Vecsim: Carrier-based, Privacy-Preserving Cellphone Contact Tracing

Avishek Mukherjee
Department of Computer Science,
Saginaw Valley State University
University Center, MI

Yaoguang Zhong
Department of Computer Science,
Florida State University
Tallahassee, FL

Zhenghao Zhang
Department of Computer Science,
Florida State University
Tallahassee, FL

Tingting Zhao
Department of Geography, Florida
State University
Tallahassee, FL

Jinfeng Zhang
Department of Statistics, Florida
State University
Tallahassee, FL

ABSTRACT

In this paper, Vecsim, a novel contact tracing method, is proposed. Vecsim determines the user proximity based on the existing log data already collected by the cellphone network carrier for network management purposes, and is transparent to the users. Compared to the existing methods that require user involvement, such as downloading an application and turning on Bluetooth, Vecsim is easier to deploy and may cover a larger population. Vecsim protects user privacy by focusing on user proximity detection, which is different from localization. In addition, proximity detection is a much easier problem than localization and can achieve higher accuracy. The key novelties of Vecsim include a simple method for distance estimation based on the similarity of two data records, as well as exploiting the massive log data to learn the discontinuity of the signal field. Vecsim has been tested with the signal field generated by a commercial ray tracing software program in a 2 km by 2 km urban area. The results show that Vecsim alerts over 96% of cellphones within 50 m, while alerting less than 4.5% of cellphones beyond 150 meters.

1 INTRODUCTION

COVID-19 has affected the entire world. Contact tracing is one of the key components in the collective effort to fight this global pandemic. By keeping track of the contact history of the infected individuals, the spread of the disease can be traced and monitored. Vulnerable individuals can be alerted or treated earlier. Lives can be saved.

In this paper, we propose a novel contact tracing method, named *Vecsim* for Vector Similarity. Vecsim is based on the analysis of the cellphone measurement report, which is a vector with many fields, and is sent by each connected cellphone to the carrier periodically for network network management purposes. Vecsim has the following main features:

- Much wider applicability than the existing solutions: Vecsim runs only at the carrier side and is transparent to the users.
- Privacy preserving: Vecsim does not localize the cellphones; instead, it only checks proximity of the cellphones. Proximity check is also much easier than localization.
- Higher accuracy than existing carrier-based solutions: Vecsim determines the contact of cellphones based on the similarity of the measurement reports, which contain very rich information.

To elaborate, Vecsim can run at the carrier side based on the data already available to the carrier, because the measurement reports are sent by the cellphones to the carrier automatically. Therefore, in Vecsim, data collection does not need any user involvement, which is in sharp contrast with existing solutions that require the users to download certain applications or turning on the Bluetooth interface.

Vecsim exploits the fact that proximity check is not localization, and, very fortunately, is both simpler and less intruding to user privacy. The localization problem can be completely circumvented, because the question of interest is whether or not a contact has occurred, not the exact location. In other words, as long as the cellphones report similar vectors, they were likely close, even though their locations are not known. In addition, without actually knowing the location of the users, user privacy is better protected.

The measurement report is accumulated continuously and contains very rich information, which allows Vecsim to achieve accuracy higher than that with the current solutions. That is, Vecsim alerts over 96% of cellphones within 50 m, while alerting less than 4.5% of cellphones beyond 150 meters. In contrast, without relying on training such as the road tests, the current solutions can typically achieve cell-level accuracy, such as within $\frac{3}{4}$ square miles [24].

The main challenge of proximity estimation, especially in urban areas, is the highly complicated wireless propagation environments, which can introduce irregularities to the signal field. That is, even when two locations are close, the signals may still be significantly different. Vecsim addresses the challenges by combining a simple method for distance estimation, as well as exploiting the massive data to learn the irregularities, which can be used to correct the errors made by the simple estimation method. The learning does not require any additional training activities such as the road tests, and instead is solely based on the data already available. The idea is to discover the discontinuity revealed by users that walk across the discontinuity locations, whose cellphones will send apparently different reports within a short time frame. At the carrier side, once such reports are observed, a pair of discontinuity locations is learned. Clearly, the learning accuracy should monotonically improve with more data, therefore, Vecsim is expected to achieve increasingly better performance over the time.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 discusses the preliminaries and background information. Section 4 explains the data generation. Section 5 explains the heuristic method for distance estimation. Section 6 explains the learning of the discontinuity of the signal field with the massive data. Section 7 evaluates Vecsim. Section 8 concludes the paper.

2 RELATED WORK

Contact tracing, cellphone localization, and urban signal field analysis are all active areas of research in recent years, which are discussed in the following.

2.1 Contact Tracing

Contact tracing has been around for a while but has recently gained a lot of traction due to the ongoing COVID-19 pandemic.

2.1.1 Contact Tracing with Cellphone Signal. In South Korea, user information like the cellphone signal, GPS, credit card transactions and CCTV footage are collectively used to determine the route an infected user may have taken [5]. The cellphone signal, however, determines only the cell the user is in, while Vecsim offers much higher accuracy.

2.1.2 Contact Tracing with QR code. In China, recent contact tracing efforts include mandating the scanning of a QR code with the user's phone, in malls and other public spaces, to create a digital footprint of every person [13]. Clearly, the accuracy of this approach is the same as the size of the public spaces. Vecsim can serve as a strong alternative and offer much higher accuracy.

2.1.3 Manual Contact Tracing. Traditional methods include centralized contact tracing by manually conducting phone interviews with exposed individuals [11], which can be challenging in areas with high population density and may be less accurate in certain cases, since it relies on the individual's memory to recreate a path map. Vecsim belongs to the family of automated contact tracing, and complements the manual contact tracing.

