Timely and Personalized Services using Mobile Cellular Data

Steven Mudda[c,*], Matteo Zignani[b,*], Sabrina Gaito[b], Silvia Giordano[a], Gian Paolo Rossi[b]

[a] *ISIN-DTI, SUPSI, Manno, Switzerland*
[b] *Computer Science Department, Universitá degli Studi di Milano, Italy*
[c] *Telepathy Labs GmbH, Zurich, Switzerland*

Abstract

The 21^{st}-century data-driven economy is rapidly evolving and large companies like Telecom operators are forced to adapt their business. They are shifting their focus from traditional but exhausted connectivity provider market towards a more services based market. Here competition is high, and other stakeholders are trying to monopolize the data-driven world of personalized services. But, Telecom operators are the custodians of Call Detail Records (CDRs), which captures mobility activities and social ties of a large number of users. Recently researchers observed that CDRs are the most valuable form of data to perform user-centric analysis, especially when related to mobility and habits.

In this paper, we demonstrate that CDRs can be used to provide personalized and timely services. Specifically, we show that it can be used to provide a recommendation service, one of the most popular personalized services. In addition, we demonstrate the advantage of leveraging human behavior characteristics for such services. Our `REGULA` recommendation algorithm, that builds on the analysis of human habits, outperforms the state of the art recommendation algorithms. We advocate that Telecom operators can leverage CDRs to provide personalized services in a data-driven world and can significantly alter the landscape of timely and personalized services.

Keywords: CDR-based LBSN, timely and personalized services, social ties

1. Introduction

In the last decade, mobile phone technology has witnessed the fastest spread of human technology and the number of mobile phone users in the world is

*indicates equal contribution
Email addresses: `steven.mudda@telepathy.ai` (Steven Mudda), `matteo.zignani@unimi.it` (Matteo Zignani), `sabrina.gaito@unimi.it` (Sabrina Gaito), `silvia.giordano@supsi.ch` (Silvia Giordano), `gianpaolo.rossi@unimi.it` (Gian Paolo Rossi)

expected to pass the five billion mark by 2019, with more than 2/3 of users owning a mobile phone[1]. Consequently, the saturation of the market is very close and has led to a radical change in the business model of Telecom companies. Since there are no people without a cell phone plan, the Telecom companies are just competing to acquire customers from their competitors. So recent years have seen a continuous push by Telecom companies to acquire new customers by offering lower prices and better communication services.

However, even this strategy has reached its limits and to attract new customers Telecom companies are nowadays moving towards providing more appealing and personalized services, apart from providing basic communication services. This is in line with the general evolution of this market. Even smartphone providers like Apple and Google have shifted their focus to provide services like Apple Streaming Service rather than just selling smartphones because of the decreasing growth rate in the number of smartphones being sold since 2016[2].

In this new competitive landscape, providing personalized services based on people's preferences is paramount to attract and retain new customers. The preferences of people can be learned by leveraging customer data. In the process of learning the preferences of people the Telecom operators are left behind by the different Online Social Networks and Media (OSNEM) platforms like Facebook, Twitter, Google, Instagram, etc. that analyze the big data collected from these platforms.

In this paper, we demonstrate that Telecom operators can utilize the big data they already have, i.e. Call Detail Records (CDRs). CDRs are a significant sink of data for analysis related to human mobility and sociality[3, 4]. They offer information about the cells (or regions) where a user performs some actions, which is usually a good approximation to reconstruct their mobility patterns and can effectively be used to provide time and location-aware services[5]. Further, CDRs give information about the direct communication links between people, i.e., the social network of users, enhancing the ability to provide personalized services. The rise of 5G and networked IoT devices will provide Telecom operators with large volume of data that can be further used to build analytical capabilities on top of their network to provide new services.

Since CDR data can capture both the mobility, the interests and the sociality of a large population [6, 7, 8], we show that CDRs can be leveraged to build an implicit Location Based Social Network (LBSN), which is a well-known type of OSNEMs (e.g., Foursquare, Facebook) that are traditionally used to provide personalized services. We call such implicit LBSN, *CDR-based LBSN* and show that it can also be used to provide timely and personalized services. An LBSN is typically a network formed by users who visit different places and share information about such places with friends in their social network. LBSNs capture both the spatial mobility of users by storing their location visits and information related to their social ties. People typically utilize LBSNs, such as Foursquare[9], to find popular places of interest close to their current location and also to extract, from their social network, the most prominent places. A *CDR-based LBSN* can be regarded as a network of users where: (i) there

are different kinds of on-phone interactions from which we can infer social relationships among people, (ii) from the spatiotemporal interaction records, we can associate people to the places they visit. The CDR-based LBSN is built to share the same LBSN modeling framework and provide a foundation for Telecom operators seeking to provide innovative services.

We argue that without obtaining the precise location information of people, Telecom operators can use a *CDR-based LBSN* to offer personalized and timely services as traditionally performed on LBSNs. Providing people a list of relevant new places to visit, accordingly to Kantar research [10], is the most popular type of LBS app (46%), followed by finding restaurants (26%), finding friends nearby (22%), checking public transportation (19%), and receiving special deals or offers from retailers (13%).

We show that traditional LBSNs and CDR-based LBSNs share a common model and thus we use our LBSN-designed model `REGULA` [11] to demonstrate the feasibility to offer novel services in CDR-based LBSN. Specifically, we provide the users with suggestion of novel places that can be visited, based on their preferences and locations. `REGULA` distinguishes from traditional algorithms by incorporating regular behaviors (e.g., people frequently visit a set of places [12]), the temporal importance of location and the real social information centered around the user. `REGULA` is based on five hypothesis about the behavior of people, and we verify them to be consistent both in standard LBSN datasets and in a CDR dataset.

An important aspect of this research direction, is the fact that these new services could be provided by Telecom operators with no significant additional legal and technical resources, as, for many operators, legal aspects are already specified in the contract with the customer and readily available CDR data can be used to offer better services.

We summarize our key contributions as follows:

- We introduce the CDR-based LBSN and formally define it. We also illustrate how to use the CDR data to build networks characterizing a CDR-based LBSN.
- We identify a set of human traits that are consistent in both CDR and LBSN data: a) people have regular visit patterns and explore locations close to their usual places; b) people go to locations recently visited by others, especially friends; c) people prefer to visit locations close to their current location.
- We use our algorithm `REGULA`, which embeds the above traits to reduce the search space for candidate locations and compare its performance against other state-of-the-art algorithms used on traditional LBSNs. We show that `REGULA` outperforms other competitors in terms of precision and recall.
- We provide a discussion on novel personalized services that can be provided with this approach.

The rest of the paper is structured as follows. In Section 2, we present our dataset and the preprocessing performed. Further, we introduce the CDR-based LBSN and describe how to construct it, and the characteristics of such networks. Then we present our assumptions on human behavior and show that they hold in Section 3. We illustrate the applicability of state-of-the-art algorithms on our CDR-based LBSN. Then we introduce our algorithm `REGULA` (in Section 4). Section 5 presents the evaluation of `REGULA` with the CDR-based LBSN dataset and its comparison with the state-of-the-art algorithms. In Section 6, we present the related work in LBSN. We then discuss the potentials of this work, and we conclude the paper along with future directions.

2. CDR Dataset

In our study, we leverage a large anonymized dataset of Call Detail Records (CDRs) [13] capturing voice calls, short text messages (SMS) and Internet traffic of about 1 million subscribers of an international mobile operator. The on-phone activities contained in the dataset are restricted to the metropolitan area of Milan for a period of 67 days, from March 26 to May 31, 2012. During this two-month period, an overall amount of more than 63 millions phone-call records, 20 million text records and more than 61 millions Internet activities took place.

Each kind of event provides us with different types of information. Specifically, on-phone communications, such as calls and text messages, provide that, when a user calls or sends a text message, the user IDs of sender and receiver, the cell ID of the handling towers and the date and time of established contacts are all recorded. The only difference between calls and texts is the duration; in fact, calls may last from seconds to hours, making it possible to extract the handling cellular tower of the callee at the end of the on-phone activity. Differently, the information about the network traffic concerns only the users doing Internet activities, i.e. when and where they connect to Internet.

Despite the above differences, we cast all three kinds of event into a single formalization. Regardless of the event type, each record in the dataset is described by the 7-ple $t_{CDR} = \langle s, r, t_{start}, t_{end}, d, loc_{start}, loc_{end} \rangle$, where s and r respectively represent the sender and the receiver of the call/text [1,2], t_{start} is the initial time of the activity (when the call starts or an SMS is sent), t_{end} is the ending time of the event, d is the duration and loc_{start} is the serving cell the user s is attached to when the activity get started or ended (loc_{end}). Note that text message and Internet activity has null duration d and empty t_{end} and loc_{end} fields, while Internet activity also has the field receiver r set to null.

Due to data confidentiality policies, we are not able to access the type and content of the Internet connection data; not enabling us to develop content-

[1] Customers' anonymity is guaranteed by a surrogate key which identifies each user.

[2] Customers' privacy is guaranteed as customers' data are used only for improving services, if agreed in contract.

based services [14], and to infer social relationships mediated by instant messaging applications, such as WhatsApp or Facebook Messenger. In our case, the latter point is not a limitation since the CDRs which we are using here, span a period when the amount of text messages equals instant messaging one [3]. So, through voice calls and SMSs, we are still able to capture the relationships mediated by mobile phones.

