail et numéro de téléphone. Ces données sont considérées comme trop privées et seraient la première source de vol d'identité. Alors que les résultats de Govani et Pashley montrent que seul 11% des élèves informent sur Facebook de leur adresses postales, une autre étude menée par Gross et Acquisti en 2005 affirme que plus de la moitié des personnes sont prêts à mettre leur adresse sur leur profil. Cette confrontation de deux sondages et résultats soulève la question des limites des sondages et déclarations faites par l'échantillon. De plus, il faudrait avoir une description plus grande sur l'échantillon pour pouvoir expliquer ces différences de résultats.
Ensuite, l'étude nous décrit, à l'aide de déclarations des étudiants, les types d'amis qui se connectent via Facebook. Une majorité, non surprenante, acceptent ou demandent en amis des personnes qu'ils côtoient à l'école ou qu'ils connaissent. Cependant une part non négligeable de ces étudiants rajoutent en amis des personnes qu'ils n'ont rencontrés qu'une seule fois ou même pas du tout. Ces derniers résultats renforcent l'étude menée par Govani et Pashley dans le but de savoir pourquoi les gens créent un profil Facebook. Ils décident de diviser ces personnes en deux catégories : A et B. La catégorie A correspond aux personnes s'inscrivant sur les réseaux sociaux car leurs amis leur ont conseillé ou parce que tout leur

entourage ont un profil. Cette même catégorie représente le haut pourcentage de personnes acceptantes des amis de l'université ou des connaissances vu à l'étude précédente. La catégorie B, elle, correspond aux individus en recherche d'amis, de relations et de réseau professionnel. On explique ainsi le pourcentage non nul de personnes qui ne se connaissent pas du tout et qui se connectent via Facebook. A l'aide de ces deux sondages, les professeurs de la CMU ont réussi à classifier les utilisateurs de Facebook en deux grandes catégories selon leur besoin et activité sur le réseau social. Enfin une dernière information importante nous est donnée dans cet article, celle concernant les réglages de vie privée. Pour introduire le problème de la loi vis-à-vis des pratiques extrêmes de Facebook en matière de vie privée, il est important de savoir si Facebook met à disposition des moyens de bloquer les informations et de les maintenir privées ou au contraire il s'agit d'un manque d'attention de la part des utilisateurs. Pour cela, nos deux économistes ont demandés à l'échantillon d'étudiants de la CMU, quels paramètres de réglages ils utilisaient. Les résultats sont sans appel, les réglages sont sous utilisés bien que disponibles. En effet, Plus de 70% des utilisateurs de Facebook ne bloquent pas leurs informations personnelles telles que leur profil entier, les évènements auxquelles ils assistent, ou les amis qu'ils ont.

Ceci pose la question du manque d'effort de Facebook sur les réglages de la vie privée et surtout de la rationalité des personnes lors de la publication de données. Ces derniers résultats nous montrent un comportement irrationnel tourné vers le gain à court terme, telle que la popularité, de mettre en ligne des données personnelles. C'est un point que je développerai plus tard en me basant sur des études de Kirman et Acquisti.

b) La rationalité face aux données personnelles:

Dans un premier temps, Alan Kirman dans « Whom or What Does the Representative Individual Represent? », nous pose les bases de la rationalité humaine. Selon lui, cette rationalité est limitée par la quantité d'informations dont ils disposent, de leur capacité cognitive, c'est-à-dire d'apprendre et de raisonner et enfin par le temps dont ils disposent pour réagir. Ce dernier paramètre est d'autant plus important lorsqu'il s'agit des réseaux sociaux où les décisions se prennent en une fraction de seconde. La capacité cognitive engendre de tels coûts que l'individu se retrouve souvent à recourir à des solutions simples et plus rapides au détriment de stratégie optimale et réfléchie. Il aura alors tendance, dans cette société de vie en communauté, à imiter ce qu'il voit.
Par exemple, un individu peut imiter rationnellement le comportement de son entourage en matière de protection de la vie privée pour éviter de supporter les coûts cognitifs associés à de telles décisions. Si personne n'adopte de solutions, il ne se protégera pas.

