Institutions
Markets
Technologies
IMT
INSTITUTE
FOR ADVANCED
STUDIES
LUCCA

#07
2013

**IMT LUCCA CSA TECHNICAL
REPORT SERIES 07
May 2013**

RA **Computer Science and Applications**

SPoT: Representing the Social, Spatial, and Temporal Dimensions of Human Mobility with a Unifying Framework

Dmytro Karamshuk
Chiara Boldrini
Marco Conti
Andrea Passarella

IMT LUCCA CSA TECHNICAL REPORT SERIES #07/2013

Piazza San Ponziano 6, 55100 Lucca

Research Area
Computer Science and Applications

SPoT: Representing the Social, Spatial, and Temporal Dimensions of Human Mobility with a Unifying Framework

Dmytro Karamshuk
IMT Institute for Advanced Studies Lucca

Chiara Boldrini
IIT-CNR, Pisa

Marco Conti
IIT-CNR, Pisa

Andrea Passarella
IIT-CNR, Pisa

SPoT: Representing the Social, Spatial, and Temporal Dimensions of Human Mobility with a Unifying Framework

Dmytro Karamshuk, Chiara Boldrini, Marco Conti, Andrea Passarella
IIT-CNR, Pisa, Italy

Abstract

Modeling human mobility is crucial in the analysis and simulation of opportunistic networks, where contacts are exploited as opportunities for peer-to-peer message forwarding. The current approach with human mobility modeling has been based on continuously modifying models, trying to embed in them the mobility properties (e.g., visiting patterns to locations or specific distributions of inter-contact times) as they came up from trace analysis. As a consequence, with these models it is difficult, if not impossible, to modify the features of mobility or to control the exact shape of mobility metrics (e.g., modifying the distribution of inter-contact times). For these reasons, in this paper we propose a mobility *framework* rather than a mobility *model*, with the explicit goal of providing a flexible and controllable tool for modeling mathematically and generating simulatively different possible features of human mobility.

Our framework, named SPoT, is able to incorporate the three dimensions – spatial, social, and temporal – of human mobility. The way SPoT does it is by mapping the different social communities of the network into different locations, whose members visit with a configurable temporal pattern. In order to characterize the temporal patterns of user visits to locations and the relative positioning of locations based on their shared users, we analyze the traces of real user movements extracted from three location-based online social networks (Gowalla, Foursquare, and Altergeo). We observe that a Bernoulli process effectively approximates user visits to locations in the vast majority of cases and that locations that share many common users visiting them frequently tend to be located close to each other. In addition, we use these traces to test the flexibility of the framework, and we show that SPoT

is able to accurately reproduce the mobility behavior observed in traces. Finally, relying on the Bernoulli assumption for arrival processes, we provide a throughout mathematical analysis of the controllability of the framework, deriving the conditions under which heavy-tailed and exponentially-tailed aggregate inter-contact times (often observed in real traces) emerge.

Keywords: human mobility, opportunistic networks, complex networks, stochastic processes

---

1. Introduction

Due to the widespread diffusion of personal handheld devices such as smartphones and tablets, emerging wireless ad hoc networks are characterized by high user mobility, which ultimately leads to intermittent connectivity and end-to-end paths that are continuously changing or even lacking at all. Reversing the traditional approach, these potentially disconnected networks benefit from the exploitation of user mobility to bridge disconnected users in the network, and for this reason they are often referred to as *opportunistic networks* [1]. In opportunistic networks messages are routed by the users of the network (which exchange them upon encounters with other users) and are eventually delivered to their destinations. The delay experienced by messages is thus a function of the users' mobility process. In particular, pairwise inter-contact times (i.e., the time intervals between consecutive contacts of a pair of nodes) are very important, since they characterize the temporal distance between two consecutive forwarding opportunities. Inter-contact times are determined by the movement patterns of users: users visiting the same locations will meet more frequently, and their inter-contact time will be shorter. Given the dependence of the delay on inter-contact times, characterizing the inter-contact time is therefore essential for modeling the performance of opportunistic networking protocols.

The first step in modeling human mobility is to understand how users move. Recently, starting from traces of real user movements, there has been a huge research effort in order to characterize the spatio-temporal (i.e, how users travel across locations [2] [3] [4]) and social (i.e., how the nature of a social relationship impacts on, e.g., inter-contact times between two users [5] [6]) properties of human mobility. There is a general agreement that users tend to travel most of the time along short distances while only occasionally following very long paths. In addition, user movements are generally

characterized by a high degree of predictability: users tend to visit the same locations frequently, and to appear at them at about the same time. Less clear is how inter-contact times are characterised. Many hypotheses have been made (about them featuring an exponential distribution [7], a Pareto distribution [5], a Pareto with exponential cut-off distribution [8], a LogNormal distribution [6], etc.), but the problem has yet to be solved. The fact is that inter-contact times are by nature heterogeneous, and trace analysis suggests that a one-distribution-fits-all approach is probably wrong.

Building upon the above findings, the current approach to human mobility modeling has been so far based on trying to incorporate in the model the newest features of mobility properties as they came up from trace analysis. Typically, each model focuses on just a few properties of human mobility. The class of location-based mobility models aims to realistically represent user mobility patterns in space. They are typically concerned with the regular reappearance to a set of preferred locations [9] or with the length of paths travelled by the users [10]. Similarly, there are models mostly focused on the accurate representation of the time-varying behavior of users, often relying on very detailed schedules of human activities [11] [12]. Finally, the class of social-based mobility models aims to exploit the relation between sociality and movements, and to formalize social interactions as the main driver of human movements [13] [14].

The disadvantage of the current approach to modeling human mobility is that the proposed models are intrinsically bound to the current state of the art on trace analysis, and typically need to be redesigned from scratch any time a new discovery is made. In addition, with current mobility models it is typically difficult, if not impossible, to fine tune the mobility properties (e.g., obtaining inter-contact times featuring a probability distribution with controllable parameters). Overall, flexibility and controllability are currently missing from available models of human mobility. *Flexibility* implies allowing for different distributions of mobility properties (e.g., return times to locations or inter-contact times) to be used with the model. The importance of flexibility is twofold. First, it gives the opportunity to evaluate networking protocols in different scenarios, and test their robustness to different mobility behaviors. Second, it allows for changing the model upon new discoveries from trace analysis without the need to start over from clean slate. On the other hand, *controllability* relates to the capability of obtaining a predictable output starting from a given input. This can be done only at a coarse granularity with the majority of available mobility models. For exam-

ple, in social-based mobility, where social relationships determine the shape of inter-contact times, an appropriate configuration can lead to heavy-tailed inter-contact times [13]. However, there is no direct way for quantitatively controlling the parameters characterizing this distribution, and a fine tuning can be attempted only with a trial-and-error approach.

In light of the above discussion, the contribution of this paper is threefold. First, we propose (Section 3) a mobility framework (SPoT – Social, sPatial, and Temporal mobility framework) that incorporates the three dimensions of human mobility, while at the same time being flexible and controllable. SPoT takes as input the social graph representing the social relationships between the users of the network and the stochastic processes characterizing the visiting patterns of users to locations. Based on the input social graph, communities are identified and are assigned to different locations. Thus, people belonging to the same community share a common location where the members of the community meet. Then, users visit these locations over time based on a configurable stochastic process. The proposed framework thus builds a network of users and locations (called *arrival network*), where a link between a generic user i and a location l characterizes the way user i visits location l. Overall, SPoT aims at being as accurate as possible in matching the real behavior of human movements while at the same time being tractable for mathematical analysis. In addition, the fact that it links together the three dimensions of human mobility provides a complete knowledge on the main mobility drivers, which are often exploited by networking protocols for opportunistic networks. For example, SPoT is superior to the direct generation of inter-contact times, since it also provides information on the social structure of the network. Knowing that a user belongs to a specific community can be very helpful when evaluating the performance of, e.g., community detection schemes for opportunistic networks or social-aware forwarding protocols.

The second contribution of this work lies in studying the mobility behavior that emerges in a real trace of human mobility (Section 4), and in using this information to address two open points in our framework, i.e., how to characterize the way users visit locations and how to position meeting places in the considered scenario. To this aim we study three datasets of human self-reported whereabouts records obtained from the online location-based social networks (LBSN) Gowalla [15], Foursquare [16], and Altergeo [17]. LBSN applications, where people can check-in into places (e.g., restaurants, offices) and share their location with friends, have become incredibly popular with

the widespread diffusion of smartphones. From the check-in records we study the time sequences of individual user arrivals at places and reveal that for the majority of user-place pairs, i.e., from 71% to 59% of the pairs depending on the dataset, they are well approximated by a Bernoulli process, for which the intervals between consecutive arrivals feature a geometric distribution. Similarly, we show that the contact sequences between the majority of user pairs, i.e., from 80% to 94% of the total number of pairs depending on the dataset, can be approximated by a Bernoulli process. As we show later in the paper, this finding is important as the Bernoulli process features a number of properties that significantly simplifies the mathematical analysis of the framework. We also use the check-in records to study the correlation between the distances between locations and the number of regular visitors they share. This property, to the best of our knowledge, has not been studied in the literature before. We find that locations that share many common users that visit them frequently tend to be located close to each other. We use this result to realistically position meeting places in the area of the modeling scenario.

The third contribution of the paper lies in showing that the proposed framework is at the same time flexible and controllable. More specifically, in Section 5 we show that SPoT is able to accurately reproduce the features of aggregate inter-contact times observed in the Gowalla dataset. This highlights the fact that the framework can be instantiated to a desired, general mobility configuration by just changing its input parameters. On the other hand, in Section 6 we focus on the controllability of the framework, i.e., on its capability to generate a predictable output. Building upon the results of the analysis of real mobility data, we represent the way users arrive to locations as Bernoulli processes. Then, first we prove that, when the arrival processes are Bernoulli, the contact process between users is also Bernoulli, which is well-aligned with the corresponding results of the data analysis (see Section 4). Finally, we mathematically derive the conditions under which heavy-tailed and exponentially-tailed aggregate inter-contact times emerge starting from simple, but heterogenous, Bernoulli arrival processes for user visits to locations. This advances the knowledge on the dependence between aggregate and pairwise mobility statistics (explored for the first time in [18]) and confirms the main result in [18], i.e., that heterogeneity in pairwise statistics can lead to aggregate statistics that are very distant in distribution.

Please note that in this paper we focus on the ability of SPoT to produce a realistic output in terms of inter-contact times. As discussed above, inter-

contact times are extremely important for the evaluation of opportunistic networks. For this reason, most network simulators, either public platforms [19] or custom simulators [20] [21], are designed to work with contact-based traces as input. Alternatively, especially outside the opportunistic networks domain, network simulators can take as input information about node movements. This spatial output is not the main focus of the paper but, due to its relevance, in Section 7 we discuss how SPoT can be extended for generating a movement-based output. However, we leave the complete evaluation of the properties of this spatial output for future work.

2. Related Work

A comprehensive overview of the state-of-the-art in mobility modeling was presented in [22]. The work points out that the main findings in human mobility research can be classified along the three axes of *spatial*, *temporal*, and *connectivity* (or *social*) properties. Spatial properties pertain to the behavior of users in the physical space (e.g., the distance they travel), temporal properties to the time-varying features of human mobility (e.g., the time users spend at specific locations), connectivity properties to the interactions between users. One of the first significant findings in human mobility, which highlighted the difference between our movements and random motion, was documented by Brockman et al. [3], who analyzed a huge data set of records of banknotes circulation, interpreting them as a proxy of human movements. They showed that travel distances, frequently called *jump size*, of individuals follow a power-law distribution. This fits the intuition that we usually move over short distances, whereas occasionally we take rather long trips. Studying data collected tracing mobile phone users, Gonzalez et al. [2] extended the previous finding showing that the distribution of jumps was power law up to a certain point, after which the decay was exponential. In addition, they showed that individual truncated power-law trajectories co-exist with population-based heterogeneity. Thus, it was shown that the distribution of the radius of gyration - a measure which depicts the characteristic distance traveled by a user - can be approximated by a truncated power-law. This suggests that the majority of people usually travel in close vicinity to their home location, while few of them frequently make long journeys.

As for the temporal properties of human movements, Gonzales et al. [2] detected the tendency of people to return to a previously visited location with a frequency proportional to the ranking in popularity of the location

with respect to other locations. The authors also computed the *return time* probability distribution (probability of returning at time t to a selected place) and concluded that prominent peaks (at 24, 48, 72, ..., hours) capture the tendency of humans to return regularly (on a daily basis) to the location they visited before.

Connectivity properties have been extensively studied in the context of opportunistic networks research. In fact, as we have already discussed, the way users interact and get in touch with each other is crucial for message delivery. In particular, the time between two consecutive contacts of two devices contributes to the overall delay, while the duration of the contact bounds the size of the data that can be exchanged at each encounter. Typically, user interactions are measured through human-carried mobile devices, which are assumed to be proxies of real users. However, despite the great efforts, a consensus hasn't been reached yet on how to exactly characterize in probabilistic terms the connectivity metrics.

