Rapport de Stage
De-anonymisation de données et vie privée : étude d'un modèle de graphes aléatoires.

Clément Lalanne, clement.lalanne@ens.fr,

Département d'informatique de l'ENS,
École normale supérieure,
PSL Research University,
75005 Paris, France

Encadré par
Florian Simatos, florian.simatos@isae-supaero.fr,
Département d'Ingénierie des Systèmes Complexes,
ISAE-SUPAERO,
31400 Toulouse, France

Introduction

La protection de la vie privée joue un rôle primordial dans les thématiques abordées durant les dernières années en Informatique. En ce qui concerne les réseaux sociaux, ils doivent permettre à leurs utilisateurs de protéger leur vie privée. Cependant les informations qu'ils possèdent peuvent être utiles à différentes structures (gouvernementales, statistiques ...). Ainsi il faut trouver un compromis entre opacité et clarté du système. Nous étudirons ici le cas où un réseau social décide de révéler tout ou une partie de son graphe d'utilisateurs. Ceci peut être, par exemple, pour permettre des recherches sur les graphes sociaux. Cependant comme nous allons le voir de-anonymisier totalement ou partiellement ce graphe peut ne pas suffire à guarantir qu'on ne puisse pas retrouver des informations sur les utilisateurs anonymisés.

Durant mon stage je me suis d'abord penché sur deux résultats fondamentaux de [4] et [2] renseignant sur la possibilité de de-anonymiser. J'ai étendu leurs résultats à des problèmes similaires et je me suis penché sur des problèmes faisant intervenir des graphes plus complexes que ceux d'Erdös-Rényi qui sont étudiés dans ces articles. Ensuite je me suis penché sur l'étude d'un algorithme pratique de de-anonymisation présenté dans [3] et apportant certaines modifications à leurs preuves (qui étaient parfois obscures) quitte à rajouter des conditions suplémentaires. Enfin j'ai créé un simulateur en OCaml permettant de simuler des problèmes de De-Anonymisation de sorte à vérifier les résultats expérimentalement.

1 Présentation du problème

Nous considèrerons ici des graphes non orientés ayant pour ensemble de sommets l'ensemble $[n] := \{0, ..., n-1\}$ et ayant pour ensemble d'arêtes une partie de $\binom{[n]}{2}$ l'ensemble

des paires de sommets de $[n]$. Ceci signifie que nous considérons des graphes sans arêtes multiples ni boucles. L'idée du problème de de-anonymisation est qu'il existe un graphe G à n sommets qui représente le graphe absolu des relations dans un groupe d'individus (i.e. i et j sont voisins dans G si et seulement si ils sont en relation au sens du problème considéré). Le problème est que nous n'avons pas accès à G, nous n'avons accès qu'à des observations de G au travers d'un réseau social par exemple. Nous modélisons ceci par un graphe G' dans lequel les arêtes des G sont sélectionnées avec probabilité s où s dépend bien entendu de la nature de l'observation.

Nous pouvons maintenant nous attaquer à l'anonymisation : Notons S_n le groupe des permutations de $[n]$. S_n agit sur $\binom{[n]}{2}$ par A de la manière suivante :

$$\forall \sigma \in S_n, A(\sigma) : \binom{[n]}{2} \to \binom{[n]}{2}$$
$$\{i,j\} \mapsto \{\sigma(i), \sigma(j)\}$$

Anonymiser un graphe consiste à appliquer $A(\sigma)$ à toutes ses arêtes pour une permutation σ tirée uniformément et aléatoirement dans S_n (Ce qui sera par la suite décrit par la variable aléatoire Π). Dans la suite nous confondrons un graphe $G = (V, E)$ avec la fonction indicatrice de ses arêtes E.

Nous pouvons enfin définir le problème de de-anonymisation. Soit G un graphe à n sommets. Nous considérons G_a et G_b deux graphes obtenus à partir de G en sélectionant ses arêtes indépendement avec probabilités respectives s_a et s_b. Ainsi G_a et G_b sont deux observations de G via deux réseaux sociaux différents par exemple. Les arêtes de G_a et de G_b ne sont pas indépendantes. En effet la présence d'une arête dans G_a ou d'un groupe d'arêtes peut renseigner sur la présence d'arêtes dans G_b. Nous appliquons $A(\sigma)$ aux arêtes de G_a pour une permutation σ tirée uniformément et aléatoirement dans S_n. Le résultat obtenu est la variable aléatoire G_c. Un de-anonymiseur est un programme qui retrouve σ à partir de G_c et G_b ou échoue. Dans la suite nous ne considèrerons que le cas symétrique $s_a = s_b = s$ Nous pouvons aussi remarquer que la permutation ne joue en fait aucun rôle significatif. Ainsi nous pourons supposer que $\Pi = Id$.