Tout cela engendre des distorsions psychologiques qui sont à l'origine de la divulgation des données personnelles. La valence, par exemple, fait que les individus ont tendance à penser à la bonne tournure des évènements lorsqu'ils publient quelque chose en ligne plutôt que d'imaginer le pire scénario. Cela provient de la sous-estimation des risques par l'individu. Les membres de ces réseaux sociaux vont alors faire preuve d'ignorance rationnelle lors de la publication de données si bien que cela finit par ne représenter aucun risque à leurs yeux, que cela soit à court ou long terme. En effet, à force de laisser des petites traces, qui ne représentent que des risques limités, sur internet il est possible de reconstituer la vie des individus. Ce procédé étant complexe, car il faut des efforts de recherches importants de ces traces, l'individu ne va alors pas considérer ce danger en tant que tel.

Le phénomène économique important mis en valeur ici par Kirman et qui repose sur ces distorsions psychologiques est l'actualisation hyperbolique. L'actualisation hyperbolique

explique l'incohérence temporelle à l'origine des actions de publication qui privilégient les biens faits à court terme en oubliant les répercutions à long termes. Elle consiste à actualiser la valeur présente des gains futurs. Ainsi les individus s'engagent dans des activités agréables, comme la publication de photos sur Facebook, même si ça leur coûte une perte de bien-être dans le futur.

Dans un autre registre, Alessandro Acquisti, dans son étude de l'économie de la vie privée, commence par définir la rationalité parfaite et finit par dire que cette rationalité ne peut exister chez l'être humain. Selon Acquisti, la rationalité parfaite suggère que les individus ont des préférences mesurées vis-à-vis de la vie privée, autrement dit qu'ils savent définir une limite entre leur vie privée et publique. Par conséquent, ils arrivent à peser le pour et le contre de la publication en ligne dans un but de bien-être et pour finir ils arrivent à choisir quoi publier et quoi garder secret.

Cependant, Alessandro Acquisti affirme que la rationalité parfaite n'existe pas. En effet, il émet l'hypothèse que les compromis entre les coûts et les bénéfices de publier des données personnelles ne peuvent exister étant donné que l'on ne peut pas quantifier les coûts de publication de données. Pour affirmer cela, il part du principe qu'une fois que les données sont publiées, il est impossible de revenir en arrière, de savoir qui les utilise et même de contrôler ces données. En somme il est difficile de calculer les probabilités d'occurrence d'évènements futurs provoqués par le comportement présent adopté par les individus. Par exemple, comment pouvons-nous calculer la probabilité d'être victime de fraude bancaire plusieurs années après un paiement en ligne ?

c) la loi autour de ces problèmes:

Comme mentionné précédemment, Carly Brandenburg, à la fin de son étude sur l'utilisation des données personnelles lors du processus de recrutement, soulève le problème de termes et conditions de Facebook qui seraient oui ou non entravés par les employeurs. D'une manière plus générale, il est important de se poser la question juridique ayant bien trop souvent peu de réponses claires. En effet, certains économistes tels que Caroline Lancelot et Claire Gauzente qui dans leur étude « Vie privée et partage de données personnelles en ligne : une approche typologique », affirment que même si le thème du respect de la vie privée n'est pas nouveau dans la loi (Nowak et Phelps 1992), cette notion est totalement altérée par le fait que les données personnelles sont désormais devenues la première source de valeur pour les entreprises.

En ce qui concerne l'étude de Brandenburg, cette dernière nous permet de confronter les termes et conditions d'utilisation de Facebook affirmant que « You understand that . . . programs offered by us on the Site (e.g., Facebook Flyers ...), the Service and the Site are available for your personal, non-commercial use only. », c'est-à-dire que Facebook ne doit pas être utilisé à des fins commerciales. Or, les employeurs l'utilisent à des fins commerciales pour trouver le meilleur capital humain possible parmi les candidats. Cependant, toute la difficulté de ces questions juridiques réside dans l'interprétation des termes exacts de la loi ou des conditions générales. Ici, la signification « non commercial purposes » dépend de l'interpretation que l'on en fait. La loi manque vraisemblablement de modernité sur ses sujets récents si bien que même la définition de ce qui demeure privée reste sans réponse exacte du point de vue de la justice. Plusieurs auteurs se demandent si ce qui est privée le reste tant

que personne à part la personne concerné ne le sait. Autrement dit, si dès lors que j'ai transmis une information personnelle à une personne, cela signifierait que cette information n'est plus privée et que si une personne en a connaissance, tout le monde peut en avoir connaissance. C'est ce qu'il se passe sur les réseaux sociaux tels que Facebook car les utilisateurs transmettent des informations, à l'origine, uniquement à leurs amis mais la loi n'arrive pas à mettre des limites à ces données étant déjà en ligne et par conséquent elles sont utilisables par tout internautes. Même si certains cas de procès ont donné raison aux utilisateurs d'autres les ont condamnés au profit des réseaux sociaux.