From a taxonomy standpoint, the three dimensions of human mobility (spatial, temporal, and social) described above can be mapped into three different approaches to modeling human mobility: *maps of preferred locations*, *personal agendas*, and *social graphs*. The models of the first group account for the properties characterizing the regular reappearance of users at a set of preferred locations. Their general approach is to store the maps (i.e., the sets) of preferred places for each user and to explore them while deciding on the next destination for her walk. The main representatives of this group are SLAW [4] and the model proposed by Song et al. [10]. These models are able to satisfy the main spatial properties of human mobility trajectories, but they do not pay enough attention to the social and temporal aspects of human movement.

The second class of models focuses on reproducing realistic temporal patterns of human mobility explicitly including repeating daily activities in human schedules. The most comprehensive approach of this group is presented in [12]. The model incorporates detailed geographic topology, personal schedules and motion generators defined for more than 30 different types of activities. Although the model gives an extremely thorough representation of human movements in some specific scenarios, it does not explain the main driving forces of human mobility and it is too complex for analytical tractability.

The most recent and most rapidly evolving trend in modeling human mobility is based on incorporating sociality into models, thus considering

human relations as the main driver of individual movements. The main idea is that the destination for the next movement of a user depends on the position of people with whom the user shares social ties. The first models of this class of approaches were CMM [23] and HCMM [13], although others have recently been developed.

A recent work that is orthogonal to the above classifications is the work by Hossmann et al. [24]. They have found that, regardless of the modeling approach to human mobility, the contact graph (i.e., the graph whose vertices are the nodes of the network and the edge weights are given by a combination of contact frequency and aggregate contact duration) generated by most synthetic models differs from that obtained from mobility traces. More specifically, traces tend to generate bridging links (only few strong edges connecting communities) in the contact graph, while synthetic models tend to generate bridging nodes (nodes linked to many other nodes). In addition to this result, Hossmann et al. found that contacts happening outside a community location are typically synchronized. In this paper, we do not consider synchronized meetings, in order to keep the framework mathematically tractable. Due to lack of space, we also do not verify whether bridging links are generated.

With respect to the related literature, SPoT covers and links together all the three dimensions of human mobility using a flexible and controllable framework, which can be instantiated to the desired mobility scenario and which is naturally suited for mathematical analysis. This work is an extended version of our previous paper in [25]. Specifically, here we have added the analysis of three relevant datasets extracted from the location-based online social networks Gowalla, Foursquare, and Altergeo. Results from trace analysis provide a strong case for Bernoulli arrivals, which are then used as the reference assumption in the mathematical analysis of the framework. In addition, we use these datasets to test the flexibility of the SPoT framework, showing that the latter is able to reproduce the mobility behavior observed in traces. With respect to [25], we also extend the mathematical analysis of the framework with the derivation of the settings under which exponentially-tailed aggregate inter-contact times (a case frequently encountered in traces) can be obtained. Finally, here we also discuss how SPoT can be extended for producing movement-based output.

3. The Proposed Mobility Framework

In this section we introduce our SPoT mobility framework, designed around the three main dimensions of human mobility, i.e., social, spatial and temporal (see Figure 1). The social dimension is explicitly captured in the framework by taking a graph of human social relationships as an input parameter. This graph can be any well known graph, such as random graphs [26] or scale-free graphs [26], or it can be extracted from real traces. Then, the framework adds the spatial dimension to the social ties by generating an arrival network, which is a bipartite graph that connects users and meeting places. A link between a user and a meeting place in the arrival network implies that the user visits that place during its movements. We exploit the fact that the structure of communities in the social graph has a significant impact on human mobility, thus we assign users to meeting places such that communities of tightly connected users (*cliques*, in complex network terminology) share meeting places.

In order to add the temporal dimension to the model, we describe the way users visit the meeting places to which they are connected in terms of stochastic point processes [27]. A stochastic point process is a stochastic process that characterizes how events (*arrivals* at location, in our framework) are distributed over time. By sampling from the random variables representing the time between consecutive arrivals, we obtain the time sequences of the visits of a user to a given location. Then, the contact network, i.e., the network describing the contacts between nodes, can be obtained by assuming that two nodes are in contact with each other if they happen to be at the same time in the same meeting place.

3.1. The social and spatial dimensions of human mobility

Social interactions between users have emerged as one of the key factors defining human mobile behavior, because individuals belong to *social communities* and their social ties strongly affect their movement decisions [28] [29]. As anticipated, in our analysis we consider proximity-based communities, i.e., communities whose members share a common meeting place (e.g., offices, bars, apartments). Since all members of the community visit a shared meeting place, it implies that users are socially connected with all other members of the community, and, therefore, form fully connected components (i.e., cliques) in the social graph.

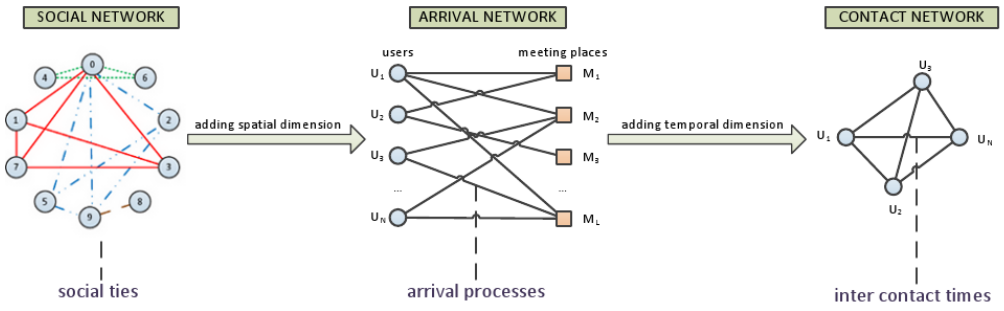

Figure 1: Framework overview

Such cliques in realistic social networks exhibit *overlapping* and *hierarchical structure* [30] [31]. Each user belongs to several overlapping cliques, representing different social circles (e.g., friends, relatives, colleagues). On the other hand, each clique is itself composed of a number of nested cliques, which share additional meeting places that are not common to all the users of a parent clique. For example, a company shares a set of offices visited by all its employees, while each subdivision has its own working places.

As anticipated, we represent the relation between the spatial and the social dimension of human mobility by means of a bipartite graph of users and meeting places, which we call arrival network. In the algorithm (summarized in Table 1) for generating the arrival network starting from the input social graph we mainly need two components: a clique finding algorithm (which also detects overlapping cliques) and a way for reproducing hierarchical cliques.

The first component corresponds to steps 1 and 2 in Table 1. In each round, the social graph is divided into a set (called *cover*) of overlapping cliques, such that each link of the social graph is assigned to exactly one clique. To this purpose, we use the BronKerbosch algorithm [32]. The cover of each round tries to capture the biggest possible cliques. For each of the newly identified cliques, we create a new meeting place and assign all members of the clique to that meeting place. In other words, we create a new meeting place vertex in the arrival network and we add links between this vertex and all members of the community. As an example, we describe in Figure 2 how cliques identified in the social graph are reflected into corresponding meeting places.

The second component (step 3 in Table 1) of the algorithm for generating the arrival network allows us to generate nested cliques. More specifically,

Table 1: Algorithm for building the arrival network - Input: social graph G and removal probability α.

---

1. Divide input social graph G into a set of overlapping cliques, such that the sizes of the cliques are maximum and each link is assigned to exactly one clique. To this aim, the BronKerbosch algorithm [32] can be used.

2. To each clique assign a separate meeting place, i.e., create a new meeting place and a set of links between this place and each member of the clique in the arrival network.

3. Remove randomly each link in the social graph with probability α, inducing emergence of new nested cliques.

4. Proceed to the next round starting from the first step, until there are no links left in the input graph.

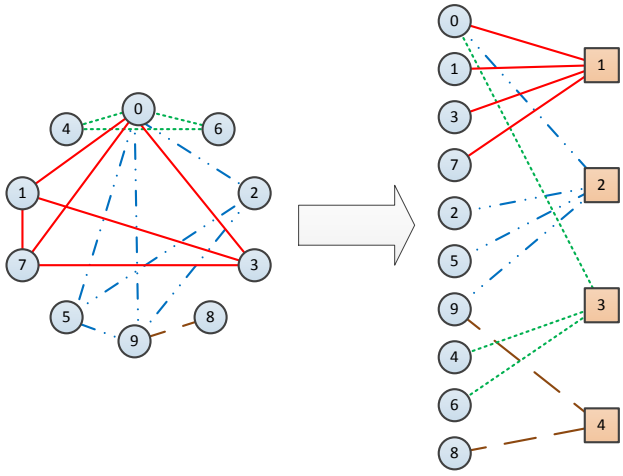

Figure 2: A round of assigning social cliques to meeting place; cliques are marked with different line styles

our algorithm tries to identify cliques of lower size nested into those identified in the previous round. To do so, cliques are split according to a very simple random process, according to which every link in the social graph is randomly removed with a constant, configurable, probability α (*removal probability*).

This leads to the emergence of smaller cliques, which are indeed nested into the original ones. This simple strategy has also the advantage of allowing for a fine control of the number of meeting places shared by users. In fact, each link participates into a geometrically distributed (with parameter α) number of rounds of meeting place assignments. As each link is assigned to at most one clique per round, also the number of cliques that includes that link will be geometrically distributed. This implies that the pair of users i, j with which this link is associated will share a number L_{ij} of cliques (and thus of meeting places) that is itself geometrically distributed with parameter α.

The algorithm for generating the arrival network stops (step 4 in Table 1) when there are no more links in the social graph to be removed.

3.1.1. From meeting places to geographical locations

The analysis of the algorithm in Table 1 reveals that the number of meeting places generated grows with the number of cliques. Thus, the more cliques in the input social graph, the more meeting places are required. The proliferation of meeting places is not of big concern as meeting places might correspond to very small geographic areas (e.g., offices). However, in order to improve the realism of the generated scenario, we combine these meeting places into a fixed number L of wider physical locations (e.g., this is equivalent to combining offices into a business center).

To assign meeting places to geographical locations we explore the observation that, intuitively, the places that share many common frequent visitors should be located geographically close to each other, like in the case of different buildings of a university campus or different offices of a company. In Section 4.3, we validate and confirm this observation using our datasets extracted from location-based online social networks. In order to quantify the closeness between two meeting places, we define the strength F_{ij} of ties between a pair of meeting places i and j as the summary co-appearing frequency across all the users the two places share. More formally, we can write F_{ij} as follows:

$$F_{i,j} = \sum_{u \in \mathcal{U}_{i,j}} f_u^i \times f_u^j \tag{1}$$

where f_u^i is the frequency of user u's visits to location i and $\mathcal{U}_{i,j}$ is the set of all users shared between place i and j. The higher the arrival frequency of user u to both places i and j, the higher the strength between the two places. We anticipate here that the result of the analysis of realistic traces in Section 4.3

suggests that the mean and the median of the distance between two places i and j decreases with the strength F_{ij}. This validate our observation and allows us to exploit it for aggregating meeting places.

Our goal now is to distribute meeting places on the 2D plane such that pairs of places with stronger ties in terms of shared visitors would be located closer. To this aim, we use a variation of the energy model for graph drawing described in [33]. In this model, the places are represented as particles, where particles connected with a link attract each other proportionally to the power of the strength of the link and inversely proportional to the power of the distance between the particles. Similarly, particles that are not connected with a link repulse each other. The final spatial positioning of the meeting places is achieved through simulation, where initial positions of the places are selected randomly in a rectangle of size $w \times h$. As a result of applying attraction and repulsing forces to the nodes, the system eventually reaches an equilibrium state in which tightly connected meeting places are situated close to each other, thus achieving our desired goal.

3.2. The temporal dimension of user visits to meeting places

The arrival network that we have built in the previous section tells us which are the meeting places visited by each user. Here we add the temporal properties of such visits. To this aim, we assign to each link in the arrival network a discrete stochastic point process A_i^l that describes the arrivals of user i to a meeting place l over time. In this work, we consider only discrete point processes, leaving the continuous case for future work. In a discrete point process, the time is slotted. During a time slot, each node visits a set of locations, where this set is determined by the evolution of the arrival processes.

In this paper we assume that processes A_i^l are independent. In real traces, contacts can be synchronized [24], but coordination between nodes may drastically complicate the mathematical analysis of mobility frameworks. For this reason, keeping in mind our target of controllability, we decided to limit the scope of the paper to independent arrival processes. Please note, however, that In Section 5 we show that, despite this simplification, the framework is able to reproduce aggregate inter-contact times observed in the realistic traces.

Once we have characterized the time at which users visit their assigned meeting places, we can build the contact graph of the network (Figure 1). In fact, a contact between two users happens if the two users appear in the

same meeting place at the same time slot. A contact duration is measured as the number of consecutive time slots in which two users have at least one commonly visited location. The contact graph can be fully mathematically characterized (we provide an example of this characterization in Section 6 for the case of arrival processes being heterogenous Bernoulli processes) or it can be obtained from simulations.

4. Analysis of real user movements

As discussed in the previous section, the SPoT framework takes as input the social graph of the network users and the arrival processes describing how users visit locations. While the properties of the user social graph have been extensively studied in the literature [34, 35], thus making their configuration easy, the statistical characterization of user arrivals has been little explored, especially for what concerns the individual user-pair behavior. In order to address this open point in the framework (i.e, which arrival process is best indicated to describe how users visit places in reality), in this section we consider three datasets of real user movements, extracted from the location-based online social networks Gowalla [15], Foursquare [16], and Altergeo [17]. In location-based online social networks, users check-in at places (e.g., restaurants, offices) and share their location with their friends. Thus, the concept of check-ins is very similar to the arrivals considered in the SPoT mobility framework. In fact, both notions represent records of the time at which users visit particular venues. For this reason, we chose to take check-ins as proxies for user arrivals at places and to use them to measure the temporal characteristics of arrival sequences.

