Roofstreet : système de traitement de données provenant d'équipements connectés en garantissant la sécurité des données et la protection de la vie privée

Anh-Dung Nguyen, Toussaint Tigori, Malek Hassani, Angel Gomez
{prenom.nom}@roofstreet.io

Résumé--

Le présent article concerne le domaine du traitement en grande masse des données personnelles provenant d'équipements connectés. La solution proposée concerne en particulier un système informatique capable de collecter et traiter des données relatives aux habitudes des individus, tout en garantissant la confidentialité des données personnelles de chaque individu.

I. INTRODUCTION

La connaissance et la compréhension des activités humaines, notamment de la mobilité humaine sont indispensables dans de nombreux domaines à savoir la gestion du trafic routier, la ville intelligente, la construction et l'aménagement urbain, la santé publique, la gestion des points de vente, le géomarketing, etc. La révolution numérique a permis aux acteurs économiques de collecter assez facilement un volume massif de données personnelles, dans le but de produire des analyses statistiques sur l'ensemble de la population. En effet, les équipements connectés de nouvelle génération, tels que les téléphones, les ordinateurs portables, les tablettes ou les montres intelligentes, sont dotés de fortes capacités de détection, de stockage et de traitement. Ceux-ci permettent de collecter un large spectre de données personnelles. La plupart des systèmes de collecte et d'analyse de données personnelles actuels suivent l'approche dite de « détection opportuniste », dans laquelle les participants (les détenteurs des équipements connectés) ne sont pas forcément conscients de la collecte [1]. Néanmoins, cela pose un sérieux problème concernant la sécurité des données et la vie privée des participants.

En effet, aujourd'hui, la collecte et le traitement des données personnelles, telles que les données de géolocalisation sont réalisés par des systèmes de type « cloud computing ». Dans ce type de systèmes, les données brutes collectées (souvent une date, une position et une statistique à mesurer par exemple la vitesse), via les capteurs d'un équipement connecté, sont envoyées à un centre de données distant (le cloud) pour être ensuite traitées et agrégées afin de produire des statistiques. Toute la partie de calcul, par exemple l'apprentissage automatique dans le cas des systèmes intelligents, est réalisée au niveau du « cloud ». Même si le but est de calculer des statistiques agrégées sur un ensemble d'équipements connectés, les serveurs peuvent très bien lier une donnée avec un équipement connecté et reconstruire tous ses déplacements. Connaitre les déplacements d'une personne permet de détecter ses points d'intérêt (lieux de résidence, de travail), ses habitudes, voire même sa religion, ses préférences politiques [2] et sa santé [3]. En conséquence, la vie privée de la personne est exposée. Ce problème a initialement été adressé par les techniques dites de pseudo-anonymisation, où le but consiste à supprimer tout lien entre l'identité d'une personne et ses données (l'identifiant de l'équipement connecté dans l'exemple précédent). L'anonymisation des données de géolocalisation est en soi un problème très difficile. Des études par la suite ont montré que ces techniques de pseudo-anonymisation ne suffisent pas [4]. Par exemple, dans une base de données de géolocalisation, quatre points spatio-temporels suffisent pour identifier une personne [5]. Il y a donc une nécessité de concevoir un système de traitement de données numériques provenant d'équipements connectés qui protège, par design, la vie privée des participants.

Garantir la sécurité des données et la protection de la vie privée des participants dans un tel système, surtout dans le contexte d'un service basé sur la géolocalisation, est un défi technique très complexe car cela concerne toutes les composantes du système dont le mode de collecte, la base de données, la communication et l'authentification. En effet, le système doit assurer :

- la protection des données contre une attaque de ré-identification au niveau de la base de données [6],
- l'anonymat des équipements connectés dans la communication [7],
- l'anonymat et le chiffrement des échanges entre les parties [8],
- l'authentification et la certitude des informations

transmises entre les parties [9].

Des techniques existantes résolvent partiellement ce problème et à notre connaissance, aucune d'entre elles ne proposent une solution complète et satisfaisante couvrant tous les points mentionnés précédemment.

II. TRAVAUX CONNEXES

Concernant le mode de collecte, alors que dans une approche centrée sur le cloud, les données brutes issues des équipements connectés sont envoyées directement dans le cloud, ce qui pose un risque de ré-identification du participant, il existe une solution qui consiste à construire un modèle d'apprentissage partagé entre les équipements connectés. Dans cette configuration, le procédé d'apprentissage est toujours exécuté au niveau du serveur, chaque équipement connecté télécharge le modèle actuel, le met à jour avec ses données brutes et renvoie le modèle mis à jour au serveur. Cette solution, décrite dans les brevets [10] et [11], est utilisée par exemple pour l'apprentissage automatique des mots tapés par des participants sur un terminal mobile afin de faire des prédictions et corrections. Dans le contexte d'un service basé sur la localisation, cette solution ne garantit pas la vie privée d'un participant, car le serveur peut toujours identifier une personne dans un endroit si elle est la seule personne qui émet la mise à jour.

Parmi les techniques de protection des données personnelles, en particulier des données de géolocalisation, celles qui sont basées sur la notion de k-anonymisation [12], sont les plus répandues. Cette notion garantie que, dans le cas d'une attaque d'une base de données, l'attaquant ne peut distinguer un participant particulier parmi k autres participants. Dans le contexte des données de géolocalisation, ceci se traduit par la dissimulation de la position exacte d'un participant dans une zone où il y a k autres participants. Les brevets [13], [14], [15], [16] décrits ci-dessous, exploitent cette technique. Nous verrons par la suite pourquoi le k-anonymisation ne suffit pas pour garantir entièrement la sécurité et la vie privée.

