Dissemination in Opportunistic Mobile Ad-hoc Networks: the Power of the Crowd

Gjergji Zyba and Geoffrey M. Voelker
University of California, San Diego
9500 Gilman Drive, Dept 0404
La Jolla, CA 92092-0404, USA
Email: {gzyba, voelker}@cs.ucsd.edu

Stratis Ioannidis and Christophe Diot
Technicolor
735 Emerson Street
Palo Alto, CA 94301, USA
Email: {stratis.ioannidis, christophe.diot}@technicolor.com

***Abstract*—Opportunistic ad-hoc communication enables portable devices such as smartphones to effectively exchange information, taking advantage of their mobility and locality. The nature of human interaction makes information dissemination using such networks challenging. We use three different experimental traces to study fundamental properties of human interactions. We break our traces down in multiple areas and classify mobile users in each area according to their social behavior: Socials are devices that show up frequently or periodically, while Vagabonds represent the rest of the population. We find that in most cases the majority of the population consists of Vagabonds. We evaluate the relative role of these two groups of users in data dissemination. Surprisingly, we observe that under certain circumstances, which appear to be common in real life situations, the effectiveness of dissemination predominantly depends on the number of users in each class rather than their social behavior, contradicting some of the previous observations. We validate and extend the findings of our experimental study through a mathematical analysis.**

I. Introduction

Independently of what technology they rely on, opportunistic mobile ad-hoc networks will allow users of portable devices such as smartphones and netbooks to communicate in a natural and effective way, taking advantage of locality and mobility to increase information exchange opportunities. The potential of epidemic dissemination is huge, enabling, for instance, a wide range of mobile ad-hoc communication and social networking applications supported entirely through opportunistic contacts in the physical world [1]. However, communication in such opportunistic mobile ad-hoc networks is challenging due to the volatility of contacts, communication technologies, and resource limitations (*e.g.*, batteries, communication opportunities, wireless data transmission technologies). Communication is also strongly impacted by human mobility, which is driven by user social behavior.

Despite substantial work in the area, both theoretical and experimental, our understanding of these networks is limited. Progress in understanding opportunistic mobile ad-hoc networks is mainly limited by the difficulty to collect complete traces, and to model large systems with realistic assumptions (which is linked to the absence of large experimental data sets). The main difficulty in the experimental approach is to collect traces that (i) contain enough information about each device (in particular its mobility, social profile of its owner, *exhaustive* list of contact opportunities, duration of contacts and communication technology impact) and (ii) are not biased by constraints due to experimental conditions.

In particular, there is a need to collect and consider data that encompasses the behavior of all devices in a population—not just experimental devices—to have a complete view of the experimental environment. Indeed, most data sets collect information in a pre-defined experimental population, such as participants carrying GPS receivers [2], Bluetooth sensors [3] and smartphones [4], and WiFi PDAs [5]. These data sets have at best a partial view of the environment, and of the role non-experimental devices could play in data dissemination. This situation is best illustrated by the Hong Kong trace explored in [6] where the experimental devices have strictly no direct contact with each other, yet they contact thousands of external devices that could play an important role in data dissemination but for whom it is not possible to collect data.

We use publicly available traces to improve the understanding of information dissemination in opportunistic mobile ad-hoc networks. We overcome the limitations identified above by choosing traces that collect information about all devices in an area (and not only a limited set of experimental devices). We further process these traces by subdividing each trace based on a specific social or professional geographical area of interest. We observe that a significant amount of devices appear rarely within a given area, and because of their large population, we explore their impact on information dissemination. In each sub-trace, we define two classes of populations with different presence characteristics, namely *Socials* and *Vagabonds*. Socials are individuals who return periodically to a specific area (analogous to the experimental devices in the discussion above, or to community members). Vagabonds instead are seen more rarely and randomly (*i.e.*, the external devices that are in general not measured, or removed from traces because of partial information). A device can be a Vagabond in one area, and Social in another as well as change its role over time, thus exhibiting both spacial and temporal characteristics.

The first contribution of our work is to study, for the first time, data dissemination spanning a large range of Social and Vagabond compositions. Previously, most studies consider Socials only and ignore Vagabonds entirely, or have just a partial knowledge of them because of experimental conditions.

Second, we observe that the efficiency of content propagation is not only a consequence of the devices' social status, but also a consequence of the number and density of devices. We see that in many cases, due to their large population, Vagabonds are more effective in spreading a message, even though they are considered unimportant. They therefore play a key role in information dissemination and they should not be ignored. This result contrasts previous works that focused only on the effect of social properties on dissemination [6]–[8].

Third, we study both experimentally and analytically the "tipping" point beyond which the population size becomes more significant than the social status. We do so by observing this behavior on our traces but also by developing an analytic model that formally characterizes the relationship between population size and the social behavior of users. Our analysis confirms our experimental results and identifies a simple formula for determining when data dissemination through Vagabonds outperforms dissemination through Socials.

Section II reviews related literature, and Section III describes the data sets we use in this study. Section IV introduces three possible definitions of the Social and Vagabond groups, and analyzes their properties in each area. Using the most promising definition, we study the mobility characteristics of Socials and Vagabonds in Section V. Then we analyze the impact and role of each group on content propagation using trace-driven simulations in Section VI. Finally, we formulate an analytical model that captures Social and Vagabond mobility properties to explain and extend our results in Section VII, and conclude in Section VIII.

II. Related work

Exploiting social behavior in opportunistic mobile networks has recently received considerable attention. Routing protocols such as SimBet [8], [9], Bubble Rap [6] and PeopleRank [7] use social-based metrics derived from contacts between devices (such as betweenness centrality and neighborhood similarity) to make opportunistic forwarding decisions with low overhead. Protocols using explicit knowledge of friendship relationships have also been proposed and shown to improve efficiency over socially agnostic protocols [10], [11].

All of the above protocols route over "strong ties" among mobile users, inferred either from contact behavior or declared friendships. Our work extends these previous efforts, exploring the role and potential of non-social, vagabond devices for communication and data dissemination. Previous routing protocols ignore such devices and, to the best of our knowledge, our work is the first to study their effect on data dissemination.