In this section we also use the three datasets for studying the features of pairwise inter-contact times in this real scenario. The pairwise results we will be later compared against the mathematical results in Section 6, showing that data and model predictions are totally in agreement. Finally, we also study the geographic distribution of the meeting places sharing common visitors, whose results we used in Section 3.1.1 when defining the algorithm for aggregating meeting places into locations.

4.1. Collecting data

Here we consider the datasets of check-ins collected from the three online location-based social networks Gowalla, Foursquare and Altergeo. Each check-in record is stored as a tuple $\omega = (U, V, T) \in \Omega$ where U represents

the user, V the venue and T the time of the check-in. For a pair of user U_i and place V_l we consider a sequence of check-ins $\Omega_i^l = \{(U, V, T) \in \Omega : U = U_i \text{ and } V = V_l\}$ and denote the number of check-ins in a sequence as n_i^l ($n_i^l = |\Omega_i^l|$). We denote the total number of user-place pairs in the dataset with Q. We use venues as proxies of meeting places without performing any aggregation. Even in case two venues have similar coordinates, we treat them as two different meeting places.

4.1.1. Gowalla

The first dataset used in this paper comprises check-ins of Gowalla users collected via public API [24]. Launched in 2007, Gowalla was a pioneer location-based social network available via mobile app for most of the major platforms (Android, iPhone, etc.). The Gowalla service was bought by Facebook in December 2011 and eventually shut down in 2012. The dataset considered in this paper accounts for $|\Omega_{GO}| = 27$M check-in records collected from $|U_{GO}| = 619$K users at $|V_{GO}| = 2.4$M venues in the period of time from 21 January, 2009 to 7 July, 2011.

4.1.2. Foursquare

Foursquare was launched in 2009 and it has quickly become the most popular location-based service, with more than 35 million users as of January 2013 [16]. Similarly to Gowalla, Foursquare users receive bonuses for check-ins at places. Recently, Foursquare is becoming more and more focused on being a tool for exploring nearby places, e.g., finding restaurant, hotel, nightclub etc. Per user Foursquare check-in data are not directly accessible. However, users can opt to share their check-ins publicly on Twitter. Using the Twitter's streaming API, it was possible to crawl publicly available check-ins [36]. Note that we can only access those check-ins that users explicitly choose to share on Twitter, although users have the possibility to set this option as default. In this paper we consider a dataset of $|\Omega_{FS}| = 23$M check-in records collected from $|U_{FS}| = 494$K users at $|V_{FS}| = 2.3$M venues across the United States in the period of time from 21 January, 2009 to 07 July, 2011.

4.1.3. Altergeo

Altergeo is an alternative online location-based social networking service focused on Russian speaking countries. Launched in 2008, Altergeo has recently reported the audience of 1M+ users, mostly from big cities of Russia and Ukraine [17]. Similarly to Foursquare and Gowalla, the Altergeo service

is available to its users as a check-in app. The service also explores check-in data for personalized food recommendation in a mobile phone app called Gvidi [37]. In the current paper we explore the dataset of $|\Omega_{AG}| = 700$ K check-ins that we collected from $|U_{AG}| = 49$K Altergeo users at $|V_{AG}| = 94$K places in the period of time between 12 February 2010 and 12 February 2012.

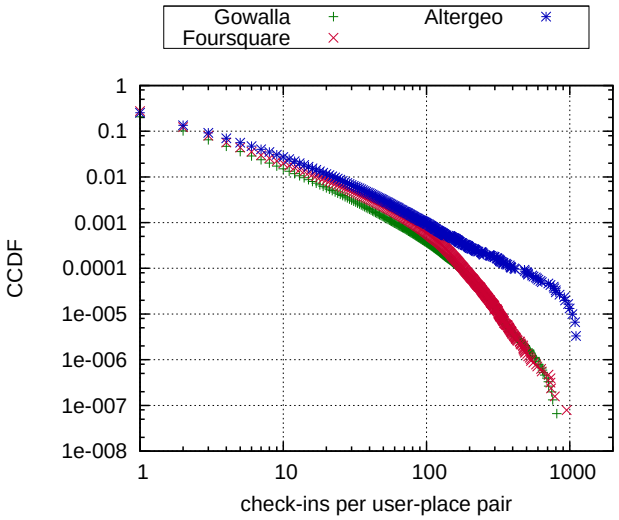

Figure 3: Distribution of number of check-ins n_i^l per user-place pair for the three considered datasets

In Figure 3 we plot the distribution of the number n_i^l of check-ins per user-place pair across all considered datasets. As the plot suggests, the distribution of the number of check-ins for individual user-place pairs is extremely heterogeneous: while 80% of check-ins in places are never repeated, i.e., $n_i^l = 1$, there exist user-place pairs with a number of repeated check-ins higher than 800, i.e., $n_i^l > 800$. In order to deliver a reliable analysis from the statistical standpoint, we discarded those pairs with a small number of check-ins. Thus, we explore the datasets of $Q'_{GO} = 94$K, $Q'_{FS} = 90$K and $Q'_{AG} = 2.4$K user-place pairs each containing at least 20 check-in records, i.e., $n_i^l \geq 20$. We note that the resulting datasets account for $C_{GO} = 46$K, $C_{FS} = 34$K and $C_{AG} = 998$ contact pairs correspondingly. The summary statistics for all three considered datasets are summarized in Table 2.

4.2. Analysis of individual inter-arrival times and inter-contact times

We now describe the methodology that we exploit to characterize the distribution of individual inter-arrival times and individual inter-contact times.

Table 2: Statistics for the three considered datasets

Characteristic	**Gowalla**	**Foursquare**	**Altergeo**
Check-ins, $\|\Omega\|$	27M	23M	0.7M
Users, $\|U\|$	619K	494K	49K
Places, $\|V\|$	2.4M	2.3M	94K
User-Place pairs, Q	15M	13M	0.3M
Arrival sequences, Q'	94K	90K	2.4K
Contact pairs, C	46K	34K	998

From a preliminary analysis we observed that across a significant population of user-place pairs the distribution of inter-arrival times has the shape of a straight line in lin-log scale, which roughly corresponds to a geometric distribution in the discrete case. Similarly, a preliminary observation of the pairwise inter-contact time yielded again a geometric distribution. We aim to validate this hypothesis by fitting individual inter-arrival time and inter-contact time distribution to a geometric distribution and evaluating the goodness of fit across all user-place pairs and user-user pairs, respectively, in the dataset.

The fitting is performed using Maximum Likelihood Estimation [38], which, in the case of a geometric distribution with success probability ρ, yields an estimator $\hat{\rho} = \frac{n_i^l}{\sum_{k=1}^{n_i^l} \tau_k}$ (where $\tau_1, \tau_2, ..., \tau_{n_i^l}$ are the n_l^i observations in the sample). Once we have fitted our data to a geometric distribution, we test whether it is plausible that our data come in fact from such fitted distribution. To this aim, we rely on one of the most popular goodness of fit tests, the *Pearson's chi-squared test* [38], which works well for discrete distributions. In the Pearson's chi-squared test, the test statistic TS is calculated as a sum of differences between observed and expected outcome frequencies (that is, counts of observations), each squared and divided by the expectation:

$$TS = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k} \tag{2}$$

where n is the number of observations, O_k is an observed frequency for a bin k of values, E_k is an expected frequency for a bin k. The test statistic TS follows, approximately, a chi-square distribution with $K = (n - c - 1)$

degrees of freedom (i.e, $TS \sim \chi^2_K$), where n is the number of non-empty bins and c is the number of estimated parameters for the distribution. In the case of a geometric distribution, $c = 1$, thus it follows that $K = n - 2$. If we denote with $q_{\chi^2_K,1-\alpha}$ the $1-\alpha$ quantile of χ^2_K, then the test rejects the geometric hypothesis at level α when $TS > q_{\chi^2_K,1-\alpha}$. In our analysis we set α to 0.001, which corresponds to a 0.001 probability of making a Type I error (i.e., rejecting the hypothesis when it is actually true). To estimate the goodness of fit across the population of individual inter-arrival times and inter-contact times, we calculate the percentage Q^{geom} of the user-place pairs and user-user pairs, respectively, for which the hypothesis of geometric distribution is not rejected.

4.2.1. Characterizing individual inter-arrival times

In Figure 4 we plot the inter-arrival times distribution (blue dots) for three characteristic user-place pairs along with the corresponding fitted geometric distributions (red crosses) as estimated with the methodology described above. In the first two cases the chi-squared test brings no evidence against the assumption of geometric distribution of the inter-arrival times, as the calculated chi-square statistics $TS_{(a)} = 0.09$ and $TS_{(b)} = 12.65$ are smaller than the corresponding quantiles for the chi-square distribution $q_{\chi^2_{K_{(a)}},1-\alpha} = 13.82$ and $q_{\chi^2_{K_{(b)}},1-\alpha} = 22.46$, with $K_{(a)} = 2$ and $K_{(b)} = 6$ degrees of freedom and statistical significance level $\alpha = 0.001$. In opposite, in the latter case the assumption is rejected, since value $TS_{(c)} = 79.93$ is bigger than the corresponding quantile $q_{\chi^2_{K_{(c)}},1-\alpha} = 16.27$ for the chi-square distribution with $K_{(c)} = 3$ degrees of freedom.

We further calculate the percentage of user-place pairs for which the assumption on the geometric distribution of inter-arrival times is not rejected. Thus, we observe that for the majority of pairs across all datasets, i.e., $Q^{geom}_{GO} = 70\%$, $Q^{geom}_{FS} = 79\%$, $Q^{geom}_{AG} = 57\%$, the inter-arrival time distribution follows a geometric distribution. This result is important as a geometric distribution of inter-arrival times can be modeled with a simple Bernoulli arrival process, which, as we discuss in Section 6, is very convenient for mathematical analysis. The implication behind Bernoulli arrivals is that users tend to visit places with a fixed rate. This matches the common finding [9] that users tend be quite regular in their movements.

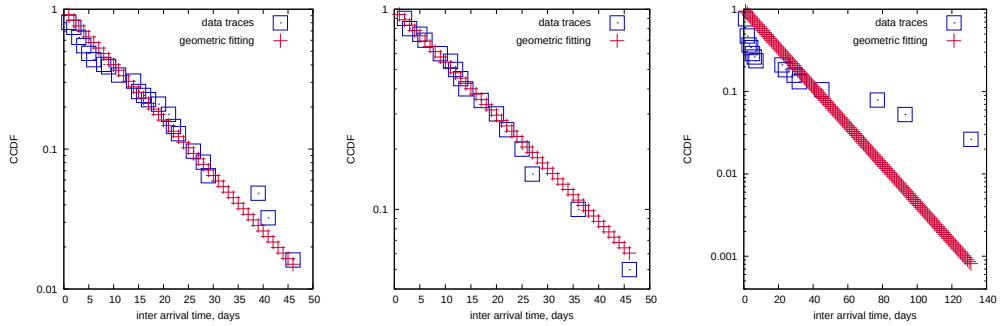

Figure 4: Individual inter-arrival time distribution from traces (blue) vs geometric fitting (red) a,b) for the cases when the assumption of geometric distribution is not rejected c) for the case when that assumption is rejected

4.2.2. Characterizing individual inter-contact times

We now analyze the pairwise inter-contact time sequences measured between consecutive contacts of the users in our datasets. In order to have statistically reliable results, we discarded pairs that have less than 20 contacts. The main obstacle in computing inter-contact times in our datasets is that there are no check-out records, i.e., no records of the time when users leave places. For this reason, we have to make some assumptions about the duration of the sojourn time at a location. In [24], the inter-contact times for the Gowalla trace (the exact same trace that we consider in this work) were measured assuming that a contact between two users happen if they have checked-in less than 1 hour apart at the same place. The rationale for this choice lies behind the nature of location-based online social networks like Gowalla, Foursquare, and Altergeo. In fact, these applications capture mostly users going out for eating or entertainment, for which the 1-hour choice appears reasonable. Thus, also in this work we keep the 1 hour assumption.