Carly Brandenburg détaille plusieurs procès afin d'illustrer la question suivante : Une fois que l'information est disponible à certains, est-elle ouverte à tous ?. Plusieurs individus ont souffert de ce type de problème. Par exemple, le cas de Nader contre General Motors illustre bien ce thème. Nader poursuivit General Motors pour entrave à la vie privée car la firme avait tenté de l'intimider suite à la parution de son livre dénonçant les pratiques dangereuses de GM. General Motors avait alors questionné l'entourage de Nader afin de récolter un maximum d'informations sur lui en prétendant aux amis de ce dernier que la firme le faisait à des fins professionnelles. Bien que Nader pense être en position de force, la loi en a décidé autrement en donnant raison à la société car selon elle, les faits partagés avec d'autres ne sont plus privés et par conséquent peuvent être connus de tous.

Dans le même cas, la cour de justice de l'état d'Ohio a déclaré que l'employé, ayant révélé à quatre collègues ses actes sexuels avec son enfant, ne pouvait pas prétendre à profiter des droits protégeant sa vie privée. La raison derrière cette décision de justice est que si un secret est révélé à certains, il est exploitable par tous.

Enfin, la littérature étant encore déficiente en termes de protections légales des données personnelles (la plupart des économistes se posent des questions sans trouver de réelles réponses), j'ai choisis d'étudier un support de travail de deux universitaires Naryshkin et Wills et d'un chercheur travaillant pour AT&T, Krishnamurthy. Leur travail repose essentiellement sur le gap qui se creuse entre la divulgation/fuite des données personnelles en ligne et les mesures de protection juridiques actuellement en place.

Tout d'abord, ils ont commencé par analyser les différents types de fuites d'informations en élaborant une classification. Pour ce faire, ils utilisent une palette de données telles que les requêtes http ou les login grâce à certains outils tels que www.alexa.com. Après avoir énuméré les fuites, ils ont travaillé sur les mesures de protection existantes à ce jour pour éviter de telles fuites. Et enfin, ils ont confronté ces mesures à l'ampleur des fuites d'informations afin d'en tirer une conclusion sur le niveau de protection actuel et proposer des solutions.

Leur étude commence par la liste exhaustive des étapes susceptibles d'être l'objet de fuites de données personnelles :

1- La création d'un compte : la plupart des sites web requiert la création d'un compte utilisateur renfermant des données personnelles telles que le nom, mail, téléphone
2- La confirmation de la création de ce compte : l'échange de mail confirmant la création du compte peut facilement être dévié car le mail passe souvent par un autre serveur lors de son acheminement entre le site et la boîte de réception de l'utilisateur.
3- L'édition du profil : une fois le compte crée, les utilisateurs ont tendance à modifier leur profil. Certains sites présentent alors des données en titre de la page profil (nom, prénom) et ainsi les rendent vulnérables à certains moyens de scrapping (code Javascript).
4- La mise en ligne de contenu : Les utilisateurs peuvent poster du contenu sur leur profil. En étudiant l'acheminement du contenu posté, les auteurs de cette étude ont détecté

que la requête utilisé est HTTP GET au lieu de POST. De plus si la page présente du contenu d'un tiers parti. Ceci encourage les fuites.

5- Recherche de termes sensibles (exemple : « Santé » ou « Race ») : Cela concerne des mots recherchés sur une page qui sont particulièrement sensible si bien que l'utilisateur est redirigé vers un autre site et un autre serveur.

Ensuite, ils énumèrent les mesures de protection existantes :

1- Blocage de requêtes venant d'une tierce partie : Des outils tels que AdBlock permettent de bloquer les requêtes de pages publicitaires
2- Paramétrage du navigateur pour bloquer les cookies
3- Désactiver les codes Javascript pour empêcher les outils de scrapping de voler les informations sur la page
4- Rendre l'IP anonyme
5- Utiliser des outils permettant de sortir du tracking

Enfin ils confrontent leurs deux listes afin de voir s'il existe un retard au niveau des mesures de protection de la vie privée des individus en fonction de l'ampleur des fuites.

