Measuring Serendipity: Connecting People, Locations and Interests in a Mobile 3G Network

Ionut Trestian
Northwestern University
Evanston, IL, USA
ionut@northwestern.edu

Aleksandar Kuzmanovic
Northwestern University
Evanston, IL, USA
akuzma@northwestern.edu

Supranamaya Ranjan
Narus Inc.
Mountain View, CA, USA
soups@narus.com

Antonio Nucci
Narus Inc.
Mountain View, CA, USA
anucci@narus.com

ABSTRACT

Characterizing the relationship that exists between people's application interests and mobility properties is the core question relevant for location-based services, in particular those that facilitate serendipitous discovery of people, businesses and objects. In this paper, we apply rule mining and spectral clustering to study this relationship for a population of over 280,000 users of a 3G mobile network in a large metropolitan area. Our analysis reveals that (*i*) People's movement patterns are correlated with the applications they access, *e.g.*, stationary users and those who move more often and visit more locations tend to access different applications. (*ii*) Location affects the applications accessed by users, *i.e.*, at certain locations, users are more likely to evince interest in a particular class of applications than others irrespective of the time of day. (*iii*) Finally, the number of serendipitous meetings between users of similar cyber interest is larger in regions with higher density of hotspots. Our analysis demonstrates how cellular network providers and location-based services can benefit from knowledge of the inter-play between users and their locations and interests.

Categories and Subject Descriptors

C.2.3 [**Computer-Communication Networks**]: Network Operations
C.4 [**Performance of Systems]:** Measurement techniques

General Terms

Measurement, Human Factors, Experimentation

IMC'09, November 4–6, 2009, Chicago, Illinois, USA.

Keywords

Location based services, Human mobility, Cellular network, Hotspot, Application interest, Mobile Network, Serendipity

1. INTRODUCTION

Recent advances in technology allow mobile devices to find their physical location via a multitude of methods: cell tower look up, cell tower triangulation, cell tower and wi-fi access point triangulation [6] and Global Positioning System (GPS), with varying accuracy levels. Besides the use of location estimation for navigation services, a new class of applications, 'serendipitous' location-based services, have also gained tremendous popularity. These services and applications allow users to serendipitously, *i.e.*, accidentally discover people, businesses and other locations around them that match their interests.

For instance, location-aware social networking applications such as Loopt [3] or Pelago [4] enable 'serendipitous meetings' between friends who discover that they are in the same neighborhood or city and hence may decide to meet. Some applications [5] even facilitate formation of new relationships by allowing users to share their location and profile information with the entire user base. A user can look up who else is in proximity and depending on common interests, can decide to communicate. Furthermore, location tagging services allow a user to leave interesting tags about a particular geographic location, *e.g.*, photos [1] or snippets about current events, *etc.*, and then other users who are in vicinity of that location could be automatically prompted with those geo-tags. Finally, location-based advertising [2] allows a retailer to send advertisements to users on detecting that a user, who previously opted-in to the service, has entered the 'geo-fence' area around the retailer.

These serendipitous location-based services are typically designed to work even when users provide only coarse-grained location information. This is due to several reasons. First and perhaps the most important reason is privacy related. Users may be more comfortable sharing their location with friends and businesses at coarse-levels such as neighborhood or city as opposed to finer-levels such as street address or GPS coordinates. Second, a vast majority of mobile phones in a 3G network are still not GPS enabled and hence their location needs to be obtained via other techniques such as cell tower look up.

Regardless, these applications still provide a meaningful experience in the face of coarse location information. In social networking, it is sufficient for two friends to decide to meet up if they know that they are in the same neighborhood - they do not necessarily have to know each other's exact latitude and longitude to decide on meeting up. Similarly, for location-based tagging as well as advertising, if someone is within a radius of a few hundred meters to a few miles, that can be sufficient for prompting the user with geo-tags left by others or advertisements from businesses in the neighborhood.

The fundamental questions relevant to these serendipitous location-based services remain yet unanswered. For instance, how likely is it to meet in our daily lives, and where, with people who share similar interests in cyber domain? What role does our physical location play in terms of what we access online from there?

We answer these questions by considering an underlying, yet even broader question: *what is the relationship between one's mobility properties and affiliations towards given applications in cyber domain?* We answer this question by systematically and methodically studying the user mobility and Web access patterns for over 280,000 clients of a 3G mobile network in a large metropolitan area. Using a one-week network trace, we obtain the application interest expressed by a user by classifying URLs accessed in to broad categories such as social networking, dating, music, gaming, trading, *etc.* We obtain mobility patterns by extracting the time-sequence of base-stations accessed by users.

To the best of our knowledge, we are the first to systematically study the relationship between mobility patterns and application affiliations at such a large scale. This is our main contribution. Among a number of insights that we provide, the key one is that we present a first-of-its-kind evidence suggesting a strong application affinity at certain locations irrespective of time of day, i.e., certain locations inspire people to access a specific application type.

To understand the relationship between mobility and the corresponding application usage, we apply an association rule-mining approach [25] to extract the most prominent behavior in mobility and applications. Our analysis confirms previously reported results [20] on the high predictability of human movement. For example, we find that 70% of users return back to at least one common location every day over a one week period. In addition, we find strong correlation and anti-correlation between some applications and mobility. For example, we find that listening and downloading music prevails for stationary users. For mobile users, bandwidth- and battery-intensive applications (such as music) fade away, while e-mail prevails.

We find that users spend most of their time within their 'comfort zone' consisting of three top-most locations, including home and work. The access behavior inside and outside the 'comfort zone' differs. For example, dating applications are mostly accessed from within the 'comfort zone', but neither from home nor from work. On the other hand, users who leave the 'comfort zone' exhibit the tendency of staying connected by accessing social networking sites, reading e-mail and news.

Next, we explore the relationship among locations and applications accessed at them. To achieve this, we extend the rule-mining approach to identify location hotspots. We define four types of hotspots based on the time of day when they are active, *i.e.*, day, noon, evening, and night. We find that there is a strong bias towards applications accessed by people at the locations at which hotspots occur. Because the hotspots we define are time dependent, we explore whether the root cause of the observed application skew is the time of day or the location itself. We find that in majority of scenarios, it is *location* that dominantly impacts which applications are accessed.

Finally, to explore the probability that people with similar cyber affinities meet each other in the real world, we proceed as follows. Using a spectral clustering approach [12], we split the metropolitan area into smaller regions. We find that the frequency with which one meets others who share the same cyber interests is determined by the density of hotspots in a given region, *i.e.*, fraction of locations that are hotspots in the region.

This paper is structured as follows. In Section 2, we provide details about the trace, and explain how we mine the desired mobility and application usage information. In Section 3, we define our rule-mining based approach and provide insights about the relationship between mobility, applications, and locations. In Section 4, we explore hotspots, perform the regional analysis in Section 5 and in Section 6, we present related work. We summarize our findings in Section 7 and discuss potential benefits to cellular networks and location-based services.

2. PRELIMINARIES

Here, we provide details about the dataset. Then, we explain how we extract users' mobility properties and interests in the cyber domain, *i.e.*, affiliation towards given Internet applications and services.

2.1 Trace Description

We use an anonymized trace collected from the content billing system for the data network of a large 3G mobile service provider. The trace contains information about 281,394 clients in a large metropolitan area of 1,900 square miles (approx. 5,000 square kilometers) during a seven day period. It preserves user privacy as all identifiers such as users' phone numbers, email addresses and ip-addresses were anonymized.

The trace provides details of a *packet data session* defined as beginning from the time the user is authenticated by the authentication, authorization and accounting (AAA) protocol by the Remote Authentication Dial in User Service (RADIUS) [24] server to the time the user logs off. In between, a user's packet data session consists of HTTP and Multimedia Messaging Service (MMS) sessions[1] initiated by the user.

When a customer logs on, the serving Packet Data Serving Node sends a RADIUS Access-Request to the RADIUS server. If the user is successfully authenticated, the RADIUS server returns an Access-Accept message which contains a 'correlation identifier' which will be used to uniquely identify the user through the entire packet data session. Next, the Packet Data Serving Node uses the RADIUS accounting protocol (RADA) [23] for communicating events that involve data usage to the RADIUS server [23].

These accounting messages contain the following relevant information: *local timestamp, anonymized user identifier (phone number or email address), anonymized ip-address* assigned to the user, *correlation identifier*, and the *base-station* that was currently serving the user besides other in-

[1]Whenever we use the term session later in the paper, we refer to the packet data session, unless stated otherwise.