2.1.4 Contact Tracing based on Bluetooth. In recent years, there has been a rise in automated decentralized contact tracing methods which make use of the networking capabilities of IoT devices to identify individuals who may have been exposed to an infected person. These existing contact tracing solutions, such as FluPhone [28], TraceTogether [1], Privacy-Preserving Contact-Tracing (PPCT) [3] requires the users to download an application to the phones and turn on the Bluetooth interface, such that, when two phones are in close contact, the Bluetooth modules can discover each other. Clearly, solutions like these depend on the wide adoption of the application, which may be difficult unless certain mandates are issued by the government. Further, these applications are susceptible to a variety of network attacks as outlined in [2]. VecSim is different from the Bluetooth-based contact tracing, since it does not require any additional actions from the user, and therefore will be much easier to deploy and may cover a larger population.

2.1.5 Contact Tracing based on Social Networks. On a similar note, [30] explores the possibility of using social network data for contact tracing. The system combines user activity from different social networks like Facebook, WeChat to detect spread of infections among other users. Naturally, these types of systems rely on the user to be active on these platforms, which is not always a reasonable assumption. VecSim is based on cellphone data that is automatically logged constantly, and will likely achieve higher accuracy.

2.1.6 Contact Tracing based on Wi-Fi. There has also been recent studies [15] analyzing the effectiveness of using Wi-Fi for contact tracing using MAC addresses for identification and the signal strength to estimate the proximity of two devices. The paper also explores contact tracing using cellular networks and GPS and uses the user location to determine possible exposure. Vescim is different because it focuses on the cellphone network data, which has more ubiquitous coverage than Wi-Fi; in addition, Vecsim does not rely on the user location.

2.1.7 Contact Tracing based on Wearable Devices. Finally, there are some contact tracing solutions that require the user to wear some kind of wristband or wearable [27], typically in workplaces that have a high risk of infections.

Vecsim is clearly different as it does not require the user to wear any additional device.

2.2 Cellphone Localization

Localization based on cellphone signals is an interesting proelem. As mentioned earlier, the current solution can typically locate a cellphone within $\frac{3}{4}$ square miles [24]. LTE log data has been recently used for user localization [18–20]. Vecsim completely circumvents the challenging problem of localization, and focuses on proximity detection, which is a completely different and easier problem. In addition, Vecsim adopts a completely different solution, which makes use of the vast amount of unlabeled log data, i.e., the MRO files, while the existing localization methods train with labeled data obtained by cellphones that report their locations. The unlabeled data is more universally available, while the labeled data may be available only for a few targeted areas.

2.3 Urban Signal Field Analysis

Urban wireless signal field has attracted constant interest in the research community due to its importance to network planning and other applications. Measurement studies have been conducted [6, 16]. As signal propagation is governed by laws of physics, which can be understood and approximated, *ray tracing* [4, 17, 23, 26, 29] has also been used for various types of environments [7, 9, 12, 25], with recent interest in speeding up the calculation by utilizing parallelism with GPU [8, 22] or with special processor support [14]. In general, ray tracing has lead to encouraging results, such as within 10 dB of the measured signal, provided that the environment information is sufficiently accurate. Vecsim does not depend on the detailed information of the signal field, because it focuses only on proximity estimation, which can be achieved by examining the similarity of the data records.

3 BACKGROUND

Cellphone carriers build towers to cover the service areas. Each tower typically mounts 3 base stations, each faces a different direction and is associated with an identification number called *PCI*. The network continuously logs the data, such as the *MRO* file, which contains records of all connected cellphones. An *MRO record* is reported by a connected cellphone every 5.12 seconds, which is a high dimensional vector, including:

- *Tadv*: A non-negative integer, denoted in this paper as γ. The system treats $\gamma\zeta$ as the length of the signal propagation path, where ζ is 78 m. Note that the signal may undergo reflections and $\gamma\zeta$ may not be the distance of the cellphone to the base station, especially in urban areas. Tadv is typically in the range of $[0, 10]$, especially in urban areas, because a cellphone prefers connecting to the strongest base station, while the base station deployment in urban areas can be very dense.
- *AoA* : Angle of Arrival, which is a real value within $[0, 2\pi]$ and is denoted in this paper as θ. AoA is the angle of the signal from the cellphone observed at the base station.
- *RSRP* and *nRSRP*: RSRP is the signal strength of the base station the cellphone is connected to; nRSRP is that of a neighboring base station. There is only one RSRP but potentially multiple nRSRPs in an MRO record. In this paper, as RSRP and nRSRP are treated the same, with a slight abuse of notation, they will be both referred to as RSRP, and denoted as ω. ω at 20 and 70 are very weak and very strong signals, respectively.

If the cellphone is not actively used by the user, it still passively measures the network. If the cellphone detects a significant change of the measurement, it will send a message, which contains an MRO report. If the measurement does not change, the cellphone still pings the base station periodically. The interval between two pings is configured by the carrier, and can be 8 hours. Although this could mean that no updates from a cellphone for an extended period of time, we note that, if an update has not been triggered, the change of the MRO data should be bounded. The exact threshold to trigger an update, as well as the ping interval, are configured by the carrier. If a carrier-based solution like Vecsim can be adopted, the carrier should have agreed to cooperate. It might also be possible to slightly change the configuration, such as changing the update threshold, such that more frequent MRO records can be obtained, especially when the user is moving.

4 DATA GENERATION

Ideally, the data should be collected in the real-world with location labels. As such data is not available, we used Wireless Insite [21], a ray-tracing software program, to generate the signal field in an urban area.