En somme, de nombreuses études concernant les mesures de protections et les lois qui existent sont disponibles, cependant, la justice a encore du mal à appliquer ces lois dans des contextes technologiques avancés. La loi tarde à se moderniser ce qui rend les enjeux de protections de données en ligne parfois encore plus important afin de combler la faiblesse en termes de juridiction.

II – L'expérimentation

Notre étude consiste à concevoir un outil de scraping des profils Facebook pour en retirer les informations divulguées.

Sur plus de 3000 profils scrapés, nous pouvons émettre des hypothèses et en vérifier la pertinence avec l'interprétation des résultats. **En revanche il reste une certaine part d'inconnue dans l'interprétation des données sachant que l'une des difficultés pratiquement insurmontable est de savoir si la personne ne dévoile pas une information simplement parce qu'il ne peut pas remplir la catégorie (par exemple pas de lieu d'études antérieur) ou parce qu'il a volontairement choisi de le bloquer grâce aux réglages de vie privée offerts par Facebook.**

Ainsi, pour la suite de notre étude, nous considérerons qu'une information divulguée correspond à une information renseignée par l'utilisateur et également diffusée grâce aux réglages de vie privée.

a) Le protocole:

Le protocole d'expérimentation se distingue par deux étapes :

Tout d'abord, l'élaboration du script de scraping codé en Python et utilisant des librairies variées comme Beautifulsoup ou Mechanize afin d'organiser le code html d'une page web et en suite la définition de groupes de personnes à comparer en fonction de critères distinctifs.

En ce qui concerne le script, il fallait tout d'abord veillez à faire en sorte de scraper les profils Facebook en tant que personne connecté. En effet, les informations disponibles lorsque l'on n'est pas connecté sur le réseau sont très maigres et donc peu intéressantes.
Une fois connecté, mon outil doit me permettre d'identifier si un individu a dévoilé les informations suivantes :

- Lives in (Lieu de résidence)
- Born (date de naissance)
- Studies (Nom de l'institution actuelle)
- Studied (Ancienne formation)
- From (Ville de provenance)

Ces critères sont les plus fréquents parmi les personnes qui renseignent le plus de données. Mon script peut également renvoyer le contenu de ces cases mais l'intérêt est moindre pour notre étude.

Pour les groupes de personnes, j'ai choisi de comparer :

- Etudiants vs. Quadragénaires
- Promo école vs. Promo antérieure
- Ecole d'ingénieur vs. Ecole de commerce
- Etudiantes vs. Etudiants
- Femmes vs. Hommes

En veillant à faire le scraping d'une centaine de profils à chaque fois pour pouvoir comparer les résultats. Le nombre de profils scrapés sera indiqué lors de la comparaison des résultats.

J’ai voulu, dans cet exercice, confronter plusieurs catégories de personnes selon leur milieu professionnel ou scolaire. De plus, mon outil me permettant de distinguer les sexes, il m’est paru intéressant de visualiser les différences de comportements selon le genre de la personne.

Remarque : Plusieurs études comparatives n’ont pas pu aboutir à cause du manque de groupes publics concernant les profils recherchés. En effet beaucoup de groupes sont dits fermés et par conséquent rendent impossibles le scraping des url.
Par exemple, en voulant étudier la comparaison de comportements entre les personnes issues de l’industrie et celles du monde de la mode et du textile, je me suis rendu compte qu’il n’existait aucun groupe de plus de 20 personnes relié à des entreprises de l’énergie, par exemple, comme Areva ou GDF. Les groupes sont en général privés.

Précautions expérimentales :

- Avant d'écrire l'outil de scraping, veiller à introduire un script vous permettant de vous logger au préalable à un compte Facebook. Si vous ne le faites pas, vous scraperez des profils vus par des personnes n'étant pas inscrits au réseau social et par conséquent accéderez à très peu d'informations

- Créez un nouveau compte facebook à partir duquel vous scraperez. Cela vous évitera de scraper avec votre compte personnel et donc de tomber sur des profils « amis » avec vous, faussant alors les résultats. En effet, vous verrez bien plus d'informations sur vos amis que sur des profils publics.

