This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2015.2393302, IEEE Transactions on Mobile Computing

Opportunistic Sampling for Joint Population Size and Density Estimation

Farid Movahedi Naini, Olivier Dousse, Patrick Thiran, *Fellow, IEEE,* and Martin Vetterli, *Fellow, IEEE*

Abstract—Consider a set of probes, called "agents", who sample, based on opportunistic contacts, a population moving between a set of discrete locations. An example of such agents are Bluetooth probes that sample the visible Bluetooth devices in a population. Based on the obtained measurements, we construct a parametric statistical model to jointly estimate the total population size (e.g., the number of visible Bluetooth devices) and their spatial density. We evaluate the performance of our estimators by using Bluetooth traces obtained during an open-air event and Wi-Fi traces obtained on a university campus.

Index Terms—Population size and density estimation, opportunistic sampling, Bluetooth sampling

1 Introduction

Estimating population size and population density finds applications in various fields. For example, ecologists and biologists are interested in estimating the population sizes of certain animal species (refer to [1], [2], [3] for a review). In the field of urban analysis, estimating population density is important, e.g., to create evacuation paths, to plan new locations for department stores (refer to [4], [5], [6], [7], [8] and references therein). Social networking applications such as activity-hotspot detection [9], make use of population density to pinpoint night-life hotspots to users of the application.

In the above mentioned examples, the measurements for population size and density estimation are obtained using various techniques. For example, in the case of population size estimation of certain animal species, traps are constructed to capture the animals. Upon capture, the animals are marked and then released. This method, known as the capture-recapture method [10], uses the number of times that animals are recaptured to infer the population size. To estimate crowd density for urban analysis, surveillance cameras are installed in different locations in a given area; computer-vision techniques are then applied to the captured data to count the people and to estimate the density of the crowd in these different locations [7], [11]. Most of the measurement techniques for estimating population size and density, as will be reviewed in Section 2, come with at least one of the following drawbacks or requirements: (i) investment in hardware (e.g., installing cameras in different positions, which also raises privacy issues [6]), (ii) deployment constraints (e.g., requiring everyone to carry RFIDs [12]), and (iii) proprietary issues (e.g., requiring cooperation with another party such as a GSM operator [8], [9]).

The (spatial) density of a population is a measure of the number of people present in different locations within an area of interest (e.g., a campus or a city). By normalizing the spatial densities such that they sum to one, we obtain the *relative* spatial density that measures the relative 'popularity' of different locations. Most existing methods essentially count the number of people in an area of interest by using some 'agents' that 'monitor' the locations (e.g., surveillance cameras). In practice, some locations might not be monitored for some periods of time, because there might be fewer agents than the number of locations for many reasons such as those outlined above, or because agents might be mobile and some locations might occasionally be empty of agents. If the total population size in the area of interest is known, then we can calculate the number of people in the non-monitored locations: the total population size minus the number of people in monitored locations gives us the number of people present in non-monitored locations; this gives us the overall density in the non-monitored locations. However, if the total population size is unknown, then estimating the spatial density becomes more challenging, as we do not know the number of people present in the non-monitored locations.

In this work, we consider the joint estimation of population size and density for the case where measurements are obtained based on *opportunistic* contacts between some *agents* that monitor the population and the population members. In particular, we consider the join estimation of population size and density based on Bluetooth measurements. Nearly every current mobile phone is equipped with a Bluetooth radio interface, each with a unique MAC address. This technology includes a detection functionality, where enabled devices can *discover* (detect) each other within a small radius (typically 10-20m), which is refered to as their *proximity*. It has also been observed [13], [14], [15], [16], [17] that a non-trivial fraction of mobile phone users leave the detection feature of their phone turned on constantly ("discoverable (visible) mode"). A particularly interesting feature is that when they are in visible mode, phones broadcast their MAC address, which makes them *uniquely identifiable*. This possibility enables us to use mobile phones as sensing devices and to evaluate different features related to the mobility patterns of the population.

Our contributions are as follows. On the theory side, we

- *F. M. Naini, P. Thiran, and M. Vetterli are with EPFL, Lausanne, Switzerland.*
  E-mail: {farid.movahedinaini, patrick.thiran, martin.vetterli}@epfl.ch.
- *O. Dousse is with HERE, Berlin, Germany.*
  E-mail: olivier.dousse@here.com.

develop a parametric estimator for the joint estimation of population size and density. Our estimator is a "minimally sufficient" estimator of population size and densities, i.e., an estimator that uses optimally all available information collected from the agents. It extends the population size estimator that we introduced in [14]. On the practical side, we use an opportunistic sampling of the population (e.g., Bluetooth measurements) in contrast to other works where sampling is systematic following a predefined planning. We only consider the case where measurements are performed by mobile agents who move normally in the area of interest (i.e., agents were given no specific movement instructions). On the empirical side, as is explained later, instead of directly using the detection patterns of the Bluetooth devices by the agents, we use the *contact* patterns, which profoundly impacts our estimator. In this setting, several questions need to be answered: What kind of information do the agents need to collect in order to estimate the density? When is it possible to estimate with good accuracy the population density from such traces? To the best of our knowledge, our work is the first effort to use such measurements of opportunistic nature for the joint estimation of population size and density. The main theoretical challenge of our approach is incorporating the mobility of the agents, which makes the computations more involved. A practical challenge in our approach is knowing the percentage of visible Bluetooth devices in the population. On average, close to 8.2% of people carry Bluetooth devices with an activated detection functionality [13], [14], [15], [16], [17], which is large enough to make possible density estimations from Bluetooth measurements.

The rest of this paper is organized as follows. After a brief literature review in the next section, we describe the experiment we conducted at the Paléo Music Festival and the obtained measurements in Section 3. We first revisit in more detail the population size estimation model of [14] in Section 4, and apply it on the Paléo measurements in Section 5. We then extend our model to the joint-estimation model of population size and density in Section 6, and apply it to the Paléo measurements in Section 7. We apply our estimators to a second dataset based on Wi-Fi technology in Section 8, and finally conclude the paper in Section 9.

2 Related Work

The problem of the estimation of population size has a long history (refer to [1], [2], [3] for a review); perhaps one of the first estimators of population size is the Turing-Good estimator presented in [2]. An important line of work in estimating the population sizes of certain animal species is the *capture-recapture* methods. In these techniques, traps are set up to capture some individuals of the animal population, after which they are marked and released. All the animals are vulnerable to the sampling process by these traps during the experiment. In the recapture process, some of the animals are captured again and the number of previously marked animals will provide information that is used to infer about the population size [10], [18]. In contrast to these works, we do not place monitoring devices or traps at given places, and we cannot start and stop the measurement campaign at given times. In our case, the "sensing devices" are carried by *individuals* from the population, with an uncontrolled, random, mobility pattern and who arrive and leave the monitored area at different, random times. The individuals are thus exposed to the sampling process for different random times. Moreover, some methods [1], [18] when applied to our problem, only account for whether an individual has been discovered by an agent or not. Whereas, in our estimator we make the best use of the information available, e.g., we process the pattern of contact between an individual and an agent.

In the field of information theory, *alphabet-size estimation* [19], pattern-likelihood maximization [20], and sequence-probability estimation [21], [22] also address related problems. These works usually assume that an observed *sequence* is drawn in an independent and identically distributed (i.i.d.) manner from a source with an unknown underlying distribution, and an unknown alphabet size, that is to be estimated. In order to apply these methods to our measurements, we would have to consider an agent as a source, and her Bluetooth traces as such a sequence. We would then have several (more precisely, a number equal to the number of agents) sequences drawn from several sources, that have the same alphabet. However, to the best of our knowledge, there is no methodological way to deal with multiple sequences/sources where the sources have the same alphabet. Furthermore, the estimator in [19] assumes that the underlying distribution of the source is uniform, which is not true in our case, because individuals have different probabilities of being detected by an agent due to their diverse mobility patterns.

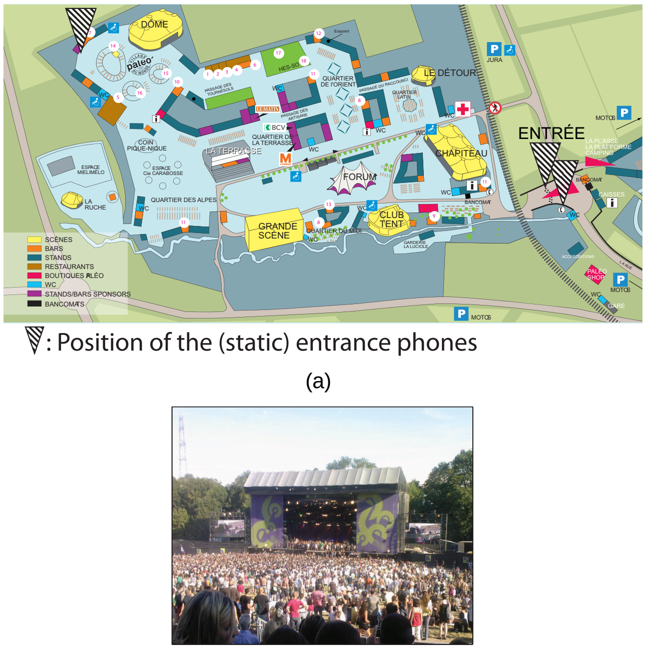
Recently, social networking applications have generated interest in developing methods for estimating the number of nodes in graphs based on some *sampling of the graphs*; the interested reader can refer to [23] and references therein. In these methods, at each step of the sampling process, exactly one node in the graph is sampled (e.g., by one random walker on the graph acting as an agent), and local information of the node (such as its neighboring nodes) is measured. After the sampling process runs for a certain number of steps, the obtained measurements are used to estimate the total number of nodes in the graph. In contrast, in our measurement process, at a step of the sampling process, i.e., Bluetooth scanning by an agent, it could happen than (i) no Bluetooth device is detected, or (ii) more than one Bluetooth device is detected. These two cases occur when there are zero and several, respectively, Bluetooth devices present in the proximity of the agent when the Bluetooth scanning is performed.