The plots for three significant pairs (blue dots) and the corresponding fitted geometric distributions (red crosses) are shown in Figure 5. As we can guess from the plot, in the first two cases the assumption about geometric distribution of the inter-contact times is not rejected. In fact, in this case the chi-square statistics $TS_{(a)} = 17.94$ and $TS_{(b)} = 3.88$ are smaller than the corresponding chi-square distribution's quantiles $q_{\chi^2_{K_{(a)}},1-\alpha} = 24.32$ and

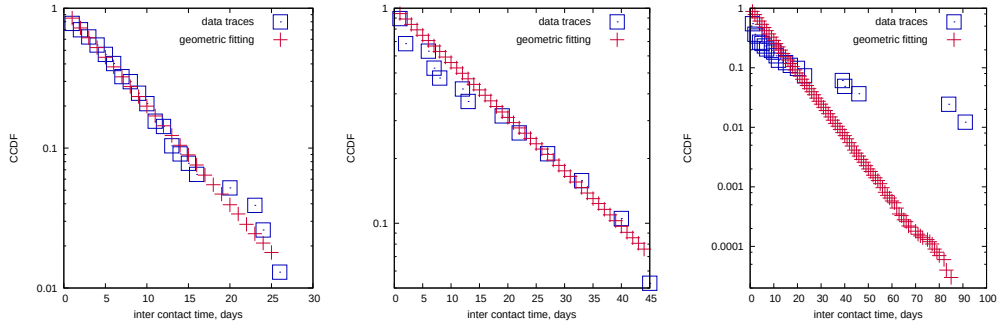

Figure 5: Individual inter-contact times distribution from the data traces (blue) vs geometric fitting (red) a,b) for the cases when the assumption of geometric distribution is not rejected c) for the case when that assumption is rejected

$q_{\chi^2_{K_{(b)}},1-\alpha} = 18.47$, with degrees of freedom $K_{(a)} = 7$ and $K_{(b)} = 4$ and statistical significance $\alpha = 0.001$. In the third case (Figure 5.c), instead, the assumption on the geometric distribution of individual inter-contact times is rejected, since $TS_{(c)} = 104.99$ is bigger than the corresponding quantile $q_{\chi^2_{K_{(c)}},1-\alpha} = 20.52$ from the chi-square distribution with $K_{(c)} = 5$ degrees of freedom. Summarizing, the chi-squared test does not reject the geometric hypothesis for $Q^{geom}_{GO} = 80\%$, $Q^{geom}_{FS} = 94\%$, $Q^{geom}_{AG} = 91\%$ of pairs in our datasets.

4.3. Relative positioning of meeting places sharing common visitors

We now study the relationship between the relative positions of meeting places and the frequency of user visits to those places. More specifically, we investigate whether places that share many common users that visit them frequently happen to be located close to each other. To this end, we measure the social strength between pairs of meeting places in our dataset, exploiting the definition of social strength that we provided in Section 3.1.1. Please recall that the social strength between places i and j measures the co-appearing frequency across all users the two places share, and it is defined as $F_{i,j} = \sum_{u \in \mathcal{U}_{i,j}} f^i_u \times f^j_u$, where f^i_u is the frequency of user u's visits to location i and $\mathcal{U}_{i,j}$ is the set of the users shared between places i and j. Intuitively, the social strength is higher if two places share a lot of common users that frequently visit both places.

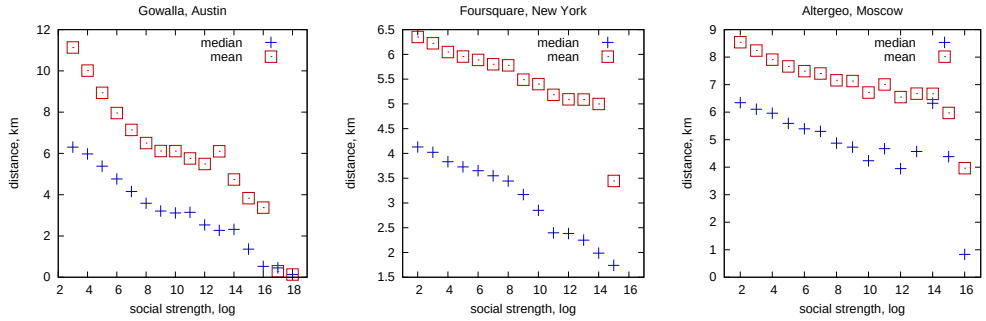

Figure 6: Median and mean values of the geographic distance between venues in the city for different values of social strength $F_{i,j}$ grouped in logarithm bins for Gowalla, Austin (left), Foursquare, New York (center) and Altergeo, Moscow (right) datasets

In Figure 6 we plot the median and mean values of the geographic distance for different values of social strength $F_{i,j}$ between venues in the biggest cities of each dataset, i.e., Austin for Gowalla, New York for Foursquare and Moscow for Altergeo. As we can see from the plot, the distances between places tend to decrease with the strength, therefore suggesting that the places that share a lot of frequent users tend to be located closer to each other. For instance, half of the venues with social strength between 2^3 and 2^4 are situated more than 6km away from each other in Austin and Moscow and more than 4km away in New York, whereas half of pairs of venues with very high social strengths of $2^{15} - 2^{16}$ are placed not farther than $1 - 2$km away from each other across all datasets.

5. Testing the framework flexibility

While in previous sections we have introduced the SPoT framework and we have used real data from location-based online social networks to address the open points in the framework, in this section we start the evaluation of SPoT. More specifically, here we study the flexibility of SPoT, i.e., its capability to reproduce a desired, general, mobility behavior, while in Section 6 we test its controllability.

Our goal in this section is to show that the framework, once configured for the settings observed in a real mobility trace, generate the same aggregate characteristics as those seen in traces. Please note that in the following we

are not validating those aspects that we directly derived from traces (e.g., Bernoulli arrivals). Instead, we aim to evaluate if the proposed generative algorithm based on the creation of the arrival network is able to produce an output that matches the distribution derived from traces.

Here we focus on the aggregate-inter contact times that SPoT generates, which are an important metric often used in the related literature [5, 2]. Aggregate inter-contact times are important for several reasons. First, sometimes the aggregate can be used instead of the pairwise distribution (as suggested by [18]) without loss of accuracy. Second, and more important, sometimes the aggregate distribution may be the only distributions that can be characterised with sufficient statistical confidence in a trace, as too few samples may be available for determining with sufficient statistical confidence each and every individual ICT distribution. Third, the literature has studied aggregates much more than pairwise distributions. So we believe that it is extremely useful to compare with existing results in the literature and to show that our framework is able to generate aggregate ICT distributions similar to those observed in real traces.

In order to use the framework, we need to configure the following quantities: the social graph G, the removal probability α, and the arrival processes A_i^l for each user i visiting a location l. We extract this information from the data traces themselves, relying on the same subset of users (those with at least 20 check-ins) that we have used in Section 4. We take users and friendship records from the dataset to construct the social graph G. In order to estimate the removal probability α from the trace, we recall that this probability is the reciprocal of the average number of places L_{ij} shared between a pair of users i and j, i.e., $\alpha = \frac{1}{E[L_{ij}]}$ (see Section 3 for more details). From the analysis of the traces we compute the sample mean $\hat{E}[L_{ij}^{GO}] = 1.22$ for Gowalla, $\hat{E}[L_{ij}^{FS}] = 1.60$ for Foursquare and $\hat{E}[L_{ij}^{FS}] = 1.64$ for Altergeo. From this we calculate the corresponding removal probabilities $\hat{\alpha}_{GO} = 0.82$, $\hat{\alpha}_{FS} = 0.63$ and $\hat{\alpha}_{AG} = 0.61$. In order to configure arrival processes A_i^l we exploit the result from Section 4 and we set them to be Bernoulli processes. We configure the rates of such processes so that they match the empirical rate distribution derived from the trace (Figure 7). More specifically, the rate distribution in Figure 7 is obtained aggregating the arrival rates for each user-location pair extracted from traces. The use of these rate distributions allows us to maintain the statistical properties of arrival process, regardless of the actual number of user or locations that we actually simulate.

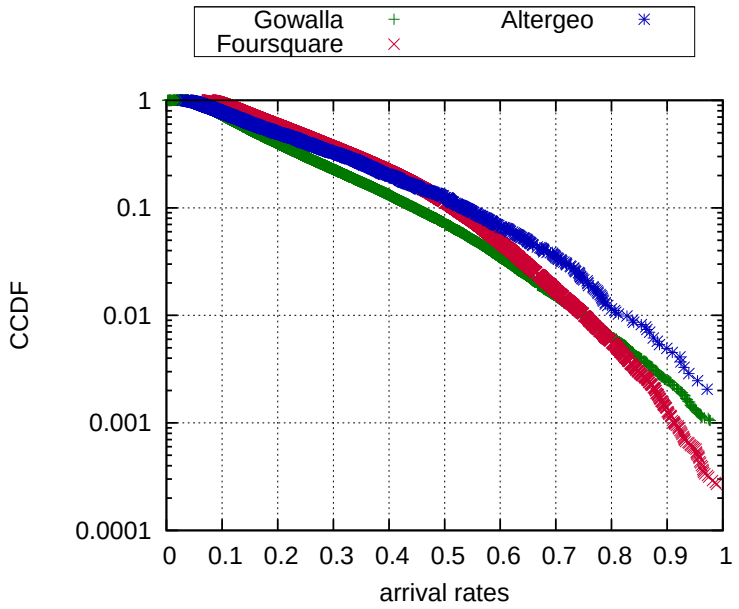

Figure 7: Distribution of arrival rates in the Gowalla, Foursquare and Altergeo traces

In Figure 8 we show the aggregate inter-contact time generated by SPoT against those observed in the traces. As we can see from the plot, the aggregate behavior observed in traces (red squares) is in good agreement with the corresponding results from the simulation (blue crosses). This confirms the flexibility of the framework to capture a desired realistic behavior seen in real traces.

6. Testing the framework controllability

In this section we show mathematically how the SPoT framework is able to produce different, controllable outputs depending on its initial configuration. To this aim, we exploit the data analysis results and we focus on Bernoulli arrivals, which we have shown in Section 4 to represent the behavior of the majority of user-place pairs. Using the Bernoulli assumption, in this section we fully characterize the pairwise dynamics of the framework and we also analytically derive the conditions under which heavy-tailed and exponentially-tailed aggregate inter-contact times, two cases often observed in real traces, emerge.

In our analysis we use the term *contact process* to describe how users

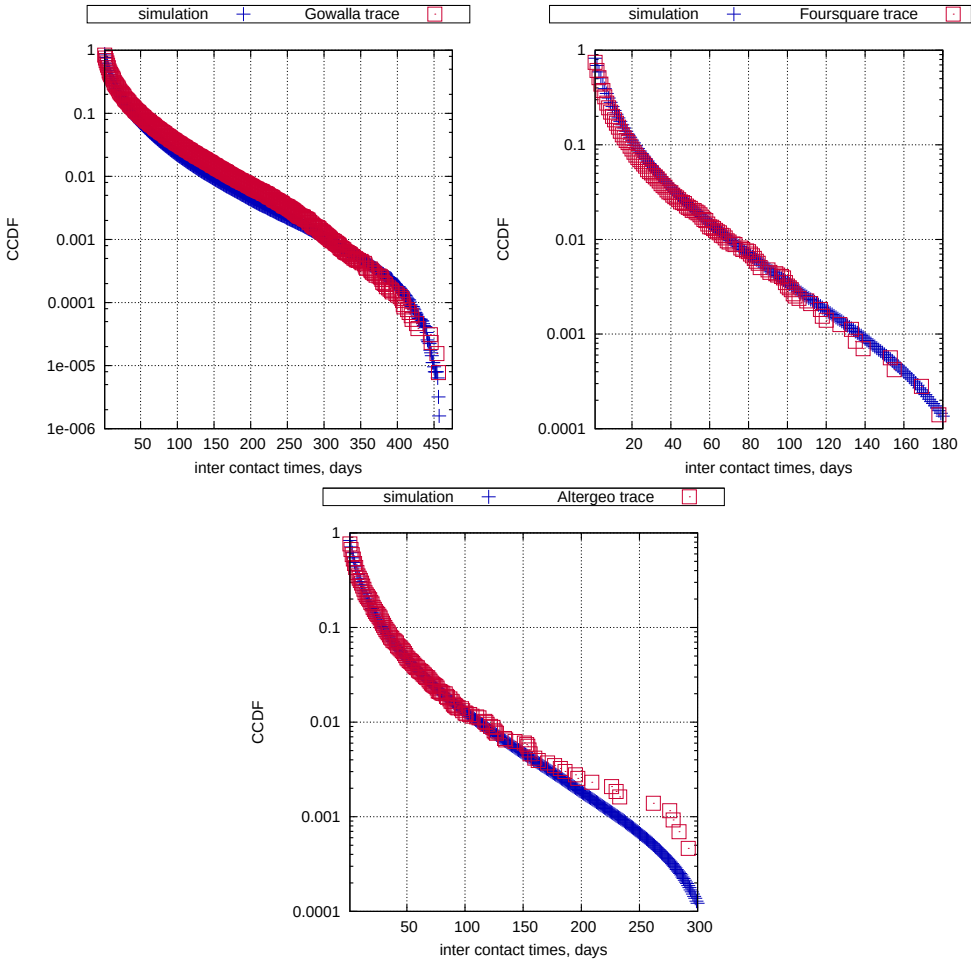

Figure 8: Aggregate inter-contact times obtained from traces (red squares) and from simulations (blue crosses)

meet with each other. Assuming that two users U_i and U_j can meet at L_{ij} distinct meeting places, the contact process between users i and j comprises all contacts happening at all L_{ij} shared meeting places. The time between consecutive contacts in the contact process defines the inter-contact times between the pair of nodes. In the following we also characterize the *single-place contact process*, as the contact process between users U_i and U_j limited to a specific meeting place M_l.

As anticipated, in this analysis we model arrival processes as Bernoulli processes, since they feature geometric inter-arrivals like those seen in traces

(Section 4.2.1). In a Bernoulli process, the probability of an arrival at a given time slot is constant and corresponds to the rate of the process. We show that, if the individual arrival processes are Bernoulli processes, then the contact process and the single-place contact process are also Bernoulli processes for any pair of users. As inter-arrival times for a Bernoulli process feature a geometric distribution, we obtain that from geometric inter-arrival times to specific meeting places (corresponding to Bernoulli arrivals) a geometric distribution of pairwise inter-contact times follows, exactly as seen in traces.

