Limitations of scanned human copresence encounters for modelling proximity-borne malware

James Mitchell*, Eamonn O'Neill*, Gjergji Zyba†, Geoffrey M. Voelker†,
Michael Liljenstam‡, András Méhes‡ and Per Johansson†

*Department of Computer Science, University of Bath, Bath, UK
†University of California, San Diego, USA
‡Ericsson Research, Stockholm, Sweden

***Abstract*—Patterns of human encounters, which are difficult to observe directly, are fundamental to the propagation of mobile malware aimed at infecting devices in spatial proximity. We investigate errors introduced by using scanners that detect the presence of devices on the assumption that device copresence at a scanner corresponds to a device encounter. We show in an ideal static model that only 59% of inferred encounters correspond to actual device copresence. To investigate the effects of mobility, we use a simulator to compare encounters between devices with those inferred by scanners. We show that the statistical properties of scanned encounters differ from actual device encounters in ways which impact malware propagation dynamics, a form of aggressive data dissemination. In addition to helping us understand the limitations of encounter data gathered by scanners in the field, our use of virtual scanners suggests a practical method for using these empirical datasets to better inform simulations of proximity malware outbreaks and similar data dissemination applications.**

I. Introduction

Understanding the dynamics of propagation and assessing the effectiveness of countermeasures in outbreaks of self-replicating computer malware relies on understanding two factors: the mechanism by which the malware infects a susceptible host, and the patterns of contact between hosts. For network-borne malware these contacts are practically instantaneous and enabled by network topology rather than spatial relationships. For malware targeting cellphones, propagation may take place over the infrastructure network, but also through direct "proximity connections". Recently, malware has propagated over short-range radio connections, such as Bluetooth [1], [2]; at the same time, organizers of large events increasingly encourage Bluetooth activation for advertisements and crime prevention [3], increasing the risk of such malware threats. In the case of such proximity connections, the patterns of contact between people carrying the devices are critical in developing an understanding of the propagation dynamics.

When attempting to understand and model proximity-based propagation the availability of relevant and generalizable empirical data is limited. Here, we consider the approach of deploying scanners which use the same radio technology as devices carried by users. These scanners connect to users' devices when they pass within range and store information about the detected devices. Devices which are detected simultaneously by a given scanner are considered to be spatially co-located (corresponding to observed "encounters" within the scanner's radio range). The main benefit of such an approach is that once the scanners are deployed, large amounts of data can be gathered easily and at low cost, allowing longitudinal comparisons of encounter patterns. However, there are also some drawbacks when such scanner data is used as a basis for inferring proximity-based malware propagation dynamics: *(i)* Scanner deployments in the real world tend to be of relatively low density, typically covering a small fraction of an area under consideration, such as a campus or a part of a city. Hence, within this already limited area, the majority of encounters between devices will take place out of range of the scanners. *(ii)* Moreover, as we limit the area of consideration, we would expect the frequency with which particular devices appear at any scanner to decrease, artificially lengthening device inter-contact times.

Despite these obvious limitations, if scanned data is used carefully (i.e., accounting for the effects of missed encounters) it would still appear to be a good source of empirically-derived data on human encounters. Superficially, the encounters which *are* captured should consist of a subset of the actual device encounters taking place in the area under study at a particular time. In fact, we find that the process of inferring copresence encounters between pairs of devices based on empirical evidence of simultaneous sightings by third-party scanners leads to the introduction of errors. Here, we investigate the extent to which errors are introduced, and make the following contributions.

- We derive analytical results on errors introduced in scanner-based measurements for a simplified case where all scanners and devices are static, and where radio propagation details are omitted. We examine the differences between device copresence as inferred by the scanners and actual copresence between the devices, and classify the discrepancies.
- Based on this classification, we then derive the probabilities with which each type of discrepancy will occur. Using simulation we validate our analytical finding that approximately 41% of copresence encounters inferred by scanners do not correspond to actual device copresence.

978-1-4673-0298-2/12/$31.00

- Also using simulation, we demonstrate the extent and impact of errors when device mobility is included. As a concrete application, we study the effect on proximity-based malware propagation, an example of flooding-based data dissemination which depends heavily on the patterns of device encounters. We find that, in addition to the expected cases of missed and spuriously inferred encounters, the set of encounters inferred from scanners differs from the actual encounters simulated in the model in terms of duration distribution and probability of encountering previously unmet devices. While the magnitude of these errors increases when simulated mobility is more diffusive, in all the cases we considered malware propagation models showed slower propagation using scanned encounters compared to actual encounters for devices with the same mobility characteristics.

II. Related Work

Several studies have considered propagation dynamics of, and defence strategies against, malware using proximity-based propagation [1], [4]–[7]. However, access to empirical data on which to base such studies is currently limited. What is required is either direct data on malware propagation, which is not generally available, or information about device encounters, with which one can model propagation of malware.

In the absence of direct encounter data, human mobility data can also be used to infer encounters by considering spatial proximity between individuals. While human mobility data can be captured at a fine resolution under certain conditions, e.g., by using GPS traces, requiring users to record their movements is typically considered intrusive and onerous. As a result, studies which attempt to gather such data have typically involved dozens or, at most, hundreds of users in a limited geographical area [8], [9].

At the other end of the spectrum in terms of number of users and data granularity, recent work has utilized information from mobile network operators, where cell phone connectivity to the nearest base station can be tracked [10]–[12]. This provides coarse mobility data for large numbers of users in a potentially large geographic area, but does not make it possible to determine when individual devices encounter each other and, for instance, are able to connect via Bluetooth.

Another option, when interested in Bluetooth connectivity, is to directly collect data by deploying an application on participants' Bluetooth devices which periodically scans for discoverable Bluetooth devices in range. However, as in the case of gathering GPS data, this approach is challenging in terms of both user effort and privacy. The Reality Mining project collected Bluetooth traces from approximately 100 users over a period of nine months [13], using an application installed on cellphones; however, the devices scanned only once every five minutes, which is likely to have led to shorter copresence encounters going undetected. In [14], portable Bluetooth scanners were carried by up to twelve users for a period of five days, with scans being conducted every two minutes.

Some of the earliest papers on Bluetooth malware included empirical tests where Bluetooth devices were carried around to collect data on other discoverable Bluetooth devices encountered [1], [15]. In both of these studies, however, the volume of data gathered directly from devices was not sufficient to be used as a veridical source of encounter data for modelling a malware outbreak. Instead, [1] used encounter data from [13], while [15] used characteristics derived from their empirical data to parameterise a mobility trace generator based on social network theory [16].