Computer vision techniques have been widely used to estimate population density [6], [7], [12], [24], [25]; the methods presented use surveillance cameras to capture images of crowds in order to count the number of people and estimate density. Their performance is usually affected by factors such as background lighting, the density of the crowd, and the view angles of the camera. The difficulties of the computer vision approaches, besides their cost, are in finding places to install the cameras, and in avoiding privacy issues. In addition, these methods are able to estimate the density only in the monitored locations, as the cameras act as *static* agents.

RFID-based techniques [26], [27], [28], [29] require the population to wear special RFID tags. These tags are later localized in order to analyze the spread of the population over different locations in order to estimate the density of the population in a given region. Some techniques exploit the wireless networking infrastructure in order to perform a passive-density estimation [30], [31], [32]. These methods model the change of the RSSI in the system in order to infer the density of people. In order to estimate the density, some algorithms use measurements (obtained via cellular phone operators) that indicate the position of cellular subscribers in different locations [8]. CitySense [9] clusters GPS and WiFi data to indicate the hotspots of activity in San Francisco. There even exist hand-counting methods [4], which need investment in personnel, and are intractable for large areas of interest such as those considered in our experiments. In [33] the authors used Bluetooth probes to estimate crowd density at the town center of Kaiserslautern. Their approach is based on the comparison and fusion of collected data from different probes. In [34] static Bluetooth sensors are deployed in a music festival in order to study group formation and music preferences of attendees.

3 The Paléo Experiment

In this section, we describe the experiment that we conducted at Paléo Music Festival, which took place in July 2010 in Nyon, Switzerland.



(a)

(b)

Fig. 1. (a) Paléo Music Festival map. The surface covers around $280,000$ m^2. Position of the entrance phones is indicated by dark triangular markers. (b) Attendees listening to a concert in 'Grande Scene'.

3.1 Experiment Description

The Paléo Music Festival is one the major music festivals in Europe: it attracts several tens of thousands of people per day. It is an open-air festival, which allows for GPS coverage, and takes place within a closed area with fixed entrance and exit points. The surface of the festival covers around $280,000$ m^2. These characteristics make this festival a good environment for performing experiments related to population sampling via Bluetooth. In order to have a better understanding of the environment of the festival, a map and a snapshot of attendees listening to a concert are shown in Figure 1. Our idea is to sample the population by sending some attendees as "agents" inside the festival. Each agent is equipped with a mobile phone (Nokia N95) that is programmed to regularly scan for Bluetooth devices within its range (10-20 m). The phones then collect the Bluetooth MAC addresses of mobile devices that have their Bluetooth visibility turned on. Bluetooth MAC addresses are unique to each device and can be used as the identifiers of attendees. The purpose is to use this information to estimate the population size and density of the attendees (more precisely, of the subset of those who carry visible Bluetooth devices).

In order to have the ground truth of the number of visible Bluetooth devices at the festival, a Bluetooth scanning is done at the entrances. Two mobile phones are installed at the main entrance of the festival, and another phone is installed at the back entrance. The position of these three mobile phones is shown by markers in Figure 1(a). The same gates are used both for the entrance and exit of attendees. We refer to these three phones as the *entrance phones*.

In our experiment, ten people (acting individually) took part as agents. The agents' phones and entrance phones were programmed to perform Bluetooth scanning every 80 seconds. The agents' phones were also programmed to record GPS positions. The experiment was performed during one day of the festival, and the duration of the festival (opening/closing of the entrance/exit gates) on that day was 13 hours 15 minutes; from 15h00 until 4h15 on the following day.

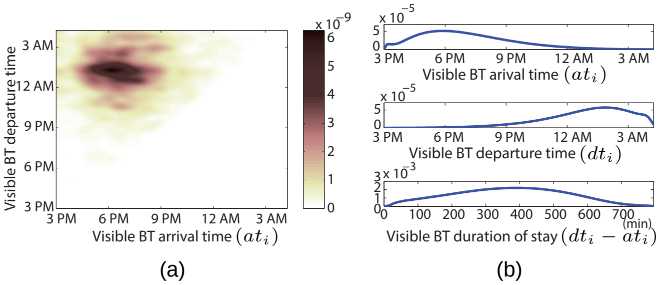
3.2 Obtained Measurements

In this section we discuss the measurements obtained in the experiment.

3.2.1 *Measurements at the Entrances*

For the entrance phones, we consider only the Bluetooth traces that were collected during the opening hours of the festival. In total, 3326 different Bluetooth devices were discovered at the entrances. The estimated number of attendees (obtained on the basis of the number of tickets sold and the tickets punched at the entrance gates), which was provided to us by the organizers of the festival, is 40,536. From these two numbers, we get 8.2% as the approximate percentage of attendees who have visible Bluetooth devices. This ratio depends on many factors such as the characteristics of the population (e.g. age). Other estimated ratios reported in the literature are as follows: 4.7% to 7% in a campus bar [13], 8% to 12.5% in an airport [17], 11% in a cultural and theater festival [15], and 13% in a sports event [16].

As the entrance phones can discover all the visible Bluetooth devices upon their arrival and departure, we can compute the empirical arrival/departure time distribution of the visible Bluetooth devices, which is plotted in Figure 2(a). The empirical marginal distributions of the visible



(a) (b)

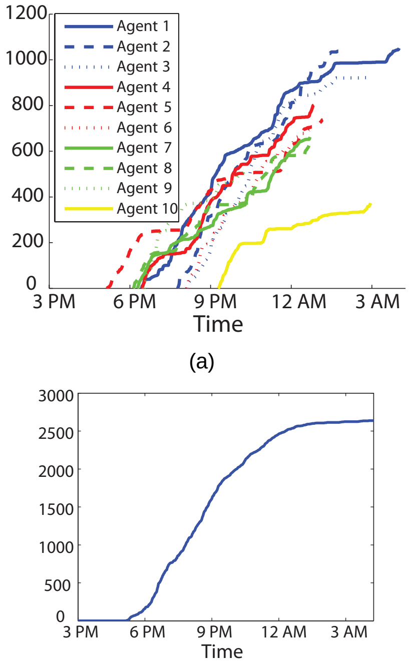
Fig. 2. (a) Empirical distribution of the visible Bluetooth devices' arrival/departures times to/from the Paléo Music festival. (b) Empirical marginal distributions of the visible Bluetooth devices' arrival times (top), departure times (middle), and the duration of stay on the festival grounds (bottom).

Bluetooth devices' arrival times, departure times, and the duration of stay on the festival grounds are plotted in Figure 2(b). The average (the 90% confidence interval) of the visible Bluetooth devices' arrival times is 18h56 (respectively, $[15\text{h}00, 21\text{h}22]$). The value for the visible Bluetooth devices' departure times is equal to 01h01 (respectively, $[22\text{h}25, 03\text{h}58]$). The average (the 90% confidence interval) of the visible Bluetooth devices' duration of stay on the festival grounds is equal to 365.4 min (respectively, $[102.4 \text{ min}, 591.6 \text{ min}]$).

3.2.2 *Measurements by Agents*

For the agents' phones, we consider only the Bluetooth traces that were collected during the period when the agents were on the festival grounds; we can determine these periods by using either the obtained GPS traces from the agents' phones. The 10 agents were able to discover 2637 out of 3326 of the Bluetooth devices discovered at the festival entrances. This corresponds to 79.3% of the visible Bluetooth devices. This ratio is referred to as the *coverage* in the literature on animal species estimation [1]. We expect this ratio to be less than 100%, because there were only a few agents present for a short period of time at the festival and the mobile phones have a short Bluetooth range. Nevertheless, the coverage percentage is rather large.

Here we analyze in more detail the Bluetooth measurements obtained by each agent. Figure 3(a) shows the evolution of the total number of discovered Bluetooth devices as a function of time for each agent. The agents are numbered in a decreasing order, according to the total number of discovered Bluetooth devices. By looking at the evolution of the curves in Figure 3(a), we observe that as soon as the agents arrived at the festival, they began discovering Bluetooth devices rapidly. All agents have however periods during which the slope of the curve is quite flat. These periods correspond, for example, to periods when the agents were listening to the main concerts, hence, were not moving; during these periods, they kept detecting the same Bluetooth devices, but discovered fewer new Bluetooth devices. Figure 1(b) shows one such situation. Figure 3(b) shows the accumulated total number of discovered Bluetooth devices by all the agents, the final value of the curve is equal to 2637.



(a)

(b)

Fig. 3. Evolutions as a function of time of (a) the total number of different Bluetooth devices discovered by each agent, and (b) the accumulated total number of discovered Bluetooth devices by all the agents.

4 Population Size Estimation Model

Our goal is to estimate the total number of visible Bluetooth devices and their spatial density at the festival, based on agents' traces. We start by defining our notation and introducing the population size estimation model.

4.1 Data Structure and Notation

4.1.1 *Population*

The population is comprised of attendees with visible Bluetooth devices. We call the population members *individuals*, use variable i to index them, and use masculine pronouns to refer to them. We assume that every individual carries one Bluetooth device with him at the festival. Denote the population size by N and the festival duration by T_{fest}. We shift the time origin such that the festival opening time is at 0 and its closing time is at T_{fest}. Let at_i and dt_i denote, respectively, the arrival and departure times of individual i to/from the festival; these variables will be treated as random variables. We assume that the tuple (at_i, dt_i) for every individual i is drawn in an i.i.d. fashion from the probability density function (pdf) $f(at, dt)$, on which we will elaborate later. The empirical estimate of $f(at, dt)$ is shown in Figure 2(a).

Note that in order to empirically verify the i.i.d. assumption of tuples (at_i, dt_i) and (at_j, dt_j) for individuals i and j, we would need several realizations of the two tuples (e.g. the measured arrival/departure times of the two individuals across different days of the festival). As we only have one set of measurements of the individuals' arrival/departure times, we cannot empirically verify the i.i.d. assumption; but we argue that some individuals tend to arrive/depart to/from the festival in groups. In such situations, the arrival/departure times of the individuals within a group will not be independent. The group sizes in which attendees arrive/depart to/from the festival were rarely greater than 6, yielding a probability of having more than one visible Bluetooth device in the group less than 0.08.