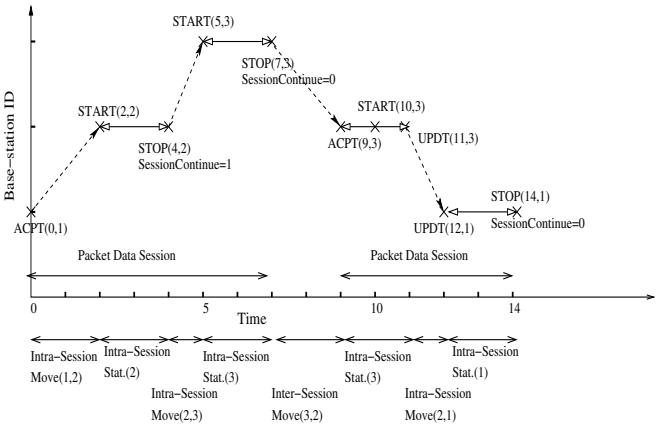


Figure 1: Sequence of locations for a user.

formation. These accounting messages can be of type Start, Update and Stop and there can be any number of these messages within a packet data session. Start messages are used to indicate the beginning of a new accounting activity, *e.g.*, when the user starts a new application in the current data session. Update messages are generated periodically to indicate the current accounting status of the data session [7]. The Stop message contains an attribute, 'Session Continue' which when set to 'false' is indicative of the end of the session. Changes in user's currently associated base-station are reported either in an Update message or via a Stop message immediately followed by a Start containing the new base-station.

Each HTTP session contains the following details: *user's anonymized ip-address, the URL accessed and the local timestamp.* Because the Multimedia Messaging Service (MMS) is run over HTTP in this network, the trace provides the same records for MMS as well.

We reconstruct a user's entire packet data session as follows. Using RADIUS and RADA messages, we build an association between a user identifier (phone number of email address) and his currently assigned ip-address. Then, we identify the applications accessed by a user by grouping the HTTP and MMS sessions that occur after a RADIUS session and have the same ip-address as was assigned to the user.

The trace provides the location of a user in terms of the base-station. In the trace, we have a total of 1,196 base stations for the large metropolitan area. The area serviced by a base-station in this network varies from hundreds of square meters (in densely populated areas) to several square miles (in sparsely populated areas). On average a base station services 4 square kilometers. In the remainder of the paper, we use the term location to refer to the area serviced by a specific base-station. Thus, while our trace does not provide GPS-level fine-grained location information, we will show later that location information at the level of base-stations is still invaluable from the perspective of the serendipitous location-based services. In particular, we will show how we can infer generic user mobility properties (Sections 2.2,3.1), as well as correlate locations with application usage (Section 3.2).

2.2 Extracting Mobility Properties

Here, we explain how we extract mobility patterns from the trace and present preliminary results about human mobility.

Table 1: Trace statistics

	Mean	**90%ile**	**Max.**
Session duration	40 min	60 min	3 days 20 hrs
Number of sessions per user	11.2	24	4,442
Number of unique base stations per user	4.2	8	128

Figure 1 shows an example for the location sequence of a user across two different packet data sessions. We use the RADIUS accounting packets of type Start, Update, and Stop to extract the sequence of locations or base-stations accessed by a user along with the timestamps at which the user was present at those locations.

There can be two kinds of movements for a user. (*i*) *(i) Intra-session* movement happens when the user's location changes within a packet data session due to hand-offs, *e.g.*, between (Accept, 0, 1) and (Start, 2, 2). (*ii*) *Inter-session* movement happens when the location changes during the inactive time, *i.e.*, when the user is not active in the mobile network, *e.g.*, between (Stop, 7, 3) and (Accept, 9,3).

We consider a user to be stationary if the base-station he is associated with does not change. That is even if the user did physically move within the base-station, for our purposes, we consider him as stationary. Similarly, there can be two kinds of stationary events for a user: (*i*) *intra-session* when a user's location stays the same within a packet data session and; (*ii*) *inter-session* when a user's location remains the same between two consecutive sessions.

In our seven-day long trace, we record 3,162,818 packet data sessions, generated by 281,394 users. Table 1 provides a few representative statistics for the trace.

2.2.1 Basic Mobility Observations

While we are able to detect a user's access to the mobile data network and accurately estimate a given location (above) and characterize accessed applications (below), an important underlying question is if we are able to accurately estimate user mobility patterns. In particular, there is on average a gap of 6 hours and 11 minutes between two consecutive sessions from the same user. On separating users as those who move and those who stay stationary between two consecutive packet data sessions, we obtain the average inter-session move and stationary times as 8 hours and 23 minutes, and 4 hours and 25 minutes respectively. In comparison, the average time spent by a user session moving is 9.3 minutes (intra-session movement) and stationary is 31 minutes (average intra-session stationary).

Necessarily, intra-session movements provide more information about user movement than inter-session movement. This is because in intra-session movements, we are capable of tracing all locations visited by a user while he was online. In this regards, we look in to whether inter-session movements still capture adequate information when compared to intra-session movements. We compare the two movements from the perspective of displacement probability.

Given a time difference ΔT, we identify all inter-session and intra-session movements where the user has changed his location within the time gap: $\Delta T \pm 0.05\Delta T$. Figure 2 shows cumulative distribution function (CDF) of user displacement, *i.e.*, how far a user moves in the given time interval, with intervals ΔT ranging from 20 minutes to one day. We compute the distance between two locations as the geodesic or great-circle distance between them, which takes

Table 2: Classifying URLs in to Interests

Interest	Keywords	Interest	Keywords	Interest	Keywords
Dating	dating, harmony, personals, single, match	**Gaming**	poker, blackjack, game, casino	**Mail**	mail
Music[3]	song, mp3, audio, music, track, pandora	**Maps**	virtualearth, maps	**MMS**	mms
News	magazine, tribune, news, journal, times	**Photo**	gallery, picture, photo, flickr	**Ringtones**	tones
Trading	amazon, ebay, buy, market, craigslist	**Search**	google, yahoo, msn	**Weather**	weather
Social netw.	facebook, myspace, blog	**Travel**	vacation, hotel, expedia, travel	**Video**	video

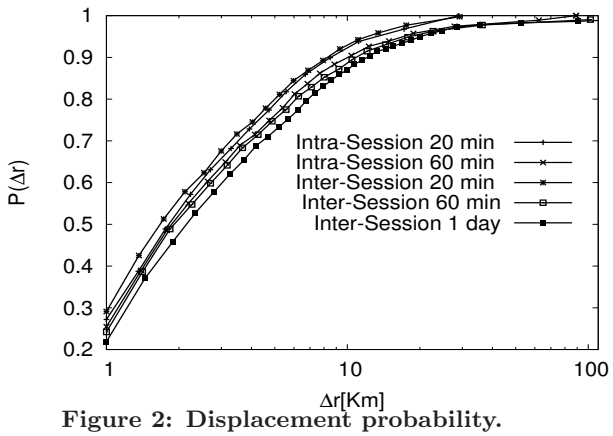


Figure 2: Displacement probability.

the earth's sphericity in to account[2]. We have two points to make. First, most human movements occur over short distances, however, humans are also likely to travel large distances, albeit with smaller probabilities. This result is in line with previous findings [9, 20]. Second, inter-session movements, despite containing less information about users' locations than intra-session, still exhibit similar displacement probabilities. Hence, lack of information about locations visited by the user when he was offline, does not undermine our ability to estimate mobility patterns.

2.3 Extracting Application Interests

Next, we show how we extract a user's application interest by classifying the URL visited. We only have access to the first part of the URL and not the actual parameters that are being exchanged with the website. Hence, user privacy is preserved in this respect as well.

Consequently, we classify the URLs accessed by users into application interests via keyword mining over the URL. We distinguish the following categories: *dating, gaming, mail, maps, MMS, music, news, photo, ringtones, search, social networking, trading, travel, video and weather*. A comprehensive list of the classification rules we employ is provided in Table 2. Some keywords, *e.g.*, google, yahoo and msn represent portals from where users can access different services (e-mail or search). Hence, in order to distinguish between e-mail (keyword: mail) and search (keywords: google, yahoo, msn) we apply the mail rule first.

Furthermore, every time we see a URL accessed by the user, we extract the last location that the user was seen at. Each of these application accesses are also encapsulated in the corresponding packet data session by considering the times at which the user logged on and off from the network.

In the rest of the paper, we present results for only the following interests: dating, social networking, music, e-mail, trading, and news for the following reasons: (*i*) categories social networking, dating and music represent interests and goals which can serve as triggers for users to arrange for a serendipitous meeting; (*ii*) categories e-mail and news represent the urge to stay connected to friends and world events and; (*iii*) category trading represents a potential location-based market place where people interested in buying and selling goods in same geographic area could be matched up. We opportunistically emphasize other applications as and when necessary.

