

# A Robustness Analysis of Phone-Pair Co-usage Evaluation Methods using Behavioral Modeling

Louise Leibbrandt

Delft University of Technology





# A Robustness Analysis of Phone-Pair Co-usage Evaluation Methods using Behavioral Modeling

by

Louise Leibbrandt

to obtain the degree of Master of Science  
at the Delft University of Technology,  
to be defended publicly on Thursday April 4, 2024 at 8:30 AM.

Thesis advisor: Huijuan Wang  
Daily supervisor: Robbert Fokkink  
External supervisor: Rolf Ypma

*TU Delft*  
*TU Delft*  
*Netherlands Forensic Institute*

Student Number: 4700864  
Programme: MSc Computer Science, Delft  
Project Duration: September, 2023 - April, 2024

# A Robustness Analysis of Phone-Pair Co-usage Evaluation Methods using Behavioral Modeling

## ABSTRACT

In criminal investigations, individuals may be connected to illicit activities by linking their personal phone to an otherwise anonymous, crime-related phone. Several methods have been published that use cell tower registrations to differentiate between same-user and different-user scenarios for the two phones. However, criminals may deviate in movement patterns and phone usage from the test subjects on which the methods are developed and evaluated. Whether the proposed methods are robust to such different behavioral profiles is unclear. The scarcity of readily available datasets on criminals' movements and phone usage further complicates this issue.

Lacking precise knowledge of the behavior of the population of interest, we propose a robustness analysis. Here, we present a tool for generating synthetic datasets, based on well-established models for the movement of individuals. We used the tool to generate data for a range of behavioral properties, encompassing variations in both underlying movement and phone usage. We evaluated three existing methods using our synthetic data. The first is a discriminatory approach that learns typical movement patterns and phone usage from a reference dataset. The second approach uses a model of cell tower behavior, making minimal assumptions on user behavior by choosing pairs of registrations close in time. The third is a generic statistical method for comparing event data. Additionally, we present a fourth method that combines the latter two, as conceptually, they use different aspects of the data.

Our analysis reveals that the discriminatory method performs best in a baseline scenario but is most sensitive to behavioral deviations. The cell tower method shows the lowest baseline performance yet exhibits the strongest resilience to variations. The generic model appears intermediate in terms of performance and sensitivity. Given the importance of robustness in evaluating evidence, we recommend using the combined approach, which is both reliable and effective across our defined variations.

## 1 INTRODUCTION

In criminal investigations, the use of anonymous burner phones and encrypted devices complicates the task of linking suspects to their illicit activities. These devices, designed for anonymity, are chosen by criminals for discreet communication. Despite this, they still produce identifying traces. Notably, Call Detail Records (CDRs), capturing interactions with cell networks, can serve as a proxy for the device's location. Often, criminals also carry a second legitimate phone registered under their name. By comparing CDRs from both the anonymous and registered devices, it is possible to assess *co-location*—phones being in close proximity over a period of time. Such patterns suggest *co-usage*, potentially exposing the anonymous device's user. Several methods have been suggested to evaluate the CDRs under the hypotheses  $H_{su}$ : the phones were carried by the same user,  $H_{du}$ : the phones were carried by independently traveling users.

The spatiotemporal patterns in the CDRs are shaped by two main processes: the behavioral process, which dictates the user's location and phone usage, and the technical telecom process, which determines the cell tower the phone registers to. Although the latter will be very similar for criminal and non-criminal users, the former may be distinct. For instance, individuals with a standard 9-to-5 office job exhibit movement patterns that are markedly different from those of professional hitmen. Likewise, an office worker's routine interactions with their work phone will result in widely different usage patterns than a hitman's use of a burner phone when coordinating an assassination. This issue has been acknowledged but not solved in previous work [1, 2]. One recent publication proposes a solution by only evaluating one pair of cell registrations close in time, arguing that distances observed are then primarily influenced by the technical telecom process rather than the underlying travel and phone usage behaviors [3]. It is currently unclear how well these methods perform if 'real' behavioral patterns deviate from assumed patterns.

To address this, we've developed a simulation-based approach for creating synthetic CDRs to capture underlying behavioral properties. Our method involves two main steps: initially modeling user movements and then generating CDRs for phones traveling along these paths. For modeling user movement, we implement the Exploration and Preferential Return mobility model that characterizes movement as a power-law distributed time spent at a location (waiting time), and a power-law distributed distance between consecutive locations (travel distance), complemented by a return strategy to revisit locations. For phone usage modeling, we describe the inter-arrival times between phone activities as being exponentially distributed, and we characterize various usage dependencies between phones carried by the same user. Connecting cell towers are chosen using an open-source cell tower location database, integrated with a statistical coverage model of the corresponding service area.

We use this tool to generate CDR data across various behavioral scenarios, using our model parameters to guide us. Starting from a basic baseline using general population movement data and assuming that phones carried by the same user are used independently, we outline three experimental scenarios: the first two examine the effects of parameter changes in the mobility model, and the third assesses the same-user phone-usage dependency. Specifically, in the first scenario, we explore the behaviors of individuals moving more dynamically or within a more restricted area by adjusting the mobility model's waiting time and travel distance distributions. In the second scenario, we simulate individuals frequenting many locations by adjusting the model's return strategy. In the final scenario, we investigate the impact of phone usage being dependent on either time or location, e.g., only using a phone at a certain time or location and otherwise using a separate phone.

We evaluate three proposed co-usage likelihood ratio systems using our simulated data. These comprise a discriminative approach,

which requires a reference dataset for parameter estimation; a close pair approach, removing any movement assumptions by investigating one pair of registrations close in time; and a generic method for evaluating event data that assesses global spatial patterns by tallying the categories of registrations. We propose a fourth approach, combining the close pair and categorical count methods.

## 2 BACKGROUND AND RELATED WORK

CDRs are generated by telecom providers, recording antenna registrations triggered by mobile phone activities such as phone calls, text messages, or data sessions. Each record in a CDR includes information like the timestamp, duration, caller and/or callee details, and the identifier of the antenna handling the event. The location of the antenna provides an approximate location of the person's phone at that time. However, challenges such as the various factors influencing which cell a phone connects to mean CDRs provide a proxy for phone location at best [4] and should ideally be handled in a probabilistic manner.

Current approaches for co-usage strength evaluation between phone pairs employ the likelihood ratio (LR) framework, which is widely recognized by the forensic science community [5]. The LR is formulated as

$$\frac{P(E|H_{su})}{P(E|H_{du})},$$

where  $H_{su}$  and  $H_{du}$  denote the hypotheses of same-user and different-user, respectively, while  $E$  represents the evidence: the CDRs. An LR greater than 1 suggests that the evidence supports the same-user hypothesis, and less than 1 supports the different-user hypothesis.

### 2.1 Statistical approaches for co-usage

For our analysis, we investigate three methods. The first two methods, produced by the *Netherlands Forensic Institute*, investigate differences in timing and geographical locations for consecutive phone registrations, termed *switches* [1, 3]. Both methods are designed explicitly for the phone-pair co-usage use case and evaluated on the same validation dataset of employees working at the forensics institute. The third method we investigate, produced by the *University of California*, is more general and can be applied to any form of two sets of user-generated event data [2]. In the case of geolocation data, the authors give the example of Twitter data, where the location of tweets originating from two accounts can be used to determine whether they belong to the same user. This methodology can straightforwardly be applied to the CDR co-usage use case and is therefore included in our analysis.

We outline each method in detail, highlighting their respective strengths and weaknesses.

**2.1.1 Discriminative approach.** The first method we investigate, proposed by Bosma et al. in [1], is completely data-driven. The method extracts consecutive pairs of registrations from differing phones, termed *switches*, from a reference CDR training dataset. For each switch, the method calculates three features:

- (1) the distance;
- (2) the time difference;
- (3) the speed, defined as the distance divided by the time difference.

