Année 2020-2021

RAPPORT DE PROJET - CDSI M1 S8

PROJET CYBERSÉCURITÉ "TNP - Twitter Network Profiler"

Projet réalisé par
Thomas BAUDUIN
Mathis ENGELS

Projet encadré par
Jérôme RIDET

GitHub

Université Polytechnique HAUTS-DE-FRANCE

Sommaire

Présentation du profilage

Définition du profilage

La notion de profilage représente l'ensemble des formes de traitement automatisé de données à caractère personnel consistant à évaluer certains aspects personnels relatifs à une personne physique, notamment en faisant des analyses ou des prédictions de ces aspects.

Exemples d'aspects

Rendement de travail

Il arrive que certaines entreprises utilisent le profilage pour vérifier les performances de leurs salariés dans leur travail.

Situation économique

Cet aspect est beaucoup utilisé dans les banques. En effet, chaque personne étant affiliée à une banque est en droit de faire une demande de prêt. L'utilisation du profilage va permettre aux banques de vérifier la situation économique de la personne ayant fait la demande, et d'accepter ou non la demande de prêt.

Santé

En analysant les achats d'un individu, on peut définir un profil. Prenons l'exemple d'une personne achetant des médicaments pour la toux, on peut suspecter que cet individu souffre de maux de gorge. Un maux de gorge peut être insignifiant, mais ce même principe peut être exploité pour trouver des maladies graves.

Centres d'intérêts

Cet aspect est très utilisé dans l'environnement du marketing. Les centres d'intérêts des personnes vont être décortiqués (ex : les sites web sur visités). Puis, lorsque l'ensemble des données sur leurs centres d'intérêts sera récupéré et analysé, les commerciaux/équipes marketing proposeront des activités liées au centre d'intérêts.

Localisation

En général, on va faire du profilage de localisation pour prédire la position d'une personne à un moment T. On récupère un maximum d'informations sur sa localisation ces derniers temps. D'après les différents endroits où la personne, qui a été profilée, est allée, on va pouvoir déterminer vers quelle direction elle va se diriger. Cette technique est très utilisée par la police pour retrouver d'éventuels criminels.

Les avantages et les inconvénients du profilage

Avantages

Marketing

Le profilage est un avantage majeur pour le marketing.
Les réseaux sociaux en sont un exemple, une grande partie de leur rémunération est basée sur les publicités apparaissant sur leurs sites web.

Ils vont réaliser une analyse sur chaque profil utilisateur, c'est-à-dire leurs publications, leurs recherches (sur le réseau social et en dehors du réseau social). On dit qu'ils construisent des profils publicitaires. Lorsque l'ensemble de ces informations est collecté par les réseaux sociaux, ils vont ensuite faire du ciblage de publicité, autrement dit, proposer des publicités qui ont un rapport avec les publications et les recherches que l'utilisateur a précédemment réalisé.

Grâce à ce profilage, les réseaux sociaux savent que les publicités proposées ont de grandes chances de plaire et d'augmenter la probabilité de cliquer sur la publicité pour en savoir plus, ce qui augmentera la rémunération du réseau social.

Sécurité

Le profilage permet d'améliorer la sécurité dans certains cas notamment pour les structures publiques, comme par exemple les Etats qui vont l'utiliser dans les domaines suivants :

- La sécurité publique
- La surveillance
- Le renseignement

Prenons l'exemple de la France qui a subi un nombre important d'attaques vigipirates, ces dernières années. L'utilisation du profilage a été primordiale pour retrouver les auteurs de ces crimes, c'est notamment en utilisant l'ensemble des informations qu'ils ont pu récupérer sur ces personnes, les images des crimes et la localisation de ceux-ci. Les auteurs de ces actes ont, dans la majorité des cas, pu être retrouvés.

Réduire les risques dans les entreprises

Le profilage permet de réduire les risques possibles dans les entreprises.

Reprenons l'exemple de la banque.
Lorsqu'une banque fait un prêt à l'un de ses clients, elle lui donne une partie de son argent dans l'optique que le client lui rembourse par la suite avec un léger taux d'intérêt. Cependant, si elle n'étudie pas la situation financière du client, elle va être exposée à un risque, qui est que le client demande un prêt alors qu'il n'a pas les moyens de rembourser la banque.
Dans ce cas là, la banque se retrouverait dans une impasse car elle aurait prêté de l'argent à une personne qui va être dans l'incapacité de les rembourser, provoquant une perte d'argent pour la banque.

Le fait d'utiliser le profilage va permettre à la banque de pouvoir consulter la situation économique du client et par conséquent d'être apte à connaître si le client possède les fonds et le profil nécessaire pour pouvoir rembourser la somme d'emprunt.
En fonction du résultat du profilage, la banque pourra accepter ou rejeter la demande du client.

Si maintenant on se base sur l'exemple d'un directeur d'entreprise qui ressent un dérèglement dans l'activité de son entreprise. En utilisant le profilage, cela va lui permettre d'étudier le rendu et l'investissement de ses salariés.
Grâce à cette analyse, le directeur pourra plus facilement trouver la source du problème (ex: un salarié qui a un manque d'investissement).
Le directeur pourra agir en conséquence pour résoudre le problème.

Certaines entreprises vont également utiliser le profilage de données pour étudier la qualité de celles-ci, vérifier si elles ne sont pas obsolètes, incomplètes ou encore impertinentes.
Détenir des données qui comportent ces problèmes peut représenter des risques pour l'entreprise.

Inconvénients

Le profilage ne possède pas forcément d'inconvénients particulier si ce n'est le grand nombre de lois s'opposant à l'utilisation du profilage. En effet, les outils de profilage sont considérés comme étant des outils puissants, ils sont donc soumis à de nombreuses régulations.

Beaucoup de personnes et d'entreprises utilisent le profilage à mauvais escient, avec de mauvaises intentions. C'est l'une des raisons pour laquelle il y a autant de régulations concernant le profilage et qu'elle est un droit réservé.

L'utilisation à mauvais escient des outils de profilage ne va pas aller en s'arrangeant. Notre monde devient de plus en plus connecté. Par conséquent, la quantité des données ne cesse d'augmenter et incite encore plus de personnes malveillantes à utiliser des outils de profilage pour soutirer des informations sensibles.

Exemple d'utilisations illicites du profilage

Comme dit précédemment, le profilage est un gros avantage pour le marketing car il permet d'optimiser le rendement des publicités grâce à la récolte des centres d'intérêts des personnes.
Cependant, certaines personnes vont faire du profilage pour récupérer des informations personnelles, plus sensibles de certains individus afin de les revendre à des entreprises intéressées, sans le consentement des personnes concernées (ex : des informations sur l'état de santé de certaines personnes).

Quelques outils de profilage

Les outils présentés sont des outils de profilage ne manipulant aucune donnée personnelle. Ils permettent l'analyse, l'optimisation et la correction de problèmes.

Pyroscope

Pyroscope est un logiciel de profilage de données. L'objectif de ce logiciel va être d'analyser de manière dynamique le code ou le comportement d'un programme grâce aux données que le programme va collecter lors de l'exécution.
L'outil va permettre de faire un compte rendu sur les potentielles optimisations qui vont pouvoir être réalisées pour améliorer la rapidité d'exécution et la réactivité de l'application, mais aussi de diminuer la consommation de mémoire et de ressources.

Astera Centerprise

Astera Centerprise est un logiciel de profilage qui permet de vérifier la qualité des données d'une entreprise.
La qualité de données d'une entreprise est primordiale, notamment pour favoriser son bon développement. Des données erronées ou incomplètes pourraient avoir des conséquences négatives sur le bon fonctionnement d'une entreprise.
L'outil Astera Centerprise donne un résultat par rapport à la qualité des données que l'entreprise possède, c'est-à-dire qu'il vous informe si certaines de vos données ont des valeurs manquantes (des valeurs qui sont nulles), si des données sont des doublons, ...

Les lois concernant le profilage

De nos jours, l'utilisation des données personnelles est un sujet sensible. Beaucoup de scandales ont eu lieu concernant une utilisation illégale de données personnelles. En effet, il n'est pas possible d'utiliser des données personnelles sans respecter les règles du RGPD.

Qu'est-ce que le RGPD ?

Le RGPD (**R**ègle**m**ent **G**énéral sur la **P**rotection des **D**onnées), qu'on peut également appeler GDPR, est un règlement qui est apparu le 27 avril 2016 et est arrivé en Europe le 25 Mai 2018. Leur but va être de maximiser la protection des personnes qui sont concernées par un traitement de leurs données personnelles et de responsabiliser les acteurs de ces traitements. D'après le RGPD, une donnée personnelle est considérée comme toute information qui se rapporte à une personne identifiée ou identifiable.

Par conséquent, chaque personne, chaque organisation et chaque entreprise ayant pour but d'utiliser des informations personnelles se doivent de respecter le RGPD. En cas de non respect des ces règles, l'acteur en question pourrait encourir de lourdes sanctions (ex: sanctions financières) de la part de l'autorité de contrôle compétent de son pays. En France, cette autorité de contrôle compétente n'est autre que la CNIL (**C**ommission **N**ationale de l'**I**nformatique et des **L**ibertés).

Exemple d'entreprise ayant eu des problèmes avec les règles du RGPD

Prenons l'exemple de Facebook.
Le siège de Facebook se trouvant à Dublin, en Irlande, l'entreprise est soumise aux règles du RGPD depuis le 25 Mai 2018, date où les règles du RGPD sont arrivées en Europe.

Facebook est un réseau social qui est amené à utiliser très souvent les données personnelles de ses utilisateurs. Il est donc suivi de très près par le RGPD.
Facebook a connu de nombreux conflits avec le RGPD, notamment avec les nombreuses fuites des données qu'il y a eu chez eux ces dernières années.
La dernière fuite de ses données date d'avril 2021, avec plus de 500 millions de comptes utilisateurs fuités.
La DPC (**D**ata **P**rotection **C**ommission), qui est l'autorité de contrôle compétent en Irlande (rappelons que le siège de Facebook se trouve en Irlande), a ouvert une

enquête pouvant infliger de nombreuses sanctions pour le premier réseau social du monde.

Le profilage et le RGPD

Concernant notre projet, nous devons donc avant de commencer toute activité, vérifier si notre projet respecte les règles du RGPD.
Comme nous l'avons vu, pour faire du profilage, nous allons être amenés à utiliser les informations personnelles des profils que nous voulons traquer.
Mais cette méthode a des limites et peut enfreindre les règles du RGPD.

Dans les règles du RGPD, il est stipulé qu'il est interdit de faire du profilage.
Cependant, il existe quelques cas exceptionnels où le profilage est autorisé.

Autorisation par contrat

En général, ce sont les entreprises qui bénéficient de ce genre de dérogation pour pouvoir faire du profilage.
Il arrive que certaines entreprises doivent étudier plusieurs informations personnelles de personnes afin de pouvoir prendre des décisions.

Reprenons l'exemple de la banque.
Lorsqu'un client vient pour demander un prêt, la banque va profiler la situation financière de ce client pour déterminer s'il est judicieux de faire un prêt à la banque.
La banque peut donc se permettre de faire cela car elle possède une dérogation de contrat.

Autorisation par loi

Ce genre de dérogation peut être une loi nationale par exemple.
Si nous nous basons sur le profilage criminel. Ce profilage est utilisé pour retrouver des criminels qui sont dans la nature et qui sont recherchés.
En général, on va essayer d'utiliser les informations de localisation, pour essayer de prédire l'endroit où ils pourront potentiellement le retrouver. Ces méthodes sont majoritairement utilisées par la police, qui est aux ordres de l'État.
Elle est donc autorisée à faire du profilage pour que le pays en question soit sûr pour ses concitoyens.

Pour ces deux premiers cas exceptionnels, on dit que les profileurs ont fait une prise de décision automatisée, c'est-à-dire que les acteurs ont pris la décision d'utiliser les informations personnelles d'une personne sans avoir son accord.

Autorisation par consentement

Nous sommes autorisés à faire du profilage avec des personnes qui connaissent vos intentions, à partir du moment où ils sont d'accord avec ce que vous allez faire de leurs informations.
De plus, les personnes profilées consentantes vont avoir plusieurs droits :

Droit d'opposition

Ce droit signifie que même après avoir consenti à l'utilisation de ses informations personnelles, la personne a le droit de changer d'avis et de décider de ne plus accepter que l'on utilise des informations personnelles.

Droit d'accès

Le droit d'accès permet que la personne profilée est en droit de demander aux personnes qui la profilent, l'ensemble des informations qu'ils détiennent sur elle. Ce droit représente l'article 15 du RGPD.

