Group Mobility Detection and User Connectivity Models for Evaluation of Mobile Network Functions

Masaki Suzuki, *Member, IEEE,* Takeshi Kitahara, Shigehiro Ano, and Masato Tsuru, *Member, IEEE*

***Abstract*—Group mobility in mobile networks is responsible for dynamic changes in user accesses to base stations, which eventually lead to degradation of network quality of service (QoS). In particular, the rapid movement of a dense group of users intensively accessing the network, such as passengers on a train passing through a densely populated area, significantly affects the perceived network QoS. For better design and operation of mobile network facilities and functions in response to this issue, monitoring group mobility and modeling the access patterns in group mobility scenarios are essential. In this paper, we focus on fast and dense group mobility and mobile network signaling data (*control-plane data*), which contains information related to mobility and connectivity. Firstly, we develop a lightweight method of group mobility detection to extract train passengers from all users' signaling data without relying on precise location information about users, e.g., based on GPS. Secondly, based on the same signaling data and the results obtained by the detection method, we build connected/idle duration models for train users and non-train users. Finally, we leverage these models in mobile network simulations to assess the effectiveness of a dynamic base station switching/orientation scheme to mitigate QoS degradation with low power consumption in a group mobility scenario. The obtained models reveal that train users consume 3.5 times more resources than non-train users, which proves that group mobility has a significant effect on mobile networks. The simulation results show that the dynamic scheme of base station improves users' perceived throughput, latency and jitter with small amount of additional power consumption in case of a moderate number of train users, but its ineffectiveness with larger number of train users is also shown. This would suggest that group mobility detection and the obtained connection/idle duration models based solely on control-plane data analytics are usable and useful for the development of mobility-aware functions in base stations.**

***Index Terms*—Mobile network, Mobile data analysis, Group mobility, Utilization model, Mobile network simulation.**

I. Introduction

Mobile networks are expected to provide an ever-increasing number of users with satisfactory quality of service (QoS) even in challenging environments. With meticulous analyses of growth trends and thorough planning, networks can adapt to variable and heterogeneous needs. However, as current mobile networks remain considerably static due to the high cost of temporary reshaping, they are still enable to adapt well to changes in demand that are not related to global trends but are instead dynamically caused.In general, sudden changes in demand are especially difficult to predict and handle efficiently, and thus result in either QoS degradation or extra expenditure. If base stations are designed to handle peak levels of changing demand, a large proportion of the resources goes unused during normal demand periods. Otherwise, users perceive degraded QoS during the peak demand period. A typical and important example causing sudden changes in demand is the movement of dense groups of users intensively accessing the network that requires the reallocation of resources to a large number of devices at the same time. From the mobile network operator's (MNO's) point of view, a new solution should be developed for QoS degradation due to group mobility that emphasizes cost and energy efficiency.

As a solution for the above-mentioned issues, in this paper, we consider a dynamic base station switching/orientation scheme. The assessment of such a new scheme involves the following procedures. Firstly, group user mobility should be modeled based on monitoring, which includes detecting the movement of groups of users and analyzing/understanding the users' utilization of the network. Then, using the results of the monitoring and modeling, an initial evaluation of the scheme should be conducted, which is needed prior to scheme deployment in the real network and directly affects the decision on deployment. However, these procedures are challenging due to the following conditions and regulations:

- ***Difficulty of collecting accurate locations of users:*** Detecting a group of users is straightforward if the accurate locations of all the users are available in general. Such information is however both difficult and/or costly to obtain and not desirable in terms of privacy issues. Therefore, the detection of moving groups of users should not require the accurate location of individual users.
- ***Difficulty of using training-based methods:*** As metropolitan areas are in a state of constant evolution with, e.g., the construction of new buildings and the modification of urban facilities, MNOs must frequently adapt their networks by adding base stations and changing their parameters including the orientation of their antennas. In addition, since the radio conditions are very sensitive, the coverage of base stations relies on the weather. If a group mobility detection method relies on some training according to the network conditions, it would need incessant training phases and would not be practical. Therefore, the method should not require any form of training.
- ***Need for frequent updates:*** How the network is used basically depends on how customers use their devices. Thus, network utilization/connectivity patterns are likely

M. Suzuki and T. Kitahara are with KDDI Research, Inc., Saitama 356-8502, Japan (E-mail: masaki-suzuki@kddi-research.jp; kitahara@kddi-research.jp).

S. Ano is with KDDI Corporation, Tokyo 102-8460, Japan (E-mail: sh-ano@kddi.com).

M. Tsuru is with Kyushu Institute of Technology, Fukuoka 802-8502, Japan (E-mail: tsuru@cse.kyutech.ac.jp)

to change with time as new devices and applications are released, so any analyses should be updated on a weekly or daily basis. Therefore, the detection method and consequent analyses should be lightweight.

- ***Difficulty of collecting and handling user data:*** All IP traffic in LTE networks is separated into user data and control data. User data contains application data and control data contains messages for handling connections for user data. Network utilization/connectivity is often analyzed based on user data. However, user-data-based analyses generally require a large amount of processing power and time. Moreover, user data is not desirable for privacy reasons. Therefore, the analyses should function without user data.
- ***Difficulty of on-field experiments:*** An on-field experiment with active performance test will require a number of testers who take trains carrying their own devices. While the public transportation service covers a variety of locations with different environmental conditions in a metropolitan area, such a costly experiment is difficult to conduct on all locations. Furthermore, the experiments should be repeated due to frequent changes of the environmental conditions as noted above. Therefore, the process should be based on passively collected data in a lightweight manner instead of actively collected ones in on-field experiments.

In this paper, we discuss a general process consisting of the following three tasks to evaluate a resource allocation scheme relating to group mobility in mobile networks by taking into account the above-mentioned challenging requirements. Firstly, we develop a group mobility detection method using signaling data in an actual mobile network, which leverages the spatial and temporal locality dimensions passively monitored in the network. The location accuracy of the method only relies on the locations of base stations that are naturally figured out by MNOs. The method is not affected by changes in environmental conditions or the configuration of base stations. The method can also be validated by consistency with the ground truth of actual train timetables and train track locations. Secondly, we construct connectivity models expressed by connected/idle durations. The connected/idle durations for train users and non-train users are extracted from the same signaling data used in the detection method, and are used to build and validate the models for this task. We assume that network accesses are triggered by either human activities or background applications activities, and then describe the models with a mixture of two different distributions. The models are validated in a way similar to the leave-one-out cross-validation technique. Finally, we leverage the models to conduct an initial evaluation of newly introduced functions. We assess the impact and effectiveness of a dynamic base station switching/orientation scheme in group mobility scenarios by network simulation. The scheme is evaluated in terms of the user-perceived QoS and additional power consumption. Since the connectivity models used in the simulation are built based on the data in a specific area and period of time, we vary the parameters of the models in order to simulate a range of group mobility with different conditions. We can obtain a relative evaluation, e.g., results revealing the sensitivity of the scheme to the model parameters.

In the above-mentioned process, the first and second tasks based on our previous conference paper [1] and the third task are integrated to demonstrate the availability and usefulness of the detection method and the models for group mobility. The main contribution of this paper is three-fold in the application of Big Data analytics:

- We develop and validate a detection method to monitor group mobility, which achieves decent performance with a precision of 0.70 and a recall of 0.75.
- We build and validate connected/idle duration models for both train users and non-train users, which reveal that train users consume 3.5 times more resources than the others.
- Through simulation based on the obtained models, we evaluate a dynamic base station switching/orientation scheme that dynamically adjusts base stations' capability in synchronization with the group mobility of users. The simulation results suggest that the new scheme is feasible in real-world commercial environments. It can improve users' perceived throughput, latency and jitter, which are 962%, 18% and 32% compared to those without the additional base stations implementing the scheme, with small amount of additional power consumption of 2.0%.

The rest of this paper is organized as follows. The position of this paper is clarified in Section II. The group mobility detection method is proposed and its performance is evaluated in Section III. In Section IV, network utilization by train users and non-train users are modeled and the proposed models are validated. The availability and usefulness of the models are demonstrated by being applied to a practical use-case of the initial evaluation phase of mobile network functions in Section V. Section VI finally concludes this paper.

II. Related works

Our work deals with mobile users communications in a metropolitan area with trains from MNO's view point, and consists of (i) detecting the group mobility of train users; (ii) building users' connectivity model using the results of (i); and (iii) assessing some new base station functions through simulations driven by the models obtained in (ii). As described in Section I, they face the five types of difficulties from practical conditions and regulations. In this section, related works are reviewed for each technical component, i.e., (i), (ii), and (iii), to show the appropriateness of our design choices.

A. User mobility extraction

User mobility has been extensively researched originally in the ad-hoc network field [2], [3]. [2] introduced typical mobility patterns and evaluated their effects on the performance of routing protocols. [3] proposed a routing-based location management scheme for wireless mesh networks. They mathematically described clients mobility using stochastic Petri net techniques, then they evaluated their proposed method. Those

mobility models have been improved and applied to cellular networks for cell planning [4], and for mobility-aware functions [5], [6]. However, while those characterizations of mobility are useful for creating sufficiently realistic simulations, they are not necessarily useful for monitoring or detecting the movement of groups of users in a real network in an efficient manner and for subsequently constructing connectivity models of users according to their fashion of locomotion.

Independently of ad-hoc or cellular networks, the nature of human mobility itself has been studied in [7]. [7] reported that human mobility displays similar features to Levy-walks, including heavy-tail flight and pause-time distributions and the super-diffusion followed by subdiffusion. Roughly speaking, human mobility is likely to be directed random walk rather than pure random walk. This report, however, does not include group mobility, and our objective is to label users with their way of movement.