Additionally, we show that the rates of the contact processes depend on the rates of the arrival processes. Starting from this dependence, we are able to derive analytically also the aggregate inter-contact times as a function of the arrival rates of users to meeting places. Although this dependence is not trivial in the general case, we show that different shapes of the aggregate inter-contact distribution can be obtained starting from simple Bernoulli arrival processes. More specifically, we focus on the two cases frequently reported in the related literature, namely, when the the aggregate inter-contact time has a power law or an exponential tail. We show that the latter emerges in homogeneous networks when all the rates of individual Bernoulli processes are equal, and the former when the rates feature a specific distribution.

Before proceeding to the details of our analysis, we first introduce the notation used throughout the section. We consider an arrival network made up of N users and L meeting places. We assume that each user U_i visits place M_l according to a Bernoulli process A_i^l with rate $\rho_{A_i^l}$. For each meeting place M_l and for each pair of users U_i and U_j we characterize the single-place contact process C_{ij}^l (of rate $\rho_{C_{ij}^l}$) and the contact process C_{ij} of rate $\rho_{C_{ij}}$, aggregated over the L_{ij} shared meeting places. The latter defines the distribution of pairwise inter-contact times. We denote the complementary cumulative distribution function (CCDF) of the pairwise inter-contact times of rate ρ with $F_\rho(\tau)$, and that of the aggregate inter-contact times with $F(\tau)$. $F(\tau)$ is obtained as a function of the probability density function (PDF) of the rates of individual inter-contact times $f_P(\rho)$. The notation is summarized in Table 3. The complete proofs for the results shown in this section, when not provided inline, can be found in the appendix.

6.1. Contact process for a pair of users

In this section, assuming Bernoulli arrivals to locations, we analytically characterize the contact process between a pair of users. To this aim, consider two Bernoulli processes A_i^l and A_j^l, describing arrivals of users U_i and U_j in

Table 3: Table of Notation

N	number of users in the arrival network
L	number of meeting places in the arrival network
U_i	user i
M_l	meeting place l
L_{ij}	number of shared meeting places between users U_i and U_j
A_i^l	arrival process of user U_i to meeting place M_l
C_{ij}^l	single-place contact process between users U_i and U_j at meeting place M_l
C_{ij}	contact process between users U_i and U_j
$\rho_{A_i^l}$	rate of arrival process A_i^l
$\rho_{C_{ij}^l}$	rate of single-place contact process C_{ij}^l
$\rho_{C_{ij}}$	rate of contact process C_{ij}
$E[P]$	expectation of the rate of pairwise inter-contact times
$F_\rho(\tau)$	CCDF of individual inter-contact times τ between a pair of nodes whose rate is equal to ρ
$f_P(\rho)$	PDF of the rates of individual inter-contact times
$F(\tau)$	CCDF of the aggregated inter-contact times

a shared place M_l. For a Bernoulli process, the probability $0 < \rho \leq 1$ of an arrival in a time slot τ is constant (i.e., does not depend on τ), and is called the *parameter* or the *rate* of the process. Moreover, time intervals between arrivals are independent geometrically distributed random variables.

We assume that individual arrival processes are independent, and that a contact between two users happens if both decide to visit place M_l in the same time slot. Thus, the single-place contact process C_{ij}^l between user pair U_i, U_j at meeting place M_l can be obtained from the intersection of the individual Bernoulli arrival processes of users U_i and U_j at meeting place M_l. An example of the intersection of individual arrival processes is provided in Figure 9. In the following lemma we prove that the single-place contact process C_{ij}^l is also a Bernoulli point process.

Lemma 1 (Single-place contact process). *The single-place contact process C_{ij}^l resulting from independent Bernoulli arrival processes A_i^l and A_j^l, of rates $\rho_{A_i^l}$ and $\rho_{A_j^l}$ respectively, is a Bernoulli process of rate $\rho_{C_{ij}^l} = \rho_{A_i^l} \times \rho_{A_j^l}$.*

Proof. The probability of a contact at meeting place M_l is equal to the

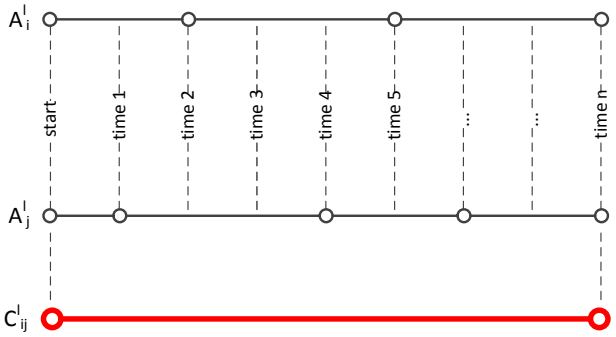

Figure 9: The single-place contact process as an intersection of arrival processes

probability that both users are at meeting place M_l in the same time slot. This can be obtained as the product $\rho_{A_i^l} \times \rho_{A_j^l}$, recalling that, for a Bernoulli process, the rate of the process is equal to the probability of an arrival in a time slot. A discrete stochastic process in which arrivals happen with constant probability $\rho_{C_{ij}^l} = \rho_{A_i^l} \times \rho_{A_j^l}$ is again a Bernoulli process of rate $\rho_{C_{ij}^l}$. □

In the following we focus on the contact process between a pair of users U_i, U_j, i.e., on their contacts in the L_{ij} shared meeting places. A contact happens between the two users in a given time slot if they meet at least in one of the L_{ij} meeting places that they share. Thus, the contact process between users U_i and U_j can be obtained merging (as shown in [27]) their single-place contact processes (Figure 10). In the following theorem we show that if single-place contact processes are Bernoulli, then also the contact process is Bernoulli.

Theorem 1 (Contact process). *The contact process C_{ij} between contacts resulting from a number L_{ij} of individual place contact processes C_{ij}^l, which, in their turn, emerge from Bernoulli arrival processes A_i^l and A_j^l of rates $\rho_{A_i^l}$ and $\rho_{A_j^l}$, is a Bernoulli process of rate $\rho_{C_{ij}} = 1 - \prod_{l=1}^{L_{ij}} (1 - \rho_{A_i^l} \times \rho_{A_j^l})$.*

Proof. The probability of at least one contact in a time slot can be computed as one minus the probability of no contact in that time slot. The probability of no contact in the time slot is equal to the probability that the two users do not meet in any of their shared meeting places. As it follows

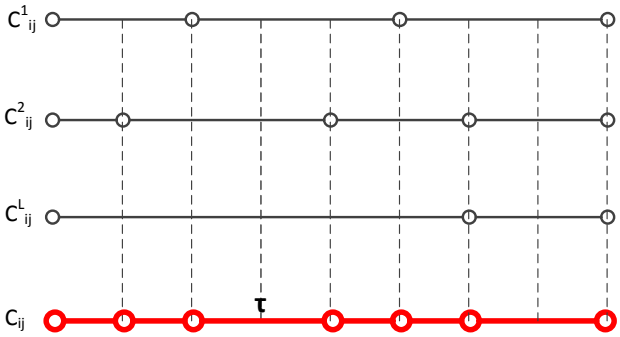

Figure 10: The compound contact process as a merging of single-place contact processes

from Lemma 1, the probability of a contact in a single shared place is constant and equal to $\rho_{A_i^l} \times \rho_{A_j^l}$. Therefore, the probability of at least one contact in a time slot is also constant and equal to $\rho_{C_{ij}} = 1 - \prod_{l=1}^{L_{ij}} (1 - \rho_{A_i^l} \times \rho_{A_j^l})$. It then follows that the sequence of time slots with at least one contact form a Bernoulli process of rate $\rho_{C_{ij}}$. □

The contact process described in Theorem 1 also defines the time intervals between consecutive contacts of a pair of users. Specifically, for a Bernoulli process the distribution of inter-contact times is geometric. We summarize this result in the following corollary.

Corollary 1 (Pairwise inter-contact times). *The inter-contact times distribution between a pair of users U_i and U_j, meeting at a number L_{ij} of meeting places, and whose arrivals to these meeting places are described as Bernoulli arrival processes A_i^l and A_j^l of rates $\rho_{A_i^l}$ and $\rho_{A_j^l}$, is geometric with the following rate:*

$$\rho = 1 - \prod_{l=1}^{L_{ij}} (1 - \rho_{A_i^l} \times \rho_{A_j^l}). \tag{3}$$

Please note that the above result is perfectly in agreement with what we have seen in traces (Section 4).

6.2. Aggregate contact process

In this section we describe how to derive the aggregate inter-contact times starting from pairwise inter-contact times featuring a geometric distribution. More specifically, we solve two cases by providing the conditions on the

Bernoulli arrival processes of users to locations such that the resulting aggregate inter-contact time distribution is either heavy-tailed or exponential. The two cases are important as they have often emerged from the analysis of real mobility traces [5][7]. Our derivation shows how these different aggregate behaviors can result from simple heterogenous Bernoulli arrival processes, which are very convenient to deal with for mathematical analysis. This result also confirms the main finding of [18]: very different aggregate statistics can emerge from the heterogeneity of simple pairwise statistics.

In order to derive the aggregate inter-contact times, we exploit the result in [18], which describes the dependence between the aggregate inter-contact time distribution and the inter-contact time distributions of individual pairs of users. Specifically, the authors consider a heterogeneous scenario, where pairwise inter-contact times distributions are all of the same type (e.g., exponential), but whose parameters (the rates, in the exponential example) are unknown a-priori. The rates of the individual contact sequences are drawn from a given distribution, which, therefore, determines the specific parameters of each pair's inter-contact times. The model described in [18] shows that both the distribution of the rates and the distributions of pairwise inter-contact times impact on the aggregate distribution. For the convenience of the reader we recall this result in Theorem 2.

Theorem 2. *In a network where the rates of pairwise inter-contact times are distributed according to a continuous random variable P with density $f_P(\rho)$, the CCDF of the aggregate inter-contact time is as follows:*

$$F(\tau) = \frac{1}{E[P]} \int_0^\infty \rho f_P(\rho) F_\rho(\tau) \mathrm{d}\rho, \tag{4}$$

where $F_\rho(\tau)$ denotes the CCDF of the inter-contact times between a pair of nodes whose rate is equal to ρ.

Please note that, while originally derived for inter-contact times featuring a continuous distribution, Theorem 2 can be used also for discrete inter-contact times. In fact, the integral in Equation 4 depends on ρ, which was continuous in [18] and it is still continuous here. Thus, discrete inter-contact times do not change the expression for $F(\tau)$, except that now Equation 4 only holds for discrete values of τ.

In Corollary 2, we extend the finding in Theorem 2 to our network of interest, where pairwise inter-contact times depend on their corresponding

arrival processes. We have shown in Corollary 1 that, for the case of independent Bernoulli arrival processes, the distribution of individual inter-contact times is geometric. In other words, the shape of the pairwise inter-contact time distribution $F_\rho(\tau)$ is fixed in our model and, thus, the resulting aggregate inter-contact times characteristic is controlled by the distribution of the rates of individual inter-contact times $f_P(\rho)$. This distribution, in turn, depends on the distribution of the corresponding arrival rates. This dependence may not be trivial in the general case.

In order to apply Theorem 2 to our case of pairwise inter-contact times featuring a geometric distribution, we note that a discrete random variable X featuring a geometric distribution with rate ρ can be expressed in terms of a discrete random variable Y featuring a discrete exponential distribution. More specifically, the CCDF[1] of the geometric distribution of the pairwise inter-contact times, i.e., $F_\rho(\tau) = (1-\rho)^\tau$, $\tau \in \{1, 2, 3, ...\}$, can be re-written in a discrete exponential form, i.e., $F_\lambda(\tau) = e^{-\lambda\tau}$, $\tau \in \{1, 2, 3, ...\}$, by substituting $\rho = 1 - e^{-\lambda}$, where $\lambda \in (0, \infty)$. Variables X and Y are thus exactly the same, but *written in a different form.* Using this substitution, we derive the following corollary of Theorem 2.

Corollary 2. *In a network where pairwise inter-contact times feature a geometric distribution with rate ρ, or, equivalently, a discrete exponential distribution with parameter $\lambda = -\ln(1-\rho)$, the CCDF of the aggregate inter-contact time is given by the following:*

$$F(\tau) = \frac{\int_0^\infty (1-e^{-\lambda})e^{-\lambda\tau} f_\Lambda(\lambda)\mathrm{d}\lambda}{\int_0^\infty (1-e^{-\lambda}) f_\Lambda(\lambda)\mathrm{d}\lambda}. \tag{5}$$

In the above equation, function $f_\Lambda(\lambda)$ denotes the density of the parameters of pairwise inter-contact times.

In the remaining of the section, we show under which arrival rate distribution it is possible to obtain heavy-tailed and exponentially-tailed aggregate inter-contact times, two specific cases frequently reported in the literature.

[1]Please note that the corresponding probability mass function is given by $e^{-\lambda\tau}\left(1-e^{-\lambda}\right)$ and adds up to one, thus showing that the discrete exponential is a properly defined distribution.