Given the difficulty in gathering sufficient data directly from devices, some work has been done using fixed Bluetooth scanners to collect data on other Bluetooth devices passing by, including inference of what we call "copresence encounters" from the simultaneous presence of devices within the scanner's range [2], [17], [18].

The Bluetooth scanning approach enables the collection of data for large numbers of devices over long periods of time, but has some problems in studying proximity-based malware propagation. These issues have not been systematically examined in the literature.

III. From device contacts to inferred encounters

The data captured by scanners is not necessarily an accurate representation of the real contacts taking place between mobile devices—even if we consider only the subset of real contacts taking place within the scanner's range. Scanners infer copresence encounters when a device pair is simultaneously sighted at the same scanner. We assume throughout that the scanner has the same radio range as the devices. Simple geometry indicates that a scanner, if capable of the same radio range as the devices moving around it, will be able to make contact with pairs of devices which are simultaneously within range of the scanner, but not within range of each other. This effect, which we term "bridging", leads to the incorrect inference of encounters between devices which did not actually meet (see Figure 2).

A. Static Analysis

To begin to understand the relationship between scanned encounters and actual contacts between mobile devices, we first consider a simplified case: a single time instance in which all devices and scanners are static. We derive simple expressions for the expected number of different encounter types as seen by an array of fixed scanners.

We consider n devices, each equipped with a short-range radio. We assume that this radio behaves ideally, producing a disc of constant signal strength with radius r. The devices are uniformly distributed over a rectangular area of size $a \times b$, and are observed by m scanners using the same radio technology and placed in the same area. We assume that the coverage areas of scanners do not overlap.

Let $\mathbf{X}_i$ $(i = 1, \ldots, n)$ denote the position of each device. We assume that these 2-dimensional random vectors $\mathbf{X}_i$ are independent and identically uniformly distributed (*iid*) over the rectangle $[0, a] \times [0, b]$. Using a simplified "perfect disc"

radio propagation assumption, we say that two devices i and j are in contact if they are within radio range, that is $\|\mathbf{X}_i - \mathbf{X}_j\| \leq r$. With a slight abuse of notation, we write $r(\mathbf{X}_i, \mathbf{X}_j)$ as a shorthand for this relation, and $\overline{r}(\mathbf{X}_i, \mathbf{X}_j)$ as a shorthand for its negation (i.e., $\|\mathbf{X}_i - \mathbf{X}_j\| > r$).

Similarly let $\mathbf{y}_k$ $(k = 1, \ldots, m)$ denote the scanner positions. Then, using our shorthand notation, the event that device i is in range of the k^{th} scanner can be expressed as $r(\mathbf{X}_i, \mathbf{y}_k)$.

In our model, a scanner k registers an *inferred contact* between two devices, i and j, if both devices are simultaneously within the scanner's range; or, more formally:

$$r(\mathbf{X}_i, \mathbf{y}_k) \wedge r(\mathbf{X}_j, \mathbf{y}_k).$$

Note that an inferred contact need not correspond to an actual device contact, since two devices may both be in range of the same scanner without being in range of one another (the "bridging effect"). On the other hand, not all actual device contacts will be inferred by a scanner, since either one or both of the devices involved in a contact with each other may be outside scanner range. The following analysis classifies all possible relationships between device contacts and the contacts inferred by scanners.

1) Types of Contacts Considered: Not all device contacts can be inferred by scanners, and not all contacts which the scanners do observe correspond to actual contacts between devices. We set out to derive expressions for the expected number of:

- real contacts between devices,
- contacts inferred by scanners (consisting of):
  - correctly inferred contacts ("inferred real"),
  - incorrectly inferred contacts ("inferred fake"),
- real device contacts missed by scanners, because
  - one device is outside coverage ("missed one"),
  - both devices are outside coverage ("missed two").

Figure 1 shows a diagram relating these contact types. We assume a sufficiently sparse scanner arrangement to preclude missed contacts where both devices are in scanner range, but the two devices are in the range of two *different* scanners.

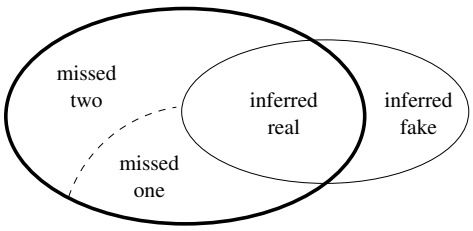


Fig. 1. Contact types: real (thick ellipse) and inferred (thin ellipse) contacts

2) Real Contacts: Intuitively, the expected number of device contacts (without regard to which of these are inferred or missed by the scanners) for n *iid* devices should equal the number of possible device pairs, $\binom{n}{2}$, times the probability of contact between any two devices. In the following derivations $\mathbb{P}$ denotes probabilities. Let us denote the ratio of a coverage area to the total observation area by

$$p \doteq \frac{r^2\pi}{ab}. \tag{1}$$

Now, one can write the expected number of real contacts as

$$c_{\text{real}} = \binom{n}{2} p(1-\delta) \tag{2}$$

where δ is an error term accounting for border effects. In the model, the border effects can be removed, e.g., by having edge wrap-around (effectively forming a torus), in which case the error term can be omitted.

3) Observed Contacts: For a single scanner at position $\mathbf{y} \in [r, a-r] \times [r, b-r]$, we have

$$\mathbb{P}\left[r(\mathbf{X}_i, \mathbf{y})\right] = p, \qquad \forall i.$$

Thus, the expected number of inferred contacts for this scanner equals $\binom{n}{2}p^2$; i.e., the number of device pairs times the probability that both devices "independently" fall inside the scanner's coverage area. For m scanners, whose coverage areas do not overlap and lie completely inside the measurement area, the above probability for a single scanner is simply multiplied by m to yield the expected number of inferred contacts.

$$c_{\text{inferred}} = m\binom{n}{2}p^2 \approx mp\ c_{\text{real}} \tag{3}$$

The result may appear intuitively satisfying, as it suggests that the scanners capture the fraction of real contacts corresponding to their combined coverage area. Unfortunately, this intuition is somewhat misleading, since a sizable portion of these inferred contacts are in fact "fake" and result from bridging, as we show next.