4.1.2 Agents

We denote the number of agents by M, use variable j for indexing them, and use feminine pronouns to refer to them. Let at_j^A and dt_j^A denote the arrival and departure times of agent j to the festival. Note that agents' arrival and departure times, unlike those of the individuals, are known to us. Let $t^j_{at_i,dt_i}$ denote the duration of time between the arrival and departure of individual i, which is overlapped with the arrival and departure of agent j. We have

$$t^j_{at_i,dt_i} = \max\Big(\min(dt_j^A, dt_i) - \max(at_i^A, at_i), 0\Big). \quad (1)$$

Intuitively, the chance that agent j discovers individual i increases as the value of $t^j_{at_i,dt_i}$ increases. We further assume that when individuals arrive at the festival, they stay on the festival grounds until they depart from the festival.

4.1.3 Detection

As described in the beginning of Section 3, the agents perform a Bluetooth detection every 80 seconds and record a list of the MAC addresses of the visible Bluetooth devices in their proximity, i.e., their Bluetooth communication range. The data that each agent provides consists therefore of a list of MAC addresses detected by the agent, together with the corresponding detection times during her stay at the festival. Denote by S the total accumulated number of discovered MAC addresses by all the agents and map the discovered MAC addresses to the set $\{1, \ldots, S\}$. In our experiment, the value of S is equal to 2736.

Consider an individual who stays in the proximity of an agent for some continuous period of time, but not before or after that period. In this case, we say that the individual is *in contact* with the agent during that period of time. Such contact periods can be identified from the detection pattern of the individual in the following way. During a contact between the individual and the agent, he will be detected by the agent at every 80 second interval when the Bluetooth scanning is performed, resulting in a *burst* of consecutive detections, each shifted by 80 sec. Thus every burst of consecutive detections represents one contact period; a contact starts at the first detection of a burst and finishes at the last detection of the burst.

It will be clear later why we focus on the contacts that occur between the agents and the individuals rather than the detections. In particular, we denote by k_{ij} the number of times that individual i is *contacted* by agent j. We denote by $k_i = \sum_{j=1}^{M} k_{ij}$ the total number of times that individual i is contacted. Note that individual i is discovered (i.e., is among the S discovered individuals) if and only if $k_i > 0$ (if he has been contacted by at least one of the agents).

4.2 Model Assumptions

In this work we adopt a parametric approach to the estimation of population size and density. Following our intuition, k_{ij} is an increasing function of $t^j_{at_i,dt_i}$ given in (1), and for a fixed agent j, we expect to observe different values of k_{ij} across the individuals. This is because individuals have diverse mobility patterns hence some of them are more easily *contactable* by the agent than the other individuals. In our population size estimation model, we assume the following.

- Poisson contacts: The number of times agent j contacts individual i, i.e., k_{ij}, is Poisson distributed with mean equal to $\lambda_i t^j_{at_i,dt_i}$, where λ_i is called the *contact rate* of individual i.
- Independence: The random variable k_{ij} is independent from $k_{i^{'}j^{'}}$ for $i \neq i^{'}$ and/or $j \neq j^{'}$.

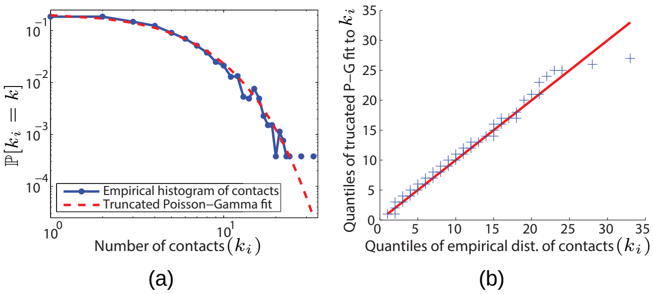
In other words, we set the mean number of contacts of individual i by agent j to be proportional to the amount of time during which both individual i and agent j are on the festival grounds ($t^j_{at_i,dt_i}$), following our intuitive expectation, and to the specific contact rate (λ_i) of individual i:

$$k_{ij} \sim \text{Poisson}\left(\lambda_i \cdot t^j_{at_i,dt_i}\right). \quad (2)$$

From (1), parameter $t^j_{at_i,dt_i}$ is a function of agent j and individual i's arrival/departure times to/from the festival. Consequently, if individual i's exact arrival/departure times are known, then the exact value of $t^j_{at_i,dt_i}$ can be calculated. Otherwise, if the distribution for individual i's arrival/departure times is known, then the distribution of $t^j_{at_i,dt_i}$ can be computed. The *contact rate* λ_i represents how easily an individual *puts himself in a contactable position* on the festival grounds, which is analogous to the concept of *abundance level* in the literature on the species estimation problem [1]; some individuals place themselves in one location which is not frequently visited by others, and in particular by the agents; others move from one place to another. Hence, we assume that for individual i, λ_i is a random variable drawn from a Gamma distribution with unknown parameters α and β, independently from other individuals and from his arrival and departure times. We use the Gamma prior, because it is a flexible distribution and it is the conjugate prior of the Poisson distribution. The probability density function of λ_i is $f_{\lambda_i}(\lambda_i; \alpha, \beta) = \beta^{\alpha} e^{-\beta\lambda_i} \lambda_i^{\alpha-1}/\Gamma(\alpha)$, where $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$. Its first two moments are $\mathbb{E}[\lambda_i] = \alpha/\beta$, and $\sigma^2_{\lambda_i} = \alpha/\beta^2$.

4.2.1 Assumptions Verification

Before going into more detail about our Poisson-Gamma model, we first verify it against the measurements. Based on the Poisson-contact assumption in (2) and the independence assumption, and given the values of λ_i, at_i, and dt_i for individual i, k_i is also Poisson distributed, i.e., $k_i = \sum_{j=1}^{M} k_{ij} \sim \text{Poisson}(\lambda_i \sum_{j=1}^{M} t^j_{at_i,dt_i})$. Now consider the values of $k_1, k_2, \ldots, k_S$ for the S discovered individuals; these values follow a *truncated* Poisson distribution, because the value of k_i for individual i must be nonzero in order for the individual to be observed. The solid curve in Figure 4(a) shows the empirical distribution of the observed k_i for $i = 1, 2, \ldots, S$. As we have access to the arrival/departure times of every individual (thanks to the entrance phones in the Paléo dataset), we can compute



(a)

(b)

Fig. 4. The goodness of fit of truncated Poisson-Gamma distribution to the measurements. (a) the probability distribution function and (b) the Q-Q plot with respect to the empirical distribution of contacts.

$t^j_{at_i,dt_i}$ from Equation (1) for $i = 1, \ldots, S$. Based on the model, we fit a truncated Poisson-Gamma distribution to the measurements (see details in Section 4.3). The dashed curve in Figure 4(a) shows the analytical distribution of k_i based on the truncated Poisson-Gamma fit to the measurements. Figure 4(b) shows the Q-Q plot of the two distributions. These two figures verify that our Poisson-Gamma model fits well to the observed measurements based on the number of contacts. In particular, the two probability distributions have similar tail behaviors.

We now consider the case where, instead of modeling the number of contacts, we model the number of detections. We repeat the above steps but replace k_{ij} with the number of detections of individual i by agent j, i.e., number of detections instead of the number of bursts of detections. In fact, the fit of the Poisson-Gamma model to the measurements based on the number of detections is not good. In particular, the empirical histogram of the detections and the corresponding fitted Poisson-Gamma distribution have very different tail behaviors. The Pearsons chi-squared test for the equality of the two distributions gives a p-value of 8.61×10^{-9}. We reject the equality assumptions of the two distributions (i.e., the empirical histogram of the detections and the corresponding fitted Poisson-Gamma distribution) when the p-value is smaller than the predetermined significance level α, which in practice is set to 0.01 or 0.05 [35]. Due to the lack of space, corresponding plots similar to those of Figure 4 are given in [36].

The reason for the poor fit when we model the detections is that we are taking into account the duration of contacts (i.e., duration of the bursts) between the agents and the individuals. The duration of a contact between an agent and an individual is a complex variable driven, for the Paléo Festival, by the duration of the concerts (which explains the shallow parts of the curves in Figure 3(a)). By taking into account all the detections in our model, we are taking into account these factors, which complicates the observed measurements, and as a result we obtain a poor fit. By taking into account only the number of contacts of the individuals with the agents, and thus neglecting the number of detections, we capture the *mixing* that happens on the festival grounds and discard irrelevant factors that complicate the process.

The Poisson-Gamma model has previously been used in the literature to address problems related to population size estimation [37], [38]. In these methods, all the population members (e.g., animals) are vulnerable to the sampling process (e.g., traps) for the entire duration of experiment. However, in our experiment, this assumption does not hold, and we account for this by using the pdf $f(dt, at)$. Some other methods [1], [18] could be applied to this problem, but they will only account for whether individual i has been discovered by agent j or not. In other words, they only take into account the indicator function $\mathbb{1}_{\{k_{ij}>0\}}$, but not k_{ij} itself. These methods address the problem by modeling the discovering probability of an individual, and come with the limitation that this discovering probability of an individual does not scale linearly with time, hence the effect of time cannot be readily included. In contrast, in the proposed Poisson model, the average number of times agent j contacts individual i scales linearly with time, as we would expect. Moreover, parameters λ_i and $t^j_{at_i,dt_i}$ have meaningful interpretations. Finally, we use all the information of the values taken by k_{ij}, and not only by $\mathbb{1}_{\{k_{ij}>0\}}$.