Every switch is labeled with whether the two phones were from the same or different users. A trained logistic regression model attempts to predict this label from the features. It outputs a score ranging from 0 to 1 for each switch, where scores closer to 1 suggest the same-user case. To aggregate these scores for a larger set of switches found for two phones, the method bins the corresponding individual switch scores into ten bins with 0.1 range, normalizes these counts to ensure their sum equals 1, and inputs these normalized vectors into a second logistic regression model to determine a final similarity score between 0 and 1.

Bosma et al. calculate the final LR by computing scores  $s$  for all same-user and different-user phone pairs in a separate calibration dataset, thus producing an empirical probability density function for both hypotheses. They proceed to apply Kernel Density Estimation with a Gaussian kernel to obtain their final smoothed density functions, and the ratio of these densities,

$$LR = \frac{P(s|H_{su})}{P(s|H_{du})},$$

represents the final LR. For a new pair of phones, a similarity score is derived by applying the two regression models, and based on this score and the two density functions, the LR is calculated.

The main drawback of this approach is that it strongly relies on a reference CDR dataset, implicitly modelling user behaviour. The authors currently suggest carefully assessing the reference data for each case, as behavioral discrepancies may influence applicability.

**2.1.2 Close pair.** Addressing the data dependency limitation of the discriminative approach, Bosma et al. introduced a novel method that aims to remove implicit assumptions about underlying behavioral patterns [3]. This approach works by assessing one pair of registrations originating from the two phones that are close together in time. The idea is that pair registrations with temporal proximity would suggest spatial proximity for two phones traveling together. The observed distance between cell connections is then mainly attributed to network factors, not user movements between registrations. The method proposes a statistical model of an antenna's coverage, specifying the likelihood that two particular cells were registered given that the two phones were in proximity. The same model is also used to determine the likelihood of observing the cell registered by the reference phone, given its typical movement behavior.

The method operates as follows: the two CDRs originating from the reference phone  $a$  and the illicit phone  $b$  are divided into two periods—a short evaluation period, typically 24 hours to represent a day, and a longer reference period comprising all other registrations, denoted as  $R_a$  and  $R_b$ . The method then attempts to find a single well-chosen pair of registrations within the evaluation period. This is defined as a registration,  $c_{a,t}$  from phone  $a$  at time  $t$  and a registration  $c_{b,t+\delta t}$  from phone  $b$  at time  $t + \delta t$ , such that this pair is close in time (within a maximum of 2 minutes) and the location for the illicit phone's registration  $c_{b,t+\delta t}$  is least occurring in  $R_b$ .

The authors then apply a closed-form LR to this pair of registrations. The LR is formulated as

$$LR = \frac{P(c_{a,t}|c_{b,t+\delta t}, l_{a,t} = l_{b,t})}{\frac{1}{|R_a|} \sum_{c_{a,u} \in R_a} P(c_{a,t}|c_{a,u}, l_{a,t} = l_{a,u})},$$



where the actual location of phone  $a$  and  $b$  at time  $t$ ,  $l_{a,t}$  and  $l_{b,t}$ , is explicitly conditioned upon in both numerator and denominator. This method proposes a coverage model for an antenna's service area to estimate these probabilities. We will utilize this same model to sample our antenna locations.

The LR can be interpreted as follows: The numerator models the probability of the two cell registrations given that the devices were in the same location at  $t$ . The closer the registrations are in time and space, the higher this probability should be. The denominator averages over the registrations  $c_{a,u}$  in  $R_a$ . This term becomes larger when the registration  $c_{a,t}$  is close to a cell that has been connected to more often in  $R_a$ , therefore modeling the coincidence of the reference phone  $a$  being in the same location as phone  $b$  at time  $t$ .

The advantage of this method over the discriminative approach is that it relies less heavily on movement and usage time patterns observed in a training data set. This should make the method more robust to a mismatch between the reference population of phone users and the actual population of interest. A clear disadvantage is that it only evaluates a single pair of registrations. This means the method misses much of the information and may, on average, perform worse than the discriminative approach.

**2.1.3 Categorical count.** Longjohn et al. follow a different line of work, in which two sets of user-generated event data are modeled as categorical count vectors [2]. For geolocation data, event categories are defined by a map segmentation, i.e., partitioning a geographic region into smaller sub-regions. Each item in the count vector corresponds to a sub-region, with counts representing the number of events originating from that sub-region. They derive a closed-form LR to evaluate the similarity between two count vectors produced by a known and unknown user.

Longjohn et al. assume that the count vectors originating from  $K$  categories follow a multinomial distribution to reach a closed-form LR. Given that the known-user vector  $r_1 = (r_{11}, r_{12}, \dots, r_{1K})$  is generated by distribution with parameters  $\theta_1 = (\theta_{11}, \theta_{12}, \dots, \theta_{1K})$ , then under the same-user hypothesis,  $H_{su}$ , the unknown-user vector  $r_2 = (r_{21}, r_{22}, \dots, r_{2K})$  follows the same distribution. Under the different-user hypothesis,  $H_{du}$ , this method assumes  $r_2$  again follows a multinomial distribution but with different parameters  $\theta_2 = (\theta_{21}, \theta_{22}, \dots, \theta_{2K})$ . The method treats  $\theta_1$  and  $\theta_2$  as unknown parameters and assumes that these follow a Dirichlet distribution with prior  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_K)$ . The special case where  $\alpha_k = 1$  for  $k = 1, 2, \dots, K$  is known as the uniform Dirichlet distribution and is used by Longjohn et al. in their experimentation as non-informative prior. We will also utilize this formulation in our experimentation.

The closed-form LR is then given by

$$LR = \frac{B(\alpha + r_1 + r_2)B(\alpha)}{B(\alpha + r_1)B(\alpha + r_2)},$$

where  $B(\cdot)$  denotes the multivariate beta function. Intuitively, this LR captures the similarity of the two count vectors while accounting for variations in count sizes. For example, two vectors with high similarity and high counts result in an LR greater than two vectors with high similarity and low counts.

This method is particularly suited for identifying global spatial patterns in cell tower registration locations. However, it does so at the expense of removing temporal information, for instance, not attributing added significance to registrations that occur close

in time. This means the method will probably perform worse for shorter interval data, which may not reflect the typical distribution of locations a user visits. Additionally, the method relies on the assumption that users typically use their phones in similar locations.

## 2.2 Human mobility and CDR generation

We investigate a bottom-up approach to generating synthetic CDRs. We model user movements before generating phone usage patterns and connecting antennas for phones traveling along these paths. This strategy is particularly suited for the phone-pair co-usage use case, as it enables the modeling of CDRs for multiple phones consistent with a single user's path. It also allows for behavioral modeling of both the underlying movement pattern and phone usage behavior.

**2.2.1 Mobility modelling.** We adopt the Exploration and Preferential Return (EPR) model proposed by Song et al. in [6] to model user movement. EPR is recognized as a general mobility framework resulting in paths that reflect broad patterns observed in human mobility rather than individual specific properties [7]. The model distinguishes between two primary behaviors: discovering new locations (exploration) and returning to previously visited locations based on personal preference (preferential return). EPR defines exploration as a random walk process, and the assumption is made that an individual's inclination to explore new places diminishes over time. Therefore, the paths produced by EPR first exhibit random walk properties before displaying the more predictable visitation patterns inherent to human mobility.

Paths generated by the EPR model consist of a series of steps, each characterized by a waiting time and a corresponding location. The model employs two probability density functions:  $\phi(\Delta t)$  for waiting times and  $f(\Delta x)$  for travel distances, with waiting times sampled independently from  $\phi(\Delta t)$  at each step. An exploration probability,  $\rho S^{-\gamma}$ , where  $S$  is the number of unique locations visited and  $\rho$  and  $\gamma$  are constant shape parameters, determines the choice of subsequent locations. With probability  $\rho S^{-\gamma}$ , a new, previously unvisited location is selected, at a distance, sampled from  $f(\Delta x)$ , from the current location. Conversely, with probability  $1 - \rho S^{-\gamma}$ , the next location is a return to one of the previously visited locations, chosen in proportion to the frequency of past visits, reflecting the preference to return towards familiar locations.