Selon l'approche de k-anonymisation, le brevet [13] présente une méthode d'anonymisation du mouvement spatial d'un périphérique mobile permettant d'assurer la vie privée de son utilisateur. Elle s'applique dans le cadre des services basés sur la localisation et permet à un utilisateur sollicitant ce type de service, par exemple une recherche de restaurants à proximité, d'éviter de divulguer sa position exacte et de préserver ainsi sa vie privée en se basant sur le principe du k-anonymisation. Cette technique consiste à noyer la position d'un équipement connecté dans une large zone. Initialement, l'espace géographique ou la zone considérée est subdivisé en plusieurs cellules. Ensuite, le mobile sollicitant un service qui nécessite sa position interagit avec un serveur qui lui fournit un historique comportant le nombre de mobile contenu dans chaque cellule. Une fois l'historique reçu du serveur, au moyen d'un algorithme récursif, le mobile détermine une sous-zone contenant au moins k mobiles dans le but de satisfaire la condition du k-anonymisation. Cette sous zone est ainsi utilisée pour solliciter le service nécessitant la position du téléphone. De cette manière, l'anonymisation de la position du téléphone est toujours garantie. Toutefois, cette architecture deviendrait vulnérable si le serveur parvenait à être corrompu. Dans ce cas, le serveur enverrait des historiques erronées au mobile de sorte que le mobile renvoi une zone qui ne garantirait pas le k-anonymisation et ainsi compromettre sa vie privée.

Contrairement à la technique proposée précédemment, les auteurs du brevet [14] font l'hypothèse de la sureté d'un serveur d'anonymisation qui permet en fonction de la position exacte du mobile et d'un critère k qui exprime le niveau de k-anonymisation souhaité par le mobile de calculer une zone englobant la position du mobile et contenant au moins k mobiles. Une fois cette zone calculée, elle est envoyée au service nécessitant la localisation pour traiter la demande du mobile et ainsi garantir sa vie privée. Le principal problème de cette technique intervient lorsque l'hypothèse de la sûreté du serveur d'anonymisation est levée, car si le serveur subit une attaque interne ou externe, les positions exactes des mobiles pourront être divulguées.

Le brevet [15] présente une méthode pour l'anonymisation des données de géolocalisation d'un mobile. Les auteurs proposent plusieurs solutions qui consistent à ajouter un serveur de confiance pour assurer l'anonymisation des données. Concernant la première solution, lorsqu'une application requiert la position d'un mobile, ce dernier envoie d'abord sa position à un serveur d'anonymisation de confiance pour définir le niveau d'anonymisation en garantissant le k-anonymisation. Après avoir définit le niveau d'anonymisation, le serveur renvoie la localisation anonymisée au mobile puis ce dernier la transmet à l'application qui a sollicité sa position. Vu que cette solution présente une faille de sécurité en divulguant la position du mobile au serveur dit de confiance, une seconde solution a été proposée. Cette solution consiste à faire tout le traitement d'anonymisation au sein du mobile, puis à vérifier le niveau d'anonymisation par le serveur d'anonymisation de confiance. Une fois le niveau d'anonymisation validé par le serveur d'anonymisation, le mobile peut dès lors divulguer sa position anonymisée. Le problème de cette solution intervient dans le cas où le serveur de confiance est corrompu et donc susceptible de fournir des informations erronées au mobile pour qu'il divulgue sa position sans respecter le k-anonymisation.

Le brevet [16] présente une autre méthode de k-anonymisation et de protection de données spatiales pour les applications mobiles nécessitant ce type de données vis-à-vis de leurs utilisateurs pour effectuer un ensemble de services. Dans l'architecture présentée, le serveur de l'application requiert la position spatiale d'un participant. Mais pour des raisons d'anonymat et de vie privée, le participant ne renvoie pas directement sa position précise au serveur. Une entité tierce d'anonymisation qui peut être soit un serveur ou un équipement connecté se charge de définir une cellule géographique comportant au moins un certain nombre de participants et englobant la position exacte dudit participant. La cellule géographique ainsi que les adresses de chaque participant s'y trouvant sont envoyées audit participant. Ledit participant sélectionne un participant et lui transfert via son adresse la requête de sa position provenant du serveur de l'application. Le

participant sélectionné à son tour envoie directement sa position au serveur de l'application. De cette manière, ledit participant effectuant une demande de service via l'application n'enverra jamais sa position exacte au serveur de l'application et pourra protéger sa vie privée. Par contre, la vie privée d'un participant est compromise si l'entité tierce d'anonymisation parvient à être corrompu car celle-ci a accès à la position exacte, ainsi que les adresses, de tous les participants.

Il existe également des techniques de protection de la vie privée basées sur le principe de la confidentialité différentielle. Ce principe protège les données d'un individu dans une base de données statistique. Un algorithme est dit de confidentialité différentielle si la modification ou l'exclusion des données d'un participant donné, a un impact négligeable sur l'agrégation des données de toute la base [17]. Dans ce cas, il s'agit donc de l'indiscernabilité d'une donnée. Souvent, les techniques basées sur la confidentialité différentielle consistent à ajouter un bruit aléatoire et contrôlé aux données avant de faire l'agrégation. Le défi de ces techniques est d'assurer un bon compromis entre l'utilité et le niveau de confidentialité de la donnée.

Dans le contexte des services basés sur la géolocalisation, en particulier des services fournissant une information sur les points d'intérêt à proximité de la position actuelle d'une personne, une technique basée sur la confidentialité différentielle pour protéger la localisation de la personne consiste à envoyer au serveur une requête couvrant une zone de recherche plus large. Les résultats retournés par le serveur peuvent être ensuite filtrés au niveau du terminal mobile pour raffiner la recherche. Ainsi, la personne a quand même des résultats précis, sans compromettre sa position exacte [18]. Cette technique n'est pas adaptée pour la collecte des données de trajectoires car la corrélation temporelle des données, même bruitées, permet quand même de reconstruire les habitudes de déplacement d'un participant.

