Génération de traces cellulaires réalistes

Anne Josiane Kouam, Aline Carneiro Viana, Aurélien Garivier, Alain Tchana

Génération de traces cellulaires réalistes

Anne J. Kouam[1,2] et Aline C. Viana[1] et Aurélien Garivier[3] et Alain Tchana[3]

[1]*INRIA Saclay, France* [2] *Ecole Polytechnique, France* [3] *ENS de Lyon, France*

Du fait de la confidentialité des données que contiennent les traces cellulaires CDRs (Call Data Records), leur obtention est soumise à des lois strictes qui limitent leur accessibilité aux chercheurs, et leur utilisabilité même lorsqu'ils sont disponibles. Pour pallier à ce manque, dans cet article, nous nous intéressons à la production de traces CDRs réalistes sur la base d'un ensemble de caractérisations de CDRs réels obtenus auprès d'un opérateur.

Mots-clefs : CDR réalistes, Réseaux cellulaires, Simulation, Préservation de la vie privée

1 Introduction

Un nombre considérable d'appareils mobiles connectés interagissent aujourd'hui avec l'infrastructure des réseaux cellulaires au travers de services divers ; principalement des données mobiles, des appels ou des SMS. Les événements géoréférencés associés sont enregistrés par les opérateurs, constituant ainsi des jeux de données appelés CDRs. Les CDRs sont une riche source de connaissances utile pour de nombreux communautés de recherche et cas d'utilisation tels que la mobilité humaine ou l'étude des performances réseaux. Cependant, ils contiennent des informations sur les déplacements, les réseaux de contacts et les habitudes des individus, ce qui relève de la vie privée. Leur obtention se fait donc exclusivement par le biais de partenariats avec les opérateurs et est soumise à des lois strictes qui limitent leur accessibilité aux chercheurs. En outre, même lorsqu'ils sont disponibles, les CDRs sont anonymisés ou agrégés, ce qui crée de nouveaux défis pour l'exploitation et limite leur utilisation à plein potentiel. Pour combler ce manque, nous présentons dans ce papier le premier générateur de CDRs complets (incluant des attributs de mobilité et de trafic, i.e., données et appels) et réalistes basé sur une modélisation de CDRs réels. Nous concevons pour cela un simulateur de réseaux cellulaires pour assurer la complétude et la flexibilité des jeux de données générés. Les CDRs réels sur lesquels nous nous basons ne contiennent pas d'information de localisation des utilisateurs, mais uniquement du trafic (données et appels). Ainsi nous produisons premièrement des traces de mobilité par une adaptation d'un modèle de mobilité de la littérature (Cf. Sec. 2). Ensuite, nous utilisons les CDRs réels pour créer un modèle de reproduction du trafic (Cf. Sec. 3). Enfin, la Section 4 présente l'algorithme de la simulation qui combine ces deux modèles, ainsi que les résultats préliminaires.

2 Génération de la mobilité

Nous produisons des traces de mobilité par une adaptation du modèle Working Day Mobility Model (WDM) [EKKO08], qui est calqué sur la mobilité humaine réelle. WDM modélise les déplacements des individus pendant les jours ouvrables de la semaine. Il combine trois sous-modèles représentant les principales phases d'activités au cours d'une journée de travail : à la maison, au travail, et après le travail ; ainsi que les déplacements entre ces sous-modèles. WDM est particulièrement adapté à la génération de CDR car il capture la régularité de la mobilité humaine dans le temps, les interactions sociales, et considère des déplacements réalistes sur la base d'une carte géographique réelle. Nous apportons des améliorations à WDM, pour avoir des traces de mobilité plus réalistes, sur la base de charactérisations de traces de mobilité réelles, pouvant être adaptées en configuration pour des zones spécifiques de simulation :

Profilage de l'exploration : Dans une simulation WDM, les nœuds décident de sortir pour une activité du soir ou de rentrer directement chez eux, après les heures de travail, en fonction d'une probabilité p_{ex} commune à un groupe de nœuds de taille configurable. Pour définir cette taille, nous nous appuyons sur le profilage du phénomène d'exploration réalisé dans [AVCL20], analysant des traces de mobilité réelle.