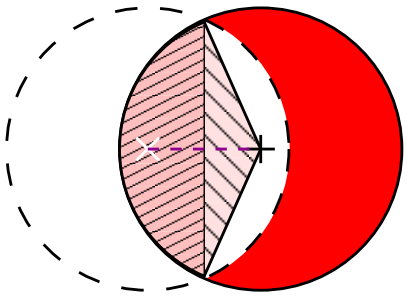


Fig. 2. Bridging probability. The position and coverage area of the scanner are shown by a + and the solid circle, while those of the first device by an × and a dashed circle. If the position of the second device falls inside the dark-shaded area to the right, the scanner will infer a contact while none actually occurs.

The bridging probability β is the conditional probability that two devices both inside the same scanner's coverage area are not within range of one another. Pictorially speaking, this corresponds to the average fraction of the darker half-moon-shaped area to the right in Figure 2. While we omit the full derivation for brevity, we have

$$\beta \doteq \mathbb{P}\left[\overline{r}(\mathbf{X}_1, \mathbf{X}_2) \,|\, r(\mathbf{X}_1, \mathbf{y}_1), r(\mathbf{X}_2, \mathbf{y}_1)\right] = \frac{3\sqrt{3}}{4\pi} \tag{4}$$

In other words, over 41% of inferred contacts are fake, introduced by bridging; and, consequently, only about 59% of inferred contacts correspond to real device contacts. The resulting formulas for the expected number of real and fake inferred contacts can be written as follows.

$$c_{\text{inferred,real}} = m\binom{n}{2}p^2(1-\beta) \qquad \approx 0.5865\ c_{\text{inferred}} \tag{5}$$

$$c_{\text{inferred,fake}} = m\binom{n}{2}p^2\beta \qquad \approx 0.4135\ c_{\text{inferred}} \tag{6}$$

4) Missed Contacts: As noted earlier, due to our assumptions on scanner placement, we consider only two types of missed contacts. In the first, one of the devices is inside scanner range while the other one is outside; and in the second type, both devices are outside scanner range.

Another careful look at Figure 2 reveals that the conditional probability of a device outside a given scanner's range being in range of another device inside the same scanner's range also equals the bridging probability β (as this case corresponds to the unshaded half-moon-shaped area on the right in the figure). Given that, viewed as an ordered pair, either of the two devices in this type of missed contact could be inside or outside scanner range, the following formula obtains:

$$c_{\text{missed,one}} = 2m\binom{n}{2}p^2\beta = 2\ c_{\text{inferred,fake}} \approx 0.827\ c_{\text{inferred}} \tag{7}$$

where one factor of p corresponds to $r(\mathbf{X}_1, \mathbf{y}_1)$ (i.e., the probability that the "first" device is in scanner range) and the other to $r(\mathbf{X}_1, \mathbf{X}_2)$ (i.e., that the two devices are within range of one another).

When both devices are outside scanner range, the analysis naturally splits into two sub-cases: one that accounts for border effects around the edges of the observation area, and another that accounts for similar effects near each scanner, when the scanner's and the devices' coverage areas intersect, as illustrated in Figure 3.

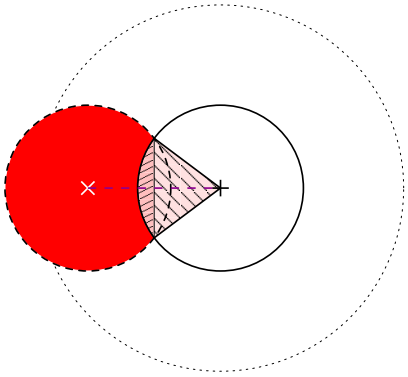


Fig. 3. Border effect near a scanner. If the first device is within distance $2r$ of a scanner (dotted circle), the position of the second device must fall inside the dark-shaded area to the left in order to guarantee that both devices end up outside scanner range.

The geometry of the dark shaded area in Figure 3 is essentially the same as the geometry of the dark shaded area in Figure 2, with the notable exception that the distance between the scanner's and the device's position varies between r and $2r$ (compared to 0 and r for the bridging probability). Consequently, λ_k, the corresponding conditional probability for scanner $\mathbf{y}_k$, can be obtained in the same fashion, giving:

$$\begin{aligned}\lambda_k &\doteq \mathbb{P}\left[\overline{r}(\mathbf{X}_2, \mathbf{y}_k) \,|\, r(\mathbf{X}_1, \mathbf{X}_2), 2r(\mathbf{X}_1, \mathbf{y}_k), \overline{r}(\mathbf{X}_1, \mathbf{y}_k)\right] \\ &= 1 - \frac{\sqrt{3}}{4\pi} = 1 - \frac{\beta}{3}\end{aligned} \tag{8}$$

Using equation (8) and subject to the constraints that:

a) the distance between any two scanners is at least $4r$, which avoids simultaneous interactions with multiple scanners, and
b) the distance between any scanner and the edge of the observation area is at least $3r$, excluding interactions between scanners and edge effects,

the expected number of missed contacts due to both devices being outside scanner coverage can be derived as follows.

$$\begin{aligned}c_{\text{missed,two}} &= \binom{n}{2}\mathbb{P}\left[r(\mathbf{X}_1, \mathbf{X}_2), \bigwedge_l \overline{r}(\mathbf{X}_2, \mathbf{y}_l), \bigwedge_l \overline{r}(\mathbf{X}_1, \mathbf{y}_l)\right] \\ &= \binom{n}{2}\left(\mathbb{P}\left[r(\mathbf{X}_1, \mathbf{X}_2)\,\middle|\, \bigwedge_l \overline{2r}(\mathbf{X}_1, \mathbf{y}_l)\right](1-\delta-4mp) \right. \\ &\qquad \left. + \sum_k \lambda_k p(4p-p)\right) \\ &= \binom{n}{2}\left(p(1-\delta-4mp) + m(4p-p)p(1-\frac{\beta}{3})\right) \\ &= \binom{n}{2}p\left(1-\delta-mp(1+\beta)\right) \qquad (9) \\ &= c_{\text{real}} - \left(c_{\text{inferred,real}} + c_{\text{missed,one}}\right) \qquad (10)\end{aligned}$$

B. Validation

To validate our analysis, we performed a simple simulation. Using the same constraints as above on scanner arrangement, we randomly placed 5,000 devices and 144 scanners, both with a range $r = 10m$, within a simulated area of size $500m \times 500m$. We then compared the encounters recorded directly by the devices with the encounters inferred by the scanners over 100 simulation runs.