Great efforts have been made to characterize and understand human mobility based on actual call detail records (CDR) or base station logs [8]–[12]. CDR are data records of communication i.e., phone call, Internet connection, short messaging service, etc., which contain subscriber IDs of endpoints of communication, the start time, the duration of communication, etc.. Since CDR contain the privacy sensitive information such as subscriber ID, the data should be carefully treated. Base station logs are data records that contain the user device IDs and the times that they connect to the base station. It sometimes contains the radio resource control logs. In [8], by monitoring the distribution of CDR during business hours, they estimated the mobility of commuters that travel daily across the geographical gap between users' residences indicated by billing ZIP and their places of employment. They further estimated the carbon emissions and traffic volume as in transportation. [9] proposed a route classification method and showed the practical capability of deriving the trajectories using handoff patterns of CDR: not only identifying the train lines but also detecting the train tracks matching cellular handoff patterns to routes. According to their evaluation, the handoff patterns are robust and distinguishable even if small changes occur in route, speed, direction, phone model, and weather conditions. [10] proposed a human travel estimation method to specify the means of movement and the destination. To finely estimate the travel using the coarse location information of base station logs, they made use of supplementary information. An efficient pattern creation method was proposed in [11]. [11] classified the ways of movement of train users in relation to train lines, which requires training in the beginning of the classification. However, those methods are not easily applied to our analysis. This is because [8]–[11] require supplementary information such as map information, etc. In addition, [9] and [11] require a training phase with considerable overhead in the initial part of the detection process.

The concept of group mobility has been especially highlighted since group mobility is likely to cause inefficiency of resource utilization and bursts of demand [13]–[16]. [13] introduced a variety of group movement models and evaluated the effect of types of group mobility on connectivity and throughput of an ad-hoc network. [14] introduced an aggregation and separation movement model for groups of people. [15] proposed very microscopic group mobility level classification and group structure recognition for pedestrians, using sensors on the pedestrians' smartphones. [16] took into account spatial and temporal locality dimensions using sensor data to quantify the correlation between the mobility of users. Another viewpoint regards group movement as one of many group events. Group event detection has much progressed in the participatory sensing field [17], [18]. [17] proposed a detection method for the boundaries of concurrent events to efficiently improve the quality of information and to appropriately define the incentive for sensing. The method uses a combination of coarse- and fine-grained boundary detection but it requires accurate location information as observed by devices. [18] proposed the combination of two centralized event detection algorithms that make use of the Min-cut theory and SVM pattern matching technique. The intuition of our approach is similar to that of [18], which is "real-world events usually exhibit some spatiotemporal patterns." However, those methods are not easily applied to our analysis. This is because the characterizations of group mobility in [13], [14] are useful for creating simulation, but are not necessarily useful for monitoring or detecting the group movement of users in real network. [15]–[18] require accurate and fine-grained location information of users but it is difficult to collect and use sensor data observed on user devices. Also, while [18] intended to deal with general events, we focus specifically on group movement. In order to eliminate the effects of behavior other than group movement, we directly make use of characteristics of group movement.

The core technique of our group movement detection is clustering. Many clustering algorithms have been developed specially for data streams where the processing speed and efficiency should be high [19]–[21]. Data streams are characterized as data that is generated continuously and simultaneously by millions of devices, e.g., phone call records, and the data is usually massive amount. [19] proposed CLARA, which is based on a partitioning around medoids algorithm. In the initial part of the process, CLARA randomly chooses data samples to calculate medoid candidates then repeatedly refines the accuracy of those medoids. [20] proposed the BIRCH algorithm, which builds a hierarchical data structure (CF-tree) to compress the amount of data and cluster leaves with a k-means clustering algorithm. [21] proposed the STREAM algorithm, which splits the data into chunks then locally clusters the data in each chunk and extracts the center of each cluster in a rapid way (LSEARCH). After the local clustering, it globally clusters all centers using a k-means algorithm. These algorithms have a great advantage in processing speed but are not sufficiently flexible for our purpose; the clustering should provide a corresponding relationship of each resulting cluster to a way of movement. In addition, they require to estimate an appropriate number of medoids in advance. However such an estimation is difficult since we do not know accurate train movement schedule all the time. It is also too time- and resource-consuming to introduce the appropriate number of medoids with an optimization function such as the Silhouette value [22].

B. Network connectivity modeling

For users' behavior analysis, user-plane data is often used. User-plane data is any data directly generated by user application, i.e., audio, video, text and so. Although it can be used to analyze the traffic volume, the time duration, and the transferred contents of each user activity, it often requires a large amount of processing power and time. Moreover, since user-plane data usually contains privacy sensitive data, it is necessary to obtain users' permission for analysis. A number of studies for user connectivity modeling are found in literature [23]–[25]. [23] analyzed traffic patterns using cellular base station trace data. They revealed the traffic patterns depending on cell-towers, the geographical context of traffic patterns. Then, they modeled the traffic patterns in terms of their time domain and frequency domain aspects. [24] analyzed user data on the commercial 3G network and defined categories of machine-to-machine devices. They investigated daily and weekly patterns of the traffic volume, frequency of data generation, geographical distribution, comparison with smartphones, application usage, and network performance. [25] deeply investigated the events which put a temporally but extremely high demand for communication capacity on cellular networks with commercial voice and data traces. They revealed the performance degradation both pre- and post-connection, then proposed and evaluated an adequate mitigation method.

On the other hand, a number of studies using control-plane data are also found in literature [26]–[28]. Control-plane data, also known as network signaling data, is convenient and useful for MNOs. This is because it consists of very smaller volume compared with user-plane data, is easy to retrieve from a mobile network perspective, and contains information related to the mobility and the connectivity of all mobile devices. Compared with user-plane data, analyses based on control-plane data are expected to be fast and practical. Analyses based on control-plane data have been studied to investigate the interactions of control-plane protocols [26], so as to improve the management of mobility [27] and even to evaluate the effect of a signaling storm on a mobile network [28].

The approach proposed in this paper is based on monitoring and analyzing control-plane data. This is firstly because we cannot analyze user-plane data due to both of computing resource and privacy issues. Secondly, we just focus on users mobility and connectivity and are not interested in user content or very microscopic users' behavior. Furthermore, we uniquely focus on the difference of users' connectivity corresponding to the users' way of movement, which requires sufficiently wide area of data but its analysis should be sufficiently fine-grained. To the best of our knowledge, this is the first challenge to investigate user connectivity models based solely on network signaling data that are captured behind base stations. The proposed network connectivity models for train and -non-train users are explained in detail in Section IV.

C. Network simulation and assessment for network functions

To quantitatively analyze and assess a complicated system in a specific circumstance, simulation-based evaluation is essential in many cases where in-lab experiments are not at scale and real-world experiments are too costly. In particular, it is of practical importance to initially assess some system functions with parameter tuning by simulation before development and deployment of the functions. In general, there are two types of simulation; model-driven simulation that uses simulation-input data generated from models built and tuned for the targeted environment [5], [6], and trace-driven simulation that uses simulation-input data generated from actual data obtained from real experiments [4], [25]. Model-driven simulation can be easily applied to any form of simulation but its adequateness depends on how well the model reflects the reality. While it possibly includes a gap between simulation and real world so that it usually requires on-field experiments afterwards, model-driven simulation is suitable for the initial evaluation with synthetic simulation. [5] evaluated the effects of adaptive virtual network function on the changes of latency with varied speed of user devices. [6] evaluated the accuracy of mobility estimation using linear dynamic system model of mobility.

On the other hand, trace-driven simulation is capable of realistic simulation by conducting the actual users' behavior but is not likely to be flexible to apply to general simulations. Thus, it is difficult to verify the generality of simulation results. [4] computed pseudo-user mobility in order to evaluate cell planning efficiency using a census data for user movement in specific but relatively large area. [25] evaluated the congestion mitigation schemes using the actual radio resource control logs and TCP header in data traffic, which are observed in the specific events, i.e., football games, and public demonstrations.

In this paper, we evaluate the mitigation methods for QoS degradation caused by user mobility through simulation based on train and non-train users connectivity models. To investigate how the mitigation methods generally impact on the fundamental network performance, we conduct a synthetic simulation. Therefore, we cannot directly apply the actual trace-data to simulation and, instead, we apply the connectivity model based on actual signaling data. In order to mitigate QoS degradation caused by mobility or group mobility, a variety of adaptive functions have been proposed [5], [6], [29]–[31]. To enhance the conventional base stations that are designed to have sufficient resources for peak demand periods and are always on even in normal demand periods, [29] proposed a smart scheme to switch on/off the base stations depending on the traffic load in order to reduce extra power consumption. The scheme could reduce the power consumption by 60% on weekdays and by 80% on weekend. [30] introduced the fundamental idea of moving antenna to steer the orientation of antenna targeting a moving user device. [31] further developed the idea of dynamic orientation focusing on a high speed train scenario. It discovered a new type of spatial-temporal correlation between the base station and moving antenna on the roof-top of train. We define dynamic base station functions essentially based on the ideas discussed in [29] and [30] as we call dynamic base station switching/orientation scheme hereafter. We evaluate the fundamental performance metrics with two dynamic base station functions: dynamic switch on when trains come close and switching off when they go, dynamic orientation of directional antenna tracing the group

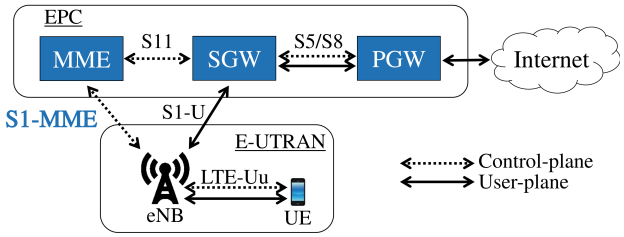


Fig. 1. Simplified LTE network architecture.