Beyond routing, social networking concepts have been used in mobile opportunistic applications such as publish/subscribe systems [12], [13], newsfeed [14] and query propagation [15], [16], and multicasting [1], [17]. These systems make use of social networking concepts like node centrality [13], friendship relationships [1], [13], [15], hotspots [16], contact usefulness [12] and edge expansion [14]. Our analysis focuses mostly on epidemic message dissemination; nevertheless, our understanding on the effect of Vagabonds motivates further study of their effect on the behavior of applications like the ones described above.

Data Set	Pop.	Length	Area	Pop. type	Log Freq.
San Francisco	483	24 days	City	Cabs	1–3 mins
Dartmouth	4248	60 days	Campus	Devices	Instant
Second Life	2713	10 days	Small	Avatars	1–3 mins

TABLE I: Basic characteristics of the data sets: population size, trace length, type of area, population type and logging frequency. The population size is the number of devices that have at least one contact with another device.

III. Data Sets

We use traces from three data sets.[1] We specifically chose these traces because they represent distinct and considerably different mobile environments. We avoid using traces of experimental devices only (*e.g.*, participants in a conference) unless all existing devices (even the ones not seen by the experimental devices) are monitored. We refer to these data sets as *Dartmouth*, *San Francisco (SF)* and *Second Life (SL)*, according to the location where they were collected. We further subdivide Dartmouth and SF into smaller geographical areas which have different social behavior characteristics. Table I summarizes the basic characteristics of the three data sets we consider. We discuss below the features of each data set and our motivation for using them.

a) Dartmouth: The Dartmouth data set comprises logs of association and disassociation events between wireless devices and access points at Dartmouth College [18]. The logs span 60 days and include events from 4920 devices. Of these, 4248 have at least one contact with another device, and we focus our study on these devices. As with many previous studies using WiFi traces (*e.g.,* [3], [5]), we assume that two devices are "in contact" when associated with the same access point.

We identify three areas within the Dartmouth campus likely visited by different social communities: *Engineering* (300m×200m), and *Medical* (300m×300m) are specific schools while *Dining* (150m×150m) corresponds to the main food court of the Dartmouth College campus where we expect all students to mix. The main features of this data set are that (1) it logs *all* WiFi devices on campus, as opposed to only preselected experimental devices in prior work [6]–[8], and (2) each region represents different social behavior in a university environment. However, the assumption that contacts take place between any two devices associated to the same access point may introduce a bias compared to real contact opportunities.

b) San Francisco: The San Francisco data set consists of GPS coordinates of 483 cabs operating in the San Francisco area [19], collected over a period of three consecutive weeks. We assume that any two cabs can communicate whenever their distance is less than 100 meters, a realistic range for WiFi transmissions.[2] We select three regions of San Francisco in which we expect cabs to exhibit different behavior. We refer

[1] Two available through CRAWDAD at http://crawdad.cs.dartmouth.edu

[2] We tried other values and observed no significant difference for ranges of 100–300 meters in our results.

to these areas as *Sunset* (2km×6km), *Airport* (0.7km×1km), and *Downtown* (2km×2km).

Our cab population is not exhaustive but represents all vehicles in a cab company comprising a large proportion of the San Francisco cabs, which number around 1500 [20]. The interest of this trace is that it represents the behavior of taxi drivers in different parts of a city where some of them live, park their cab, or simply decide to wait for customers because of their friends or social habits. Their social behavior is clearly impacted (and possibly dominated) by customer requests and the lack of information about customers is clearly a limitation of this trace. Nonetheless, the SF trace is very interesting as it is representative of a community behavior across the different areas we study, and it is the only environment where the ratio of Social and Vagabond varies significantly. Last, it is worth noting that mobility in this trace is mostly defined by traffic conditions and speed limits.

c) Second Life: The last data set captures avatar mobility in the Second Life (SL) virtual world [21], [22]. The data set consists of the virtual coordinates of all 3126 avatars that visit a virtual region during 10 days. We assume that two avatars are able to contact each other and exchange data when they are within a vicinity of 10 meters, a reasonable range for close-proximity communication such as Bluetooth [23]. The number of avatars which engage in contacts is 2713, and as with the Dartmouth trace we study only these avatars. It has recently been shown that the social network defined by such contacts between SL avatars resembles real-life social networks [24].

We do not define sub-areas in this data set as the virtual region is small (300m×300m). This limitation is balanced by the exhaustive user population captured, where Socials are people returning on regular basis and Vagabonds are occasional visitors that come only once in most cases.

IV. Socials and Vagabonds

We first classify users according to their social mobility behavior. To do so, we divide the user population in each trace into two distinct groups: *Socials* and *Vagabonds*. Intuitively, Socials are the devices that appear regularly—and, therefore, predictably—in a given area. In contrast, Vagabonds are devices that visit an area rarely and unpredictably.

Based on the above intuitive definition, we propose three different methods for classifying users into Vagabonds and Socials, and we apply these methods to the selected areas of the three data sets we presented in the previous section. By definition, the classification of a user as a Social or a Vagabond will depend on the area one considers. For example, it is possible that a user is a Social in the Engineering area of Dartmouth, and a Vagabond in the Medical area.

A. Identifying Vagabonds and Socials

The first method classifies users based on how long they stay in a given area. The other two methods classify users based on the regularity of their appearance in an area.

The results shown in this section focus on a five-day consecutive weekday period, as we expect Vagabonds and Socials to exhibit different behavior between weekdays and weekends. We have verified, however, that our definitions and behavioral properties hold on all other five-day weekday periods in all traces.

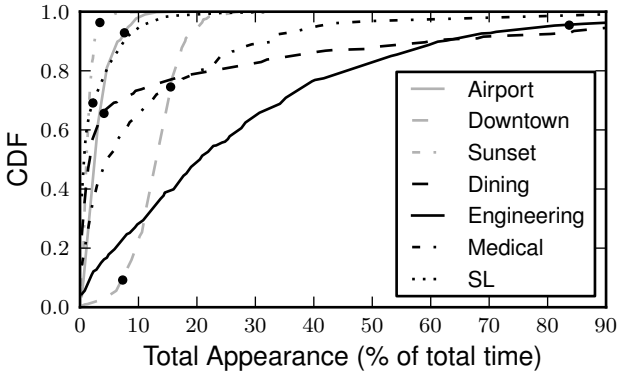


Fig. 1: Total time the population appears in each area. The black dots represent the knees of the CDF curves as found by the linear regression method.