Le brevet [19] présente une autre méthode d'anonymisation basée sur la confidentialité différentielle d'une base de données déjà établie contenant des informations spatiales. Cette technique permet de répartir les données spatiales dans des grilles puis pour chaque grille, un bruit est calculé puis appliqué à l'ensemble des positions spatiales qui s'y trouvent afin de les anonymiser. Le problème de cette technique concerne l'exploitabilité des données anonymisées car leur précision est impactée par le processus d'anonymisation. Dans le cas d'un système de collecte de données statistiques, anonymiser les données via cette technique pourrait biaiser fortement les statistiques issues de ces données.

De tout ce qui a été mentionné précédemment, nous pouvons déduire que les différentes techniques basées sur la confidentialité différentielle ne sont pas adaptées pour la collecte statistique des données de déplacement.

Une autre solution pour garantir l'anonymat des participants dans un système de collecte est d'utiliser un serveur dit serveur d'anonymisation jouant le rôle de médiateur entre les participants et le serveur de collecte. Ledit serveur d'anonymisation est dit de confiance pour masquer l'identité du participant à chaque échange de ce dernier avec le serveur de collecte. Cette technique souffre du même problème de ré-identification que les techniques de pseudo-anonymisation face aux données de haute dimension, telles que les données de géolocalisation. Ceci dû à la forte unicité de ce type de donnée.

Selon ladite approche, le brevet [20] décrit un procédé pour collecter des profils des utilisateurs mobiles d'une manière anonymisée, dans le but de diffuser des messages ciblés, par exemple une publicité personnalisée. La technique repose sur un serveur d'anonymisation interposé entre l'équipement connecté et le serveur de collecte. Ledit serveur d'anonymisation est la seule entité qui connaît l'identité du participant. Il se charge de chiffrer l'identité de l'équipement connecté à chaque échange entre ce dernier et le serveur de collecte. Ce schéma permet audit serveur de collecte de connaître le profil du participant sans pouvoir connaître l'identité de ce dernier. Néanmoins, cette technique ne suffit pas pour anonymiser des données de haute dimension ou des données de géolocalisation car l'unicité de ce type de donnée permet de retrouver assez facilement l'identité d'un participant.

Le brevet [21] propose une méthode pour anonymiser l'historique de localisation d'un mobile en se basant sur la détection des points stationnaires, la solution consiste à stocker les localisations dans un premier temps dans une base de données temporaire (en quarantaine) pendant une semaine par exemple avant de les exporter sur la base de données long terme. Après avoir récolté des localisations sur un mobile donné sur la base de données de quarantaine, une procédure de détection de tous les points stationnaires (maison, travail, etc.) sensibles en matière de vie privée est effectuée. Une fois ces points stationnaires détectés, un processus de nettoyage permettant de les remplacer par une seule position est effectué. Les données traitées sont enfin exportées vers la base de données long terme. La vulnérabilité de cette technique se caractérise par l'accès à la base de données de quarantaine par une personne malhonnête qui parviendrait à avoir accès aux données personnelles des utilisateurs.

Le brevet [22] présente une technique qui consiste à appliquer une modification, par exemple, une troncature ou un arrondissement, à la géolocalisation d'un utilisateur en prenant en compte l'indication de son degré de confidentialité. Le degré de confidentialité est défini en fonction soit d'une mesure de distance, d'une mesure d'angle ou d'une valeur sélectionnée par l'utilisateur. Cette technique ne peut garantir l'anonymisation de la collecte statistique des données de déplacement car la corrélation temporelle entre les données de géolocalisation consécutives peut révéler l'identité d'un participant, comme mentionné précédemment. D'ailleurs, la technique ne permet pas d'obtenir un bon compromis entre l'utilité de la donnée et le niveau de protection de la vie privée.

Dans le cas des services basés sur la localisation, le brevet [23] présente une technique qui consiste à associer la position réelle de l'utilisateur à un ensemble de fausses positions générées à partir de celle-ci. Toutes les positions y compris celle de l'utilisateur sont envoyées au serveur pour exécuter le service demandé. Une fois le service traité, les réponses associées à chaque position sont renvoyées à l'utilisateur qui à partir de sa position réelle identifie la réponse adéquate. Cette technique permet certes de garantir la confidentialité de la

position d'un utilisateur vis-à-vis du serveur mais n'est pas adaptée dans le cadre d'une collecte statistique. En effet, utiliser cette technique reviendrait à envoyer au serveur de collecte des fausses données qui biaiseront les statistiques.

Malgré de nombreuses solutions proposées pour protéger la confidentialité des données, la plupart des solutions souffrent du même problème technique. En effet, elles partent du principe que les serveurs d'anonymisation ou de calculs sont des entités de confiance sans répondre vraiment à la question sur l'authentification et la certitude des calculs réalisés. Il est nécessaire de se protéger contre toute attaque qui pourrait provenir des serveurs malintentionnés.

Une solution serait de vérifier la certitude de toutes les informations issues de ces serveurs. Plusieurs travaux [24], [25] se sont intéressés à ce type de problème et ont proposé des méthodes d'authentification des services basés sur la localisation. Ces méthodes se focalisent sur l'utilisation des signatures électroniques basées sur la cryptographie asymétrique où l'idée est de signer toutes les informations qui sont transmises au serveur par les équipements connectés. De cette manière, le serveur ne contient que des informations certifiées ou signées par les équipements connectés. Alors, lorsqu'une requête est effectuée par un équipement connecté au serveur, le serveur lui retourne en plus du résultat de la requête toutes les preuves permettant de prouver l'authenticité dudit résultat.