In our simulation an average of 41.347% (SD=1.29%) of encounters inferred by the scanners were "fake", that is, they did not correspond to pairs of devices which were in range of each other. We further found that the mean number of pairwise encounters missed by scanners because only one device was in range, divided by the total number of encounters which were inferred by scanners, was 0.829 (SD=0.044). These values compare to the expectations of 41.35% (Eq. 6) and 0.827 (Eq. 7) which were derived in our previous analysis.

IV. Scanner Errors in Mobile Data

In Figure 1 we defined the four possible classifications of contact types in a static scenario: *inferred-real*, *inferred-fake*, *missed-one*, *missed-two*. These contact types describe all the

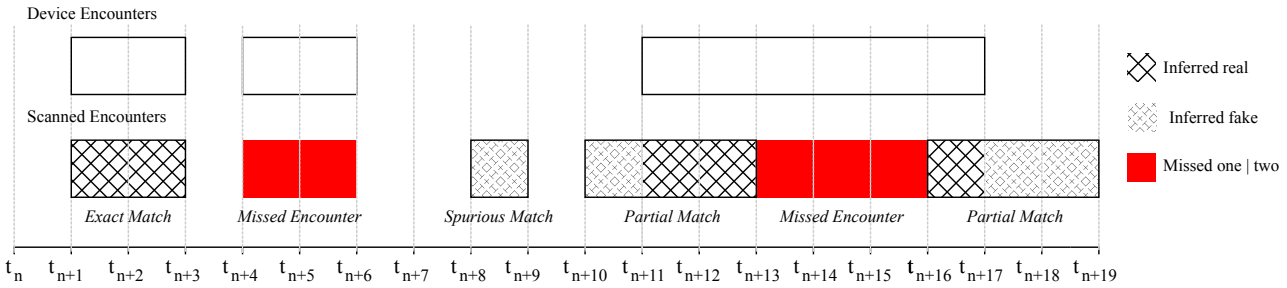


Fig. 4. Relating device contacts to scanned encounters

ways in which a scanner might (or might not) register an encounter between two devices at a particular moment in time.

When mobility and time are introduced, the way in which a scanner infers (or misses) an encounter can be thought of as a sequence of successive static contacts over a period of time. The relationship between the device encounter and what is inferred (or not) by the scanner falls into four possible classifications, each characterised by which of the four contact types makes up the sequence. We show the relationship between device encounters and scanned encounters, and their corresponding contact types, in Figure 4. The top time-series shows the ground truth of the encounters of two devices, and the bottom time-series shows the different ways a scanner may infer (or miss) device encounters. The shading on each scanned encounter relates to the contact types from which it is composed over time. While we separately considered the distinct probabilities of *missed-one* and *missed-two* in our static analysis, the distinction is not useful from here on; in both cases a device encounter goes undetected by all scanners.

- **Exact matches.** Encounters between two devices that take place entirely within scanner range. Devices are only within range of a scanner for the period of time that they are within range of each other. These encounters are composed entirely of *inferred-real* contacts, as in the encounter in Figure 4 beginning at time t_{n+1}.
- **Missed encounters.** Encounters between two devices that take place beyond the range of scanners are composed of *missed-one* or *missed-two* contacts. In Figure 4 this is shown by an unbordered solid area at time t_{n+4}.
- **Spurious matches.** Encounters between two devices when the devices are each within range of a scanner, yet are never actually in range of each other. These encounters are composed entirely of successive *inferred-fake* contacts, as in the encounter in Figure 4 at time t_{n+8}.
- **Partial matches.** Encounters between pairs of devices which are composed of more than one of the four types of contact (other than a mixture of *missed-one* and *missed-two*). Compared to the actual encounter the scanner may infer one or more longer or shorter encounters which partially overlap in time with the actual encounter. In Figure 4 two such encounters are shown beginning at times t_{n+10} and t_{n+16}.

The extent to which the encounters inferred (or missed) by scanners fall into each of these categories is critical when considering the degree of error which is likely to appear in malware propagation models based on them.

Consider an ideal scanner, which was somehow able to accurately infer all device encounters which passed within its range. For this scanner, all of its inferred encounters would fall into one of the first two categories above. As such, they would be composed entirely of *inferred real* and *missed one* or *missed two* contacts. By definition, the encounters inferred by this scanner would be a subset of the total set of device encounters which took place in the area under study. While those encounters which never passed in range of the scanner would be missed, if we assume that devices are dispersed uniformly around the area then we would expect the selection of encounters which the scanner did infer to be unbiased. As such, a deployment of a number of these ideal scanners would together collect a subset of total device encounters (once de-duplicated to account for encounters which pass through multiple scanners). We would expect the relative numbers of encounters inferred to reflect the ratio of scanner coverage area to the total area under study. If we compared the property distributions (e.g., encounter duration) of the scanner and device encounters, we would expect to see no difference. As a consequence, once we adjusted for the lower encounter rates in the scanner data, we would expect to see identical dynamics of propagation between the two sources of encounter data.

In practice, we can easily identify ways in which scanners deviate from ideal behaviour. In our previous analysis, we have shown that bridging leads scanners to incorrectly infer copresence in around 41% of device pair sightings. These incorrect inferences will give rise to *partial matches* and *spurious matches* appearing in the scanner data, which differ from, or do not actually appear in, the device encounters. In addition, when we consider moving devices over time, even in the absence of bridging, scanners can report encounters which differ from actual device encounters. For example, a pair of devices may remain in range of each other while moving on equal vectors. They may pass in and out of the ranges of a number of scanners, which will report numerous shorter encounters between the pair, instead of one continuous meeting. In this particular example, the scanner data will contain a number of *partial match* encounters. If we again

consider the distribution of encounter duration in the scanner encounters, a disparity will clearly be introduced compared to the device encounters.

Having shown the circumstances in which the use of scanners to infer copresence can introduce errors, the remainder of the paper investigates how often these errors occur, and the extent to which they lead to inaccurate estimates of proximity borne malware propagation as one concrete application.

A. Methodology

Compared to our analytical solution for the static case, a similar analytic model for the dynamic case of devices that are mobile over time is substantially more complex. On the other hand, obtaining empirical data with which to compare the incidence of errors in scanned encounters is also difficult. To understand the extent of errors introduced by the use of scanners in the case of mobile devices, we require data on both the real encounters (as detected by devices themselves) and the encounters inferred from scanners for the same set of devices at the same time. Most datasets consist of either high volume scanned data or relatively low volume GPS trace data, but not both.