1) Least Total Appearance: We define total appearance as the total time spent by each device within an area during the five-day period. Figure 1 shows the CDF of the total appearance time of the population for the first week of each area. In almost all areas (excluding Engineering) more than 75% of the population appears less than 20% of the time, with even lower appearance time being the common case. Thus, few devices stay within an area for longer periods and, intuitively, such devices would be the Socials of this particular area.

We define the least total appearance (LTA) threshold as the first inflection point ("knee") of the CDF of the total appearance for an area. This threshold separates Vagabonds from Socials, and is specific to each area.

To objectively identify such inflection points in the CDF curves of Figure 1, we employ a technique for detecting significant changes in curvature [25]. Each curve is iteratively approximated by a straight line using linear regression in the range $[0, t]$, where $t > 0$. The iteration stops when there is a significant error in the approximation. We assume that there is a significant error in approximation when the correlation coefficient r is such that $r^2 < 0.9$. This point identifies the knee in the CDF, and thus also signifies the threshold that we should use in the LTA method.

Figure 1 also shows the inflection points for the different sub-areas as dots on their respective CDF curves. Although Sunset has a single clear "knee", Downtown and Engineering do not. Downtown has two possible inflection points, and LTA selects the lowermost. Engineering has no distinctly apparent knee. Its curvature varies slowly across the full distribution, and LTA eventually selects a point as the CDF levels off.

2) Fourier: Our second classification method, *Fourier*, detects periodicity. It relies upon the Fourier transformation and the autocorrelation of the appearance of a user in an area, approaches used in signal processing to detect periodicity.

We employ a technique by Vlachos *et al.* [26], and Figure 2 shows an example of applying this technique to a device in

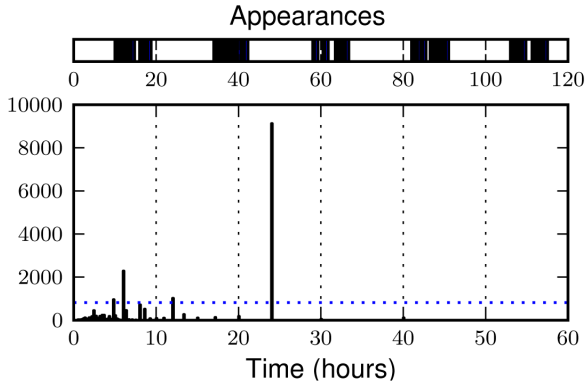


Fig. 2: An example of a social device detected using the Fourier method.

Dartmouth Medical. The top graph shows the appearance of the device throughout a five-day period. The next graph shows the Fourier transform of this signal into the frequency domain.

The Vlachos technique determines a threshold on the frequency coefficients in the Fourier transform. If the transform has coefficients above the threshold, the device appearance is periodic and corresponds to a social user. Otherwise, the device is a Vagabond. The bottom graph shows the threshold for the example device with a horizontal dashed line. Several Fourier coefficients exceed this threshold, and hence the device is Social.

For social devices, the technique identifies the inverse of the highest frequency coefficient as a potential period of the device appearance. The technique subsequently uses autocorrelation to improve the accuracy of the period estimate.

The Fourier method is problematic for nodes that appear very infrequently (*e.g.*, once or twice). The spectrum of such nodes would be roughly uniform (*e.g.*, white noise), making the selection of an appropriate threshold difficult. Consequently, almost half of devices that appear once or twice in certain areas were labeled as Socials by this method, which is clearly a mischaracterization. As a result, we investigate an additional method that focuses on periodicity.

3) Bin: Our third method, termed *Bin*, is motivated by the observation that people's mobility patterns exhibit a diurnal behavior [27]. Our traces also confirm this behavior, as the most frequent period detected by the Fourier method was 24h. Based on this observation, Bin detects if a user appears *every day* in an area, and *consistently* during the same time period.

For each trace we divide our measurement period into bins of equal size b, corresponding to the length of the "time period" during which a user frequents the area. We then represent the appearance of each device over time as a binary string, where each bit corresponds to a time bin. For each device, we flag a time bin with "1" if the device appears in the area during the period corresponding to this bin, and "0" otherwise.

We then consider a device to be periodic if it appears every day, at a specific period of the day. For a given bin size b, a device whose corresponding string has a "1" every $\frac{24}{b}$ bits is periodic. For flexibility, we identify a device as periodic even when an exact bin is not flagged but a neighboring (either previous or next) bin is. If a device is "periodic" by this definition we consider it Social, otherwise it is a Vagabond.

In experiments using the Bin method in this paper, we use bin sizes of 3 hours. We believe that this is representative of the time variance of the diurnal behavior of users from one day to the next. We obtained very similar results when repeating the experiments with a bin size of 4 hours, suggesting that around this time granularity the results are not very sensitive to the bin size.

B. Classifying Vagabonds and Socials

Table II shows the percentage of Vagabonds in each area according to each classification method. We observe that, under all methods, in most of the areas Vagabonds represent the majority of the population. The Downtown area in SF is an obvious exception: as expected, most cabs visit the downtown area frequently enough to be characterized as Socials by all three methods.

Area	Total	LTA	Bin	Fourier
Airport	451	92.7%	44.1%	70.3%
Downtown	455	7.3%	9.9%	39.3%
Sunset	436	96.1%	89.0%	81.7%
Second Life	1563	60.7%	96.7%	62.0%
Dining	404	61.6%	75.5%	58.4%
Engineering	940	95.3%	51.3%	27.4%
Medical	207	72.0%	79.2%	40.1%

TABLE II: Percentage of Vagabond devices in the areas.

We observe that LTA classifies a much higher number of Vagabonds than the other two methods in the Engineering area. Since the total appearance curve for this area is not amenable to partitioning the population into Vagabonds and Socials (Figure 1), the threshold selection method for LTA does not work well for this area.