Waiting times and travel distances are standard mobility distributions and are widely investigated in human travel data. Analyses done on geolocation datasets, including GPS, CDR, and Dollar bill data, have found the travel distances and waiting times in human mobility follow a power-law distribution [6, 8, 9]. More specifically these findings indicate  $f(\Delta x) \sim |\Delta x|^{-1-\alpha}$  and  $\phi(\Delta t) \sim |\Delta t|^{-1-\beta}$ .

While more sophisticated adaptations of the EPR model exist—such as those that factor in the recency of visited locations [10], or integrate a gravity model to allow *preferential exploration* [11]—these enhancements primarily focus on improving the realism of the generated paths. We argue that EPR equips us with the essential parameters to define a range of movement behaviors for our analysis. By adjusting the underlying parameters, we can simulate a broad spectrum of variations in both time and space scalings while also capturing the predictability characteristic of human movement.

**2.2.2 Generating synthetic CDRs.** To generate synthetic CDRs, our methodology incorporates two primary components: modelling the temporal patterns of phone usage and accurately sampling antenna locations based on phone positioning.

In line with our mobility model, we treat phone usage events as temporally independent, focusing on general patterns rather than individual-specific behaviors. The inter-event times between phone activities are modeled using an exponential distribution, in line with the global traffic patterns identified in large CDR datasets [12]. Due to limited research on behaviors that are associated with using multiple phones simultaneously, we propose dependency models aimed at robustness testing. These models produce datasets that vary in complexity from simple to difficult to differentiate between the two hypotheses, enabling us to test the robustness of co-usage methods without striving to mimic real-life phone usage dependencies.

To sample realistic antenna locations, we use the coverage model implemented by Bosma et al. in [3]. This model was trained on a coverage dataset containing GPS locations and connected antennas in the Netherlands, collected between February and June 2021. The dataset was collected by Police surveillance cars from different regions, carrying prepared phones from varying Dutch telecom providers. This dataset resulted in 4,699 usable pairs of GPS locations  $l$  and corresponding connecting antenna locations  $c$ . The authors model the probability  $p(c|l)$  of connecting to a cell given a location with isotonic regression, taking the distance to the cell tower and the angle between the transmission direction of the antenna and the line connecting the two locations as features. Isotonic is a flexible, non-parametric method that makes the hard assumption that the probability will decrease with increasing distance and angle. Logistic regression is used to reduce the two features to a single one. The authors verified their coverage model and found no signs of miscalibration.

### 3 MATERIALS AND METHODS

To construct the evaluation datasets for our study, we make two main assumptions: first, that each phone’s CDR corresponds to a single individual, and second, that this individual continually carries this phone along a specific path. At a discrete observation time  $t$ , an agent  $a$  is assumed to be at an exact location  $x^a$  forming a path  $P = \{t_i^a, x_i^a\}_{i=1,\dots,n}$  of size  $n$ . We then use this path to generate CDRs, where an activity on phone  $m$  at time  $t$  results in a connection to a cell tower antenna  $c$  at location  $x^c$ , yielding a CDR  $C = \{t_j^m, x_j^c\}_{j=1,\dots,k}$  of length  $k$ . While each agent follows a unique path, multiple distinct CDRs can be derived, representing phones traveling along this route. For our use case, we model users carrying two phones.

The primary objective of this research is to generate CDR data under different scenarios, thereby assessing the robustness of the co-usage methods evaluated. We detail the development of the behavioral modeling and CDR simulation tool in Section 3.1. Our study focuses on three scenarios of user behavior: 1) *dynamic/local*: individuals move more dynamically and/or confined; 2) *many locations*: individuals visit more distinct locations; 3) *dependence*: cell phone usage is dependent on time or location (Section 3.2). In Section 3.3, we describe details about the co-usage methods, including

the three existing methods and a fourth combination approach. To ensure consistent evaluation of the co-usage methods, we evaluate them using the same simulated datasets, CDR preprocessing, and performance metrics (detailed in Section 3.4).

#### 3.1 User behavior model

To generate CDR datasets, we have developed a generic user behavior model and corresponding simulation tool<sup>1</sup>. The model operates in two stages: initially, it applies EPR upon a realistic street network with connecting buildings to generate user location data (see Figure 1a). To create corresponding CDRs for two phones traveling along this route (illustrated in Figure 1b), the second stage implements a phone usage model that describes when users use each of their two mobile phones. To sample connecting antenna locations for these usage times, known antenna locations are combined with a coverage model of their service area.

**3.1.1 User location.** For user location simulation, we extend the Geo-Mesa Agents and Networks framework [13] to incorporate EPR capabilities. We utilize the framework’s interface to include several adjustable parameters. These include the number of agents, the agent moving speed, and the simulation step duration. We also include the EPR-specific parameters, which include:

- Truncated power law parameters for waiting time distribution  $P(\Delta t)$ :  $\beta$ ,  $\Delta t_{\min}$ , and  $\Delta t_{\max}$ ,
- Truncated power law parameters for travel distance distribution  $P(\Delta x)$ :  $\alpha$ ,  $\Delta x_{\min}$ , and  $\Delta x_{\max}$ ,
- Probability of exploration constants:  $\rho$  and  $\gamma$ .

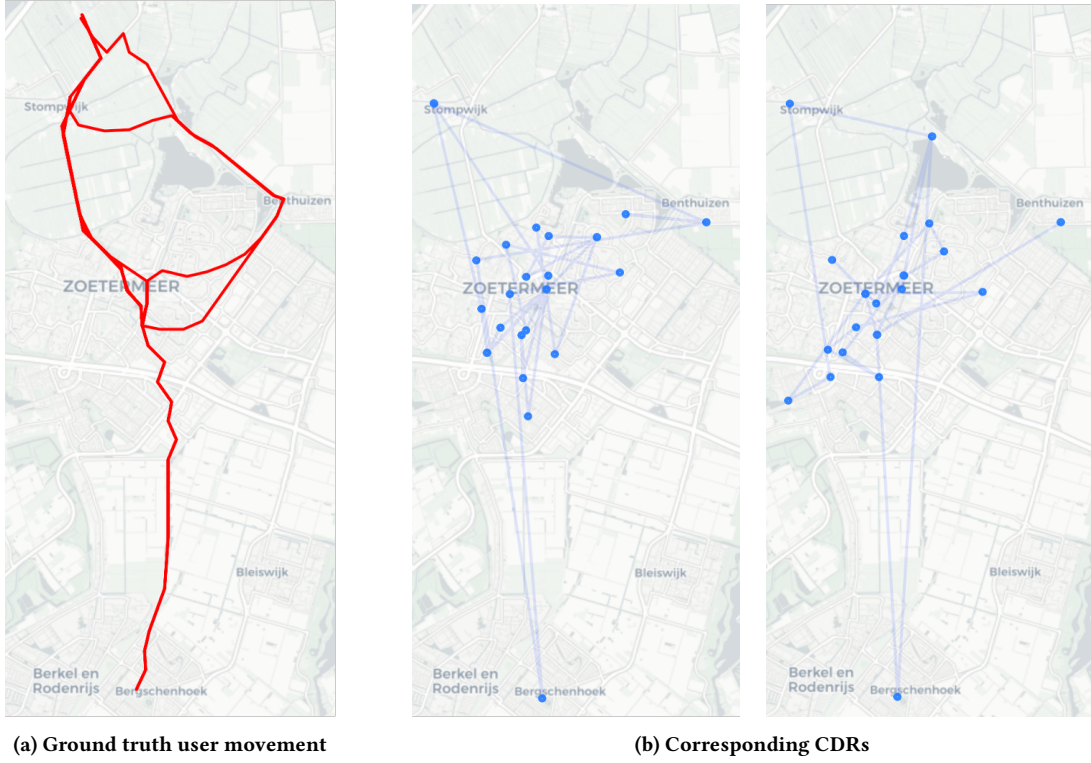
The simulation utilizes data from an open-source geographic database<sup>2</sup> to import a street network and building locations into the Mesa GeoSpace. Initially, each agent is allocated a random building as their starting point and given a set of visited locations and corresponding visitation frequencies  $S(t=0)$ , initialized with their starting location marked with a visitation frequency of 1. Agents receive an initial waiting time sampled from  $P(\Delta t)$ , then proceed with an exploration step to a new location. Once at a new location, agents sample a new waiting time. In subsequent moves, agents choose to explore a new location with probability  $\rho S^{-\gamma}$ , where  $S$  is the number of unique locations visited, or return to a previously visited location with complementary probability  $1 - \rho S^{-\gamma}$ . The exploration and return steps are detailed below. This cycle repeats until the simulation is stopped.