4.1 The Study Area and the Signal Field

Fig. 1 shows a 4 km by 4 km urban area. A total of 13 towers are simulated, which contains 39 base stations. As the signal decays fast, to limit the computation complexity, for each base station, only the 2 km by 2 km area centered at the base station is simulated. For each base station, the Wireless Insite software is configured to calculate the rays received at locations on 4 m by 4 m grids. The bandwidth is 20 MHz. Fig. 2 shows the signal field in a birds-eye view, where the field of different base stations are shown in different colors. Note that the field of a base station may only be partly shown if it has some overlap with another base station plotted afterwards. The 2 km by 2 km area in the center of Fig. 1 is used as the study area, because the signal near the edge of the complete 4

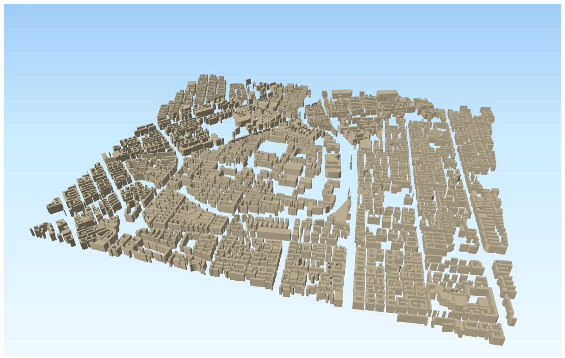

Figure 1: The 4 km by 4 km urban area.

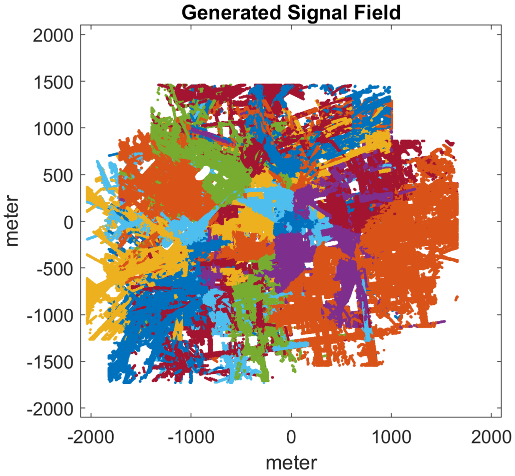

Figure 2: The birds eye view of the generated signal field.

km by 4 km area does not have sufficient coverage due to the locations of the base stations.

As we do not have the information of the building materials, the building is modeled as concrete blocks in the software. We therefore do not have indoor data, and limit this study to outdoor areas only. However, the basic principles of Vecsim should also apply to indoor cases.

4.2 Generating the MRO Data

For each location, an MRO record is created. Typically, multiple rays may be received from a single base station, each traveling along a different path. The total power from all rays is used to calculate the RSRP of this base station, because even if two rays may cancel at a certain frequency, they may add up constructively at another. The delay and angle of the strongest ray is used to calculate the Tadv and AoA, respectively. As a location may receive from multiple base stations, the strongest base station is assumed to be the connected base station. Fig. 3 shows the RSRP of three base stations in the study area.

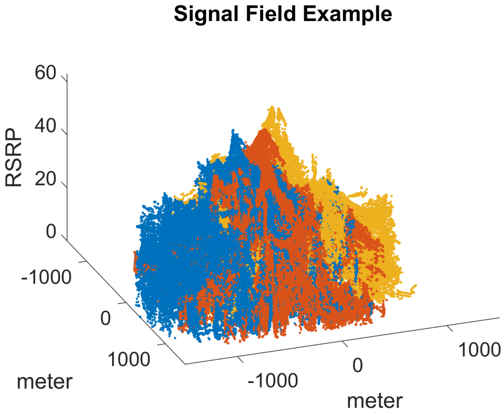

Figure 3: The RSRP of three base stations in the study area.

5 DISTANCE ESTIMATION

The distance estimation heuristic is the core of Vecsim, which gives an estimate of the distance of two MRO records.

5.1 Initial Estimate

Given the two MRO records, the initial estimate, denoted as $\hat{D}$, is simply

$$\hat{D} = \left[\gamma_1^2 + \gamma_2^2 - 2\gamma_1\gamma_2 \cos\left(\theta_1 - \theta_2\right)\right]^{\frac{1}{2}} \zeta,$$

where γ_m and θ_m are the Tadv and AoA in MRO record m, respectively, for $m = [1, 2]$. Clearly, this estimate is correct when the signal propagation paths to both locations are Line-of-Sight (LOS), in which case the distance is simply the length of an edge in a triangle, where the other two edges are of length $\gamma_1\zeta$ and $\gamma_2\zeta$, respectively, with angle $\theta_1 - \theta_2$ in between.

Note that the LOS assumption is likely true in rural areas. Even in urban areas, where the LOS assumption is not true, the initial estimate may still give a reasonable estimate in certain cases. Fig. 4 shows an example, in which the paths have undergone a reflection, such that the initial estimate is basically the distance between locations A and B in the figure, which are mirrors of the actual cellphone locations. However, fortunately, the distance between the mirrors is exactly the actual distance between the cellphones. The example also shows the challenges in localization, because a localization method that naively combine the AoA and Tadv may likely believe that the mirrors are the actual cellphone locations.

Clearly, the initial estimate will likely deviate more from the actual distance in more complicated propagation environments with more reflections. It is still a good starting point in

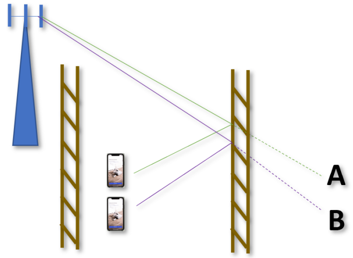

Figure 4: An illustration of the initial estimate based on Tadv and AoA.

many cases, because a cellphone connects to a base station likely because there exists a strong path, while strong paths typically do not undergo many reflections.

5.2 Maximum Likelihood Estimate

Due to the limitation of the initial estimate, additional adjustments are made based on the RSRP values according to the Maximum Likelihood principle. The likelihood value of a distance value is calculated based on $\hat{D}$ as well as the RSRP values, and the one with the largest likelihood is chosen as the estimate.