4.3 Likelihood Function

In order to derive the estimator for N, we compute the probability of observing the obtained measurements under the model described above with parameters N, α, β. This is usually called the *likelihood function*. We then choose the set of parameters, in particular N, that maximize this likelihood. The likelihood function has the following form:

$$L(N, \alpha, \beta) = \underbrace{\binom{N}{S} \left(1 - p_{dsc}(\alpha, \beta)\right)^{N-S}}_{L_1(N,\alpha,\beta)} \cdot \underbrace{\prod_{i=1}^{S} \mathbb{P}_i}_{L_2(\alpha,\beta)}, \quad (3)$$

where p_{dsc} and $\mathbb{P}_i$ are given below. The first term (L_1) is the likelihood of the undiscovered individuals, and the second term (L_2) is the likelihood of the pattern of the discovered individuals. Below, we discuss each of the two components of the likelihood function. The computation makes use of the following property of the Gamma distribution: for all real positive $x, y > 0$,

$$\mathbb{E}_\lambda \left[e^{-\lambda x} \lambda^y \right] = \frac{\Gamma(\alpha + y)\beta^\alpha}{\Gamma(\alpha)(\beta + x)^{\alpha + y}}. \quad (4)$$

4.3.1 Likelihood of the Undiscovered

Let $p^{(at,dt,\lambda)}_{dsc}$ be the probability that at least one of the M agents discovers an individual having contact rate λ, and arrival and departure times at, dt. Using the Poisson contact assumption, we have

$$p^{(at,dt,\lambda)}_{dsc} = 1 - \prod_{j=1}^{M} e^{-\lambda t^j_{at,dt}} = 1 - e^{-\lambda \sum_{j=1}^{M} t^j_{at,dt}}. \quad (5)$$

As the contact rate λ is a random variable, by taking the expectation over λ using (4) we have

$$p^{(at,dt)}_{dsc}(\alpha, \beta) = \mathbb{E}_\lambda[p^{(at,dt,\lambda)}_{dsc}] = 1 - \left(\frac{\beta}{\beta + \sum_{j=1}^{M} t^j_{at,dt}} \right)^{\alpha}, \quad (6)$$

and by computing the expectation of this probability over the joint distribution of the arrival and departure times $f(at, dt)$, we get

$$p_{dsc}(\alpha, \beta) = 1 - \mathbb{E}_{at,dt} \left[\left(\frac{\beta}{\beta + \sum_j t^j_{at,dt}} \right)^{\alpha} \right]. \quad (7)$$

The likelihood of the undiscovered individuals is equal to the probability of not discovering $N - S$ of the individuals:

$$\begin{aligned} L_1(N, \alpha, \beta) &= \binom{N}{N-S} \left(1 - p_{dsc}(\alpha, \beta)\right)^{N-S} \\ &= \binom{N}{S} \left(\mathbb{E}_{at,dt} \left[\left(\frac{\beta}{\beta + \sum_{j=1}^{M} t^{j}_{at,dt}} \right)^{\alpha} \right] \right)^{N-S} . \end{aligned} \tag{8}$$

4.3.2 Likelihood of the Discovered

We first compute the probability of the observed pattern of contacts by each agent for one of the discovered individuals. Given that individual i has contact rate λ and arrival and departure times at, dt, the probability for him to be contacted k_{ij} times by agent j for $j = 1, \ldots, M$, is

$$\mathbb{P}_i^{(at,dt,\lambda)} = \prod_{j=1}^{M} e^{-\lambda t^{j}_{at,dt}} \frac{(\lambda t^{j}_{at,dt})^{k_{ij}}}{k_{ij}!}. \tag{9}$$

By taking the expectation over λ using (4) and after some manipulations we have

$$\mathbb{P}_i^{(at,dt)} = \frac{\Gamma(\alpha + k_i)\beta^{\alpha}}{\Gamma(\alpha)(\beta + \sum_{j=1}^{M} t^{j}_{at,dt})^{\alpha + k_i}} \prod_{j=1}^{M} \frac{\left(t^{j}_{at,dt}\right)^{k_{ij}}}{k_{ij}!}, \tag{10}$$

and by taking the expectation over at, dt we get

$$\mathbb{P}_i = \mathbb{E}_{at,dt} \left[\frac{\Gamma(\alpha + k_i)\beta^{\alpha}}{\Gamma(\alpha)(\beta + \sum_{j=1}^{M} t^{j}_{at,dt})^{\alpha + k_i}} \prod_{j=1}^{M} \frac{\left(t^{j}_{at,dt}\right)^{k_{ij}}}{k_{ij}!} \right]. \tag{11}$$

The second part of the likelihood is equal to the probability of the observed pattern for all the discovered individuals. Using the independence assumption we have

$$L_2(\alpha, \beta) = \prod_{i=1}^{S} \mathbb{P}_i. \tag{12}$$

4.3.3 Maximum Likelihood Estimator

The full likelihood is the product of the two likelihoods in Equations (8) and (12):

$$\begin{aligned} L(N, \alpha, \beta) = &\binom{N}{S} \left(\mathbb{E}_{at,dt} \left[\left(\frac{\beta}{\beta + \sum_{j=1}^{M} t^{j}_{at,dt}} \right)^{\alpha} \right] \right)^{N-S} \\ &\times \prod_{i=1}^{S} \left\{ \frac{\Gamma(\alpha + k_i)\beta^{\alpha}}{\Gamma(\alpha)} \mathbb{E}_{at,dt} \left[\frac{\prod_{j=1}^{M} \frac{\left(t^{j}_{at,dt}\right)^{k_{ij}}}{k_{ij}!}}{(\beta + \sum_{j=1}^{M} t^{j}_{at,dt})^{\alpha + k_i}} \right] \right\}. \end{aligned} \tag{13}$$

We define the maximum likelihood estimators for N, α, β as

$$(\widehat{N}, \widehat{\alpha}, \widehat{\beta}) = \underset{N,\alpha,\beta}{\arg\max} \log L(N, \alpha, \beta), \tag{14}$$

where $L(N, \alpha, \beta)$ is the full likelihood given by (13). $\widehat{N}$ is the maximum likelihood estimator for the population size. In the next section, we apply the above maximum likelihood estimator on the Paléo measurements and compare its performance with other existing methods.

5 Results of Population Size Estimation

5.1 Input Measurements for the Population Size Estimator

For clarification purposes, we list in Table 1 the input to our estimation model of population size, which is called a *statistic* in usual terminology. We have the following theorem, where proof is given in [36] because of lack of space.

Theorem 1. The input quantities in Table 1 are minimally sufficient statistics for estimating the population size in our model. □

Theorem 1 means that the input contains the minimally sufficient information for estimating the population size. In other words, any more information is irrelevant for estimating N, and removing any information from the statistic deteriorates the estimation of N based on our model.

Here we elaborate on the choice of the model for arrival and departure times $f(at, dt)$. As mentioned before, the individuals' arrival/departure times to/from the festival are not known in general. We use three different arrival and departure-time distributions, that we discuss below.

5.1.1 Deterministic

One extreme choice for $f(at, dt)$ is a deterministic arrival time and departure time for all the individuals. We choose $f_1(at, dt) = \delta(at)\delta(dt - T_{fest})$, where $\delta(\cdot)$ is the Dirac function. This distribution assumes that all the individuals enter at the beginning of the festival (time 0) and leave at the end of the festival (T_{fest}), as in the studies in [18], [37], [38].

5.1.2 Actual Distribution

The opposite extreme choice for $f(at, dt)$ is to use the Bluetooth traces obtained from entrance phones to estimate the actual distribution of $f(at, dt)$. This information is in general not available, but is used in our experiment for benchmarking purposes. We computed the empirical distribution of $f(at, dt)$, shown in Figure 2(a).

5.1.3 Low Informative

In practice, we do not have sufficiently detailed information about arrival and departure times to estimate $f(at, dt)$. We assume that we have access to the first two moments of individuals' arrival/departure times. We then approximate individuals' arrival and departure times by two independent Gaussian distributions centered at the corresponding mean arrival and departure times with the corresponding standard deviations (refer to Fig. 2(b)). A tuple (at, dt) is valid if both elements fall within time period $[0, T_{fest}]$ and if $dt > at$.

5.2 Estimating the Population Size

For each of the three pdfs $f(at, dt)$ described above, we computed the maximum likelihood estimator of population size given in (14). The result is given in Table 2. We observe that the naive choice of deterministic arrival and departure times gives a relatively large undershoot. The explanation for this underestimation is that with deterministic (at, dt), all the individuals arrive at the beginning and leave at

- k_{ij} for $i = 1, 2, \ldots, S$; $j = 1, 2, \ldots, M$ (resp. $k_{ij}^{(l)}$ for $i = 1, 2, \ldots, S$; $j = 1, 2, \ldots, M$; $l = 1, 2, \ldots, K$),
- $\left(at_j^A, dt_j^A\right)$ for $j = 1, 2, \ldots, M$ (resp. $\left(at_j^A, dt_j^A\right)$ for $j = 1, 2, \ldots, M$, and agents' trajectories),
- Individuals' exact arrival/departure times to/from the festival, or their distribution $f(at, dt)$ or some approximation of the distribution.

TABLE 1
Input measurements for the population size estimator (resp. for the population size and density estimator)

Choice of $f(at, dt)$	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{p}_{dsc}$	$\widehat{N}$	$(N - \widehat{N})/N$
$f_1(at, dt)$	1.583	1670.5	0.849	3106	6.61%
$f_2(at, dt)$	1.868	1345.3	0.796	3311	0.45%
$f_3(at, dt)$	1.961	1774.5	0.805	3275	1.53%

TABLE 2
Comparison of the estimated population size with the ground truth (3326) for three different distributions of arrival/departure times.