One limitation of our study lies in the fact that our trace does not contain device type information. Indeed, certain devices have characteristics which make them attractive for a specific purpose, for example they can be easily used as navigation tools or for sending e-mails. Such extra features that are device-dependent can have a bias on our analysis. Also, recently, mobile service providers have started commercializing modems that can be used with personal computers such as laptops. Users can use these devices to connect to the Internet from anywhere within the cellular network. One concern is that these devices do not constrain the user (in terms of application accesses) in the way a limited resource platform such as a mobile phone might. Such modem devices exist in the network that we have analyzed, although in a relatively small proportion compared to the total devices; they number a few hundred as informed by the provider.

3. FROM HUMAN MOVEMENT TO APPLICATION USAGE

Here, we explore basic human mobility patterns, and then study the relationship between movement and locations on one hand and application usage on the other.

3.1 Basic mobility patterns

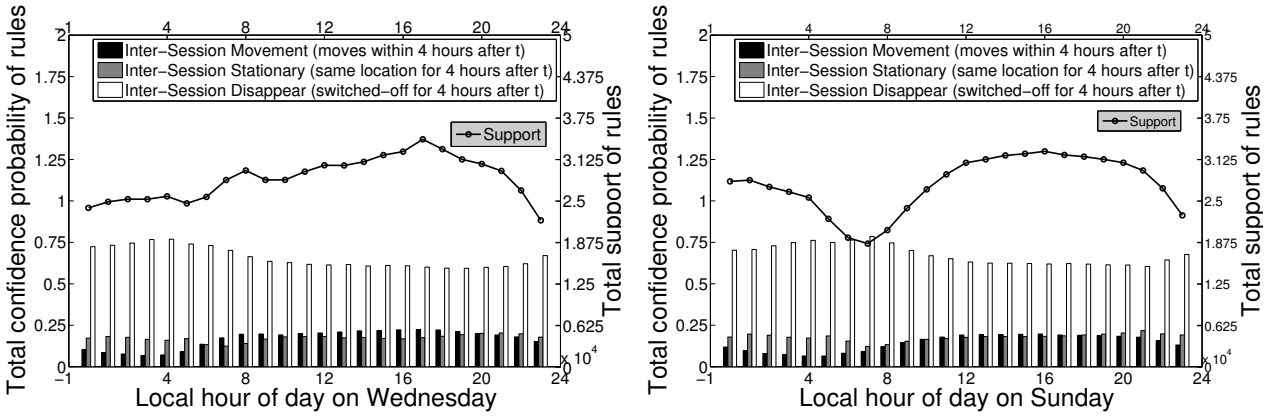
We first develop a methodology to extract basic mobility patterns exhibited by users. Then we provide initial insights about inter- and intra-session movement properties. Finally, we study movement predictability.

3.1.1 Binary trajectory rules

We develop a methodology based on association rule mining [25] to extract *binary rules* which group movement and stationarity events with one antecedent and one consequent, *e.g.*, users who are present at one location who then move to another location. This methodology also allows us to identify location boundaries (source to target location) that become popular at certain times of day, *e.g.* due to work commutes as well as locations which are popular with stationary users, *e.g.* residential and work areas. We will later use this methodology for identifying hotspots in Section 4.

First, note that a user who is accessing the mobile data network from a certain location (base-station) has the fol-

[2]As earth is not a perfect sphere, our calculations are an approximation, which however is sufficient for our purposes.
[3]Note, that the music interest category comprises of both downloads as well as streaming music.



(a) Weekday (Wednesday) (b) Weekend (Sunday)

Figure 3: Hours of day when inter-session rules are active

lowing three possible *exit states*: (1) user moves to a new location either while staying connected via the same session (intra-session movement) or logs-off and logs back again via a new session (inter-session movement); (2) user stays at the same location (intra- or inter-session stationary) and; (3) user switches-off his mobile device (disappear) and does not re-appear for some time. Equivalently, a user can be defined to have the following three *entry states* with respect to a location: (1) user appears in the location for the first time via a new session; (2) user stayed in the location from the past and; (3) user entered the location from a different one. We will explore the entry states of a user later in Section 5. Here, we define the following three rules to group users with respect to their exit states:

Definition 1. **Movement Rule:**
Group all users u_i from the user-set U who are present at location x_i, from where the user accessed the network via a session at time $t_i \in$ time window w and whose next location is $x_j \neq \emptyset$ either within the same session or via a new session before time $t_i + \delta$.

$$\text{Given } \delta,\ \forall u_i \in U,\ \text{obtain groups } (x_i, x_j \neq \emptyset, w, \delta) \text{ s.t.}$$
$$\exists t_j \in (t_i, t_i + \delta] : x_i, t_i \in w \implies x_j, t_j.$$

Definition 2. **Stationary Rule:**
Group all users u_i from the user-set U who are present at location x_i, from where the user accessed the network via a session at time $t_i \in$ window w and has since been present at the same location x_i either via the same session or a new session, since time t_i up until the time $t_i + \delta$.

$$\text{Given } \delta,\ \forall u_i \in U,\ \text{obtain groups } (x_i, x_i, w, \delta) \text{ s.t.}$$
$$\forall t_j \in (t_i, t_i + \delta] : x_i, t_i \in w \implies x_i, t_j.$$

Definition 3. **Disappear Rule:**
Group users who were present at location x_i at time $t_i \in$ window w and who since then have switched-off their device up until $t_i + \delta$ seconds.

$$\text{Given } \delta,\ \forall u_i \in U,\ \text{obtain groups } (x_i, \emptyset, w, \delta) \text{ s.t.}$$
$$\forall t_j \in (t_i, t_i + \delta] : x_i, t_i \in w \implies \emptyset, t_j.$$

Define **support** for a rule as the number of users that follow the antecedent, *i.e.*, that were present at location x_i within the time window w. Define **confidence** for the rule as the number of users who follow the rule, *e.g.*, for the inter-session movement rules, those who move from location x_i to location x_j. Thus, **confidence probability** for a rule is defined as the probability that users who have followed the rule antecedent so far will follow the consequent as well and is given by: $confidence/support$. We consider time windows of length one hour in this paper and hence the time window variable w takes values at the hour boundaries.

3.1.2 Inter-session Movement

Next, we consider the movement and stationary events which occur across sessions. In order to group users by their inter-session movements and stationarity *alone*, we only consider the movements and stationarity shown by a user across two consecutive sessions in Definitions 1-3.

We next identify the times of day when rules of a particular type are active. To achieve this, we cluster all inter-session movement (stationary, disappear) rules that occur in the same time window w irrespective of locations x_i or x_j associated with them. In this case, we choose δ as four hours, which is close to the average inter-session stationary time (4 hours 25 minutes). Figure 3 shows the averaged confidence probability of each rule type over an hour window for two days, one during the week and another over the weekend. The total support in terms of number of users present across all locations at a given hour is also plotted (see y2 axis). First, the total confidence probabilities of all the rules at a given hour add to 1.0. Second, confidence for disappear rule dominates.

We derive the following insights: (i) Stationary rules have a larger average confidence probability than movement rules during the hours of the night, 10 pm-5 am for Wednesday and 8 pm-7 am for Sunday. This is indicative of users being more stationary during the night than day and during the weekend than during the week; (ii) Finally, two local peaks at 8 am and 5 pm in Figure 3(a) shows increased group behavior on a workday, *i.e.*, moving towards work and back. No such local peaks occur on the weekend. Note that the confidence probability is by definition normalized by the values of the support at the considered time interval. Therefore the ratio between the values of confidence for the hours of the day and the hours of night is larger than the ratio between the values in the confidence probability for the hours of the day and the hours of the night.

3.1.3 Intra-session Movement

Considering all the sessions generated over seven days, we quantify how much mobility is captured within a user's packet data session. While a majority 84% of sessions stay stationary (that is stay within one base-station), and 6%

move for 15 minutes or less, the 99%-ile is 3.5 hours and the maximum time that a session spends in motion is one day. A user could divide his time within a session as both moving and being stationary. Once again, about 84% of sessions are completely or 100% sedentary while 6% of sessions are completely mobile. Finally, about 10.2% of sessions spend more time moving than staying within a session, *i.e.*, percent time spent within a session moving is larger than staying.

3.1.4 Daily Trajectories

Next, we investigate the predictability in users' behavior. In this regards, we identify the number of locations that a user visits every day across all the days that he is seen in the trace. First, for each user we build a *daily trajectory* by combining the sequence of locations that are visited by the user starting from the midnight of a day up until midnight of the next day. We use the locations corresponding to both intra- and inter-session movements for a user while building the daily trajectory. Next, given a user, we use his daily trajectory to extract the set of unique locations that he accessed that day. Say, a user was seen over 3 days such that he accessed the following sets of locations: $\{A, B, C\}$, $\{A, B\}$ and $\{A, C\}$ on those 3 days respectively. Then we compute the overlapping set of locations for this user as: $\{A, B, C\} \cap \{A, B\} \cap \{A, C\} = \{A\}$. Interestingly more than 70% of the mobile users visit at least one common location on every single day that they access the network, suggesting that users regularly revisit their usual locations. We will explore this affinity of users to certain locations in the rest of this section.