To be more specific, the actual distance estimate of Vecsim, denoted as η, is a discrete value within $[1, N]$, which represent evenly spaced distances with step δ meters, where δ is currently 25. N is currently 8, because distances above 200 m are believed to be equivalent to 200 m in the context of contact tracing. For each distance value i, two likelihood functions, denoted as $L_1(i)$ and $L_2(i)$, are calculated, which will be explained shortly. The discrete distance estimate is

$$\eta = \underset{1 \le i \le N}{\arg\max} L_1(i) L_2(i)$$

5.2.1 $L_1()$ – *Likelihood Based on the Initial Estimate.*

Let d be the discrete initial estimate, i.e., d is the smallest integer larger than $\frac{\hat{D}}{\delta}$. Let a be the discrete actual distance, i.e., a the smallest integer larger than $\frac{D}{\delta}$, where D is the actual distance. Let $\Phi_1^1()$ to $\Phi_N^1()$ be N Probability Density Functions (PDF), where $\Phi_a^1(d)$ represent the probability that d is discrete initial estimate, when the discrete actual distance is a. $\Phi_a^1()$ is set according to a simple heuristic:

- $a = 1$:
  - $\Phi_a^1(1) = 0.9$,
  - $\Phi_a^1(2) = 0.1$.
- $1 < a < N$:
  - $\Phi_a^1(a-1) = 0.1$,
  - $\Phi_a^1(a) = 0.8$,
  - $\Phi_a^1(a+1) = 0.1$.
- $a \ge N$:
  - $\Phi_a^1(N-1) = 0.1$,
  - $\Phi_a^1(N) = 0.9$.

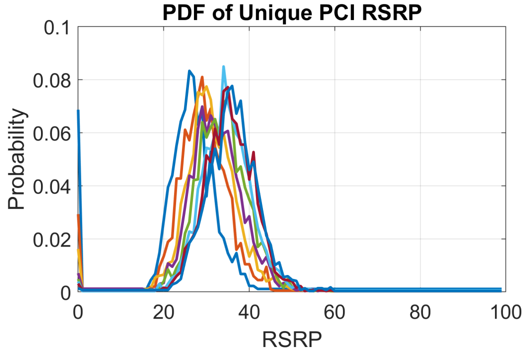

Figure 5: The PDF of the unique PCI RSRP used in Vecsim.

$L_1(i)$ is simply

$$L_1(i) = \Phi_i^1(d).$$

5.2.2 $L_2()$ – *Likelihood Based on the Unique Base Stations.*

Let $\Phi_1^2()$ to $\Phi_N^2()$ be another set of PDFs for the unique base stations, where a base station is unique if it is in only one of the MRO records. To be more specific, let ω^* a value to represent the unique base stations:

- if there is no unique base station: $\omega^* = 0$,
- if there is one unique base station: ω^* is the RSRP of this base station,
- if there are two or more unique base stations: ω^* is the average of the top 2 RSRP values of the unique base stations.

$\Phi_a^2(x)$ is the probability that $\omega^* = x$, when the discrete actual distance is a. Fig. 5 shows the set of PDF currently used in Vecsim, which is learned based on the simulated data. $L_2(i)$ is simply

$$L_2(i) = \Phi_i^2(\omega^*).$$

5.3 Discussions

The distance estimation in Vecsim relies on more robust features of the MRO data, namely the Tadv, the AoA, which are physical layer measurements that are fairly accurate even when the signal is very weak. The RSRP, on the other hand, can be more noisy, therefore, is taken with more precaution, e.g., only the maximum RSRP values of the unique base stations are used, which more sensitive to the distance than other featured that have been tested. Still, the distance estimation is a heuristic and can lead to errors, especially in case of signal field discontinuity, which is addressed in Section 6.

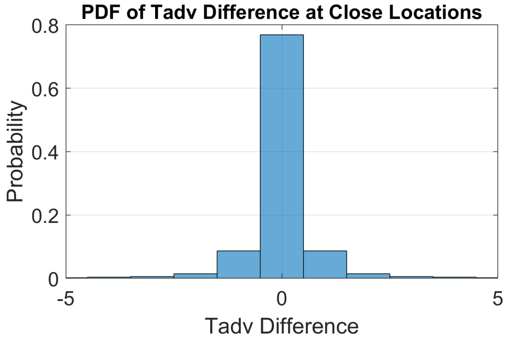

Figure 6: The Tadv difference of close pairs.

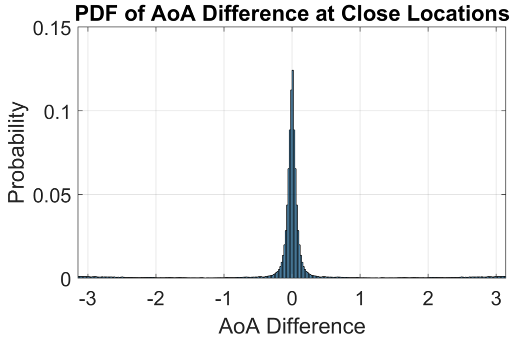

Figure 7: The AoA difference of close pairs.

6 COPING WITH SIGNAL DISCONTINUITY

The distance estimation in Section 5 is based on the assumption that the signal field is continuous, which is true in areas like an open field. However, in cities with many buildings which may block and reflect the signals, the assumption may not hold at certain locations. For example, the signal may experience a sudden change once someone emerges from the alley between two buildings into a major road. Vecsim solves this problem by exploiting the massive MRO data to learn the discontinuity of the signal field.

6.1 Discontinuity Pair

We checked every grid point in the study area and its neighboring grid points within a radius of 25 m. Fig. 6, Fig. 7, and Fig. 8 show the PDFs of the differences of Tadv, AoA, and RSRP, respectively, for points connected to the same base station. It can be seen that although the majority of the differences are small, there exist a non-negligible number of cases with large differences, such as those with the Tadv difference greater than 1, or those with the AoA difference greater than $\pi/4$. Given the close proximity of the locations, such large differences can only be caused by rays from different *families*, where two rays are from the same family if they experience reflections at the same set of surfaces before reaching the destinations.