Dans l'article [26], les auteurs stipulent que les techniques citées précédemment sont insuffisantes car elles permettent certes de vérifier l'exactitude des informations mais pas leur complétude car le serveur peut délibérément omettre des informations sans le notifier. De plus, les auteurs indiquent que ces techniques supposent que l'équipement connecté est de confiance car les informations qui lui sont retournées pour l'authentification des résultats sont signées mais non chiffrées. Dans ce cas de figure, un équipement connecté a donc la possibilité d'accéder aux données personnelles d'autres équipements connectés et cela pourrait compromettre la vie privée de ceux-ci. Ils présentent alors une technique permettant aux équipements connectés de vérifier à la fois la complétude et l'exactitude des résultats retournés par le serveur. Les données personnelles sont d'abord signées par une entité de confiance, puis chiffrées par le serveur via une technique de chiffrement homomorphe. En recevant les données chiffrées et signées qui servent de preuve, l'équipement connecté peut vérifier la complétude et l'exactitude du résultat, sans avoir accès aux données personnelles des autres équipements connectés.

Malgré tout, les techniques citées précédemment permettent seulement de résoudre le problème de l'authentification des informations provenant du serveur. Elles ne permettent pas de garantir la confidentialité des données personnelles dans le cas d'un serveur compromis ou malintentionné car ce dernier a bien accès aux données personnelles, en clair, des équipements connectés.

Toujours selon l'hypothèse des serveurs compromis ou malintentionné, des techniques de cryptographie dite techniques de chiffrement homomorphe ont été proposées [27], [28]. Elles sont plus utilisées dans les systèmes de type cloud computing et permettent d'assurer la confidentialité des données qui sont envoyées au serveur à des fins de calculs. En effet, la particularité de ces techniques se base sur la capacité du serveur distant à pourvoir réaliser des opérations mathématiques, dont le type est spécifique à chaque technique, sur des données chiffrées. De cette manière, le serveur n'a aucune connaissance sur les données qu'il manipule. Néanmoins, cette technique nécessite d'être complémentée par d'autres techniques, afin de résoudre les problèmes mentionnés.

Le problème de l'anonymat d'une communication peut être résolu en utilisant des réseaux de communication anonymes. En effet, ces réseaux décentralisés permettent d'anonymiser l'origine (l'adresse IP) d'un paquet en le chiffrant plusieurs fois à travers des routeurs d'un chemin aléatoire [29]. Cette technique peut être utilisée par exemple pour anonymiser les échanges entre les équipements connectés et les serveurs.

Parmi les solutions qui tentent de résoudre plusieurs problèmes mentionnés ci-dessus, on trouve PrivStats [30]. Les auteurs proposent un système de collecte et d'agrégation des données de géolocalisation en se basant sur un modèle de menace dit « zero trust », qui consiste à ne faire confiance à aucune des parties du système, y compris le serveur de collecte et les équipements connectés. En utilisant ladite technique de chiffrement homomorphe, le serveur de collecte agrège les données chiffrées envoyées via un réseau anonyme par les équipements connectés. Le serveur doit ensuite faire appel à certaines entités désignées possédant la clé de déchiffrement, qui peuvent être des équipements connectés, pour récupérer le résultat de l'agrégation qu'il a effectué. Ce schéma repose sur une phase de « boostrap » permettant de désigner ces derniers et de partager entre eux les clés de chiffrement homomorphe. Chaque équipement connecté (client) joignant le système reçoit la clé de chiffrement via une entité désignée. Les équipements connectés désignés ont également la capacité d'auditer le serveur de collecte, afin de vérifier la certitude des agrégations effectuées par ce dernier. Néanmoins, cette solution n'empêche pas l'envoi de données dont l'unicité permettrait d'identifier une personne durant un déplacement [5]. En effet, les données sont transmises sans garantie de k-anonymisation ou de confidentialité différentielle. Par ailleurs, cette solution n'explique pas comment distribuer les clés de chiffrement homomorphe aux équipements connectés sans risquer de les exposer à une partie malveillante.

III. ARCHITECTURE DU SYSTEME

Notre solution consiste en une combinaison de moyens techniques (équipements connectés portées par les participants et serveurs informatiques et de traitement informatique des données) pour satisfaire les conditions suivantes :

- (i) Dans le système, le participant ne fait confiance ni au serveur qui collecte des statistiques, ni aux autres participants. Il ne veut divulguer aucune information liée à son identité et à sa vie privée. Le participant ne fait confiance au serveur que pour agréger des mesures sur l'ensemble des participants et calculer les

statistiques.

- (ii) Inversement, le serveur ne fait pas confiance aux participants, pour éviter qu'ils biaisent ses statistiques.
- (iii) Toutes les données sensibles liées à l'identité ou à la vie privée d'une personne ne sortent jamais de l'équipement connecté.
- (iv) Toutes les données transmises aux serveurs sont anonymisées.
- (v) La seule chose que le serveur puisse apprendre des participants est une agrégation des mesures provenant de l'ensemble des participants.
- (vi) L'équipement connecté a le contrôle sur ce qu'il souhaite divulguer. Le système est responsable de l'obtention d'une meilleure précision possible des statistiques, en fonction de ce que les participants divulguent.

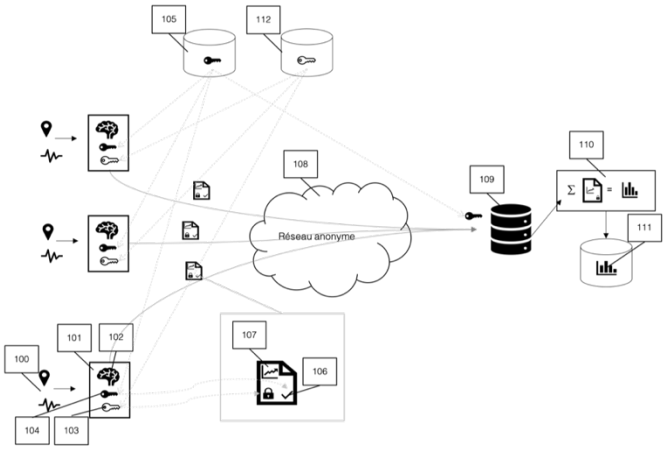

FIG.1. Architecture du système

La FIG.1 illustre notre solution. Le système qui décide de collecter une statistique (par exemple, la vitesse moyenne des personnes sur tous les axes routiers en France), active le module de collecte de ladite statistique au niveau des équipements connectés (101) et leur envoie des informations concernant cette collecte. Toutes les communications entre le serveur de collecte (109) et le participant s'effectuent via un réseau anonyme (108). Avantageusement, cela évite que le serveur (109) identifie la source de chaque message qu'il reçoit afin de traquer un équipement connecté (101), en utilisant par exemple les techniques de suivi d'adresse IP. En conséquence, ceci satisfait partiellement les conditions (i), (iii), (iv) et (v).