Il existe des variantes naturelles de ce problème :

- Le problème de de-anonymisation avec graine ou "seed" est similaire à celui présenté plus haut à la différence que nous fournissons de plus au de-anonymiseur une restriction de σ.
- Le problème de $(1-\epsilon)$ de-anonymisatio, un de-anonymiseur cherche une permutation σ' qui coincide avec σ sur une proportion de taille au moins $(1-\epsilon)$ de son ensemble de définition.
- Le problème de de-anonymisation active, le programme peut influer sur G_a et G_b avant le processus d'anonymisation. Par exemple créer de nouveaux utilisateurs sur un réseau social qui pouront potentielement devenir amis avec d'autres avant qu'une version anonymisée de ce réseau social ne soit publiée. Ce problème est étudié en détail dans [1]

2 Mise en perspective des différents résultats

Dans un premier temps nous étudierons le problème sur les graphes d'Erdös-Rényi. Ainsi $(G_a, G_b) \sim ER(n, p, s)$ si G est un graphe à n sommets et dont chaque arête est

présente indépendement avec probabilité p et que ces mêmes arêtes sont sélectionnées indépendement dans G_a et G_b avec probabilité s.

Le premier article fondateur dans le domaine est [4]. Il utilise une notion naturelle qui est l'observation de la structure des graphes et la minimisation de la différence entre les deux. Ainsi leur de-anonymiseur est le suivant :

$$PG((g_c, g_b)) = argmin_\pi \Delta(g_c \circ A(\pi^{-1}), g_b)$$

avec

$$\Delta(g_1, g_2) = \sum_{e \in \binom{[n]}{2}} |g_1(e) - g_2(e)|$$

Leur résultat principal est le suivant :

Théorème 1 *Pour le problème de de-anonymisation avec $s = \omega(1)$ et $p \to 0$, si $ps\frac{s^2}{2-s} \geq 8\frac{\log n + \omega(1)}{n}$ alors asymptotiquement presque sûrement (a.p.s.) PG réussit à retrouver l'identité.*

Dans leur article, [2] arrivent à améliorer ce résultat par un facteur 2 et obtiennent le résultat suivant :

Théorème 2 *Pour le problème de de-anonymisation avec $s = \omega(1)$ et $p \to 0$, si $ps^2 \geq 2\frac{\log n + \omega(1)}{n}$ alors asymptotiquement presque sûrement (a.p.s.) PG réussit à retrouver l'identité.*

Cependant leur plus gros résultat porte sur une impossibilité de de-anonymiser. Pour y arriver ils utilisent la notion de maximum à postériori pour estimateur (MAP). Pour comprendre leur raisonnement commençons par écrire la probabilité qu'un identificateur I réussisse.

$$P(I \text{ réussit}) = \sum_{(g_c, g_b)} P((G_c, G_b) = (g_c, g_b)) P(\Pi = I((g_c, g_b)) | (G_c, G_b) = (g_c, g_b))$$

Donc en prenant pour MAP la définition suivante :

$$MAP((g_c, g_b)) = argmax_\pi P(\Pi = \pi | (G_c, G_b) = (g_c, g_b))$$

Nous avons que si MAP réussit avec probabilité x alors il existe un de-anonymiseur qui réussit avec probabilité x et de plus tout de anonymiseur réussit avec probabilité au plus x. En particulier si MAP réussit avec probabilité $o(1)$ alors tout de-anonymiseur réussit avec probabilité $o(1)$. Ils arrivent à relier l'estimateur MAP à la fonction Δ de la manière suivante :

Lemme 1 *Si $(G_a, G_b) \sim ER(n, p, s)$ alors :*

$$P(\Pi = \pi | (G_c, G_b) = (g_c, g_b)) \propto \left(\frac{p_{10}p_{01}}{p_{11}p_{00}}\right)^{\frac{1}{2}\Delta(g_c \circ A(\pi^{-1}), g_b)}$$

avec

$$p_{11} = ps^2$$
$$p_{10} = p_{01} = ps(1-s)$$
$$p_{00} = 1 - p(2s - s^2)$$

Ainsi à partir d'un certain rang MAP et PG coïncident. Remarquons maintenant que dès lors qu'ils coincident alors MAP réussit avec probabilité au plus $\frac{1}{|Aut(G_a \cap G_b)|}$, en effet tout automorphisme permet un score identique à celui de l'identité via la fonction Δ. Ainsi en comptant les sommets isolés ils arrivent au résultat suivant

Théorème 3 *Si* $(G_a, G_b) \sim ER(n, p, s)$ *avec* $p \to 0$ *et* $ps^2 \leq \frac{\log n - \omega(1)}{n}$ *alors tout de-anonymiseur réussit avec probabilité* $o(1)$

Ce résultat montre notamment la précision des résultats de [4] et [2] car seul un facteur 2 diffécencie les domaines de possibilité et impossibilité dans le régime $\frac{\log n}{n}$. Nous nous sommes intéressés aux adaptations de ces résultats dans les différentes variantes du problème présentées plus haut.

2.1 De-Anonymisation avec graine

Dans le cas de la de-anonymisation avec graine nous avons obtenu deux résultats notoires, le premier fait intervenir une graîne quelconque.