Let $\xi(R_A, R_B)$ denote the discrete distance estimate calculated according to the heuristic in Section 5 for two MRO records R_A and R_B, which are for locations A and B, respectively. Let $D(A, B)$ be the distance between locations A and B. A and B are called a *discontinuity pair*, if $D(A, B) \leq 25m$ but $\xi(R_A, R_B) > 4$. Fig. 9 shows the distribution of $\xi(R_A, R_B)$ for locations pairs A and B within 25 m, where it can be seen that the fraction of discontinuity pairs are significant. Note that the percentage of large distance estimates is more than than the percentage of large Tadv and AoA differences, because the estimate can be large even with the same Tadv, as long as the AoA difference and the Tadv are large. Although it may be possible to further tune $\xi()$ to reduce such errors, we chose to use the current $\xi()$ due to its simplicity; more importantly, any tuning will still only work to a certain degree but not completely eliminate the problem. Instead, we solve this problem by learning from the MRO records.

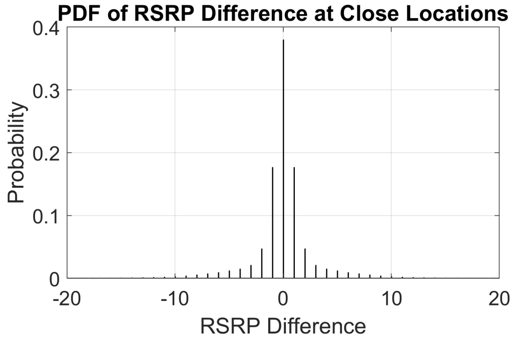

Figure 8: The RSRP difference of close pairs.

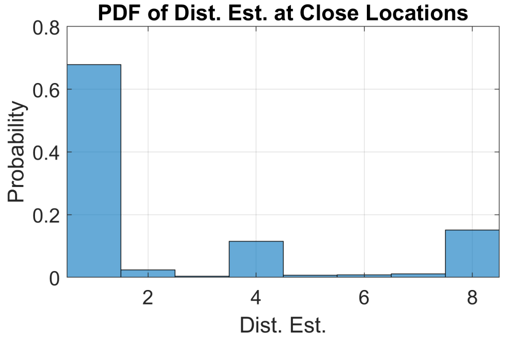

Figure 9: Distance estimation of the close pairs.

6.2 Learning the Discontinuity Pair Database (DPD)

The discontinuity pairs, fortunately, can be learned with the MRO data. The MRO data is reported by each connected cellphone to the base station continuously. Over the time, the amount of collected data is massive. Although it cannot be expected that a cellphone would report its location, the MRO data does very rich information that can be exploited. Of particular interest are those from cellphones carried by walking users, such as when the user is making a phone call or surfing the web while walking. The key observation is that *if the path of the user visits a discontinuity pair* (A, B)*, the MRO records reported at A and B will show high difference, but will appear within a few timestamps to each other*. Given that consecutive timestamps are only 5.12 seconds apart, it can be assumed that the two MRO records are likely from a discontinuity pair. Therefore, Vecsim examines the MRO records. If two MRO records, denoted as R_A and R_B, are reported by the same cellphone within 10 timestamps, but $\xi(R_A, R_B) > 4$, (R_A, R_B) may be add to the Discontinuity Pair Database (DPD).

6.3 Online DPD Compression

Let the *complete DPD* refer to the list that includes all the discontinuity pairs. Clearly, the size of DPD should be limited. Therefore, when learning the DPD, a simple online form of compression is used, which adds a pair only if it is significantly different with those currently in the DPD. To be more specific, suppose a pair of MRO records, say, $(R_{A'}, R_{B'})$, is found to be a candidate pair to be added to the DPD. Vecsim still checks the existing DPD, and discards $(R_{A'}, R_{B'})$, if another pair currently in the DPD, say, (R_A, R_B), is *similar* to $(R_{A'}, R_{B'})$ with *degree* 1, i.e., $\xi(R_{A'}, R_A) \le 1$ and $\xi(R_{B'}, R_B) \le 1$.

6.4 A Simulation Study based on Random Walk

To test the number of walking user cases needed to learn the complete DPD, we generated random walks in the 2 km by 2 km study area, where the walk is limited to the 4 m by 4 m grid points. From any point (x, y), where x and y are the indices of the grid points, the walk can make one of the 4 possible moves, namely, $(1, 0)$, $(0, 1)$, $(-1, 0)$, and $(0, -1)$, to neighbors $(x + 1, y)$, $(x, y + 1)$, $(x - 1, y)$, and $(x, y - 1)$, respectively. The maximum length of the walk is 100. For each random walk, first, a starting point is randomly selected from the area. A quadrant is then randomly selected as the main walking direction. For example, if quadrant 1 is selected, the allowable moves are $(1, 0)$ or $(0, 1)$; that is, with the starting point as the origin, any point on the walk can only be in quadrant 1. Another random number is selected as the *move probability*, that is, still assuming the selected quadrant is 1, the move probability is the probability whether $(1, 0)$ or $(0, 1)$ is selected as the move to be attempted first. Another limitation is that the walk can only visit points with service, i.e., can receive signal from some base stations. Therefore, if a move leads to a point with no service, this move fails, and the other move will be attempted. If both move fails, the walk terminates.

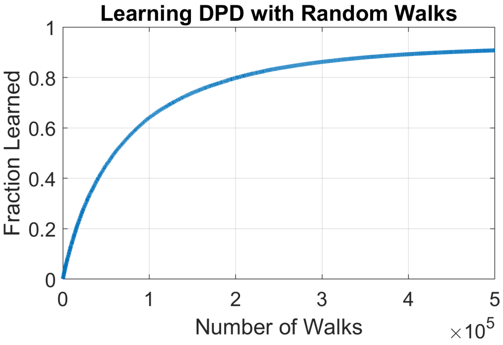

Figure 10: Fraction of discontinuity pairs visited as a function of random walks.

Fig. 10 shows the percentage of discontinuity pairs learned as a function of the number random paths, where 10 typical paths are shown in Fig. 11. It can be seen that over 90% of the pairs can be learned with 500,000 paths. The growth rate slows down as the number of paths increases, which is expected because this is in many ways similar to the coupon collector problem [10]. Also, certain pairs cannot be learned, if there does not exist a short path between them, such as when they are blocked by a building. Note that while the number of paths may be high, it will be eventually reached, as the MRO data continues to accumulate.