2.1. CDR-based Location Based Social Network

Call Detail Records represent a valuable data source which enables us to capture the social relationships among the operator's customers. In fact, they are considered an essential tool to analyze social dynamics of a large population. Also, CDRs provide the location of the users, enhancing the connection between communications mediated by mobile devices and habits and relationships in the real world. So, mobile phone data are the most valuable form of data to perform user-centric analysis, especially when related to mobility and sociality.

The combination of social networking and geographic information is also typical of location-based social networks (LBSN), so we can cast our CDRs dataset into the LBSN modeling framework. Usually, LBSNs are characterized by three graphs: *i*) a *location-location graph* which expresses a different kind of relationships among the locations visited by the users; *ii*) a *user-location graph*, modeling and summarizing the visit patterns of the users; and *iii*) a *user-user graph* that represents the relationships between users.

In this work, we focus on the last two networks. Specifically, we define the graph $G = (U, L, E_f, E_m)$, where U and L represent the set of users and serving cells/locations, respectively. $E_f \subseteq U \times U$ is the set of links representing social relationships between users, and $E_m \subseteq U \times L$ contains the links which indicate a person $u \in U$ has been attached to the cell $loc \in L$ at least once in two months. To better describe the visit patterns of the locations in L, we introduce a mapping v that associates to each link $e = (u, loc) \in E_m$ a list of the timestamps when the user u has visited the location loc, transforming E_m into a temporal network. The above definition includes both the user-location graph and the user-user graph and supports the algorithm we will introduce in the following sections. In the following, we will also describe how we infer the user-user graph E_f from call and text message records, and how we define the set E_m by examining call, text message and Internet activities.

2.2. Building the friendship graph

Text message and call records are the primary data source to infer the social interactions mediated by on-phone communications; interactions which are going to form the friendship graph (U, E_f). However, deciding whether or not an interaction has a social value depends on the purpose of communication. In fact, calls or text messages are noisy due to advertisements and commercial messages

[3] https://www.agcom.it/documents/10179/2681146/AGCOM+-+Annual+Report+2012_02_The_communications_sector_in_Italy/de68cbc2-0860-42c2-9915-4102dff2feb1

or communications issued by call centers and ask for a pre-processing to extract the relevant interactions only. To this aim, we apply a bunch of heuristics to filter out 'not-social' traits. First, according to the literature on mobile phone data cleaning [15, 16, 17, 18], we keep only reciprocated communications, i.e., there is a link from A to B if and only if A calls B and vice-versa. Furthermore, we filter out spurious and not persistent interactions by discarding the pairs of users whose sum of call duration is less than one minute or whose total number of on-phone interactions is lower than 4. Second, we filter out calls/-texts involving other mobile operators' customers. This way we eliminate the bias given by the limited amount of information on the interactions/locations of extra-operator customers. After the cleansing process, we obtain the undirected graph (U, E_f) made up by about 460,000 customers and 1,430,000 connections, which generate almost 7 millions calls, 317.000 hours of conversations and 2 million texts.

In previous works [4], we have observed that the so built friendship graph exhibits typical characteristics of social networks: it is a scale-free network that exhibits small-world properties and tightly clustered groups; the same structural characteristics revealed by traditional LBSN social networks [19].

2.2.1. Building the user-location graph

The realization of the user-location graph (U, L, E_m) relies on all the three types of record. Unlike the creation of E_f, we exploit all the records so that we obtain the mobility traces of the operator's customers as much detailed as possible. Specifically, we extract the locations of the callee/caller at the beginning and at the end of the call, the locations of the sender/receiver of the text message and the position of the user when s/he reaches a fixed amount of data traffic. By these data we create the set of the locations L, corresponding to all the cells visited by at least one user, and define the temporal connections (u, loc) in E_m.

The mobile operator has provided the location of the above activities in terms of area names of the zones , i.e. a group of cells, but without information about cells coverage area and their exact positioning [20]. Thus, to estimate the effective cell size distribution, we applied the following method. First, we obtained the position of the cells by querying the `LocationAPI` web-service offered by *UnwiredLabs*[4], which provides the cell center along with the estimating error. For half of the cells the estimation error returned by the service is below 300 meters, but the service does not release any detail on the algorithm they use to infer the location of the cells. Second, we assume each cell $cell_i$ being a circle with center c_i and coverage radius r_i since we do not have any information about the strength of the signals and the interference caused by buildings. Then, we compute r_i as half the average distance between c_i and the centers of the cells surrounding $cell_i$. In Figure 1 we report the cumulative distribution function (CDF) of the cell radius as a function of the cell position. Due to the

[4]Website `http://unwiredlabs.com/`

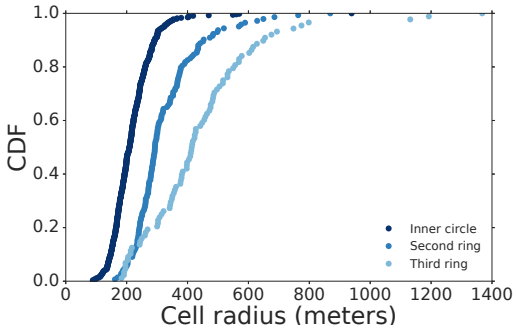


Figure 1: The cumulative distribution functions of the radius of the cells in the three city regions.

concentric topology of Milan, we group cells into three city regions: the inner circle of 3 Km radius - the city center; a second ring from 3 Km to 4 Km from the city center; and a third ring, in the range of 4 Km to 5 Km. We detect 538, 143, and 88 cells inside each region, respectively. The figure shows that the radius of the cells increases as we move farther from the city center, in fact, the average radius of the inner circle, second ring, and third ring are 217, 325 and 446 meters, respectively.

Given this small coverage radius, we can provide a good approximation of the mobile user's position. In fact, unlike most of the previous localization studies which derive location by cell towers covering as wide as a few kilometers zones, in the city space cells have a very small coverage, of one or very few hundred meters, that approximately corresponds to a building block.

2.3. Discussion

Although CDRs and LBSNs share a common model, they present a few differences which may affect applications typical of LBSNs. The first difference lies on how data about locations are sampled. In LBSNs the location can be linked to published contents (geo-tagged media based) - passively by the device or explicitly by a user's labeling - or the venue is the key point of the service (point-location based) and the subject of all the user's activities, e.g., Foursquare or Yelp. By contrast, CDR data are not point-location based - the serving cell is an additional information released by the billing system. Neither they are geo-tagged-media based, since they do not focus on specific content or interest, such as photos, videos; and users do not explicitly set the location.

Second, there are differences about the precision of the localization and the role of locations within the system. As for the former, in most LBSN the geo-localization is GPS-based, thus providing high precision and the possibility to associate a specific interest to the position of the user. By contrast, the localization in CDRs has a cell-granularity which makes the identification of the semantic of the location more difficult.

As for the role of locations, LBSNs are social- and purpose-driven platforms and locations are the key points to identify an interest common to many people and to promote offline/online social relationships. Conversely, the location in CDRs is a side-information of calls, text messages and Internet traffic. In this case, the platform is not built around the concept of location rather the system exploits this information to manage the services it offers.

3. Timely and personalized services in CDR

In the previous section, we have shown how CDR and typical LSBN share the same model. We thus advocate that the timely and personalized services performed on LBSNs, can be applied on CDRs (CDR-based LBSN), leveraging on their rich mobility information. However, to the best of our knowledge, CDRs have not yet been exploited for such services. We can argue that this is because in most of existing CDRs the granularity of user's localization was too large to provide any relevant service.

We are going to introduce 5 hypothesis. In some forms, some of them where demonstrated already several years ago, either in general, or for CDR, or both. This is the case for H1 and H2, demonstrated in the general case in [21, 22, 23] and for CDR in [3], or H4, demonstrated in general in several works, as discussed in [24]. In the other cases, as far as we know, such hypotheses have not been demonstrated, but are observations derived by human psychological behaviour.

In order to quantify such LBSN nature of CDR, we present an analysis of the main characteristics of two standard LBSN datasets, Gowalla and Brightkite, and compare them with the characteristics of our CDR-based LBSN.

3.1. Observations from datasets

In this section, we analyze our CDR-based LBSN dataset and compare it with two standard and publicly available LBSN datasets, Gowalla and Brightkite [25], that contain check-in records of people, to understand their regular mobility patterns and temporal behavior. Based on multiple studies [21, 12] about human mobility patterns, we formulate and test the following five hypothesis:

H1: Regularity - Users regularly (or habitually) visit a set of locations i.e., their Frequently Visited Locations (FVLs).

H2: Vicinity - Users visit places in the vicinity of their FVLs.

H3: Recency - Users are more likely to go to places that were visited recently by others.

H4: Sociality - Users are more likely to go to places that were visited recently by their friends.

H5: Inertia - Users are more likely to go places geographically close to their present location.

Table 1: Symbols

Symbol	**Description**
U, L, T	user set, location set, check-in time set (in days)
u, l	user $u \in U$, location $l \in L$
$t_{(u,l)}$	time at which u visited location l, $t_{(u,l)} \in T$
C	set of all check-ins $\{< u, l, t_{(u,l)} >\}$
E_f	set of all friendship links between U users
G	friendship graph of U users and E_f edges
VL_u	unique locations visited by user u, $VL_u \subseteq L$
VL_u^{lastk}	*lastk* locations visited by user u, $VL_u^{lastk} \subseteq VL_u$
FVL_u	frequently visited locations of user u, $FVL_u \subseteq L$
$P_u(l)$	Preference of user $u \in U$ for location $l \in L$
F_u	friends of user u, $F_u \subseteq U$
$VL_f, f \in F_u$	locations visited by the friends of user u
l_{last}	The last location visited by a user $l_{last} \in L$
RT	time of recommendation (in days)
K	number of recommended locations
$ts(u, l)$	temporal score assigned by user u to location l
$ts(l)$	aggregated temporal score of location l
TS	set of aggregated temporal scores for all L
$ds(u, l)$	distance score assigned by user u to location l
$fs(u, l)$	friendship score assigned by user u to location l
$rs(l)$	recommendation score assigned to location l
$d(l, \hat{l})$	distance between locations $l \in L$ and $\hat{l} \in L$

Formally, we can define our assumptions as follows. Given a set U of people, with the set of friends F_u for user $u \in U$, and a set L of locations visited in the period T. Further definitions of symbols are given in Table 1.