**Exploration step:** Agents select a new destination—a building at a distance sampled from  $P(\Delta x)$  in a random direction from their current position. This location is added to  $S(t)$  with visitation frequency 1. The path to this destination is calculated as the shortest route along the street network, connecting the nearest points on the street to the buildings’ location and segmented according to the agent’s moving speed and simulation step duration.

**Return step:** Agents return to a previously visited location by sampling a location from  $S(t)$  weighted by the corresponding visitation frequencies. The route to this new location is chosen based on the shortest path, and the visitation frequency corresponding to the location in  $S(t)$  is incremented by 1.

<sup>1</sup>[github.com/louiseleibbrandt/mesa-mobility](https://github.com/louiseleibbrandt/mesa-mobility)

<sup>2</sup>[openstreetmap.org](https://openstreetmap.org)



**Figure 1: Simulation is split into two steps: (a) step 1 produces an underlying user movement, (b) step 2 produces CDRs for two phones traveling along this route.**

**3.1.2 CDRs.** In the second stage, we utilize the simulated paths produced by the mobility model to generate CDRs for phones traveling along these routes. We create two CDRs for each path, reflecting the scenario where each user carries two phones. The sampling happens in two phases: first, determining the phone usage times  $\{t_j^m\}_{j=1,\dots,k}$ , closely related to user behavior, and second, modeling the location of the corresponding connecting antenna,  $\{x_j^c\}_{j=1,\dots,k}$ , driven by the technical telecom process of cell tower connectivity.

In the first phase, we model when users use their mobile phones. As each user carries two phones, we implement two sampling strategies modeling whether the phones are used independently or dependently from one another. For both strategies, we assume that the inter-event times between consecutive phone usages follow an exponential distribution with a rate of one per hour. Assuming starting time  $T_0$ , set to the starting time of user location simulation, we define the sampling process for usage times of the two phones,  $t_j^1$  and  $t_j^2$ , sampled at position  $j$  under each dependency type. Under independent sampling, inter-event times ( $\Delta T$ ) for each phone are sampled separately, with usage times defined as follows:

$$t_j^1 = T_0 + \sum_{i=1}^j \Delta T_{i,1},$$

$$t_j^2 = T_0 + \sum_{i=1}^j \Delta T_{i,2},$$

where  $\Delta T_{j,1} \sim \text{Exp}(1)$  and  $\Delta T_{j,2} \sim \text{Exp}(1)$ . Under dependent sampling, rather than sampling inter-event times from a separate distribution per phone, we sample inter-event times from a shared distribution per user. A switch condition  $s(j)$  then dictates the assignment to one of the two phones. This switch condition could, for example, be at a time interval during the day. This would result in one phone being used between certain hours; otherwise, the user uses the other phone. The usage times then follow:


$$\text{if } s(j) \text{ then } t_j^1 = T_0 + \sum_{i=1}^j \Delta T_i,$$

$$\text{otherwise } t_j^2 = T_0 + \sum_{i=1}^j \Delta T_i,$$

where  $\Delta T_j \sim \text{Exp}(1)$ . The resulting usage times sampled in the first phase,  $\{t_j^m\}_{j=1,\dots,k}$ , are mapped to the closest point in time in the underlying movement path, the corresponding locations at these times,  $\{x_j^m\}_{j=1,\dots,k}$ , then represent the locations of the phone at these usage times. In the next phase, we utilize these phone locations to sample connecting cell locations  $\{x_j^c\}_{j=1,\dots,k}$ .

To determine the locations of the connecting antenna, we aim to replicate the complex process of cell connections. We utilize the antenna coverage model detailed in Section 2.2.2, which predicts the probability of connecting to an antenna from a given location. Figure 2 provides an illustrative example of how this coverage model works. We extract antenna locations and azimuth directions



0.1	0.1	0.1	0.1	0.1	0.15	0.2
0.1	0.1	0.1	0.1	0.2	0.35	0.5
0.1	0.1	0.1	0.1	0.5	0.75	0.65
0.1	0.1	0.1		0.9	0.8	0.7
0.1	0.1	0.1	0.1	0.5	0.75	0.65
0.1	0.1	0.1	0.1	0.2	0.35	0.5
0.1	0.1	0.1	0.1	0.1	0.15	0.2

**Figure 2: The coverage model specifies the probability of connecting to an antenna  $c$  from a given location  $x$ :  $p(c|x)$ . It is a basic model that only considers the distance to the cell tower and the azimuth, the angle with its transmission direction. We use this model to sample from possible antennas  $c_i$  by drawing each antenna with probability  $\frac{p(c_j|x)}{\sum_{c_i} p(c_i|x)}$ .**

from a Dutch open-source antenna database<sup>3</sup>. We assume all agents utilize phones connected to the fourth-generation network (4G, also known as LTE) and only extract antennas in the LTE group. We build a coverage model for each of the antennas, and for each phone location  $x_j^m$  in  $\{x_j^m\}_{j=1,\dots,k}$ , we sample a connecting antenna  $c_j$  proportional to

$$\frac{p(c_j|x_j^m)}{\sum_{c_i} p(c_i|x_j^m)}.$$

We use the locations of connecting antennas,  $x_j^c$ , to form our connecting cell locations  $\{x_j^c\}_{j=1,\dots,k}$ .

## 3.2 Defining the scenarios

We take one set of parameters as a baseline scenario. Here, we simulate individuals that move in the Rotterdam-The Hague area and carry two independently used phones. We then adjust the parameters to get at three specific scenarios, examining user movement and phone usage variations. As our goal is to perform a robustness analysis, we purposefully simulate scenarios that are on the extreme side rather than maximally realistic scenarios. For example, in scenario *many locations*, we simulate agents that constantly travel to new places rather than simply returning to more locations than baseline agents.

**3.2.1 Baseline parameters.** For our baseline scenario, we use parameters for waiting time and travel distance distributions estimated by Song et al. using GSM data [6]. This analysis investigated a year-long CDR dataset from 3 million anonymized users and a smaller, two-week GPS study tracking 1,000 users. The waiting

times follow a truncated power-law distribution  $P(\Delta t)$  with parameters  $\beta = 0.8$ ,  $\Delta t_{min} = 20min$ , and  $\Delta t_{max} = 17hrs$ , and the travel distances follow a truncated power law distribution  $P(\Delta x)$  with parameters  $\alpha = 0.55$ ,  $\Delta x_{min} = 1km$ , and  $\Delta x_{max} = 100km$ . Our baseline model assumes a more regular pattern of movement, setting the exploration probability  $\rho S^{-\gamma}$  with  $\rho = 1$  and  $\gamma = 2$ . The registration times of the phones follow a Poisson process, where inter-event times are sampled from an exponential distribution with a rate of one per hour.