6.5 Discussions

The random walk simulation clearly has limitations, because the simulated walk may not match exactly the walks in the real world. However, we note the following facts that should make the learning of the DPD likely successful in the real world. First, in practice, there are likely high traffic areas and low traffic areas. The discontinuity pairs in high traffic areas can be quickly learned, which is important, because users are more likely to appear in high traffic areas than low traffic areas. Second, the more challenging areas, i.e., areas with many buildings that lead to complicated signal propagation paths, tend to be high traffic areas, because the number of buildings is likely correlated with the population size.

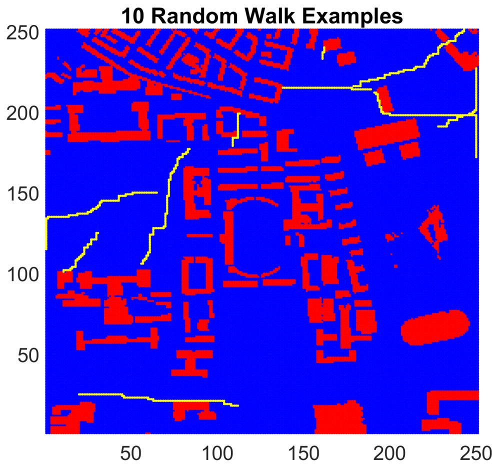

Figure 11: 10 typical random walks in the study area.

6.6 Applying the DPD

Vecsim applies the DPD to correct its estimate. To be more specific, consider two MRO records denoted as R_A and R_B. The final estimate of Vecsim is denoted as $\Xi\left(R_A, R_B\right)$, which is initially just $\xi\left(R_A, R_B\right)$. However, if $\xi\left(R_A, R_B\right) > 4$ and there exists a pair in the DPD similar to $\left(R_A, R_B\right)$ with degree 2, $\Xi\left(R_A, R_B\right)$ is replaced with 2.

To reduce the time to find similar pairs of $\left(R_A, R_B\right)$ in the DPD, the DPD is organized based on the PCI, the Tadv of the first MRO record, and the Tadv of the first MRO record. The computation of $\xi\left(\right)$ is only for the *related* pairs of $\left(R_A, R_B\right)$ in the DPD, where a pair, say, $\left(R_{A'}, R_{B'}\right)$, is related to $\left(R_A, R_B\right)$, if both $\left(R_A, R_{A'}\right)$ and $\left(R_B, R_{B'}\right)$ are related. For example, $\left(R_A, R_{A'}\right)$ are related if

- R_A and $R_{A'}$ have the same PCI,
- R_A and $R_{A'}$ have the same Tadv,
- the difference of AoA values of R_A and $R_{A'}$ is no more than $\frac{\pi}{4}$, and
- the difference of RSRP values of R_A and $R_{A'}$ is no more than 20.

7 EVALUATION

In this section, Vecsim is evaluated based on the signal field generated by the Wireless Insite software [21] in the city area discussed in Section 4.

7.1 Estimation Accuracy

A total of 1000 locations are randomly selected in the study area. For each location, the clean MRO is first calculated based on the output of the Wireless Insite software [21]. Gaussian noise with zero mean is added to the AoA and RSRP, where the standard deviations of the noise for AoA and RSRP are 0.05 and 5, respectively. Tadv is a discrete number and assumed to be accurate. The DPD learned from the 500,000 random walks discussed in Section 6.4, after the online compression discussed in Section 6.3, is used. The distance estimate is calculated for every pair of locations that connect to the same base station. In total, 34,567 pairs were tested.

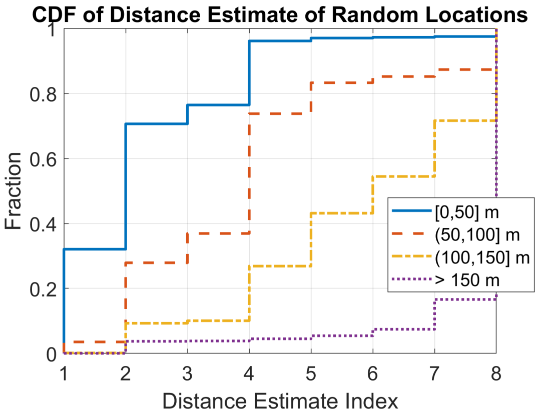

Figure 12: CDF of the discrete distance estimate.

Fig. 12 shows the Cumulative Density Function (CDF) of the discrete distance estimate. In practice, a threshold, denoted as κ, should be set, i.e., any discrete distance estimate equal to or less than κ should trigger an alert. It can be seen that if $\kappa = 4$, Vecsim will alert over 96% of the users that were close, i.e., within 50 m, but will alert only 4.5% of the pairs that were far and have little chance of infection, i.e., beyond 150 m. Therefore, Vecsim achieves satisfactory accuracy in practice, considering that it does not need any training, such as road tests or MRO data with attached GPS locations.

7.2 The Effectiveness of DPD

Fig. 12 illustrates the effectiveness of the DPD, which shows the CDFs of the discrete distance estimates with or without correcting the estimate with DPD. It can be seen that the DPD significantly improves the performance, i.e., with $\kappa = 4$, the alert probability increases from 76% to 96% for all pairs within 50 m, at the same time only slightly increases the alert probability of the pairs beyond 150 m from 0.01% to 4.5%.

7.3 Computation Complexity

The main complexity of Vecsim is to look up the DPD, which is a linear scan of the related pairs until a similar pair is found, or until all related pairs have been compared. Fig. 14 shows the the number of comparisons, where it can be seen that the complexity is reasonably low, thanks to the compression of the DPD, as the median of the number of comparisons is only 2.