3.2 Application Usage

Here, we explore the following two questions: (*i*) What is the relationship between mobility and application usage? (*ii*) What is the relationship between users' location and the applications they access from there?

3.2.1 Mobility and Applications

To answer the first question, we first correlate movement and stationarity exhibited within a session (intra-session) with application accesses. We obtain the following groups of packet data sessions: completely stationary, *i.e.*, those which spend 0 minutes in movement and sessions with varying levels of mobility. For the latter, we consider Definition 1 while restricting to only intra-session movement events, and obtain groups: (x_i, x_j, w, δ) for δ ranging from [10-60] minutes. Recall from definition that a group (x_i, x_j, w, δ) contains those sessions which were present at location x_i at some time in the hour window w and then move to x_j *within* δ time. Hence, by definition, for the same location pair x_i, x_j and hour window w, for $\delta_{lo} < \delta_{hi}$, the following relationship holds: $(x_i, x_j, w, \delta_{lo}) \subset (x_i, x_j, w, \delta_{hi})$. Thus, a group with a higher value of δ contains the sessions from lower δ values as well as additional sessions which spent larger amount of time in movement than those considered previously.

We find that the top applications accessed by stationary sessions is social networking, music and e-mail which comprise 29%, 21% and 21% of all application accesses respectively. Interestingly, as sessions become more mobile, we observed that users access less music, *e.g.*, in the highly mobile sessions with δ of 60 minutes, percentage accesses for music reduces to 9%. On the other hand, as sessions become more mobile, they comprise of more e-mail accesses.

Mobility span. To explore this issue in more depth, we explore how does mobility span, *i.e.*, the number of locations that a user visits, impact the applications he accesses? Figure 4 shows the application access probability as a function of discrete mobility spans. In particular, we explore groups of users that have been seen at that particular number of locations during the seven day period. For each of the points on x-axis, the sum of normalized access probabilities on y-axis equals to one.

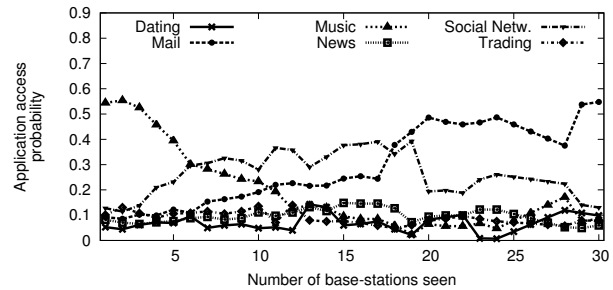


Figure 4: Application usage breakdown for people seen in the given number of locations

Figure 4 shows high correlation (and anti-correlation) between the mobility span and applications that people access. In particular, for the 'stationary' users (number of base-stations seen equals 1), music dominates. We explain this phenomenon later in the text, in the context of the 'comfort zone' that we introduce later.

In contrast, e-mail shows completely opposite trend. Indeed, the more stationary users are, the less they access e-mail on their mobile devices. This is most likely because they use other devices (*e.g.*, a home computer) to access e-mail. However, the more people move, the more e-mail starts dominating the applications. Indeed, for those who have a large mobility span, e-mail is by far the most accessed application, more than 50% of time. Indeed, those who move a lot have their mobile phones as their primary communication devices.

Finally, social networking shows highly intriguing behavior. It lags far behind the leading applications both within highly stationary group (lags behind music) and within highly mobile group (lags behind e-mail). Yet, for the medium mobility span group, for which music starts to fade due to mobility, and e-mail still does not start to dominate fully, social networking is the leading application.

Weather and maps patterns. Here, we correlate the inter-session rules with applications as follows. We use Definition 3 to obtain the users who disappear, *i.e.* switch-off their devices for δ time. Next, in Definitions 1 and 2, we consider only inter-session (and no intra-session) changes and also restrict the set U to only those users who are accessing the application being correlated. Thus, given an application say, maps, we only consider those sessions where the user accessed an online maps website and then compute the various inter-session movement and stationary rules with δ values varying as [1-8] hours to capture varying proportions of those users. Since, we are only interested in aggregate statistics, we cluster all the movement and stationarity events that occur within the same δ value irrespective of the locations and hours of day involved.

Next, we identify the rule type which captures the maximum percentage of accesses to an application type. We identified two applications, weather and maps as highly correlated with inter-session stationary and movement respectively. First, amongst all accesses to weather applications,

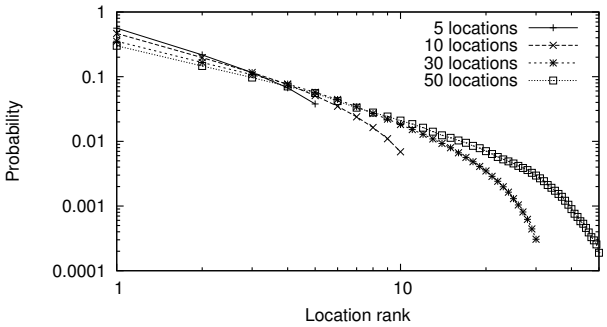


Figure 5: Probability to 'see' users at different location ranks (for users who span 5, 10, 30, and 50 locations during one week)

a majority (69%) of them are grouped as inter-session stationary with a δ of 6 hours. In other words, after a weather query, users are *not* inclined to movement. This directly implies that users are more interested in weather when they are sedentary for a long time. Second, amongst all accesses to a maps application, a majority (67%) of them are grouped by inter-session movement rules with a δ of 3 hours. This suggests that after accessing a maps website, users typically switch-off and move (possibly to the location they looked-up) and log-on to the network again after 3 hours.

3.2.2 The Role of Locations

Here, our goal is to understand the relationship between a user's physical location and the applications he accesses from there. To answer this question comprehensively, we answer the following two related questions: (*i*) What is the distribution of locations that users visit in terms of time spent in these locations? (*ii*) What applications do users access at these different locations?

To answer the first question, we proceed as follows. We first find the groups of users that have been detected in 5, 10, 30, and 50 locations during one week period. Then, we rank the locations at which the users reside based on the time spent in each of the locations.

Figure 5 plots the probability to 'see' given users as a function of the location rank. The key insight from the figure is that users spend the vast majority of time in the top three locations. For example, users who span five locations (base-stations) during one week, spend 89.5% of their time in the top three locations. Likewise, users who visit as many as 50 different base-stations during a week, spend more than 55% of their time in the top three locations. Hence, we call the top three location ranks as the user 'comfort zone' — the area where they spent the most of their time.

The second question we want to address is what applications do users access at differently ranked locations, in particular with respect to the 'comfort zone.' Figure 6 plots the cumulative distribution functions of the probabilities to access the six applications at the given ranked locations. The vertical line at location rank 3 marks the 'comfort zone' border. Here, we show the statistics for all users, not only the subgroups we discussed above.

In Figure 6, the higher the curve is, the more the given application is accessed within the 'comfort zone'. For example, more than 85% of music accesses happen within the 'comfort zone', while less than 15% outside it. We hypothesize that because music (and video likewise) is bandwidth and battery consuming, it is less likely to be accessed outside the

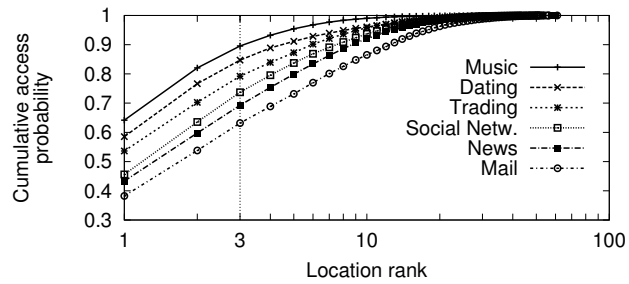


Figure 6: Cumulative access probability as a function of location rank

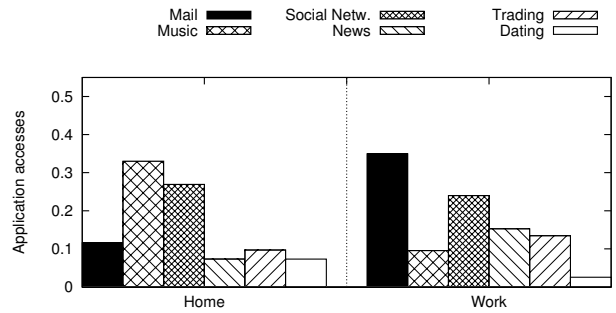


Figure 7: Home vs. Work

'comfort zone'. The second interesting result is that dating applications are mostly accessed from within the 'comfort zone', only 18% outside it. We explain this result later in the text.