Ce travail définit trois profils d'individus pour l'exploration : Les *scouters* qui sont plus enclins à explorer et découvrir de nouvelles zones, Les *routiners* qui explorent rarement et restent dans leurs lieux familiers peu nombreux, et les *regulars* au comportement intermédiaire, qui alternent entre exploration et routine. En appliquant ce profilage sur le jeu de données CDR ChineseDB [AVCL20] incluant 642K utilisateurs, nous avons obtenu 20.27% de *scouters*, 54.75% de *regulars*, et 24.98% de *routiners*. Par conséquent, nous avons configuré WDM pour diviser la population en trois groupes de tailles pondérées par ces pourcentages respectifs. Ainsi, p_{ex} est fixé à 0.8 pour les *scouters*, 0.5 pour les *regulars*, et 0.2 pour les *routiners*.

Clustering et popularité : Nous modifions WDM pour considérer chaque coordonnée de lieu (maison, travail, lieu d'activité) comme le centre d'un cluster de forme rectangulaire et de taille configurable. Un nœud se voit d'abord attribuer un cluster domicile/bureau, puis choisit son emplacement exact au hasard à l'intérieur du cluster. De plus, nous avons ajouté la notion de popularité de clusters, qui définit la probabilité pour un nœud de choisir un cluster donné comme cluster domicile/bureau ou, dans le cas des points de rencontre, la probabilité de se rendre à cet endroit pour son activité du soir. Cette modification est essentielle ici car détermine la couverture spatio-temporelle des emplacements des individus tout en permettant de retrouver les zones à forte densité d'habitation (e.g., les zones résidentielles), les zones à forte densité commerciale (e.g., quartiers d'affaires) et les lieux de loisirs populaires ou les points d'intérêt (POI).

Profilage de la distance : WDM permet de définir des communes ou des quartiers artificiels dans une ville pour reproduire le monde réel, comme l'illustre la Fig. 1a. Pour simuler de tels cas, il faut définir des groupes de nœuds se déplaçant dans des zones uniques (c'est-à-dire A, B, C ou D), des zones conjointes (par exemple, A et B, A et C) ou dans la carte entière. Pour déterminer la taille de ces groupes, nous avons caractérisé la distance maximale parcourue par les nœuds, ce qui permet de déduire la couverture spatiale des déplacements des individus. Nous utilisons le jeu de données Geolife[ZXM10], qui décrit la mobilité de 182 utilisateurs pendant 64 mois. Ainsi, nous avons choisi de créer trois profils d'utilisateurs pour représenter les déplacements dans une seule zone (*profil 1 : 72%*), dans deux zones (*profil 2 : 19%*), ou sur toute la carte (*profil 3 : 9%*), par souci de simplicité.

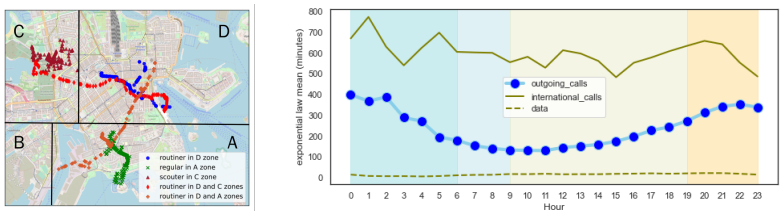
La simulation enregistre dans un fichier la position (latitude, longitude) de tous les utilisateurs avec une résolution temporelle que nous avons fixée à 15 min. Nous traitons ensuite les traces de sortie en *(1)* ramenant la position de chaque trace à la résolution de l'id de cellule correspondant, en *(2)* les filtrant pour ne garder que les traces correspondant aux déplacements des utilisateurs, et en *(3)* y ajoutant un horodatage pour chaque trace. Nous obtenons une trace de mobilité au format *Timestamp, userId, cellId*. La Fig. 1a illustre les traces de mobilité produites dans la région d'Helsinki. Nous pouvons distinguer cinq silhouettes de mobilité d'individus en fonction des profils susmentionnés. Nous remarquons que la mobilité des *scouters* est beaucoup moins structurée que celle des *regulars*, et des *routiners*.