Droit à l'oubli

La personne visée par le profilage à la possibilité de demander aux profileurs de ne plus utiliser certaines ou toutes informations personnelles qu'ils possèdent sur elle et de les supprimer.

Droit de rectification

Ce droit octroie, à la personne visée par le profilage, la possibilité de demander aux profileurs de modifier certaines des informations qu'ils possèdent sur elle. Cela peut être dû au fait que les informations qu'ils possèdent sont complètement fausses, sont incomplètes ou encore doivent être mise à jour.

Même dans ces cas exceptionnels, **il est formellement interdit de faire du profilage sur des enfants.**

Les informations concernant la santé d'un individu ne peuvent être récupérées que si cette personne est consentante.
Par exemple, si une entreprise est en droit de faire du profilage (autorisation par contrat), elle ne pourra tout de même pas récupérer les informations de santé des personnes qu'elle va profiler.

Pour résumer

D'après les règles du RGPD, nous ne sommes pas en droit de faire un outil de profilage, car nous ne possédons aucune des autorisations nécessaires :

- Nous ne possédons aucun contrat.

- Aucune loi ne nous permet de faire du profilage.

- Nous pourrions tester le profilage sur des personnes étant consentantes mais nous ne voulions pas que notre futur outil soit testé sur une minorité de personnes.
  De plus, faire du profilage sur des personnes consentantes est risqué puisqu'elles peuvent changer d'avis à tout moment, ce qui pourrait compromettre notre projet.

Elaboration de notre futur outil

Problématique : Comment créer un outil de profilage sans enfreindre les règles du RGPD et sans dérogation ?

Environnement

Dans un premier temps, nous avons cherché un environnement dans lequel nous pouvions nous baser pour réaliser notre futur outil de profilage. Nous avons décidé de nous baser sur le réseau social Twitter.

L'avantage d'utiliser Twitter, ou un réseau social de manière générale, comme base pour notre futur outil de profilage est que nous avons accès à une grande quantité d'informations qui sont rendues publiques par les utilisateurs eux-mêmes.
Par conséquent, l'utilisateur est conscient que les informations qu'il a divulgué sont publiques et utilisables par tous.
Twitter étant un grand réseau social, cela nous permet d'utiliser notre outil de profilage sur un large panel de personnes.

Informations

Après avoir choisi notre milieu, nous avons réfléchi par rapport aux différentes informations que nous pouvions extraire d'un utilisateur pour ensuite déterminer si ce sont des informations qui sont potentiellement exploitables et si nous sommes en droit d'exploiter ces informations.
Nous avions donc la possibilité d'utiliser les informations suivantes :

Extrait du compte Twitter de François Hollande

Pseudonyme

Nous avons décidé d'utiliser cette information car on ne peut pas la considérer comme une donnée sensible, pour la simple et bonne raison que ce n'est pas une information nominative.
Par exemple, si je demande à un parfait inconnu qui est Dark_Hacker59, il ne saura pas me répondre. Le seul moyen de récupérer des informations depuis un pseudonyme est de le rentrer dans la barre de recherche Twitter. Cela revient donc à rechercher une personne sur Twitter, ce qui est complètement légal.

Localisation

Nous n'avons pas retenu cette information car nous la considérons comme sensible. En effet, connaître la localisation, combinée avec d'autres informations telles qu'une date de naissance, par exemple, pourrait nous permettre d'identifier une personne, représentant donc une atteinte à la vie privée, ce qui est vivement sanctionnée par le RGPD.

Date de naissance

Contrairement à la localisation, la date de naissance, sans la croiser avec d'autres informations, est une donnée que nous pourrions utiliser, puisqu'elle ne met pas en danger la vie privée de la personne profilée.
La seule information intéressante que nous pouvons recueillir sur la date de naissance est : “La personne est-elle mineure ou majeure".
Nous avons donc décidé de ne pas utiliser cette information car nous la considérons comme inutile pour notre utilisation.

Publications

Ce type d'informations sera utilisé car les publications sont des données qui sont disponibles publiquement.
Si nous allons sur un compte Twitter, nous allons avoir accès à l'historique de ses publications et ne révèle pas l'identité de son auteur. Par conséquent, l'utilisation des publications du compte que nous souhaiterions profiler n'enfreint pas les règles du RGPD.
De plus le contenu des publications peut être intéressant et nous donner des informations supplémentaires utiles comme par exemple : les comptes Twitter identifiés dans les publications.

Abonnés

Nous n'utiliserons pas cette information car cela représente les comptes d'utilisateurs qui suivent la cible. Ces personnes n'ont donc aucun rapport avec le compte visé et représentent juste ceux qui suivent le compte car ils apprécient son contenu. Cette information ne nous sera donc d'aucune utilité pour faire notre profilage.

Abonnements

Cette information aurait pu être un élément intéressant, notamment par rapport au lien qu'a le compte profilé par rapport aux comptes auxquels il est abonné. Cette liste représente la liste des comptes pour lesquels le compte ciblé a un intérêt.
Cependant, il ne nous a pas été possible de récupérer cette information. C'est pour cela que nous avons décidé de ne pas utiliser cette information.

Comptes privés

Il faut maintenant gérer un dernier problème, qui est celui des comptes utilisateurs privés. Le concept de ces comptes est de ne partager ses informations et ses publications seulement avec les comptes qui font partie de la liste d'abonnés et de la liste d'abonnements du compte.
Par conséquent, si nous essayons de faire du profilage sur des comptes utilisateurs privés, cela sera considéré comme une atteinte à la vie privée, puisque nous ne sommes pas censés avoir accès à ces informations.
Afin de rester dans les règles, nous avons décidé de faire du profilage uniquement sur des comptes publics.

Exploitation des informations

Maintenant que nous savons dans quel milieu nous allons utiliser notre futur outil de profilage et que nous savons quelles informations nous allons traiter sans enfreindre les règles du RGPD, nous avons dû déterminer de quelles manières nous pourrions exploiter ces informations.

Utilisation du pseudonyme

Nous savions que grâce au pseudonyme du compte de l'utilisateur, nous pouvions récupérer le contenu de l'ensemble de ses publications et retweets. Grâce à ceci, nous pouvons exploiter les publications et les retweets.

Utilisation des publications

Dans une publication twitter, vous avez la possibilité d'identifier quelqu'un. Nous nous sommes dit que nous pouvions décortiquer les publications des comptes que nous profilons, et de récupérer l'ensemble des comptes qu'ils ont identifiés dans leurs publications. En fonction du nombre de fois où un même compte a été identifié, nous pouvons déterminer le lien qu'il y entre deux comptes.

Exemple

On utilise notre futur outil de profilage sur un compte d'utilisateur qui a pour pseudonyme : Alice01.
Après avoir décortiqué ses tweets, nous avons remarqué que Alice01 avait identifié 16 fois Bob02, 4 fois Charlie03 et une fois David04. On va donc en conclure que Alice01 a un meilleur lien avec Bob02 que Charlie03 et David04.

Utilisation des retweets

Nous avons également la possibilité d'avoir accès aux retweets du compte visé par le profilage. Un retweet est une publication venant d'un autre compte que le compte visé a décidé de partager sur le sien. Nous allons donc nous baser sur le même principe que celui de l'identification des publications, nous allons décortiquer chaque retweet du compte visé, et regarder quel est le compte utilisateur de base qui a publié le retweet. En fonction du nombre de fois où nous trouvons des retweets provenant du même compte, nous pourrons déterminer les liens entre les deux comptes.

Exemple

On utilise un outil de profilage sur un compte utilisateur qui a pour pseudonyme Alice01. Après avoir décortiqué ses retweets on trouve que 5 de ses retweets viennent du compte de Bob02, 2 de ses retweets viennent de Charlie03 et 1 vient de David04. Cette information est un critère nous montrant que Alice01 a plus de liens avec Bob02 que Charlie03 et David04. On peut potentiellement supposer que Alice01 et Bob02 ont des centres d'intérêts en commun.

L'ensemble de cette réflexion nous a permis de déterminer le but final de notre outil de profilage : **trouver les comptes avec lesquels le compte cible a le plus d'interactions.**

La question qu'il faut maintenant se poser est la suivante :
Sous quelle forme allons-nous afficher le résultat de notre recherche ?

Affichage de nos résultats

Première idée

La première manière que nous allions utiliser pour afficher les résultats de notre profilage était de tout simplement afficher une liste nous donnant les comptes utilisateurs dans l'ordre décroissant ayant un lien (identification sur une publication, retweet, c'est-à-dire interaction) avec le compte cible.

Exemple

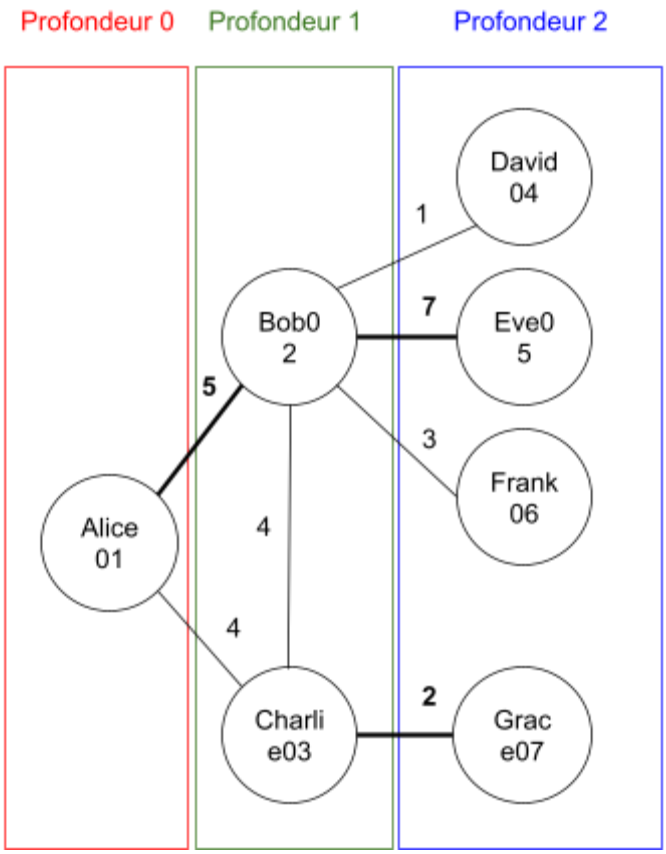
On lance un profilage sur Alice01 qui a fait deux tweets en identifiant Bob02 et Alice01 a retweeté 3 de ses tweets. Alice01 a également fait trois tweets en identifiant Charlie03 mais n'a retweeté aucun de ses tweets et Alice01 a fait 1 tweet à David04 et a retweeté un de ses tweets

Nous aurions eu le résultat suivant :
1. Bob02 : 5 (2 tweets + 3 retweets)
2. Charlie03 : 3 (3 tweets + 0 retweets)
3. David04 : 2 (1 tweet + 1 retweet)

Finalement, nous avons trouvé que cet affichage était simpliste et que cette manière n'offrait pas assez de visibilité concernant les liens entre les comptes.

Seconde idée

Puis, nous avons eu une seconde idée d'affichage qui nous permettait d'avoir une meilleure visibilité du résultat, un affichage un peu plus épuré, dynamique et la possibilité d'obtenir plus de contenu dans les résultats. Nous avons décidé d'afficher les résultats sous forme d'un graphe de trois profondeurs.



Graphe à 3 profondeurs

Ci-dessus, nous pouvons voir un exemple de la forme dans laquelle le résultat du profilage de Alice01 va s'afficher.

Nous avons le nœud de départ, se situant dans la profondeur 0, qui va représenter la personne profilée (dans notre exemple il s'agit de Alice01).

Ensuite, le graphe va nous montrer les personnes que Alice01 a identifiées ou qu'elle a retweetées. Ces personnes vont se situer dans la profondeur 1.
Dans notre exemple, nous avons Bob02 et Charlie03.
Le nombre qui se situe au-dessus des liaisons qui sont entre les nœuds représente l'addition des publications où Alice01 a identifié Bob02 et le nombre de fois où Alice01 a retweeté un tweet de Bob02. La liaison ayant la plus grande valeur sera représentée en gras et déterminera le compte avec lequel Alice01 a le plus d'interactions.

Puis, nous voyons que le graphe continue avec une profondeur supplémentaire.
Dans cette profondeur (profondeur 2), nous allons retrouver les comptes avec lesquels Bob02 et Charlie03 ont interagi (identification dans une publication et retweet).