While email, news, and social networking are again more likely to be accessed within the 'comfort zone', they are accessed much more frequently outside the zone relative to the other applications. For example, e-mail is accessed 40% of time outside the 'comfort zone', while social networking is accessed 30% of time outside the 'comfort zone'. Indeed, it appears that users have the tendency to 'stay socialized and connected' while outside the 'comfort zone': they access social networking sites, they read e-mail, and access news.

Home vs. work. Here, we want to understand which applications do users access at different locations within their 'comfort zone'. Because both home and work locations are amongst two of the three top-most locations in the vast majority of scenarios, our first goal is to identify locations corresponding to a user's home and work locations. Then, we provide insights about the applications accessed at the two locations.

To identify home and work locations, we proceed as follows. We consider a location as a user's home if we observe the user spending most of his time between 10 PM and 6 AM at this location. Further, in order to identify the most likely work place of a user, we consider the time interval between 10 AM in the morning and 12 PM (noon) and the time interval between 2 PM and 5 PM (during weekdays) and determine the top location that the user has been present at. If the two locations are the same, we are unable to accurately distinguish between home *vs.* workplace. This can happen either because the user does not work or the user lives and works within the area covered by the same base-station, *etc.* If the two locations are different, we assume the one accessed during night is home and the other is work.

Figure 7 shows the application access statistics. Our find-

ings are the following. First, at home, users are most likely to access music. This is in-line with our findings above, as music dominates among applications accessed in the 'comfort zone'. Moreover, for the group of users for which we were unable to distinguish home *vs.* work location as it overlaps, music is again the leading application (result not shown in the figure). Second, at work, users are most likely to access e-mail. Third, at both home and work locations, social networking is the second most popular application. Most interestingly, dating is the least accessed application at both home and work. However, given that dating is frequently accessed in the 'comfort zone' (see Figure 6), it follows that it is mostly accessed outside home and work, yet within the 'comfort zone'. We shed more light on this phenomenon in the next section.

4. HOTSPOTS

In this section, we study the effects of user movements on locations and how locations evolve as a result. First, by following an approach outlined in [26], we identify which locations have a high net change in userbase at a certain time and hence become *hotspots*. After identifying and classifying hotspots, we then study the interests of the userbase that is attracted to the hotspot as well as the actual applications that users access while they are present there.

We use the binary rule Definitions 1-3 for detection of hotspots as follows. We consider time windows of one hour and the change time δ as one hour as well. Let x and h represent a location and an hour window respectively. Let the number of users who first switched-on their mobile devices at location x in the hour window h be denoted as: $n_a(x,h)$. Next, let $n_d(x,h)$ denote the users who log-off from the network at location x in the hour window h. Finally, let those who entered the location x within the hour window h from some other location be denoted as $n_e(x,h)$, those that left it in that hour as $n_l(x,h)$ and those that continued staying there for that hour as $n_s(x,h)$.

The number of users entering a location x at a given hour h can be computed as the total confidence of all the movement rules which have this location as a target as follows: $n_e(x,h) = \sum_{y \neq x} conf(y,x,h,1)$. Next, the number of users leaving a location at a given hour can be computed as the total confidence of all the movement rules which have this location as the source as follows: $n_l(x,h) = \sum_{y \neq x} conf(x,y,h,1)$. Next, the number of users that stay stationary at a location at a given hour is given directly by the confidence of the stationary rule involving this location as follows: $n_s(x,h) = conf(x,x,h,1)$. Similarly, the users who first appear or finally disappear at a location at a given hour are given directly by the confidence of the appear and disappear rules respectively as: $n_a(x,h) = conf(\emptyset,x,h,1)$ and $n_d(x,h) = conf(x,\emptyset,h,1)$.

For each location, the total number of users who were present in an hour window, $N(x,h)$ can be described by considering all the exit states of those users, *i.e.*, by counting all the users who disappeared, those who left for some other location and those who stayed stationary: $N(x,h) = n_s(x,h) + n_l(x,h) + n_d(x,h)$. Now, for the same location at the next hour window $h+1$, the total number of users is given by those who stayed back from the past window as well as those who first switched-on their devices at this location and those who moved from some place else: $N(x,h+1) = n_s(x,h) + n_a(x,h+1) + n_e(x,h+1)$. Thus, the net change in users at a location across two consecutive hour windows h and $h+1$ is composed of two components, a *net inflow* and *net outflow* and is obtained as *inflow - outflow* or: $N(x,h+1) - N(x,h) = \{n_a(x,h+1) + n_e(x,h+1)\} - \{n_l(x,h) + n_d(x,h)\}$.

Hence, we determine if a location becomes a hotspot at a certain hour as follows. When the *net inflow* at a location during a certain hour contributes to the total number of users at the location at that hour by more than a fraction γ_{in}, then we tag it as a *sink*, *i.e.* $\frac{n_e(x,h)+n_a(x,h)}{N(x,h)} \geq \gamma_{in}$. Similarly, when the *net outflow* at a location during a certain hour contributes to the total number of users at the location at that hour by larger than a fraction γ_{out}, then we tag the location as a *source*, *i.e.*, $\frac{n_l(x,h)+n_d(x,h)}{N(x,h)} \geq \gamma_{out}$. Finally, when the number of users who stayed at a location within an hour, $n_s(x,h)$ contributes to the total number of users at the location at that hour by more than a fraction, γ_s, we tag it as a *stationary location*. Note that a location could be both a source and a sink at the same hour in some cases *e.g.*, base-stations located next to freeways.

We select the values of thresholds as the 90%-ile for each of the fractions across the entire trace duration. Thus, the threshold γ_{in} is chosen as the 90%-ile of the fractional contribution of net inflow across all the base-stations over the entire trace and similarly, for the other two thresholds. This yields values of $\gamma_{in} = 0.7$, $\gamma_{out} = 0.7$ and $\gamma_s = 0.3$.

Next, we use sinks, sources, and stationary locations to detect hotspots. In addition, we characterize the hotspots by looking for most likely causes for their creation. In particular, we use the available geographic (*e.g.*, downtown vs. suburb) and other properties (*e.g.*, residential *vs.* business area) of given areas that we obtain from publicly available sources. Hence, we characterize the hotspots as follows.

Day hotspots. These locations are sinks during early morning (8 AM-10 AM), stationary locations during the day and become sources in early evening(6 PM-7 PM). They are dominated by people at work, who reside at their offices during business hours.

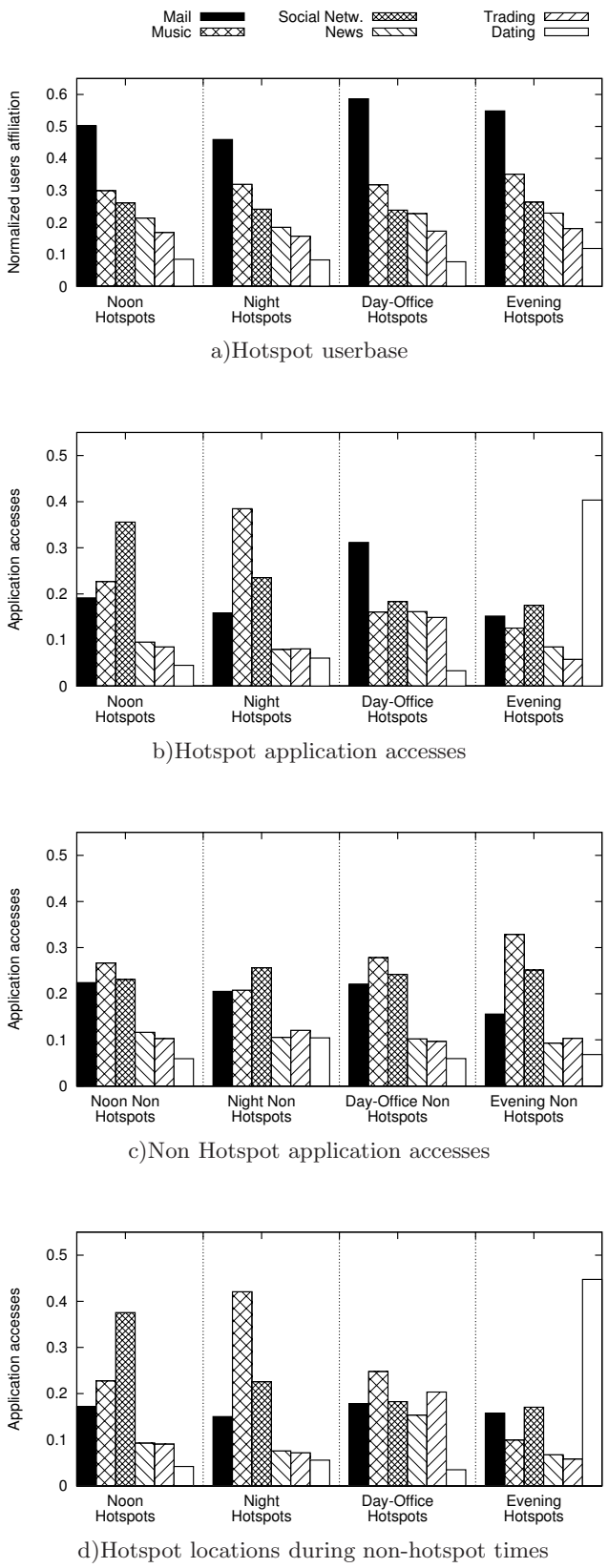
Noon hotspots. These locations become sinks during the afternoon (12 AM-1 PM) and sources shortly after (2 PM-3 PM). They are dominated by people taking a noon (lunch) break.