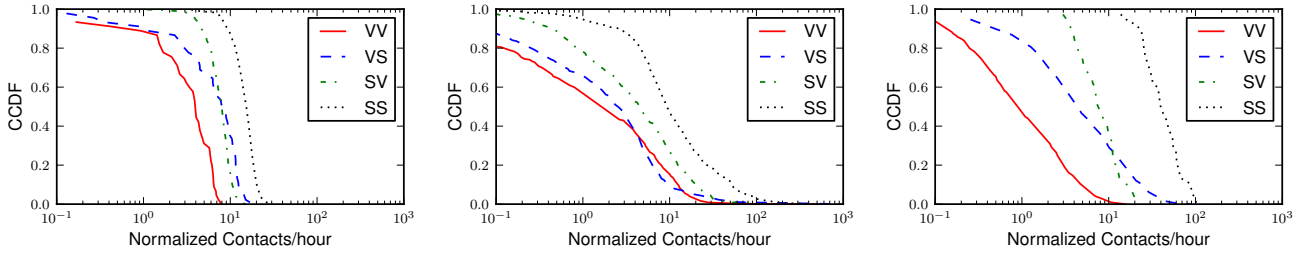
We also conduct a pairwise comparison of the results of the three methods to determine to what extent they agree on device classifications. We use the fraction of users for which the methods make the same decision as the metric of similarity.

Table III compares the three methods. We observe that the overlaps are similar for LTA and Bin yet surprisingly different for Fourier, even though Bin and Fourier are both based on periodicity detection.

Area	LTA & Fourier	LTA & Bin	Bin & Fourier
Airport	71.8%	51.4%	53.4%
Downtown	60.0%	90.9%	60.4%
Sunset	81.4%	91.5%	78.4%
Second Life	84.3%	63.0%	63.1%
Dining	64.6%	82.7%	59.7%
Engineering	23.2%	56.0%	55.5%
Medical	49.8%	85.0%	52.2%

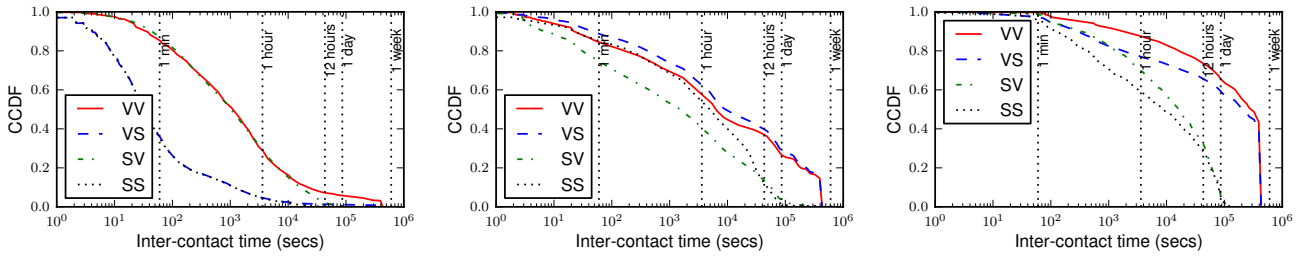
TABLE III: Percentage of devices for which the classification methods agree.

For the remainder of the paper, we use the Bin method to classify Socials and Vagabonds. Bin strikes a balance between the simplicity of LTA and the rigidity of Fourier. Although



(a) Downtown (b) Engineering (c) Second Life

Fig. 3: Contact rate distributions in three areas.



(a) Downtown (b) Engineering (c) Second Life

Fig. 4: Inter-contact time distributions.

LTA is simple, the single dimensionality of appearance time is not flexible enough to capture essential differences in Social behavior across the full range of areas. Fourier, however, requires Socials to appear according to a strict period and regimented schedule. Bin goes beyond LTA by incorporating appearance frequency and periodicity, but with a flexibility that better matches human behavior.

V. Contact properties

We know from previous work [3], [7] that contact characteristics are key in the effectiveness of opportunistic ad-hoc communication. We examine three different contact metrics: the *contact rate*, the *inter-contact time*, and the *contact duration*. We study these metrics for four different contact scenarios: *Social-meets-Socials* (SS), *Vagabond-meets-Socials* (VS), *Social-meets-Vagabonds* (SV) and *Vagabond-meets-Vagabonds* (VV). For example, the contact rate for VS is the rate at which a given vagabond device meets any social devices.

Our main observation is that Socials have significantly higher contact rates than Vagabonds, indicating that they have more opportunities for data dissemination, while inter-contact times are heavier tailed for Vagabonds. This observation is in accordance with our expectations based on our definition of Vagabonds and Socials and provides a validation point for the classification method that we chose. However, we have also seen in Section IV that Vagabonds considerably outnumber Socials in most regions. We later study how these two factors interact to affect data dissemination in Section VI.

A. Contact rate

For each device, we compute the number of contacts per hour with other devices in the social or vagabond group. We normalize this metric to remove the bias introduced by the size of the target population. Figure 3 shows the CCDF of the normalized contact rates for representative areas of the three traces. We also chose these areas because they span the spectrum of Social and Vagabond combinations: Socials dominate Downtown SF, Vagabonds dominate Second Life, and they are balanced in Dartmouth Engineering. The results for the other areas are similar to these, and we omit the corresponding graphs for space considerations.

We observe in *all* areas that the SS contact rate is an order of magnitude higher than the VV contact rate, with the VS and SV contact rates somewhere in between. The distribution shape appears to be driven by the region characteristics and by the nature of the source device. The tail of the distribution is longer when the source device is a Vagabond (VS and VV contact rates), while SS and SV contact rate distributions decay faster and have short tails. This indicates that there are few Vagabonds that have higher contact rates than the rest of the vagabond population. This is possibly due to our method for selecting Socials and Vagabonds. Social devices exhibit quite homogeneous contact rates on the other hand.

B. Inter-contact time

The inter-contact time of a device is the time interval that starts with the end of a contact and ends with the beginning of the next contact, whatever the device encountered is. This quantity is very interesting as it characterizes the periods during which a device cannot forward any content to other devices. The inter-contact distribution has been shown to be heavily tailed [3], which makes it impossible to estimate the delivery performance in such a network.