**3.2.2 scenario dynamic/local.** For scenario *dynamic/local*, we investigate variations in the underlying mobility distributions used in EPR, encompassing the waiting time  $P(\Delta t)$  and travel distance  $P(\Delta x)$  distributions. We adjust the scale parameters of  $P(\Delta t)$  and  $P(\Delta x)$ , i.e., the  $\Delta t_{min}$ ,  $\Delta t_{max}$ , and  $\Delta x_{min}$ ,  $\Delta x_{max}$ , by a factor of 10 whilst maintaining the shape parameters at  $\beta = 0.8$  and  $\alpha = 0.55$ . We examine the impact of significantly reduced waiting times for the waiting time distribution. The baseline scenario, with  $\Delta t_{min} = 20min$ ,  $\Delta t_{max} = 17hrs$ , depicts our *static* agents, while the adjusted version,  $\Delta t_{min} = 2min$ ,  $\Delta t_{max} = 1.7hrs$ , models our *dynamic* agents. Regarding travel, considering the baseline parameters already accommodate large movements, we test the effects of reduced travel distances: the baseline scenario, with  $\Delta x_{min} = 1km$ ,  $\Delta x_{max} = 100km$ , characterizes our *regional* agents. The adjusted scenario, with distances from  $\Delta x_{min} = 100m$ ,  $\Delta x_{max} = 10km$ , defines our *local* agents. We investigate all agent combinations, i.e., *static local*, *dynamic local*, *static regional* (=baseline scenario), and *dynamic regional*.

**3.2.3 Scenario many locations.** In scenario *many locations*, we simulate individuals with a less predictable movement pattern. We define two agent categories, *returners* with predictable movements and *explorers* with unpredictable movements. The exploration probability in the EPR model is  $\rho S^{-\gamma}$ , where  $S$  is the number of unique locations visited. The *returners* are kept at  $\rho = 1$  and  $\gamma = 2$ , corresponding to our baseline parameters, while *explorers* have  $\rho = 1$  and  $\gamma = 0$ . The choice of parameters ensures that *returners* will converge to an average of roughly 13 visited locations after 40 days of simulation, whereas *explorers* will always remain in a state of exploration.

**3.2.4 Scenario dependence.** Scenario *dependence* examines the impact of a different usage pattern for the two phones: using one phone exclusively during the day or at home and otherwise using the other. Agents are either independent or dependent. The independent case corresponds to our baseline parameters in which we sample usage times independently for the two phones. Under dependent sampling, we define a switch condition to determine which phone is being used by the user. We examine two types of switch functions: time-based and location-based. The time-based switch occurs at 09:00 and again at 17:00; the user uses phone one during the hours 09:00-17:00 and phone two from 17:00-9:00. The location-based switch occurs when the agent enters or exists within a 500-meter radius of their home location; using phone one within a 500-meter radius of the home and phone two outside of this radius. To assign home locations to each agent, we assign a random location with a visitation frequency of 10 at the start of the simulation, ensuring regular return visits.

<sup>3</sup>antenneregister.nl

### 3.3 Defining the methods

We evaluate four co-usage methods, the three existing approaches detailed in Section 2.1 and a fourth combination approach, combining the LRs from the close pair and categorical count methods. We provide the implementation-specific decisions for each method in the subsequent subsections.

**3.3.1 Discriminative.** The discriminative approach requires a training dataset for fitting model parameters. To obtain these, we simulate separate CDR datasets with parameters equal to those used to generate the evaluation datasets. This results in 7 separate training datasets of equal size and behaviour parameters as those in the evaluation datasets described in Section 3.2. We train models on each unique training dataset, followed by assessments across all evaluation datasets. This process allows us to examine the effect of performance when behavioural patterns in the underlying training dataset align or misalign with those in the evaluation dataset used for model validation.

**3.3.2 Close pair.** The close pair method requires a statistical model of the service area of an antenna. We utilize the coverage model detailed in Section 2.2.2, from [3].

**3.3.3 Count.** The categorical count method requires a geographic segmentation to define event categories. We use an open source dataset<sup>4</sup> that partitions the Netherlands based on the first four digits of the Dutch six-character postal codes. We limit these to the postal areas covering our bounding box, resulting in 620 partitions. We assume a non-informative (symmetric) prior,  $\alpha = (1, 1, \dots, 1)$ , assuming categories are equally likely to occur. As this method provides unrealistically large LRs [2], we bound the LRs using the input dataset size as a cut-off. We set  $n$  to the smaller size of the two input CDR vectors, bounding the LR to lower bound  $1/n$  and upper bound  $n$ . This simple bounding improves performance over the unbounded method on all generated evaluation datasets (see Appendix A).

**3.3.4 Close pair x count.** We propose a fourth method that combines the likelihood ratios produced by the close pair and categorical count methods through multiplication. The rationale behind this approach is that these methods provide (nearly) independent assessments; the close pair method considers the evidence contained in a pair of registrations occurring close in time, whilst the categorical count considers the broad spatial patterns in all registrations whilst ignoring the temporal dimension. The two methods will not be fully independent, e.g. because the former estimates the rarity of an antenna registration based on the overall spatial pattern of that phone. Thus, it is interesting to see if the combined method performs better than the separate approaches.

### 3.4 Evaluating the methods

We use the same metric to evaluate all methods using the same generated datasets and corresponding phone pair combinations. Details of this process are described below. We follow the terminology of [1], referring to all phone registrations in a 24-hour period as a *track*, and combinations of tracks as *track pairs*.

**3.4.1 Simulation parameters.** For all simulated evaluation datasets, we simulate 100 agents within a bounding box encompassing Rotterdam and The Hague (coordinates: 4.2009, 51.8561 to 4.5978, 52.1149). The simulations last 40 days, but we analyze only the last 30 to minimize any initial start-up biases. Agents move at 14 m/s (roughly 50 km/h), and user location is recorded every 60 seconds.

**3.4.2 CDR preprocessing.** To evaluate the chosen phone-pair co-usage methods, we utilize telcell<sup>5</sup>, a collection of scripts developed by the *Netherlands Forensic Institute*. This code base provides a pipeline for evaluation. We split the antenna registrations into 24-hour intervals called *tracks* for all evaluation datasets. Tracks are matched from the same date to form the track pairs, with same-user track pairs matched if they originate from the same user. For the different-user track pairs, we match phones originating from different users, and we sample a subset of track pairs matching the size of same-user pairs. This sampled subset is utilized across all methods to ensure uniform conditions.

**3.4.3 Measuring performance.** To evaluate the performance of the aforementioned co-usage assessment methods, we utilize the log-likelihood ratio cost ( $C_{llr}$ ), as recommended for forensic evidence analysis [14]. This metric assesses both the discrimination and calibration of the method—it penalizes misleading LRs and penalizes them more strongly when they deviate further from 1. Given a set of same and different user observations,  $O_{su}$  and  $O_{du}$ , sampled under hypotheses  $H_{su}$  and  $H_{du}$  respectively, and with size  $N_{su}$  and  $N_{du}$ , the  $C_{llr}$  is defined as

$$C_{llr}(O_{su}, O_{du}) = \frac{1}{2} \left( \frac{1}{N_{su}} \sum_{o \in O_{su}} \log_2 \left( 1 + \frac{1}{LR(o)} \right) + \frac{1}{N_{du}} \sum_{o \in O_{du}} \log_2 (1 + LR(o)) \right),$$

where  $LR(o)$  is the LR produced by the system for observation  $o$ . As the  $C_{llr}$  is a cost function, a lower score indicates higher performance. A perfect system would result in a  $C_{llr}$  of 0, whereas a non-informative system, such as one always returning  $LR = 1$ , would yield a  $C_{llr}$  of 1. Scores below 1 signify that the LR system will improve decision-making on average; above 1, it will make decisions worse on average [15].