H1: Regularity - $\forall u \in \hat{U}, \hat{U} \subseteq U, \hat{U} \gg \hat{U}^c$

$$P_u(l) > P_u(l') \qquad l \in FVL_u, l' \notin FVL_u, FVL_u \subseteq L$$

H2: Vicinity - $\forall u \in \hat{U}, \hat{U} \subseteq U, \hat{U} \gg \hat{U}^c$

$$\min_{l_f \in FVL_u}(d(l, l_f)) < \min_{l_f \in FVL_u}(d(l', l_f)) \quad \Rightarrow \quad P_u(l) > P_u(l')$$

H3: Recency - $\forall u \in \hat{U}, \hat{U} \subseteq U, \hat{U} \gg \hat{U}^c$

$$l \in VL_{\hat{u}}, \hat{u} \in U \ \ \& \ \ l' \notin VL_{\bar{u}}, \forall \bar{u} \in U \quad \Rightarrow \quad P_u(l) > P_u(l')$$

H4: Sociality - $\forall u \in \hat{U}, \hat{U} \subseteq U, \hat{U} \gg \hat{U}^c$

$$l \in VL_f \ \ \& \ \ l' \notin VL_f, f \in F_u \quad \Rightarrow \quad P_u(l) > P_u(l')$$

H5: Inertia - $\forall u \in \hat{U}, \hat{U} \subseteq U, \hat{U} \gg \hat{U}^c$

$$d(l, l_{last}) < d(l', l_{last}), l_{last} \in VL_u \quad \Rightarrow \quad P_u(l) > P_u(l')$$

Table 2 presents the characteristics of the two standard LBSN datasets: Gowalla and Brightkite, and our CDR-based LBSN dataset. To validate our hypothesis, we conducted tests at different points in time. When a test is conducted at time t, a subset of complete dataset up to time t is used to evaluate the hypothesis. For Gowalla, the analysis was conducted at days $t = \{90, 120, .., 510\}$ and for Brightkite at $t = \{90, 120, .., 870\}$. For our CDR-based LBSN dataset, the analysis was conducted on two days $t = \{30, 45\}$. For all tests on LBSN datasets, we only consider users who have at least one new check-in in next 30 days, i.e., $[t, t + 30)$. The filter to select users in Gowalla and Brightkite is low because the dataset was collected over a long period ($>$ 2 years) and some users join or leave the network during this collecting period. For all tests on our CDR-based LBSN dataset, we consider only users, who have at least 10 check-ins in next 15 days. The threshold to select users in CDR-based LBSN is higher because the dataset spans a shorter period (2 months).

3.1.1. H1: Regularity

Large-scale studies [21] on human mobility patterns have shown that humans exhibit regular mobility patterns, i.e., they visit a few set of locations repetitively like a favorite pizzeria, McDonald's near home, etc. We refer to any location that a user has visited more than once as one of her/is relevant Frequently Visited Location (FVL). Figure 2 shows the fraction of users with at least one FVL at different testing times in Gowalla, Brightkite and CDR-based LBSN datasets. We observe that a significant number of users regularly visit at least one or more set of locations. Further, in Gowalla and Brightkite datasets, we find that the fraction of users with at least one FVL increases with time, that suggests that users tend to regularly visit the same set of locations.

Table 2: Dataset Characteristics

	Gowalla	**Brightkite**	**CDR-based LBSN**
Covered Period of Check-ins	569 Days	929 Days	68 Days
Number of Users	196,591	58,228	468,466
Number of Locations	1,280,956	772,933	901
Number of Check-ins	6,264,203	2,627,870	118,752,518
Number of Friendship Links	950,327	214,078	1,432,938

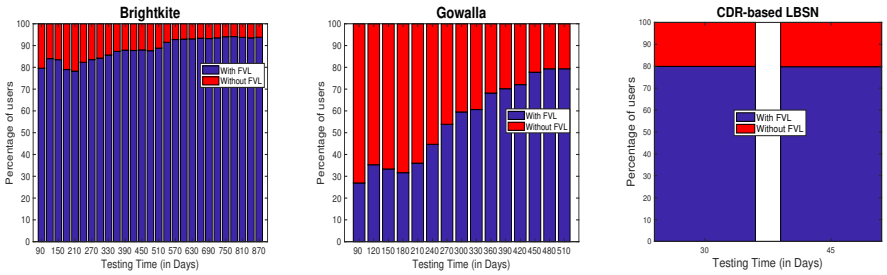


Figure 2: For each snapshot of dataset at different times (days), fraction of users with atleast one FVL in Gowalla, Brightkite and CDR-based LBSN datasets.

3.1.2. H2: Vicinity

Prior analysis done by other researchers on different LBSN datasets has shown that people exhibit a lot of inertia and are unwilling to travel long distances from their current location [22, 23]. These works also show that more than 70% of new check-in locations are within 10km of the previous check-in location. Based on these studies, and on the fact that people spend most of their time in the FVLs [21, 3], we hypothesize that people usually tend to explore new locations that are close to their frequently visited locations such as places close to their favorite pizzeria, etc. We call a location "new" for a person when such location has never been visited by this person before. Formally:

$$\forall u \in U, l \in L \text{ is } new \text{ for } u \text{ if } l \notin VL_u$$

We test this hypothesis by measuring the fraction of users with new check-ins around their FVLs.

Figures 3, 4 and 5 present the fraction of users who have visited at least one new location within a fixed range of their FVLs in Gowalla, Brightkite and CDR-based LBSN datasets. We observe that more than 80% of users visit locations close to their FVLs in CDR-based LBSN dataset. While in Gowalla and Brightkite datasets the fraction of users that visit locations close to their FVLs keeps growing with time. In order to reduce the impact that could be generated by well-known ping-pong effects that characterize mobile device associations to base stations, we further repeated the analysis while removing the neighbourhood of the FVL base stations. What we found is a very similar distribution, demonstrating that such effect has a very minor impact for recom-

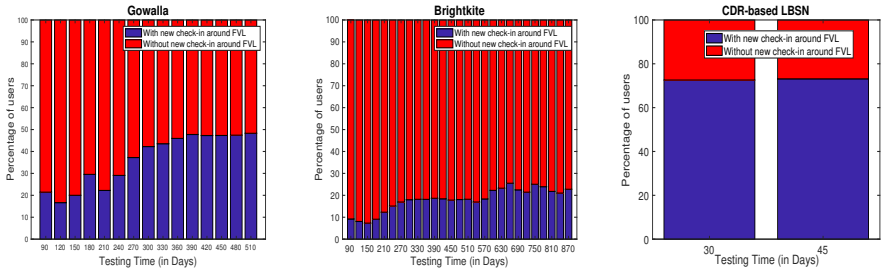


Figure 3: For each snapshot of dataset at different times (days), fraction of users with new check-ins within a distance of 1km from their FVLs at different times (days) in Gowalla, Brightkite and CDR-based LBSN datasets.

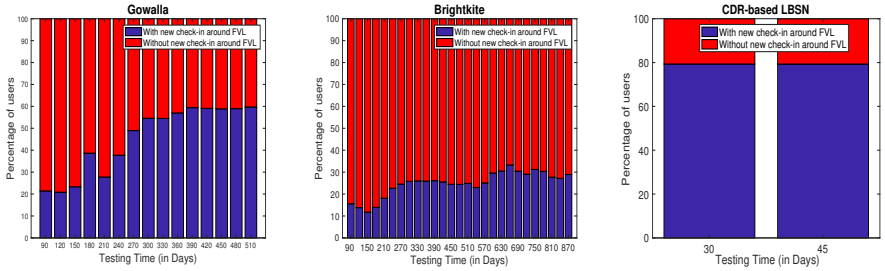


Figure 4: For each snapshot of dataset at different times (days), fraction of users with new check-ins within a distance of 2.5km from their FVLs at different times (days) in Gowalla, Brightkite and CDR-based LBSN datasets.

mendation. Therefore, we can utilize this observation to provide more timely and personalized services, for example to give correct recommendations to users by suggesting them new locations around their FVLs.

3.1.3. H3: Recency

We hypothesize that people go to new places that were recently visited by others. To test this hypothesis, we measure the number of new check-in locations that were visited by others at most m days back. The maximum possible value of m for Gowalla, Brightkite and the CDR-based dataset is the duration of the period they cover, i.e., 569, 929 and 68 days respectively (see Table 2).

Figure 6 presents the most recent day of visit for all new check-ins in Gowalla and Brightkite datasets. We observe that a significant number of users go to places that were visited by others in the last 30 days. Therefore, we can say that users mostly ignore locations that were visited a long time back. This analysis confirms, if necessary, the importance of time while providing more timely and personalized services.