Figure 4 shows the CDF of the inter-contact time by social group of devices for the representative areas in each data

set. We observe two different parts in each curve: the main body (roughly below 12 hours) and the tail of the distribution (above 12 hours). In the main body of the distribution, inter-contact is similar for Socials (respectively Vagabonds), independently of what type of device they encounter. This part of the distribution characterizes the mobility patterns that are specific to each area. The tails of the distribution though are always much longer when the device met is a Vagabond, independently of the nature of the source, which characterizes the vagabond devices and not the mobility in the area. This heavy-tailed inter-contact with Vagabonds will help us explain later why Vagabonds are not individually as effective at content dissemination.

C. Contact duration

The amount of data that can be transmitted between two devices depends both on contact durations and on the communication technology (*e.g.*, WiFi or Bluetooth). Therefore, contact duration is difficult to interpret and does not characterize the performance of communication in opportunistic ad-hoc networks. Contact duration is mostly a characteristic of the mobility in the area. As a consequence, we find that Socials and Vagabonds experience comparable contact characteristics and their distributions are very similar; as a result, we do not plot their distributions. In the Dartmouth data set, contacts last longer due to the stationary nature of the devices. Contacts are uniformly distributed between a couple of minutes and 3 hours. In San Francisco, the contact duration is defined by the road traffic condition in each area (with most of the cabs experiencing contacts between one second and one minute). In Second Life, avatar mobility is defined by social events or points of interest, which leads to the majority of contacts lasting between one minute and one hour.

VI. Data Dissemination

We now analyze the impact of each social group of devices on data dissemination using trace driven simulations. We replay each trace multiple times using only Socials, only Vagabonds, or any device to propagate messages, while all devices can receive messages.

Our main observation is that, in areas in which Vagabonds outnumber Socials significantly, dissemination using Vagabonds outperforms dissemination using Socials, despite the lower contact rate experienced by Vagabonds. Further, we observe in most traces that there is a simple law by which we can predict which population is going to be more effective at propagating information.

A. Methodology

We simulate message dissemination using flooding. Since the outcome depends on the start time of the simulation, we repeat the simulation by uniformly sampling many start times between the beginning of the selected week (Sunday midnight) and the middle of that week (Wednesday noon). At the start of each simulation only one device carries the message, and for each randomly chosen start time we simulate dissemination starting from each of the devices in the trace. Simulations last 2.5 days to ensure they all complete within the week-long trace. The number of simulations is determined by the standard deviation of the results of the completed simulations. For each point in time we calculate the average value and standard deviation of the number of devices receiving the message for all the completed simulations. We perform as many simulation runs as necessary so that each sampled point is within a 95% confidence of its expected value.

We also assume that message transfers are instantaneous. This simplification overestimates transmission opportunities, but it does not introduce a bias between Socials and Vagabonds as they exhibit similar contact durations characteristics.

The metric characterizing message dissemination that we study is *contamination*. Contamination is the number of devices that receive a given message as a function of time. It reflects how effective a given population is at disseminating information in an area.

B. Evaluation

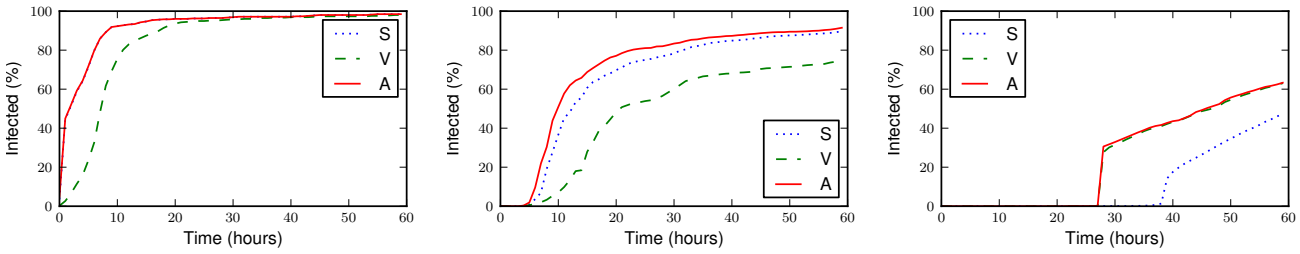
To understand the role that Socials and Vagabonds play in transmitting a message to the population of an area we first examine the number of devices that the message can reach relying only on Vagabonds or Socials. Note that we only account for message transmissions that take place through contacts that occur within the boundaries of the area. If devices make contact outside the area, we do not consider it to be a transmission opportunity since that situation does not reflect the contamination properties of a specific group of devices (the nature of a device being potentially different in each area).

Figure 5 shows the contamination result for the three different representative areas that we used previously. The curves represent the median across all simulations of the percentage of all devices reached.

The general observation is that Socials outperform Vagabonds in areas where they are the majority (SF Downtown) or of comparable population size (Dartmouth Engineering). However, in areas where Vagabonds largely dominate, they exhibit better contamination characteristics than Socials (Second Life). We also observed the same effect in all the other areas where Vagabonds form a clear majority (Dartmouth Dining and Medical, SF Sunset).

Individually, Socials contaminate more effectively than Vagabonds because they have a higher contact rate and more frequent contacts. In contrast, Vagabonds experience long periods of time without an opportunity to forward a message. However, we observe that large populations of Vagabonds can achieve the same contamination performance as Socials. Each Vagabond has a lower contact rate, but with many Vagabonds the total number of contacts is as high as what Socials would achieve with a smaller number of devices.

To explore the relationship between the number of devices and social behavior further, we simulate message dissemination while varying the population sizes of each group by taking random subsets. We decrease the number of Socials when the social group performs better in an area, or similarly decrease



(a) Downtown (b) Engineering (c) SL

Fig. 5: Contamination within an area when using Vagabonds (V), Socials (S), or any device (A) to propagate messages.