Lacking empirical data, we instead use a mobility simulator to produce complete traces of mobile devices moving within a simulated two-dimensional space. To obtain a baseline for the actual encounters between devices, we process the mobility traces generated by the simulator to identify the encounters between pairs of devices over time. Using the same definition as in the static case above, we say that a pair of devices i and j with positions given by two-dimensional vectors $\mathbf{X}_i$ and $\mathbf{X}_j$ are within radio range while $\|\mathbf{X}_i - \mathbf{X}_j\| \leq r$. As the devices move over time, we say that they are in a pairwise encounter for any contiguous time period during which they remain in radio range, i.e., for an encounter between times t_m and t_n (where $m < n$):

$$\|\mathbf{X}_i(t_s) - \mathbf{X}_j(t_s)\| \leq r \qquad \forall s : m \leq s \leq n$$

We then simulated the deployment of "virtual scanners" in the simulated area to generate encounters inferred from scanner observations. From the perspective of a scanner, two devices have an encounter when both devices are simultaneously within range of the scanner for a specified time period.

With these two data sets, we can then compare the baseline "actual" encounters with the inferred "scanner" encounters to understand the nature and frequency of encounter errors introduced by the use of scanners. The use of simulation also allows us to investigate the effects of scanner density on the accuracy and completeness of scanned data by deploying up to thousands of scanners per square kilometer.

1) Mobility simulator: We employ a mobility simulator which implements the Lévy walk mobility model described in [9]. We note that a considerable variety of synthetic mobility models have been proposed over time, including models proposed after the Lévy walk model (e.g., SLAW [19], SWIM [20], and individual-mobility [21]). We sidestep debates about the "best" mobility model, and instead observe that the Lévy model has the merits of validation with large realistic traces [9] and is relatively popular and increasingly well understood (e.g., [22]). Other models might result in different absolute values for malware propagation times and encounter distributions, but, given the inherent approach of using scanners to infer device encounters, we believe that the effects we observe are illustrative of the problem and not the mobility model.

We consider a number of agents, each carrying a device with a radio range of $10m$ (a typical range for proximity communication using Bluetooth), and we consider a pair of devices to be copresent if both are within the other's radio range. As before, we make the simplifying assumption that radios produce a sharply-demarcated disc of constant signal strength. The agents move in steps, with each step being comprised of a *flight* — motion in a single direction θ (randomly chosen from a uniform distribution such that $0° \leq \theta \leq 360°$) — followed by a *pause*, during which the agent is stationary. For each step, the flight length and pause time are chosen randomly from two Lévy distributions respectively having scale factors α (flight length) and β (pause time). Additionally, for flight length and pause time values, we apply unit scale factors c and d, and maximum values t_f and t_p.

As in [9] flight time (and hence velocity) is related to flight length to reflect the greater probability that longer flights use a mode of transportation other than walking. Flight length is given by $t_f = kl^{1-p}$, where k and p are constants and $0 \leq p \leq 1$. For flights of less than $500m$, we use values of $k = 18.72$ and $p = 0.79$. For longer flights over $500m$, we use values of $k = 1.37$, $p = 0.46$. We set unit scale factors for flight length c of $10m$, and pause time d of 1 second in all simulations.

To reduce the effects of reflection on device mobility patterns, we define a large square area within which devices move ($3000m \times 3000m$). If devices reach the edge of this area, their flights reflect off the outer boundary and continue their current flight step. We also define a smaller central inner area ($1000m \times 1000m$) as the area of device interaction, and consider encounters between devices only within this area.

Inside the central inner area we deploy "virtual scanners" at fixed locations. These static scanners, like the mobile devices, have a $10m$ radio range and are placed at least $20m$ apart to avoid overlapping coverage areas. For simplicity of implementation, the scanners were placed on a square lattice, resulting in scanner coverage of 79% of the total area.

To investigate whether differing mobility parameters affected the extent and nature of errors in the scanned data, we use three sets of parameters for the scale factors of flight length (α) and pause time (β) distributions in the Lévy walk model. Each of these three pairs of α and β values represent simulation parameters found to fit well with empirical GPS datasets gathered from sets of walkers in three separate locations [9]: San Francisco ($\alpha = 0.75$, $\beta = 1.68$), NCSU ($\alpha = 0.86$, $\beta = 0.99$) and KAIST ($\alpha = 0.97$, $\beta = 0.45$). For each of the three mobility parameter sets we performed 25 simulation runs lasting one week of simulated time, each for 900 devices. We assume all devices are susceptible, corresponding to malware

propagation among mobile users who share devices with the same platform (and are a subset of all mobile users [2]). In each case, we deployed 2,500 scanners within the inner area ($1000m \times 1000m$) of the simulation.

B. Simulation results

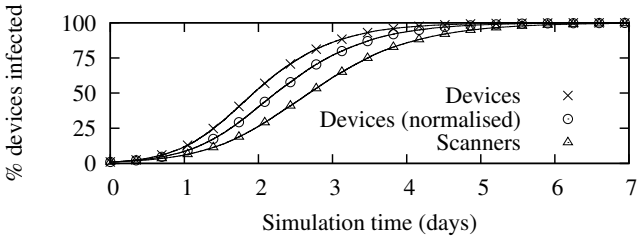
Our simulations produced datasets containing, for each mobility trace, a set of device encounters sensed by the mobile devices themselves, and a set of scanned encounters inferred by the "virtual scanners". In comparing the two sets of encounters, our aim was to highlight the errors introduced by incorrect inferences leading to *partial match* and *spurious match* scanner encounters and their impact on simulations of malware propagation models using the encounter data.

1) Comparing malware propagation dynamics: As an initial test of our assertion that the use of scanned encounter data may lead to inaccurate estimates of malware propagation, we performed a simple malware propagation simulation using the encounter data from our simulator. At a high level, proximity-based malware propagation is a form of data dissemination in opportunistic ad-hoc networks. As such, since such malware propagation strongly depends upon the distributions of device contacts and contact durations, it is particularly useful for evaluating the sensitivity of such data dissemination applications to errors in device encounter data. As discussed above, our aim is to investigate the errors which arise in encounter data as a result of the scanners' deviation from ideal behaviour. A deployment of ideal scanners would infer a subset of device encounters, selected without bias, whose size is related to the proportion of area under scanner coverage.