Théorème 4 *Si* $(G_a, G_b) \sim ER(n, p, s)$ *avec* $ps^2 \leq \frac{\log n - c_n}{n}$ *et* $c_n \to \infty$ *alors tout identificateur utilisant une seed de taille au plus* $\frac{1}{2}\exp(\frac{c_n - ps^2 \log n}{1 - ps^2}) - 1$ *réussit avec probabilité au plus* $\frac{1}{2}$*. Si de plus la taille de la seed est en* $o(\exp(\frac{c_n - ps^2 \log n}{1 - ps^2}))$ *Alors l'identificateur réussit avec une probabilité* $o(1)$*.*

Pour le second résultat nous avons examiné un cas que nous avons trouvé dans la litérature. Dans la réalité certains utilisateurs décident volontairement de lier leurs comptes (Youtube, Facebook ...) et ceci avec une certaine probabilité. Ainsi nous avons examiné le cas d'une graîne dans laquelle apparaît chaque utilisateur indépendement avec probabilité l. Nous avons obtenu le résultat suivant :

Théorème 5 *Si* $(G_a, G_b) \sim ER(n, p, s)$ *avec* $ps^2 \leq \frac{\log[(1-l)n] - c_n}{n}$ *et* $c_n \to \infty$ *alors tout identificateur utilisant une seed pour laquelle chaque sommet y apparaît avec probabilité l indépendement des autres réussit avec probabilité* $o(1)$*.*

Ce résultat montre que la connaîssance même d'une fraction du graphe ne modifie que peu la borne d'impossibilité.

2.2 $(1 - \epsilon)$-De-Anonimisation

En appliquant les mêmes méthodes nous sommes parvenus à un résultat d'impossibilité de $1 - \epsilon$)-de-anonymiser.

Théorème 6 *Soit* $\epsilon \in]0, \frac{1}{2}[$*. Si* $(G_a, G_b) \sim AER(n, p, s)$ *avec* $ps^2 \leq \frac{\log(\frac{1}{2(\delta+\epsilon)})}{n}$ *pour* $\delta \in]\epsilon, \frac{1}{2}[$ *alors tout* $(1 - \epsilon)$*-de-anonymiseur réussit avec probabilité* $o(1)$*.*

Cependant ici nous changeons de régime pour un régime qui décroît beaucoup plus vite.

3 De-Anonymisation pratique

Bien que les résultats de [4] et [2] donnent des bornes de possibilité de de-anonymisation, le calcul de PG est bien souvent impossible car il demande d'examider les $n!$ permutations de $[n]$. Nous avons alors décidé d'étudier un algorithme glouton présenté dans [3]. Cet algorithme prend en entrée un mapping partiel qui met en relation une fraction l des sommets et retourne un mapping étendu du premier. Il se base sur la notion de témoins de similarité.

Définition 1 *Une paire de sommets (u_a, u_b) est avec $u_a \in G_a$ et $u_b \in G_b$ est dit témoin de similarité de la paire (v_a, v_b) avec $v_a \in G_a$ et $v_b \in G_b$ si $u_a \in N_a(v_a)$, $u_b \in N_b(v_b)$ et u_a a été associé à u_b.*

Algorithm 1 Matching glouton de Korula et Lattanzi

Require: Deux graphes G_a et G_b, Un mappilg partiel S, le degré maximum D du graphe, un score minimum de matching T (en pratique $t \geq 2$ ou 3) et un nombre d'itérations maximal k.
for $i = 1, ..., k$ **do**
  for $j = logD, ..., 1$ **do**
    for All the pairs (u, v) with $u \in G_a$ and $v \in G_b$ and such that $d_{G_a}(u) \geq 2^j$ and $d_{G_b}(v) \geq 2^j$ **do**
      Assign to (u, v) a score equal to the number of similarity witness between u and v.
    end for
    if (u, v) is the pair with highest score in which either u or v appear and the score is above T **then**
      add (u, v) to L
    end if
  end for
end for
return L

Voici les résultats présentés dans [3] pour des graphes d'Erdos-Renyi. Ces résultats peuvent paraître ambigus, nous expliquerons dans la partie suivante pourquoi ils le sont et nous apporterons quelques modifications de manière à obtenir des résultats plus restrictifs mais dont la véracité est correctement prouvée.

Lemme 2 *Si $p > \frac{24}{s^2 l} \frac{\log n}{n-2}$ alors a.p.s. le nombre de témoins de similarité entre u et $\sigma^{-}1(u)$ lors de la première phase de l'algorithme est au moins $\frac{(n-1)ps^2 l}{2}$. Inversement le nombre de témoins de similarité entre u et $v \neq \sigma^{-}1(u)$ est au plus $\frac{(n-1)ps^2 l}{2}$ a.p.s.*

Lemme 3 *Si $p \leq \frac{24}{s^2 l} \frac{\log n}{n-2}$ alors a.p.s. l'algorithme ne met jamais en relation deux noeuds u et v si $v \neq \sigma^{-}1(u)$.*

Théorème 7 *L'algorithme identifie une fraction $1 - o(1)$ des noeuds a.p.s.*

3.1 Petites rectifications

Cependant il semble difficile d'identifier une fraction $1-o(1)$ des utilisateurs dans tous les régimes de probabilités pour la raison qu'il est à priori possible d'avoir une fraction constante (et supérieure à l) des sommets qui sont isolés. C'est l'idée qui est à l'origine du théorème suivant :