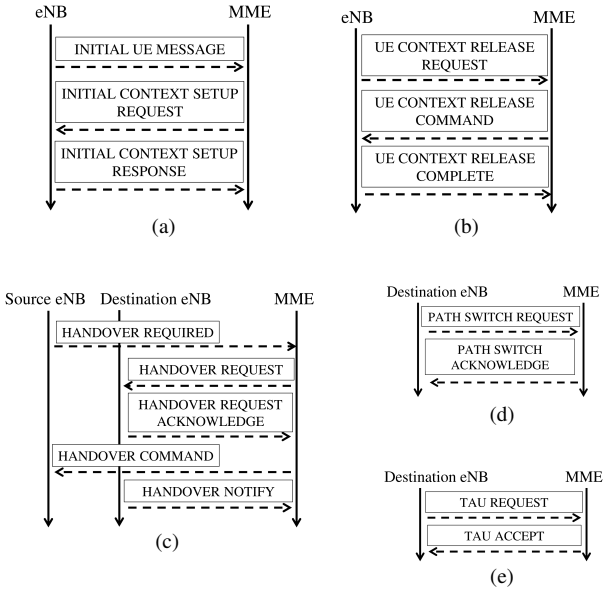


Fig. 2. S1AP signals related to connectivity and mobility. Signals related to connectivity are described in (a) and (b). S1-handover signals, X2-handover signals, and TAU signals are respectively detailed in (c), (d) and (e).

movement in Section V.

III. Signaling-based group mobility detection

We propose a group mobility detection method consisting of three steps: signaling data analysis, handover clustering, and moving group extraction. The core idea of our group mobility detection lies in the fact that users moving together will often change the associated base station at the same time and around the same location. The first step in this method is to detect changes of base station by users, called *handover events*. The second step aims to identify groups of users by clustering these handover events by location and time. The locations here do not correspond to users' locations but to the locations of base stations. Finally, by taking advantages of the time dimension to filter and gather the clusters, we can retrieve the users moving together over time with a high-level of confidence.

A. Signaling data analysis

When devices move or access the network, control-plane signals are triggered aiming at the management of mobility and connectivity. By capturing and analyzing these signals, we can obtain access information about the mobile network and its devices without interfering with the network entities themselves. Our analysis focuses on the control signals carried over the *S1-MME* interface of the LTE architecture, described in Fig. 1. Located between the evolved NodeBs (eNBs), i.e., the base stations of LTE networks, and the mobility management entity (MME), this interface uses the S1 Application Protocol (S1AP) [32] to carry signals related to connectivity and mobility.

Each time an idle device tries to access the network, a sequence of three messages described in Fig. 2(a) is exchanged between its serving eNB and the MME. When a connected device causes use of the network, it returns to an idle state after the exchange of a sequence of three other messages, described in Fig. 2(b). Analyzing these two sequences of messages allows us to detect when a device goes connected or idle, and thus to determine the connectivity state of the device.

In order to improve the strength of the radio signals, a connected device can change the serving eNB by performing a *handover*. These phenomena frequently occur when users are moving and can be split into two categories according to the way in which the process is executed: *X2-handovers*, in which the process is performed almost entirely between the two eNBs, and *S1-handovers*, in which the process is performed by the MME. The mobility of *connected* devices can thus be determined by detecting the handovers. Besides, in order for the mobile network to be able to redirect mobile-terminated calls and packet transfers using paging, the evolved packet core (EPC) must remember the tracking area (TA), i.e., the group of eNBs, containing the eNB last contacted by the device. When a device changes TA, a signaling procedure called a *tracking area update* (TAU) is emitted on the S1-MME interface to notify the EPC, *even when the device is idle*. The messages related to handovers and TAUs that we analyze in our method are detailed in Figs. 2(c), 2(d), and 2(e).

Although enormous, the volume of signaling data is relatively small compared with user-plane data. Our implementation in C/C++ shows that it is possible to finish analyzing the numbers of S1AP packets faster than the rate of data acquisition. We output events related to the connectivity of the devices to queue C_q and events related to mobility to queue M_q. As all mobility events represent a change of base station from source base station s_a to destination base station d_a, all mobility events will now be called handover events. Each handover event will be denoted by $h = (s_a, d_a)$.

B. Handover clustering

The second step of the detection method is responsible for revealing the movement of groups of users by clustering the previously computed handover events by location and time. We define (λ, ϕ) as the location of handover $h = (s_a, d_a)$ as the middle point of s_a and d_a, where λ and ϕ respectively refer to the longitude and latitude of h. (λ, ϕ) is a rough approximation of the location of the device initiating the handover.

At each time $t = k\Delta t$ $(k \in \mathbb{N})$, with Δt being a constant interval in seconds, we remove and retrieve all the handovers currently present in M_q and cluster their locations using a regular grid of interval γ bounded by minimum and maximum

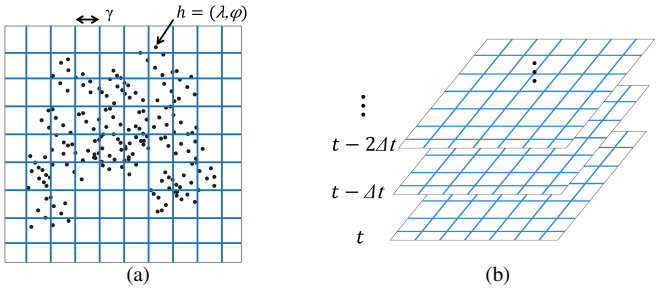


Fig. 3. Handover clustering. The clustering realized at one specific time is illustrated in (a), and (b) represents the sliding window storing the clusters of several times.

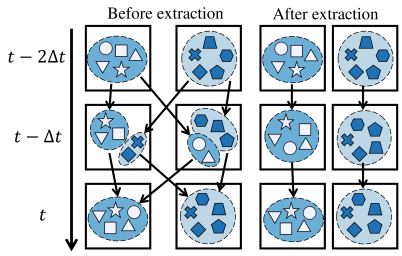


Fig. 4. Group extraction with $P = 3$ and $N = 2$. For instance, device ● is in the same clusters as device ★ $N_{com} = 2$ times and $N_{com} \geqslant N$; therefore, they are considered to be moving together at $t - 2\Delta t$, $t - \Delta t$, and t.

longitudes and latitudes λ_{min}, λ_{max}, ϕ_{min}, and ϕ_{max}. Let $r = \lceil(\phi_{max} - \phi_{min}/\gamma\rceil$, and $c = \lceil(\lambda_{max} - \lambda_{min})/\gamma\rceil$ be the number of rows and columns of the grid and $n = rc$ the number of clusters, respectively. Clusters $C(t) = (C_0, \ldots, C_{n-1})$ computed at time t are defined in Eq. (1).

$$C_i = \left\{ h \in M_q \middle| \left\lfloor \frac{\lambda - \lambda_{min}}{\gamma} \right\rfloor r + \left\lfloor \frac{\phi - \phi_{min}}{\gamma} \right\rfloor = i \right\}. \quad (1)$$

This clustering method is primitive. However, since the locations of handovers are particularly rough, our method can be useful without more sophisticated clustering. It also has the advantage of being linear in the number of handovers, which preserves the processing speed of the method. Figure 3(a) illustrates the clustering realized at time t. A sliding window $W = (C(t+i\Delta t))_{-P<i\leqslant 0}$ of size $P \in \mathbb{N}$ is shown in Fig. 3(b). The values of Δt and γ have a direct impact on the size of the clusters and the frequency at which they are computed, which will be discussed in more detail in Section III-D.

C. Moving group extraction

Let us define the following relation between devices: two devices u and v are *clustered together* at time t *if and only if* they initiate handovers that are located in the same cluster at t. Due to the roughness of the handover locations and the irregular accesses to the network by different devices, two devices actually moving together may not always initiate handovers at the same time and same location and thus may not always be *clustered together*. Besides, two devices crossing paths may at some point initiate handovers at the same location and same time and thus may be wrongly *clustered together*. In the last step of the detection method, we use sliding window W to filter and merge clusters in order to only retrieve groups of devices truly moving together.

We define $G(t) = (G_1, G_2, \ldots)$ as the list of groups of devices truly moving together at time t. We decide that two devices are actually moving together *if and only if* they are *clustered together* at least N times out of P, where $N \in \{1, \ldots, P\}$ is a constant parameter. At time t and for each pair of devices $\{u, v\}$, we retrieve from W the lists of clusters $C_u = (C_{u,t+i\Delta t})$ and $C_v = (C_{v,t+i\Delta t})$ that u and v have been clustered in for $-P < i \leqslant 0$. We then compute $N_{com} = |C_u \cap C_v|$, the number of clusters that are common in C_u and C_v. Let $t_{first} \leqslant t$ and $t_{last} \leqslant t$ be the times of the first and last common clusters, if they exist. If $N_{com} \geqslant N$, we consider that u and v were truly moving together during interval $[t_{first}, t_{last}]$ and thus we put them in the same group of devices for each time $t_g \in [t_{first}, t_{last}]$. This implies that groups $G(t_{first})$ are modified at time $t \geqslant t_{first}$: groups of previous times can be modified as new clusters are computed. This choice is correctly made to gather devices that have just started to move together and cannot otherwise be considered to be moving together. Figure 4 illustrates this step with $P = 3$ and $N = 2$: each geometrical symbol represents the handovers initiated by one device, each large square represents one cluster, and the dashed ellipses show which devices are truly moving together. As ● and ▲ are in the same clusters as ★, ■, and ▼ at times $t - 2\Delta t$, and t, we have $N_{com} = 2$ and as $N_{com} \geqslant N = 2$, they are considered to be moving together at $t - 2\Delta t$, $t - \Delta t$, and t. By repeating the same procedure for all devices, groups of devices moving together are extracted and the result is depicted on the right-hand side of Fig. 4.