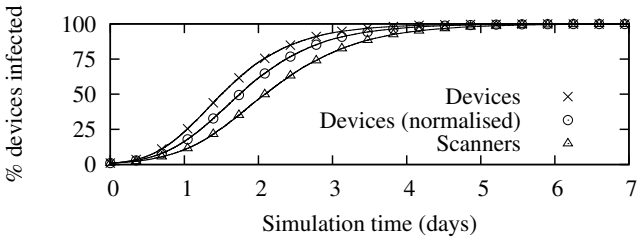
Since scanner coverage in our simulation was incomplete, we would not expect propagation between the scanned and device encounters to match, even in the unlikely event that our virtual scanners behaved ideally. To control for the effects of incomplete scanner coverage, we created a normalised set of device encounters for use in our propagation model. This dataset consists of a subset sampled at random from the set of device encounters such that, for each mobility trace, the subset contains the same number of encounters as the corresponding set of scanned encounters. While we make no attempt to match the particular encounters taking place at scanner sites in this subset, we would expect the aggregate characteristics of the encounters to match those which our scanners would have inferred had they behaved ideally.

Figure 5 shows mean propagation over time over 100 runs on each mobility trace. Each simulation run assumed one initially infected device in a standard susceptible-infected (SI) model, with all devices susceptible and a latency for propagation of 30 seconds. For all three of the mobility parameter sets we see, as expected, that propagation proceeds more slowly in the "normalised" subset of device encounters.

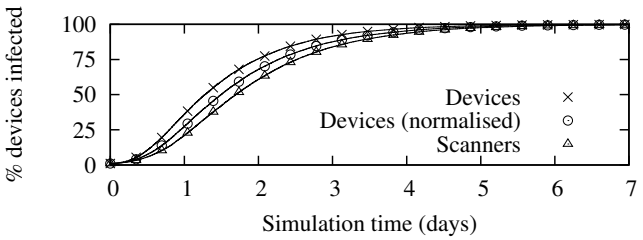
It is also apparent that when the scanned encounter sets are used in the model, despite having the same number of encounters as the normalised subsets of device encounters, propagation is slower still in all three mobility traces. The difference between propagation using the normalised device encounters and the scanner encounters is large. After two days of simulation time, the respective mean proportions of infected devices are: SF 40.2% vs. 26.4%, NCSU 61.6% vs. 46.4%, and KAIST 67.9% vs. 60.8%.



(a) San Francisco mobility parameters



(b) NCSU mobility parameters



(c) KAIST mobility parameters

Fig. 5. Malware propagation, comparing device encounters, normalised device encounters and scanned encounters.

This experiment shows that the use of scanned encounters, when compared to subsets of device encounters, leads to an underestimation of propagation rates in proximity malware models. The deviation suggests that our scanners are not behaving ideally, and are introducing wrongly inferred *partial match* and *spurious match* encounters. The extent of these errors of inference is sufficient to alter the characteristics of the whole set of device encounters.

2) Encounter overlaps: Our malware propagation simulation shows that the sets of scanned encounters differ from those detected directly by the mobile devices, and that this difference is attributable to incorrect inferences by the scanners which lead to the reporting of erroneous encounters. To better understand the nature of these erroneous encounters, we directly compared the scanner and encounter data from each individual mobility trace. By considering the encounters between each device pair which met (or was inferred to have met) at least once during the simulation, we show the proportion of encounter types present in the scanned encounter

data (including *missed encounters*).

For each encounter between devices, we determined whether, in the set of scanned encounters, an encounter between the same two devices existed with the same start and end time (an *exact match*), or whether one or more *partial matches* existed which overlapped it in time. Device encounters where no match or overlap was found correspond to *missed encounters* (combining *missed-one* and *missed-two* static contacts). Repeating the process from the perspective of the scanned encounters revealed the *spurious matches* which did not overlap any device encounters.

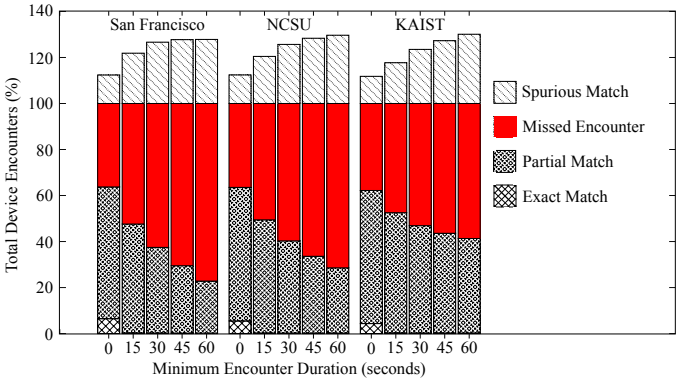


Fig. 6. Relationships between device and scanner encounters by encounter type.

We first compared all encounters in the device and scanned encounter sets for each mobility parameter set (i.e., a minimum encounter duration of zero). We see that approximately two-thirds of device encounters have a corresponding *exact match* or *partial match* in the scanned encounters. There is little difference in this proportion between the three mobility sets. The incidence of exact matches is low, representing 6.5%, 5.5% and 4.4% of total device encounters in the San Francisco, NCSU and KAIST mobility parameters, meaning that almost all of the encounters actually inferred by the scanners are either *partial matches* or *spurious matches*.

Given the sensitivity of proximity-based malware propagation to encounter duration (since infection latency may be 30 seconds or more), we repeated the experiment using subsets of the device and scanner encounters with successively higher minimum durations. For all of the mobility traces, we first see that, once a minimum threshold of 15 seconds is imposed, the proportion of exact matches becomes vanishingly small. Where exact matches do occur, they are typically between encounters of short duration. This result is not unexpected, and suggests that very short encounters may simply offer less opportunity for erroneous inferences to take place.

We also see that the proportion of missed encounters and spurious encounters rises for all three mobility traces as the minimum encounter duration threshold increases. A calculation of correlation between the proportion of missed encounters to total device encounters and minimum latency threshold suggests that a strong relationship exists in all cases (SF $r^2 = 0.96$, NCSU $r^2 = 0.96$, KAIST $r^2 = 0.92$). A similar calculation of correlation between the proportion of spurious encounters and minimum encounter duration showed a less strong relationship which appeared to strengthen in the less diffusive mobility parameter sets (SF $r^2 = 0.71$, NCSU $r^2 = 0.82$, KAIST $r^2 = 0.89$).