Théorème 8 *Si $l < \frac{1}{2}$ et qu'il existe $\delta \in]0, \frac{1}{2} - l[$ tel que $p \leq \frac{\log(\frac{1}{2(l+\delta)})}{n}$ alors l'algorithme ne peut pas identifier une fraction $1-o(1)$ des sommets.*

Cependant le résultat reste vrai lorsque $p > \frac{24}{s^2 l}\frac{\log n}{n-2}$. Ainsi nous avons le théorème suivant :

Théorème 9 *Si $ps^2 \geq \frac{24}{l}\frac{\log n}{n-2}$ alors $\frac{R}{n} \to_P 1$ où R est le nombre de sommets correctement identifiés par l'algorithme.*

4 Cas du "Stochastic block model"

Le modèle d'Erdös-Renyi, bien que instructif et facile à étudier ne reflète en réalité que peu la réalité. Premièrement pour la loi des degrés qui se rapproche asymptotiquement d'une loi de Poisson alors que les observations ont montré que la loi des degrés des réseaux sociaux est en réalité une loi de puissance. Ensuite car le modèle d'Erdös-Renyi ne montre pas la notion de comunautés. Nous allons ici étudier le cas du stochastic block model (SBM) à deux communautées. Nous définissons le problème de de-anonymisation sur le "stochastic block model" de la manière suivante. $(G_a, G_b) \sim SBM(n, k, p_1, p_2, s)$ si G_a et G_b sont obtenus à partir d'un graphe G en sélectionnant ses arêtes indépendement suivant une loi de Bernouilli de paramètre s. De plus le graphe G est construit sur un ensemble de n sommets qui sont séparés entre deux sous ensembles de k et $n-k$ sommets de sorte que la probabilité de présence d'une arête dans le graphe est p_1 si les deux sommets sont dans la même composante et p_2 sinon.

4.1 Problème de De-Anonimisation

Pour étudier la de-anonymisation du SBM nous allons utiliser une approche similaire à celle utilisée pour l'étude précédente. Nous pouvons voir le problème comme une superposition de deux graphes, le graphe intra-composantes (1) et le graphe inter-composantes (2) Nous avons alors le résultat suivant :

$$P(\Pi = \pi | (G_c, G_b) = (g_c, g_b)) \propto \left(\frac{p_{10}^{(1)} p_{01}^{(1)}}{p_{11}^{(1)} p_{00}^{(1)}}\right)^{\frac{1}{2}\Delta^{(1)}(g_c \circ A(\pi^{-1}), g_b)} \left(\frac{p_{10}^{(2)} p_{01}^{(2)}}{p_{11}^{(2)} p_{00}^{(2)}}\right)^{\frac{1}{2}\Delta^{(2)}(g_c \circ A(\pi^{-1}), g_b)}$$

En étudiant le nombre de sommets isolés de $G_a \cap G_b$ il est alors possible d'obtenir un théorème d'impossibilité de de-anonymisabilité.

Théorème 10 *Si $(G_a, G_b) \sim SBM(n, k, p_1, p_2, s)$ avec $\max(p_1, p_2) \to 0$ et $\max(p_1, p_2)s^2 \leq \frac{\log n - \omega(1)}{n}$ alors tout de-anonymiseur réussit avec probabilité $o(1)$*

4.2 Algorithme de De-Anonymisation

En reprenant la preuve de validité de l'algorithe pratique de de-anonymisation, il est possible d'adapter la preuve pour obtenir un résultat similaire à celui deja obtenu.

Théorème 11 *Si* $(G_a, G_b) \sim SBM(n, k, p_1, p_2, s)$ *avec* $\max(p_1, p_2) \to 0$ *et* $\min(p_1, p_2)s^2 \geq \frac{24}{l}\frac{\log n}{n-2}$ *alors* $\frac{R}{n} \to_P 1$ *où* R *est le nombre de sommets correctement identifiés par l'algorithme.*

5 Simulateur OCaml

Tout mon code source se trouve sur le dépot Git suivant :
https ://github.com/ClementLalanne/De-Anonymization
Pour la création de ce simulateur je me suis inspiré du travail qui a été fait sur la librairie Ocamlgraph d'Opam. J'ai essayé d'utiliser au plus le système de modules et foncteurs d'OCaml pour me permettre par la suite de pouvoir faire des modifications dans l'implémentation de diférentes parties sans avoir à tout changer. Deux problèmes algorithmiques intéressants se sont alors posés que je vais détailler ici

5.1 Générer dynamiquement S_n

Dans mes premiers tests j'avais besoin de calculer la fonction Δ et pour ce faire j'avais besoin de regarder toutes les permutations de S_n. Cependant il est impossible en pratique de construire statiquement S_n. J'ai donc du trouver une solution pour générer dynamiquement S_n (ceci revient à mettre un ordre sur S_n avec un plus petit élément et un plus grand élément et à donner une fonction "suivante" qui à partir d'une permutation calcule la permutation suivante. Pour ordre j'ai choisi l'ordre lexicographique sur les tableaux représentant les permutations avec pour plus grand élément l'identitée et pour plus petit élément l'identitée renversée. La fonction "suivante" est calculée avec l'algorithme suivant :