## 4 RESULTS

We first show results for the baseline (Section 4.1) and three scenarios (Section 4.2) and conclude with a comprehensive analysis combining all datasets to analyze general model trends in Section 4.3. See Appendix B for an overview of simulated dataset sizes and global dataset metrics for the baseline and three scenarios.

### 4.1 Baseline

Table 1 presents the  $C_{llr}$  scores obtained by each method on the evaluation dataset simulated using baseline parameters. Although  $C_{llr}$  scores may be sensitive to the dataset evaluated, we compare method scores to those reported in the papers that introduced them as a sanity check. Both the discriminative and close pair methods were previously assessed on the *Netherlands Forensic Institute* (NFI)

<sup>4</sup>hub.arcgis.com

<sup>5</sup>github.com/NetherlandsForensicInstitute/telcell

Method	Train	Baseline
Discriminative	<i>Baseline</i>	<b>0.097</b>
Close Pair	-	0.580
Count	-	0.471
Close Pair * Count	-	0.418

**Table 1: The log likelihood ratio cost ( $C_{llr}$ ) for the evaluation dataset with baseline parameters.**

curated dataset [1, 3], and the count method on a Twitter dataset [2].

The discriminative approach achieves  $C_{llr} = 0.097$ , outperforming the  $C_{llr}$  of 0.47 it achieves on the NFI dataset. This improvement likely stems from the simulation training data’s larger size and greater similarity to the corresponding validation data. The close pair method scores a  $C_{llr} = 0.580$ , more closely aligning with its  $C_{llr} = 0.71$  on the NFI dataset. The improved performance indicates that our baseline parameters favor the close pair model more. This is potentially due to the NFI dataset’s composition, which includes employees commuting to the same location at similar times, resulting in a higher incidence of co-locations for different users.

The count method scores a  $C_{llr}$  of 0.471, improving upon the  $C_{llr} = 0.778$  achieved on the Twitter data under non-informative prior. However, this does not account for the boundary adjustment we applied to improve calibration. Without this adjustment, the method’s performance drops to  $C_{llr} = 0.667$  on the simulated data, more in line with its result on the Twitter data. These results may be less comparable as we only examine data originating from a single day for each LR, unlike the months-long observation period used in the Twitter data. Combining close pair and count yields a  $C_{llr} = 0.418$ , surpassing their individual performances and indicating combined evidential strength on the baseline parameters.

## 4.2 Scenarios

**4.2.1 Scenario dynamic/local.** Table 2 provides the results for variations in the scaling of the waiting time and travel distance distributions. While the discriminative method demonstrates resilience across most training and evaluation configurations, it struggles when trained on *local* agents with short travel distances and assessed on *dynamic regional* agents with larger travel distances and short waiting times. This discrepancy is likely due to the system learning to recognize shorter travel distances belonging to the same

user while also learning from stationary agents. Therefore, it only fails on datasets marked both by greater travel distances and minimal stationary behavior.

The close pair method showcases robustness across the scaled waiting time distributions, achieving consistent outcomes for both *static* and *dynamic* settings when the travel distance distribution remains the same. However, its performance is sensitive to scaling in the travel distances, with *regional* agents performing better than *local* ones. This sensitivity likely stems from the *local* agents traveling smaller distances and connecting more often to a smaller subset of antennas. This increased repetitiveness of antenna connections increases the likelihood of coincidental co-locations for same-user pairs, resulting in less confident and accurate same-user likelihood ratios.

The count method’s performance diminishes with shorter waiting times and larger travel distances, showcasing sensitivity to scaling in both distributions. Scaling these parameters broadens the spread of the location data, reducing the accuracy and confidence of likelihood ratios produced by the count method for same-user track pairs. Given that the close pair and count methods offer compensatory benefits across the variations, the combined approach seems to provide a stabilizing effect, positioning the combined performance between the outcomes of the individual methods.

**4.2.2 Scenario many locations.** In scenario *many locations*, we investigate the impact of location predictability on method performance, detailed in Table 3. The results indicate that the discriminative approach appears robust to these variations, most likely because the method does not consider precise location information but investigates the time and space distance between consecutive pairs of registrations.

Method	Train	Returners	Explorers
Discriminative	<i>returners</i>	<b>0.097</b>	0.077
	<i>explorers</i>	0.133	<b>0.075</b>
Close Pair	-	0.580	0.299
Count	-	0.471	0.623
Close Pair * Count	-	0.418	0.235

**Table 3: The log likelihood ratio cost ( $C_{llr}$ ) for the evaluation datasets in scenario *many locations*. The dataset “Returners” corresponds to our baseline parameters.**

Method	Train	Static Local	Dynamic Local	Static Regional	Dynamic Regional
Discriminative	<i>static local</i>	<b>0.126</b>	0.136	0.281	1.193
	<i>dynamic local</i>	0.127	<b>0.137</b>	0.263	1.540
	<i>static regional</i>	0.156	0.188	<b>0.097</b>	0.396
	<i>dynamic regional</i>	0.245	0.252	0.142	<b>0.251</b>
Close Pair	-	0.789	0.725	0.580	0.517
Count	-	0.250	0.308	0.471	0.744
Close Pair * Count	-	0.298	0.333	0.418	0.538

**Table 2: The log likelihood ratio cost ( $C_{llr}$ ) for the evaluation datasets in scenario *dynamic/local*. The dataset “Static Regional” corresponds to our baseline parameters.**