Method	$\widehat{N}$	$(N - \widehat{N})/N$
PML [39]	3129	5.95%
M_{th} in [18]	3013	9.46%
[19]	2676	19.54%

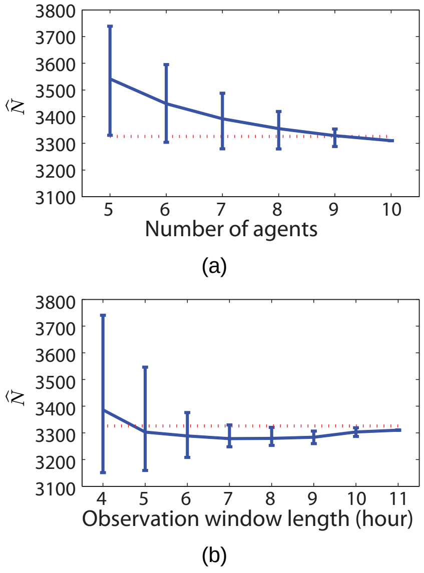
TABLE 3
Comparison results with other estimators.

the end of the festival, and hence the overlap time between agents and individuals is overestimated. The discovering probability is overestimated, which results in an undershoot. By using a non-deterministic $f(at, dt)$ instead, individuals are in contact with the agents on average for a smaller time duration, hence the discovering probability decreases and we have an increase in the estimated population. We also observe that by estimating $f(at, dt)$ from the entrance-phone traces, we get surprisingly close to the true value ($N = 3326$). This is expected as the model fits well to the observed measurements based on Fig. 4(a) and Fig. 4(b).

We also compare our method with the capture-recapture method described in [18], with the pattern-maximum likelihood (PML) method in [39], and with the method in [19]. The results are shown in Table 3; all methods exhibit an undershoot, which is explained as follows. Recall that the time duration that each individual is vulnerable to the sampling process is random (according to his arrival and departure time), which is not taken into account in [18]. Therefore, the result has an undershoot similar to our method for the choice of $f_1(at, dt)$. The method in [18] assumes uniform sampling of the population, which is not valid in our experiment and is the reason for the undershoot. We remark that the approximation used in the estimator in [19] is not valid for our measurements, thus we have used the exact expression provided in [19]. PML, a nonparametric method described in [39], gives the best result among the three. As explained in Section 2, no method in the state of the art, copes with the randomness both in the sampling process and in the arrival/departure times of the actual measurement setting. For comparison purposes, in particular with respect to the effect of these additional random factors, it is however useful to evaluate how they would perform on this dataset.

5.3 Results of Population Size Estimation by Using a Subset of the Information

In this section, we apply our population size estimator to a subset of the measurements. We use the estimator given in (14) and use the actual distribution of individuals' arrival and departure times ($f_2(at, dt)$). In the first part, we consider the measurements obtained by a subset of size m of the agents. For each subset of size m, we consider all the possible combinations of the agents and estimate the population size for each combination of m agents. The average estimated population sizes and their 90% confidence intervals for all the combinations of size m are shown in Figure 5(a) for $m = 5, 6, \ldots, 10$. In the second part, we consider the measurements obtained by all agents during an observation window of length w smaller than the festival duration. Consider the agents' arrival and departure times shown in Figure 3(a); the observation window starting from the moment when the first agent arrives at the festival (17h09 for agent 5), until the moment when the last agent departs from the festival (4h01 for agent 1) is approximately 11 hours. We partition this interval into slots 10-minutes in length. For an observation window of length w, we consider all the consecutive 10-minute slots with total length w, and we estimate the population size based on the measure-



(a)

(b)

Fig. 5. Estimated population size as a function of (a) the number of agents and (b) the observation window length. The solid lines and the bars show the average and the 90% confidence interval, respectively. The dashed line shows the ground truth for the population size.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2015.2393302, IEEE Transactions on Mobile Computing

ments obtained during these slots. The average estimated population sizes and their 90% confidence intervals are shown in Figure 5(b) for $w = 4, 5, \ldots, 11$ hours. Note that for the right-most point in both figures, there is only one estimate for the population size, specifically, the one based on the entire measurements. Thus, the calculated confidence interval is zero when $m = 10$ (Figure 5(a)) and $w = 11$ hours (Figure 5(b)). We observe that as the number of agents decreases, the averaged estimated population size increases. In [36] we have plotted the average and the 90% confidence intervals of the number of discovered individuals (S) as functions of m and w.

5.4 Estimating the Total Number of Attendees

Remember that N in (14) is the number of attendees who carry visible (i.e., discoverable) Bluetooth devices. In order to estimate the entire number of attendees (with or without a visible Bluetooth device), the ratio of attendees who carry visible Bluetooth devices needs to be estimated. One way to estimate this ratio is to compare visual counts of attendees entering the gates with the number of discovered Bluetooth devices during the same period. Several of these counts can be performed at different time periods and the resulting ratios averaged [15]. The same idea is used in the works of [40], [41] to propose techniques to estimate the ratio of discovered Bluetooth devices.

Let N_{Tot} be the total number of attendees and let r be the ratio of attendees carrying visible Bluetooth devices, i.e., $r = N/N_{Tot}$. Recall from Section 3.2.1 that in our experiment, $N_{Tot} = 40,536$ and $r = 0.082$. Let $\widehat{N} = N(1 + \Delta N)$ and $\widehat{r} = r(1 + \Delta r)$ be the estimates for N and r, respectively, with relative errors equal to ΔN and Δr. We have

$$\widehat{N}_{Tot} = \frac{\widehat{N}}{\widehat{r}} = \frac{N(1 + \Delta N)}{r(1 + \Delta r)}. \tag{15}$$

If $|\Delta N| \ll 1$ and $|\Delta r| \ll 1$ then, $\widehat{N}_{Tot} \approx N_{Tot}(1 + \Delta N - \Delta r)$, which means that in the worst case, the relative error in estimating the total number N_{Tot} of attendees is approximately equal to the sum of the relative errors ΔN and Δr in estimating N and r. Hence, in a setting where Δr is relatively small (e.g., smaller than 20%), the choice of $\widehat{N}$ in Tables 2 and 3 has a measurable impact on the final error in estimating $\widehat{N}_{Tot}$. For instance, in [41] the authors estimate the ratio of visible Buetooth devices carried by pedestrians in a high-traffic area that is similar to the entrance gates at Paléo. They report a 9.2% visibility ratio with a relative error of about 20%. Under such a setting, the relative error in estimating $\widehat{N}_{Tot}$ in (15) by using f_3 and f_1 in Table 3 would be 21.5% and 26.7%, respectively.

Although estimating the entire population size of attendees requires the knowledge of r, some population characteristics, such as the relative density of attendees at different locations, scales linearly with the size of the subset of visible Bluetooth devices. Hence, studying this subset of attendees gives us insight into the behavior of the entire population. This is the topic of the next section.

6 Joint Population Size/Density Estimation

In this section, we extend our population size estimation model to the joint-estimation model of population size and density. Recall that the agents' phones also record GPS positions; hence the approximate locations where the agents contact the individuals are known. Nevertheless, our population size estimation model does not differentiate among the locations where the individuals are contacted by the agents; the algorithm only processes the number of times that the agents contact the individuals. In other words, the information of the exact location where an agent contacts an individual is not relevant for estimating the population size. However, as we approximate the locations where individuals are contacted, we are able to use this extra information to infer the spatial density of different locations. For example, consider the case where most of the contacts happen in a small subset of the locations; in this case, we can conclude that locations in this subset are more popular than the rest of the locations. In this section, we present a model that takes into account the location where the individuals are contacted by the agents in order to jointly estimate population size and density. Our idea is to *split* the contact rates of the individuals into a set of *location-dependent* contact rates. Before describing the model in more detail, we introduce a new notation.

We partition the area of interest (e.g., the festival area) into K locations $S_1, S_2, \ldots, S_K$, which determines the granularity of the density estimation. Consider a density (popularity) vector $\boldsymbol{\pi}^* = [\pi^*(1), \pi^*(2), \ldots, \pi^*(K)]$, where the non-negative density value of $\pi^*(l)$ is associated with location l for $l = 1, 2, \ldots, K$, such that $\sum_l \pi^*(l) = 1$. We later elaborate on what these density values represent. We denote by $k_{ij}^{(l)}$ the total number of times that agent j contacts individual i *in location* l; this value can be computed using the agents' trajectories. The random variable $k_i^{(l)}$ denotes the total number of times that individual i is contacted in location l; $k_i^{(l)} = \sum_{j=1}^{M} k_{ij}^{(l)}$. We denote by $t_{at_i,dt_i}^{j,(l)}$ the overlap time between individual i's presence at the area of interest (the individual could be in any location) and agent j's presence in location l; $t_{at_i,dt_i}^{j} = \sum_{l=1}^{K} t_{at_i,dt_i}^{j,(l)}$.

6.1 Model Assumptions

Our joint-estimation model of population size and density is based on these assumptions:

- Poisson contacts: The number $k_{ij}^{(l)}$ of times that agent j contacts individual i in location l is Poisson distributed with mean equal to $\lambda_i^{(l)} t_{at_i,dt_i}^{j,(l)}$, where $\lambda_i^{(l)}$ is the contact rate of individual i for location l,
- Independence: The random variable $k_{ij}^{(l)}$ for the triplet of individual i, agent j, and location l is independent from that for all other triplets of individuals, agents, and locations.

In this model, in contrast with our previous model for population size estimation, we differentiate among the locations where an agent contacts an individual. Parameter $\lambda_i^{(l)}$ represents how easily the individual *puts himself in a contactable position* in location l:

$$k_{ij}^{(l)} \sim \text{Poisson}\left(\lambda_i{}^{(l)} \cdot t_{at_i,dt_i}^{j,(l)}\right). \tag{16}$$

Parameter $t_{at_i,dt_i}^{j,(l)}$ is a function of agent j's trajectory and individual i's arrival/departure times to/from the area of interest. Consequently, if individual i's exact arrival/departure times are known, then the exact value of $t^{j,(l)}_{at_i,dt_i}$ can be calculated. Otherwise, if only the distribution for individual i's arrival/departure times is known, then the distribution of $t^{j,(l)}_{at_i,dt_i}$ can be computed.