Evening hotspots. These locations are sinks during the evening (7 PM-8 PM) and sources shortly after(10 PM-11 PM). They are dominated by people going out in the evening.

Night hotspots. These locations are sinks in early evening (6 PM-8 PM), stationary locations during the night and become sources in the early morning (7 AM-9 AM). They are dominated by the people at their homes during night.

By applying the above analysis we identify 23 day hotspots, 28 noon hotspots, 8 evening hotspots, and 62 night hotspots. A majority (95%) of hotspots get classified by one label only. Our next goal is to understand what online applications do people access at these locations. More precisely, we want to answer the following related questions: (*i*) What *general* application affiliations do people who gather in these hotspots have, *i.e.*, what is the hotspots' *userbase*; and (*ii*) what applications do users access when they are present at hotspots?

First, we define the userbase of a hotspot (or, any location) as the breakdown in applications accessed by the users who were present at the hotspot, while considering all the applications that they have accessed during the seven day trace period, *i.e.*, not necessarily just the applications they accessed while they were present at the hotspot. Figure



a)Hotspot userbase

b)Hotspot application accesses

c)Non Hotspot application accesses

d)Hotspot locations during non-hotspot times

Figure 8: Hotspots application statistics

8(a) shows the results regarding hotspots' userbase. It plots the normalized user affiliation for given applications at different hotspots. For example, the figure for noon hotspots shows that 50% of users present at noon hotspots use e-mail (during the seven day interval, not necessarily in the given hotspot), 30% access music, *etc.* Because users can have more than one application affiliation, the sum of normalized affiliations does not equal to one. The key insight from the figure is that *all* hotspots have exactly the same userbase. The majority of people at these locations access e-mail on their mobile phones, and the least number engages in dating. Indeed, the trend over all locations (hotspots and non-hotspots) is exactly the same (the result not shown in the figure). Hence, it appears that there is nothing specific about users who enter these hotspots relative to those who do not access them.

However, on considering the actual applications accessed at the hotspots, things are quite different (see Figure 8(b)). The figure clearly shows that there is a strong correlation between the hotspot type and the primary application that people access at these hotspots. In other words, people with the same general application affiliations show highly skewed and biased group behavior towards a *single application* at the considered hotspots. In particular, social networking is the dominant application among those at noon hotspots; music among those at night hotspots; e-mail among those at office hotspots; and dating among those at evening hotspots. Thus, the given locations are not hotspots only in terms of a significant number of users present at them at a specific time. Stunningly, these locations are application hotspots as well — large groups of people show common 'cyber' behavior at them.

4.1 Time-of-day or Location?

Given that hotspots happen at detected locations, yet hotspots are time-of-day dependent, the next question is: what determines the bias shown by hotspots' users: time-of-day effects or the locations themselves?

Figure 8(c) shows the time of-day effect; we plot the application accesses by clients *outside* the hotspots, yet at the same time periods when a hotspot happens. For example, consider the locations which are not a hotspot at noon, thereafter referred to as noon non-hotspots. For these noon non-hotspots, we collect statistics about user accesses in the same time period, 12-2 PM. We can see a different trend than the one shown in Figure 8(b). As an example, at noon, social networking is not accessed as frequently at the noon non-hotspots. In the evening, dating is the least accessed application outside evening hot spots, *etc.* Hence, we conclude that time-of-day does not dominantly affect the accesses at hotspots.

Figure 8(d) shows that *location* itself dominantly determines the bias in application accesses observed at hotspots. In particular, we plot the number of accesses at the hotspot locations *outside* the time period that characterizes the given hotspot. The results clearly show the same trends as observed at hotspots in Figure 8(b). In particular, music is the leading application at locations corresponding to night hotspots even during the daytime as well; social networking is the leading application at locations corresponding at noon hotspots even outside noon intervals; dating is the leading application at locations corresponding to evening hotspots even during non-evening periods. Only in the case of day (office) hotspots, the leading application is no longer e-mail,

but music. Music prevails in these areas as they are dominated by residential customers during nights.

5. REGIONAL ANALYSIS

Individual users do not span the entire metropolitan area. Hence, the probability for one user to meet another user from a different part of the area might be small. Thus, to fully understand the potential for serendipitous location-based services, we first attempt to split the metropolitan area into smaller regions, *i.e.*, by clustering groups of people who access the network from similar locations (base-stations). Then, we explore the interactions within these regions. In this context we study three user interests which are representative of serendipitous location-based services: dating, social networking, and music. We currently only use broad interest identifiers, assuming this can be a sufficient trigger for users to meet up, *e.g.*, users interested in music may be prompted when they are in proximity and may decide to meet up. We leave an exploration of fine-grained interests, *e.g.*, users interested in classic rock likely to meet up, to future work. Regardless, the broad interest categories allow us to provide an important upper-bound on likelihood for users with similar interests to meet up.

Our first task is to identify regions (composed of locations) and also to determine people belonging to a certain region based on the time they spend in it. We model this problem as a bipartite graph between users and locations, and then perform a co-clustering across users and locations such that there is a one-to-one correspondence between a cluster of users and that of locations. In this regards, co-clustering can be thought of as a graph partitioning problem. To solve this NP-hard graph partitioning problem, several heuristics such as Kernighan-Lin [18] have been proposed, which, however, only consider the local minima while partitioning. In contrast, spectral clustering has been shown to be global and can obtain a semi-optimal cut [14]. Authors in [11, 14] show that the second eigenvector of a graph's Laplacian matrix gives a guaranteed approximation to the optimal cut. Other approaches [21, 12] use multiple eigenvectors to obtain a k-way partitioning of a graph. We adopt one such multi-way partitioning approach that was proposed to obtain a co-clustering of words and documents [12].

We begin by defining a bipartite graph G between users and locations. Let $u = 281,394$ be the total number of users, and let $l = 1,196$ be total number of locations. The vertices of graph G comprise of all users and locations, for a total of $(u+l)$ vertices. In the graph G, an edge connects a user i to a location j if the user has spent time in that location (936,280 edges). Each edge is given a weight $W(i,j)$ as the amount of time (seconds) spent by a user in that location and a weight 0 if a user has never visited a location. By definition of the bipartite graph, there are no edges between vertices of the same type *i.e.*, between users or between locations. Denote A as the user-by-location matrix of dimension $u \times l$ with values $A(i,j) = W(i,j)$. The multi-partitioning algorithm for co-clustering users and locations is as described in Algorithm 1.

The number of connected components in the graph G is given by the number of trivial singular vectors of the graph Laplacian [12]. We obtain only one trivial singular vector, implying that the dynamics of human movement connects the entire metropolitan area in to one giant connected component. Still within this giant connected component, location clusters exist on account of the fact that a corresponding cluster of users spends majority of its time within a location cluster. We run the multi-partitioning algorithm 1 with different values for the number of desired clusters k and across multiple runs of the algorithm, we always identified the same five significant regions.

Algorithm 1 Multi-partitioning users and locations in to k clusters each.

- Define the Laplacian of the Graph G as: $L = \begin{pmatrix} D_1 & -A \\ -A^T & D_2 \end{pmatrix}$ where, the squared diagonal matrix D_1 of size u^2 and D_2 of size l^2 are the following: $D_1(i,i) = \sum_{j=0}^{l} A(i,j)$ and $D_2(i,i) = \sum_{j=0}^{u} A(j,i)$.
- Construct matrix: $A_n = D_1^{-1/2} A D_2^{-1/2}$.
- Perform singular value decomposition on the A_n matrix and starting from the second largest singular vectors (since the first one solves the decomposition trivially) obtain $\lceil log_2 k \rceil$ singular left and right vectors each, and form matrices U and V respectively.
- Construct the following matrix, on which we run K-means to obtain k clusters each for users and locations: $\begin{pmatrix} D_1^{-1/2} U \\ D_2^{-1/2} V \end{pmatrix}$.

Table 3: User and location clusters.