6.2.1. Modeling heavy-tailed distribution of aggregate inter-contact times

In this section we study under which arrival rate distribution heavy-tailed aggregate inter-contact times are obtained. To this aim, using Corollary 2, we first derive in Lemma 2 the pairwise contact rate distribution that leads to heavy-tailed aggregate inter-contact times.

Lemma 2. *In a network where pairwise inter-contact times have a discrete exponential distribution of the form* $F_\lambda(\tau) = e^{-\lambda\tau}$, $\tau \in \{1, 2, 3, ...\}$, *and parameters* λ *are drawn from an exponential distribution with rate* a, *the aggregate inter-contact time distribution is as follows:*

$$F(\tau) = \frac{a + a^2}{(\tau + a)(\tau + a + 1)} \quad \left(x \to \infty \Rightarrow F(\tau) \sim 1/\tau^2\right). \tag{6}$$

The complete proof for the above Lemma and for all results introduced below can be found in the Appendix.

Lemma 2 says that the aggregate inter-contact times distribution decays proportionally to the power $\gamma = -2$ of τ, i.e., $F(\tau) \sim 1/\tau^2$, if the distribution of the parameters λ of individual inter-contact times is exponential. In the rest of the section we develop this case and show how the exponential distribution of the parameter of individual inter-contact times emerges in the arrival network with independent Bernoulli arrival processes.

As we have already shown, the distribution of the parameters of pairwise inter-contact times depends on the distribution of the corresponding arrival rates. This dependence is described by Equation 3, which after substitution of $\rho_{C_{ij}}$ with λ, according to what we discussed above, takes the form $\lambda = \sum_{l=1}^{L_{ij}} -\ln(1 - \rho_{A_i^l} \times \rho_{A_j^l})$. From this dependence, we find a distribution of arrival rates $\rho_{A_i^l}$ such that the conditions of Lemma 2 are satisfied, i.e., the distribution of parameters λ of the individual inter-contact times is exponential. To this aim, we prove the following lemma.

Lemma 3. *If individual arrival processes are independent Bernoulli point processes, the rates* $\rho_{A_i^l}$ *of the processes are drawn such that* $\rho_{A_i^l} = e^{-\frac{1}{2}Y^2}$, *where* Y *is a standard normal random variable, and the number of shared meeting places* L_{ij} *between pairs of users is a geometric random variable with parameter* α, *then the resulting pairwise inter-contact times parameters* λ *are exponentially distributed with parameter* α.

A condition for Lemma 3 to be applicable is that the number of shared meeting places between pairs of users is geometrically distributed. Recall that this type of distribution is secured by the arrival network generating algorithm described in Section 3. Therefore, the result of Lemma 3 can be applied to the networks generated by the mobility framework. Finally, we combine the results of Lemma 2 and Lemma 3 in the following theorem.

Theorem 3 (Heavy-tailed aggregate inter-contact times). *If individual arrival processes are independent Bernoulli point processes, the rates $\rho_{A_i^l}$ of the processes are drawn such that $\rho_{A_i^l} = e^{-\frac{1}{2}Y^2}$, where Y is a standard normal random variable, and the number of shared meeting places L_{ij} between pairs of users is a geometric random variable with parameter α, the CCDF of the aggregated inter-contact times is given by Equation 6.*

6.2.2. Modeling exponential distribution of aggregate inter-contact times

In this section we show that the aggregate inter-contact time distribution have an exponential decay if the arrival processes are homogeneous. To this end, we firstly consider the case when the number of shared meeting places L_{ij} between pairs of users is constant and prove that in these conditions the aggregate inter-contact times results in a discrete exponential (i.e., geometric) distribution. Formally, we get the following result (proof can be found in the appendix):

Theorem 4 (Exponential aggregate inter-contact times). *If individual arrival processes are independent Bernoulli point processes with homogeneous rates $\rho_{A_i^l} = \beta$ and the number of shared meeting places L_{ij} between pairs of users is constant, i.e., $L_{ij} = L$, then the aggregated inter-contact times feature a discrete exponential (i.e., geometric) distribution with CCDF:*

$$F(\tau) = e^{-\gamma\tau} \tag{7}$$

where $\gamma = -L\ln(1-\beta^2)$.

In the above case we have shown that the exponential inter-contact time distribution emerges if we put additional constraints on the number of shared meeting places L_{ij}, i.e., we assume that L_{ij} is constant across all pairs of users. Below we consider a more general scenario when the number of shared meeting places L_{ij} between pairs of users is a geometric random variable with parameter α. Recall that this case is secured by the arrival network

generating algorithm described in Section 3. In the following theorem (proof can be found in the appendix) we show that also for this case the aggregated statistics has an exponential decay in the tail of the distribution.

Theorem 5 (Exponentially-tailed aggregate inter-contact time). *If individual arrival processes are independent Bernoulli point processes with homogeneous rates $\rho_{A_i^l} = \beta$ and the number of shared meeting places L_{ij} between pairs of users is a geometric random variable with parameter α, then the CCDF of the aggregated inter-contact times has an exponential tail, i.e.,:*

$$F(\tau) \sim e^{-\delta\tau}, \tau \to \infty \tag{8}$$

where $\delta = -\ln(1-\beta^2)$.

In this section we have studied arrival networks in which links between users and places correspond to Bernoulli processes. We have shown that the pairwise contact sequences in such networks are described by Bernoulli processes, for which the inter-contact times feature a geometric distribution. We have also shown that the rate of the resulting inter-contact times distribution can be derived from the rates of arrival processes. Thus, the pairwise inter-contact times in such networks, firstly, feature a geometric distribution, secondly, have rates distributions controllable by the distribution of arrival rates. As both components, i.e., individual inter-contact times distribution and distribution of inter-contact times rates, have been shown [18] to have impact on the aggregate inter-contact times distribution, we were able to derive different forms of the latter from different distributions of arrival rates.

6.3. Validation

In this section, we support the results obtained above comparing analytical predictions against simulation results. Please note that this validation is needed since Theorems 3 and 5 provide an approximation for the tail of the distribution of inter-contact times, not an exact analytical prediction.

In order to instantiate the proposed framework, we need to define its input parameters: the social graph G, the removal probability α, and the arrival processes A_i^l for each user i visiting a location l. We use the state-of-the-art Barabási-Albert model [39] to generate input social graphs with realistic characteristics (e.g., scale-free degree distribution, short average path length). Thus we consider the two graphs G_{n_1,m_1} and G_{n_2,m_2} of $n_1 = 500$ and $n_2 = 1000$ users and growth parameters $m_1 = 50$ and $m_2 = 30$. The

graph generating algorithm starts with m randomly connected nodes and adds nodes to the network one at a time. Each new node is connected to m existing nodes with a probability that is proportional to the number of links that the existing nodes already have. As a result heavily linked nodes tend to accumulate even more links, while nodes with only a few links are unlikely to attract a lot of new links. This mechanism of "preferential attachment" has been shown to govern the evolution of realistic social networks [39].

We evaluate both graphs G_{n_1,m_1} and G_{n_2,m_2} when the removal probability used by the algorithm for generating the arrival network is $\alpha_1 = 0.5$ and $\alpha_2 = 0.2$. These settings correspond to an average number of locations shared by a pair of users (which are geometrically distributed) equal to $1/\alpha_1 = 2$ and $1/\alpha_2 = 5$, correspondingly. As a result, we obtain four arrival networks with different structural parameters which we explore in simulations. For each of these arrival networks, we study the resulting inter-contact times obtained changing the characteristics of the arrival processes A_i^l of users to meeting places. More specifically, we focus on two cases discussed in the previous sections, namely, when the arrival processes are homogeneous and when the arrival rates features specific distribution that leads to the heavy-tailed aggregate inter-contact times. Simulations are run for 10000 time units of simulated time, and results are shown with a confidence level of 99.9%.

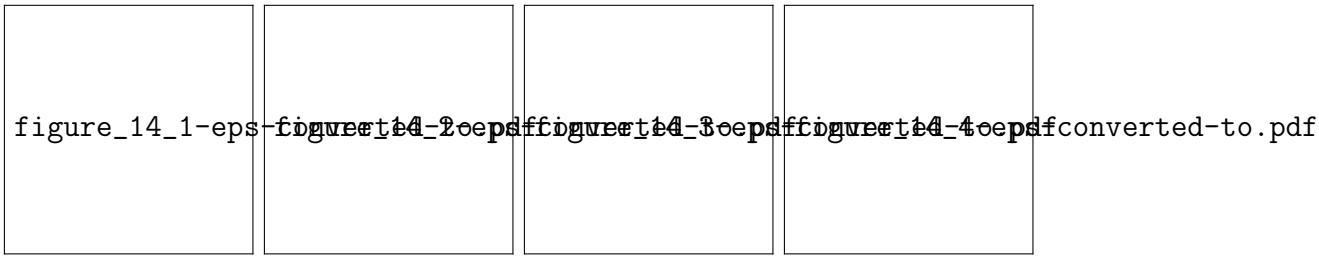

Figure 11: The aggregate inter-contact times distribution for different arrival networks

We assign rates $\rho_{A_i^l}$ of the Bernoulli arrival processes such that $\rho_{A_i^l} = e^{-\frac{1}{2}Y^2}$, where Y is a standard normal random variable. These settings correspond to the case which we mathematically characterized in Section 6.2.1. Figure 11 depicts the result of simulations for each of the arrival networks. For instance, Figure 11.a depicts simulation results for the network with parameters $n = 500$, $m = 50$ and $a = 0.5$. As we can see from the figure, the resulting aggregate inter-contact time CCDF for this network decays as a power law of exponent $\gamma = -2$, i.e., $F(\tau) \sim \tau^{-2}$. In the other arrival net-

works we observe similar results, which are in agreement with the theoretical predictions from Theorem 3.

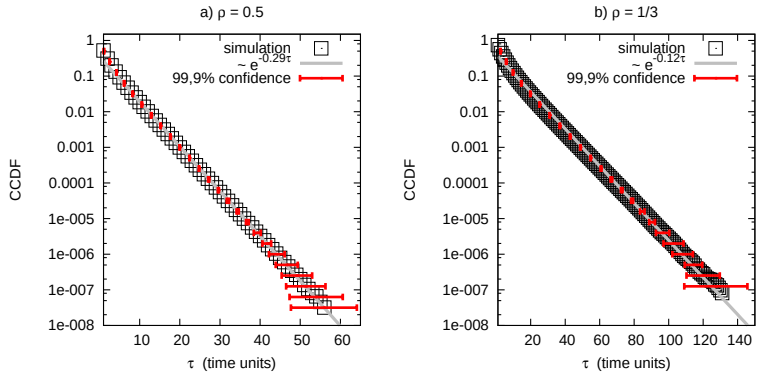

Figure 12: The aggregate inter-contact times distribution for arrival network with identical arrival rates

In the second experiment we simulate arrival networks where arrival processes are Bernoulli processes, like in the first experiment, but this time with identical rates. These settings correspond to the case which we mathematically characterize in Section 6.2.2. More specifically, we model two networks with same parameters $\{n = 500,\ m = 50,\ a = 0.5\}$, in which all the rates of arrival processes are identical and equal to $\rho^{(1)}_{A^l_i} = 1/2$ for the first network, and $\rho^{(2)}_{A^l_i} = 1/3$ for the second. Recall that the rate of the arrival process is the reciprocal of the average of the inter-arrival times. Therefore, the first case corresponds to the network where the average inter-arrival time for all processes is equal to $1/\rho^{(1)}_{A^l_i} = 2$ time units, and the second case to the network with average inter-arrival time for all processes equal to $1/\rho^{(2)}_{A^l_i} = 3$ time units. From Figure 12 we can see that the resulting distribution of the aggregate inter-contact times decays as an exponential function with exponent $\delta_{(1)} = 0.29$ in the first case and $\delta_{(2)} = 0.12$ in the second. This result is in agreement with the theoretical prediction ($\delta = -\ln(1-\rho^2)$, where ρ is the rate of the arrival process) from Theorem 5.

7. Extending SPoT for generating a spatial output

The main focus of the previous sections was on the ability of SPoT to produce a realistic output in terms of inter-contact times. As previously

discussed, inter-contact times are extremely important for the evaluation of opportunistic network and, for this reason, most network simulators (general simulation platforms [19] or custom simulators [20, 21]) are designed to work with contact-based traces as input. Outside the opportunistic networks domain, network simulators [40] often take as input information about nodes' movements instead of (inter-)contact times. In order to make SPoT more general, in this section we discuss how it can be extended for generating a movement-based output. We do not intend to provide an exhaustive analysis of the problem, but just to sketch the main steps for generating a movement-based output. Due to lack of space, we leave the complete evaluation of the properties of this spatial output for future work.

7.1. Generating user trajectories from arrival sequences

In order to obtain a movement-based output, we need to derive trajectories from arrival sequences. To explain the mechanism of transformation we consider a scenario in which a user U_i visits a set of places $\{M_1, M_2, \ldots, M_l\}$ in a time slot T. The order in which user U_i visits individual locations M_j can be defined through a sequence of arrival times $\{T_1, T_2, ..., T_l\}$ where T_j is the time inside time slot T when the user arrives at location M_j. Then, the trajectory of user movements can be reconstructed by connecting places in the order defined by the sequence of arrival times $\{T_j\}$.