5.2 Tirer aléatoirement une permutation

Pour les simulations j'avais besoin de pouvoir tirer uniformément en indépendement une permutation dans S_n. Cependant S_n est de taille $n!$. Il n'est donc pas envisageable de construide entièrement S_n pour en tirer ensuite un élément uniformément et indépendement. J'ai donc utilisé l'algorithme suivant qui construit une permutation de $\{0...n-1\}$ qui suit une loi uniforme et ceci de manière indépendante des tirages précédents. Une permutation σ est représentée par un tableau t de sorte que $t.(i) = \sigma(i)$

6 Tentative de construction d'un modèle de graphe plus réaliste

Comme nous l'avons vu, le modèle d'Erdos-Renyi présente un problème qui est lié à la faible queue de la loi des degrés. Cependant le fait que les arêtes soient indépendantes les unes des autres facilite l'étude du problème. Un modèle qui se rapproche de celui d'Erdos-Renyi est celui des graphes inhomogènes. Chaque sommet se voit affecter un poids

Algorithm 2 Algorithme de successeur

Require: n et p la permutation dont on veut calculer la permutation suivante.
  s := [|p.(0)... ;p.(0)|] de taille n
  j := 0
  for $k = 0, ..., n-2$ **do**
    s.(k) := p.(k)
    s.(k+1) := p.(k+1)
    if p.(k) < p.(k+1) **then**
      j := k
    end if
  end for
  k := j
  for i = j+1 ... n-1 **do**
    if p.(j) < p.(i) **then**
      k := i
    end if
  end for
  Echanger s.(j) et s.(k)
  renverser s de j+1 à n-1
  return s

Algorithm 3 Tirage d'une permutation dans S_n

Require: n.
  t := [|0 ;... ;n-1|]
  for $i = n-1, ..., 1$ **do**
    k := un entier tiré aléatoirement et uniformément dans $\{0, ..., i-1\}$
    Echanger t.(i) et t.(k)
  end for
  return t

p_i et une arête entre deux sommets i et j est présente indépendement avec probabilité $f(p_i, p_j)$ où f est une fonction spécifique. Nous allons nous inspirer de ce modèle pour construire le notre. Nous avons besoin de plus d'une loi sur $\mathbb{N}^*$ correspondant à peu près à la loi des degrés (nous l'appelerons loi des poids). Chaque sommet se voit ensuite affecter un poids selon la loi des poids. Comme la loi des poids correspond à peu près à la loi des degrés nous allons choisir une loi de puissance pour la loi des poids. Il nous faut aussi un paramètre de contrôle pour ajuster le poids moyen en fonction de la taille du graphe.
Donnons-nous une suite K_n vérifiant $\omega(1) \leq K_n \leq o(n)$. Pour tout n considérons la loi suivante : $P_n(p = k) = \frac{1}{\zeta(\alpha_n)k^{\alpha_n}}$ avec $\alpha_n = \frac{1}{K_n\zeta(2)} + 2$. Un graphe à n sommets sera alors construit en prenant pour loi des poids P_n et pour fonction f $f(p_i, p_j) = \frac{p_i p_j}{nK_n + p_i p_j}$.

6.1 Premières propriétés

Lemme 4 *Commençons tout d'abord par quelques rappels sur les séries de Riemann et la fonction zéta en cas de convergence. La série $\sum \frac{1}{k^\alpha}$ est une série de Riemann qui converge si et seulement si $\alpha > 1$. En cas de convergence nous définissons $\zeta(\alpha)$ qui est égal à la somme de cette série. La fonction ζ est alors continue sur $]1, \infty[$. De plus en ∞ nous avons $\zeta(x) = 1 + 2^{-x} + o(2^{-x})$ et en 1 nous avons $\zeta(x) = \frac{1}{x-1} + o(\frac{1}{x-1})$. Enfin nous aurons besoin d'un équivalent du reste en cas de convergence, nous avons le résultat suivant : $R_n(\alpha) = \frac{1}{(\alpha-1)n^{\alpha-1}} + O(\frac{1}{n^\alpha})$*

Proposition 1 *Le poids moyen d'un sommet est équivalent à K_n.*

Démonstration.

$$\begin{aligned} E_n(p) &= \sum_{k=1}^{\infty} kP_n(k) \\ &= \frac{\zeta(\alpha_n - 1)}{\zeta(\alpha_n)} \\ &= \frac{\frac{1}{\alpha_n - 2} + o(\frac{1}{\alpha_n - 2})}{\zeta(\alpha_n)} \\ &= \frac{K_n\zeta(2) + o(K_n)}{\zeta(\alpha_n)} \\ &= K_n + o(K_n) \end{aligned}$$

Proposition 2 *Soit u_n une suite telle que $u_n = \omega(1)$ alors $P_n(d \geq u_n) \sim \frac{1}{u_n}$*