Cluster	1	2	3	4	5
Nr. users	54,589	41,845	40,569	82,389	17,148
Nr. locations	162	216	194	257	118
Day Hot.	0	3	4	15	1
Noon Hot.	9	2	5	10	2
Evening Hot.	3	0	0	4	1
Night Hot.	27	3	4	26	2

Table 3 presents the five regions (clusters) and the corresponding statistics. Cluster 4 is the largest. It covers the downtown area, and it clusters together around 82k users. Cluster 1 and Cluster 3 are suburbs that immediately border the downtown, with cluster 1 being more urban. Clusters 2 and 4 are suburbs located farther away from the downtown. Indeed, the average number of users per base-station clearly reveals the more urban nature of clusters 4 and 1 relative to other clusters. The average number of users per base-station in these two clusters is above 320, while for the other three clusters it is below 200 on average. Indeed, the density of users is higher in urban areas.

The urban nature of clusters 4 and 1 is further revealed via the number of hotspots that occur in these regions. For example, as many as 55 and 39 hotspots reside in clusters 4 and 1, respectively. To the contrary, less than 10 hotspots on average reside in the other three regions. Interestingly enough, the number of day hotspots is as high as 15 in region 4, while it is *zero* in region 1. As we mentioned above, cluster 4 covers the business part of the downtown area, and all day hotspots reside there. Although cluster 1 is urban, it is more residential; hence, no day hotspots occur.

Table 4 presents the statistics for inner- and outer-cluster user movement, as given by the binary rule Definition 1 in Section 3. Inner- and outer-cluster movement are defined by whether the two endpoints involved in a movement rule belong to the same cluster or not, respectively. Our observations are the following. First, as expected, the majority of users move within their clusters, as the percents on the

Table 4: Breakdown of movement by users in a cluster and across clusters.

Movement [%]	Src. 1	Src. 2	Src. 3	Src. 4	Src. 5
Dest 1	70.4	1.7	2.3	13.1	0.2
Dest 2	1.5	63.5	18.9	1.4	22.5
Dest 3	2.1	21.7	60.6	7.3	5.1
Dest 4	25.9	2.9	16.3	77.7	2.3
Dest 5	0.1	10.2	1.9	0.5	69.9

diagonal positions in the table are the largest for each of the columns. Second, even if people move outside their cluster, they are most likely to visit the neighboring urban area (25% from 1 to 4, and 13% from 4 to 1). Third, people from suburbs rarely visit downtown; they are more likely to visit neighboring suburbs. Indeed, most of the movement shown in the table reflect geographic relationships *i.e.*, (1 and 4 are neighbors; so are 2 and 3, 2 and 5, as well as 3 and 4).

5.1 The Potential for Location-Based Services

Here, we explore how probable is it, and what determines the probability, for people who share the same interests in the cyber domain to meet as part of their daily lives? To answer these questions, we focus on the following interest categories: social networking, dating, and music, for their potential to trigger serendipitous interactions. Given a set of users with the same interest, *i.e.*, those who have accessed websites relevant to the interest type, either at current time at current location or at some time before reaching the current location, we compute the following two interaction metrics.

Time-independent interactions. We consider the overlap in trajectories between users of the same interest, irrespective of the actual time of overlap. This is relevant for location-based tagging services [1] where users leave geo-tags for a location which can be picked by other users who are in its vicinity.

Time-dependent interactions. In this more restrictive type of interaction, we consider that users with same interest are present in the same location at the same time instance. This type of interaction is the basis for location-aware mobile social networking, and other location-based services.

We consider two versions of the above questions: (*i*) How many *unique* people sharing the same cyber interests are likely to meet each other (in both time-dependent or independent manner)? (*ii*) How many *interactions* are people who share the same interest likely to have?

The first insight (not shown in a figure due to space constraints) is that the *number of unique people* sharing the same interests that meet each other is larger in region 4 (downtown) than in region 1 (neighboring urban area). This holds true both for time-independent and time-dependent interactions, and for all applications of interest. This is because the number of users is much larger in region 4 (82k) than in region 1 (54k). Hence, even if the user mobility patterns are similar in both regions, the probability of meeting *different* people is larger in a more populated region.

On the other hand, the results are reversed when considering the *number of interactions* with people who share similar interests, as we will show below. In particular, for time-independent interactions, we not only detect that two users met, but also count the number of places they met at. For time-dependent interactions, we count not only that two users met each other once at a location, but count all such contacts.

Figure 9 shows the number of time-independent and -dependent interactions as a function of different locations. We plot the curves in Figure 9 based on the decreasing number of meeting events. Hence, the order of locations (base-stations) on the *x-axis*, while similar, is not identical. Our insights are the following.

First, the number of interactions for time-independent interactions is necessarily larger than for time-dependent interactions since the probability to meet a person at a given location and at the same time is smaller than the probability that the two trajectories overlap. As a result, the scale on the y-axis in Figure 9(b) is an order of magnitude larger than that of Figure 9(c). Also, social networking and music curves are above dating, because these applications are more popular.

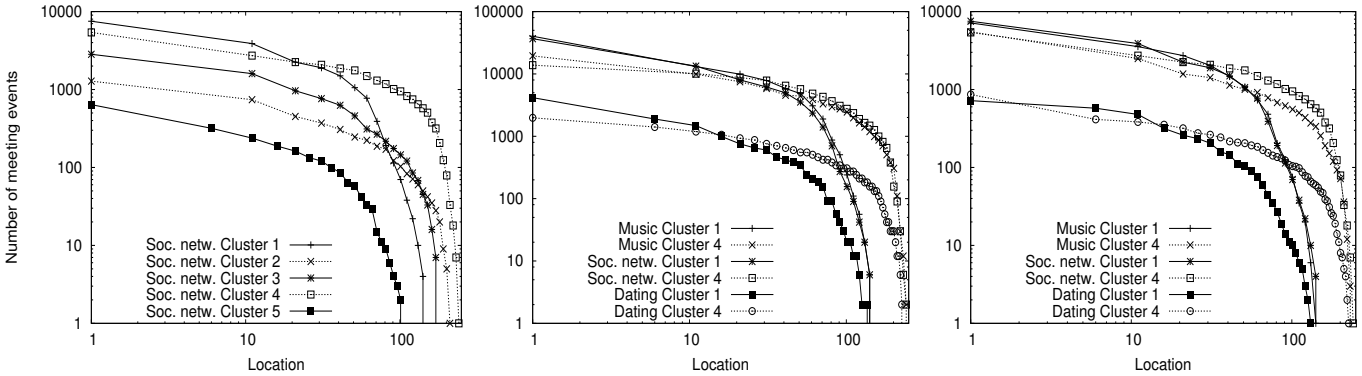
Second, for both time-independent and -dependent interactions, the regions that provide most interactions for either of the interests, social networking, dating and music, are ordered in descending order as: 1, 4, 3, 2 and 5. To see why this happens, first note that the top-most locations in Figure 9(a) for regions 1 and 4 are hotspots. The urban regions 1 and 4 have higher number of interactions than other regions mainly because of the large number of hotspots in these regions as shown in Table 3. Most interestingly, even though region 4 (downtown) contains larger user population than region 1, the order between them is reversed in terms of interactions. We explore the reasons for this further.

The key reason is the density of hotspots in a region, defined as a fraction of locations in a region that are hotspots. A region with higher hotspot density provides more chances for interactions. For instance, consider music interest. Based on Table 3, region 1 has 27 night hotspots from 162 base-stations, and hence a night hotspot density of 17%. The regions in descending order in terms of night hotspot density are: 1, 4, 3, 5 and 2. The same order amongst regions is found for noon (social networking) and evening (dating) hotspots as well. Hotspot density is able to explain interactions as the top three regions in terms of interactions are also 1, 4 and 3. For regions 2 and 5, the number of hotspots is small, hence, non-hotspots influence interactions as well. Hence, for the same mobility properties, the probability of accessing a hotspot is larger in region 1, and thus the number of interactions increases.

Table 5: Interactions per user class.

Event type	Mobile users	Static Users (Hotsp.)	Static Users (Non-Hotsp.)
Social netw.	704	604	424
Music	828	565	319
Dating	253	188	96

The final question that we explore is the following: given the impact that hotspots have on interactions, who will 'experience' a larger number of social interactions: a mobile user or a stationary user present at a hotspot? For this experiment, we cluster out 3 categories of users: (*i*) mobile users that have been seen in at least 20 locations, (*ii*) static hotspot users that have spent at least 6 hours in a hotspot, and (*iii*) static non-hotspot users who have spent at least 6 hours in a non-hotspot. Table 5 shows the results. Most interactions are observed by mobile users, since they meet more users than others. Still, the results show that it pays off to spend a considerable amount of time at a popular location. Indeed, the result shows that static hotspot users are



a) Social networking across clusters b) Time independent interactions c) Time dependent interactions

Figure 9: The number of meeting interactions as a function of locations

close behind highly mobile users. As expected, static non-hotspot users experience smallest number of interactions.