In our CDR-based LBSN dataset, we observe that all users go to places that were visited by others in the previous day. Since the ratio of the number of users to the number of location is very high (519), i.e., on an average each location was visited by around 519 users, the probability that a user had visited a given

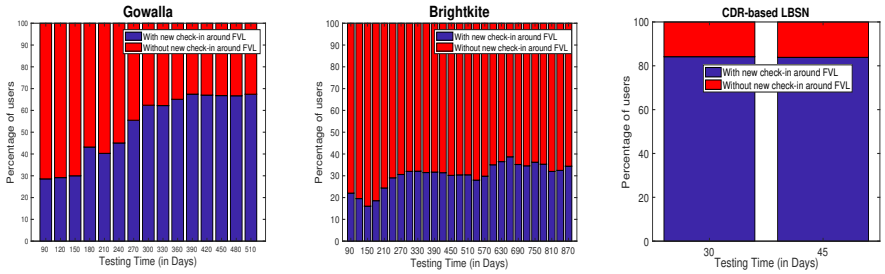


Figure 5: For each snapshot of dataset at different times (days), fraction of users with new check-ins within a distance of 5km from their FVLs at different times (days) in Gowalla, Brightkite and CDR-based LBSN datasets.

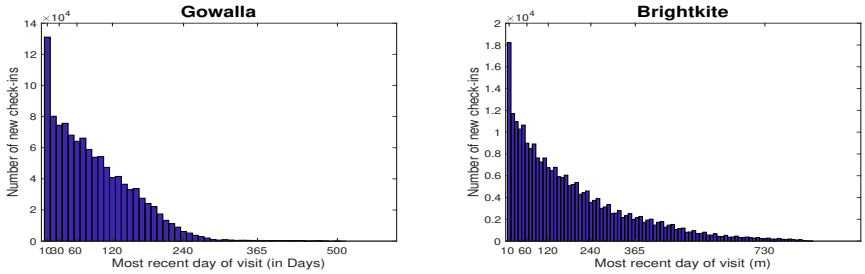


Figure 6: Most recent day of visit for all new check-ins.

location on the previous day is very high. Thus, while confirming Gowalla and Brightkite analysis, the distribution is not significant.

3.2. H4: Social Strength

Humans are social animals and heavily influenced by their friends and family. We conducted tests to check whether a user in our CDR-based LBSN dataset is affected by her/is social network friends. Specifically, we measure the percentage of users visit new places after their friends have visited them.

Figure 7 presents how friends influence the choice of new places visited by a user. For each user, we find out the total number of unique new location visits and then measure what percentage of them were recently visited by their friends. We observe that a significantly large number of users have gone to at least one new location that was previously visited by their friends. Among them, the number of users who do not visit a single location after their friends is only about 2% of the total number of users i.e., 688,302. Even if this number is very low, in order to avoid a causality relationship, in Section 5, we use a holistic scoring function that takes into account friendship, temporal and distance scores.

3.3. H5: Inertia

Our last hypothesis is that people exhibit a lot of inertia and often tend to go to nearby places. For example, tourists are more likely to eat close to the

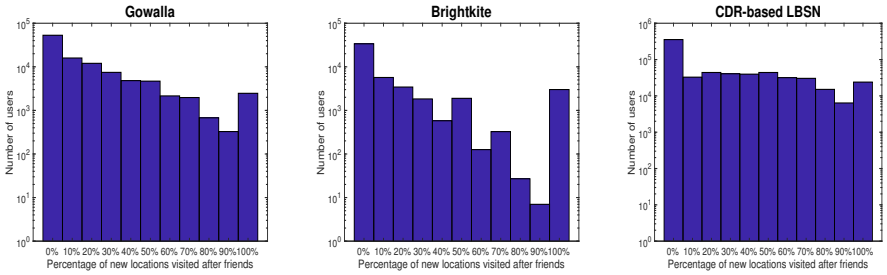


Figure 7: Distribution of users who have visited a new location after their social friends.

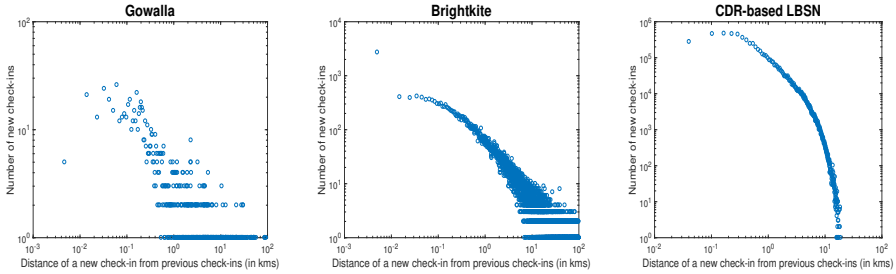


Figure 8: Distribution of how far do people travel to visit new locations.

monuments or other attractions they visit. Thus, in addition to our hypothesis that people move closer to their FVL (H2), we argue that, when they are offered a new service, they are going to accept if this requires a minimal effort. For example, in the case of recommendation for new place to visit, they are likely to go there if the place is geographically close to their present location. We conducted tests where, for each new location visited by a user in our CDR-based LBSN dataset, we measure its geographical distance from locations visited before to see whether they exhibit inertia or not.

Figure 8 presents a distribution of distance traveled for a new location visit by users in Gowalla, Brightkite, and our CDR-based LBSN dataset. We observe that almost 90% of new locations visited by users in CDR-based LBSN are within a range of less than 10km. Further, more than 84% of new locations visited by users in Gowalla and Brightkite are within a range of less than 10km. It confirms our hypothesis that people prefer going to places geographically close to their present location.

Having validated our hypothesis on our CDR-based LBSN dataset along with two major LBSN datasets, we can state that CDR-based LBSNs have similar characteristics to standard LBSNs for timely and personalized services purposes. We then formulate, from the five hypothesis, the following observations.

We use these five observations to develop our recommendation model `REGULA` that will be used as example of timely and personalized service for Telecom operators.

Observation 1	People regularly visit a certain set of places (i.e., frequently visited locations).
Observation 2	People usually visit places in the vicinity of their frequently visited locations.
Observation 3	People usually visit places recently visited by others.
Observation 4	People usually go to places visited by their friends.
Observation 5	People usually go to places close to their own recently visited places.

4. REGULA: Recommendation Algorithm

As discussed in the introduction, the primary objective of this work is to demonstrate that CDR-based LBSN datasets can be used to provide timely and personalized services. As example of such services, we use recommendation, which is a typical service in LBSNs, and we obtain precise and accurate recommendations by exploiting the specific features described in Section 3. Therefore, after we have shown that *CDR-based LBSNs* are qualitatively and quantitatively LBSNs, we want to demonstrate that we can apply standard recommendation algorithms to recommend new places or regions a person is more likely to visit. For this reason, we present in this section our recommendation model `REGULA` that we will compare with state-of-the-art algorithms. The original algorithm, presented in [11], was designed to get benefits from the presented hypotheses, but in a limited scenario with fixed parameters. Here we relax such constraints, and produce a more exhaustive analysis, performing a grid search of each parameter within a range, and measuring its impact on the performance metrics. Specifically, we provide such extensive evaluation to measure the impact of the different features of REGULA on the recommended performance for CDR-based LBSN.

Given the check-in history C of U users at L locations, the goal of `REGULA` is to recommend a list of K new locations (out of L) to any user u. In this section, we first describe three functions used to assign scores to all L locations. Later, we present our `REGULA` algorithm that utilizes these scoring functions to provide location recommendations. The symbols used in the scoring functions are given in Table 1.

4.1. Temporal Scoring Function

Our temporal function to assign score to a location l by user u is given by the following equation:

$$ts(u,l) = \begin{cases} \frac{t_{(u,l)}}{RT} & l \in VL_u \\ 0 & l \in L \setminus VL_u \end{cases}, \quad ts(u,l) \in [0,1]$$

Let us assume that we have the check-in history of a user $a \in U$. User a has visited locations $p \in L$ and $q \in L$ at timestamp $t_{(a,p)}$ and $t_{(a,q)}$ respectively ($t_{(a,p)} < t_{(a,q)}$). Based on the *Observation 3* (refer to Section 3.1), the recently visited location q must be assigned a higher score compared to location p. Since $t_{(a,q)}$ is greater than $t_{(a,p)}$ the above equation ensures that the temporal score assigned to location q is greater than location p. Therefore, our generalized temporal scoring function ensures that locations that were visited in distant past are assigned smaller scores compared to recently visited locations.

The aggregated temporal score of a location l driven from all user check-ins, can be computed as follows:

$$ts(l) = \sum_{\forall u \in U} ts(u,l) \quad , \quad ts(l) \in [0, |U|] \tag{1}$$

4.2. Distance Scoring Function

Utilizing *Observation 5*, the distance scoring function assigns scores to locations based on the distance to last few check-ins of a user. Let VL_u^K be the set of last K locations visited by user u. The following equation gives the function to assign a score to location l by user u:

$$ds(u,l) = \begin{cases} 0 & l \in VL_u \\ \frac{1}{\min_{l' \in VL_u^{lastk}} \text{dist}(l',l)} & l \in L \setminus VL_u \end{cases} \tag{2}$$

$$\text{dist}(l', l) - \text{Euclidean distance between locations } l' \text{ and } l$$

This distance scoring function ensures that new locations closer to the last K visited locations of user u are assigned higher scores compared to other locations.