3 Modélisation du trafic

Nous effectuons trois catégories de modélisation pour reproduire respectivement (1) le timing de génération des évènements, (2) la structure sociale des CDRs, et (3) les paramètres associés à la génération d'évènements (e.g., durée des appels). Par la suite nous montrons comment combiner l'ensemble de ces modèles pour la simulation des CDRs (Cf. Sec. 4). Comme base exploratoire, nous utilisons des CDRs réels qui contiennent environ 3 millions de lignes générées par 186,738 utilisateurs sur une durée d'un mois.

Timing de génération d'évènements Nous décrivons ici la modélisation de 3 types d'évènements : les appels locaux émis, les appels internationaux émis et reçus, et les données. Les appels locaux reçus ne sont pas modélisés car étant induits par les appels locaux émis. Nous adoptons une régression non paramétrique pour caractériser le temps d'inter-arrivée (IAT) pour chacun de ces types d'évènement. Avec U l'ensemble des utilisateurs de la trace réelle et $T_{i,u}$ l'instant auquel l'utilisateur u génère son i-ième évènement, nous extrayons comme **modèle 1** pour chaque type d'évènement $\{\lambda_h,\ \forall\ heure\ h \in \{0, ..., 23\}\}$ tel que $IAT_h \sim Exp(\lambda_h)$ avec $IAT_h = \{\forall u \in U, T_{i+1,u} - T_{i,u} \setminus heure(T_{i,u}) = h\}$. La Fig. 1b illustre la régression obtenue.

Par ailleurs, nous subdivisons l'échelle de temps en 4 intervalles ($Inter = \{[0h - 6h], [6h - 9h], [9h - 19h], [19h, 23h]\}$) pour capturer les dynamiques de la génération d'évènements en fonction du moment de la journée. Ainsi, pour chaque évènement nous calculons en tant que **modèle 2** la matrice de probabilité conditionnelle $P[T_{i+1,u} \in B \setminus T_{i,u} \in A], \forall\ A, B \in Inter$. Ainsi $\forall\ T_i / heure(T_i) = h_{T_i} \in A$ le **modèle 2**



(A) **(B)**

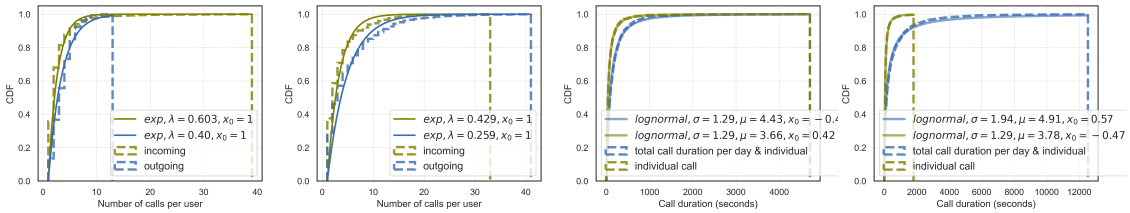
FIGURE 1 : (A) Mobilité simulée dans Helsinki (mieux en couleur) (B) Régression non paramétrique de l'IAT

est premièrement utilisé pour fixer à priori l'intervalle $B/T_{i+1} \in B$. A partir de cela, on pourra déduire les valeurs IAT_{min} et IAT_{max} de $IAT_{h_{T_i}}$, et utiliser le **modèle 1** pour faire un échantillonage de l'exponentielle tronquée $Y_{h_{T_i},min,max} = IAT_{h_{T_i}}/IAT_{h_{T_i}} \in [IAT_{min}, IAT_{max}]$.