Regarding the contact rates, we assume that for every individual i, $\lambda^{(l)}_i$ is drawn independently from all other contact rates from a Gamma distribution with unknown parameters $\alpha^{(l)}$ and β. To take into account the density of each location, we modulate the parameter $\alpha^{(l)}$ for location l with $\pi^*(l)$ as follows. We assume that $\alpha^{(l)} = \alpha\pi^*(l)$; α, β, N, and $\boldsymbol{\pi}^*$ are unknown constants similar to the population size estimation model. This particular choice of the prior for the contact rates guarantees that $\lambda_i = \sum_{l=1}^{K} \lambda^{(l)}_i$ has a Gamma distribution with parameters α and β. Hence, we split the contact rate of every individual i into location-dependent contact rates $\lambda^{(l)}_i$. If location l is a popular location, then individuals spend on average more time in that location. Consequently, the contact rates for the individuals in location l will be larger than other locations (because there is a higher chance of contacting individuals in location l, as they spend more time there). This will be reflected in the estimated Gamma distribution for location l by having a large value of $\alpha^{(l)}$, which means a large value of $\pi^*(l)$, and thus a large popularity for the location.

6.2 Likelihood Function

The likelihood function of the joint-estimation model of population size and density is obtained by a similar reasoning as in Section 4.3, and reads [36]:

$$L = \binom{N}{N-S}\left(1 - \mathbb{E}_{(at,dt)}\left[\prod_{l=1}^{K}\left(\frac{\beta}{\beta + \sum_{j=1}^{M} t^{j,(l)}_{at,dt}}\right)^{\alpha^{(l)}}\right]\right)$$
$$\times \prod_{i=1}^{S} \mathbb{E}_{(at,dt)}\left[\prod_{l=1}^{K} \frac{\beta^{\alpha^{(l)}}\Gamma\left(\alpha^{(l)} + k^{(l)}_i\right)}{\Gamma(\alpha^{(l)})\left(\beta + \sum_{j=1}^{M} t^{j,(l)}_{at,dt}\right)^{\alpha^{(l)}+k^{(l)}_i}}\right.$$
$$\left.\times \prod_{j=1}^{M} \frac{\left(t^{j,(l)}_{at,dt}\right)^{k^{(l)}_{ij}}}{k^{(l)}_{ij}!}\right]. \tag{17}$$

The above likelihood function is maximized in order to obtain the maximum likelihood estimates of the parameters N, $\alpha^{(l)}$ for $l = 1, 2, \ldots, K$, and β. The maximum likelihood estimate for the spatial density $\pi^*(l)$ of location l will then be equal to $\alpha^{(l)}/\sum_j \alpha^{(j)}$. Note that maximizing the above likelihood function is performed over $K+2$ parameters (N, $\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(K)}$, β), whereas in the case of population size estimation (Eq. (13)), it is performed over 3 parameters (N, α, β).

7 Results of Joint Population Size and Density Estimation

7.1 Input Measurements for the Joint Estimator of Population Size and Density

We list in Table 1 the input to our joint-estimation model of population size and density. Similar to Theorem 1, we have the following theorem; the proof is given in [36].

Theorem 2. The input quantities in Table 1 are the minimally sufficient statistics for jointly estimating the population size and density in our model. □

This means that based on our model, any more information is irrelevant for estimating N and $\boldsymbol{\pi}^*$, and removing any information from the input will deteriorate estimation of N and $\boldsymbol{\pi}^*$.

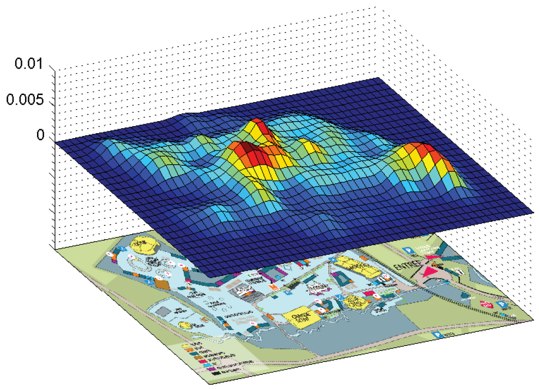


Fig. 6. Reconstructed relative spatial density (i.e., $\boldsymbol{\pi}^*$ defined in Section 6) at Paléo.

7.2 Result on the Paléo Dataset

Here we apply our joint-estimation model of population size and density to the Paléo dataset. In order to estimate the spatial density at the Paléo music festival, we partition the area into K locations. Using the GPS traces of the agents, we can reconstruct their trajectory and determine the time duration that each agent spends in every location. Then by processing agents' Bluetooth measurements, we can determine the location where each contact occurs. This will give us the values of $k^{(l)}_{ij}$, for $l = 1, \ldots, K$, $j = 1, \ldots, M$, and $i = 1, \ldots, S$. We partition the festival area into squares of approximate size $15m \times 15m$, which is comparable with communication range of Bluetooth; this gives us $K = 1200$ locations. We use the actual arrival/departure time distribution of the individuals in our computation. The maximum likelihood estimate of the population size is equal to 3062, which has 7.93% undershoot. The maximum likelihood estimate of the density after being smoothed by a low-pass Gaussian filter is shown in Figure 6.

The error in the estimated population size is larger than the errors in Table 2. One reason for the larger error is that here we are estimating many more parameters than before ($K + 2 = 1202$ versus 3), and the additional estimation of density parameters introduces error in the estimation of the population size. However, the error is still small compared to the values in Table 3. Although the Poisson-Gamma model succeeds in modeling the *mixing* that happens *globally* on the festival grounds when it is considered as one location, it fails to model the mixing that happens *locally* inside every location when the festival is partitioned. Regarding the estimated density, contrary to the population size, we do not have the ground truth of the density. Nevertheless, the estimated density shown in Figure 6 matches well with the popular locations of the event.

Dataset	Population size ground truth	Spatial density ground truth	Agents type	Comm. technology
Paléo	available	not available	real	Bluetooth
EPFL	available	available	simulated	Wi-Fi

TABLE 4
The properties of the datasets used in the experiments.

8 Result on the EPFL Campus Wi-Fi Dataset

So far we applied our estimators of population size and density to the Paléo measurements, for which only the ground truth of population is known. In this section, we estimate the population size and spatial density of people on the École Polytechnique Fédérale de Lausanne (EPFL) campus. In contrast to the Paléo measurements, these measurements are obtained using Wi-Fi technology, and both ground truths of population size and density are known. Table 4 summarizes the properties of the datasets used in our experiments.

8.1 Dataset Description

We consider the EPFL campus, which consists of several buildings and hundreds of wireless access points (APs). The main wireless network on the campus requires authentication, and can thus be accessed only by members of the university (students, faculty, etc.). The history of connections of the users to the network is recorded in the following way: Whenever a device (user) connects to the network, its (anonymized) MAC address, the time of start of the connection, and the identification (ID) of the AP to which it connects are stored in a log file (with a precision to the second). Moreover, when the device moves across the campus and gets connected (roamed) to a new AP, the time of this new connection and the ID of the AP are similarly stored. However, a device that loses its connection or disconnects, does not lead to the storage of any entry in the log file. Wireless network administrators of EPFL provided us with a log file for one weekday.

8.2 Preprocessing of the Data

We want to reconstruct the trajectories of all the users on the campus by using the log file. There are two main sources of error: First, all the wireless devices (laptops, smart-phones, etc.), connected at any time to the network, appear in the log file; therefore, if a user does not connect his device to the network, or does not carry the device everywhere he goes, his true trajectory cannot be reconstructed. Second, whenever a user leaves the campus (disconnects), the time of disconnection is unknown. To compensate for these sources of error, we consider only devices that arrive and depart within the time period between 6h00 and 24h00. In addition, we assume that a device remains connected to the same AP until the time when it is connected to a new AP (based on the log file entries). When a device is connected to an AP, it stays in the communication range of the AP (typically 50-100 m), specifically, in the access point's *vicinity*. We discard the last connection of each device, because we do not know when it terminates.

In summary, we assume that the arrival and departure times of a user to (from) the campus are equal to the first (respectively, the last) connection time associated with the user inside the log file. This will allow us to approximate users' trajectories at an access-point level of granularity. Although the reconstructed trajectories are affected by the above mentioned sources of error, they are reconstructed based on actual wireless connection logs. The empirical marginal distributions of the users' arrival times, departure times, and durations of stay on the campus are shown in [36]. The average and the 90% confidence interval of the users' arrival times are 10h37 and $[7\text{h}35, 14\text{h}36]$, respectively. The respective values for the users' departure times are equal to 17h29 and $[11\text{h}51, 22\text{h}14]$. The average (the 90% confidence interval) of users' duration of stay on the campus is equal to 412.2 min (respectively, $[30.2 \text{ min}, 669.6 \text{ min}]$).

8.3 Experiment Description

After applying the above preprocessing, 5834 devices remain for which we reconstruct the trajectories at an access-point level of granularity. Thus the ground truth for population size is $N = 5834$ individuals (users). For our experiment, we assume campus security personnel to act as our M agents, and we simulate their trajectories as follows. Every agent arrives at 6h00 at departs at 24h00. During her stay on the campus she visits different buildings uniformly at random, and the sequence of her durations of stay in each building is drawn i.i.d. from a Gaussian distribution with 1-hour mean and 10-minute standard deviation (more precisely, the distribution is a truncated Gaussian that removes negative values). We further assume that when an agent enters a building, she goes through every floor of the building consecutively from the bottom floor to the top floor. She then *visits* every AP inside each floor of the building in a random order, and she equally spreads her staying time in the building among all its APs. By visiting an AP we mean that the agent stays in the vicinity of that AP. We consider an agent to be *in contact* with an individual whenever they are both located in the vicinity of the same AP. To obtain building-level granularity we further process the trajectories by aggregating the access points within each building of the campus. This means that an individual and an agent who are inside the same building, are in contact with each other when they are in the vicinity of the same access point. The campus has 21 main buildings that, in total, consist of 680 access points, thus $K = 21$ in our experiment. To compute the ground truth for spatial density we proceed as follows. Let $\tau^{(l)}(i)$ be the duration of time that individual i spends in location (building) l; then the (relative) spatial density of location l is equal to $\pi^*(l) = \sum_{i=1}^{N} \tau^{(l)}(i) / \sum_{j=1}^{K} \sum_{i=1}^{N} \tau^{(j)}(i)$.