Area	Better	Socials	Vagabonds	V / S
Airport	S	99	199	2.01
Downtown	S	22	45	2.04
Sunset	V	48	220	4.58
Dining	V	99	205	2.07
Engineering	S	229	482	2.10
Medical	V	43	140	3.26
Second Life	V	37	215	5.81

TABLE IV: Vagabond and social population sizes when contamination is comparable using either of the two groups.

the number of Vagabonds when they perform better, until we observe a similar contamination ratio for dissemination using each group. Table IV reports these results. To have comparable contamination ratios, Vagabonds need to number two to six times more than Socials, depending on the area. Of course, these results are just one point in the parameter space balancing population sizes and social class—but they hint at the possibility of a deeper relationship. In the next Section, we formally present a model that develops a general "law" for this relationship.

We learn here two major properties of communication in opportunistic ad-hoc networks. First, *the effectiveness of contamination is more a matter of contact "density" in an area than an issue of social behavior*. Second, *Vagabonds have an important role in dissemination of information and should not be ignored or removed when studying propagation in opportunistic networks.*

VII. Analysis

Section VI indicates that the performance of data dissemination depends both on the density of devices as well as their contact rate. As a result, even though Vagabonds have on average an order of magnitude fewer contact opportunities than Socials, they can achieve similar dissemination performance in areas with 3–4 times more Vagabonds than Socials.

The goal of this section is to formally characterize the relationship between the population size and the social behavior of users under which such phenomena occur. Our approach relies on a so-called "mean field" limit applied to epidemic dissemination.

A. Model Description

1) Vagabonds and Socials: We consider N mobile users visiting an area A, partitioned into the two classes of Vagabonds and Socials. Let N_v and N_s be the number of Vagabonds and Socials, respectively. Users in each class enter and exit the area A as follows. Time is slotted, and at each timeslot a Vagabond enters A with probability ρ_v, independently of previous slots and of other users. Similarly, a Social enters A with a probability ρ_s. We call ρ_v and ρ_s the *occupancy rate* of Vagabonds and Socials, respectively, and we assume that $\rho_v \ll \rho_s$, *i.e.*, Vagabonds spend less time in the area than Socials.

Note that the occupancy rate of each class captures the "social" behavior of the class, as it indicates whether its users frequent this area or not. The expected number of Vagabonds and socials present in the area—*i.e.*, the density of each class—is given by $\rho_v N_v$ and $\rho_s N_s$, respectively.

2) Contacts between users and data dissemination: At each timeslot, we select two users uniformly at random among all (unordered) pairs of the N users in the system. If both of these users are within the area A then a contact takes place between them. If at least one of them is outside A, then no contact takes place within this timeslot. Note that, with $\rho_v \ll \rho_s$, the contact rate (average number of contacts per timeslot) of a Social is higher than the contact rate of a Vagabond, as the latter is far less likely to be inside A at a given timeslot. This is consistent with our empirical observations in Section VI.

Data dissemination starts with an initial number of users (Vagabonds or Socials) carrying a message. Each time a user carrying the message contacts a user that does not, a message transfer occurs with a probability that depends on whether the two users are Vagabonds or Socials. As with the simulations in Section VI, we focus on the two cases where either Vagabonds or Socials (but not both) are message forwarders, while all devices can receive a message. In particular, denote by λ_{vv}, λ_{vs}, λ_{sv}, and λ_{ss} the probabilities that transmissions succeed across and within classes; for example, λ_{sv} is the probability that the message transfer succeeds when a Social contacts a Vagabond. We focus on the following two cases: (a) only vagabond users forward the message, *i.e.*,

$$\lambda_{vv} = \lambda_{vs} = 1, \text{ and } \lambda_{sv} = \lambda_{ss} = 0, \quad (1)$$

and (b) only social users forward the message, *i.e.*,

$$\lambda_{vv} = \lambda_{vs} = 0, \text{ and } \lambda_{sv} = \lambda_{ss} = 1. \quad (2)$$

3) Main Result: Our analysis yields the following theorem, which quantifies when the "power of the crowd" dominates social behavior.

Theorem 1: For large enough N, the epidemic dissemination using Vagabonds eventually dominates dissemination using Socials if and only if $N_v\rho_v^2 > N_s\rho_s^2$.

Recall that Vagabonds occupy the area less frequently than Socials and are thus at a disadvantage w.r.t. epidemic dissemination. Thm. 1 implies that, when relative population sizes result in $N_v \gg N_s$, propagation using Vagabonds may outperform propagation using Socials. The necessary and sufficient condition is that the ratio of the two populations exceeds the square of the ratio of their occupancy rates. For instance, if Socials appear 10% of the time in the area, while Vagabonds appear only 5% of the time, Vagabonds will outperform Socials if their population is 4 times the population of Socials.

B. Proof of Theorem 1

1) A fluid limit: Let $r_v = N_v/N$, $r_s = N_s/N$, be the corresponding fractions of the total population belonging to each class. We refer to users that carry the message as *infected* and users that do not as *susceptible*. We denote by I_v, I_s the number of infected Vagabond and Socials, respectively, and by $i_v = I_v/N$, $i_s = I_s/N$ the corresponding fractions over all users. We also denote by S_v, S_s the number of susceptible Vagabond and Socials, respectively, and by $s_v = S_v/N$, $s_s = S_s/N$ the corresponding fractions.

Under the assumptions of Section VII-A, the evolution of the vector $\vec{i}(t)$, $t \in \mathbb{N}$, representing the number of infected users in each class, is a stochastic process. Nonetheless, as N tends to infinity, we can approximate the evolution of the system through a deterministic process, also known as a "fluid" or "mean field" limit. In particular, for large enough N, $\vec{i}(t)$ can be approximated with arbitrary accuracy through the solution of the following ordinary differential equation (ODE):

$$di_v/dt = \rho_v^2 i_v(r_v - i_v)\lambda_{vv} + \rho_v\rho_s i_s(r_v - i_v)\lambda_{sv} \tag{3a}$$

$$di_s/dt = \rho_s\rho_v i_v(r_s - i_s)\lambda_{vs} + \rho_s^2 i_s(r_s - i_s)\lambda_{ss} \tag{3b}$$

where the initial conditions $i_v(0)$ and $i_s(0)$ are set equal to the initial fractions of infected vagabonds and social users. Note that the above ODE is essentially the classical susceptible-infected model (see, *e.g.*, [28]) applied, in this case, to two infectious classes.