The groups of devices extracted at time $t - P\Delta t$ can be modified until t and thus can only be fully determined at t. This induces a delay of $d = P\Delta t$, which is the theoretical limitation of our approach. This delay is irrelevant to our current objective but could be important in other applications. At time t, we add the list of groups of devices $G(t - P\Delta t)$ to queue G_q. At the end of this final step, G_q contains for each time $t = k\Delta t$ ($k \in \mathbb{N}$) list $G(t)$ of groups of devices moving together at t.

D. Evaluation and discussion

As our objective is to detect dense moving groups of users for network performance evaluation, we incorporate our method into an actual LTE network and evaluate its performance by applying it to the detection of train movements. The location of a detected group of devices can be approximated by the average location of the handovers it contains. Comparing the locations of the detected groups with real train locations allows us to evaluate the method. Using train timetables and train track locations, we estimate the locations of real trains at each time t. Let $A(t)$ be the set of real trains present in the considered area at time t and let $(a_{t,1}, \ldots, a_{t,|A(t)|})$ be their actual locations. We filter $G(t)$ to only keep large groups of devices, whose locations are denoted by $(d_{t,1}, \ldots, d_{t,|G(t)|})$. We match $(a_{t,1}, \ldots, a_{t,|A(t)|})$ and $(d_{t,1}, \ldots, d_{t,|G(t)|})$ using the Hungarian algorithm [33] to minimize the squared distance

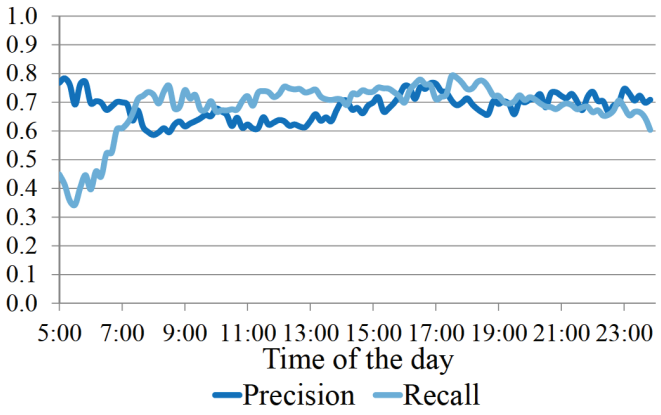


Fig. 5. Precision and recall according to the time of day.

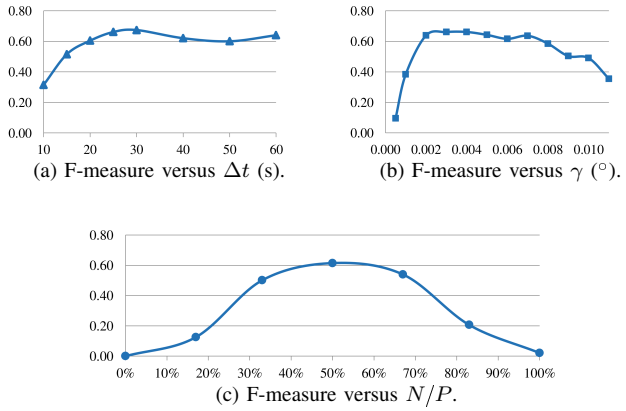


Fig. 6. Influence of the parameters on the performance.

between each matched location. Matched locations represent detected groups that are located near real trains, i.e., *true positives*, and unmatched locations represent either detected groups located far from any real train, i.e., *false positives*, or real trains that have not been detected, i.e., *false negatives*. Let $M(t)$ be the set of matched locations; we can compute the precision, the recall, and F-measure F_1 of the method for each time t as shown in Eqs. (2) and (3).

$$\text{precision}(t) = \frac{|M(t)|}{|G(t)|}, \qquad \text{recall}(t) = \frac{|M(t)|}{|A(t)|}, \tag{2}$$

$$F_1(t) = 2\frac{\text{precision}(t) \times \text{recall}(t)}{\text{precision}(t) + \text{recall}(t)}. \tag{3}$$

This evaluation is realized on the actual anonymized control-plane data corresponding to a period of four weeks in a large metropolitan urban area in Japan. A total of 25 billion packets are analyzed, trains are detected, and their locations are evaluated. The data used were captured at a specific area and period of time but the area and period are sufficiently typical in terms of human activities. Figure 5 shows the average performance of our method according to the time of day with the parameters described in (4). These values were chosen as they gave the best results amongst a wide range of parameters. Except in the early morning, the results show that our method achieves decent performance with a precision of 0.70 and a recall of 0.75. Considering the roughness of the handover locations, this performance is satisfactory and sufficient for further statistical analyses. The drop in recall between 5 a.m. and 7 a.m. is due to the small number of people during that period of the day: fewer people leads to fewer devices clustered together and thus fewer detected groups. The recall increases progressively between 5 a.m. and 7 a.m. as more and more people are present in the area. This issue is irrelevant to our objective as we focus on challenging situations for mobile networks and are only interested in the busiest periods of the day (between 7 a.m. and midnight).

$$\Delta t = 30s, \ \gamma = 0.003°, \ P = 8, \ N = 4. \tag{4}$$

To determine the best parameters for the method, we choose the F-measure as a balanced metric, i.e., the performance, to evaluate both the precision and the recall according to the parameters. Figures 6(a), (b), and (c) respectively illustrate the influence on the performance of Δt, γ and the performance of ratio N/P representing the proportion of time that two devices must be clustered together to be considered to be actually moving together. When Δt is small, devices have less time to move between two clustering steps and they thus trigger fewer handovers, which reduces their chance of being clustered with other devices and leads to a drop in performance. A high value increases the number of handovers and thus the performance but also increases the delay $d = P\Delta t$ of the method. γ directly affects the size of the clusters: too small a value will result in extremely small clusters and fewer devices will be clustered together, whereas too high a value will result in disproportionate clusters including devices that do not move together. Finally, N/P represents the minimum similarity of the movements of two devices moving together. A good balance must be chosen as a low similarity will incorrectly mix together devices moving differently, while a high similarity will only group together devices that always trigger handovers at the same time, leading to great precision but low recall and thus low performance. With these considerations in mind, we evaluated our method for a wide range of parameters and selected the best set of parameters introduced in (4).

As our method relies solely on the base stations contacted by the users, it does not require the accurate locations of the users and it can be generalized to any type of mobile network. Furthermore, it does not necessitate any training phase depending on the base stations and thus works even if the topology of the network is modified. The processing delay $d = P\Delta t = 4$ min induced by the method is insignificant because the following analyses described in Section IV are expected to be updated on a weekly or daily basis. Moreover, the volume of data is large enough to include diverse information and to allow reliable analyses based on the data. Therefore, the proposed detection method fulfills the conditions and regulations described in Section I, and is applicable to the following analyses.

IV. Signaling-based users' connected/idle duration models

We propose users' connected/idle duration models based on signaling data to estimate user-level network utilization by the

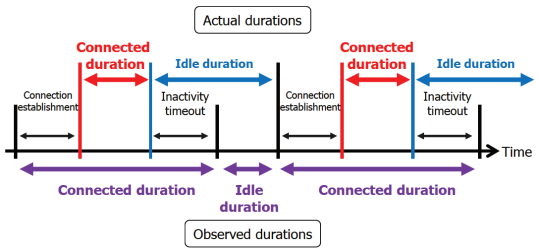


Fig. 7. Relation between observed durations and actual durations.

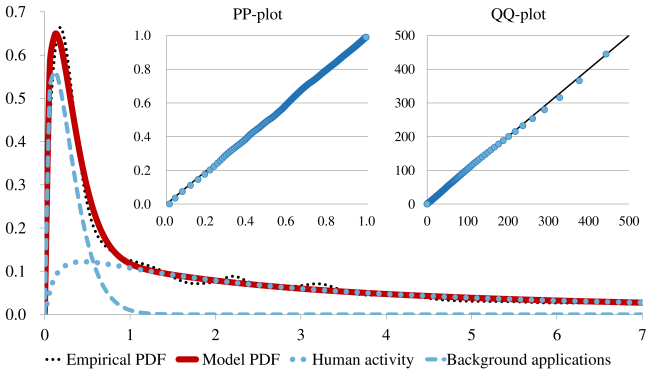


Fig. 8. CNT model: empirical PDF $\tilde{f}_{CNT}$, the two components of the model, and the model PDF f_{CNT} according to the connected duration (in seconds). The PP-plot and QQ-plot evaluating the model are also depicted.

detected moving groups of users. We call users detected by our detection method *train users* and all the other *non-train users*. We then model the time spent using the network by each user.

A. Users' connected/idle duration models

Network utilization time can be estimated by the durations of connected and idle states of each user. We retrieve the connectivity events stored in queue C_q defined in III-A and compute the list of connected/idle durations of each device. Using our group mobility detection method, all users are divided into train users and non-train users. Four different durations are recorded in lists D_{CT}, D_{CNT}, D_{IT}, and D_{INT} containing respectively the durations of *connected train users* (CT), *connected non-train users* (CNT), *idle train users* (IT), and *idle non-train users* (INT). As mobile networks use a timeout to detect the inactivity of users, a device does not actually use the network for the duration of this timeout and network utilization is thus better approximated by subtracting the value of the timeout from the observed connected duration. In addition, a device does not actually use the network while it is establishing a connection. We estimate the connected/idle durations in our model from the observed durations as illustrated in Fig. 7.