Clearly, there are many possible orders in which U_i can visit places $\{M_j\}$ and, therefore, many possible instantiations of the sequence $\{T_j\}$. By design, the SPoT framework assumes that all pairs of users who arrive at time slot T in a place M_j meet with each other. This means that all visitors of M_j should be at M_j during a common time interval. The problem of scheduling the meetings such that everyone attends but also visits other places on their agendas can be transformed into a graph coloring problem [41]. There are numerous algorithms available in the literature to solve graph coloring problems efficiently [42, 43, 44]. Here we consider a graph as composed of meeting places and links between those pairs of places that appear in the agenda of at least one user (see Figure 13). The goal of a graph coloring (or meeting scheduling) algorithm is to assign a color to each vertex, i.e., an arrival time T_j to a place M_j, such that the vertices at the ends of each edge are assigned different colors, i.e., meetings do not overlap in time. In this way, meetings at places with the same colors must be scheduled at the same time, whereas meetings at places with different colors (i.e., sharing

common visitors) must be scheduled at different times. More schematically, the trajectory generating algorithm proceeds as described in Table 4.

Table 4: A step of the trajectory generating algorithm.

1. The pairs of places that are on the agenda for time slot T (an example is shown on the left in Figure 13) of at least one common user are connected with links in the graph of places (on the right in Figure 13).
2. A graph coloring algorithm assigns different colors to vertices that share a common link.
3. Arrival times T_j are assigned to places M_j, such that meeting places with the same color are assigned the same T_j.
4. Individual trajectories are generated by connecting places on individual user agendas according to the order defined by the sequence $\{T_j\}$.

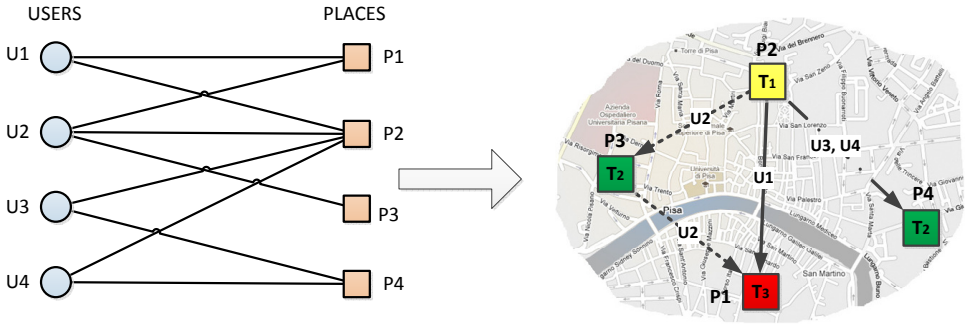

Figure 13: An example of the transformation from arrival sequences generated by the framework to trajectories of movements. The bipartite graph on the left describes users arrivals to places at a time slot T. The resulting trajectories of users are shown on the right with different arrow styles: $P_2 \to P_3 \to P_1$ for user U_2, $P_2 \to P_1$ for user U_1, $P_2 \to P_4$ for users U_3 and U_4.

Please note that the coloring process does not include any notion of "sequence", i.e., taking for example two places with different colors, the coloring algorithm does not tell us anything about whether the first place should be

visited before the other one, or vice versa. One possible option is to preserve the same order of visits across all time slots, thus producing repeating sequences in the way people visit different locations (a property that has been observed in real traces [45]). This can be simply achieved by tagging all meeting places in the arrival network with numeric IDs, and ordering colors at each time slot T in increasing (or decreasing) order of IDs. Clearly, this is just one of the possible ways for assigning visiting times to meeting places, and we leave to future work a more extensive evaluation of the problem.

7.2. Discussion

The realism of the trajectories generated by the approach proposed in this section to a big extent depends on the parameters of the arrival network. For instance, the number of places that a user visits per time slot depends on the number of places he is connected to and on the arrival rates to those places. The former, in turns, depends on the structure of the initial social graph, whereas the latter depends on the distribution of arrival rates $f_P(\rho)$. In general, a user cannot visit all meeting places of the network (unless he is a member of all the cliques identified when running the algorithm for generating the arrival network, event that is extremely unlikely in realistic scenarios), but only a subset. When selecting the locations to be visited in a time slot, each user takes a subset of the set of meeting places he can visit (according to the outcome, e.g., of the Bernoulli selection process). The size of this subset depends on the arrival rates defined for users in the arrival network, since the higher the rates, the higher the number of places selected to be visited in a time slot. We note that the majority of the rates in the real traces which we have considered in Section 5 has small values (i.e., 0.1 for more than 80% of user-place pairs) and, thus, in general, each user visits a small number of places in each time slot.

The fact that users are bound to visit in a time slot all the selected meeting places introduces a consistency problem: at what speed should the users move to visit all these meeting places, and is this speed realistic? There are two main parameters that can be tuned to guarantee realistic user movements: the duration of a time slot T and the size of the scenario considered. As for the latter, it is clear that in a city-wide area visiting multiple locations does not pose great challenges as these multiple locations can be reached, in the worst case, at bus speed in quite a short time (e.g., an hour or so). For larger scenarios (which are typically not considered for opportunistic networks), obtaining realistic movements can be more challenging, and further

investigation is required to address this point. The time slot T can also be helpful. In fact, the larger the time slot, the higher the chances that multiple meeting places can be reached using realistic speed.

8. Conclusion

In this paper we have proposed SPoT, a mobility framework that incorporates the spatial, social, and temporal dimensions of human mobility. The social and spatial dimensions are added imposing that people belonging to the same social community are assigned to the same location, which is where the people of that community meet. Then, the way users visit their assigned locations over time (corresponding to the temporal aspects of mobility) is described by means of a stochastic process.

In order to provide a realistic instantiation of two building blocks of the framework, namely, the arrival process of users to meeting places and the aggregation of meeting places into larger locations, we have analyzed three datasets containing traces of human check-ins at real locations, extracted from the online location-based social networks Gowalla, Foursquare, and Altergeo. The analysis of these datasets has revealed that human arrivals to places can be reasonably approximated, for the majority of user-place pairs, by Bernoulli processes. In addition, we have found that meeting places sharing a lot of common users visiting them with high frequency are typically located close to each other (thus, they should be aggregated).

In the third part of the paper we have focused on the flexibility and controllability of the framework. First we have shown that the SPoT framework can be easily instantiated to accurately reproduce the mobility behavior seen in the Gowalla, Foursquare, and Altergeo traces. Second, as far as the controllability is concerned, we have analytically derived the conditions under which aggregate heavy-tailed and exponentially-tailed inter-contact times emerge, and we have shown that these analytical predictions are totally in agreement with simulation results.

SPoT produces as output a contact-based trace, which can be fed to the vast majority of simulators for opportunistic networks. In the last part of the paper we have discussed how SPoT can be extended to generate a movement-based trace, which can be useful for using SPoT together with simulators such as NS3. In the future we plan to fully investigate the properties and constraints of this spatial output.

Acknowledgements

This work was partially funded by the European Commission under the SCAMPI (FP7-FIRE 258414), RECOGNITION (FET-AWARENESS 257756), EINS (FP7-FIRE 288021), and MOTO (FP7 317959) projects.

References

[1] L. Pelusi, A. Passarella, M. Conti, Opportunistic networking: data forwarding in disconnected mobile ad hoc networks, Communications Magazine, IEEE 44 (11) (2006) 134–141.

[2] M. Gonzalez, C. Hidalgo, A. Barabási, Understanding individual human mobility patterns, Nature 453 (7196) (2008) 779–782.

[3] D. Brockmann, L. Hufnagel, T. Geisel, The scaling laws of human travel, Nature 439 (7075) (2006) 462–465.

[4] C. Song, T. Koren, P. Wang, A. Barabási, Modelling the scaling properties of human mobility, Nature Physics 6 (10) (2010) 818–823.

[5] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, J. Scott, Impact of human mobility on opportunistic forwarding algorithms, IEEE Transactions on Mobile Computing (2007) 606–620.

[6] V. Conan, J. Leguay, T. Friedman, Characterizing pairwise inter-contact patterns in delay tolerant networks, in: Proceedings of the Autonomics'07, 2007.

[7] W. Gao, Q. Li, B. Zhao, G. Cao, Multicasting in delay tolerant networks: a social network perspective, in: MobiHoc '09, ACM, 2009, pp. 299–308.

[8] T. Karagiannis, J. Le Boudec, M. Vojnovic, Power law and exponential decay of intercontact times between mobile devices, Mobile Computing, IEEE Transactions on 9 (10) (2010) 1377–1390.

[9] W. Hsu, T. Spyropoulos, K. Psounis, A. Helmy, Modeling time-variant user mobility in wireless mobile networks, in: Proceedings of IEEE INFOCOM, Citeseer, 2007, pp. 758–766.

[10] K. Lee, S. Hong, S. Kim, I. Rhee, S. Chong, Slaw: A new mobility model for human walks, in: INFOCOM 2009, IEEE, IEEE, 2009, pp. 855–863.

[11] F. Ekman, A. Keränen, J. Karvo, J. Ott, Working day movement model, in: Proceeding of the 1st ACM SIGMOBILE workshop on Mobility models, ACM, 2008, pp. 33–40.

[12] Q. Zheng, X. Hong, J. Liu, D. Cordes, Agenda driven mobility modelling, International Journal of Ad Hoc and Ubiquitous Computing 5 (1) (2010) 22–36.

[13] C. Boldrini, A. Passarella, HCMM: Modelling spatial and temporal properties of human mobility driven by users' social relationshipss, Computer Communications 33 (9) (2010) 1056–1074.

[14] V. Borrel, F. Legendre, M. De Amorim, S. Fdida, Simps: Using sociology for personal mobility, IEEE/ACM Transactions on Networking (TON) 17 (3) (2009) 831–842.

[15] Gowalla.
URL `http://blog.gowalla.com/`

[16] Foursquare.
URL `https://foursquare.com/about`

[17] Altergeo.
URL `http://www.crunchbase.com/company/altergeo`

[18] A. Passarella, M. Conti, Analysis of individual pair and aggregate inter-contact times in heterogeneous opportunistic networks, IEEE Transactions on Mobile Computing [available online] doi: 10.1109/TMC.2012.213.

[19] A. Keränen, J. Ott, T. Kärkkäinen, The ONE Simulator for DTN Protocol Evaluation, in: SIMUTools '09, 2009.

[20] P. Hui, J. Crowcroft, E. Yoneki, Bubble rap: Social-based forwarding in delay-tolerant networks, Mobile Computing, IEEE Transactions on 10 (11) (2011) 1576–1589.

[21] C. Boldrini, M. Conti, A. Passarella, Exploiting users social relations to forward data in opportunistic networks: The hibop solution, Pervasive and Mobile Computing 4 (5) (2008) 633–657.

[22] D. Karamshuk, C. Boldrini, M. Conti, A. Passarella, Human mobility models for opportunistic networks, Communications Magazine, IEEE 49 (12) (2011) 157–165.

[23] M. Musolesi, C. Mascolo, Designing mobility models based on social network theory, ACM SIGMOBILE CCR 11 (3) (2007) 59–70.

[24] T. Hossmann, T. Spyropoulos, F. Legendre, Putting contacts into context: Mobility modeling beyond inter-contact times, in: Twelfth ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 11), ACM, Paris, France, 2011.

[25] D. Karamshuk, C. Boldrini, M. Conti, A. Passarella, An arrival-based framework for human mobility modeling, in: World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2012 IEEE International Symposium on a, IEEE, 2012, pp. 1–9.

[26] M. Newman, The structure and function of complex networks, SIAM review (2003) 167–256.

[27] M. Zukerman, An introduction to queueing theory and stochastic teletraffic models (2007).

[28] J. Silvis, D. Niemeier, R. D'Souza, Social networks and travel behavior: Report from an integrated travel diary, in: 11th International Conference on Travel Behaviour Reserach, Kyoto, 2006.

[29] E. Cho, S. Myers, J. Leskovec, Friendship and mobility: user movement in location-based social networks, in: SIGKDD'11, ACM, 2011, pp. 1082–1090.

[30] M. Newman, M. Girvan, Finding and evaluating community structure in networks, Physical review E 69 (2) (2004) 026113.

[31] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (7043) (2005) 814–818.

[32] C. Bron, J. Kerbosch, Algorithm 457: finding all cliques of an undirected graph, Communications of the ACM 16 (9) (1973) 575–577.

[33] T. M. Fruchterman, E. M. Reingold, Graph drawing by force-directed placement, Software: Practice and experience 21 (11) (1991) 1129–1164.

[34] T. Hossmann, G. Nomikos, T. Spyropoulos, F. Legendre, Collection and analysis of multi-dimensional network data for opportunistic networking research, Computer Communications.

[35] S. Catanese, P. De Meo, E. Ferrara, G. Fiumara, Analyzing the facebook friendship graph.

[36] C. Brown, V. Nicosia, S. Scellato, A. Noulas, C. Mascolo, Where online friends meet: Social communities in location-based networks, in: Proc. Sixth International AAAI Conference on Weblogs and Social Media (ICWSM 2012), Dublin, Ireland, 2012.

[37] Gvidi.
URL `http://gvidi.ru`

[38] E. Lehmann, J. Romano, Testing statistical hypotheses, Springer, 2005.

[39] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Reviews of modern physics 74 (1) (2002) 47.