Formally, consider the following extension of the discrete time stochastic process $\vec{i} : \mathbb{N} \to [0,1]^2$ to a continuous time process $\underline{\vec{i}} : \mathbb{R}_+ \to [0,1]^2$. Define $\tau_k = \frac{k}{N}$, and, for all $k \in \mathbb{N}$,

$$\underline{\vec{i}}(\tau_k) = \vec{i}(k), \text{ and}$$

$$\underline{\vec{i}}(\tau_k + s) = \vec{i}(k) + s\frac{\vec{i}(k+1) - \vec{i}(k)}{\tau_{k+1} - \tau_k}, \text{ for } 0 < s < \frac{1}{N}.$$

Our main lemma states that the continuous version $\underline{\vec{i}}(\tau)$ of the fraction of infected users can be approximated with arbitrary accuracy through the solution of the ODE (3).

Lemma 1: Let $\vec{\xi}(\tau)$, $\tau \in [0,\infty)$, be the solution of the ODE (3) with initial condition $\vec{\xi}(0) = \vec{i}(0)$. Then, for every $T \geq 0$,

$$\lim_{N\to\infty} \sup_{0\leq\tau\leq T} \|\xi(\tau) - \underline{\vec{i}}(\tau)\| = 0, \text{ in probability.}$$

The proof can be found in [29]. Intuitively, the above lemma implies that the trajectory of $\vec{i}(t)$, for $0 \leq t \leq T \cdot N$ (*i.e.*, in an ever increasing interval), can be arbitrarily well approximated by the trajectory of the solution $\xi(\tau)$ of (3) in the interval $[0,T]$. For N large enough, the probability that the stochastic process $\vec{i}(t)$ strays too far from the deterministic trajectory $\vec{\xi}(\tau)$ is arbitrarily small.

2) Solution of the ODE (3)*:* The following lemma, whose proof can be found in [29], determines the evolution of $\vec{i}(t)$, as given by (3), under a single infectious class.

Lemma 2: The ODE

$$dx/dt = \alpha(A - x)x \tag{4a}$$

$$dy/dt = \beta(B - y)x \tag{4b}$$

with initial conditions x_0, y_0, has the solution

$$x(t) = A - (A - x_0)A/\left(x_0 e^{\alpha A t} + (A - x_0)\right) \tag{5a}$$

$$y(t) = B - (B - y_0)\left[A/\left(x_0 e^{\alpha A t} + (A - x_0)\right)\right]^{\beta} \tag{5b}$$

Using the above, we establish that the condition of Thm. 1 implies the domination of propagation through Vagabonds.

Lemma 3: Let i^{vo} and i^{so} be the fractions of infected users under ODE (3) when either (1) or (2) hold, respectively. If $\rho_v^2 r_v > \rho_s^2 r_s$, then $\lim_{t\to\infty}(1 - i^{vo}(t))/(1 - i^{so}(t)) = 0$.

The proof of this lemma can also be found in [29]. Theorem 1 therefore follows directly from Lemmas 3 and 1. To summarize, it implies that, if $\rho_v^2 r_v > \rho_s^2 r_s$, the propagation using vagabonds eventually dominates the propagation using social users, in spite of the fact that Vagabonds show up in the area much less frequently than Social users.

C. Numerical Validation

Figure 6(a) illustrates the performance of epidemic propagation under our model, evaluated through the ODE (3). We consider population ratios N_v/N_s ranging between 0.1–10 and occupancy rates ρ_s, ρ_v ranging between 1–10%. Circles correspond to cases for which propagation using Socials infects 97% of the population faster, and crosses are cases when propagation using Vagabonds is faster. The dashed line corresponds to a balance in propagation speeds between Vagabonds and Socials, as predicted by the inequality in Thm. 1.

Note that Thm. 1 is asymptotic: it states that when $N_v\rho_v^2 > N_s\rho_s^2$, Vagabonds will *eventually* dominate Socials. Figure 6(a) shows that the theorem correctly predicts which class reaches the 97% contamination threshold in most cases. The cases for which the theorem does not correctly predict the outcome are due to insufficient time for the asymptotic behavior to manifest; indeed, we repeated these evaluations with higher thresholds and observed a decrease in misclassified points.

Recalling the simulations in Section VI, none reached more than 95% of the total population, so it is difficult to compare the analytic results in Figure 6(a) with our simulation results. Instead, Figure 6(b) shows the relative propagation performance of Vagabonds and Socials after 60 hours of message propagation. Circles correspond to cases where, after 60 hours, the simulated propagation using Socials infected more

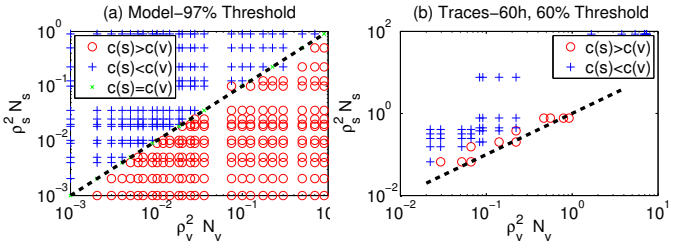


Fig. 6: Validation of Thm. 1: (a) relative performance of epidemic propagation using Vagabonds and Socials under our model; (b) relative propagation performance using Vagabonds and Socials from the Dartmouth, SF, and SL traces. The dashed lines indicate the threshold above which, according to Thm. 1, Vagabonds outperform Socials.

users than the propagation using Vagabonds, while crosses correspond to the converse. To exclude simulations not in the asymptotic regime, we show only the cases where either simulation reached more than 60% of the total population. Although many of these points are far from the asymptotic propagation behavior, Thm. 1 correctly predicts the outcome in most cases.

In summary, we proposed a model incorporating the population sizes of Vagabond and Social devices, as well as their social behavior. We have identified a law determined by these two parameters that governs the asymptotic efficiency of epidemic dissemination. Though our focus was on asymptotic efficiency, our ODE approach in general applies to more complicated interactions between users, including, *e.g.*, transmissions that fail with class-dependent probabilities or re-infections introduced after a received message expires.