The probability density functions (PDFs) of the durations contained in each list are estimated using a kernel density estimation (KDE) [34] with Gaussian kernel $K \sim \exp\left(-x^2/2\right)$ and bandwidth $h = 0.1s$. If $(d_1, \ldots, d_n)$ are n durations, the empirical PDF $\tilde{f}$ of the distribution of their durations is computed as in Eq. (5). Figure 8 illustrates the empirical PDF of the CNT distribution.

$$\tilde{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - d_i}{h}\right). \tag{5}$$

Some network accesses are caused by the diverse communication activities of users, generally with relatively long durations. Some other network accesses are triggered by background applications causing an exchange of small amounts of data in a short duration, e.g., to retrieve the latest news, to check for emails, or to receive notifications from a server. Therefore, we model the CT and CNT distributions using a mixture of one log-normal distribution representing *human activity* and one Weibull distribution representing *background applications*.

Idle durations are directly linked with the inactivity of users but are also affected by the activity of other users. Indeed, if two users A and B are texting each other and user A sends a message to user B before going to idle, user B's answer will activate user A and the idle duration of user A will be closed in a short time. This phenomenon is intensified by the inactivity timeout we previously mentioned: if the inactivity timeout equals 5s and user B takes 6s to reply, user A will stay in idle for 1s only. Therefore, we also model the IT and INT distributions using a mixture of one log-normal distribution for *human inactivity* and one Weibull distribution for idle durations closed by other user activities and applications.

We choose the log-normal distribution that is often used in the description of natural phenomena with a relatively long tail. The Weibull distribution is chosen as it fits peaks well.

Let f_{CT}, f_{CNT}, f_{IT}, and f_{INT} be the PDFs of these models. The common equation for them is described in Eq. (6), where σ and μ are the log-normal parameters, k and λ are the Weibull parameters, and α is a weighting factor between the two components.

$$\begin{aligned} f(x) = \alpha \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right) \\ + (1 - \alpha)\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{x}{\lambda}\right)^k\right). \end{aligned} \tag{6}$$

Let F be the cumulative distribution function (CDF) of f and $\tilde{F}$ be the CDF of empirical PDF $\tilde{f}$. The five parameters of each model are determined by maximizing the objective function described in Eq. (7), where $X = (\sigma, \mu, k, \lambda, \alpha)$ is the vector of the parameters, and S is a dataset of samples of the duration. ${R_P}^2$, ${R_C}^2$, and ${R_Q}^2$ are the coefficients of determination respectively indicating: how well PDF f fits PDF $\tilde{f}$, how well CDF F fits CDF $\tilde{F}$ (PP-plot), and how well the quantiles of f correspond to the quantiles of $\tilde{f}$ (QQ-plot). This process is chosen for the model PDF with parameters X to fit the shape of the empirical PDF derived from samples S well, quantified by the coefficient of determination ${R_P}^2$ of two PDFs, while conserving the same statistical properties as the data, quantified by ${R_C}^2$ and ${R_Q}^2$. As the PP-plot (quantified by the coefficient of determination ${R_C}^2$ of two CDFs) magnifies errors in the center of the distribution and the QQ-plot (quantified by the coefficient of determination ${R_Q}^2$

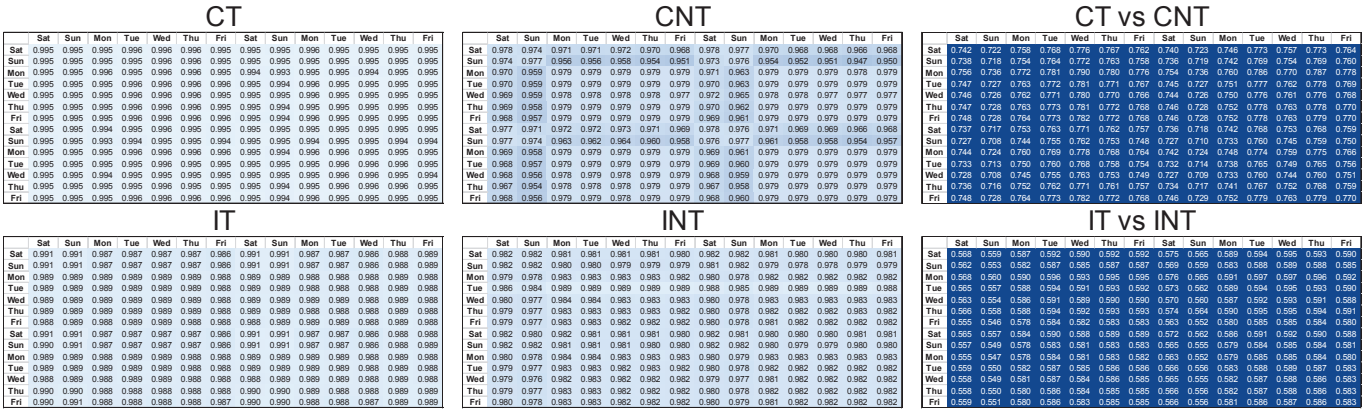


Fig. 9. Evaluation of the CT, CNT, IT, and INT models. Each row shows how well the model computed from one day's worth of data fits the data of all the other days. Light areas represent good fitting and dark areas poor fitting.

TABLE I
MODELS: PARAMETERS AND EVALUATION.

	σ	μ	k	λ	α	${R_P}^2(X_q, S_q)$	${R_C}^2(X_q, S_q)$	${R_Q}^2(X_q, S_q)$	$\text{obj}(X_q, S_{pq})$
CT	1.31	2.44	1.32	0.34	0.86	0.985	0.999	0.999	0.994
CNT	1.59	1.72	1.35	0.30	0.77	0.962	0.999	0.999	0.987
IT	0.73	3.65	1.08	6.16	0.42	0.970	0.999	0.996	0.988
INT	1.65	3.34	1.21	1.93	0.87	0.981	0.990	0.978	0.983

of two sets of quantiles) magnifies errors in the tail of the distribution, their association is expected to be effective.

$$obj(X,S) = \frac{1}{3}({R_P}^2(X,S) + {R_C}^2(X,S) + {R_Q}^2(X,S)). \quad (7)$$

We build four types of duration models for CT, CNT, IT, INT, based on fourteen days' worth of data of duration samples coming from the same dataset used in evaluating the group mobility detection method as described in Section III-D, which consists of different days of the week. Let S_{pq} be the samples of date p and type $q \in \{CT, CNT, IT, INT\}$, and $S_q = \bigcup_p S_{pq}$. To build a good model for given type q, we determine parameters X_q so as to maximize the objective function as Eq. (8). Maximization is performed using the L-BFGS-B algorithm [35].

$$X_q = \underset{X}{\text{argmax}} \ \ obj(X, S_q). \quad (8)$$

As there are fewer train users than non-train users, this corresponds to about 1 million samples each for the CT and IT models and 25 million samples each for the CNT and INT models for each day. Table I shows the parameters and the averaged fitting performance of the *most general* models for CT, CNT, IT, and INT, respectively. Figure 8 shows the two components of the CNT model, its PDF, PP-plot, and QQ-plot. The PDFs of the four models are depicted in Fig. 10(a) and will be discussed in IV-C.

B. Evaluation of the models

In order to evaluate our models and their computation process, we use an approach similar to the leave-one-out cross-validation technique used in statistical model validation [36].

For each day of our dataset, we compute a model using this day's worth of data and then evaluate how well it fits the data of all the other days. This allows us to ensure that the computed models do not overfit the data. The evaluation is performed on two weeks of data using the objective function, $obj(X_{pq}, S_{p'q'})$ where X_{pq} is the maximized parameters (9) for date p and $q \in \{CT, CNT, IT, INT\}$, $S_{p'q'}$ is the comparison samples of date p' and $q' \in \{CT, CNT, IT, INT\}$.

$$X_{pq} = \underset{X}{\text{argmax}} \ \ obj(X, S_{pq}). \quad (9)$$

Figure 9 shows the values of the objective function for each model type, where one row represents the evaluation of one day's model by all the other days. The values are shown by indicating the gradient of intensity from $obj \leqslant 0.8$ (dark) to $obj = 1.0$ (light). The average value is 0.995 for CT, 0.969 for CNT, 0.988 for IT, and 0.981 for INT. As a comparison, the two last tables of Fig. 9 show the results obtained when evaluating the CT models with the CNT data (average of 0.755) and when evaluating the IT models with the INT data (average of 0.578) and show that these models are very different.

We can deduce from these results that, independently of the day, all models fit remarkably well the data of all other days. This proves that our models are generalized well to other datasets and thus that they do not overfit the data. Also, the slightly darker areas that can be observed in the CNT table correspond to Saturdays and Sundays. As their evaluation averages 0.965, although these models still fit the data, it is suggested that the behavior of users on weekend is slightly different from weekdays.