Le fait d’avoir 3 profondeurs dans notre graphe nous donne un meilleur aperçu des liens qu’il y a autour de Alice01. Cela nous permettra également de faire certaines suppositions.
Par exemple, si nous reprenons l’exemple ci-dessus, nous avons Alice01 qui a un lien important avec Bob02, et Bob02 possède à son tour un lien important avec Eve05. On peut donc supposer que Alice01 et Eve05 pourraient potentiellement bien s’entendre s’ils se connaissaient.

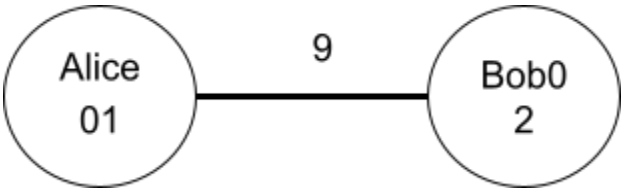
A partir de l’exemple, nous pouvons aussi dire qu’il y a un groupe d'amis/qu'il se fréquente. Le groupe serait constitué de Alice01, Bob02 et Charlie03 car ils communiquent tous ensemble.

Le graphe que nous avons décidé d’afficher est un graphe unilatéral.
Les données que nous affichons dans le graphe vont dans le sens de la personne que nous avons ciblée.

Exemple

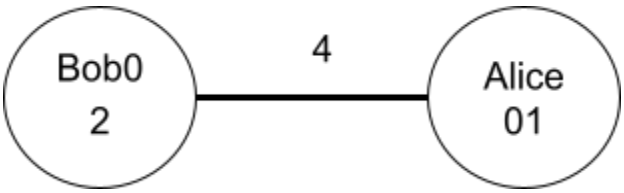
Alice01 a identifié Bob02 5 fois dans ses publications et Alice01 a retweeté 4 de ses publications. On peut donc dire qu’il aura eu 9 interactions avec Bob02.
Maintenant on suppose que Bob02 a identifié 2 fois Alice01 dans ses publications et que Bob02 a retweeté 2 de ses publications. On en déduit que Bob02 aura donc eu 4 interactions avec Alice01.

Si maintenant nous essayons de profiler Alice01, les deux premières profondeurs du graphe ressembleront à ça :



La valeur 9 représente le nombre de fois où Alice01 a eu des interactions avec Bob02. Nous avons donc ici une seule valeur qui est affichée à cause de l'unilatéralité du graphe.
Le graphe étant centré sur Alice01 (la personne profilée), ce sont ses valeurs d'interactions qui seront affichées.

Si maintenant nous réalisons notre profilage sur Bob02, nous allons obtenir le résultat suivant :



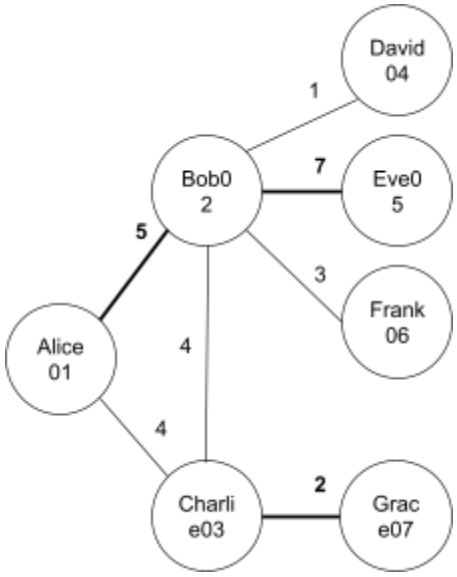
Ici, comme nous avons fait le profilage sur Bob02, le graphe va être centré sur lui.
Par conséquent, la valeur d'interaction qui va être affichée entre Bob02 et Alice01 va être de 4,soit le nombre d'interactions que Bob02 a eu avec Alice01.

Fonctionnalité de changement de profondeur

Nous avons voulu ajouter une fonctionnalité qui va nous permettre de modifier la profondeur du graphe.
Les futures personnes notre futur outil de profilage pourront alors chercher des informations différentes.
Par exemple, une personne peut utiliser l'outil de profilage seulement pour savoir avec quels comptes, la cible interagit. Dans cette situation, un graphe avec deux profondeurs est suffisant.

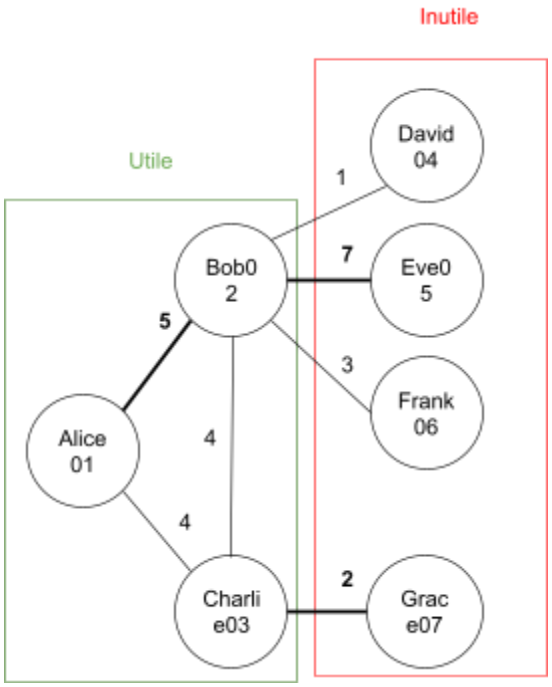
Exemple

Nous supposons que nous cherchons à récupérer les personnes avec qui Alice01 a interagi. Imaginons que le résultat de ce profilage soit représenté par un graphe de profondeur 3, cela nous donnerait un résultat de ce type :



Graphe centré sur Alice01 de profondeur 3

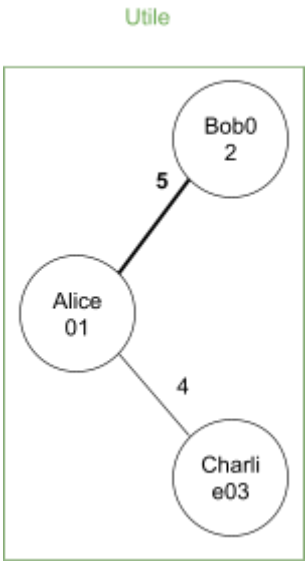
Nous remarquons que tous les noeuds qui sont présents sur ce graphe ne vont pas être utile pour l'utilisateur :



Si on utilise une profondeur de 3, certaines données du graphe vont être générées en trop par rapport aux attentes de l'utilisateur.

En permettant à l'utilisateur de choisir sa profondeur par rapport à sa demande, cela va lui permettre d'éviter de générer des données inutiles.
Cela va également réduire le temps d'exécution de l'outil de profilage puisque le nombre de nœuds générés va être réduit.

Dans cet exemple, en utilisant une profondeur de 2, l'utilisateur aura toutes les informations nécessaires répondant à ses attentes.

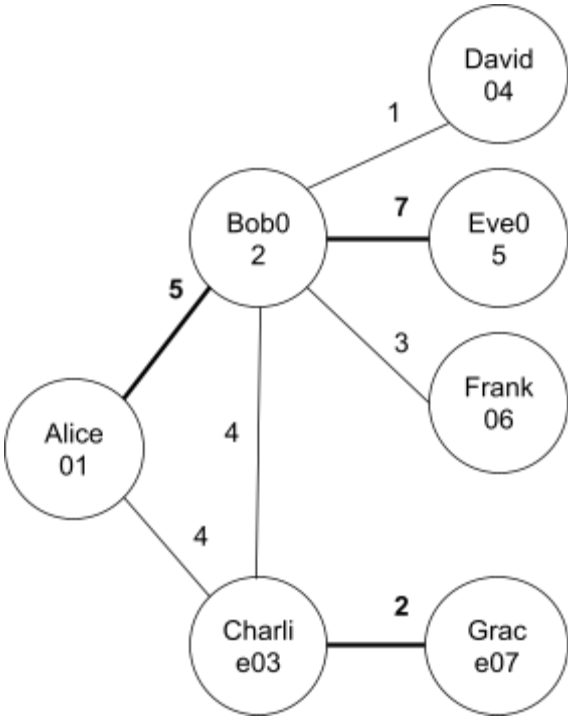


Nous voyons bien que dans ce graphe de profondeur 2, aucune des données affichée ne va être inutile.

Le deuxième type d'utilisateur voulant obtenir des informations depuis notre outil de profilage cherchera à obtenir le réseau de la cible. Pour cela il va avoir besoin d'un graphe avec une profondeur supérieure à 2.
C'est donc pour cela que nous avons décidé de rajouter cette fonctionnalité qui va permettre à l'utilisateur de choisir entre un graphe avec une profondeur de 2 et un graphe avec une profondeur de 3.

Fonctionnalité de recentrage

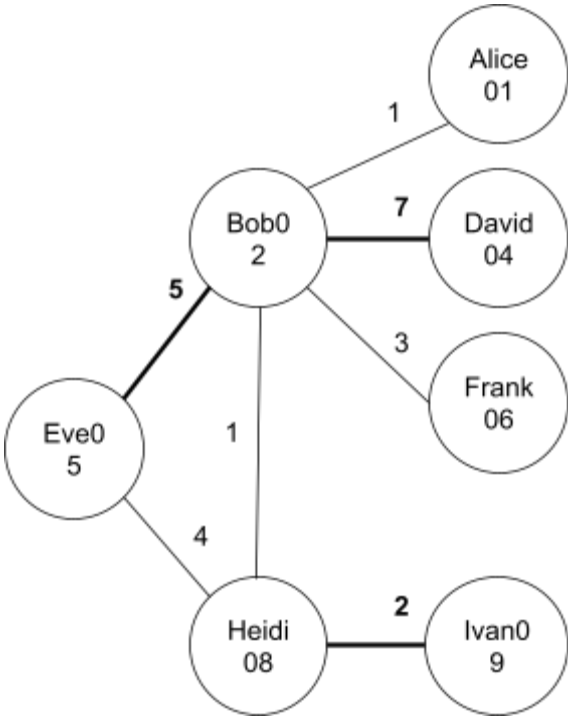
Nous avons eu l'idée de rajouter une possibilité dans le graphe permettant de centrer le graphe sur n'importe quel nœud affiché. Il suffira de cliquer sur le nœud à partir duquel nous aimerions recentrer le graphe.
Pour expliquer cette fonctionnalité, nous allons reprendre le graphe précédent.



Graphe centré sur Alice01 de profondeur 3

Imaginons que nous ayons fait un profilage sur Alice01 et que nous voyons que le lien le plus fort que Alice01 possède est avec Bob02 de 5 interactions.
On voit également que la personne avec qui Bob02 possède le plus gros lien est Eve05 avec 7 interactions.
On pourrait donc essayer d'en savoir plus concernant Eve05. On va donc avoir la possibilité de cliquer sur le nœud Eve05, ce qui va automatiquement recentrer le graphe sur Eve05 et nous montrer les différentes liaisons qu'elle a avec les autres utilisateurs.

Le nouveau graphe pourrait ressemblait à :



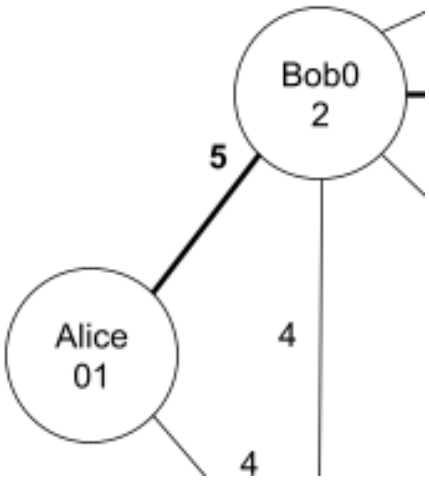
Graphe probable centré sur Eve05 de profondeur 3

Donc ici, nous voyons que le graphe est recentré sur Eve05. Nous voyons donc que maintenant le nœud de Eve05 se trouve en profondeur 0.

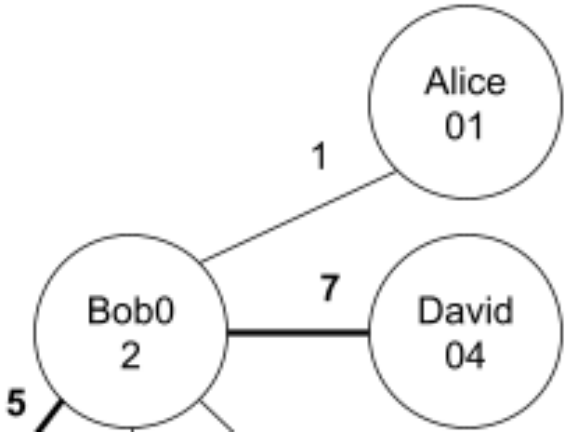
Quel va être l'utilité de cette fonctionnalité?