TABLE I
ENCOUNTERS EXCEEDING EXAMPLE MALWARE PROPAGATION LATENCIES (IN SECONDS)

	Device encs. (m)	Scanned encs. (m)	$\frac{P(D)}{P(S)}$	$\frac{n(D)}{n(S)}$
SF				
All	3.20 (100.0%)	2.84 (100.0%)	0.0%	-11.1%
>15	0.38 (11.9%)	0.28 (10.0%)	-15.9%	-25.2%
>30	0.14 (4.3%)	0.09 (3.2%)	-25.6%	-33.8%
>45	0.57 (1.8%)	0.03 (1.2%)	-34.7%	-41.9%
>60	0.02 (0.7%)	0.01 (0.4%)	-42.6%	-49.0%
NCSU				
All	2.53 (100.0%)	2.24 (100.0%)	0.0%	-11.4%
>15	0.42 (16.5%)	0.32 (14.2%)	-13.8%	-23.6%
>30	0.17 (6.8%)	0.12 (5.3%)	-22.1%	-31.0%
>45	0.80 (3.2%)	0.05 (2.3%)	-28.4%	-36.6%
>60	0.04 (1.5%)	0.02 (1.0%)	-33.2%	-40.9%
KAIST				
All	1.40 (100.0%)	1.19 (100.0%)	0.0%	-15.3%
>15	0.37 (26.1%)	0.29 (24.0%)	-8.0%	-22.0%
>30	0.19 (13.9%)	0.15 (12.4%)	-10.6%	-24.3%
>45	0.12 (8.6%)	0.09 (7.6%)	-11.8%	-25.3%
>60	0.08 (5.7%)	0.06 (5.0%)	-12.4%	-25.8%

The relationship between encounter length and proportion of missed encounters appears counter-intuitive. While practically all encounters over 15 seconds for all mobility traces do not match exactly between the scanner and device encounters, we had expected that longer encounters would experience a higher level of partial overlaps, if only by chance. Our experiments showing the opposite to be true suggest that there are differences between the distribution of encounter durations in the scanner and encounter datasets, with the scanner datasets simply including fewer long encounters to match against the device encounters.

3) Encounter duration: The correlation between increasing encounter duration and incidence of missed and spuriously matched encounters in the scanner data led us to investigate the distribution of encounter duration between the device and scanner encounter sets. The duration of encounters is important when modelling the propagation of proximity-borne malware, where propagation between devices might occur only during uninterrupted connections of 30 seconds or more.

For each of the three mobility parameter sets, all simulation runs were combined to produce large sets of device encounters and scanner encounters. We calculated the proportion of encounters within each encounter set which were longer than a set of latency thresholds for proximity malware transmission. As Table I shows, in all cases a smaller proportion of the scanned encounters exceeds the latency thresholds. In other words, the scanned encounters underestimate the duration of

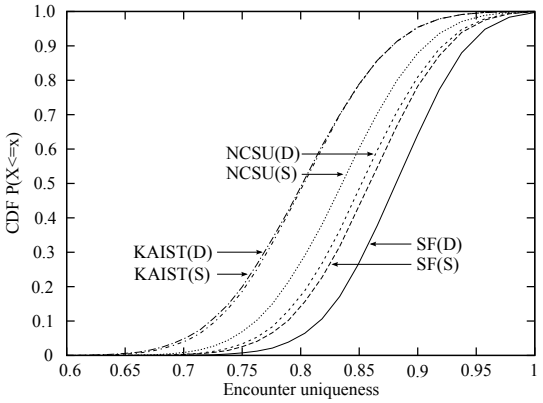


Fig. 7. "Uniqueness" of device vs. scanner encounters

the device encounters considerably. For longer but still realistic latency values of one minute, the proportion of scanned encounters lasting long enough to allow malware propagation to occur is lower than the device encounters by between 12% and 42% across the three sets of mobility parameters.

4) Encounter "uniqueness": We have shown that the scanned encounters generated by our simulator differ from the device encounters by having characteristics which lead to underestimation of malware propagation when these scanned encounters are used as source data. To further investigate the extent to which the scanned encounters differ from the device encounters, we compared the distribution of contact degrees of each set of encounter data. The distribution of contact degrees is a key driver of malware propagation: the dynamics of epidemic spread in networks with heavy-tailed distributions of encounter degrees differ significantly from "fully-mixed" models in which all agents are equally likely to meet [23].

However, contact degree is related to encounter rates, and the incomplete coverage of the area provided by scanners will likely result in lower encounter rates. As a result the contact degrees of device and scanned encounters cannot be directly compared. To address this, we calculated a normalised metric, *encounter uniqueness*, which is the proportion of unique devices within the total devices encountered in a given period. In the case where a device meets each other device only once, all its encounters can be described as *unique*, giving a value of 1.0. As the proportion of encounters with previously-seen devices increases the ratio of unique encounters falls. For encounters with similar distributions of encounter duration, we would expect higher encounter uniqueness to correspond to increased rates of malware propagation.

To ensure comparability across our simulation data, we calculated the uniqueness values for encounters from the simulation start until each device in the simulation had met a given number of unique devices. We repeated this process for each of the mobility traces across all three mobility parameter sets. Figure 7 shows the distribution of the uniqueness ratio for the device and scanned encounters.

As expected, the more diffusive mobility parameter sets (NCSU, SF) show a higher encounter uniqueness. Longer flight lengths mean that devices are less likely to repeatedly encounter devices they have previously met. However, the two more diffusive mobility parameter sets are also most affected by underestimation of encounter uniqueness in the scanned encounters, while the least diffusive mobility parameter set (KAIST) shows very little difference in encounter uniqueness between the device encounters and the scanned encounters.

V. Conclusions and Future Work

Our detailed examination of errors induced by inferring device encounters from third party scanners suggests caution in the use of such data sets, for instance, for the study of flooding-based data dissemination applications like proximity malware propagation. However, it is also suggestive of a potential way forward.

We have demonstrated the circumstances in which "bridging" errors between pairs of out-of-range devices occur. Further, we have shown, under assumptions of equal and homogeneous communication ranges, that over 41% of device encounters inferred from simultaneous scanner sightings were incorrect. In the case of mobile devices, these incorrect inferences have a complex effect on the accuracy of scanned encounters as they accumulate over time. As well as encounters which are missed or spuriously inferred by considering sightings at scanners, unreliable inference results in inferred encounters which have shorter durations than the actual encounters between devices, and underestimates the extent to which the devices encounter new, unmet devices.