[40] Ns3.
URL `http://www.nsnam.org/`

[41] T. R. Jensen, B. Toft, Graph coloring problems, Vol. 39, Wiley-Interscience, 1994.

[42] H. Kierstead, A simple competitive graph coloring algorithm, Journal of Combinatorial Theory, Series B 78 (1) (2000) 57–68.

[43] D. Brélaz, New methods to color the vertices of a graph, Communications of the ACM 22 (4) (1979) 251–256.

[44] P. Galinier, J.-K. Hao, Hybrid evolutionary algorithms for graph coloring, Journal of combinatorial optimization 3 (4) (1999) 379–397.

[45] C. Song, Z. Qu, N. Blumm, A.-L. Barabási, Limits of predictability in human mobility, Science 327 (5968) (2010) 1018–1021.

Appendix A. Proofs

Corollary 2. *In a network where pairwise inter-contact times feature a geometric distribution with rate ρ, or, equivalently, a discrete exponential distribution with parameter $\lambda = -\ln(1-\rho)$, the CCDF of the aggregate inter-contact time is given by the following:*

$$F(\tau) = \frac{\int_0^\infty (1-e^{-\lambda})e^{-\lambda\tau} f_\Lambda(\lambda)\mathrm{d}\lambda}{\int_0^\infty (1-e^{-\lambda}) f_\Lambda(\lambda)\mathrm{d}\lambda}. \tag{3}$$

In the above equation, function $f_\Lambda(\lambda)$ denotes the density of the parameters of pairwise inter-contact times.

Proof. The proof is based on adapting Theorem 2 for the case when the pairwise contact sequences are modeled by the corresponding Bernoulli arrival processes. As it follows from Corollary 1, in this case the distribution of individual inter-contact times is geometric, i.e., $F_\rho(\tau) = (1-\rho)^\tau$, $\tau \in \{1,2,3,...\}$. We note that a discrete random variable X featuring a geometric distribution with rate ρ can be expressed in terms of a discrete random variable Y featuring a discrete exponential distribution. More specifically, the CCDF of the geometric distribution of the pairwise inter-contact times, i.e., $F_\rho(\tau) = (1-\rho)^\tau$, $\tau \in \{1,2,3,...\}$, can be re-written in a discrete exponential form, i.e., $F_\lambda(\tau) = e^{-\lambda\tau}$, $\tau \in \{1,2,3,...\}$, by substituting $\rho = 1-e^{-\lambda}$, where $\lambda \in (0,\infty)$. Variables X and Y are thus exactly the same, but written in a different form. Using this substitution the distribution of the pairwise inter-contact times $f_P(\rho)$ in Equation 4 can be rewritten in the form:

$$f_P(\rho) = \frac{\mathrm{d}F_P(\rho)}{\mathrm{d}\rho} = f_\Lambda(\lambda)\frac{\mathrm{d}\lambda}{\mathrm{d}\rho}, \tag{A.1}$$

which follows from the following

$$F_P(\rho) = P(1-e^{-\Lambda} \leq \rho) = F_\Lambda(-\ln(1-p)). \tag{A.2}$$

The expectation $E[P]$ of the rates of the pairwise inter-contact times can be rewritten as:

$$E[P] = \int_0^\infty \rho f_P(\rho)\mathrm{d}\rho = \int_0^\infty (1-e^{-\lambda}) f_\Lambda(\lambda)\mathrm{d}\lambda \tag{A.3}$$

Therefore, after substituting of A.1 and A.3 in Equation 4, Equation 5 follows.

$$F(\tau) = \frac{\int_0^\infty (1 - e^{-\lambda})e^{-\lambda\tau} f_\Lambda(\lambda)\mathrm{d}\lambda}{\int_0^\infty (1 - e^{-\lambda}) f_\Lambda(\lambda)\mathrm{d}\lambda} \tag{A.4}$$

□

Lemma 1. *In a network where pairwise inter-contact times have a discrete exponential distribution of the form $F_\lambda(\tau) = e^{-\lambda\tau}$, $\tau \in \{1, 2, 3, ...\}$, and parameters λ are drawn from an exponential distribution with rate a, the aggregate inter-contact time distribution is as follows:*

$$F(\tau) = \frac{a + a^2}{(\tau + a)(\tau + a + 1)}. \tag{4}$$

Proof. The proof of this lemma exploits the result in Corollary 2. Basically, when the distribution of λ is exponential with parameter a, we have that $f_\Lambda(\lambda) = ae^{-a\lambda}$. Then Equation 6 simply follows from substitution. □

Lemma 2. *If individual arrival processes are independent Bernoulli point processes, the rates $\rho_{A_i^l}$ of the processes are drawn such that $\rho_{A_i^l} = e^{-\frac{1}{2}Y^2}$, where Y is a standard normal random variable, and the number of shared meeting places L_{ij} between pairs of users is a geometric random variable with parameter α, then the resulting pairwise inter-contact times parameters λ are exponentially distributed with parameter α.*

Proof. As it follows from Corollary 1, the rate of the pairwise inter-contact times for the case of Bernoulli arrival processes depends on the rates of the arrival processes as described by Equation 3. By substituting rates ρ of the pairwise inter-contact times with λ parameters, i.e., $\rho = 1 - e^{-\lambda}$, Equation 3 can be rewritten as:

$$\lambda = \sum_{l \in L_{ij}} -\ln(1 - \rho_{A_i^l} \times \rho_{A_j^l}). \tag{A.5}$$

In the following we show that the exponential distribution of λ follows from Equation A.5 if the rates $\rho_{A_i^l}$ of the arrival processes are drawn such that $\rho_{A_i^l} = e^{-\frac{1}{2}Y^2}$, where Y is a standard normal random variable, and the number of shared meeting places L_{ij} between pairs of users is a geometric random

variable with parameter α. To this purpose we, first, analyze random variable X_l, defined as follows:

$$X_l = -\ln(1 - \rho_{A_i^l} \times \rho_{A_j^l}) = -\ln(1 - e^{-\frac{1}{2}({Y_i^l}^2 + {Y_j^l}^2)}) \tag{A.6}$$

In the above equation, Y_i^l and Y_j^l are i.i.d. random variables with a standard normal distribution. Note also that equality $\lambda = \sum_{l \in L_{ij}} X_l$ holds.

In the following, we show that X_l is an exponential random variable with parameter $\beta = 1$. More specifically, if random variables ${Y_i^l}^2$ and ${Y_j^l}^2$ have moment generating function $M_{Y^2}(t)$, then random variable $Z = {Y_i^l}^2 + {Y_j^l}^2$ has moment generating function $M_Z(t) = M_{Y^2}^2(t)$. Particularly, if Y_i^l and Y_j^l are standard normal random variables, then $M_{Y^2}(t) = (1-2t)^{-\frac{1}{2}}$, and, thus, $M_Z(t) = (1-2t)^{-1}$. This corresponds to an exponential random variable Z with parameter $\frac{1}{2}$, i.e., $F_Z(z) = 1 - e^{-\frac{1}{2}z}$. Then the CDF of random variable X_l can be obtained as follows:

$$\begin{aligned} F_{X_l}(x) &= P(-\ln(1 - e^{-\frac{1}{2}Z}) \leq x) = \\ &= P(Z \geq -2\ln(1 - e^x)) = \\ &= 1 - F_Z(-2\ln(1 - e^x)) = 1 - e^{-x}. \end{aligned}$$

Thus, X_l is distributed as an exponential random variable X_l with parameter $\beta = 1$.

To derive the distribution of random variable $\lambda = \sum_{l \in L_{ij}} X_l$, we explore the fact that a random sum of L_{ij} i.i.d. random variables X_l with moment generating function $M_{X_l}(t)$, has moment generating function $M_\Lambda(t) = G_{L_{ij}}(M_{X_l}(t))$, where $G_{L_{ij}}(z)$ is a probability generating function of a discrete random variable L_{ij}. Particularly, if X_l are i.i.d. exponential random variables, i.e., $M_{X_l}(t) = (1 - \frac{t}{\beta})^{-1}$, and L_{ij} is a geometric random variable, i.e., $G_{L_{ij}} = \frac{\alpha z}{1 - z(1-\alpha)}$, then random variable λ has a moment generating function $M_\Lambda(t) = (1 - \frac{t}{\alpha \times \beta})^{-1}$. This corresponds to an exponential random variable λ with parameter $\alpha \times \beta$. In our case $\beta = 1$, therefore, the distribution of λ is exponential with parameter α. □

Theorem 3. *If individual arrival processes are independent Bernoulli point processes, the rates $\rho_{A_i^l}$ of the processes are drawn such that $\rho_{A_i^l} = e^{-\frac{1}{2}Y^2}$, where Y is a standard normal random variable, and the number of shared meeting places L_{ij} between pairs of users is a geometric random variable with parameter α, the CCDF of the aggregated inter-contact times is given by Equation 6.*

PROOF. The theorem combines the results of Lemma 2 and Lemma 3. More specifically, from Lemma 3 it follows that for the case when individual arrival processes are independent Bernoulli point processes, the rates $\rho_{A_i^l}$ of the processes are drawn such that $\rho_{A_i^l} = e^{-\frac{1}{2}Y^2}$, where Y is a standard normal random variable, and the number of shared meeting places L_{ij} between pairs of users is a geometric random variable with parameter α, the resulting pairwise inter-contact times parameters λ are exponentially distributed with parameter α. This allows us to apply Lemma 2 which says that the CCDF of the aggregated inter-contact times in this case is given by Equation 6, then Theorem 3 follows. □

Theorem 4. *If individual arrival processes are independent Bernoulli point processes with homogeneous rates $\rho_{A_i^l} = \beta$ and the number of shared meeting places L_{ij} between pairs of users is constant, i.e., $L_{ij} = L$, then the aggregated inter-contact times feature a discrete exponential (i.e., geometric) distribution with CCDF:*

$$F(\tau) = e^{-\gamma\tau} \tag{5}$$

where $\gamma = -L\ln(1-\beta^2)$.

PROOF. The proof is based on adapting the result of Theorem 2 for the case of homogeneous arrival processes and constant number of meeting places L_{ij}. Recall that Bernoulli arrivals generate inter-contact times that are geometrically (or, equivalently, discrete exponentially) distributed (Corollary 1 and Corollary 2). Below we show that, in this scenario, the parameter λ of the discrete exponentially distributed pairwise inter-contact times is constant and, thus, has a degenerate distribution, i.e.:

$$f(\lambda) = \begin{cases} 1 & \text{if } \lambda = \gamma \\ 0 & \text{otherwise} \end{cases}, \tag{A.7}$$

where $\gamma = -\alpha\ln(1-\beta^2)$. To this aim we explore Equation A.5 that relates λ with individual arrival rates and substitute rates of arrival processes $\rho_{A_i^l}$ and $\rho_{A_j^l}$ with constant β, obtaining:

$$\lambda = -L_{ij} \times \ln(1-\beta^2). \tag{A.8}$$

Particularly, if the number of shared meeting places L_{ij} is constant across all pairs of arrival processes, i.e., $L_{ij} = \alpha$, we get $\lambda = -\alpha\ln(1-\beta^2)$. It then

follows that parameter λ for the pairwise inter-contact times is constant and, thus, has a degenerate distribution as in Equation A.7, where $\gamma = -\alpha \ln(1-\beta^2)$. The proof is concluded by substituting the expressions for λ and $f(\lambda)$ in Equation 5 of Corollary 2 from which Equation 5 follows. □

Theorem 5. *If individual arrival processes are independent Bernoulli point processes with homogeneous rates $\rho_{A_i^l} = \beta$ and the number of shared meeting places L_{ij} between pairs of users is a geometric random variable with parameter α, then the CCDF of the aggregated inter-contact times has an exponential tail, i.e.,:*

$$F(\tau) \sim e^{\delta\tau}, \tau \to \infty \tag{6}$$

where $\delta = -\ln(1-\beta^2)$.

Proof. The proof is similar to the one in the previous theorem: firstly, we derive an expression for parameter λ of pairwise inter-contact times in the considered case; then we apply Equation 5 from Corollary 2 to derive the aggregate inter-contact times characteristic.

Recall from Equation A.8 that for the case of homogeneous arrival processes, parameter λ of the pairwise inter-contact times distribution can be expressed as a product of geometric random variable L_{ij} (with PDF $f_L(l) = (1-\alpha)^{l-1}\alpha$, $l = 1, 2...$) and constant $\delta = -\ln(1-\beta^2)$. Then the PDF of λ can be written as $f_\Lambda(\lambda) = P(\Lambda = \lambda) = P(L_{ij}\delta = \lambda) = f_L(\frac{\lambda}{\delta}) = (1-\alpha)^{\frac{\lambda}{\delta}-1}\alpha$, $\lambda = \delta, 2\delta, ...$ By substituting $f_\Lambda(\lambda)$ in Equation 5 we further obtain the closed-form expression for the aggregate inter-contact times:

$$F(\tau) = \frac{\alpha(e^\delta - 1 + \alpha)e^{\delta\tau}}{(e^{\delta\tau} - 1 + \alpha)(e^{\delta+\delta\tau} - 1 + \alpha)} \tag{A.9}$$

By taking the limit of $F(\tau)$ when τ goes to infinity we obtain Equation 8. This concludes the proof. □

INSTITUTE FOR ADVANCED STUDIES LUCCA

Piazza San ponziano 6, 5100 Lucca, Italy. www.imtlucca.it