Le module de collecte fait appel aux capteurs embarqués pour la collecte des données brutes (100). Une donnée brute (100) correspond à une géolocalisation horodatée associée à une mesure (dans l'exemple précédent, la vitesse). Ces données sont ensuite stockées par un moyen de stockage disponible sur l'équipement connecté (101). Un module d'intelligence artificielle (102) récupère ces données pour entrainer un modèle d'apprentissage automatique. Ce dernier permet de reconnaitre des schémas de déplacement et d'en déduire des habitudes de déplacement du participant. Le module intelligent (102) se charge également de calculer les mesures pour chaque déplacement, propres au participant, par exemple, la vitesse moyenne sur le trajet maison-travail. Ces mesures sont stockées via le moyen de stockage sur l'équipement connecté (101). En fonction des habitudes détectées, le module intelligent (102) définit lui-même le niveau de confidentialité pour chaque mesure. Plus une habitude est répétée, par exemple son lieu d'habitation, son lieu de travail ou son trajet maison-travail, plus le niveau de confidentialité devrait être élevé. Cette technique permet de satisfaire les conditions (iii) et (vi).

Pour envoyer les mesures anonymisées au serveur de collecte (109), l'équipement connecté (101) et le serveur (109) suivent un procédé d'anonymisation sécurisé et garantissant la protection de la vie privée. D'abord, toutes les mesures sont chiffrées, de préférence, par un système de chiffrement homomorphe que seuls les équipements connectés (101) sont capables de déchiffrer. Le serveur (109) est capable de faire des opérations mathématiques (110) sur ces mesures chiffrées mais ne peut en aucun cas connaître leurs valeurs exactes. Cette technique permet de garantir que le serveur (109) peut effectuer des agrégations statistiques (111) sur des mesures sans connaitre leurs valeurs (condition (v)).

La solution proposée protège la vie privée des participants par un mécanisme de généralisation (illustré par FIG.2). L'anonymat d'un participant par rapport à une mesure donnée est garanti par le principe de k-anonymat (illustré par FIG.3). En effet, le procédé d'anonymisation garantit que pour chaque mesure envoyée, il existe au moins k autres participants qui ont envoyé la même mesure. Chaque mesure envoyée est généralisée, par exemple : une vitesse de 27 km/h est généralisée par une plage de valeurs allant de 20 à 30 km/h. En particulier, les données de géolocalisation sont généralisées par un système de zones hiérarchiques (illustré par FIG.2) dont le plus bas niveau correspond à une petite zone (par exemple une rue) et le plus haut niveau correspond à une zone très large (par exemple un pays). Chaque équipement connecté (101) envoie d'abord sa mesure avec le plus haut niveau de généralisation. Il n'accepte de descendre à un niveau inférieur que si le serveur lui prouve qu'il existe déjà k autres équipements connectés avec une mesure similaire.

Pour prévenir le cas d'un serveur malintentionné qui pourrait tromper les équipements connectés (101) en leur envoyant des fausses informations, un équipement connecté (101) doit pouvoir vérifier et authentifier les informations que le serveur (109) lui envoie. Pour cela, un mécanisme d'authentification basé sur la cryptographie asymétrique et une entité tierce (105) est mis en place pour assurer que chaque mesure envoyée est authentifiable. Plus précisément, chaque équipement connecté (101) possède, en plus des clés de chiffrement homomorphe, des clés publiques et privées (104), générées et distribuées par l'entité tierce (105). L'équipement connecté (101) certifie l'origine de chaque mesure en la signant (106) par sa clé privée. Le serveur de collecte (109) doit prouver à un équipement connecté (101) l'exactitude du résultat de l'agrégation effectuée sur des mesures en lui envoyant, en plus dudit résultat de l'agrégation (111), k différentes mesures chiffrées signées (107) par d'autres équipements connectés. Avantageusement, cette dernière technique permet de garantir entièrement les conditions (i), (iii), (iv), (v).

IV. GENERALISATION DES MESURES

Afin de garantir une protection totale des données personnelles d'un participant, les différentes mesures produites doivent être anonymisées avant leur envoi au serveur de collecte (109).

L'anonymisation de ces mesures est basée sur une technique dite de généralisation permettant aux équipements connectés (101) de préserver leur anonymat. Un procédé GÉNÉRALISATION, pour chaque mesure produite précédemment, procède à sa généralisation, puis à l'enregistrement de la mesure généralisée via le moyen de stockage. Ladite généralisation est réalisée comme suit :

- Généralisation :

La généralisation est un procédé qui vise à réduire la précision d'une mesure en modifiant son échelle ou son ordre de grandeur respectif. Elle s'applique aussi-bien sur des mesures définies sur une dimension que sur des mesures définies sur plusieurs dimensions.

Considérons une mesure $P \in \mathbb{S}^n$, où n est le nombre de dimension, la généralisation de P est définie par

$$Gen(P) = S$$

où S est un espace de n dimensions tel que $P \subset S$ et S appartient à $\mathbb{S}^n$.

Par exemple :

- une généralisation d'une vitesse de 40 km/h consiste en l'intervalle 20-60 km/h,
- une généralisation d'une présence à 9h32 à la Mairie de Paris consiste en la présence dans le Premier arrondissement de Paris entre 9h et 10h.