Proposition 3 *Soit u_n une suite telle que $u_n = \omega(1)$ alors $P(\max_{i=0\ldots(n-1)} p_i \geq u_n) = 1 - o(1)$ si $u_n = o(n)$ et $P(\max_{i=0\ldots(n-1)} p_i \geq u_n) = o(1)$ si $u_n = \omega(n)$*

Conclusion

Durant ce stage j'ai pu aborder, avec l'aide de mon encadrent, différents aspects de la de-anonymisation de graphes. Tout d'abord j'ai effectué un travail de documentation. puis nous avons repris certaines preuves et élargit certains résultats. En parallèle j'ai développé un simulateur pour pouvoir faire des tests de de-anonymisation. Enfin nous avons essayé

de créer un modèle de graphes qui soit à la fois réaliste et qui permette une étude facile du problème de de-anonymisation mais le temps nous a manqué pour poursuivre dans cette direction. Je tiens à remercier Monsieur Florian Simatos ainsi que l'ensemble du DISC de l'ISAE-Supaero pour leur accueil pendant ces quelques mois.

Annexe 1 : Preuve du théorème 4

Commençons par remarquer que $G_a \cap G_b$ est un graphe d'Erdos-Renyi de probabilité ps^2. Notons X la variable aléatoire qui compte le nombre de sommets isolés de $G_a \cap G_b$. Si $X \geq k+2$ où k est la taille de la graine alors il y aura au moins deux permutations de de-anonymisation qui seront indistinguables. D'après l'inégalité de Chebyshev :

$$P(X \leq \frac{1}{2}E(X)) \leq 4\frac{E(X^2) - E(X)^2}{E(X)^2}$$

Nous allons calculer le terme de droite : Un sommet est isolé avec probabilité $(1-ps^2)^{n-1}$ alors $E(X) = n(1-ps^2)^{n-1}$. De plus la probabilité qu'une paire de sommets soit une paire de sommets isolés est $(1-ps^2)^{2n-3}$ donc $E(\binom{X}{2}) = \binom{n}{2}(1-ps^2)^{2n-3}$. Alors :

$$\begin{aligned}
\frac{E(X^2) - E(X)^2}{E(X)^2} &= \frac{2E(\binom{X}{2}) + E(X) - E(X)^2}{E(X)^2} \\
&= \frac{(n^2-n)(1-ps^2)^{2n-3}}{n^2(1-ps^2)^{2n-2}} + E(X)^{-1} - 1 \\
&= (1-ps^2)^{-1} - n^{-1}(1-ps^2)^{-1} + E(X)^{-1} - 1 \\
&\leq 2ps^2 + E(X)^{-1}
\end{aligned}$$

Il nous reste donc à calculer $E(X)$ ou plus exactement à minorer $E(X)$.

$$\begin{aligned}
E(X) &= n(1-ps^2)^{n-1} \\
&= n(1 + \frac{ps^2}{1-ps^2})^{-(n-1)} \\
&\geq n(\exp(\frac{ps^2}{1-ps^2}))^{-n} \\
&= \exp(\log n - \frac{nps^2}{1-ps^2}) \\
&\geq \exp(\frac{c_n - ps^2 \log n}{1-ps^2})
\end{aligned}$$

La dernière inégalité provient du fait que $ps^2 \leq \frac{\log n - c_n}{n}$. Comme $ps^2 \leq \frac{\log n}{n}$ et $c_n \to \infty$ alors $ps^2 \to 0$, $ps^2 \log n \to 0$ et $E(X) \to \infty$. Finalement nous obtenons les deux résultats suivants :

$$P(X \leq \frac{1}{2}E(X)) \to 0$$

$$E(X) \geq \exp(\frac{c_n - ps^2 \log n}{1-ps^2})$$

Examinons alors le cas $k \leq \frac{1}{2}\exp(\frac{c_n - ps^2 \log n}{1-ps^2}) - 1$. L'événement $(X \leq k+1)$ est inclus dans l'événement $(X \leq \frac{1}{2}E(X))$ dont l probabilité tend vers 0. Ainsi asymptotiquement presque surement (a.p.s.), $G_a \cap G_b$ a au moins deux sommets isolés indépendement du choix de la graine. Ainsi tout de-anonymiseur réussit avec probabilité au plus $\frac{1}{2}$.

Examinons maintenant le cas $k = o(\exp(\frac{c_n - ps^2 \log n}{1-ps^2}))$. L'événement $(X - k \leq \frac{1}{3}E(X))$ est inclus à partir c'un certain rang dans l'événement $(X \leq \frac{1}{2}E(X))$. Donc a.p.s. $X - k \to \infty$. Ainsi tout de-anonymiseur réussit avec probabilité $o(1)$.

Annexe 2 : Preuve du théorème 5

La preuve de ce théorème ressemble à celle du théorème précédent. Nous notons ici X la variable aléatoire qui compte le nombre de sommets de $G_a \cap G_b$ qui sont isolés et non sélectionnés dans la graine. Si $X \geq 2$ alors tout de-anonymiseur réussira avec probabilité au plus $\frac{1}{2}$. Si $X \to \infty$ alors tout de-anonymiseur réussira avec probabilité au plus $o(1)$.