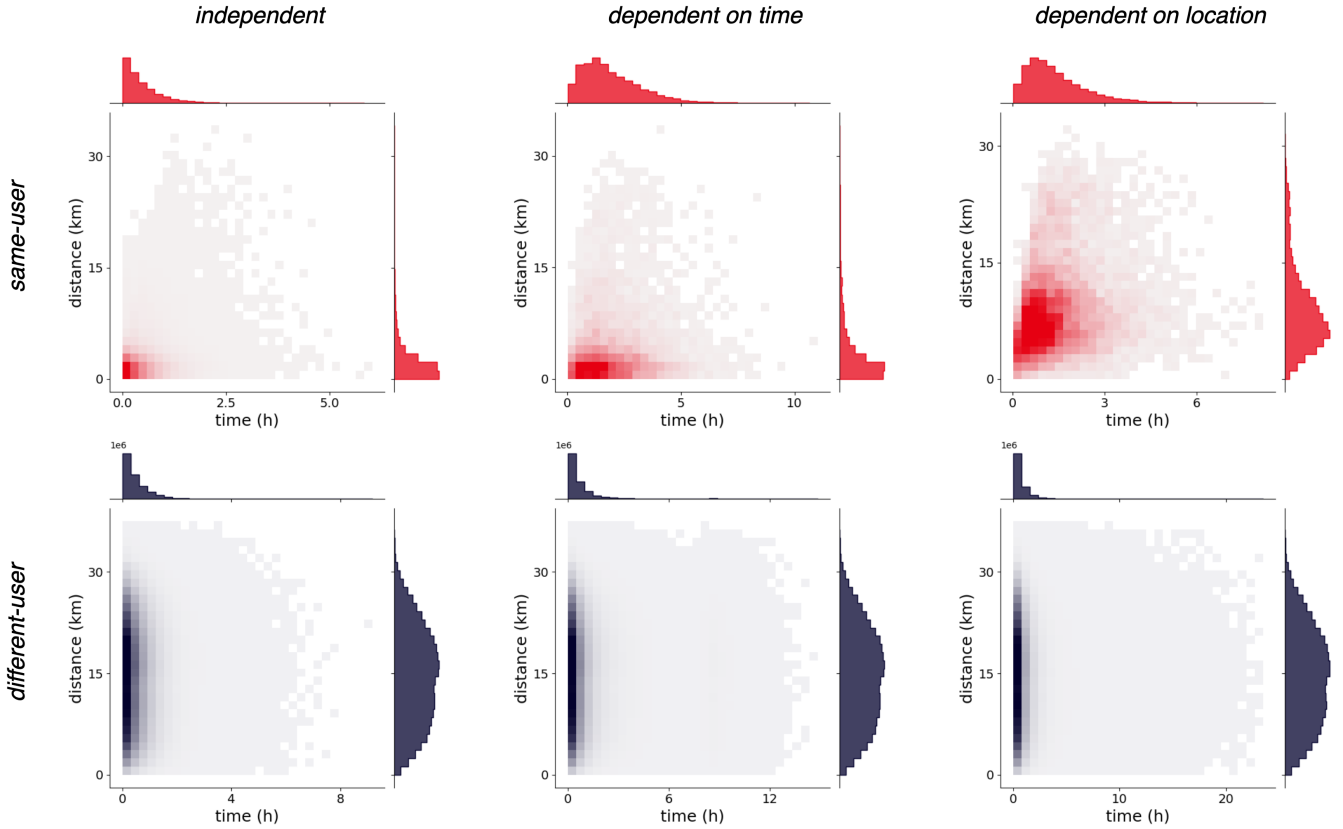


Figure 3: Heatmap and marginal distributions of distance and time difference for features extracted from consecutive phone-pair registrations, for (top) same-user and (bottom) different-user track pairs for the three-time sampling cases: (left) independent (middle) dependent on time and (right) dependent on location.

The close pair method performs better on the many locations visited in the *explorers* dataset. This trend mirrors scenario *dynamic/local*, where user locations, and therefore corresponding antenna connections, are more repetitive under the predictable locations in the *returners* dataset. This repetitiveness, again, results in agents connecting to a smaller subset of antennas with higher visitation frequency, increasing the likelihood of coincidental co-location for same-user pairs.

Unlike the close pair approach, the count method’s performance decreases on the *explorers* dataset. Mirroring scenario *dynamic/local*, users visiting many locations produce a wider dispersion of points, reducing the count methods performance for same-user classifications. Similar to baseline findings, merging the close pair and count methods improves upon the individual performances, showcasing combined strength under these parameters.

**4.2.3 Scenario dependence.** In Table 4, we provide model performance for the varying dependency sampling techniques in scenario *dependence*. The discriminative method’s sensitivity becomes apparent when trained on either the *independent* or *dependent* time dataset and then evaluated on the *dependent* location dataset. To explain this sensitivity, we include the distribution of features (used

for training) for the same- and different-user pairs across the varying dependencies in Figure 3. These distributions reveal that location dependence fundamentally changes the time and distance distributions between the same-user pairs (as seen in the top right plot in Figure 3). This change makes the distribution of features between the same- and different-user cases more similar, making it

Method	Train	Ind.	Dep. Time	Dep. Loc.
Discriminative	<i>ind.</i>	<b>0.097</b>	0.607	3.151
	<i>dep. time</i>	0.161	<b>0.413</b>	1.503
	<i>dep. loc.</i>	0.361	0.643	<b>0.752</b>
Close Pair	-	0.580	0.412	0.228
Count	-	0.471	0.838	1.242
Close Pair * Count	-	0.418	0.556	0.425

Table 4: The log likelihood ratio cost ( $C_{llr}$ ) for the evaluation datasets in scenario *dependence*. The dataset “Ind.” corresponds to our baseline parameters. Abbreviations: Ind. = Independent, Dep. = Dependent, Loc. = location.

Method	Train	Static Local	Dynamic Local	Baseline	Explorers	Dynamic Regional	Dependent Time	Dependent Location
Discriminative	<i>static local</i>	<b>0.126</b>	0.136	0.281	0.165	1.193	1.611	4.497
	<i>dynamic local</i>	0.127	<b>0.137</b>	0.263	0.244	1.540	1.129	4.458
	<i>baseline</i>	0.156	0.188	<b>0.097</b>	0.077	0.396	0.607	3.151
	<i>explorers</i>	0.287	0.220	0.133	<b>0.075</b>	0.403	0.653	2.752
	<i>dynamic regional</i>	0.245	0.252	0.142	0.086	<b>0.251</b>	0.594	2.766
	<i>dependent time</i>	0.185	0.214	0.161	0.152	0.344	<b>0.413</b>	1.503
	<i>dependent location</i>	0.515	0.455	0.361	0.387	0.494	0.643	<b>0.752</b>
Close Pair	-	0.789	0.725	0.580	0.299	0.517	0.412	0.228
Count	-	0.250	0.308	0.471	0.623	0.744	0.838	1.242
Close Pair * Count	-	0.298	0.333	0.418	0.235	0.538	0.556	0.425

**Table 5: The log likelihood ratio cost ( $C_{llr}$ ) for all datasets featured in our experimental analysis, including all train and evaluation combinations for the discriminative approach. The ordering of datasets is according to the performance of the count method.**

more difficult to distinguish between them. When the discriminative model is fit on easier-to-distinguish data, it learns to classify simple patterns as originating from the same user; if then validated on a much harder evaluation dataset, the method will naively assign any non-simple patterns as being from different users, resulting in many overconfident and incorrect different-user likelihood ratios.

The close pair model consistently achieves high and stable performance across the dependencies. In both dependent samplings, the time between consecutive registrations is much higher than under independent sampling, as seen in Figure 3’s temporal distributions. As this method attempts to find a pair of registrations within a 2-minute window, it can only create likelihood ratios for a small set of same-user track pairs under dependency sampling, something not reflected by the  $C_{llr}$  metric.

The count method’s performance declines under dependency sampling, particularly for the location dependence. Both dependencies create a segmentation of the location data, which is especially pronounced under location sampling. As the count method only investigates spatial information, this complicates its ability to recognize same-user pairs, resulting in many overconfident and incorrect results. The combination approach stabilizes the count method’s low performance, as the close pair filters out many of its incorrect same-user likelihood ratios.

### 4.3 Comprehensive Dataset Analysis

We combine previous results into Table 5, including results for all combinations of train and evaluation datasets used in the discriminative approach. As expected, the discriminative approach achieves its highest performance when trained and evaluated on datasets with shared parameters, as seen by the results on the diagonal axis in Table 5. The general behavior is that fundamentally changing the time difference and distance between pair registrations results in worse  $C_{llr}$  values.  $C_{llr}$  values even surpass 1 when the resulting patterns for the hypotheses are easy to distinguish in the training set but hard in the evaluation, i.e., in the top right corner of the table.

The general pattern for the close pair approach is that it is robust to all scenarios, achieving only  $C_{llr}$  values below 1. Its performance declines under more bounded or repetitive movement; however, it

manages high results under dependency sampling due to the close time constraint when choosing a pair of registrations to evaluate.

Contrary to the close pair approach, the count method improves for bounded and repetitive movements yet deteriorates under dependency sampling. Given this general trend of compensatory benefits between the close pair and count methods, the combined approach results in a stabilizing effect, achieving relatively high performance across the variations. Combining the likelihood ratios increases the individual performances for 2 out of the 7 investigated datasets, indicating that these methods may achieve combined evidential strength for some datasets.

## 5 DISCUSSION AND CONCLUSION

Our study examined the robustness of various systems for assessing phone-pair co-usage from Call Detail Records (CDRs). We designed a systematic evaluation method using simulated data to observe the impact of travel and phone usage behaviors on the performance of various systems. We explored a discriminative approach trained on a reference CDR dataset, a close pair method analyzing a select registration pair close in time, and a categorical count method that assesses global spatial information. Additionally, we introduce a hybrid approach combining the close pair and categorical count methods, supported by their complementary potential. Our experimentation focused on three behavioral scenarios: alterations in temporal and spatial distributions modeling user movement, variations in location visit predictability, and different phone usage dependency samplings.