La forme géométrique de S peut-être par exemple circulaire, rectangulaire, hexagonale ou des régions administratives comme Paris, Ile-de-France, France. Une adresse précise à 1 rue d'Uzès peut être généralisée par le 2ème arrondissement de Paris ou Paris ou Ile-de-France ou France.

La période S peut-être une minute, une heure ou une période définie par convention comme le weekend ou le matin.

La généralisation concerne un descriptif de personne, une donnée comme l'âge S d'un participant peut être une tranche d'âge.

- Généralisation multiniveau :

Une mesure peut être généralisée sur différentes échelles de manière à obtenir des mesures généralisées plus ou moins précises. Soit l'espace $\mathbb{S}$ structurée en H niveaux hiérarchiques, $\mathbb{S}$ est d'abord partitionné en n_H sous-espaces S_H^i contiguës, où $i \in [1, n_h]$. Pour chaque niveau $h \in [2, H]$, un sous-espace S_h^i est subdivisé en n_{h-1}^i sous-espaces S_{h-1}^i contiguës. La généralisation multiniveau d'un point P au niveau h est définie par

$$Gen_h(P) = S_h^i ,$$

où S_h^i est le sous-espace du niveau h qui contient P.

Dans l'exemple précédent, si on considère $\mathbb{S}$ comme la France, son partitionnement sur quatre niveaux consiste en des régions, des départements, des communes et des arrondissements. La position à 1 rue d'Uzès est généralisée au niveau 4 par France, au niveau 3 par Ile-de-France, au niveau 2 par Paris et au niveau 1 par le 2ème arrondissement.

Ce partitionnement hiérarchique est calculé préalablement par l'application mobile au démarrage de la collecte, puis stocké dans une base de données. Afin d'effectuer une généralisation de donnée, le procédé GÉNÉRALISATION prend comme entrée le paramètre h et applique une recherche dans ladite base de données pour trouver le sous-espace au niveau h qui contient la mesure. Le procédé GÉNÉRALISATION remplace ladite mesure par ce sous-espace. Il sauvegarde, le cas échéant, cette généralisation par le moyen de stockage.

Un exemple de généralisation multiniveau est illustré par FIG.2.

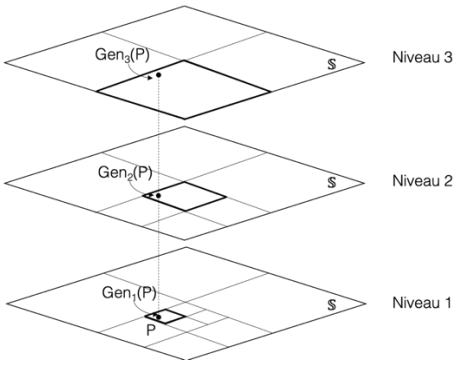

FIG.2. Généralisation multiniveau d'une mesure

- Généralisation suivant le concept du k-anonymat :

La généralisation multiniveau est couplée avec une technique d'anonymisation. La généralisation à elle seule ne peut garantir une anonymisation complète des mesures produites. En effet, si l'on suppose que l'équipement connecté (101) envoie toujours des mesures généralisées uniques au serveur de collecte (109), par exemple un participant résidant dans un village avec très peu d'habitant qui envoie ses comptages de passage. A partir de la base de données du serveur de collecte (109), lesdites mesures pourront servir à isoler le participant dudit équipement connecté (101) afin de l'identifier. Par conséquent, les données personnelles du participant sont compromises.

Pour résoudre ce problème, un procédé nommé ANONYMISATION protège les données personnelles du participant en exploitant le concept du k-anonymat. Ledit procédé garantit que dans la base de données des mesures reçues par le serveur de collecte (109), il existe au moins k participants qui soumettent la même mesure. De cette manière, une mesure généralisée ne peut être isolée afin d'identifier l'équipement connecté (101), et donc le participant, qui l'a envoyé au serveur de collecte (109).

Considérons les mesures P généralisées $Gen_h(P)$, reçues des différents participants par le serveur de collecte (109). L'anonymat d'un participant à cette collecte par rapport à P est garanti si et seulement s'il existe un niveau h tel que

$$|Gen_h(P)| \geq k.$$

Le niveau de confidentialité dudit système de collecte est donc défini en fonction de k.

Le niveau de confidentialité k d'un participant peut être défini en fonction des mesures, par exemple, ses données sensibles telles que les lieux d'habitation et de travail peuvent

être protégées avec un plus grand k. D'autre part, le niveau de confidentialité d'un participant peut être défini par lui-même. Il peut être également estimé d'une manière automatique et dynamique par le procédé ANONYMISATION, décrit ci-dessous, à partir des habitudes de déplacement du participant détectés.

Un exemple du k-anonymat est illustré par FIG.3.

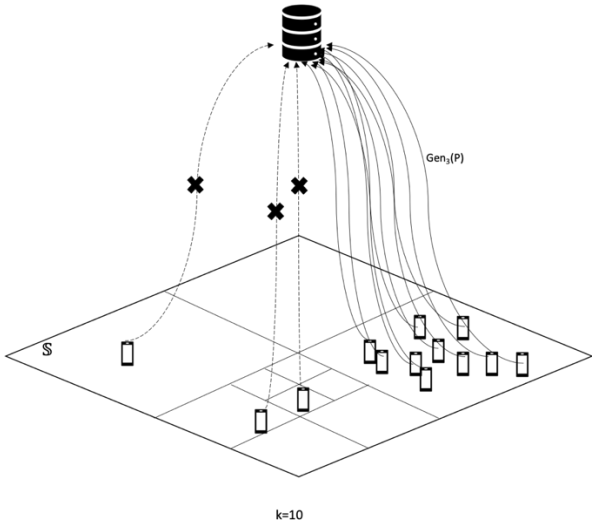

FIG.3. Principe du k-anonymat

V. ANONYMISATION DES MESURES

Dans un mode de réalisation particulier, le procédé ANONYMISATION consiste à envoyer une mesure audit serveur de collecte (109) tout en garantissant la sécurité et l'anonymat des données échangées. Dans ce procédé, il est supposé que chaque équipement connecté (101) e possède deux paires de clés publiques et privées (PK_e, SK_e) et (PK_m, SK_m) et que le serveur dispose d'une paire de clés publique privée (PK_s, SK_s) générées au démarrage du système.