4.3. Friendship Scoring Function

Utilizing *Observation 4*, this function assigns scores to locations based on check-in history of the friends of a user. Our friendship function to assign score to location l by user u is given by the following equation:

$$fs(u,l) = \begin{cases} 0 & l \in VL_u \\ \sum_{v \in F_u} (\frac{t_{(v,l)}}{RT} \cdot \alpha^{\frac{t_{(v,l)}}{RT}}) & l \in L \setminus VL_u \end{cases} \tag{3}$$

$$\alpha > 0, \text{ a constant weighting factor}$$

This friendship scoring function ensures that new (or unvisited) locations visited by friends of a user u are assigned higher scores compared to other locations.

Finally, for every user u we define a recommendation score assigned to any unvisited location l as follows:

$$rs(u,l) = \; ts(l) + fs(u,l) + ds(u,l) \qquad l \in L \setminus VL_u \tag{4}$$

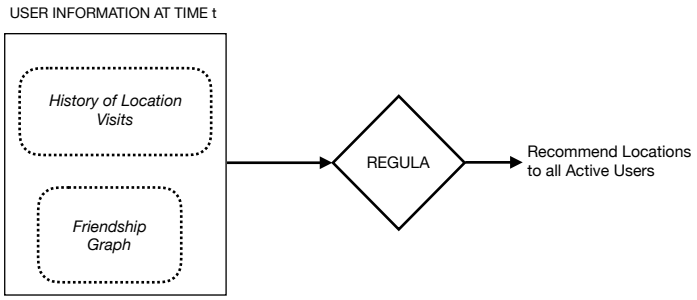


Figure 9: The overall process to recommend locations using `REGULA` at time t

The pseudo-code of `REGULA` recommendation algorithm is depicted in Algorithm 1. Based on *Observation 1*, we first find out the most frequently visited locations (FVLs) of a user u (Line 3). Later, for each FVL, we obtain a list of all unvisited locations within a certain geographical region or bounded box (Line 9) (Utilizing *Observation 2*). Further, for every unvisited location l_k of u, we compute the temporal score (Line 12). Since each unvisited location is within a bounding box around FVL, we increment the temporal score of each unvisited location with the temporal score of associated FVL (Line 14). For every unvisited location, we also compute the distance score (Line 16) and friendship score (Line 19). The final recommendation score ($rs(l_k)$) assigned to l_k is the sum of three scores: 1) aggregated temporal score, 2) distance score and, 3) friendship score (Lines 12-19). Top N locations with highest scores ($rs(l_k)$ are recommended to a user. Figure 9 presents the complete recommendation process of `REGULA`.

5. Evaluation

In this section, we demonstrate that *CDR-based LBSNs* can be successfully used for a timely and personalized service such as recommendation, and that taking into consideration the five observations we presented in Section3 is advantageous. To do so, we present the quantitative evaluation of `REGULA` and compare its performance with other algorithms in recommending a location to users based on our Call Detail Records. In particular, we show that the regular mobility habits of a person and distance to a recommended location significantly impacts the accuracy of the recommendations.

We evaluated the performance on our *CDR-based LBSN* dataset (refer to Table 2). Since we have the temporal mobility of users, we evaluated sequentially on two different days: 30th and 45th. The training data considered for a test on day t are all locations visited in the interval of $(0, t)$. The testing data was all locations visited in the interval of $[t, t + 15days)$. In each test, `REGULA` was used to provide K locations to every user who has at least one new check-in in the testing data.

1: `REGULA` $(u, TS, C, G, K, B, M, lastk)$
Input : user $u \in U$; Set of temporal scores $TS := \{(l, ts(l)\}, l \in L$; Set of all check-ins C; Friendship Graph, $G = (U, E_f)$; K: Number of recommended locations: ; B : Bounding box size; M: Maximum number of FVLs; *lastk* : Number of last locations
Output: A set of K recommended locations, R
2: Set Bounded Box size $D = B$(in km)
3: Get M most frequently visited locations of u using C, L_{fvl}
4: Get direct friends of u using G, F_u
5: Get *lastk* locations visited by u using C, VL_u^{lastk}
6: $R = \emptyset$
7: **foreach** $l' \in L_{fvl}$ **do**
8: **if** l' $\exists$ TS **then**
9: Get unvisited locations in bounded box of size D around l', $Box(l', D)$
10: **foreach** $l_k \in Box(l', D)$ **do**
11: `/* Assign temporal score of` l_k `*/`
12: $rs(l_k) = TS.ts(l_k)$
13: `/* Add temporal score of FVL` l' `*/`
14: $rs(l_k) = rs(l_k) + TS.ts(l')$
15: `/* Compute and add distance score of` l_k `using Equation 2 */`
16: $rs(l_k) = rs(l_k) + ds(u, l_k)$
17: `/* Compute and add friendship score of` l_k `using Equation 3 */`
18: **if** $F_u \neq \emptyset$ **then**
19: $rs(l_k) = rs(l_k) + fs(u, l_k)$
20: **end**
21: **if** $\{(l_k, rs(l_k))\} \in R$ **then**
22: Update R if new $rs(l_k)$ is larger
23: **else**
24: $R \leftarrow R \cup \{(l_k, rs(l_k))\}$
25: **end**
26: **end**
27: **end**
28: **end**
29: Sort R in descending order of scores $rs(l_k)$
30: $R \leftarrow$ Get top N locations in R
31: return R

Algorithm 1: Pseudo Code of `REGULA` Recommendation Algorithm

We evaluated the performance of REGULA using two widely used metrics i.e., precision at k ($p@k$) and recall at k ($r@k$) and are defined as,

$$\text{p@k} = \frac{1}{N}\sum_{u=1}^{N}\frac{|S_u(k) \cap V_u|}{k}, \text{ r@k} = \frac{1}{N}\sum_{u=1}^{N}\frac{|S_u(k) \cap V_u|}{|V_u|}$$

$S_u(K)$ is the set of top K locations recommended to a user u and V_u is the set of locations visited. $p@K$ metric measures how many locations (out of K recommendations) were visited by users. $r@k$ metric captures how many locations visited by the user were part of the recommendation. We also would like to highlight that, as the precision and recall of a recommender system are defined as the average precision and recall over all the users, the recommendations provided by the system to a user does not change in a given instance of evaluation. Thus, there is no variance of such metrics in the set of locations recommended to a given user.

5.1. Evaluation of REGULA

We vary the parameters of our REGULA model and measure its impact on the performance metric. The different parameters of REGULA are:

1. FVL – The number of Frequently Visited Locations control the regions from where candidate locations are selected. We perform a grid search over FVL $\in \{10, 20, 30, 40, 50\}$ and measure its impact on the performance metrics.
2. *bbox* – The Bounding box (bbox) represents a rectangular region of interest. Bounding boxes placed around every FVL capture the set of candidate locations. Therefore, the total number of candidate locations to rank depends on the size of Bounding Box. A larger bounding box will capture more candidate locations than compared to a smaller bounding box. We perform a grid search of bounding box size over $bbox \in \{1km, 2.5km, 5km\}$ and measure its impact on the performance metric. The bounding box size is the diagonal length of the rectangle.
3. *lastk* – This parameter controls how many of the locations visited in recent past contribute to the distance score of a candidate location (refer to Section 3.1.3). We perform a grid search over $lastk \in \{10, 20, 30, 40, 50\}$ and measure its impact on the performance metrics.
4. α – It controls the impact of friendship ties between users on the ranking of candidate locations. We perform a grid search over $\alpha \in \{10, 50, 100, 200\}$ and measure its impact on the performance metrics.

For each experiment, we vary any of the above four independent parameters, together with K - the number of locations recommended to a user u and measure its impact on the performance metric. Results of experiments performed on the 30th day and 45th day are shown in Figures 10, 12, 14 and Figures 11, 13, 15 respectively.

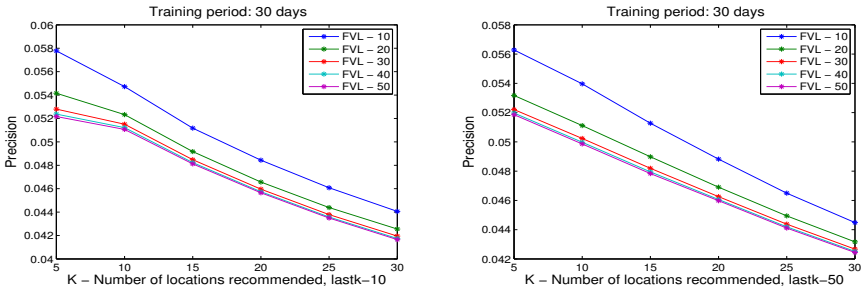
In the following sections, we demonstrate that REGULA can utilize information about the frequently visited locations of a user to provide better recommendations compared to the baseline algorithms. We also show that considering new locations within a small distance from the FVLs lead to better recommendations.

5.1.1. Impact of FVL on Precision

Figures 10 and 11, show how the performance of REGULA varies for different FVLs and a fixed bounding box of $1km$. They also show the difference in performance when $lastk$ is varied from 10 to 50. For brevity, we only show the performance when $lastk$ is 10 or 50 (for all intermediate values of $lastk$ refer to Figures in Appendix). We observe that for a fixed K as the number of FVLs is increased from 10 to 50, the set of candidate locations obtained within a fixed bounding box increases along with difficulty to correctly rank them. Adding more FVLs increases the noise to rank candidate locations. The phenomena behind this effect have been presented in Barabasi's work [21]: people tend to be inertial. Therefore, they will hardly visit places far away from current location. Thus, places that are close to far away FVLs are unlikely to be visited. Further, when we increase the number of locations recommended K, the total number of locations that need to be ranked correctly also increases. Therefore, as expected, we observe that with increasing K the performance decreases, independently from the other parameters.