- Ici, le contenu des critères (de lieu de résidence, d'études…) est peu intéressant. Vous pouvez alors vous limiter à détecter si oui ou non le mot apparait sur le profil. Dans mon cas, mon code m'indique « -1 » si le mot 'Lives' par exemple n'apparaît pas. Et me renvoie la place du mot dans la page s'il y apparait. Ci-dessous un exemple de données que je scrappe :

Lives	Born	Studies	Studied	From
142782	-1	-1	-1	143706
-1	-1	-1	-1	-1
-1	-1	-1	-1	-1
-1	-1	-1	-1	-1
-1	139629	-1	138548	-1
136876	-1	-1	135823	137799
116065	-1	-1	-1	117017
135302	-1	-1	134251	136225
155017	-1	-1	-1	155942
72105	-1	-1	-1	73027
140385	-1	-1	-1	141317
-1	-1	-1	-1	-1

Ici, par exemple, le premier profil indique son lieu de vie et d'où il vient mais pas le reste. Et ainsi de suite.

- Enfin, comptabilisez les cases « -1 » et diviser par le nombre de profils scrapés. Vous aurez alors le pourcentage de personnes n'indiquant pas le critère en question.

Voici un exemple :

	Lives	Born	Studies	Studied	From
N.A	50,00%	97,69%	98,85%	66,54%	59,62%
Renseigné	50,00%	2,31%	1,15%	33,46%	40,38%

b) Les résultats:

Voici tout d'abord le résumé des données scrapés. Il s'agit là des pourcentages de personnes ayant renseigné leurs informations personnelles. Celles qui n'ont pas divulgués correspondent simplement au tableau complémentaire.

Renseigne	membres	Lives	Born	Studies	Studied	From
Etudiants	517	**59,96%**	**2,51%**	**3,87%**	**35,40%**	**47,58%**
Quadra	111	**62,16%**	**1,80%**	**1,80%**	**5,41%**	**49,55%**
Epitech2013	216	**50,00%**	**2,31%**	**1,15%**	**33,46%**	**40,38%**
Epitech2015	319	**57,86%**	**1,89%**	**23,58%**	**25,16%**	**46,86%**
Supelec	327	**45,26%**	**0,61%**	**7,34%**	**33,94%**	**36,39%**
ESCP	502	**55,38%**	**0,40%**	**5,98%**	**26,89%**	**42,23%**
Dauphine Hommes	104	**58,65%**	**3,85%**	**20,19%**	**25,96%**	**44,23%**
Dauphine Femmes	212	**52,36%**	**0,47%**	**7,08%**	**22,17%**	**41,04%**
Hommes	544	**50,88%**	**1,33%**	**5,31%**	**30,09%**	**41,59%**
Femmes	226	**47,79%**	**0,92%**	**6,07%**	**28,86%**	**41,91%**
Total	**3078**	**53,7%**	**1,4%**	**7,5%**	**28,9%**	**43,0%**

- Etudiants vs. Quadragénaires

Dans un premier temps, nous allons nous pencher sur les différences de comportements entre les étudiants et les quadragénaires sur Facebook. Pour cela, j'ai choisi de scraper les profils des membres du groupe Challenge Centrale Lyon qui regroupe tous les étudiants participant à cet évènement sportif. Et un groupe s'appelant People That Look Good Over 40 qui regroupe des quadragénaires.

Remarque : les quadragénaires sélectionnés ont une attirance prononcée pour la beauté et le soin du corps et par conséquent, on pourrait s'attendre à un comportement sur les réseaux sociaux différents de personnes âgées de plus de 40 ans moins « trendy ».

Voici un récapitulatif des données traitées :

	membres	Femmes	Hommes
Quadragénaires	111	76	35
Etudiants	517	120	397

La proportion de femmes au sein des deux groupes s'inverse. Les étudiants participant au challenge sont en effet principalement issues d'école d'ingénieurs et par conséquent en

majorité de sexe masculin. Alors que les quadragénaires rejoignant les groupes de beauté tel que celui que j'ai sélectionné sont plus des femmes.

Attente :

Dans un premier temps, j'ai pensé que les quadragénaires auraient plus de difficulté à publier leur date de naissance. Cependant avec le contenu du groupe que j'ai choisi, très axé sur les personnes de plus de 40 ans faisant attention à leur apparence, montrer leur âge pourrait s'avérer être une stratégie gagnante pour ces membres. Nous verrons dans les résultats s'ils renseignent leur âge autant que les étudiants.
De plus, en ce qui concerne le lieu d'étude antérieur, on pourrait deviner que les quadragénaires rempliraient plus cette case et la publieraient alors que les étudiants moins, car très peu de ces étudiants ont étudié autre part avant leur école actuelle. Ainsi pour les quadragénaires on aurait une majorité de **Studied** et peu de **Studies** et l'inverse pour les étudiants qui seront plus enclins à indiquer leur école actuelle.