Initialement, en fonction des habitudes de déplacement détectées par l'intelligence artificielle (102) (par exemple, maison, travail, lieu de passage fréquent, etc.), une valeur k est définie par chaque équipement connecté (101) pour chaque mesure. Par exemple, la valeur de k définie pour publier une mesure caractérisant une zone d'habitation d'un participant sera plus élevée que celle définie pour publier une mesure caractérisant un lieu de passage. La valeur donnée à k a une influence sur le niveau de confidentialité d'une mesure. Ainsi, si k est élevé (resp. bas), le niveau de confidentialité est beaucoup plus haut (resp. bas).

Le procédé ANONYMISATION assure en effet à travers une phase de négociation qu'une mesure généralisée transmise au serveur par un équipement connecté (101) garantit l'anonymat du dit équipement connecté (101) selon ledit concept du k-anonymat.

La phase de négociation du procédé ANONYMISATION peut s'effectuer suivant plusieurs manières dont :

- une négociation partant du niveau de généralisation le plus bas de la mesure à envoyer,
- une négociation partant du niveau de généralisation le plus haut de la mesure à envoyer,
- une négociation simultanée de plusieurs niveaux.

Considérons le mode de réalisation dans lequel la négociation s'effectue à partir du niveau de généralisation le plus haut de la mesure.

Soit M une mesure à envoyer au serveur de collecte (109) par un équipement connecté (101) e.

Initialement, l'équipement connecté (101) réalise une généralisation au niveau H, $Gen_H(M)$, de sa mesure M via le procédé GENERALISATION. La mesure généralisée est ensuite chiffrée au moyen de la clé de chiffrement homomorphe PK_m pour obtenir

$$Chiffre\big(Gen_H(M)\big),$$

puis signée par la clé de chiffrement asymétrique SK_e pour obtenir

$$Signe_e\big(Chiffre(Gen_H(M))\big).$$

Ce dernier est ensuite envoyé au serveur de collecte (109) via un réseau anonyme (108).

A la réception de la mesure, le serveur de collecte (109) vérifie son authenticité au moyen de la clé de chiffrement asymétrique PK_e via l'entité tierce de confiance (105). Une fois l'authentification terminée, le serveur de collecte (109) obtient

$$Chiffre\big(Gen_H(M)\big),$$

puis enregistre la paire

$$Chiffre\big(Gen_H(M)\big), Signe_e\big(Chiffre(Gen_H(M))\big),$$

en utilisant un moyen de stockage, par exemple, une base de données.

Supposons qu'après ladite réception, le serveur a reçu des mesures

$$Chiffre\big(Gen_H(M)\big)$$

provenant des équipements connectés (101) $e \in \{1, \dots, N\}$. Selon divers modes de réalisation, le serveur de collecte (109) effectue une agrégation (110), par exemple, une somme, à savoir

$$Agg\left(Chiffre\big(Gen_H(M)\big)\right) = \sum_{e=1}^{N} Chiffre\big(Gen_H(M)\big).$$

Ledit serveur de collecte (109) enregistre le résultat de ladite agrégation via un moyen de stockage.

Après ledit envoi, l'équipement connecté (101) sollicite auprès du serveur de collecte (109) le résultat de ladite agrégation. Le serveur signe alors le résultat de l'agrégation au moyen de sa clé de chiffrement asymétrique SK_s, à savoir

$$Signe_s\left(Agg\left(Chiffre\big(Gen_H(M)\big)\right)\right).$$

Lorsque l'équipement connecté (101) reçoit ledit résultat, il vérifie d'abord son authenticité au moyen de la clé de chiffrement asymétrique PK_s via l'entité tierce de confiance (105). Ainsi, l'équipement connecté (101) obtient

$$Agg\left(Chiffre\big(Gen_H(M)\big)\right).$$

Dans le cas où l'agrégation est un comptage, l'équipement connecté (101) vérifie si ladite agrégation est supérieur ou égal à k prédéfini. Si

$$Agg\left(Chiffre\big(Gen_H(M)\big)\right) < k,$$

l'équipement connecté (101) recommence une autre phase de

Algorithme 1 procédé *ANONYMISATION*

```
ENTREES
- Mesure M
- Équipement connecté e
- Serveur de collecte s

DEBUT
  // Généralisation au niveau H
  h := H;
  Gen_h(M) ;
  Envoyer  Signe_e(Chiffre(Gen_h(M))) ;
  Attendre T seconds;
  Demander et recevoir Signe_s(Agg(Chiffre(Gen_h(M)))) ;
  Authentifier et obtenir Agg(Chiffre(Gen_h(M))) ;

  // Négociation
  tant que Agg(Chiffre(Gen_h(M))) ≥ k et h > 1
     h := h − 1;
     Gen_h(M) ;
     Envoyer  Signe_e(Chiffre(Gen_h(M))) ;
     Attendre T seconds ;
     Demander et recevoir Signe_s(Agg(Chiffre(Gen_h(M)))) ;
     Authentifier et obtenir Agg(Chiffre(Gen_h(M))) ;
  Fin tant que

  // Envoi final
  si h ≠ H
      si h ≠ 1 ou (h == 1 et Agg(Chiffre(Gen_h(M))) < k)
         h = h + 1 ;
      Fin si
     Demander k signatures pour Agg(Chiffre(Gen_h(M))) ;
     k_anonymat := true ;
     pour chaque Signe_i(Chiffre(Gen_h(M))) reçue
          Authentifier Signe_i(Chiffre(Gen_h(M))) ;
          si échec k_anonymat := false ;
     Fin pour
     si k_anonymat == true
          Envoie Gen_h(M) ;
     Fin si
  Fin si
FIN
```

négociation avec le même niveau de généralisation après un certain temps, par exemple, 24 heures.