In Figures 12, 13 we show the performance of REGULA for different FVLs and a fixed bounding box of size $2.5km$. When K is varied from 5 to 30, we observe a significant change in the pattern of performance of REGULA compared to the $1km$ bounding box. For a user with a certain number of FVLs, when the bounding box size is increased from $1km$ to $2.5km$ the number of potential unvisited locations of user increases because a bigger bounding box around FVL will contain more number of candidate locations than compared to a smaller one. Analyzing our CDR-based LBSN dataset, we observe that, on an average, each user in the tests conducted on 30th and 45th day have visited 15 locations, that is: the average value of $|V_u|$ is 15. When K increases from 5 to 10, REGULA's performance improves because it can correctly rank the top-k locations. However, when K is increased beyond 10, the performance reduces because, as we said above, with a higher value of K the total number of locations that need to be correctly ranked increases, but the $2.5km$ bounding box is unable to capture all potential candidate locations.

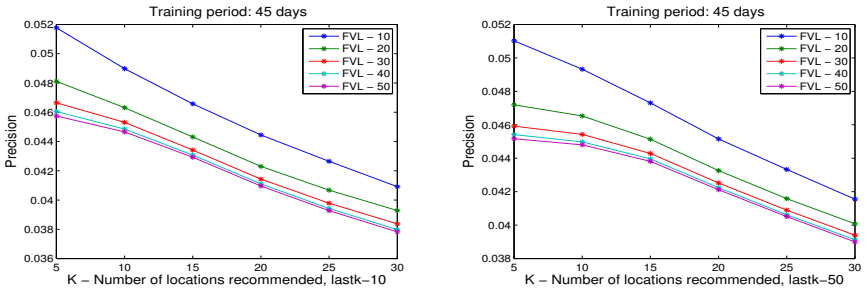
In Figures 14, 15 we show the performance of REGULA for different FVLs and a fixed bounding box of size $5km$. We observe that there is no difference in performance for different FVLs because the $5km$ bounding box captures all the potential candidate locations for all users. Therefore, whether we consider 10 or 50 FVLs of a user the total number of candidate locations to rank is same, as the large $5km$ bounding box already includes all unvisited locations. When K is increased from 5 to 15 the performance of REGULA increases because it can correctly rank the top-K locations. Since each user in our CDR-based LBSN dataset has visited on an average 15 locations, the theoretically best



(a) (b)

Figure 10: Performance of REGULA on CDR-based LBSN for different FVLs, α=10, Box=1km, and 30 days for training.

performance will be at $K = 15$, where the set of recommended locations and visited locations could be same. Therefore we observe the best performance of `REGULA` when K is 15. Beyond 15, the performance of `REGULA` drops because the intersection between the set of recommended locations and visited locations is maximum at 15, while the increase in K or number of recommended locations negatively impacts the performance.

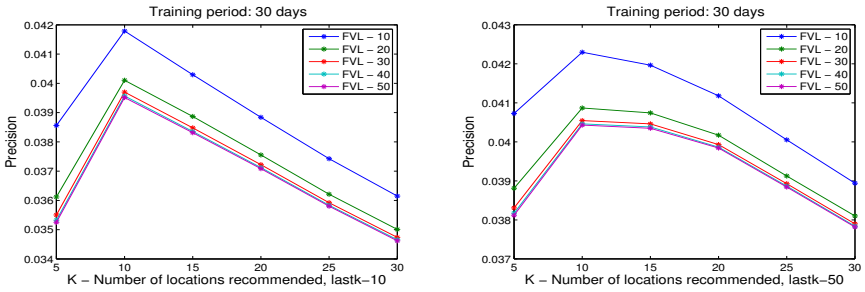


(a) (b)

Figure 11: Performance of REGULA on CDR-based LBSN for different FVLs, α=10, Box=1km, and 45 days for training.

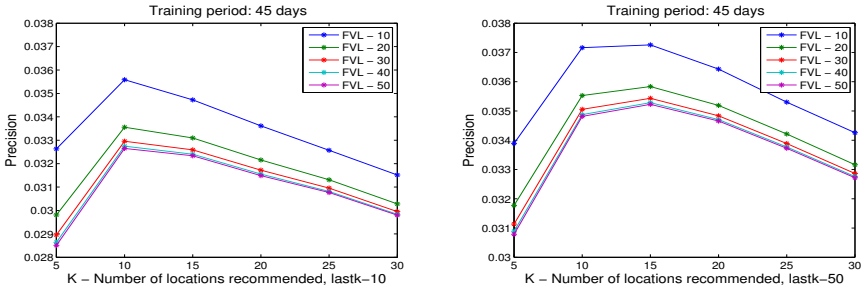
5.1.2. Impact of Bounding Box on Precision

We highlight here the impact of distance that is reflected by the bounding box parameter ($bbox$) already seen in the previous subsection by fixing $FVL = 10$, $\alpha = 10$ and $lastk = 10$, which are the values for which we observed best results. Figure 16 shows how the performance of `REGULA` varies with the size of the bounding box for experiments performed on 30th and 45th day. As the size of the bounding box ($bbox$) is increased from $1km$ to $5km$ the total number of candidate locations within the bounding box increases. It is also evident from



(a) (b)

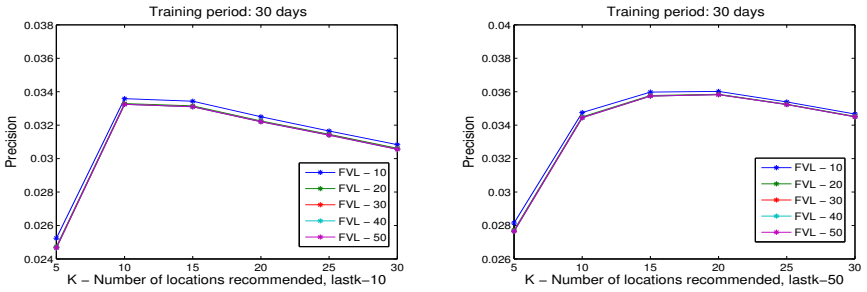
Figure 12: Performance of REGULA on CDR-based LBSN for different FVLs, α=10, Box=2.5km, and 30 days for training.



(a) (b)

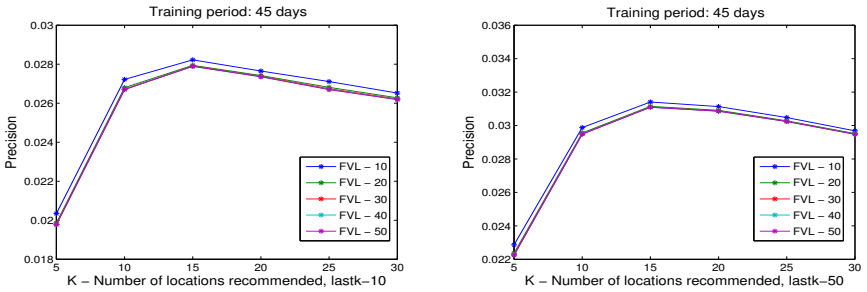
Figure 13: Performance of REGULA on CDR-based LBSN for different FVLs, α=10, Box=2.5km, and 45 days for training.

our analysis (refer to figure 3) related to Vicinity hypothesis (H2) that shows that as the size of the bounding box is increased the number of users with a new location around their FVLs increases. An increase in the size of the bounding box also increases the percentage of unrelated locations in the candidate set that are very far from the places regularly visited by the user. Locations that are far away from the FVLs might be closer to the *lastk* recently visited locations that lead to a rise in the distance score (Equation 2) and an increase in the overall recommendation score (Equation 4) assigned to it. An increase in the bounding box increases the noise in the overall ranking of candidate locations. Thus, a smaller bounding box will lead to a reasonable number of candidate locations that can be ranked efficiently.



(a) (b)

Figure 14: Performance of REGULA on CDR-based LBSN for different FVLs, α=10, Box=5km, and 30 days for training.



(a) (b)

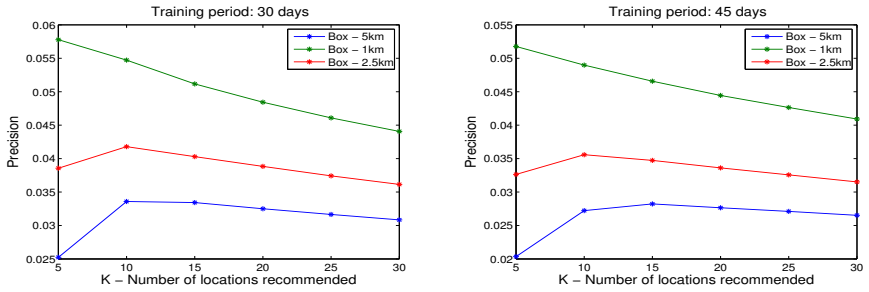
Figure 15: Performance of REGULA on CDR-based LBSN for different FVLs, α=10, Box=5km, and 45 days for training.

5.1.3. Impact of Last Visited Locations lastk on Precision

From Figures 10-15, we observe that `REGULA` performs best when we consider a small number of recently visited locations *lastk*. Increase in the parameter *lastk* adversely impacts the performance because locations visited in the distant past reduces the impact of locations that were visited most recently. Therefore the performance of `REGULA` reduces as *lastk* is increased from 10 to 50.