Lorsque nous comparons les deux graphes, nous remarquons qu'on peut y trouver des informations intéressantes.
Si l'on regarde les deux précédents graphes, les informations nous en dit beaucoup sur le lien qu'il y a entre Alice01 et Bob02, et Bob02 et Eve05

Relation entre Alice01 et Bob02



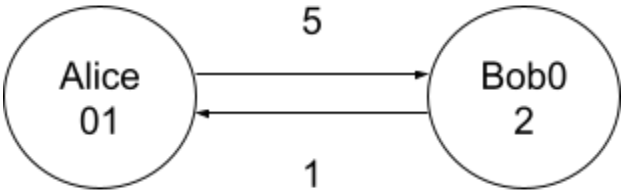
Graphe1



Graphe 2

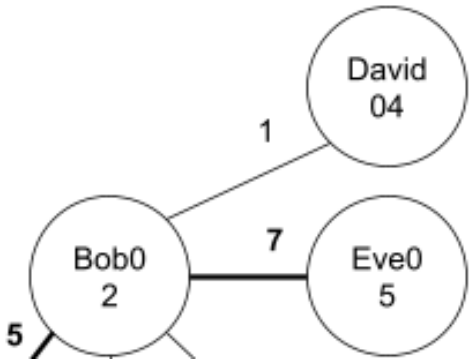
On peut voir que lorsque nous sommes centrés sur Alice01, son interaction est de 5. Alice01 a donc un gros lien avec Bob02. Cependant, lorsque nous regardons le deuxième graphe, nous voyons que Bob02 a une seule interaction avec Alice01.

En prenant le contenu des ces informations, nous pouvons les combiner pour récupérer le graphe suivant :

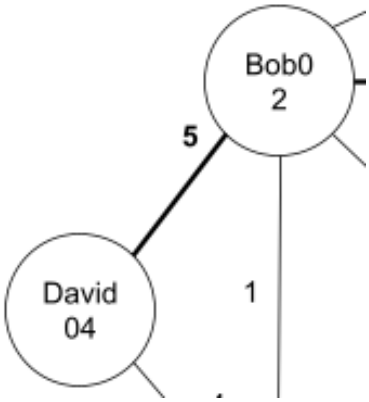


On peut donc en conclure que cette liaison n'est pas réciproque.
Si nous prenons exemple sur notre quotidien, on peut apprécier quelqu'un sans que cette personne ne nous apprécie en retour. On peut également être fan d'une célébrité (avoir une forte liaison avec elle) alors que celle-ci ne vous connaît même pas et n'a aucune liaison avec vous.

Relation entre Bob02 et David04



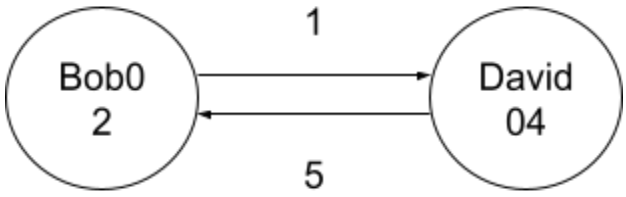
Graphe 1



Graphe 2

Lorsque nous regardons le graphe 1, nous voyons que l'interaction qu'il y a de Bob02 à David04 est très faible (1). On en conclut donc qu'il y a une faible liaison entre Bob02 et David04.
Cependant, lorsque nous regardons le deuxième graphe, nous voyons que David04 a une interaction de 5 avec Bob02 ce qui montre que David04 a une forte liaison avec Bob02.
Donc, comme avec Alice02 et Bob02, on peut en conclure que la liaison qu'il y a entre David04 et Bob02 n'est pas réciproque.

On pourrait combiner les deux parties de graphe que nous avons pour n'en faire qu'une seule :



On voit donc que le fait de centrer notre graphe sur plusieurs personnes différentes nous permet d'obtenir des informations supplémentaires qui peuvent être essentielles.

Cette fonctionnalité va également permettre à l'utilisateur de pouvoir faire le profilage de plusieurs comptes en étant parti du profilage d'un seul compte.

Grâce à cette fonctionnalité, les utilisateurs vont pouvoir en un clic obtenir des informations sur la réciprocité des liaisons entre les utilisateurs, ce qui leur évitera de perdre du temps à retaper le pseudonyme et de pouvoir interagir avec les résultats.

Fonctionnalité du nombre tweet et relation limites

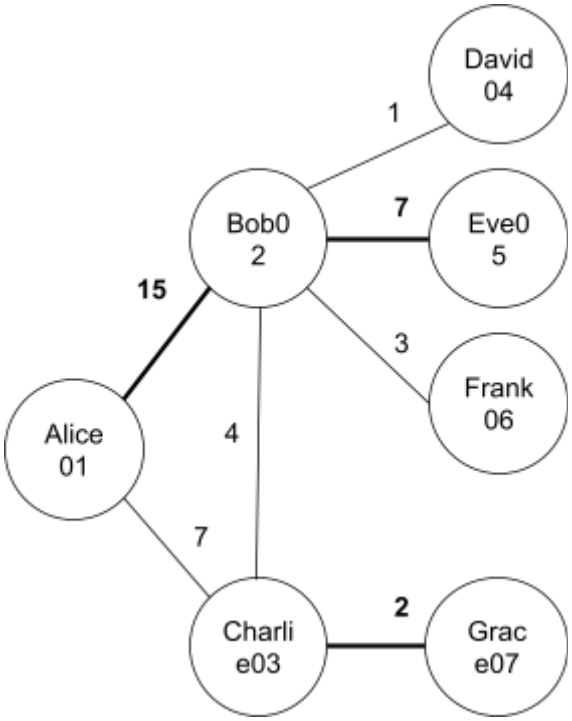
La fonctionnalité suivante que nous avons voulu implanter dans notre futur outil permet de sélectionner l'échantillon de publication sur lequel nous voulons faire notre profilage. Ceci va nous permettre de pouvoir étudier un même profil sur des périodes différentes.

Prenons l'exemple d'un utilisateur ayant le pseudonyme Alice01, ayant Twitter depuis 2010. Cela fait donc 11 ans que cette personne est sur Twitter. En 11 ans, les personnes, ou encore les comptes que Alice01 a suivis ont pu changer.
De ce fait, si nous faisons un profilage de l'ensemble de ses tweets, nous allons récupérer des tweets datant de 2010 et des tweets datant de 2021, ce qui va biaiser les résultats.
Le graphe de Alice01 va potentiellement afficher des comptes auxquels elle n'a pas parlé depuis plusieurs années et avec lesquelles elle n'a plus aucune affinité.
Si maintenant nous faisons un profilage de Alice01 sur ses 50 derniers tweets, le graphe que nous allons récupérer sur lui sera plus représentatif de ses liens actuels.

On pourrait donc se dire qu'il nous suffit de tout le temps utiliser un échantillon de 50 pour que les informations que nous allons récupérer à propos de Alice01 soit le plus à jour possible.
Il n'empêche que le fait de pouvoir tester le graphe de Alice01 sur plusieurs échantillons peut nous être utile et nous donner plusieurs renseignements.
Nous allons illustrer cela à l'aide d'un exemple.

Exemple

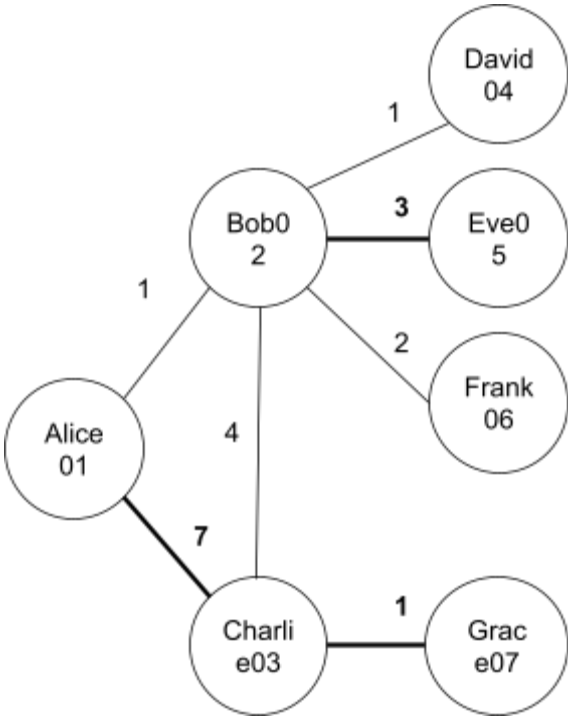
Imaginons que nous fassions deux tests de profilage sur Alice01 mais avec deux échantillons différents.
Nous allons prendre pour le premier test un échantillon de 200 tweets. Puis, pour le deuxième test, nous allons utiliser un échantillon de 50 tweets.
Ensuite, on compare les résultats des deux graphes pour voir si nous pouvons récupérer quelques informations intéressantes.



Graphe centré sur Alice01 de profondeur 3 avec un échantillon de 200 tweets

Lorsque nous étudions ce graphe, nous voyons que Alice01 a une forte relation avec Bob02, puisque la valeur de son interaction avec lui est de 15.
Cependant, on voit que l'interaction qu'il y a entre Alice01 et Charlie03 est relativement plus faible puisque la valeur de leur interaction n'est que de 7. On pourrait donc supposer qu'au jour d'aujourd'hui, Alice01 a de meilleurs liens avec Bob02 qu'avec Charlie03.
Mais ce n'est pas forcément vrai.
En effet, ces 200 tweets peuvent être répartis sur plusieurs années. Si Alice01 a fait ses 200 tweets en 5 ans, le graphe nous montre donc que sur ces 5 ans Alice01 a un meilleur lien avec Bob02 qu'avec Charlie03.
Mais cela ne signifie pas forcément que leur lien est aussi fort aujourd'hui.

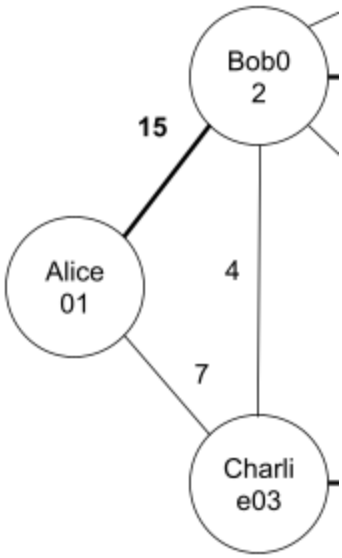
Nous allons maintenant faire de nouveau un profilage sur Alice01 mais en mettant un échantillon de 50 tweets au lieu de 200 tweets.
L'outil de profilage va donc agir sur les 50 derniers tweets.



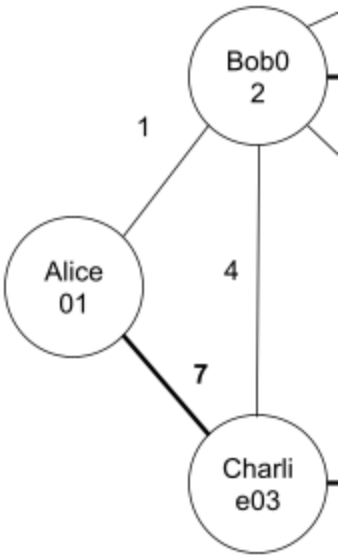
Graphe centré sur Alice01 de profondeur 3 avec un échantillon de 50 tweets

Lorsque nous regardons le résultat que nous donne l'outil de profilage, nous voyons que sur les 50 derniers tweets, Alice01 a un très gros lien avec Charlie03 puisque la valeur d'interaction avec lui est de 7.
Cependant, son lien avec Bob02 est nettement inférieur puisque la valeur de leur interaction est de 1.
D'après ce graphe, on peut supposer que Alice01 a un meilleur lien avec Charlie03 qu'avec Bob02.

Nous allons maintenant comparer les deux graphes pour essayer d'en obtenir certaines informations qui pourraient être intéressantes :



Echantillon de 200



Échantillon de 50

Lorsque nous étudions le cas entre Alice01 et Bob02, on remarque que sur les 200 derniers tweets, Alice01 a beaucoup interagi avec Bob02.
Cependant, lorsque nous regardons le deuxième graphe, on voit que sur les 50 derniers tweets, Alice01 n'a parlé qu'une fois à Bob02.
On peut donc supposer que dans le passé, Alice01 avait un certain lien avec Bob02 et que depuis, ils se sont un peu perdus de vue.