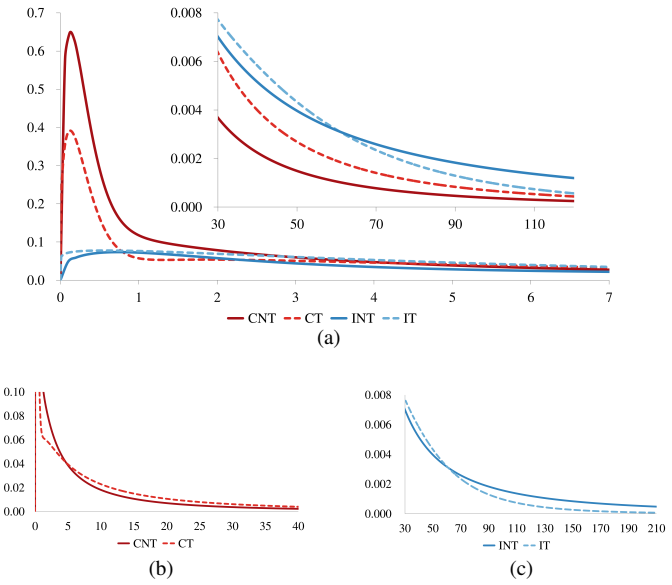


Fig. 10. (a) shows the PDFs of the CNT, CT, INT and IT models for durations in $[0s, 7s]$ and $[30s, 120s]$. (b) shows the CNT and CT PDFs between 0s and 40s. (c) shows the INT and IT PDFs between 30s and 210s.

C. Discussion

Figure 10(a) shows that connected duration distributions have an important peak of short durations corresponding to background applications, whereas idle duration distributions are much more driven by human behavior and look like log-normal distributions. Idle duration distributions also contain a significant proportion of long durations (more than 1 min), corresponding to periods when users are not using their device at all.

Let us now compare the models for train users and non-train users in order to determine the influence of a moving group of users on the mobile network. Figure 10(b) shows that CNT contains proportionally more durations between 0s and 4s than CT and fewer between 4s and $+\infty$. Moreover, the short-durations peak is higher in CNT than in CT. This means that, not on a train, more accesses to the network are made by background applications or correspond to quick accesses by users, for example to check for new messages or briefly consult an application. On a train, however, accesses generally last longer, for instance to browse the web or watch a video. This means that users tend to use their devices more on a train, which can be explained by the fact that they have more time to spare, especially when commuting. As shown in Fig. 10(c), a train trip contains more durations between 0s and 60s than INT and fewer between 60s and $+\infty$. As short idle durations correspond to frequent network accesses and long idle durations correspond to sparse network accesses, this means that users tend to access the network more frequently on a train.

Let μ_{CT}, μ_{CNT}, μ_{IT}, and μ_{INT} be the means of these distributions: we obtain $\mu_{CT} = 22.7s$, $\mu_{CNT} = 15.2s$, $\mu_{IT} = 24.5s$, and $\mu_{INT} = 96.3s$. As devices constantly alternate between one connected duration and one idle duration, the activity of users can be modeled as the couple of random variables (C, I) with $C \sim CT$, $I \sim IT$ for train users and $C \sim CNT$, $I \sim INT$ for non-train users. The proportion of time that train users spend connected is thus $\mu_{CT}/(\mu_{CT}L\mu_{IT}) = 48.1\ \%$ and the proportion of time non-train users spend connected is $\mu_{CNT}/(\mu_{CNT} + \mu_{INT}) = 13.6\ \%$. We deduce from these values that train users are about 3.5 times more active than non-train users. A large group of moving users will therefore have 3.5 times more impact on network resources than the same number of non-moving users, confirming that public transportation has a significant impact on mobile networks in metropolitan areas.

The use of control-plane signals to approximate network utilization is presented here as an example of user-level analysis that can be conducted using our group mobility detection method. Compared with analyses based on user-plane data, it has the advantage of being extremely fast and can be performed on all users without requiring data sampling. Estimating network utilization with connected and idle durations still remains an approximation.

V. Simulation-based assessment for mobile network functions

In this section, we present an example of how to evaluate base station functions by using the connectivity models. We build appropriate simulation configurations representing mobile users communications in a metropolitan area with commuter trains and train stations.

In the simulation, we focus on the situation where train users have worse QoS than non-train users. One possible solution is to put base stations near train tracks. However, those base stations would only be used when a train is passing by. In order to mitigate QoS degradation for train users without wasting extra power in base stations, we define dynamic base station functions essentially based on the ideas discussed in [29], [30]. Our purpose is to evaluate the following functions using our connectivity models and simulation environment.

- *On/off switching function:* Dynamic on/off switching of base stations when a train is getting nearer or getting farther from the base stations.
- *Dynamic orientation function:* Dynamic orientation of the base stations' antennas to follow the movement of the trains.

A. Simulation configuration

To assess the efficiency of dynamic base stations, we develop a network simulation using QualNet 7.4 [37] that can simulate packet-level behaviors of wireless communications in realistic environments.

1) City setup: Figure 11 depicts the outline of the simulation and the parameters for city setup are described in Table II. Three train stations are located in the area and each train station is surrounded by seven static eNBs that have three sectors. These eNBs are located in hexagonal grids. One train, 200m long and 3m wide, goes from west (left) to east (right) at a constant 60km/h. This speed represents the average travel speed of commuter trains in the middle of metropolitan urban city. The simulation starts when the train is at the west edge,

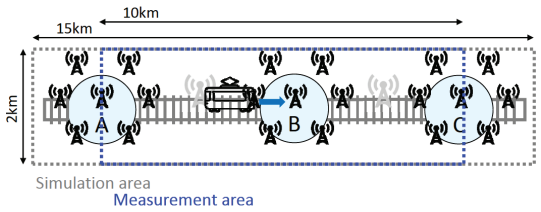


Fig. 11. Simulation field.

TABLE II
SIMULATION PARAMETERS FOR CITY SETUP.

Parameters	Values
Size of simulation area	2 km x 15 km
Size of measurement area	2 km x 10 km
Number of base stations	21 (3 train station x 7 base stations)
Number of sectors	3
Distance between train stations	5 km
Train movement	60 km/h (constant)

and the measurement starts when the train is at station A, and ends when the train is at station C. The simulation ends when the train is at the east edge. A total of 500 non-train users are distributed in the coverage of each sector and those users do not move during simulation. Varied numbers of train users are uniformly distributed in the train. The eNB selection by users is based on received signal power, e.g., train users are likely to connect to the eNB from which a user device observes the strongest signal power.

We assume one extra base station, called additional BS, is added at each intermediate point between two train stations, i.e., (A and B) and (B and C). In order to evaluate the fundamental efficiency, we use five scenarios as follows.

- [A] As a reference model, no additional BS.
- [B-1] Additional BS without on/off switching or dynamic orientation functions.
- [B-2] Additional BS with an on/off switching function but no dynamic orientation function.
- [B-3] Additional BS with no on/off switching function but with a dynamic orientation function.
- [B-4] Additional BS with both on/off switching and dynamic orientation functions.

These scenarios are differentiated by the presence or absence of each function in order to simplify the determination of the performance benefit of each function.

2) Radio setup: The parameters for radio setup are described in Table III. We adopt the 800MHz frequency band for wireless communication, as is commonly used for LTE communication. To simplify the situation, user devices use only one channel in the simulation and the band width for the communication is 10MHz. The transmission power is 45dBm which means that the packet would be transferred for an approximately 2km distance, which is assumed be the size of macro-cell ordinarily deployed in cities. The path loss model is standardized in 3GPP [38] and fading would be modeled as a random process since it can be due to multi-path propagation, shadowing from obstacles, etc.. The path loss $P_L(x)$ is described as follows:

TABLE III
SIMULATION PARAMETERS FOR RADIO SETUP.

Parameters	Values
Channel Frequency	800 MHz
LTE bandwidth	10MHz (50 RB / subframe)
LTE Scheduler type	ROUND-ROBIN
eNB transmission power	45 dBm
Pathloss model	COST231-Walfisch-Ikegami NLOS
Shadowing model	LOGNORMAL
Shadowing mean	10 dB

$$P_L(x) = \bar{P}_L(x) + N(0, \sigma^2), \tag{10}$$

$$\bar{P}_L(x) = -55.9 + 38 \times log_{10}(x) + (24.5 + 1.5\frac{f}{925})log_{10}(f), \tag{11}$$

where x is the distance from the base station and f is the frequency of the electromagnetic wave. The serving sector is simply decided by user devices seeking the highest gain, which means user devices handover as they connect to the base station that delivers the strongest signal. Dynamic switching of base stations when a train is closer/farther than the threshold from the base stations, with threshold for switching being the radio coverage of the base station. Dynamic orientation of the base station's antennas to follow the movement of the trains. The angle and direction of the antenna are designed to face the middle of the train.

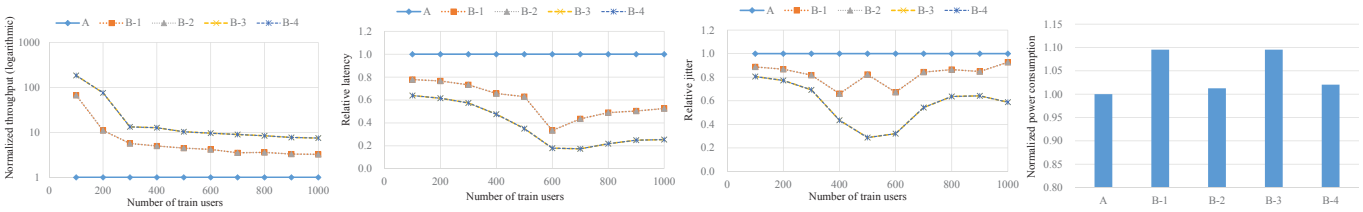
3) Application setup and evaluation metrics: In the simulation, each user device connects to the network according to the connectivity models introduced in Section IV. After the simulation starts, a user u individually starts an active duration after a random waiting duration $[0, (average(CT) + average(IT))/2]$. This draws 1-st active duration $d_{a1}^{(u)}$ for the active state from the CT/CNT model. After the active state ends, the user goes inactive and draws 1-st idle duration $d_{i1}^{(u)}$ from the IT/INT model. After the inactive state ends, the user goes active again, drawing a duration $d_{a2}^{(u)}$ from the CT/CNT model and again selecting a serving base station based on the received signal power. Each user repeats the above processes independently from each other, until the simulation ends.