5.1.4. Impact of Social Ties (α) on Precision

We performed multiple experiments on 30th and 45th day to study the impact of friendship ties by varying α. We perform a grid search of α over $\alpha \in \{10, 50, 100, 200\}$ and found that the performance of `REGULA` does not vary meaningfully. Therefore we conclude that the variable α used to compute the impact of friendship score on the recommendation does not have a significant influence on the overall performance. The poor impact of the friendship relations on the recommendation of new locations in a CDR-based LBSN w.r.t. LBSNs may be the consequence of the different nature of the two social networks. Since



(a) (b)

Figure 16: Performance of REGULA on CDR-based LBSN for different size of Bounding Box, α=10, FVL=10, lastk=10, and 30/45 days for training.

typical LBSNs are interest- and social-driven platforms, the check-in activities are strongly influenced by friends' behaviors. A factor which is further emphasized by the gamification mechanisms these platforms implement and promote. On the other hand, the social interactions captured by CDR data express different purposes, from a common interest to formal or essential communications, so just a few relationships may influence the mobility patterns of an individual.

Finally, based on the performance results presented in this section, we observe that the precision of `REGULA` in CDR-based LBSNs, which ranges between $4e^{-2}$ to $6e^{-2}$ ($k = 10$), is comparable with traditional recommendation in LBSNs. As presented in [11], precision ranges between $1e^{-3}$ to $2e^{-2}$ ($k = 10$) for large LBSN datasets. Thus, we can conclude that, in addition to sharing similar characteristics with traditional LBSNs, CDR-based LBSNs can effectively be used for recommendation services.

5.2. Comparison of REGULA with reference recommendation models

To show the importance of the five observations in Section3, we compare the performance of `REGULA`, which is based on them, with the following standard factorized based recommendation models:

- LibFM [26] is a factorization based model that estimates interactions between users and locations by means of a product of vectors, whose factors derived from the user-location visits. The users and the locations vectors lie in the same latent space. The product between a user and a location vectors represent the preference of such user for this location. In our evaluation, we perform a grid search of factor size over $Factors \in \{16, 32, 64, 128, 256\}$ to select the best parameters.
- GeoMF [27] is another factorization based model that augments the user's and location's latent factors to incorporate the spatial constraints. Matrix Factorization (MF) based models assume that each user and location can be mapped to joint latent factor space and the preference of a user for a location can be approximated by their dot product in the latent

factor space. In GeoMF the authors augment the latent factors of user with latent factors of locations visited by them and augment the latent factors of locations with latent factors of influential locations. REGULA incorporates geographical constraint using the distance score. Further, GeoMF does not incorporate time information in the matrix factorization for location recommendation, while REGULA incorporates temporal information in all the three scoring functions. In our evaluation, we perform a grid search of factor size over *Factors* $\in \{16, 32, 64\}$ to select the best parameters. We restrict the factor size to 64 due to memory constraints in the execution of GeoMF.

5.2.1. LibFM model

Figure 17 shows the performance of LibFM for the different number of recommended locations K. Based on our experiments we observe that LibFM assigns very similar scores to all the candidate locations of a user. Since there are only 900 locations in our CDR-based LBSN dataset, LibFM is unable to utilize the limited number of interactions between user and locations in training data, to correctly score the candidate locations in testing data. As K is varied from 5 to 30 the probability for a visited location to be part of the top k increases that leads to an increase in precision. However, the performance of LibFM is still very low than compared to `REGULA`. Based on our additional experiments we observe that precision of LibFM reduces beyond $K = 50$.

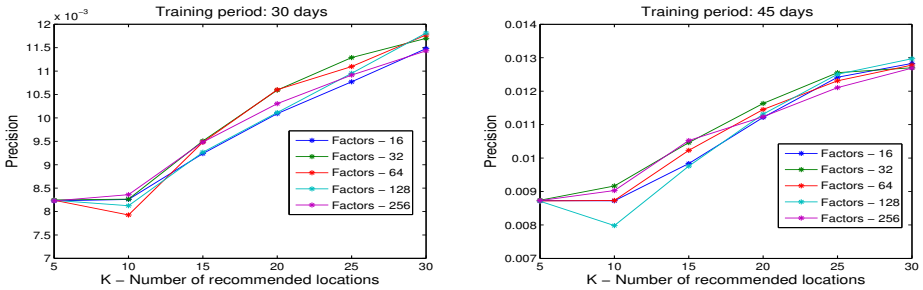


Figure 17: Precision of LibFM on CDR-based LBSN.

5.2.2. GeoMF Model

In Figure 18, we show the performance of GeoMF for the different number of recommended locations K. Similar to standard recommendation algorithms that rank all unvisited locations in the training data of a user, as K increases the precision reduces because of the increased in difficulty to correctly rank the top K locations. As K is increased from 5 to 30, we do not observe a proportional increase in the number of visited locations that are part of the top K recommendations. We find that GeoMF performs best when a latent factor of size 32 models every user and location.

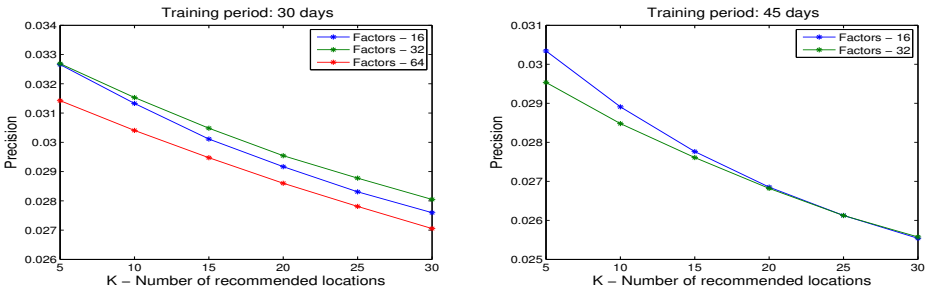


Figure 18: Precision of GeoMF on CDR-based LBSN.

5.2.3. Comparison of algorithms in CDR-based LBSNs

In Figures 19 and 20, we compare the performance of `REGULA`, LibFM, and GeoMF for experiments performed on 30th and 45th day. For true comparison, we select the best parameters for each algorithm i.e., for REGULA $\{bbox = 1km$, $FVL = 10$ and $lastk = 10\}$; for GeoMF $\{Factors = 32\}$; for LibFM $\{Factors = 128\}$. When we recommend 5 locations to each user, i.e., $K = 5$, `REGULA` performs six times better than LibFM and two times better than GeoMF. When K is increased from 5 to 30, `REGULA` still outperforms GeoMF, but we observe a reduction in the performance gap because $1km$ bounding box of `REGULA` does not capture all potential candidate locations.

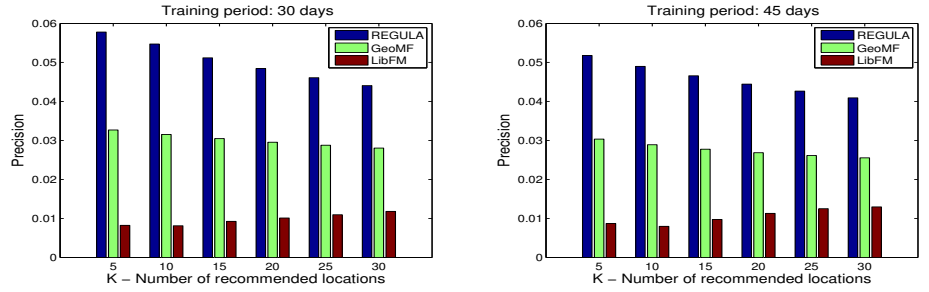


Figure 19: Comparison between the precision of REGULA, LibFM and GeoMF for their best parameters i.e., for `REGULA` $bbox = 1$km, $FVL = 10$ and $lastk = 10$; for GeoMF $Factors = 32$; for LibFM $Factors = 128$

5.3. Discussion

In this Section, we aimed to confirm the validity of using CDR data for timely and personalized services, and to quantify the importance of the five observations on human behaviour for this task. We used as example of such services the recommendation service. In order to provide a complete understanding of the impact of each observation, we have measured the impact of different independent parameters of our `REGULA` model on the task of recommending new locations to users of a CDR-based LBSN. We measured the per-

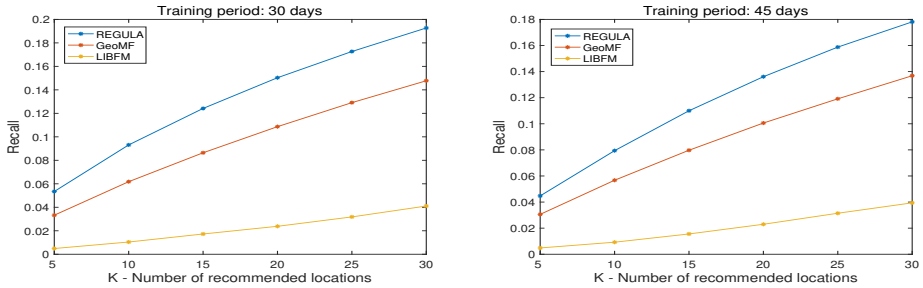


Figure 20: Comparison between the recall of REGULA, LibFM and GeoMF for their best parameters i.e., for `REGULA` $bbox = 1$km, $FVL = 10$ and $lastk = 10$; for GeoMF $Factors = 32$; for LibFM $Factors = 128$

formance of `REGULA` based on two standard metrics precision and recall. This demonstrated the effectiveness of using CDR-based LBSN for such timely and personalized service, and gave a qualitative measure of the benefit of basing the strategy on the five observations on human behaviour. Further, in order to give a quantitative measure of this benefit, we compare `REGULA` with reference recommendation algorithms, i.e., LIBFM and GeoMF and we show that `REGULA` outperforms them.