Si maintenant nous regardons l'évolution de la relation qu'il y a entre Alice01 et Charlie03, on voit dans le premier graphe que sa relation avec Charlie03 est nettement inférieure qu'avec Bob02.
Néanmoins lorsque nous nous fions au deuxième graphe, nous voyons que sur les 50 derniers tweets, Alice01 a beaucoup plus interagi avec Charlie03 qu'avec Bob02.
On peut donc supposer que ces derniers temps, Alice01 et Charlie03 sont très liés, mais qu'auparavant ils ne se côtoyaient pas.

Cette fonctionnalité nous permet donc en effet de pouvoir analyser la relation de deux utilisateurs en fonction du temps.

Pour conclure la mise en place de notre futur outil de profilage, nous avons voulu mettre en place une dernière fonctionnalité.
Cette fonctionnalité permet d'avoir la possibilité de choisir le nombre de relations que l'on veut afficher dans le graphe par compte. Il se peut que lors du résultat du profilage, certains utilisateurs possèdent un trop grand nombre d'utilisateurs avec lesquels ils ont interagi. De plus, il se peut que certains liens soient impertinents.

Si un utilisateur utilise Twitter depuis 10 ans, il va potentiellement avoir des liens très faibles avec un grand nombre de comptes. En mettant une limite de relation par personne, cela va permettre d'épurer le graphe, en évitant qu'il soit encombré par des nœuds inintéressants.

Si nous voulons afficher un plus petit nombre de relations par comptes, nous devons alors faire une sélection des relations que nous allons conserver.

3 solutions de sélection d'offrent à nous :

- Faire une sélection aléatoire : pour chaque compte, nous allons prendre un nombre limité de compte au hasard parmis ses relations.

- Faire une sélection par rapport au temps : Parmi les relations de la cible, nous sélectionnons les relations les plus récentes avec lesquelles il a interagi.

- Faire une sélection par rapport au degré de relation : Nous allons conserver les comptes avec lesquels la cible profilée a le lien le plus fort.

Nous avons décidé de ne pas utiliser la première possibilité car le fait de choisir des relations aléatoires peut réduire la qualité du résultat de notre futur outil de profilage. Il est possible que le compte avec lequel notre cible a le plus gros lien ne soit pas sélectionné parmi les relations affichées et ne sera donc pas présent dans le résultat du profilage alors qu'il représente une information importante.

Nous avons également décidé de ne pas utiliser la deuxième possibilité car le fait de choisir les relations à afficher par rapport au temps va être un avantage pour déterminer les relations actuelles du compte profilée ; nous n'aurons aucune information sur les relations que cette personne a eu auparavant.

Nous avons donc choisi la troisième possibilité, qui semble la plus appropriée pour avoir les informations les plus intéressantes sur le compte ciblé. Cette méthode va nous permettre de supprimer les relations qui ne sont pas intéressantes, tout en conservant les meilleures relations de celui-ci.

TNP - Twitter Network Profiler

D’un point de vue technique

Langages, framework, bibliothèques et fonctionnement général

Pour la réalisation de notre web app, nous avons décidé d’utiliser le framework Next.JS qui se base sur ReactJS.
Ce framework permet d’avoir le frontend et le backend dans le même projet. Il permet aussi de faire du pré-rendu des pages au moment de compilation (SSG) mais aussi au moment de la requête (SSR).
Pour le frontend, ce framework utilise ReactJS.
Pour le backend, nous utilisons ExpressJS.

Next.JS peut être développé en JavaScript ou TypeScript, HTML et CSS.
TypeScript nous semblait être le choix le plus judicieux, du fait de son typage, pour assurer un bon développement.

En plus de Next.JS, nous utilisons la bibliothèque “React-force-graph”. Cette bibliothèque assure la création du graphe et la logique des forces.
Du côté du backend, nous utiliserons des scripts Python.

Le fonctionnement général de la web app est assez classique.
Le client (frontend) rentre en paramètre les informations de son profilage. En lançant la recherche, il lance une requête au serveur (backend).
Le serveur va vérifier si les paramètres sont corrects, et va commencer à faire les calculs.
Le serveur recherche les informations, les formatent et les renvoie au client.
Le client reçoit les informations et les affiches.
Rentrons dans les détails du fonctionnement.

Récupération des données

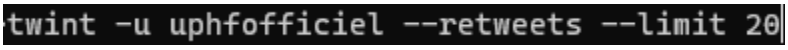
Pour récupérer les données de Twitter, deux choix s’offraient à nous :

- Utiliser l’API officielle de Twitter.
- Utiliser un outil de scraping.

Le premier choix semblerait le plus logique, nous permettant d'être sur des résultats et des données, mais il possède un gros défaut.
Les API publiques disposent généralement de limiteur de requêtes.
Dans le cas de Twitter et de son API, nous aurions été limités à 10 requêtes par secondes, 450 requêtes par 15 minutes et 25 000 requêtes par mois.

Ce nombre limité de requêtes est vraiment problématique. Prenons un exemple de graphe avec les paramètres suivants : relationMax = 60, profondeur = 2.
Le nombre de requêtes maximum sera de :
1 (Noeud profondeur 0) + (60 (Noeuds profondeur 1) * 60 (Noeuds profondeur 2))
Ce qui équivaut à 3601 nœuds, donc 3601 requêtes au maximum.
Il faudrait 135 minutes afin de récupérer les données, ce qui n'est pas envisageable.

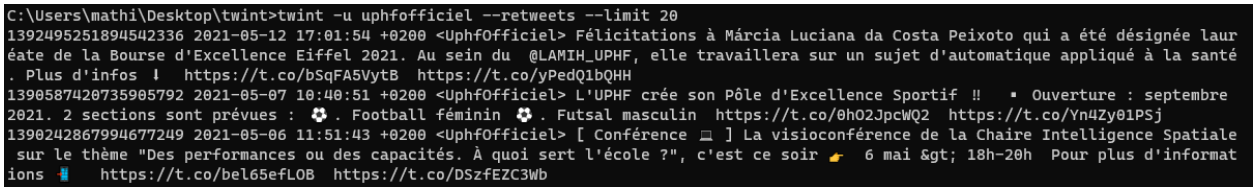
Nous nous sommes alors tournés vers des outils de scraping. Pour Twitter nous avons trouvé l'outil Twint.
Twint, écrit en Python, possède beaucoup de fonctionnalités et ne subit pas de limitations de requêtes.
Twint est simple d'utilisation, peut être utilisé en ligne de commande mais aussi comme module Python.
Voici quelques exemple de son utilisation :



Exemple d'utilisation de twint

twint : on spécifie le programme a utiliser
-u : on spécifie le compte ciblé
--retweets : on inclut les retweets dans la recherche
--limit 20 : on limite la recherche à 20 résultats.

On obtient ce résultat :



Résultat de la recherche (tweets)

Résultat de la recherche (retweet)

Comme vous pouvez le voir, on obtient un large panel d'informations. Les voici dans l'ordre :

- L'ID du Tweet
- La date du Tweet
- Le compte auteur du Tweet
- Le contenue du Tweet (contient RT au début si c'est un retweet)

Fonctionnement du backend

Le backend possède 2 routes différentes :

- Search
- Stats

Commençons par "/search", cette route prend 4 paramètres d'entrée : "target", "tweetLimit", "relationLimit" et "depth".
On vérifie si tous ces paramètres sont présents et ne sont pas nuls, puis nous lançons la première la recherche.

Le serveur va alors appeler un script Twint afin de prendre les tweets de la "target" avec une limite de "tweetLimit".

Grâce au résultat du script Twint, on va filtrer les tweets/retweets qui disposent d'un "@" dans leur contenu.

On va insérer les caractères qui suivent ce "@" dans un tableau qui permet de compter le nombre de fois qu'une valeur apparaît.
On trie ce tableau par ordre d'occurrence, et on garde seulement les X meilleures relations. X est "relationLimit".

Maintenant, on commence à créer les nœuds et les liens à partir des données de ce tableau.

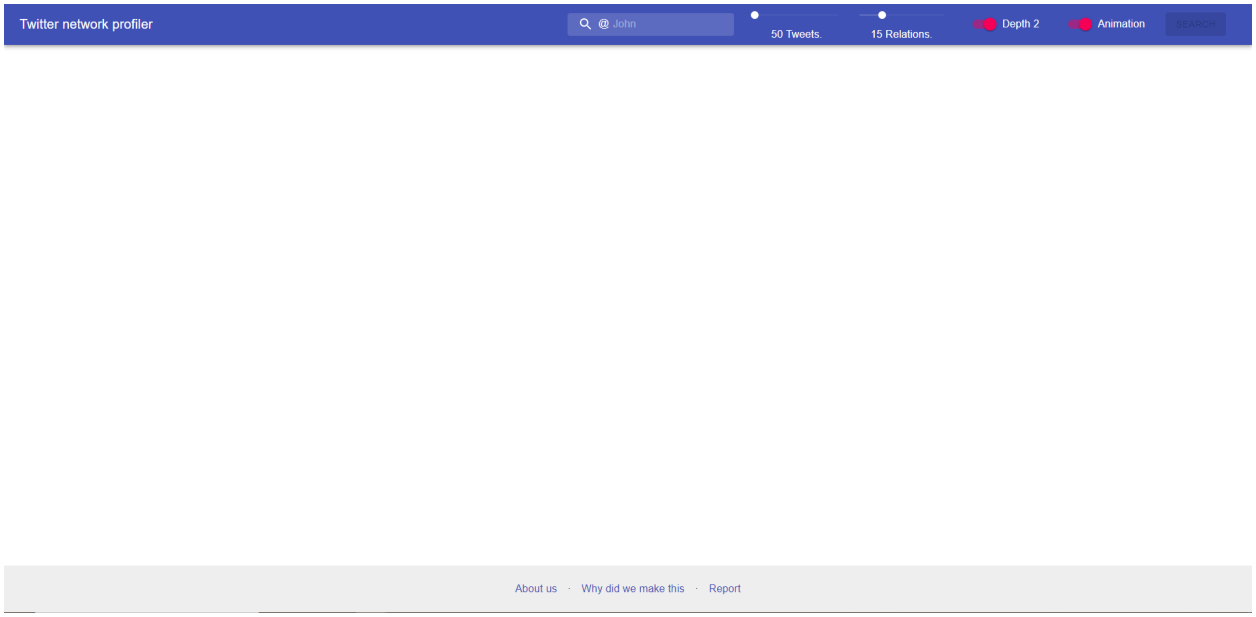
Si le paramètre "depth" est de 2, alors on prend le contenu de ce tableau et on répète l'action précédente. C'est-à-dire, prendre les tweets de la cible, les trier, garder seulement les X meilleures relations, et ajouter les nœuds et les liens.

A la fin, le serveur envoie les données (nœuds et liens).

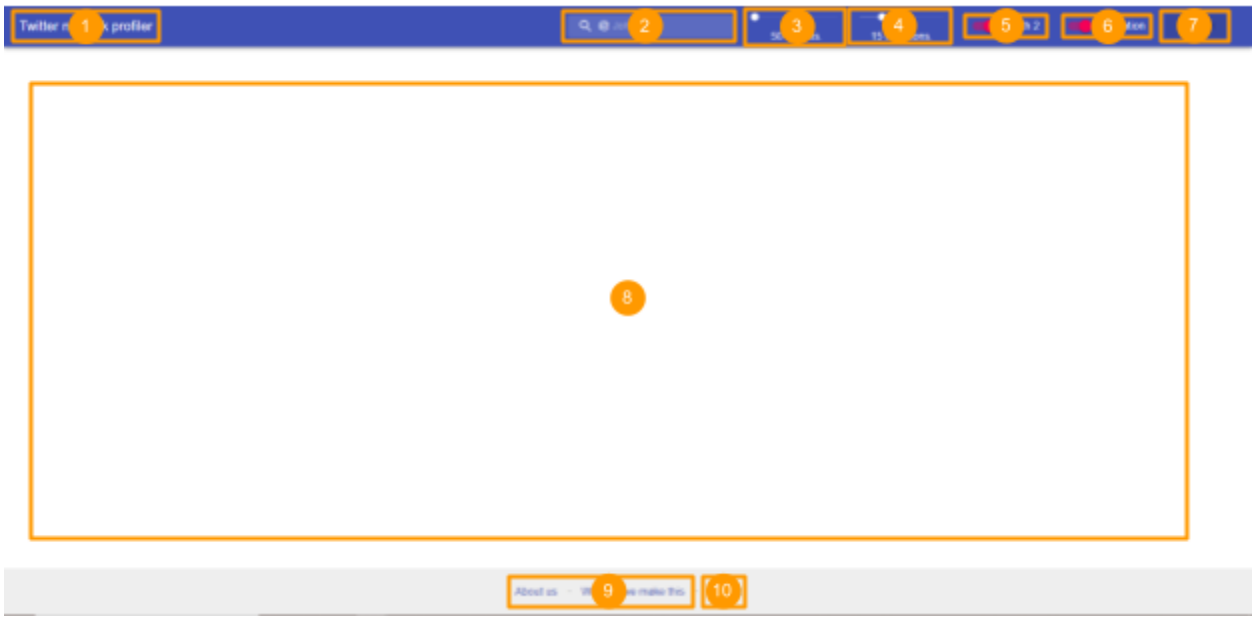
Parlons de /stats.
Cette route est appelée seulement pour les requêtes de profondeur 2. Elle permet d'estimer le temps que la requête "search".
/stats va faire exactement la même chose que /search, mais retourne juste le nombre de relations de la target, le temps moyen passé par relation (pour prendre les informations et les formatter).
Grâce à cette route, le frontend pourra faire une estimation du temps de la requête "search" afin de donner un feedback à l'utilisateur.