VIII. Conclusions

In this paper we improve our understanding of data dissemination in opportunistic mobile ad-hoc networks. By separating users into two behavioral classes, we find that, although Socials form an active population subset, most areas are dominated by Vagabonds in terms of population size. Vagabonds, often excluded as unimportant, can often play a central role in opportunistic networks. As a result, tracing efforts should strive to capture the presence of Vagabonds, and analyses of protocols and applications should not discount them.

This work is just a first step in studying the impact of social behavior of users on information dissemination. A number of interesting directions naturally follow, including studying the characteristics of inter-area message propagation, the dynamics of user social behavior (*e.g.*, Vagabonds becoming Socials in other areas), and the interactions between Vagabonds and Socials in supporting information dissemination.

References

[1] A.-K. Pietiläinen, E. Oliver, J. LeBrun, G. Varghese, and C. Diot, "MobiClique: Middleware for Mobile Social Networking," in *WOSN*, 2009.

[2] I. Rhee, M. Shin, S. H. K. Lee, and S. Chong, "Human Mobility Patterns and Their Impact on Delay Tolerant Networks," in *HotNets VI*, 2007.

[3] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of Human Mobility on Opportunistic Forwarding Algorithms," in *INFOCOM*, 2006.

[4] N. Eagle and A. Pentland, "Reality Mining: Sensing Complex Social Systems," *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255–268, 2006.

[5] M. McNett and G. M. Voelker, "Access and Mobility of Wireless PDA Users," *Mobile Computing and Communications Review (MC2R)*, vol. 9, no. 2, pp. 40–55, April 2005.

[6] P. Hui, J. Crowcroft, and E. Yoneki, "BUBBLE Rap: Social-based Forwarding in Delay Tolerant Networks," in *MobiHoc*, 2008.

[7] A. Mtibaa, M. May, C. Diot, and M. Ammar, "PeopleRank: Social Opportunistic Forwarding," in *INFOCOM Mini Conference*, 2010.

[8] T. Hossmann, T. Spyropoulos, and F. Legendre, "Know Thy Neighbor: Towards Optimal Mapping of Contacts to Social Graphs for DTN Routing," in *INFOCOM*, 2010.

[9] E. M. Daly and M. Haahr, "Social Network Analysis for Routing in Disconnected Delay-Tolerant MANETs," in *MobiHoc*, 2007.

[10] G. Bigwood, D. Rehunathan, M. Bateman, T. Henderson, and S. Bhatti, "Exploiting Self-Reported Social Networks for Routing in Ubiquitus Computing Environments," in *Workshop on Social Aspects of Ubiquitous Computing Environments*, 2008.

[11] A. Mtibaa, A. Chaintreau, J. LeBrun, E. Oliver, A.-K. Pietilainen, and C. Diot, "Are You Moved by Your Social Network Application?" in *WOSN*, 2008.

[12] P. Costa, C. Mascolo, M. Musolesi, and G.-P. Picco, "Socially-aware Routing for Publish-Subscribe in Delay-tolerant Mobile Ad Hoc Networks," *IEEE JSAC Special issue on Delay-Tolerant Networks*, 2008.

[13] E. Yoneki, P. Hui, S. Chan, and J. Crowcroft, "A Socio-Aware Overlay for Publish/Subscribe Communication in Delay Tolerant Networks," in *ACM MSWiM*, 2007.

[14] S. Ioannidis, A. Chaintreau, and L. Massoulié, "Optimal and Scalable Distribution of Content Updates over a Mobile Social Network," in *INFOCOM*, 2009.

[15] A. G. Miklas, K. K. Gollu, K. K. W. Chan, S. Saroiu, P. K. Gummadi, and E. de Lara, "Exploiting Social Interactions in Mobile Systems," in *Ubicomp*, 2007.

[16] M. Motani, V. Srinivasan, and P. S. Nuggehalli, "PeopleNet: Engineering a Wireless Virtual Social Network," in *MobiCom*, 2005.

[17] W. Gao, Q. Li, B. Zhao, and G. Cao, "Multicasting in Delay Tolerant Networks: A Social Network Perspective," in *MobiHoc*, 2009.

[18] T. Henderson, D. Kotz, and I. Abyzov, "The changing usage of a mature campus-wide wireless network," *Computer Networks*, vol. 52, no. 14, pp. 2690–2712, October 2008.

[19] M. Piorkowski, N. Sarafijanovoc-Djukic, and M. Grossglauser, "A Parsimonious Model of Mobile Partitioned Networks with Clustering," in *COMSNETS*, January 2009.

[20] SFMTA, "San francisco transportation fact sheet," November 2009, http://www.sfmta.com/cms/rfact/documents/SFFactSheet2009_November2009_FINAL.pdf.

[21] Second Life, http://www.secondlife.com.

[22] M. Varvello, F. Picconi, C. Diot, and E. Biersack, "Is There Life in Second Life?" in *ACM CoNEXT*, October 2008.

[23] A.-K. Pietiläinen and C. Diot, "Experimenting with Opportunistic Networking," in *MobiArch*, 2009.

[24] M. Varvello and G. M. Voelker, "Second Life: a Social Network of Humans and Bots," in *NOSSDAV*, 2010.

[25] M. Ley, "Does the Knee in a Queuing Curve Exist or is it just a Myth?" July 2009, http://www.cmg.org/measureit/issues/mit61/m_61_16.html.

[26] M. Vlachos, P. S. Yu, and V. Castelli, "On Periodicity Detection and Structural Periodic Similarity," in *SDM*, 2005.

[27] M. González, C. Hidalgo, and A.-L. Barabasi, "Understanding Individual Human Mobility Patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[28] K. Nelson and K. M. Williams, *Infectious Disease Epidemiology: Theory and Practice*. Jones and Bartlett Publishers, 2007.

[29] G. Zyba, G. Voelker, S. Ioannidis, and C. Diot, "Dissemination in opportunistic mobile ad-hoc networks: The power of the crowd," Technicolor, Tech. Rep. CR-PRL-2010-07-0001.