Dans le cas contraire, si

$$Agg\left(Chiffre\left(Gen_H(M)\right)\right) \geq k$$,

l'équipement connecté (101) vérifie l'intégrité du résultat de l'agrégation en demandant au serveur k différents mesures généralisées chiffrées signées

$$Signe_e\left(Chiffre(Gen_H(M))\right).$$

Lorsque l'équipement connecté (101) reçoit les k mesures généralisées chiffrées signées, il vérifie auprès de l'entité de confiance (105) leur authenticité puis valide le résultat. Une fois le résultat validé, l'équipement connecté (101) recommence une autre phase de négociation avec une mesure généralisée au niveau H-1, à savoir $Gen_{H-1}(M)$, et descend progressivement le niveau de généralisation, jusqu'à atteindre dans le cas idéal le niveau de généralisation le plus bas qui garantirait son k-anonymat. À savoir,

$$h_{optimal} = argmin_h\left(\left|Agg\left(Chiffre\left(Gen_h(M)\right)\right) - k\right|\right) .$$

Dès qu'un accord est trouvé, ledit équipement connecté (101) procède à l'envoi de sa mesure généralisée

$$Gen_{h_{optimal}}(M).$$

Si en fin de période de négociation un équipement connecté (101) ne parvient pas à obtenir un résultat de l'agrégation supérieur à k sur sa mesure généralisée la plus élevée, l'équipement connecté (101) annule l'envoi de ladite mesure généralisée au serveur.

Avantageusement, ce procédé permet de garantir le meilleur compromis entre la précision des mesures obtenues au niveau du serveur de collecte (109), et le respect de la vie privée des participants.

Pour éviter que l'équipement connecté (101) envoie plusieurs fois une même mesure dans le but de biaiser les statistiques, le serveur vérifie la duplication d'une mesure chiffrée signée. Dans le cas échéant, ledit serveur ignore les doublons.

Le procédé ANONYMISATION au niveau d'un équipement connecté (101) dans le cas d'un comptage est résumé par l'algorithme 1.

VI. CONCLUSION

Le présent article concerne un système de collecte et de traitement des données statistiques anonymisées liées aux déplacements des personnes et garantissant, par design, la protection de la vie privée. La solution proposée vise à, à la fois, obtenir une meilleure précision possible des statistiques, et respecter la vie privée des participants.

VII. RÉFÉRENCES

[1] Urban Sensing Systems : Opportunistic or Participatory, Nicholas D. Lane et. al., ACM HotMobile 2008

[2] The Long Road to Computational Location Privacy : A Survey, Vincent Primault et. al., IEEE Communications Surveys & Tutorials, 2018

[3] Using Autoencoders to Automatically Extract Mobility Features for Predicting Depressive States, Abhinav Mehrotra and Mirco Musolesi, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2, 3, Article 127, 2018

[4] Anonymization of Location Data Does Not Work : A Large-Scale Measurement Study, Hui Zang et Jean Bolot, ACM MobiCom 2011

[5] Unique in the Crowd : The privacy bounds of human mobility, Yves-Alexandre de Montjoye et. al., Scientific Reports, 2013

[6] Robust De-anonymization of Large Datasets, Arvind Narayanan et Vitaly Shmatikov, CoRR, 2007

[7] Tor : The Second-Generation Onion Router, Syverson et. al., 13ème USENIX Security Symposium, 2004

[8] A fully homomorphic encryption scheme, Craig Gentry, thèse de doctorat, université de Stanford, 2009

[9] Authenticating Location-based Services without Compromising Location Privacy, Haibo Hu et. al., ACM SIGMOD 2012

[10] Brevet WO2018057302A1

[11] Brevet WO2017066509A1

[12] Protecting privacy when disclosing information : k-anonymity and its enforcement through generalization and suppression, Pierangela Samarati, Latanya Sweeney, Technical Report, Harvard Data Privacy Lab, 1998

[13] Brevet CN104080081A

[14] Brevet US8856939B2

[15] Brevet US9736685B2

[16] Brevet US20180091942A1

[17] Differential privacy, Cynthia Dwork, ICALP 2006

[18] Geo-Indistinguishability : Differential Privacy for Location-Based Systems, Miguel Andrés et. al., Proceedings of the 2013 ACM SIGSAC conference on Computer & Communications Security

[19] Brevet US20130145473A1

[20] Brevet WO2013182639A1

[21] Brevet US20150079932A1

[22] Brevet WO2014110647A1

[23] Brevet WO2012087296A1

[24] Authenticating query results in edge computing, HweeHwa Pang et Kian-Lee Tan, ICDE 2014

[25] Aggregate and verifiability encrypted signatures from bilinear maps, Dan

Boneh et. al., EUROCRYPT 2003
[26] Authenticating Location-based Services without Compromising Location Privacy, Haibo Hu et. al., ACM SIGMOD 2012
[27] A fully homomorphic encryption scheme, Craig Gentry, thèse de doctorat, université de Stanford, 2009
[28] Public-Key Cryptosystems Based on Composite Degree Residuosity Classes, Pascal Paillier, ACM EUROCRYPT 1999
[29] Tor : The Second-Generation Onion Router, Syverson et. al., 13ème USENIX Security Symposium, 2004
[30] Privacy and Accountability for Location-based Aggregate Statistics, Raluca Ada Popa et. al., ACM CCS 2011