Structure sociale des CDRs La structure sociale renvoie à l'architecture de graphe formé par les interactions entre les utilisateurs au travers des évènements d'appels. Reproduire cette structure revient à répondre aux questions : *(Q1) combien de contacts possède chaque utilisateur ?, (Q2) comment choisir ces contacts ?, et (Q3) comment un utilisateur interagit-il avec l'ensemble de ses contacts ?*. Nous repondons à ces questions par un ensemble de modèles non paramétriques de ces processus dans les CDRs réels.
Ici, $\#c_u$ renvoie au nombre de contacts d'un utilisateur $u \in U$. Pour *(Q1)*, nous exportons comme **modèle 3** $P(\#c_u = \#c) \quad \forall\, \#c \in [1, max]$. Par ailleurs, nous définissons 4 catégories disjointes de contacts : les contacts internationaux uniquement (c_{inter}), les contacts locaux émetteurs uniquement (c_{out}), les contacts locaux récepteurs uniquement (c_{in}) et les contacts locaux émetteurs et récepteurs (c_{both}). $\forall\, u \in U\ \#c_u = \#c_{inter,u} + \#c_{out,u} + \#c_{in,u} + \#c_{both,u} = (x_{inter,u} + x_{out,u} + x_{in,u} + x_{both,u}) \times \#c_u$. Nous exportons comme **modèle 4** les valeurs moyennes $\overline{x_{cat,u}}\ \forall\ cat \in \{_{inter}, _{out}, _{in}, _{both}\},\ u \in U$. Ces deux modèles permettent d'attribuer, à tout utilisateur du CDR généré, un nombre total de contacts, réparti par catégories.
Pour *(Q2)*, nous implémentons une variante du modèle de configuration [†] qui permet de construire un graphe à partir d'un ensemble de valeurs connues de degrés des noeuds. Nous ajoutons en plus une heuristique pour choisir les contacts en fonction de la relation (voisins, collègues ou amis) entre les noeuds. Pour obtenir les noeuds ayant de telles relations nous analysons la trace de mobilité synthétique produite (Cf. Sec. 2). Les noeuds dans le même cluster maison/travail entre [01h-04h], [10h-14h] respectivement, sur toute la durée de la trace sont considérés comme voisins et collègues respectivement. De même, les noeuds du même groupe se réunissant pour les activités du soir, quand elles ont lieu, sont considérés comme amis. Ainsi, pour définir les contacts d'un noeud on choisira avec une plus grande probabilité (définie en configuration), ses voisins, collègues, et amis, et ensuite les autres contacts au hasard jusqu'à atteindre le nombre fixé de contacts.
Enfin pour *(Q3)*, nous considérons individuellement chacun des évènements e impliquant un échange avec un contact : Il s'agit des appels locaux émis, des appels internationaux émis et des appels internationaux reçus. $\forall e$, nous ordonnons l'ensemble des contacts de chaque utilisateur par ordre décroissant du nombre d'évènements e fait avec chaque contact ($\#e_{u,c}$) sur toute la durée de la trace. $\forall u \in U, e,\ c_u = (c_1, c_2, .., c_k, .., c_{\#c_u} \setminus \#e_{u,c_1} \geq \#e_{u,c_2} \geq .. \geq \#e_{u,c_k})$. Nous définissons par $P_{k,u,e}$ la proportion du nombre d'évènements e fait par l'utilisateur u à son contact c_k sur le nombre total d'évènements e qu'il a réalisé sur toute la durée de la trace : $P_{k,u,e} = \#e_{u,c_k}/\sum_k \#e_{u,c_k}$. Nous en tirons comme **modèle 5** pour chaque évènement e l'ensemble des moyennes $\{\forall k \in \{1, .., max(\#c_u)\},\ \overline{P_{k,e}} = \overline{P_{k,u,e}}\ \forall u \in U\}$. Ainsi pour chaque utilisateur de la trace synthétique u' avec $\#c_{u'}$ contacts, nous ordonnons au hasard les contacts de u' pour chaque évènement e, et ensuite considérons les proportions $(\overline{P_{1,e}}, \overline{P_{2,e}}, .., \overline{P_{\#c_{u'},e}})$ tirées du **modèle 5** que nous normalisons. Ces proportions représentent la probabilité de choisir un contact pour réaliser un évènement e durant la simulation.