To evaluate the typical performance metrics, we define the following applications and metrics. Each user device executes the applications while it is in the active state.

- *TCP-FTP.* User $u \in U$ continuously downloads files while u is connected through a TCP session. In the beginning of each active duration, user device establishes a TCP session that lasts until the end of the active duration. Using the TCP-FTP application, we measure the mean throughput Th_U as follows.

$$\mathrm{Th}_U = \frac{1}{N_U} \sum_u \frac{1}{T_u} \sum_{i=1}^{n^{(u)}} d_{ai}^{(u)} \mathrm{th}_i^{(u)}, \tag{12}$$

where N_U is the number of users, T_u is the total active duration for user u, $n^{(u)}$ is the number of active duration of user u, $d_{ai}^{(u)}$ is the i-th active duration and $\mathrm{th}_i^{(u)}$ is the throughput for the i-th active duration of user u.



(a) Th_U of train users. (b) Lat_U of train users. (c) J_U of train users. (d) Power consumption.

Fig. 12. Relative performance normalized to [A].

- *UDP-CBR.* User u periodically receives fixed size UDP packets. The size of a UDP packet is 1k byte and is sent every 2 seconds, which is sufficiently small to make it unlikely that packets will be dropped during communication due to network congestion. As for the UDP-CBR application, we measure the mean latency Lat_U and jitter J_U as follows.

$$\mathrm{Lat}_U = \frac{1}{N_U}\sum_u \frac{1}{L_u}\sum_{l=1}^{L_u} \mathrm{Lat}_l^{(u)}, \tag{13}$$

$$\mathrm{J}_U = \frac{1}{N_U}\sum_u \frac{1}{L_u - 1}\sum_{l=2}^{L_u} \left|\mathrm{Lat}_l^{(u)} - \mathrm{Lat}_{l-1}^{(u)}\right|, \tag{14}$$

  where L_u is the number of packet user u sends, and $\mathrm{Lat}_l^{(u)}$ is the latency for l-th packet user u sends.

To simplify the conditions, these two applications do not appear together. By considering two different applications, we have ten scenario variations in total. In addition, we measure the simple power consumption C_e to evaluate the efficiency of the eNBs.

$$C_e = \frac{1}{T}\sum_e s_e T_e, \tag{15}$$

where s_e is the number of sectors eNB e has, and T_e is the duration eNB e is connected by at least one user device. Here we simply assume that the power consumption for eNB is proportional to the duration that the eNB is connected to any user device.

B. Simulation result

Figure 12 depicts the relative values of TCP throughput, UDP latency, UDP jitter, and power consumption averaged over the whole measurement period of 600 secs, which are respectively normalized to those in [A]. Figures (a),(b), and (c) depict TCP-throughput, UDP-latency and UDP-jitter of train users. Figure (d) depicts the power consumption. From Figs. (a) to (c), [B-1] and [B-3] indicate the same performance as [B-2] and [B-4] respectively, which suggests the dynamic switching ON/OFF function does not affect the user perceived performance in this configuration. In Fig. (d), [B-1] indicates the same additional power consumption as [B-3] since the additional base station has no switching scheme in both scenarios.

In terms of train users' performance shown in Fig. 12(a), (b) and (c), even without the dynamic orientation function, the additional base station can clearly mitigate the degradation of throughput, latency, and jitter of train users, as indicated by [B-1]. As the number of train users increases, the normalized throughput (i.e., the ratio of mitigation from [A]) decreases because the train users share the fixed amount of the additional resources. As the number of train users increases, the normalized latency and jitter decrease first but exhibit a downward convex curve, which implies the capability of the additional BS is regulated by the number of train users and is appropriate for 600 train users.

It is clearly shown by comparing [B-3] with [B-1] that the dynamic orientation function mitigates the degradation of throughput, latency, and jitter of train users more effectively. For example, in the case of 600 train users, the normalized throughput, latency and jitter are 962%, 18% and 32% respectively with the dynamic orientation function, while they are 418%, 33% and 67% without it.

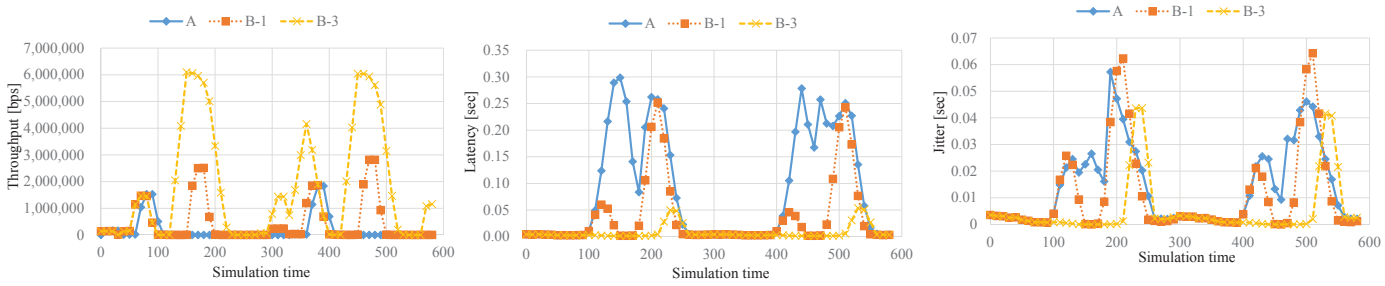
Note that, as the number of train users increases, the change of the normalized performance exhibits a similar shape to [B-1], implying that the dynamic orientation function also has the limitation regarding with the number of train users.

In terms of power consumption shown in Fig. 12(d), while the additional power consumption is 9.5% in the case of no switching scheme, the switching scheme mitigate the additional power consumption; [B-2] indicates 1.3% and [B-4] indicates 2.0%.

C. Discussion

To deeply analyze how the dynamic orientation function improves the performance, we draw time transitions of TCP throughput, UDP latency and UDP jitter of train users as depicted in the case of 1000 train users in Fig. 13. The horizontal axis is the simulation time and the vertical axis is the mean values during the corresponding 10 seconds' period. The train constantly moves and its head arrives at train station A, B and C, at 0s, 300s, and 600s respectively.

In Fig. 13(a), [A] indicates high throughput around $t = 80$ and 380, which means that the throughput for train user is unlikely to increase as soon as the train go through a train station. This suggests that TCP throughput cannot increase quickly after being connected to a congested base stations. In case of [B-1], the same phenomenon occurs at $t = 80$ and 380. Also, the additional BS improves the throughput especially when the train travels around the middle of two train stations. In this situation, since the additional BS is not used by non-train users, the throughput quickly increases as soon as train users connect to the additional BS. [B-3] indicates much



(a) TCP throughput Th_U. (b) UDP latency Lat_U. (c) UDP jitter J_U.

Fig. 13. Time transition of TCP throughput, UDP latency and UDP jitter for train users in the case of 1000 train users.

greater improvement. Since the additional BS with dynamic orientation function covers a wider area than that without the function, train users perceive earlier increase and later decrease of throughput than that in [B-1]. In addition, while the same phenomenon occurs at $t = 80$, the throughput around $t = 380$ is better than the other scenarios. This suggests that a better handover in [B-3] may contribute a higher TCP throughput in the succeeding base station.

Figure 13(b) indicates that the latency is drastically improved at $t = 150$ and 450, when the train travels around the middle of train stations. In [A], two peaks can be observed at $t = 150$ and 450, which means base stations around train stations do not sufficiently cover the train users while train is between train stations. In contrast, in [B-1], two small peaks and two large peaks can be observed around $t = 120$ and 210, and $t = 420$ and 510 respectively, which means that the additional BSs partially cover between train stations but there are still gaps between base stations around train stations and additional base stations. The difference of the size of the peak largely depends on the angle of the sector of base stations, which are likely to cover the east (right) side of each train station in the scenarios. As for [B-3], the dynamic orientation function cover almost all area between train stations. Thus, as described in Fig. 12(b), the dynamic orientation function drastically improves the latency. On the other hand, small peaks are still observed at $t = 240$ and 540 due to a small gap of the coverage of base stations, which could be improved by designing the location and angle of base stations.

In Fig. 13(c) the shapes of time transition are similar to Fig. 13(b). Although the average jitter indicates insignificant difference between [A] and [B-1], the time transition indicates a certain improvement at $t = 150$ and 450. On the other hand, two peaks in [B-3] at $t = 240$ and 540 are relatively greater than those of Fig. 13(b), which means that train users are likely to use the network so that the small gap of base station coverage cause relatively large degradation in jitter.

As a summary of this section, our simulation reveals that the dynamic base station scheme on the additional base station largely mitigates the QoS degradation in terms of TCP throughput, UDP latency, and UDP jitter, with a small volume of additional power consumption. On the other hand, our simulation suggests the limitation of dynamic orientation function, that is the capability of the additional BS is regulated by the number of train users and the coverage of the base station. The additional BS with dynamic orientation function is likely to provide better Web browsing by the improving of throughput and smooth voice call and video streaming services by improving of latency and jitter, than that without dynamic orientation function.