Voici les résultats du scraping :

- **Les étudiants :**

	Lives	**Born**	**Studies**	**Studied**	**From**
N.A	40,04%	97,49%	96,13%	64,60%	52,42%
Renseigné	59,96%	2,51%	3,87%	35,40%	47,58%

- **Les quadragénaires :**

	Lives	**Born**	**Studies**	**Studied**	**From**
N.A	37,84%	98,20%	98,20%	94,59%	50,45%
Renseigné	62,16%	1,80%	1,80%	5,41%	49,55%

Résultats :

Finalement on se rend compte, sur la base de ces résultats, qu'aucun de nos groupes de révèlent leur date de naissance (seulement 2% pour les deux groupes).
En ce qui concerne les études, les quadragénaires ne renseignent rien sur leurs études passées. Evidemment, il n'y a pas d'information sur leurs études actuelles, bien que 2% l'indiquent (Après vérification, il s'agit de personnes ayant oublié de l'enlever). Par contre, les étudiants ne renseignent pas du tout leur école actuelle mais, à 37%, divulguent leurs écoles passées. En général, il s'agit des lycées dans lesquels ils ont évolués.
Pour le reste, les proportions d'individus indiquant leur lieu de vie et d'origine sont égales. 60% divulguent leur ville de résidence et la moitié d'où ils proviennent.

- Promo 2013 vs. Promo 2015 (Epitech)

Après vérification, les promotions sont numérotées comme l'année de sortie d'école. Ainsi la promotion 2013 est déjà sur le marché de travail alors que la promotion 2015 effectue sa dernière année.

Epitech est une école post-bac d'informatique et par conséquent les groupes des promotions que j'ai scrapé contiennent des profils plutôt avisés sur l'informatique. Cela peut présenter un biais dans la comparaison avec d'autres étudiants. Ici, on ne fait pas face à ce problème car nous allons comparer les promotions d'une même école.

Voici un récapitulatif des données :

	membres	hommes	femmes
epitech2013	216	206	10
epitech2015	319	303	16

La proportion d'étudiantes est environ le même et même très bas. Ceci est logique étant donné la dominante très prononcée en informatique de cette école et attirant alors plus de garçons.

Attente :

Les deux groupes choisis pour cette étude sont les promotions d'année différente pour la même école, l'EPITECH. L'une des promotions est déjà diplômé et sur le marché du travail. L'autre effectue sa dernière année. Par conséquent, il est naturel de penser que le taux d'étudiants indiquant son établissement fréquenté auparavant (**Studied**) sera plus grand pour la promotion 2013 que 2015. Et de même pour la catégorie **Studies** qui sera divulgué par plus de membres de la promotion 2015 étant donné que la promotion déjà sur le marché n'étudie à priori plus. Pour le reste, les taux devraient rester sensiblement les mêmes.

Voici les résultats du scraping :

- **Epitech 2013 :**

	Lives	**Born**	**Studies**	**Studied**	**From**
N.A	50,00%	97,69%	98,85%	66,54%	59,62%
Renseigné	50,00%	2,31%	1,15%	33,46%	40,38%

- **Epitech 2015:**

	Lives	**Born**	**Studies**	**Studied**	**From**
N.A	42,14%	98,11%	76,42%	74,84%	53,14%
Renseigné	57,86%	1,89%	23,58%	25,16%	46,86%

Résultats :

En effet, on peut remarquer la propension des 2013 a indiqué leur éducation passée (34%). Cependant il y a une part non négligeable des 2015 (25%) qui divulguent également cette information. Il s'agit, après vérification des noms de lycées et/ou universités fréquentés lors de cursus et semestres antérieurs.
De plus, il n'y a pratiquement plus d'élèves de la promotion 2013 qui indiquent leurs études actuelles (1%) ce qui est tout à fait logique étant donné leur activité professionnelle. Et beaucoup plus de la promotion 2015 le renseigne (24%). Le fait étonnant réside dans ce faible pourcentage (24%) d'étudiants encore scolarisé et divulguant leur établissement de scolarité.