8.4 Assumption Verification

We proceed similarly as in Section 4.2.1 by first verifying our Poisson-Gamma assumption of contacts on this dataset. We set $M = 7$ agents and simulate their trajectories. The agents are able to discover $S = 4801$ out of the $N = 5834$ individuals, which corresponds to 82.3% of the total population. The solid curve in Figure 7 shows the empirical distribution of the observed number of contacts of the discovered individuals (note that the curves correspond to the particular obtained measurements). The dashed curve in Figure 7 shows the analytical distribution of the number of contacts, based

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2015.2393302, IEEE Transactions on Mobile Computing

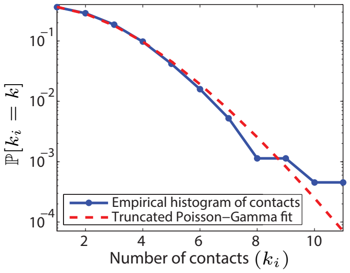


Fig. 7. The goodness of fit of truncated Poisson-Gamma distribution in the EPFL dataset.

on the truncated Poisson-Gamma fit to the measurements. We observe that our Poisson-Gamma model fits well the observed measurements, similar to the Paléo dataset (refer to Figures 4(a) and 4(b)), although the two datasets are of different nature: one is based on Bluetooth traces and the other on those of Wi-Fi. In particular, Pearson's chi-squared test for the equality of the two distributions gives a p-value of 0.41, which verifies the good fit.

8.5 Estimation of Population Size and Spatial Density

In our experiments, we perform 1000 iterations, where at each iteration we simulate trajectories for $M = 7$ agents. Similarly as for the Paléo dataset, we use three different arrival/departure time distributions: (i) the estimated actual distribution (refer to Section 8.2 for an explanation), (ii) a deterministic choice where every individual arrives at 6h00 at departs at 24h00 (similar to the agents), and (iii) a low informative choice, where we approximate individuals' arrival and departure times by two independent Gaussian distributions centered at the corresponding mean arrival and departure times with the corresponding standard deviations; a tuple (at, dt) is valid if both elements fall within time period 6h00 $-$ 24h00 and if $dt > at$.

The average and 90% confidence intervals of various quantities are shown in Figure 8. The dashed line shows the ground truth for population size $N = 5834$. The leftmost value is the number of discovered individuals (S); on average the agents discover 83.2% of the population. The second, the third, and the fourth values are, respectively, the estimated population sizes using the estimated actual distribution ($\widehat{N}_{f_4}$), the deterministic choice ($\widehat{N}_{f_5}$), and the low informative choice ($\widehat{N}_{f_6}$). Similarly to the results on the Paléo dataset (refer to Tables 2 and 3), estimating the population size by using the actual arrival/departure times distribution gives the best result among the three, whereas the deterministic choice of arrival/departure times gives a considerable undershoot. The fifth value is the result of the capture-recapture method M_{th} described in [18]; the obtained result exhibits an undershoot similar to the Paléo dataset. We also perform experiments by varying the number of agents and agents' arrival/departure times (similar to Section 5.3 for the Paléo dataset). The observed behaviors are similar to those for the Paléo dataset; due to lack of space, the results are shown in [36].

In our second experiment, at each of the 1000 iterations, we jointly estimate population size and density by using the estimated actual arrival/departure time distribution and also the low informative choice. The two rightmost values in Figure 8 show, respectively, the result for the population size using the estimated actual distribution ($\widehat{N}^J_{f_4}$) and the low informative choice ($\widehat{N}^J_{f_6}$). Similarly to Section 7.2, the error in $\widehat{N}^J_{f_4}$ is larger than that in $\widehat{N}_{f_4}$, but it is still much less than that of $\widehat{N}_{M_{th}}$ and $\widehat{N}_{f_5}$. The dashed curve in Figure 9 shows the estimated ground truth for spatial density, sorted in decreasing order of popularity. The average and the 90% confidence interval of the estimated spatial density, using the low informative choice for individuals' arrival/departure times, for each location is shown by solid curve in Figure 9. We observe that the average estimated spatial density of every location is very close to the ground truth, and that the ground truth falls within the 90% confidence interval. Figure 10 shows the reconstructed two dimensional *heatmap* for the ground truth of density and the estimated density at EPFL campus, at an access-point level of granularity. Here the estimated density is the average estimated density across all iterations (the solid curve in Figure 9), where we divide the density of every building uniformly among all of its APs.

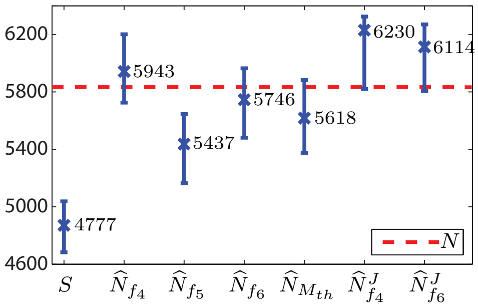


Fig. 8. Population size estimates at EPFL campus. The solid line and the bars, respectively, show the estimated average and 90% confidence intervals of various quantities (refer to Section 8.5).

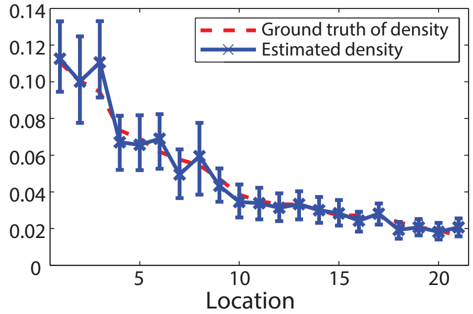
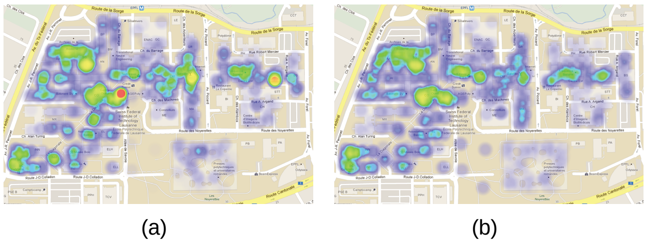


Fig. 9. Spatial density estimation at EPFL campus. The solid line and the bars show the estimated average and 90% confidence intervals, respectively. The dashed line shows the ground truth.

9 Conclusion

In this paper we have introduced a novel application that exploits the opportunistic contacts between mobile devices: we estimate population size and density by using mobile devices to sample a population. In order to test the feasibility of this method, we conducted an experiment at Paléo Music Festival. We derived a model to estimate the population of



(a) (b)

Fig. 10. EPFL Campus: Two dimensional heatmap of (a) the ground truth of density, and (b) the average estimated density (the solid curve in Figure 9). The densities are reconstructed at an access-point level of granularity.

people who carry visible Bluetooth devices, by optimally using all the available information. The resulting estimate of population size is surprisingly close to the ground truth, even with a small number of agents. We observed that by considering the contact patterns instead of the detection patterns the quality of the estimation improves. We also observed the importance of taking into account the random exposure times during which the individuals are vulnerable to the sampling process. We then extended the model to obtain joint estimation of population size and density, and applied it on both the Paléo traces and real datasets of Wi-Fi contacts over a University campus.

Although having an estimate for the number of people requires the knowledge of the ratio of visible Bluetooth devices, some population characteristics, such as the relative density of people in different time periods or in different locations within the area of interest, scale linearly with the size of the subset of visible Bluetooth devices. Therefore, the method can be used to study such population characteristics. Some open questions still remain. For example, what kind of mobility models result in contact processes that can provably be modeled with a Poisson-Gamma model? Is it better to have a large number M of agents over a short period of time T, or vice-versa? Our future work will focus also on better understanding the difference between the joint and the separate estimation of population size and density; for example, knowing the population size can there be other estimators, such as nonparametric estimators, of spatial density?

Acknowledgments

We thank Stéphane Szilagyi for his help in organizing the experiment at Paléo, Pascal Viot for providing us with an estimate of the total attendees of the festival, Julien Eberle for his help in programming the mobile phones, and the agents for participating in the experiment. We thank also Richard Timsit and Yves Despond for providing us with EPFL Wi-Fi connection logs.

References

[1] J. Bunge and M. Fitzpatrick, “Estimating the number of species: A review,” *J. Amer. Statist. Assoc.*, vol. 88, no. 421, 1993.

[2] I. J. Good, “The population frequencies of species and the estimation of population parameters,” *Biometrika*, vol. 40, 1953.

[3] C. J. Schwarz and G. A. F. Seber, “Estimating animal abundance: Review iii,” *Statistical Science*, vol. 14, no. 4, 1999.

[4] W. Cottrell and D. Pal, “Evaluation of pedestrian data needs and collection efforts,” *Transp. Res. Rec.*, vol. 1828, 2003.

[5] D. Kong, D. Gray, and H. Tao, “Counting pedestrians in crowds using viewpoint invariant training,” in *BMVC*, 2005.

[6] H. Rahmalan, M. Nixon, and J. Carter, “On crowd density estimation for surveillance,” in *IET Crime and Security*, 2006.

[7] B. Zhan, D. Monekosso, P. Remagnino, S. Velastin, and L. Xu, “Crowd analysis: a survey,” *MACH VISION APPL*, 2008.

[8] C. Ratti, S. Williams, D. Frenchman, and R. Pulselli, “Mobile landscapes: using location data from cell phones for urban analysis,” *ENVIRON PLANN B*, vol. 33, no. 5, pp. 727–748, 2006.

[9] M. Loecher and T. Jebara, “Citysense: multiscale space time clustering of GPS points and trajectories,” in *JSM*, 2009.