### 5.1 Interpretation and implications

Our findings reveal that the data-driven discriminative approach is adaptable and achieves high performance across behavioral scenarios, but only when trained on representative data. This same adaptability means that, when the parameters for training and evaluation are distinct, the model can become over- or underconfident, even leading to a system that makes decisions worse on average ( $C_{llr} > 1$ ). Therefore, we recommend that practitioners only use this system if confident that their reference dataset closely matches the alternative population in the case at hand.

We find that the close pair approach is much more robust, showing adequate performance for all scenarios tested. This makes it a safer method to use in practice. However, particularly under bounded and repetitive movement, investigators should be aware that this method may result in underconfident same-user likelihood ratios.

The categorical count method performs well under various scenarios, showing mostly low  $C_{llr}$  values. The method requires that the sampling period is long enough, such that most of the user’s movement behavior has been observed, and will thus be less useful when phones are used for only days. More importantly, the method breaks down ( $C_{llr} > 1$ ) when there is a strong mismatch in phone usage, e.g. when one phone is used at home and the other at work. The close pair approach is more favorable if such a mismatch is present.

As the close pair method looks at detailed time-dependent information and the count method looks at overall spatial patterns, their evaluations may be regarded as (nearly) independent. We therefore proposed a fourth approach that simply multiplies their LRs. This approach resulted in a well-performing system that was stable across the scenarios.

## 5.2 Limitations and future work

We employ relatively simple behavioral modeling that does not cover all possible relevant behaviors. For example, bounding-box effects limit the maximum distances and speeds covered by agents, therefore inter-city travel at high speeds is missing from our investigation. We aimed to look at a spread of feasible behavioral profiles specifically to test the robustness of methods rather than provide the definitive answer on what method works best. Given the expected and observed performance of the methods, adding more details will likely not change the overall picture of robustness we found.

Another interesting avenue for extension is the inclusion of social-based models aimed at capturing dynamics between related agents, i.e., family members or coworkers. Modeling these relations would result in harder datasets containing more examples of accidental co-locations. These datasets would not conform to the alternative hypothesis used by the method of independent movement, it would be interesting to see what this would do to model results.

Our work also points at improvements possible for the evidence evaluation methods themselves. Although the close pair method shows robustness, it is often underconfident. Incorporating more information, such as sequential data, may improve this. One way forward may be to further investigate the combination of methods that look at different aspects of the data, which we did in a very simplistic manner here by multiplying LRs.

## 5.3 Conclusion

In conclusion, no single method is preferable in any situation. When little knowledge is available on the population of interest, the close pair method seems the safest bet. For longer sampling periods, it may be well worth looking at the count method or a combination of the two, as it provides additional information. When reference data are available that can be seen as representative of the population of

interest, more data-driven approaches, such as the discriminative approach, will offer the best performance.

## A COUNT POST-HOC CALIBRATION

Table 6 provides the log likelihood ratio cost for the original and our proposed bounded count methods on the experimental evaluation datasets. Our simple bound improves results on all datasets, indicating that the count method may result in unreasonably large LRs.

Dataset	Count original	Count bounded
<i>static local</i>	0.267	0.250
<i>dynamic local</i>	0.383	0.308
<i>baseline</i>	0.667	0.471
<i>explorers</i>	0.943	0.623
<i>dynamic regional</i>	1.156	0.744
<i>dependent time</i>	0.994	0.838
<i>dependent location</i>	2.082	1.242

**Table 6: The log likelihood ratio cost ( $C_{llr}$ ) for the count method for both the original method and our proposed bounded method. For bounding, we clip the LR to lower bound  $1/n$  and upper bound  $n$  where  $n$  is the smaller length of the evaluated input tracks.**

## B OVERVIEW OF EVALUATION DATASETS

Table 7 provides an overview of the global metrics for the evaluation datasets across the baseline and three behavioral scenarios. It details the total number of days, agents, phones per agent, data points, and track pairs utilized, and it also outlines the average number of daily switches observed between same- and different-user track pairs.

Metric	Baseline	Dyn./Loc.	Many Loc.	Depend.
days	30	30	29	30
agents	100	100	100	100
phones per agent	2	2	2	2
data points	142.8k	143.9k	129.9k	71.8k
track-pairs	6k	6k	5.7k	6k
avg. $\phi_{su}$	22.7	22.5	22.0	2.35
avg. $\phi_{du}$	22.6	22.5	21.9	7.29

**Table 7: Global dataset metrics for baseline and various scenarios.  $\phi_{su}$ : switches between same-user track pairs,  $\phi_{du}$ : switches between different-user track pairs.**

Similar values are observed across datasets that utilize independent sampling, namely the baseline, the *dynamic/local* scenario, and the *many locations* scenario. In contrast, the *dependence* scenario employs dependency sampling, where we sample usage times from a single distribution per individual rather than separate distributions for each phone. As a result, datasets generated under dependency sampling are approximately half the size of those produced via independent sampling. With independent sampling, we observe a consistent average of around twenty-two daily switches



between same- and different-user track pairs. Dependency sampling, however, leads to a substantial reduction, with different-user pairs averaging about seven switches. We also use a restrictive switch condition under dependency sampling based on a specific time of day or if the user is within a particular area. This results in an average of roughly only two switches per same-user track pair.

## REFERENCES

- [1] W. Bosma, S. Dalm, E. van Eijk, R. El Harchaoui, E. Rijgersberg, H. T. Tops, A. Veenstra, and R. Ypma, "Establishing phone-pair co-usage by comparing mobility patterns," *Science & Justice*, vol. 60, no. 2, pp. 180–190, 2020.
- [2] R. Longjohn, P. Smyth, and H. S. Stern, "Likelihood ratios for categorical count data with applications in digital forensics," *Law, Probability and Risk*, vol. 21, no. 2, pp. 91–122, 2022.
- [3] W. Bosma and R. Ypma, "A feature-based lr system for evaluating phone-pair co-usage from mobility patterns."
- [4] M. Tart, "Cell site analysis: Changes to networks with time," *Science & Justice*, vol. 62, no. 3, pp. 377–384, 2022.
- [5] C. Champod, A. Biedermann, J. Vuille, S. Willis, and J. De Kinder, "Enfsi guideline for evaluative reporting in forensic science: A primer for legal practitioners," *Criminal Law and Justice Weekly*, vol. 180, no. 10, pp. 189–193, 2016.
- [6] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nature physics*, vol. 6, no. 10, pp. 818–823, 2010.
- [7] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini, "Human mobility: Models and applications," *Physics Reports*, vol. 734, pp. 1–74, 2018.
- [8] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, 2006.
- [9] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *nature*, vol. 453, no. 7196, pp. 779–782, 2008.
- [10] H. Barbosa, F. B. de Lima-Neto, A. Evsukoff, and R. Menezes, "The effect of recency to human mobility," *EPJ Data Science*, vol. 4, pp. 1–14, 2015.
- [11] L. Pappalardo, S. Rinzivillo, and F. Simini, "Human mobility modelling: exploration and preferential return meet the gravity model," *Procedia Computer Science*, vol. 83, pp. 934–939, 2016.
- [12] Y. Leo, A. Busson, C. Sarraute, and E. Fleury, "Call detail records to characterize usages and mobility events of phone users," *Computer communications*, vol. 95, pp. 43–53, 2016.
- [13] B. Wang, V. Hess, and A. Crooks, "Mesa-geo: A gis extension for the mesa agent-based modeling framework in python," in *Proceedings of the 5th ACM SIGSPATIAL International Workshop on GeoSpatial Simulation*, pp. 1–10, 2022.
- [14] D. Meuwly, D. Ramos, and R. Haraksim, "A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation," *Forensic science international*, vol. 276, pp. 142–153, 2017.
- [15] D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, and C. Aitken, "Information-theoretical assessment of the performance of likelihood ratio computation methods," *Journal of forensic sciences*, vol. 58, no. 6, pp. 1503–1518, 2013.