- Ecole d'ingénieur vs. Ecole de commerce

Dans cette partie il s'agit de comparer les comportements des étudiants en école de commerce et d'ingénieur. Pour cela, nous étudierons les profils appartenant aux groupes de SUPELEC, école d'ingénieur, et de l'ESCP, école de commerce. Les deux promotions sont issues de la même année, 2015.
Les groupes Facebook de ces deux promotions correspondent aux élèves de troisième année fréquentant l'établissement. Contrairement à la numérotation des promotions faites par les écoles, certains élèves des promotions 2014 sont également dans ces groupes car ayant fait une année de césure, ils se retrouvent mélangés aux élèves de la vraie promo 2015.

Voici un récapitulatif des données :

	membres	hommes	femmes
Supelec2015	327	259	68
ESCP2015	502	301	201

Attente :

Une intuition serait de penser que les élèves de Supelec, plus sensibilisés à l'informatique, présenteraient des résultats tout à fait différents de ceux de l'ESCP. Cependant, ces derniers sont, pour leur part, plus sensibilisés aux enjeux de communication, de données personnelles et de l'impact social de la divulgation des données. Il se pose ici le problème de savoir si une personne divulgue ces informations par souhait et stratégie pensée ou par manque de connaissance sur le contrôle technique des réglages de vie privée.

Voici les résultats du scraping :

- **Ecole d'ingénieur :**

	Lives	**Born**	**Studies**	**Studied**	**From**
N.A	54,74%	99,39%	92,66%	66,06%	63,61%
Renseigné	45,26%	0,61%	7,34%	33,94%	36,39%

- **Ecole de commerce:**

	Lives	**Born**	**Studies**	**Studied**	**From**
N.A	44,62%	99,60%	94,02%	73,11%	57,77%
Renseigné	55,38%	0,40%	5,98%	26,89%	42,23%

Résultats :

Le premier résultat qui nous interpelle concerne la rubrique ‘Lives’ qui est renseigné par une majorité des étudiants en école de commerce. Le pourcentage baisse en école d’ingénieur. Cela peut facilement s’expliquer par la ville correspondant à la situation géographique des écoles. L’ESCP est à Paris alors que SUPELEC a trois campus à Metz, Rennes ou Gif sur yvette. La très grande majorité des élèves de SUPELEC ne continuant pas à vivre dans l’une de ces trois ville, le sentiment d’appartenance à celle-ci n’est pas assez grande pour l’indiqué sur son profil Facebook. Le raisonnement est le même pour les élèves de l’ESCP indiquant la mention ‘Paris’ sous la rubrique ‘Lives’. D’autre part, la catégorie ‘Studied’ est plus renseigné par les élèves ingénieurs (à 34% contre 26%). Cela pourrait s’éxpliquer par le grand nombre de personnes ayant effectué des semestres à l’étranger en deuxième année et l’ayant donc divulgué. Cette pratique se fait plus en troisième et dernière année concernant les élèves de l’ESCP.
Pour le reste, les statistiques se valent et traduisent d’un phénomène commun à tous les profils sur Facebook, c’est à dire une très faible divulgation des rubriques ‘Born’ et ‘Studies’

- Etudiantes vs. Etudiants

L'outil permettant de scraper les profils des membres d'un groupe Facebook nous donne également le sexe de ces personnes. Ainsi, il s'avère être intéressant de comparer dans un premier temps les comportements de divulgation d'étudiantes en comparaison aux étudiants, pour une promotion de l'université Dauphine.
Dans un second temps nous généraliserons aves des personnes de tous horizons.

Voici un récapitulatif des données :

	membres	hommes	femmes
Dauphine2015	316	104	212

Attente :

Pour la rubrique 'Lives' on peut s'attendre à une plus grande propension à afficher son lieu de vie pour les étudiants voulant indiquer la ville dans laquelle ils sont pour faire plus de connaissances vivant aux alentours. Pour des raisons propres aux femmes, la date de naissance sera probablement plus renseignée en ce qui concerne les étudiants.