Fonctionnement du frontend

Nous allons décortiquer l'interface pour faciliter l'utilisation :



Interface de TNP (Twitter Network Profiler)



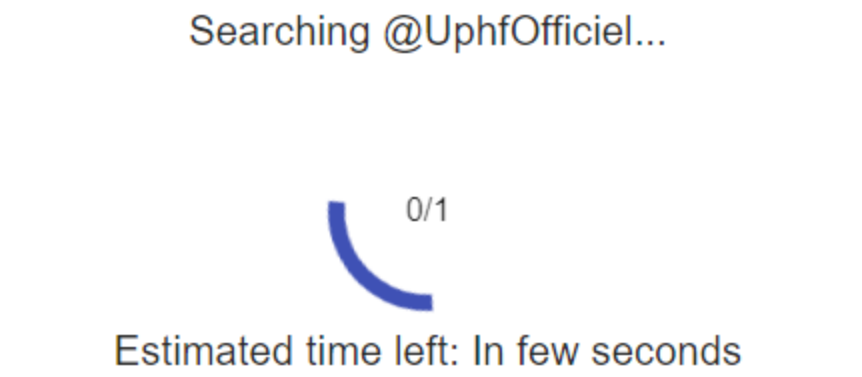
Description de l'interface de TNP

1 - Nom de l'outil.
2 - Entrée de la target
3 - Slider du nombre de tweet pris en compte
4 - Slider du nombre de relation pris en compte
5 - Activer/Désactiver la profondeur 2
6 - Activer/Désactiver l'animation du graphe
7 - Lancer la rechercher
8 - Graphe
9 - Informations complémentaires
10 - Accès à ce rapport

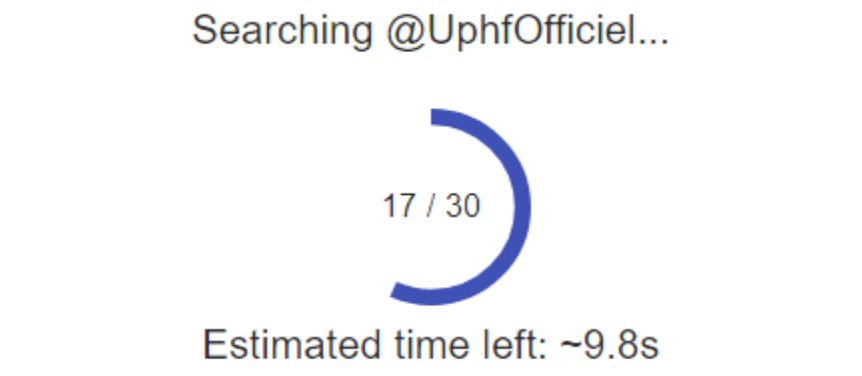
Un exemple de requête sur @UphfOfficiel :

Exemple sur @UphfOfficiel avec 155 tweets, 30 relations, profondeur 2 et animation

Lançons la recherche, une nouvelle interface va apparaître :



Interface d'attente profondeur 1



Interface d'attente profondeur 2

On remarque différentes informations qui résument notre profilage.
La target, le tweet limit, le relation limit, le depth.

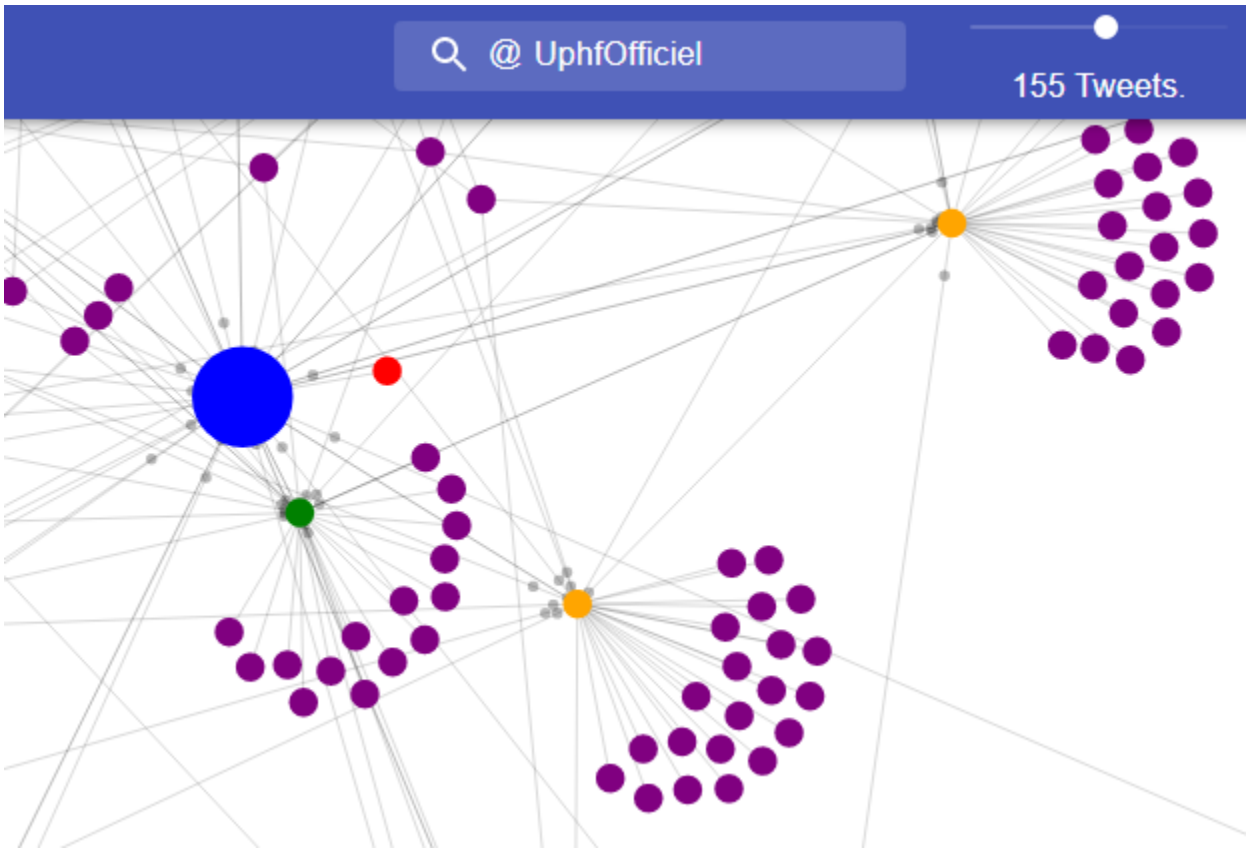
En profondeur 2, on observe le nombre de relations et le temps moyen par relation.

Bien évidemment, un compteur approximatif des relations restantes et du temps restant.

Le graphe

Le graphe est géré par la bibliothèque : react-graph-force.

Lors de l'affichage du graphe, on observe 5 couleurs différentes :



Graphe de @uphfofficiel, avec 155 tweets, 30 relations, profondeur 2 et animation

Décrivons les couleurs :

- Bleu : L'origine du graphe, la cible de notre profilage.
- Rouge/Orange/Vert : Ces couleurs sont attribuées au nœud de profondeur 1. Les nœuds rouges sont ceux qui disposent du moins d'interaction avec la cible. Les nœuds oranges, quant à eux, correspondent à une interaction médiane avec la cible. Alors que les nœuds vert sont ceux qui disposent de plus d'interaction avec la cible.
- Violet : Le violet est attribué exclusivement au nœud de profondeur 2.

En passant votre souris sur les différents nœuds, un tooltip apparaît.

Tooltip du noeud @uphfofficiel

Les tooltips pour les nœuds de profondeur 0 et 1 affichent la meilleure relation qu'ils entretiennent.

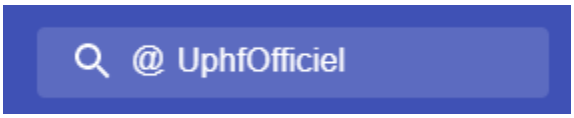
Pour tous les nœuds, une option est affichée : "Double-click to run a search on @...". C'est la fonctionnalité du recentrage. En double-cliquant sur un nœud, vous allez refaire une recherche, avec les paramètres actuels, sur le compte du nœud.

En utilisation

Exemple avec @UphfOfficiel

Nous avons voulu faire une démonstration pour montrer que TNP (Twitter Network Profiler) est bien opérationnel. Pour celle-ci, nous utiliserons le compte Twitter de l'UPHF (@UphfOfficiel) comme cible.

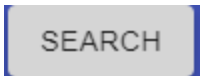
On rentre le pseudonyme de l'UPHF dans notre barre de recherche :



Dans un premier temps, nous allons lancer un profilage sur ce compte en désactivant l'ensemble des fonctionnalités disponibles :

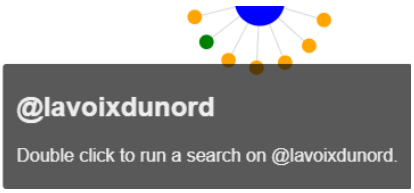
Nous allons également limiter l'échantillon de tweet de la recherche à 50 et le nombre de relations par compte à 15.

Puis nous appuyons sur le bouton "Search" pour afficher le résultat :

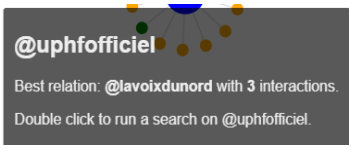


On obtient ce résultat :

Dans un premier temps, nous pouvons voir que la limite de relation est bien respectée puisque dans les réglages nous voulions que le nombre de relations ne dépasse pas 15.
Sur les 15 relations il y a un compte avec lequel l'UPHF a beaucoup d'interactions (couleur verte). Lorsque mettons notre souris sur le noeud en question, nous voyons que cela correspond au compte : @lavoixdunord



Le fait de mettre notre souris sur le noeud de l'UPHF va nous permettre de voir les informations suivantes :



Ces informations nous permettent de confirmer que @lavoixdunord est bien le compte avec lequel l'UPHF interagit le plus. Il nous permet également de savoir que le nombre d'interactions que l'UPHF a eu avec la @lavoixdunord est de 3.

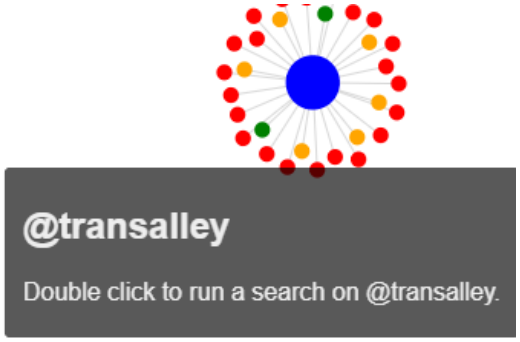
Maintenant, nous allons modifier les fonctionnalités de tweets et des relations, pour voir s'il va y avoir des changements par rapport aux résultats précédents.