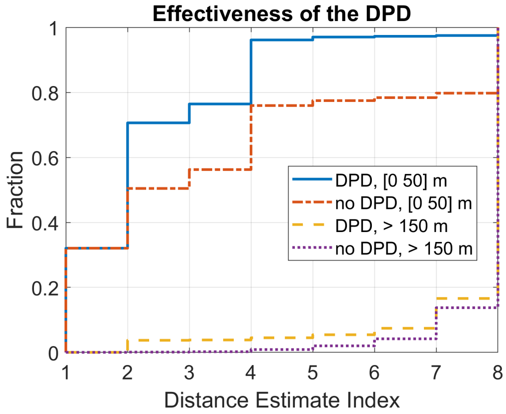

Figure 13: The performance with or without DPD.

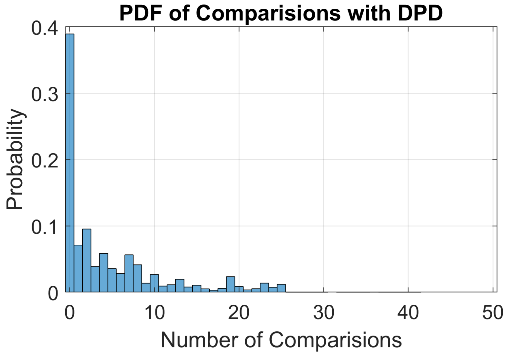

Figure 14: Number of comparisons in DPD look up.

7.4 Compression of DPD

Fig. 15 compares the performances with the uncompressed DPD and with compressed DPD. It can be seen that there is virtually no loss of performance due to compression.

7.5 Sensitivity to Noise

The last set of tests focus on the sensitivity of Vecsim to noise. Fig. 16 shows the alert probability of pairs within 50 m and beyond 150 m under various levels of AoA noise measured by the standard deviation. It can be seen that the AoA noise starts to affect the performance when the noise standard deviation is 0.1, which is understandable, because a slight change of AoA may lead to large change in the initial distance estimate if the Tadv is large.

Fig. 17 shows the alert probability of pairs within 50 m and beyond 150 m under various levels of RSRP noise measured by the standard deviation. It can be seen that, for pairs within 50 m, the alert probability almost stays the same when the noise standard deviation is 10 or less, and only starts to drop when the noise standard deviation is 15. For pairs beyond 150 m, the alert probability does not change significantly with the noise. Therefore, Vecsim appears to be highly robust to the noise in RSRP.

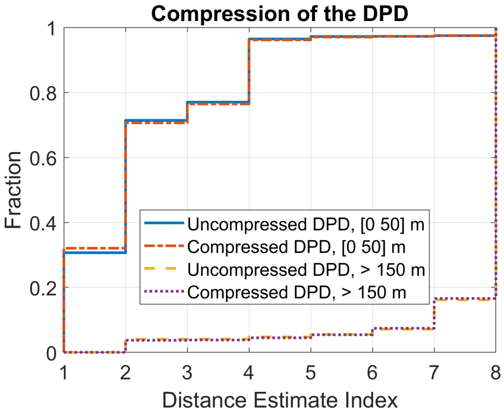

Figure 15: Comparison of the uncompressed DPD and compressed DPD.

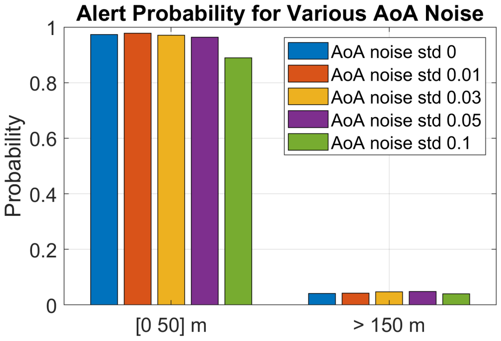

Figure 16: Vecsim performance for different levels of AoA noise.

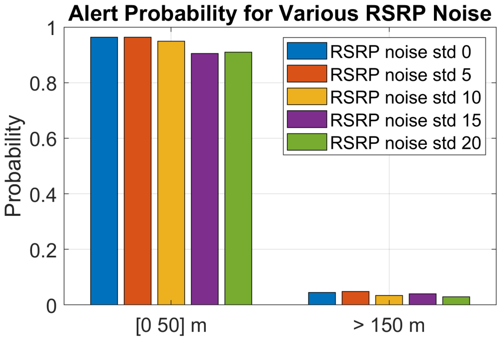

Figure 17: Vecsim performance for different levels of RSRP noise.

8 CONCLUSION

In this paper, we propose Vecsim, a carrier-based contact tracing method which preserves user privacy. Vecsim is different from contact tracing methods that require the users to install special software on their cellphones, because Vecsim is expected to be a software package run only on the carrier side. In addition, Vecsim preserves user privacy, because it does not require the location information to determine the proximity range between two users. Vecsim leverages the rich cellphone network log data that is already collected by carriers. In our current evaluation, Vecsim was able to alert more than 96% of users within 50 meters while alerting less than 4.5% of users beyond 150 meters. In contrast, the existing carrier-based solutions can provide only cell-level accuracy. We note that the performance of Vecsim continuously improves as more data is logged, which allows more discontinuity pairs to be learned. Overall, Vecsim provides a reliable alternative to traditional contact tracing methods that protects user-privacy and requires negligible financial cost to implement.

REFERENCES

[1] Singapore Government Agency. 2020. TraceTogether, safer together. www.tracetogether.gov.sg.

[2] N. Ahmed, R. A. Michelin, W. Xue, S. Ruj, R. Malaney, S. S. Kanhere, A. Seneviratne, W. Hu, H. Janicke, and S. K. Jha. 2020. A Survey of COVID-19 Contact Tracing Apps. *IEEE Access* (2020).

[3] Google Apple. 2020. Privacy-Preserving Contact Tracing. https://www.apple.com/covid19/contacttracing/.

[4] J. G. Cleary and G.Wyvill. 1988. Analysis of an algorithm for fast ray tracing using uniform space subdivision. *Vis. Comput.* 4, 2 (1988), 65–83.