[10] A. Chao, “An overview of closed capture-recapture models,” *J Agric Biol Environ Stat*, vol. 6, no. 2, 2001.

[11] P. Kilambi, E. Ribnick, A. Joshi, O. Masoud, and N. Papanikolopoulos, “Estimating pedestrian counts in groups,” *COMPUT VIS IMAGE UND*, vol. 110, no. 1, pp. 43–59, 2008.

[12] A. Chan, Z. Liang, and N. Vasconcelos, “Privacy preserving crowd monitoring: Counting people without people models or tracking,” in *CVPR*, 2008, pp. 1–7.

[13] R. Jose, N. Otero, S. Izadi, and R. Harper, “Instant places: Using bluetooth for situated interaction in public displays,” *Pervasive Comput.*, vol. 7, no. 4, pp. 52–57, 2008.

[14] F. M. Naini, O. Dousse, P. Thiran, and M. Vetterli, “Population size estimation using a few individuals as agents,” in *ISIT*, 2011.

[15] M. Versichele, T. Neutens, M. Delafontaine, and N. Van de Weghe, “The use of bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the ghent festivities,” *Applied Geography*, 2012.

[16] M. Versichele, T. Neutens, S. Goudeseune, F. Van Bossche, and N. Van de Weghe, “Mobile mapping of sporting event spectators using bluetooth sensors: Tour of flanders 2011,” *Sensors*, 2012.

[17] T. Ellersiek, G. Andrienko, N. Andrienko, D. Hecker, H. Stange, and M. Mueller, “Using bluetooth to track mobility patterns: depicting its potential based on various case studies,” in *ACM SIGSPATIAL*, 2013.

[18] S. Lee and A. Chao, “Estimating population size via sample coverage for closed capture-recapture models,” *Biometrics*, vol. 50, no. 1, pp. pp. 88–97, 1994.

[19] A. Orlitsky, N. Santhanam, and K. Viswanathan, “Population estimation with performance guarantees,” in *ISIT*, 2007.

[20] J. Acharya, A. Orlitsky, and S. Pan, “The maximum likelihood probability of unique-singleton, ternary, and length-7 patterns,” in *ISIT*, 2009, pp. 1135–1139.

[21] G. M. Gemelos and T. Weissman, “On the entropy rate of pattern processes,” *IEEE Trans. Inform. Theory*, vol. 52, 2006.

[22] A. B. Wagner, P. Viswanath, and S. R. Kulkami, “A better good-turing estimator for sequence probabilities,” in *ISIT*, 2007.

[23] M. Kurant, C. Butts, and A. Markopoulou, “Graph size estimation,” *arXiv preprint arXiv:1210.0460*, 2012.

[24] S. Lin, J. Chen, and H. Chao, “Estimation of number of people in crowded scenes using perspective transformation,” *IEEE Trans. Syst., Man, Cybern. A*, vol. 31, no. 6, pp. 645–654, 2001.

[25] A. Davies, J. Yin, and S. Velastin, “Crowd monitoring using image processing,” *Electronics & Comm. Eng.*, vol. 7, 1995.

[26] R. Want, A. Hopper, V. Falcao, and J. Gibbons, “The active badge location system,” *TOIS*, vol. 10, no. 1, 1992.

[27] L. Ni, Y. Liu, Y. Lau, and A. Patil, “Landmarc: indoor location sensing using active RFID,” *Wireless networks*, 2004.

[28] H. Wang, Q. Jia, C. Song, R. Yuan, and X. Guan, “Estimation of occupancy level in indoor environment based on heterogeneous information fusion,” in *CDC*, 2010.

[29] Z.-N. Zhen, Q.-S. Jia, C. Song, and X. Guan, “An indoor localization algorithm for lighting control using RFID,” in *Energy 2030 Conference*, 2008.

[30] M. Nakatsuka, H. Iwatani, and J. Katto, “A study on passive crowd density estimation using wireless sensors,” in *ICMU*, 2008.

[31] N. Patwari and J. Wilson, “Spatial models for human motion-induced signal strength variance on static links,” *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 3, 2011.

[32] M. Youssef, M. Mah, and A. Agrawala, “Challenges: device-free passive localization for wireless environments,” in *MobiCom*, 2007.

[33] J. Weppner and P. ukowicz, “Bluetooth based collaborative crowd density estimation with mobile phones,” in *IEEE PerCom*, 2013.

[34] J. E. Larsen, P. Sapiezynski, A. Stopczynski, M. Mørup, and R. Theodorsen, “Crowds, bluetooth, and rock’n’roll: Understanding music festival participant behavior,” in *Proceedings of the 1st ACM international workshop on Personal data meets distributed multimedia*, 2013.

[35] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*. Springer, 2005.

[36] F. Movahedi Naini, O. Dousse, P. Thiran, and M. Vetterli, “Opportunistic sampling for joint population size and density estimation: Supplementary notes,” 2014. [Online]. Available: http://infoscience.epfl.ch/record/187617

[37] A. Chao and J. Bunge, “Estimating the number of species in a stochastic abundance model,” *Biometrics*, vol. 58, 2002.

[38] J. Wang, “Estimating species richness by a poisson-compound gamma model,” *Biometrika*, vol. 97, no. 3, 2010.

[39] J. Acharya, A. Orlitsky, and S. Pan, “Recent results on pattern maximum likelihood,” in *ITW*. IEEE, 2009, pp. 251–255.

[40] T. Nicolai and H. Kenn, “About the relationship between people and discoverable bluetooth devices in urban environments,” in *Mobility Conference’07*, 2007, pp. 72–78.

[41] E. ONeill, V. Kostakos, T. Kindberg, A. Penn, D. S. Fraser, T. Jones *et al.*, “Instrumenting the city: Developing methods for observing and understanding the digital cityscape,” in *UbiComp*, 2006.

Farid M. Naini received the M.Sc. degree in Communication Systems from the Swiss Federal Institute of Technology at Lausanne (EPFL), Switzerland in 2009, and the Ph.D. degree in Communications and Computer Sciences from the same institution in 2014. His research interests include opportunistic sampling, data mining, and location privacy.

Olivier Dousse received the MSc degree in physics from the Swiss Federal Institute of Technology at Lausanne, Switzerland (EPFL) in 2000, and the PhD degree in communication systems from the same institution in 2005. From 2006 to 2008, he was with the Deutsche Telekom Laboratories in Berlin. He is currently a principal member of Research Staff at the Nokia Research Center in Lausanne. His research interests are in stochastic models for communication networks. He received the honorable mention of the 2005 ACM Doctoral Dissertation Competition and was runner up for the IEEE-Infocom Best Paper Award in 2003. He served as a guest editor of the IEEE Journal on Selected Areas in Communications in 2008 and is serving as an associate editor of the IEEE Transactions on Mobile Computing since 2011.

Patrick Thiran (S’89 - M’96 - SM’12 - F’14) received the electrical engineering degree from the Université Catholique de Louvain, Louvain-la-Neuve, Belgium, in 1989, the M.S. degree in electrical engineering from the University of California at Berkeley, USA, in 1990, and the Ph.D. degree from EPFL, in 1996. He is a Full Professor at EPFL. He became an Adjunct Professor in 1998, an Assistant Professor in 2002, an Associate Professor in 2006 and a Full Professor in 2011. From 2000 to 2001, he was with Sprint Advanced Technology Labs, Burlingame, CA. His research interests include communication networks, performance analysis, dynamical systems, and stochastic models. He is currently active in the analysis and design of wireless and PLC networks, in network monitoring, and in data-driven network science. Dr. Thiran served/s as an Associate Editor for the IEEE Transactions on Circuits and Systems in 1997-99, for the IEEE/ACM Transactions on Networking in 2006-10 and for the IEEE Journal on Selected Areas in Communications since 2014. He was the recipient of the 1996 EPFL Ph.D. award and of the 2008 Crédit Suisse Teaching Award.

Martin Vetterli received the Dipl. El.-Ing. degree from Eidgenossische Technische Hochschule (ETHZ) in 1981, the Master of Science degree from Stanford University in 1982, and the Doctoratès Sciences degree from Ecole Polytechnique Fédérale de Lausanne (EPFL) in 1986. After his dissertation, he was an Assistant and Associate Professor in Electrical Engineering at Columbia University in New York, and in 1993, he became an Associate and then Full Professor at the Department of Electrical Engineering and Computer Sciences at the University of California at Berkeley. In 1995, he joined the EPFL as a Full Professor. He held several positions at EPFL, including Chair of Communication Systems and founding director of the National Competence Center in Research on Mobile Information and Communication systems (NCCR-MICS). From 2004 to 2011 he was Vice President of EPFL for international affairs, and from 2011 to 2012, he was the Dean of the School of Computer and Communications Sciences. Since January 2013 he is President of the National Research Council of the Swiss National Science Foundation. He works in the areas of electrical engineering, computer sciences and applied mathematics. His work covers wavelet theory and applications, image and video compression, self-organized communications systems and sensor networks, as well as fast algorithms, and has led to about 150 journals papers, as well as about 30 patents that led to technology transfer to high-tech companies and the creation of several start-ups. He is the co-author of three textbooks, Wavelets and Subband Coding” (with J. Kovacevic, Prentice-Hall, 1995), ”Signal Processing for Communications” (P. Prandoni, EPFL Press, 2008) and ”Foundations of Signal Processing” (with J. Kovacevic and V. Goyal, Cambridge University Press, 2014). These books are available in open access, and his research group follows the reproducible research philosophy. His work won him numerous prizes, like best paper awards from EURASIP in 1984 and of the IEEE Signal Processing Society in 1991, 1996 and 2006, the Swiss National Latsis Prize in 1996, the SPIE Presidential award in 1999, the IEEE Signal Processing Technical Achievement Award in 2001 and the IEEE Signal Processing Society Award in 2010. He is a Fellow of IEEE, of ACM and EURASIP, was a member of the Swiss Council on Science and Technology (2000-2004), and is a ISI highly cited researcher in engineering