Un sommet est isolé et non sélectionné avec probabilité $(1-ps^2)^{n-1}(1-l)$ donc $E(X) = n(1-ps^2)^{n-1}(1-l)$. De plus la probabilité qu'une paire de sommets soit une paire de sommets isolés et non sélectionnés est $1-ps^2)^{2n-3}(1-l)^2$ donc $E(\binom{X}{2}) = \binom{n}{2}(1-ps^2)^{2n-3}(1-l)^2$.

Nous conservons alors l'inégalité :

$$P(X \leq \frac{1}{2}E(X)) \leq \frac{1}{4}(2ps^2 + E(X)^{-1})$$

Il nous reste donc à trouver une minoration pour $E(X)$.

Nous avons :

$$E(X) = n(1-l)(1-p)^{n-1}$$

Donc :

$$E(X) \geq \exp(\log[(1-l)n] - \frac{nps^2}{1-ps^2})$$

Donc en utilisant le fait que $ps^2 \leq \frac{\log[(1-l)n - c_n]}{n}$:

$$E(X) \geq \exp(\frac{c_n - ps^2 \log((1-l)n)}{1-ps^2})$$

Il vient alors que $E(X) \to \infty$ et donc que a.p.s. $X \to \infty$. Finalement tout de-anonymiseur réussit avec probabilité $o(1)$.

Annexe 3 : Preuve du théorème 6

En effet ce cas correspond à $c_n = log(2(\epsilon + \delta)n)$. Ainsi si X est la variable aléatoire comptant le nombre de sommets isolés de $G_a \cup G_b$ nous avons :

$$E(X) \geq \exp(\frac{log(2(\epsilon+\delta)n) - ps^2 \log n}{1-ps^2})$$

Donc à partir d'un certain rang l'événement $(X \leq (\epsilon + \frac{\delta}{2})n)$ est inclus dans l'événement $(X \leq \frac{1}{2}E(X))$ qui est asymptotiquement impossible. Ailsi il y a asymptotiquement et presque surement une proportion au moins $(\epsilon + \frac{\delta}{2})$ de sommets isolés ce qui rends impossible la $(1-\epsilon)$-de-anonymisation presque surement.

Annexe 4 : Preuve du théorème 8

En effet ceci correspond au cas ou $p \leq \frac{\log n - p_n}{n}$ avec $p_n = \log(2(l+\delta)n)$. Alors avec le même raisonement que pour les théorèmes d'impossibilité de de-anonymiser précédents il viens que :

$$P(X \leq \frac{1}{2}E(X)) \to 0$$

$$E(X) \geq \exp(\frac{c_n - ps^2 \log n}{1 - ps^2})$$

Ainsi à partir d'un certain rang l'événement $(X \leq (l + \frac{\delta}{2})n$ est inclus dans l'événement $(X \leq \frac{1}{2}E(X))$ dont la probabilité tend vers 0. Ainsi a.p.s. il y a au moins une fraction $(l + \frac{\delta}{2})$ des sommets qui sont isolés et ainsi il est impossible de de-anonymiser une fraction $(1 - o(1))$ des sommets car indépendement du choix de l il y a a.p.s. au moins une fraction $\frac{\delta}{2}$ des sommets qui restent isolés et non dans la graine et donc ils n'ont aucun témoin de similarité à n'importe quel instant de l'algorithme et en particulier ils en ont moins que T.

Annexe 5 : Preuve du théorème 9

Notons

$$R_i := 1_{(\text{i est reconnu})}$$

Et pour une seed S fixée notons

$$N_i := 1_{(V_0(i) \geq T) \cup (i \in S_0)}$$

et

$$N'_i := 1_{(V_0(i) \geq T)}$$

Enfin notons

$$R = \sum_{i=0}^{n-1} R_i$$

$$N = \sum_{i=0}^{n-1} N_i$$

$$N' = \sum_{i=0}^{n-1} N'_i$$

Comme nous sommes dans le cas $ps^2 \geq \frac{24}{l}\frac{\log n}{n-2}$ le lemme 1 (dont la preuve est détaillée et exacte) nous permet d'affirmer que

$$P(\min_u SW_0(\sigma(u), u) > \max_{u, v \neq \sigma^{-1}(u)} SW_0(u, v)) \to 1$$

Pour alléger les notations nous posons

$$A := (\min_u SW_0(\sigma(u), u) > \max_{u, v \neq \sigma^{-1}(u)} SW_0(u, v))$$

Nous avons alors les inégalités ensemblistes suivantes :

$$((N' > \lambda) \cap A) \subseteq ((N > \lambda) \cap A) \subseteq (R > \lambda)$$