[5] Epidemiology COVID-19 National Emergency Response Center, Korea Centers for Disease Control Case Management Team, and Prevention. 2020. Contact Transmission of COVID-19 in South Korea Novel Investigation Techniques for Tracing Contacts. (2020). www.ncbi.nlm.nih.gov/pmc/articles/PMC7045882/

[6] E. Damosso and L. M. Coreia. 1999. COST Action 231: Digital Mobile Radio Towards Future Generation Systems: Final Report. In *European Commissions*.

[7] V. Degli-Eposti, G. Lombardi, C. Passerini, and G. Riva. 2001. Wideband measurement and ray-tracing simulation of the 1900-MHz indoor propagation channel: Comparison criteria and results. *IEEE Trans. Antennas Propag* 49, 7 (Jul. 2001), 1101–1110.

[8] B. R. Epstein and D. L. Rhodes. 2010. GPU-accelerated ray tracing for electromagnetic propagation analysis. In *IEEE Int. Conf. Wireless Inf. Technol. Syst.*

[9] V. Erceg, S. J. Fortune, J. Ling, A. J. Rustako, and R. A. Valenzuela. 1997. Comparisons of a computer-based propagation prediction tool with experimental data collected in urban microcellular environments. *IEEE J. Sel. Areas Commun* 15, 4 (May 1997), 677–684.

[10] P. Flajolet, D. Gardy, and L. Thimonier. 1992. Birthday paradox, coupon collectors, caching algorithms and self-organizing search. *Discrete Applied Mathematics* 39, 3 (1992), 207 – 229.

[11] Center for Disease Control and Prevention. 2020. Identify the Primary Components of COVID-19 Contact Tracing. www.cdc.gov/coronavirus/2019-ncov/php/contact-tracing/identify-primary-components-of-contact-tracing.html.

[12] F. Fuschini, H. El-Sallabi, V. Degli-Esposti, L. Vuokko, D. Guiducci, and P. Vainikainen. 2008. Analysis of multipath propagation in urban environment through multidimensional measurements and advanced ray tracing simulation. *IEEE Trans. Antennas Propag* 56, 3 (Mar. 2008), 848–857.

[13] E. Gan and D. Culver. 2020. China is fighting the coronavirus with a digital QR code. Here's how it works. https://www.cnn.com/2020/04/15/asia/china-coronavirus-qr-code-intl-hnk/index.html.

[14] C. Gribble and J. Amstutz. 2015. Stingray: High-performance RF energy propagation modeling in complex environments. *DIAC Journal* 2, 2 (Spring 2015), 16–24.

[15] E. Hernandez-Orallo, P. Manzoni, C. T. Calafate, and J. Cano. 2020. Evaluating How Smartphone Contact Tracing Technology Can Reduce the Spread of Infectious Diseases: The Case of COVID-19. *IEEE Access* (2020).

[16] G. Liang and H. L. Bertoni. 1998. A new approach to 3-D ray tracing for propagation prediction in cities. *IEEE Trans. Antennas Propag.* 46, 6 (Jun. 1998), 853–863.

[17] H. Ling, R.-C. Chou, and S.-W. Lee. 1989. Shooting and bouncing rays: Calculating the RCS of an arbitrarily shaped cavity. *IEEE Trans. Antennas Propag* 37, 2 (Feb. 1989), 194–205.

[18] R. Margolies, R. A. Becker, S. D. Byers, S. Deb, R. Jana, S. Urbanek, and C. Volinsky. 2017. Can you find me now? Evaluation of network-based localization in a 4G LTE network. In *IEEE INFOCOM*.

[19] L. Ni, Y. Wang, H. Tang, Z. Yin, and Y. Shen. 2017. Accurate Localization Using LTE Signaling Data. In *IEEE International Conference on Computer and Information Technology*. 268–273.

[20] A. Ray, S. Deb, and P. Monogioudis. 2016. Localization of LTE measurement records with missing information. In *IEEE Infocom* (San Fransisco, CA, USA).

[21] Remcom. 2020. Wireless insite. http://www.remcom.com/wireless-insite.

[22] J. Tan, Z. Su, and Y. Long. 2015. A full 3-D GPU-based beam-tracing method for complex indoor environments propagation modeling. *IEEE Trans. Antennas Propag* 63, 6 (Jun. 2015), 2705–2718.

[23] S. Y. Tan and H. S. Tan. 1996. A microcellular communications propagation model based on the uniform theory of diffraction and multiple image theory. *IEEE Trans. Antennas Propag.* 44, 10 (Oct. 1996), 1317–1326.

[24] Minh Tran. 2015. Accurate Location Detection – 911 Help SMS App.

[25] Reinaldo A. Valenzuela. 1994. Ray tracing prediction of indoor radio propagation. In *PIMRC*.

[26] E. M. Vitucci, F. Mani, V. Degli-Esposti, and C. Oestges. 2012. Polarimetric properties of diffuse scattering from building walls: Experimental parameterization of a ray-tracing model. *IEEE Trans. Antennas Propag.* 60, 6 (Jun. 2012), 2961–2969.

[27] E Waltz. 2020. Back to Work: Wearables Track Social Distancing and Sick Employees in the Workplace. https://spectrum.ieee.org/the-human-os/biomedical/devices/wearables-track-social-distancing-sick-employees-workplace.

[28] E Yoneki. 2011. FluPhone Study: Virtual Disease Spread Using Haggle. In *Proceedings of the 6th ACM Workshop on Challenged Networks*. Association for Computing Machinery, New York, NY, USA, 65–66.

[29] Z. Yun and M. F. Iskander. 2015. Ray tracing for radio propagation modeling: principles and applications. *IEEE Access* 3 (2015), 1089–1100.

[30] K. Zhang, X. Liang, J. Ni, K. Yang, and X. Shen. 2018. Exploiting Social Network to Enhance Human-to-Human Infection Analysis without Privacy Leakage. *IEEE Transactions on Dependable and Secure Computing* (2018).