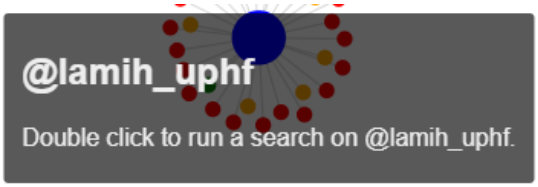
Nous allons donc mettre les valeurs suivantes :

Et nous obtenons le résultat suivant :

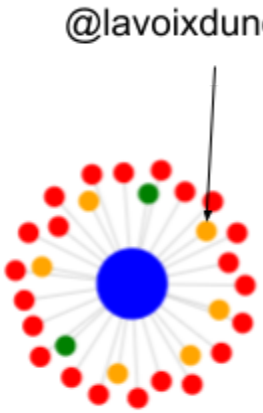
Nous pouvons voir que le nombre de relations de l'UPHF a augmenté. Cela est dû au fait que nous ayons augmenté la limite de relation. De plus, en augmentant l'échantillon de tweets sur lequel nous réalisons notre profilage, nous pouvons voir que la force des liens entre l'UPHF et les autres comptes a changé.

Lorsque nous regardons les comptes avec lesquels l'UPHF a les liens les plus forts :



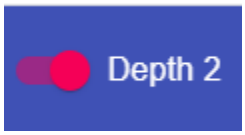


On peut donc voir qu'avec le nouvel échantillonnage de tweets, les plus fortes interactions que l'UPHF ont changées.
Quand on compare le résultat du premier profilage, nous voyons que la meilleure interaction que l'UPHF a est avec @lavoixdunord. Maintenant nous essayons de retrouver @lavoixdunord dans le graphe actuel, nous voyons que celui-ci est représenté en orange :

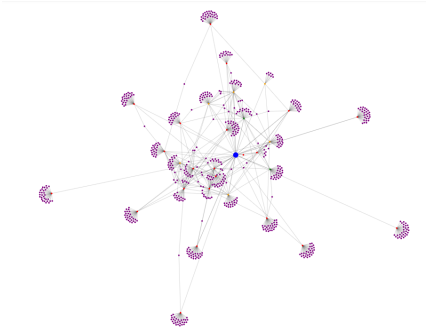


On voit donc que sur les 50 derniers tweets, l'UPHF interagit le plus avec @lavoixdunord, celui-ci ne fait pas partie des plus importants liens d'interaction dans le deuxième résultat.
On peut donc en conclure que dans le passé, l'UPHF n'avait pas une grosse relation avec @lavoixdunord, mais que ces derniers temps, leur relation s'est renforcée.

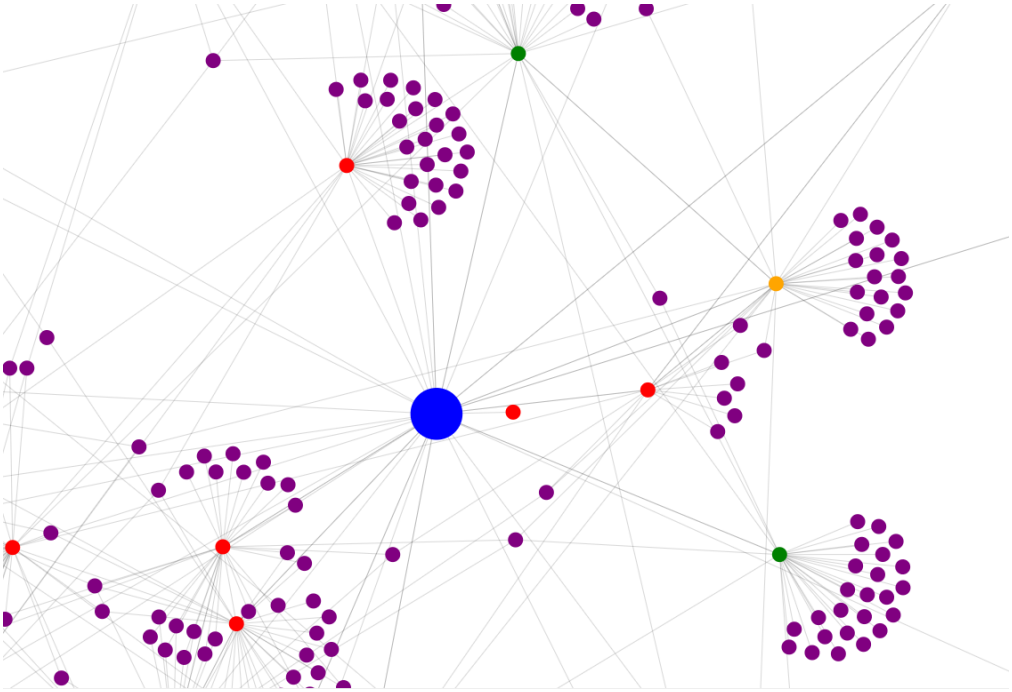
Nous allons maintenant activer la fonctionnalité “Depth 2” qui va permettre d’augmenter la profondeur du graphe de 1 :



L’outil nous donne le résultat suivant :



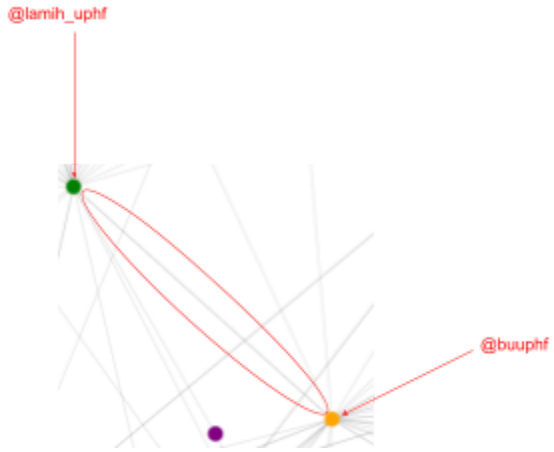
Nous allons zoomer au niveau de la profondeur 0 pour pouvoir mieux visualiser les informations :



Ici nous voyons le nœud représentant l’UPHF (nœud bleu).
Nous avons ensuite les nœuds colorés qui vont représenter les relations directes de l’UPHF (les mêmes relations que le résultat précédent).

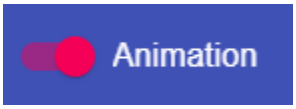
Puis nous avons les nœuds violets qui sont les relations indirectes de l'UPHF, c'est-à-dire les relations directes des relations directes de l'UPHF (profondeur 2). L'activation de cette profondeur va également nous montrer les liens entre les relations directes de l'UPHF.

Par exemple :

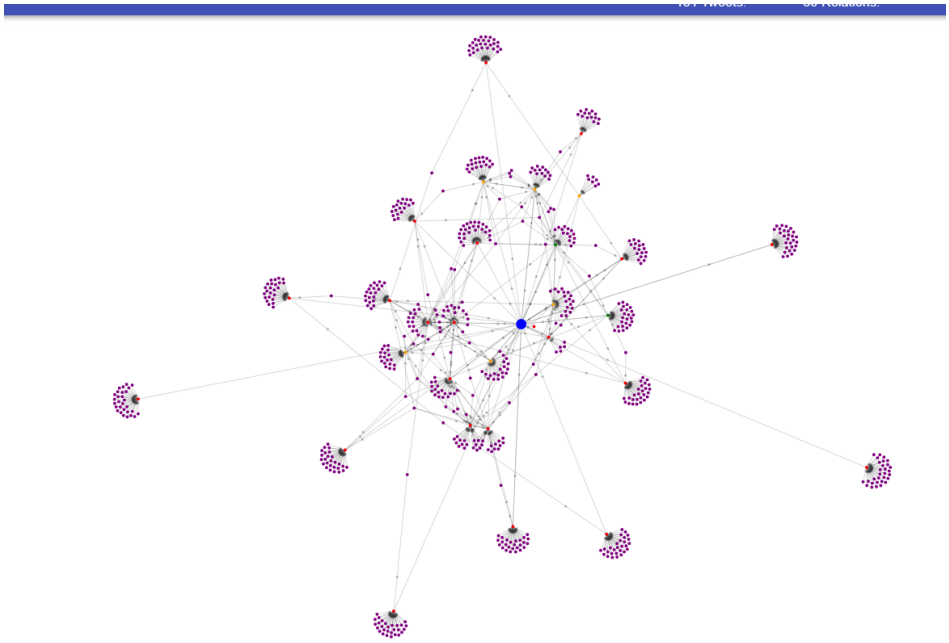


Nous pouvons voir, ci-dessus, qu'il y a une liaison entre @lamih_uphf et @buuphf qui n'étaient pas présents dans le résultat allant jusqu'à la profondeur 1.

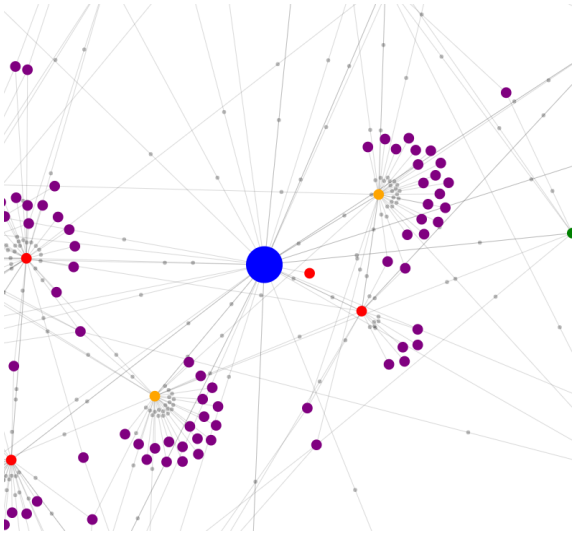
Maintenant nous allons activer l'activation des liaisons :



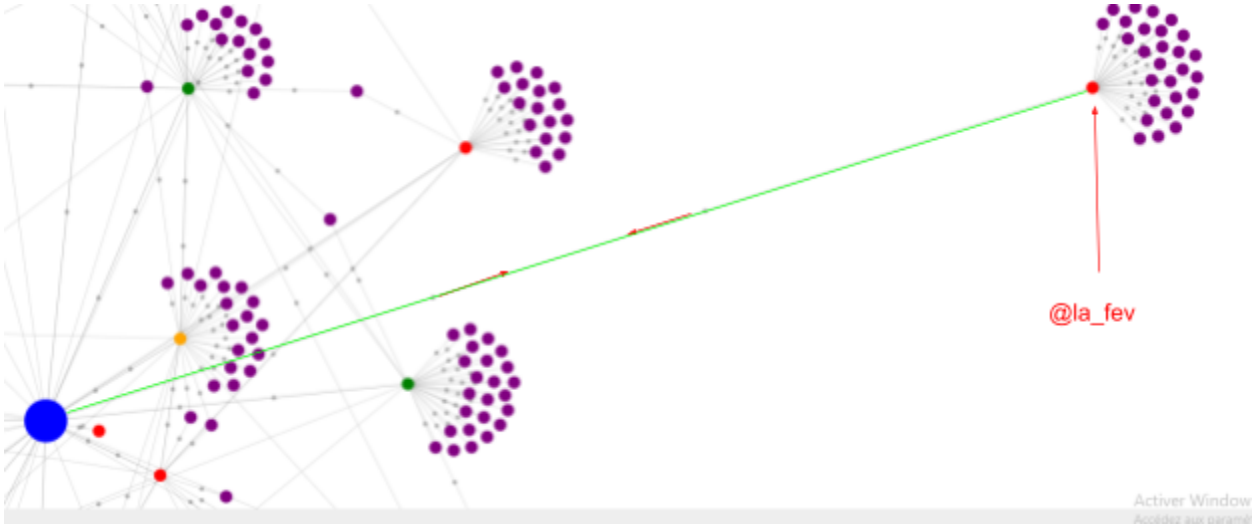
Nous obtenons le résultat suivant :



Pour une meilleure visualisation, nous allons zoomer au niveau de la profondeur 0 :



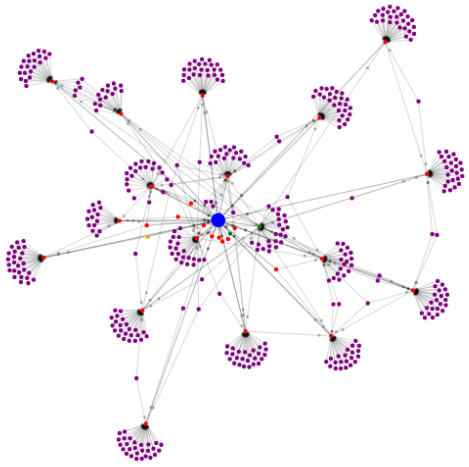
Nous pouvons voir qu'au niveau de chaque liaison, des bulles se déplacent dans un sens, le long des liens, représentant le sens d'une interaction entre deux nœuds.



Lorsque nous observons le lien qu'il y a entre l'UPHF on remarque que nous avons deux bulles qui se baladent sur la liaison dans des sens opposés.
Cela signifie donc que les deux utilisateurs se sont déjà mentionnés. On voit malgré tout que l'UPHF n'a pas beaucoup interagi avec @la_fev puisque celui-ci est représenté en rouge.