VI. Conclusion

In this paper, we have focused on fast and dense group mobility and mobile network signaling data. Fast and dense group mobility may cause a significant degradation of users perceive QoS. Signaling data is convenient and useful for mobile network operators due to its low volume and its easy accessibility. Firstly, a lightweight group mobility detection method was developed, based solely on signaling data. The method can successfully detect train movements in an actual LTE network. Secondly, based on the same data and the results obtained by the detection method, connected/idle duration models for train users and non-train users were built to characterize network utilization. The obtained models revealed that train users consumed about 3.5 times more resources than non-train users, which is consonant with the fact that public transportation induces dynamic changes in network utilization and significant affects the network resources. Finally, these models were leveraged in mobile network simulations to assess the effectiveness of a dynamic base station switching/orientation scheme to mitigate QoS degradation with low power consumption in a group mobility scenario. The simulation results showed that the dynamic switching/orientation functions improved users' perceived throughput, latency and jitter, which were 962%, 18%, 32% compared to those without the additional base stations implementing the functions, with small amount of additional power consumption of 2.0% in case of a moderate number of train users. It was also indicated that the scheme could not be very effective when the number of train users becomes larger. This would suggest that group mobility detection and the obtained connection/idle duration models based on control-plane data analytics are usable and useful for the development of mobility-aware resource allocation functions in base stations.

Our future work includes enhancing the control-plane data analytics for quickly and accurately monitoring group mobility and modeling the access patterns in group mobility. Validating the generality of the models and the simulation conditions also remains a task, and is essential for a reliable simulation-based

assessment of newly introduced mobility-aware functions before their deployment. More effective online use of control-plane data analytics should also be considered for monitoring and estimating the users' perceived QoS in the field, which allows the constant evaluation of mobility-aware functions after being deployed on base stations.

REFERENCES

[1] Q. Plessis, M. Suzuki, T. Kitahara, and S. Ano, "Group mobility in mobile networks: Signaling based detection and network utilization modeling," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec 2016, pp. 1–7.

[2] B. Malarkodi, P. Gopal, and B. Venkataramani, "Performance evaluation of adhoc networks with different multicast routing protocols and mobility models," in *ARTCom '09. International Conference on*, Oct 2009, pp. 81–84.

[3] Y. Li and I. R. Chen, "Mobility management in wireless mesh networks utilizing location routing and pointer forwarding," *IEEE Transactions on Network and Service Management*, vol. 9, no. 3, pp. 226–239, September 2012.

[4] Y. Chew, P. K. Tham, S. Nanba, B. S. Yeo, and H. Nakamura, "More Results on the Validation of Gravity Model and the Effect of User Mobility in Cell Planning," in *Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE*, Singapore, May 2008, pp. 2755–2759.

[5] T. Taleb, M. Bagaa, and A. Ksentini, "User Mobility-Aware Virtual Network Function Placement for Virtual 5G Network Infrastructure," in *Communications (ICC), 2015 IEEE International Conference on*, Jun. 2015, pp. 3879–3884.

[6] Z. Zaidi and B. Mark, "A Mobility-Aware Handoff Trigger Scheme for Seamless Connectivity in Cellular Networks," in *Vehicular Technology Conference, 2004.*, vol. 5, Sep. 2004, pp. 3471–3475.

[7] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong, "On the levy-walk nature of human mobility," *IEEE/ACM Transactions on Networking*, vol. 19, no. 3, pp. 630–643, Jun. 2011.

[8] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky, "Human mobility characterization from cellular network data," *Communications of the ACM*, vol. 56, no. 1, pp. 74–82, 2013.

[9] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, "Route classification using cellular handoff patterns," in *Proceedings of the 13th International Conference on Ubiquitous Computing*, ser. UbiComp '11. New York, NY, USA: ACM, 2011, pp. 123–132. [Online]. Available: http://doi.acm.org/10.1145/2030112.2030130

[10] Y. Yamada, A. Uchiyama, A. Hiromori, H. Yamaguchi, and T. Higashino, "Travel estimation using control signal records in cellular networks and geographical information," in *2016 9th IFIP Wireless and Mobile Networking Conference (WMNC)*, July 2016, pp. 138–144.

[11] H. Ishizuka, N. Kobayashi, M. Kurokawa, C. Ono, and T. Hara, "Traffic analysis of railways using call detail records," in *Proceedings of the main conference on the scientific analysis of mobile phone datasets 2015*, ser. Netmob'17, 2017.

[12] Y. Zhang, "User mobility from the view of cellular data networks," in *INFOCOM, 2014 Proceedings IEEE*, April 2014, pp. 1348–1356.

[13] X. Hong, M. Gerla, G. Pei, and C.-C. Chiang, "A Group Mobility Model for Ad Hoc Wireless Networks," in *Proceedings of the 2nd ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, New York, NY, USA, 1999, pp. 53–60.

[14] L. Tu, F. Zhang, F. Wang, and X. Wang, "A Random Group Mobility Model for Mobile Networks," in *Ubiquitous, Autonomic and Trusted Computing, 2009. UIC-ATC '09. Symposia and Workshops on*, Jul. 2009, pp. 551–556.

[15] H. Du, Z. Yu, F. Yi, Z. Wang, Q. Han, and B. Guo, "Recognition of group mobility level and group structure with mobile devices," *IEEE Transactions on Mobile Computing*, vol. PP, no. 99, pp. 1–1, 2017.

[16] Y. Li, M. Zhao, and W. Wang, "Internode Mobility Correlation for Group Detection and Analysis in VANETs," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 9, pp. 4590–4601, Nov. 2013.

[17] B. Zhang, Z. Song, C. H. Liu, J. Ma, and W. Wang, "An event-driven qoi-aware participatory sensing framework with energy and budget constraints," *ACM Transactions on Intelligent System and Technology.*, vol. 6, no. 3, pp. 42:1–42:19, Apr. 2015.

[18] C. H. Liu, J. Zhao, H. Zhang, S. Guo, K. K. Leung, and J. Crowcroft, "Energy-efficient event detection by participatory sensing under budget constraints," *IEEE Systems Journal*, vol. PP, no. 99, pp. 1–12, 2016.

[19] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley, 1990.

[20] T. Zhang, R. Ramakrishnan, and M. Livny, "Birch: A new data clustering algorithm and its applications," *Data Mining and Knowledge Discovery*, vol. 1, no. 2, pp. 141–182, Jan. 1997.

[21] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan, "Clustering data streams: Theory and practice," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 3, pp. 515–528, May 2003.

[22] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, no. 1, pp. 53–65, Nov. 1987.

[23] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," *IEEE/ACM Transactions on Networking*, vol. 25, no. 2, pp. 1147–1161, April 2017.

[24] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Large-scale measurement and characterization of cellular machine-to-machine traffic," *IEEE/ACM Transactions on Networking*, vol. 21, no. 6, pp. 1960–1973, Dec 2013.

[25] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, S. Venkataraman, and J. Wang, "Characterizing and optimizing cellular network performance during crowded events," *IEEE/ACM Transactions on Networking*, vol. 24, no. 3, pp. 1308–1321, Jun. 2016.

[26] G.-H. Tu, Y. Li, C. Peng, C.-Y. Li, H. Wang, and S. Lu, "Control-plane protocol interactions in cellular networks," in *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 4, 2014, pp. 223–234.

[27] F. S. D. Silva, A. J. Neto, D. B. Maciel, J. Castillo-Lema, F. d. O. Silva, and P. F. Rosa, "SDN Based Control Plane Extensions for Mobility Management Improvement in Next Generation ETArch Networks," in *Proceedings of the 18th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*. ACM, 2015, pp. 189–193.

[28] G. Gorbil, O. H. Abdelrahman, M. Pavloski, and E. Gelenbe, "Modeling and Analysis of RRC-Based Signalling Storms in 3G Networks," *IEEE Transactions on Emerging Topics in Computing*, vol. 4, no. 1, pp. 113–127, Jan 2016.

[29] E. Oh, K. Son, and B. Krishnamachari, "Dynamic Base Station Switching-On/Off Strategies for Green Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 2126–2136, May 2013.

[30] Y. Okada, H. Tsuji, H. Kagiwada, and A. Sano, "Millimeter-wave broadband wireless access system with tracking technology of moving targets," in *Vehicular Technology Conference, 1998. VTC 98. 48th IEEE*, vol. 3, May 1998, pp. 2057–2061 vol.3.

[31] Z. Zhang, C. Jiao, C. Zhong, H. Zhang, and Y. Zhang, "Differential modulation exploiting the spatial–temporal correlation of wireless channels with moving antenna array," *IEEE Transactions on Communications*, vol. 63, no. 12, pp. 4990–5001, Dec 2015.

[32] "Evolved Universal Terrestrial Radio Access Network, S1 Application Protocol (S1AP)," *document TS 36.413, V12.4.0, Release 12, 3GPP*, Sep. 2014.

[33] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[34] M. Rosenblatt *et al.*, "Remarks on some nonparametric estimates of a density function," *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 832–837, 1956.

[35] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-scale Bound-constrained Optimization," *ACM Transactions on Mathematical Software*, vol. 23, no. 4, pp. 550–560, Dec. 1997.

[36] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143.

[37] Qualnet network simulator software. (SCALABLE Network Technologies, Inc.). [Online]. Available: http://web.scalable-networks.com/qualnet-network-simulator-software

[38] "Spatial channel model for Multiple Input Multiple Output (MIMO) simulations," *document TR 25.996, V10.0.0, Release 10, 3GPP*, Mar. 2011.

Masaki Suzuki received the B.E., M.E. and Ph.D. degrees in engineering from Keio University, Japan, in 2006, 2008 and 2013, respectively. His research interests include network traffic analysis, QoS monitoring, mobile users' behavior analysis and intelligent transportation systems. He is currently a Research Engineer with KDDI Research, Inc., Japan. He is a member of IEEE, ACM, IEICE and IPSJ.

Takeshi Kitahara is a very nice person.

Shigehiro Ano is a very very nice person.

Masato Tsuru is a very very very nice person.