Soit $\mu \in]0, 1[$

$$
\begin{aligned}
P(\frac{R}{n} \leq 1-\mu) &\leq P((\frac{N'}{n} \leq 1-\mu) \cup A^c) \\
&\leq P(\frac{N'}{n} \leq 1-\mu) + o(1) \\
&\leq \sum_{S \text{ mapping initial possible}} P((\frac{N'}{n} \leq 1-\mu) \cap (S_0 = S)) + o(1) \\
&\leq \sum_{k=0}^{n} \sum_{S \text{ mapping initial possible de taille k}} P((\frac{N'}{n} \leq 1-\mu) \cap (S_0 = S)) + o(1) \\
&\leq \sum_{k=0}^{n} \sum_{S \text{ mapping initial possible de taille k}} P(S_0 = S) P((\frac{N'}{n} \leq 1-\mu) | (S_0 = S)) + o(1)
\end{aligned}
$$

Or conditionnellement à S_0 la famille (N'_i) est indépendente et chacune de ses variables suit une variable de bernouilli de paramètre $p(k) := 1-(1-ps^2)^k - kps^2(1-ps^2)^{k-1}$ où k est la taille de S_0 car $T = 2$. Il est alors possible d'utiliser la borne de Chernoff suivante :

$$
P(N' \leq (1-\delta)E(N'|S_0)|S_0) \leq \exp(-E(N'|S_0)\delta^2/2)
$$

Nous avons alors :

$$
\begin{aligned}
P(\frac{R}{n} \leq 1-\mu) &\leq \sum_{k=0}^{n} \sum_{S} P(S_0 = S) P((\frac{N'}{n} \leq 1-\mu) | (S_0 = S)) + o(1) \\
&\leq \sum_{k=0}^{n} \sum_{S} P(S_0 = S) P((N' \leq \frac{1-\mu}{p(k)} p(k) n) | (S_0 = S)) + o(1) \\
&\leq \sum_{k=0}^{n} \sum_{S} P(S_0 = S) P((N' \leq (1-(1-\frac{1-\mu}{p(k)}))E(N'|S_0) | (S_0 = S)) + o(1) \\
&\leq \sum_{k=0}^{n} \sum_{S} P(S_0 = S) \exp(-np(k)(1-\frac{1-\mu}{p(k)})^2/2) + o(1) \\
&\leq \sum_{k=0}^{n} \binom{n}{k} l^k (1-l)^{n-k} \exp(-np(k)(1-\frac{1-\mu}{p(k)})^2/2) + o(1)
\end{aligned}
$$

Finalement nous arrivons à l'inégalité suivante :

$$
P(\frac{R}{n} \leq 1-\mu) \leq \sum_{k=0}^{n} \binom{n}{k} l^k (1-l)^{n-k} (1_{[0, \lfloor \frac{nl}{4} \rfloor] \cup [\lfloor \frac{3nl}{4} \rfloor, n]}(k) + 1_{]\lfloor \frac{nl}{4} \rfloor, \lfloor \frac{3nl}{4} \rfloor[}(k) \exp(-np(k)(1-\frac{1-\mu}{p(k)})^2/2)) + o(1)
$$

Nous pouvons borner uniformément en k sur l'intervalle $]\lfloor \frac{nl}{4} \rfloor, \lfloor \frac{3nl}{4} \rfloor[$ l'expression de $p(k)$:

$$
1 - (1-ps^2)^{\frac{ln}{4}-1} - (\frac{3ln}{4}+1)ps^2(1-ps^2)^{\frac{ln}{4}-2} \leq p(k) \leq 1
$$

Or $ps^2 \geq \frac{24}{l}\frac{\log n}{n-2}$ donc en dévellopant (DL) il vient que

$$
1 - o(1) \leq p(k) \leq 1
$$

Finalement ceci prouve que

$$\sum_{k=0}^{n} \binom{n}{k} l^k (1-l)^{n-k} 1_{]\lfloor \frac{nl}{4} \rfloor, \lfloor \frac{3nl}{4} \rfloor[}(k) \exp(-np(k)(1 - \frac{1-\mu}{p(k)})^2/2) \to 0$$

Nous pouvons ensuite remarquer que

$$\sum_{k=0}^{n} \binom{n}{k} l^k (1-l)^{n-k} 1_{[0, \lfloor \frac{nl}{4} \rfloor] \cup [\lfloor \frac{3nl}{4} \rfloor, n]}(k) \to 0$$

En effet cette quantité est inférieure à $P(|X - E(X)| \geq \frac{3}{4}E(X))$, où $X \sim B(n, l)$, or cette probabilité tend vers 0 d'après les bornes de Chernoff.

Donc :

$$P(\frac{R}{n} \leq 1 - \mu) \to 0$$

Références

1. Lars Backstrom, Cynthia Dwork, and Jon Kleinberg. Wherefore art thou r3579x ? : Anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 181–190, New York, NY, USA, 2007. ACM.
2. Daniel Cullina and Negar Kiyavash. Improved achievability and converse bounds for erdos-renyi graph matching. *SIGMETRICS Perform. Eval. Rev.*, 44(1) :63–72, June 2016.
3. Nitish Korula and Silvio Lattanzi. An efficient reconciliation algorithm for social networks. *Proc. VLDB Endow.*, 7(5) :377–388, January 2014.
4. Pedram Pedarsani and Matthias Grossglauser. On the privacy of anonymized networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1235–1243, New York, NY, USA, 2011. ACM.