Voici les résultats du scraping :

- **Etudiants :**

	Lives	**Born**	**Studies**	**Studied**	**From**
NA	41,35%	96,15%	79,81%	74,04%	55,77%
Renseigné	58,65%	3,85%	20,19%	25,96%	44,23%

- **Etudiantes:**

	Lives	**Born**	**Studies**	**Studied**	**From**
NA	47,64%	99,53%	92,92%	77,83%	58,96%
Renseigné	52,36%	0,47%	7,08%	22,17%	41,04%

Résultats :

En effet, les étudiants renseignent d'avantage leur date de naissance et leur lieu de vie. Pour le reste les données restent sensiblement les mêmes. 59% d'étudiants contre 52% d'étudiantes indiquent où ils résident actuellement.
Un bon pourcentage de personnes indique d'où elles viennent (42 et 45%) et près de 25% des étudiant(e)s divulguent leur lieu d'études antérieur. Tout à fait dans la logique d'étudiants en promo 2015, effectuant donc leur dernière année d'études et ayant alors, probablement étudié dans plusieurs établissement auparavant.

- Femmes vs. Hommes

Ici, on va manipuler des membres de tous horizons : étudiants, étudiantes, sportifs, professionnels…en distinguant les hommes des femmes. Il s'agit de mes relations personnelles sur le réseau social. Le fait que le scraping se fasse à partir d'un compte quelconque rectifie le biais provoqué par la provenance de ces profils. Cependant, l'âge moyen reste autour de 25 ans et est vraisemblablement majoré par 30.

Voici un récapitulatif des données :

	membres
Femmes	226
Hommes	544

Attente :

Pour la rubrique 'Lives', on peut s'attendre à une plus grande propension à afficher son lieu de vie pour les hommes voulant indiquer la ville dans laquelle ils sont pour faire plus de connaissances vivant aux alentours. Pour des raisons propres aux femmes, la date de naissance sera probablement plus renseignée en ce qui concerne les hommes.

Voici les résultats du scraping :

- **Femmes :**

	Lives	**Born**	**Studies**	**Studied**	**From**
NA	49,12%	98,67%	94,69%	69,91%	58,41%
Renseigné	50,88%	1,33%	5,31%	30,09%	41,59%

- **Hommes:**

	Lives	**Born**	**Studies**	**Studied**	**From**
NA	52,21%	99,08%	93,93%	71,14%	58,09%
Renseigné	47,79%	0,92%	6,07%	28,86%	41,91%

Résultats :

Contrairement à ce que l'on prévoyait, et même contrairement aux tendances affichées entre les étudiants et les étudiantes, les femmes, dans notre cas, divulguent d'avantage leur lieu de résidence actuel (51% contre 47% des hommes).
La date de naissance est très peu divulguée dans les deux cas.
Le reste des catégories présentent les mêmes caractéristiques.
La catégorie 'Studied' est renseignée à près de 23% en accord avec l'âge moyen des profils (25 ans) étant déjà actifs dans le monde professionnel.

IV - Conclusion

Notre étude nous a permis de dessiner des tendances de divulgation des données personnelles sur Facebook. On y a confronté des classes de personnes différentes selon leur milieu professionnel ou leur âge.

De manière générale, les étudiants divulguent plus que les étudiantes et d'une promotion à l'autre au sein de la même école, les comportements changent.

Cependant, concernant ce dernier résultat, il serait pertinent de se demander si le changement de comportement est dû au changement de situation professionnelle ou aux tendances de divulgations qui évoluent d'une année à l'autre.

Ainsi, il serait intéressant d'étudier en dynamique les activités de divulgation des données sur les profils Facebook de manière à relier chaque activité à la période de l'année.

V – Bibliographie

- Balachander Krishnamurthy, « Privacy leakagevs. Protectionmeasures: the growing disconnect », Study Document

- Tabreez Govani, « Student Awareness of the Privacy Implications When Using Facebook », in the CMU Journal

- Long Jin, University of California, San Diego Yang Chen, Duke University, « Understanding User Behavior in Online Social Networks: A Survey », Study document

- Alessandro Acquisti, « The economics of privacy »

- Alessandro Acquisti, « Privacy in Electronic Commerce and the Economics of Immediate Gratification »,, Heinz, CMU

- F. Benevenuto, « Characterizing User Behavior in Online Social Networks »,

- Bertrand, Marianne and Mullainathan, Sendhil. "Do People Mean What They Say? Implica- tions For Subjective Survey Data." Mimeo, University of Chicago, 2000

- Kirman, Alan P. 1992, « Whom or What Does the Representative Individual Represent? », in the Journal of Economic Persepectives