Imaginons que maintenant, nous voulons voir le degré de liaison que @la_fev a avec l'UPHF, nous allons faire un double-clic sur le nœud correspondant :

Cela va nous donner le graphe suivant :



Puisque nous savons que @la_fev a un lien direct avec l'UPHF, nous allons pouvoir désactiver l'animation et la profondeur 2 :

Cela nous donne le résultat suivant :

Parmi les relations de @la_fev, l'UPHF n'en fait pas partie. Or, les relations affichées représentent les comptes avec lesquels l'UPHF a les meilleures relations.
Donc cela signifie que la relation qu'a @la_fev avec l'UPHF ne fait pas partie des 30 meilleures relations que @la_fev possède.

Nos contraintes

Lors du déroulé de notre projet, nous avons eu plusieurs contraintes qui ont pu ralentir notre avancée ou encore réduire la qualité de nos résultats. Ces contraintes sont les suivantes :

La puissance de nos machines

Lorsque nous avons fait les premiers tests de TNP, nous avons eu un résultat qui était cohérent avec ce que nous attendions. Cependant, le fait de générer une animation de graphe consommait une grande partie des ressources de notre machine.
Cette consommation de ressources impactait énormément la fluidité de l'affichage de notre graphe. Cette perte de fluidité était notamment liée au fait qu'un très grand nombre de nœuds était généré en même temps.
Pour donner un ordre d'idée, lorsque nous réalisions un profilage sur nos propres comptes, nous avions environ 20 000 nœuds qui se génèraient en même temps, ce qui explique totalement que nos machines avaient du mal à gérer cette génération de masse.
Nous avons donc dû trouver un moyen de réduire le nombre de nœuds généré en même temps pour que notre machine puisse mieux supporter la génération du graphe.

Résolution de la contrainte

Nous avons ajouté les fonctionnalités concernant le nombre de relations et le nombre de tweets auxquels nous allions réaliser le profilage. Le fait d'avoir réduit le nombre de relations par utilisateur nous a fait considérablement réduire notre nombre de nœuds générés en même temps.
De plus, nous avons modifié notre algorithme afin d'éviter les doublons.

Exemple

Si nous mettons une limite de relation de 50 lors d'un profilage sur Alice01.
Admettons que Alice01 possède 100 relations au total. Seulement les 50 meilleures relations qu'il entretient seront retenues.
Admettons un second cas ou Alice01 possède seulement 30 relations. Alors ses 30 relations seront retenues.
Si l'on calcule le nombre maximum de noeuds avec une limite de relation à 50, on obtient :

1 (Noeud profondeur 0) + (50 (Noeuds de profondeur 1) * 50 (Noeuds de profondeur 2). Ce qui revient à : 1 + (50 * 50) = **2501 nœuds maximum.**

La fonctionnalité qui permet de limiter le nombre de tweets va aussi nous permettre de réduire le nombre de noeuds dans le graphe, notamment pour deux raisons :

- Lorsque l'outil de profilage va lancer le profilage sur un compte, il va essayer de trouver toutes les interactions que celui-ci a eu dans l'ensemble de ses tweets. De ce fait, avant que l'outil affiche le résultat sous forme de graphe, il faut qu'il récupère l'ensemble de ses tweets.
  Si la personne que nous profilons a fait 50 tweets depuis qu'elle possède son compte Twitter, cela ne va pas poser de problème à l'outil qui n'aura que 50 tweets à analyser.
  Néanmoins, il existe certains comptes Twitter, qui possèdent plusieurs dizaines de milliers de tweets. Le temps que l'outil de profilage traite tous ces tweets cela pourrait grandement augmenter le temps d'exécution.
  En utilisant une limite de tweets, cela va permettre à l'outil d'avoir moins de tweets à traiter et de pouvoir s'exécuter plus rapidement.
  Prenons une cible qui a environ 30 000 tweets mais que nous limitons le nombre de tweets à 50, l'outil de profilage n'aura que 50 tweets à décortiquer, ce qui n'impactera pas le temps d'exécution.

- L'autre utilité d'avoir la possibilité de limiter le nombre de tweets traités influe aussi sur le nombre de nœuds que nous allons générer lors de notre résultat. Plus l'outil de profilage va traiter les tweets d'un compte, plus il aura de chance de trouver de nouvelles interactions avec d'autres utilisateurs. En limitant le nombre de tweets, l'outil va trouver moins d'interactions entre la cible et les autres comptes. Ce qui résulte en un nombre de nœuds plus bas et favorisant la fluidité de l'affichage du graphe.

Contrainte de droits

Le gros problème de faire un outil de profilage est qu'il y a énormément de restrictions qui nous sont imposées pour éviter d'atteindre la vie privée des personnes profilées.
Il y a donc de nombreuses informations que nous ne pouvions pas utiliser (l'adresse, l'adresse mail, la localisation).
Nous avons alors rencontré des difficultés pour trouver une solution qui pouvait nous permettre de faire du profilage en utilisant des données qui n'avait aucune atteinte à la vie privée des personnes concernées.
De plus, lors de l'affichage de notre graphes, nous ne pouvions pas nous permettre d'afficher n'importe quelles informations.
Nous ne pouvions afficher le contenu du compte Twitter de chaque compte qui faisait partie du graphe, nous nous devons de rester assez vague par rapport aux informations de chaque compte.

Évolutions de TNP

Nous sommes satisfaits du résultat final de notre projet, cependant, nous pouvons tout de même penser à d'éventuelles améliorations qui pourraient rendre TNP encore plus performant et intéressant.

Optimisation de l'outil

Nous pourrions essayer d'optimiser les ressources que va utiliser notre outil. Actuellement, à cause des ressources que l'outil utilise pour générer le résultat, nous sommes obligés de réduire le nombre d'entités affichées à l'écran pour éviter d'éventuelles pertes de fluidité.
En optimisant l'outil, cela permettrait d'augmenter les nombre d'entités affichées à l'écran pour avoir un résultat plus précis et n'avoir aucune perte de fluidité.
Cela nous donnerait la possibilité de pouvoir augmenter la limite de tweets et la limite des relations, nous donnant la possibilité d'obtenir des résultats venant d'un plus grand nombre de tweets analysés et nous permettant d'avoir plus de relations avec la cible.

Optimisation du graphe

Actuellement, notre graphe nous permet de récupérer beaucoup d'informations concernant la personne profilée. Cependant, en optimisant davantage, cela nous permettrait de récupérer plus facilement ces informations.
Lorsque nous mettons l'animation pour un graphe, celle-ci ne va pas afficher le sens des liaisons se situant dans les deux premières profondeurs du graphe. L'animation ne sera donc pas bidirectionnelle.
Par exemple, si Alice01 a interagi 5 fois avec Bob02, alors, l'animation ira d'Alice01 vers Bob02, mais pas l'inverse.
Nous ne pourrons donc pas voir depuis le graphe si Bob02 interagit également avec Alice01.

En optimisant l'animation, chaque liaison animée sera bidirectionnelle et permettra de savoir si deux personnes interagissent l'une avec l'autre.

Nous pouvons également modifier notre manière de montrer le degré d'interaction entre deux personnes. Actuellement, nous utilisons des codes couleurs.

Cependant, nous ne pouvions pas utiliser les mêmes codes couleurs pour deux profondeurs différentes car cela aurait pu porter à confusion et réduire la visibilité du graphe..
En trouvant une autre manière de représenter le degré d'interaction, cela permettrait de pouvoir donner un degré de liaison entre la profondeur 1 et la profondeur 2.

Cette nouvelle manière nous permettrait également d'afficher le degré de liaison qu'il y a entre un nœud appartenant à une certaine profondeur et un autre appartenant à une profondeur inférieure.
Par exemple, les seuls degrés de liaison qui apparaissent actuellement dans le graphe sont ceux étant entre la profondeur 0 et la profondeur 1.
Avec notre nouvelle manière de représenter les degrés de liaisons, nous aurions donc la possibilité de montrer le degré de liaison entre un nœud de profondeur 1 et un nœud de profondeur 0 ou encore entre un nœud de profondeur 2 et un nœud de profondeur 1.

Cela permettrait donc de voir la réciprocité d'une relation entre deux utilisateurs.

Evolution de l'environnement

Actuellement, TNP (Twitter Network Profiler) est basé sur le réseau social Twitter.
Il est, actuellement, impossible d'utiliser notre outil sur un autre réseau social que Twitter.
Nous pourrions faire évoluer TNP de profilage en donnant la possibilité aux utilisateurs de choisir la plateforme sur laquelle ils souhaitent faire du profilage.
On pourrait imaginer qu'on puisse réaliser du profilage sur Facebook ou encore Instagram.

Conclusion: Notre objectif est-il atteint ?

Nous rappelons que l'objectif de notre projet était réparti en deux sous-objectifs.
Le premier sous-objectif était d'en apprendre davantage sur le milieu du profilage.
Le second était de créer notre propre outil de profilage tout en respectant les règles du RGPD

Concernant, le premier sous-objectif, nous avons appris beaucoup de choses concernant le profilage, notamment au niveau des différentes restrictions. Nous savions que le profilage était très mal vu par les organisations de protections de données et qu'elle était considérée comme une pratique malveillante qui enfreint la vie privée des personnes.Cependant, nous avons également appris que le profilage était majoritairement interdit (mise à part quelques dérogations). Le fait d'avoir créé TNP nous a aussi permis de voir à quel point les outils de profilage peuvent être puissants lorsqu'ils sont à plus grande échelle et avec des informations plus sensibles divulguées.

Pour notre deuxième sous-objectif, nous avons réussi à développer un outil de profilage à partir d'informations qui ne mettaient pas en danger les personnes profilées. De plus, nous avons réussi à obtenir un résultat qui comblera nos attentes, qui nous a permis de récupérer des informations intéressantes concernant les cibles sans pour autant faire atteinte à leur vie privée ou encore enfreindre les règles du RGPD.

Nous pensons donc qu'au final, ce projet nous aura été bénéfique et nous aura permis d'en apprendre plus sur le sujet. Nous estimons donc avoir rempli les objectifs que nous nous sommes imposés.

Timeline de notre projet

Voici comment s'est déroulé le développement de notre projet :

De fin janvier à début février

Pour commencer, nous avons dû trouver le sujet de notre projet.
Nous avons eu assez vite l'idée de le faire sur le profilage car nous voulions en apprendre plus sur le sujet. Ce qu'il nous a pris le plus de temps pendant cette période était de trouver un moyen afin de respecter les règles du RGPD, et ne pas être dans l'illégalité.

De début février à mi-février

Une fois l'idée de base trouvée, nous avons fait toutes les recherches nécessaires sur le profilage (sa définition, ses utilités, ses avantages, ses inconvénients…). Nous avons également étudié l'ensemble des lois à respecter par rapport au profilage, notamment celles en rapport avec le RGPD.

De mi-février à fin février

Avec les prérequis, il était temps de commencer à développer notre projet.
Nous avons structuré TNP, choisi la manière dont nous allions l'agencer, l'interface qu'il aurait, les fonctionnalités qui seraient disponibles, la manière dont nous allions afficher les résultats.

De début mars à mi-mars

Le squelette de TNP réalisé, nous avons cherché, puis trouvé le moyen de récupérer les informations de Twitter pour ensuite pouvoir les utiliser. En parallèle, nous avons également commencé à développer notre interface.

De mi-mars à fin mars

Avec l'arrivée des données, nous avons commencé à travailler sur nos différentes fonctionnalités pour améliorer la qualité des résultats de TNP (limite de tweets, limite de relation, recentrer le graphe).

De début mai à mi-mai

Nous avons résolu quelques derniers problèmes au niveau de notre outil puis nous avons réalisé ce rapport.

Sitographie

Recherches

https://www.cnil.fr/fr/vos-droits-lintervention-humaine-face-votre-profilage-ou-une-decision-automatisee

https://www.youtube.com/watch?v=ciC6q4M80cQ&ab_channel=Dolmen

https://www.youtube.com/watch?v=9aXpiGyNUOk&ab_channel=LaQuadratureduNet

https://www.legalplace.fr/guides/profilage-rgpd/

https://www.april.org/qu-est-ce-que-le-profilage

Code

https://twitter.com/

https://www.typescriptlang.org/

https://nextjs.org/

https://fr.reactjs.org/

https://material-ui.com/

https://github.com/vasturiano/react-force-graph

https://stackoverflow.com/

https://zellwk.com/blog/async-await-in-loops/