6. RELATED WORK

The increasing ease in availability of *digital footprints* of humans via the mobile devices they carry, has led to a plethora of studies on human movement. One such group of work [10, 13, 17, 22] explores the possibility for delay-tolerant networking, *i.e.*, opportunistic peer-to-peer delivery of messages between mobile devices coming within Bluetooth or WiFi radio range of each other. Such studies are based on data sets with fine-grained meeting information, *e.g.*, [10, 17] use contact information polled from a set of various datasets which use different technologies (*e.g.* bluetooth, wi-fi, etc) to study the inter-contact times between devices while [13, 22] use bluetooth on a sample set of 100 subjects to study the chance of meeting someone. In contrast, the motivating application for our study is serendipitous location-based service, for which coarse-grained location information at the level of neighborhood or city can be a sufficient trigger for two users to decide to meet. While our study here does not directly address peer-to-peer message delivery, our findings do have indirect implications to such applications as well. In particular, our observation on the affinity between a location hotspot and an application type suggests that is at these places that such services will be most likely used.

Another body of work is focused on modeling human trajectories [19]. Recently, authors in[20] studied the trajectory of 100k anonymized mobile phone users and determined that the trajectory of humans is not as random as predicted by the earlier models (Lèvy flight and random walk models) and in fact humans exhibit a high degree of temporal and spatial regularity. Our findings regarding this are similar, in that humans are very likely to spend most of their time in their three most preferred locations. While our data set is also from a mobile carrier, our trace is primarily from the data network (HTTP and MMS) while [20] uses phone calls and SMS logs. Regardless, this points to an important evidence that human movement studies such as ours and [20] are not biased by the data source they are based on. In contrast, our goals in this paper are very different, to study human movement patterns when correlated with application interests.

A third body of work [26] has studied mobility in the context of sequential rule mining, where the goal is to extract the most frequent trajectory sequences. We adapt [25, 26] to develop our binary rule framework to identify the basic mobility patterns and then extend the same to perform a novel joint study of application and mobility. Finally, rule mining has also been used in the context of other applications, *e.g.* identifying patterns in shopping transactions [8, 25], identifying cause-effect pairs in network traffic [16], *etc.*

The fact that each user is usually associated with three locations (comfort zone) is most closely related to [15], where the authors used an anonymized data set from U.S. Census Bureau to find that a user's work and home location at the granularity of census tract (zip-code) can be used to uniquely identify about 5% of users. However, such reconstruction as suggested in [15] requires an adversary to have access to a mapping between the home/work locations and user identities, the availability of which we are not aware of; even the data set used in [15] was synthetic due to privacy concerns.

7. SUMMARY AND CONCLUSIONS

In this paper we conducted, to the best of our knowledge, the first large-scale study to characterize the relationship that exists between people's cyber interests and their mobility properties. Our key finding is that both users' mobility and locations heavily impact their application access behavior. We believe our results demonstrate significant promise for further research in this area, paving the way for many advances in understanding basic human behavior and in developing location-based services.

Summary. From the user perspective, our insights are the following: (*i*) Most users spend the vast majority of their time within the 'comfort zone' which consists of the top three locations, including home and work. (*ii*) Within the 'comfort zone', music prevails, particularly from home. Outside the 'comfort zone', the popularity of such bandwidth and battery intensive applications quickly fades. (*iii*) Dating applications are mostly accessed from within the 'comfort' zone, but neither from home nor work. (*iv*) Users who leave the 'comfort zone' have an inclination to 'staying connected' by accessing social networking sites, reading e-mail and news.

From the perspective of the most popular locations, our insights are the following: (*i*) There is a strong time-invariant bias towards specific applications at those locations at which hotspots are likely to occur. (*ii*) In most cases, such a bias remains unchanged when hotspots are created, *i.e.*, those

who join the hotspot show the same access behavior. (*iii*) Office hotspots are the only scenario in which the newly created majority manages to change the previously established application access bias.

From the user interactions perspective, our insights are the following: (*i*) The probability to meet *different* people with the same cyber interests is dominantly impacted by the number of users sharing the same interests in a given region. (*ii*) However, the frequency with which one meets with others who share the same cyber interests is dominated by the density of hotspots in a given area. (*iii*) Both mobile users and those present at popular hotspots have the potential to achieve a large number of interactions.

From the mobile provider and location-based services perspective, our insights are the following: (*i*) The observed location-based application access bias validates the enormous potential for existing location-based services, and opens the doors to a number of new ones. (*ii*) Due to the strong bias towards bandwidth intensive applications at a subset of hotspots, base-station-level caching at such locations would be very beneficial. (*iii*) There exists a significant observed anti-correlation between the use of bandwidth- and battery-intensive applications, such as music, with mobility. This finding can be a strong indicator of whether p2p-based mobile applications have a potential need or not; yet, we are unable to provide such a prediction. If the small usage is due to bandwidth concerns, then p2p mobile applications have a huge potential. Yet if battery is the concern, the result is reversed.

Acknowledgements

We would like to thank our shepherd, Laurent Mathy (Lancaster University) for his help with the final version of this paper. We are also grateful to the anonymous reviewers for their helpful comments and suggestions.

8. REFERENCES

[1] Flickr. `http://www.flickr.com/`.

[2] Location-based Advertising: Place Trumps Traditional Targeting. `http://venturebeat.com/2008/12/02/location-based-advertising-place-trumps-traditional-targeting/`.

[3] Loopt. `http://www.loopt.com/`.

[4] Pelago. `http://www.pelago.com`.

[5] Skout Brings Location-based Dating to the iPhone. `http://venturebeat.com/2009/01/21/skout-brings-location-based-dating-to-the-iphone/`.

[6] Skyhook Hybrid Positioning System: XPS. `http://www.skyhookwireless.com/howitworks/`.

[7] 3GPP2. CDMA2000 Wireless IP Network Standard: Accounting Services and 3GPP2 RADIUS VSAs, Oct. 2006. `http://www.3gpp2.org/public_html/specs/X.S0011-005-C_v3.0_061030.pdf`.

[8] R. Agrawal, and R. Srikant. Mining Sequential Patterns. In *ICDE*, Taipei, Taiwan, March 1995.

[9] D. Brockmann, L. Hufnagel, and T. Geisel. The Scaling Laws of Human Travel. In *Nature*, 439(7075), Jan. 2006.

[10] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of Human Mobility on Opportunistic Forwarding Algorithms. In *IEEE Transactions on Mobile Computing*, volume 6, 2007.

[11] D. Spielman, and S. Teng. Spectral Partitioning Works: Planar Graphs and Finite Element Meshes. In *IEEE Symposium on Foundations of Computer Science*, 1996.

[12] I. S. Dhillon. Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning. In *SIGKDD*, San Francisco, California, August 2001.

[13] N. Eagle, and A. Pentland. Reality Mining: Sensing Complex Social Systems. In *Personal Ubiquitous Computing*, volume 10, 2006.

[14] F. Chung. Spectral Graph Theory. In *American Mathematical Society, CBMS Regional Conference Series in Mathematics*, number 92, 1997.

[15] P. Golle, and K. Partridge. On the Anonymity of Home/Work Location Pairs. In *Pervasive*, Nara, Japan, May 2009.

[16] S. Kandula, R. Chandra, and D. Katabi. What's Going On?: Learning Communication Rules in Edge Networks. In *SIGCOMM*, Seattle, Washington, August 2008.

[17] T. Karagiannis, J. -Y. L. Boudec, and M. Vojnović. Power Law and Exponential Decay of Inter Contact Times Between Mobile Devices. In *MOBICOM*, Montreal, Canada, September 2007.

[18] B. Kernighan, and S. Lin. An Efficient Heuristic Procedure for Partitioning Graphs. In *The Bell System Techincal Journal*, volume 29, 1970.

[19] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong. SLAW: A Mobility Model for Human Walks. In *INFOCOM*, Rio de Janeiro, Brazil, April 2009.

[20] M. Gonzalez, C. Hidalgo, and A. Barabasi. Understanding Individual Human Mobility Patterns. In *Nature*, 453(7196), Jun. 2008.

[21] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and Texture Analysis for Image Segmentation. In *International Journal of Computer Vision*, June 2001.

[22] A. Miklas, K. Gollu, K. Chan, S. Saroiu, K. Gummadi, and E. de Lara. Exploiting Social Interactions in Mobile Systems. In *UBICOMP*, Innsbruck, Austria, September 2007.

[23] C. Rigney. RADIUS Accounting. 2000, Internet RFC 2866.

[24] C. Rigney, S. Willens, A. Rubens, and W. Simpson. Remote Authentication Dial In User Service (RADIUS). 2000, Internet RFC 2865.

[25] P. Tan, M. Steinbach, and V. Kumar. Introduction to Data Mining. Addison Wesley, 2006.

[26] F. Verhein, and S. Chawla. Mining Spatio-Temporal Association Rules, Sources, Sinks, Stationary Regions and Thoroughfares in Object Mobility Databases. In *DASFAA*, Singapore, April 2006.