The magnitude of these differences is sensitive to the underlying mobility characteristics of the devices being scanned, with more diffusive mobility correlating with increased errors. In all three sets of mobility parameters we tested (each closely matching GPS trails gathered from human movement), the extent of errors introduced through inferring copresence by simultaneous presence at scanners led to a significant underestimation of the rate at which proximity-based malware would spread amongst devices.

On the other hand, our use of a mobility simulator to compare actual encounters observed from mobility trails with encounters inferred from sightings at scanners suggests a method for mitigating erroneous inferences of copresence in data gathered by scanners deployed in the field. In cases of highly diffusive mobility, where the errors introduced by bridging appear to be most pronounced, the quality of scanned encounter data might be materially improved, leading to more accurate simulations of malware spread and countermeasures.

Estimated or observed characteristics of mobility patterns around the scanners, such as the distribution of velocities, flight lengths and pause times would be used to set initial parameters for a mobility simulator. This simulator would then be populated with virtual scanners similar to those used in the field, and used to infer simulated encounters. The mobility parameters used in the simulation could then be improved iteratively until the simulated encounters closely matched the statistical properties of those gathered from the field scanners.

The malware propagation model could then be based on the direct encounters between devices in the simulator. Since the same fundamental geometry leads to errors in simulated scanners and the real deployed scanners we would expect the incidence of bridging errors in both cases to be similar, provided the simulator's mobility parameters closely match the observed characteristics of mobility around the deployed scanners. This being the case, the direct encounters between devices in the simulator should capture the observed properties of human mobility at the scanner sites, while reducing errors from incorrect inferences — and in doing so be closer to the real human encounters which took place around the scanners.

REFERENCES

[1] J. Su, K. K. W. Chan, A. G. Miklas, K. Po, A. Akhavan, S. Saroiu, E. de Lara, and A. Goel, "A preliminary investigation of worm infections in a bluetooth environment," in *Proc. ACM WORM 2006*, 2006.

[2] G. Zyba, G. M. Voelker, M. Liljenstam, A. Méhes, and P. Johansson, "Defending mobile phones from proximity malware," in *Proc. IEEE INFOCOM 2009*, 2009.

[3] North Wales Police, "Radio One Big Weekend in Bangor," http://bit.ly/8YETHC (accessed June 2010).

[4] J. W. Mickens and B. D. Noble, "Modeling epidemic spreading in mobile environments," in *Proc. ACM WiSe 2005*, 2005, pp. 77–86.

[5] E. Yoneki, P. Hui, and J. Crowcroft, "Wireless Epidemic Spread in Dynamic Human Networks," in *BIOWIRE 2007 Cambridge, Revised Selected Papers*. Springer, 2008, p. 116.

[6] G. Yan and S. Eidenbenz, "Modeling Propagation Dynamics of Bluetooth Worms," in *Proc. IEEE ICDCS 2007*, 2007, p. 42.

[7] A. Bose, X. Hu, K. G. Shin, and T. Park, "Behavioral detection of malware on mobile handsets," in *Proc. ACM MobiSys 2008*, 2008, pp. 225–238.

[8] M. Kim, D. Kotz, and S. Kim, "Extracting a mobility model from real user traces," in *Proc. IEEE INFOCOM 2006*, 2006.

[9] I. Rhee, M. Shin, S. Hong, K. Lee, and S. Chong, "On the Levy-Walk nature of human mobility," in *Proc. IEEE INFOCOM 2008*, 2008, pp. 924–932.

[10] M. Gonzalez, C. Hidalgo, and A. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, Jun. 2008. [Online]. Available: http://dx.doi.org/10.1038/nature06958

[11] C. Song, Z. Qu, N. Blumm, and A. Barabási, "Limits of Predictability in Human Mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010. [Online]. Available: http://www.sciencemag.org/cgi/content/abstract/327/5968/1018

[12] H. Zang and J. C. Bolot, "Mining call and mobility data to improve paging efficiency in cellular networks," in *Proc. ACM MobiCom 2007*, 2007, pp. 123–134.

[13] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal and Ubiquitous Computing*, vol. 10, pp. 255–268, 2006. [Online]. Available: http://dx.doi.org/10.1007/s00779-005-0046-3

[14] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Pocket switched networks: Real-world mobility and its consequences for opportunistic forwarding," *University of Cambridge, Computer Lab, Technical Report. UCAM-CL-TR-617*, 2005.

[15] L. Carettoni, C. Merloni, and S. Zanero, "Studying bluetooth malware propagation: The bluebag project," *IEEE Security and Privacy*, vol. 5, no. 2, pp. 17–25, 2007.

[16] M. Musolesi and C. Mascolo, "A community based mobility model for ad hoc network research," in *Proc. ACM REALMAN 2006*, 2006, p. 38.

[17] E. O'Neill, V. Kostakos, T. Kindberg, A. Schiek, A. Penn, D. Stanton Fraser, and T. Jones, "Instrumenting the city: developing methods for observing and understanding the digital cityscape," in *Proc. UbiComp 2006*, 2006.

[18] V. Kostakos, E. O'Neill, A. Penn, G. Roussos, and D. Papadongonas, "Brief encounters: Sensing, modeling and visualizing urban mobility and copresence networks," *ACM Trans. Comput.-Hum. Interact.*, vol. 17, no. 1, pp. 1–38, 2010.

[19] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "SLAW: A Mobility Model for Human Walks," in *Proc. IEEE INFOCOM 2009*, 2009, pp. 2106–2113.

[20] S. Kosta, A. Mei, and J. Stefa, "Small World in Motion (SWIM): Modeling Communities in Ad-Hoc Mobile Networking," in *Proc. of the 7th IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON 2010)*, 2010.

[21] C. Song, T. Koren, P. Wang, and A. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, no. 10, pp. 818–823, 2010. [Online]. Available: http://www.nature.com/nphys/journal/v6/n10/full/nphys1760.html

[22] K. Lee, Y. Kim, S. Chong, I. Rhee, and Y. Yi, "Delay-Capacity Tradeoffs for Mobile Networks with Lévy Walks and Lévy Flights," in *Proc. IEEE INFOCOM 2011*, 2011, pp. 3128–3136.

[23] M. E. J. Newman, "Spread of epidemic disease on networks," *Phys. Rev. E*, vol. 66, no. 1, p. 016128, 2002.