Paramètres associés à la génération Les principaux paramètres associés à la génération des évènements sont : la durée de chaque évènement appel et la taille de chaque évènement données. Dans ce travail nous

†. https://en.wikipedia.org/wiki/Configuration_model



(A) Nombre d'appel généré **(B)** Nombre d'appel réel **(C)** Durée d'appel généré **(D)** Durée d'appel réel

FIGURE 2 : (A) et (B) Distributions exponentielles des CDFs du nombre d'appel par utilisateur. (C) et (D) Distributions lognormales des CDFs de la durée d'appel.

considérons uniquement le premier, du fait de l'absence du dernier dans le CDR réel que nous utilisons comme base. Pour ce faire, nous réalisons un test statistique de distributions continues pour modéliser la distribution de la durée d'un appel dans le CDR réel. Il en ressort comme **modèle 6**, la distribution de la durée des appels en secondes, $D \sim LogNormal(\mu = 89.89, \sigma^2 = (187.72)^2)$.

4 Simulation et résultats

Les modèles liés à la structure sociale des CDRs définissent pour chaque utilisateur, le nombre par catégorie de ses contacts, avant la simulation. Chaque utilisateur u a une horloge de scheduling d'évènements dirigée par les **modèles 1** et **2**. A la fin d'un évènement de type e, on schedulera le prochain évènement e associé à cet utilisateur et ce jusqu'à la fin de la simulation. Pour les évènements du type appel, le scheduling détermine en plus du temps, le contact $\#c_{u,i}$ et la durée de l'appel $d \in D$, en utilisant les **modèles 5** et **6**. Les évènements schédulés sont enregistrés dans un calendrier et exécutés de façon chronologique. Après l'exécution d'un évènement, la simulation crée une nouvelle ligne de CDR au format *Timestamp, callerPhoneNumber, eventType, callerCellId, calledPhoneNumber, calledCellId, Duration*, en complétant les champs de mobilité (Cf. Sec 2). Pour les évènements *données*, le champ *Duration* et les champs liés au contact sont vides. À l'exécution d'un appel, on vérifiera premièrement si le contact est disponible. S'il est occupé, un autre scheduling sera fait, sans ajouter une ligne de CDR. La Fig. 2 valide l'exactitude du CDR généré en constatant la correspondance des distributions obtenues de la durée (lognormale) et du nombre d'appel (exponentiel) avec celles du CDR réel. Pour celà, nous générons le trafic de 1000 utilisateurs sur 5 jours, du fait de la scalabilité actuellement limitée du simulateur ; et de même, considérons dans le CDR réel le trafic d'un échantillon de 1000 utilisateurs sur la même durée. Notons que l'équivalence des paramètres de ces distributions n'est pas le but visé. Une évaluation plus détaillée fait l'objet d'un travail en cours.

5 Conclusion

Dans cette étude, nous proposons une approche de production de CDRs par génération réaliste de mobilité et corrélation avec un trafic obtenu au moyen de modèles de reproduction de paramètres clés. Les résultats préliminaires montrent une correspondance des paramètres du CDR généré, aux mêmes distributions que ceux du CDR réel. Ce travail adresse les limitations dues à l'accessibilité et l'incomplétude des CDRs. Le simulateur résultant sera, une fois achevé, rendu disponible pour une large utilisation.

Références

[AVCL20] L. Amichi, A. Carneiro Viana, M. Crovella, and A. Loureiro. Understanding individuals' proclivity for novelty seeking. SIGSPATIAL '20. ACM, 2020.

[EKKO08] F. Ekman, 1. Keränen, J. Karvo, and J. Ott. Working day movement model. In *Proceedings of the 1st ACM SIGMOBILE Workshop on Mobility Models*, MobilityModels '08. ACM, 2008.

[ZXM10] Yu Zheng, Xing Xie, and Wei-Ying Ma. Geolife : A collaborative social networking service among user, location and trajectory. *IEEE Data(base) Engineering Bulletin*, June 2010.