6. Related Work

To the best of our knowledge, this is the first work addressing timely and personalized services for Telecom operators. Some state-of-the-art, given the service used to validate our concept, can be given for recommendation services or for place recommendation.

Recommendation in LBSNs The most popular recommender systems utilize variants of Collaborative Filtering (CF) techniques to recommend places and two most commonly used filtering methods are Location-Based and User-Based Collaborative filtering [28]. However, these collaborative filtering algorithms do not exploit complete information of users in an LBSN as they do not take into account social network based information. Specifically, a location recommender system for LBSN can utilize both the location history of all users and their social ties to provide more accurate and quicker recommendations. It will also help businesses to identify their potential customers and provide incentives based on their personal and social interests. Some recent works started to develop systems for traditional LBNSs taking into account also some behavioral and social aspects of people. LORE [29] focuses on exploiting sequential movement pattern of users to provide better recommendations in LBSN. The work done by Hao et al. [22] utilizes friendship and distance to a new location to provide recommendations. However, these works do not take into account temporal importance of recommended locations, i.e., they do not distinguish between old and new (in the time domain) popular locations. In [30] authors present a time-aware recommender system but do not take into account friendship ties to recommended location.

Factorization based models are found to perform better than collaborative filtering algorithms[31]. These models have become popular after outperforming filtering based methods in the NetFlix competition [32]. LibFM [26] is a standard factorization based model that characterizes users and locations by a vector of factors inferred from the user-location visits. The vectors are in the same latent space and, the preference of a user for a location is modeled as an inner product in that space. GeoMF [27] is another factorization based model that augments the user's and location's latent factors to incorporate spatial constraints because people mobility patterns tend to cluster around specific locations.

Recently, some proposed works, [33] and [34], utilize different Deep Neural Networks to provide recommendations of movies/items on web portals. The work in [35] provided a first example of location recommendation with Deep Neural Networks.

Place recommendation in CDR datasets. While all these algorithms have been applied to tradition LBSNs to recommend new venues or locations, to the best of our knowledge CDRs data have never been used to cope with this task. In fact, studies on forecasting the users' mobility through CDRs data have focused on the prediction of the next cell/location a user will visit to handle and manage more efficiently the cellular network infrastructure. From the seminal work by Song et al. [36] on the predictability of human mobility extracted from a CDR dataset, a plethora of methods have been proposed to solve the next location prediction task. Most of them are based on the sequential pattern of the location visits of a single user, commonly modelled by a k-order Markov chain [37, 38]; whereas other predictors include the location histories of other users to solve the cold-start problem [39, 40, 41, 42]. Although some of the previous methods can predict unvisited locations, they are mainly adopted in the prediction of the cell a user will visit next or in the next time interval.

Human mobility and social behaviors from CDR data. In the last ten years the availability of a few CDR datasets resulted into an abundant literature on human mobility and sociality of large populations. Here we focus only on results which confirm and validate the five traits we introduced in Section 3. We refer readers to more organic and exhaustive literature reviews covering both mobile phone data analysis [16] and methodological and technological aspects [13].

The people's propensity of frequently visiting a small numbers of locations and exploring places close to them has been confirmed by many studies on the regularity of human mobility. For instance, Song et al. [36] have shown that the characteristics of human mobility can be reproduced by a model mixing the frequent visits of a pool of locations and the exploration of new places. Csáji et al. [43] have identified the same trait in a CDR dataset in Portugal and found that the average number of frequently visited locations is 2, as confirmed in Papandrea et al.'s work [3]. Finally, Bagrow et al. [44] have introduced the idea of 'habitats' to capture daily and weekly mobility regularity based on the frequency of visit of locations.

The strict interplay between people interacting by mobile devices and their geographical proximity - which results into our fourth hyphotesis - has been the subject of a few recent studies. For instance, the work of Phithakkitnukoon et al. [45] has shown that most of the places a person visits are close to their friends locations, while Calabrese et al. [46] and Wang et al. [47] that the frequency of between users is highly correlated with their frequency of calls and proposed a few mobility features to improve the performance of link recommendation algorithms. Here we explore the opposite direction, in fact we exploit the sociality of the operator customers to recommend new places to visit, i.e. we use some social aspects to influence people's mobility.

7. Discussion

In Section 5, we have demonstrated the plausibility of using CDR data for timely and personalized services, and we have also shown the better performance of the `REGULA` algorithm that is based on the five observations on human behaviour described in Section 3. This opens for a novel and quite unexplored field for Telecom operator services. Two important factors make this field very promising:

- Privacy : CDR-based services are much less privacy invasive than GPS-based ones. Currently, with the new GDPR, for the latter it will be very difficult to be compliant with them.
- Competitive Costs: The costs associated to gathering data to build CDR-based LBSNs is very low. In fact such data are already automatically gathered with the traditional communication services provided by Telecom operators.

Finally, we also notice that such data would be complemented and reinforced, once a given service is started and running, by the service-specific data. The service specific data and CDR data would provide strategic advantage to Telecom operator that offers such a given service. This is particular important to build the friendship graph, as we described in Section 2.2. In this work, as we were using 2012 traces, we had the possibility to use SMS, in addition to calls, for this task. However, SMS is no longer very used. But, as indicated above, this problem will be overcome by service-specific data.

8. Conclusion and Future Work

CDR data is the most available and representative set of information on human behavior that include both physical and social data. CDR is one of the most secure and regulated data that is also legally accessible to Telecom operators as they ensure that the data storage and analysis is compliant to the strict government regulations as Telecom industry already comply with them. Further, this data is available to Telecom operators because most of the privacy and security regulations are already included in the operators' contracts.

At the same time, Telecom operator's business model is rapidly moving towards providing timely and personalized services. We anticipate a challenging research direction on such services for Telecom operators, built upon a methodology that uses CDR as traditional LBSNs. To support this promising approach, we first demonstrated that we could build an LBSN from CDR: the *CDR-based LBSN*. We then proved that CDR data could be efficiently used for such timely and personalized services: we have used some state-of-the-art algorithms on our *CDR-based LBSN* and shown that we can recommend a new location with similar performances to standard LBSNs.

We also showed that such timely and personalized services are better built considering some specific characteristics of human behavior, and we demonstrated the superior performance of this approach. Among recommendation algorithms employed, our `REGULA` algorithm has shown much better performances due to its capacity of better fostering the social and the proximity knowledge. As we have illustrated in Section 3.1, the mobility patterns of people, and consequently also LBSNs and *CDR-based LBSN*, present the characteristics of Regularity, Vicinity, Recency, Social Strength and Inertia. `REGULA` has been built to take advantage of such characteristics, and the benefits are evident as `REGULA` can recommend better locations than other state-of-the-art algorithms.

We predict that timely and personalized services from Telecom operators are going to be the next frontier of this decade and our work opens a novel and very rich research direction.

9. Acknowledgement

This work was supported by the Swiss National Science Foundation via the SwissSenseSynergy project, grant number 154458.

References

[1] Number of mobile phone users worldwide from 2015 to 2020 (in billions). https://www.statista.com/statistics/274774/forecast-of-mobile-phone-users-worldwide/.

[2] Apple shifts focus to services business. https://www.ft.com/content/68e80a44-9b28-11e6-b8c6-568a43813464.

[3] Michela Papandrea, Karim Keramat Jahromi, Matteo Zignani, Sabrina Gaito, Silvia Giordano, and Gian Paolo Rossi. On the properties of human mobility. *Computer Communications*, 87:19–36, 2016.

[4] M. Zignani, C. Quadri, S. Bernardinello, S. Gaito, and G.P. Rossi. Calling and texting: Social interactions in a multidimensional telecom graph. In *Proceedings of 10th International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2014*, 2015.

[5] I. H. Sarker and F. D. Salim. *Mining User Behavioral Rules from Smartphone Data through Association Analysis*. PAKDD, Springer, 2018.

[6] Zolzaya Dashdorj and Stanislav Sobolevsky. Characterization of behavioral patterns exploiting description of geographical areas. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXVII*, pages 159–176. Springer, 2016.

[7] C. Quadri, M. Zignani, S. Gaito, and G. P. Rossi. On non-routine places in urban human mobility. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 584–593, Oct 2018. doi: 10.1109/DSAA.2018.00075.

[8] Francesco Calabrese, Francisco C Pereira, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. The geography of taste: analyzing cell-phone mobility and social events. In *International conference on pervasive computing*, pages 22–37. Springer, 2010.

[9] Foursquare. http://foursquare.com.

[10] Blessing in disguise. http://www.tnsglobal.com/intelligence-applied/blessing-in-disguise.

[11] Steven Mudda and Silvia Giordano. Regula: Utilizing the regularity of human mobility for location recommendation. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on GeoStreaming*, IWGS '15, pages 69–77, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3971-1. doi: 10.1145/2833165.2833172. URL `http://doi.acm.org/10.1145/2833165.2833172`.

[12] M. Papandrea, M. Zignani, S. Gaito, S. Giordano, and G.P. Rossi. How many places do you visit a day? In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on*, pages 218–223, March 2013. doi: 10.1109/PerComW.2013.6529485.

[13] Diala Naboulsi, Marco Fiore, Stephane Ribot, and Razvan Stanica. Large-scale mobile traffic analysis: a survey. *IEEE Communications Surveys & Tutorials*, 18(1):124–161, 2015.

[14] Kamini Garg, Silvia Giordano, and Mehdi Jazayeri. Indigo: Interest-driven data dissemination framework for mobile networks. In *Proceedings of the 20th ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems*, MSWiM '17, pages 269–277. ACM, 2017.

[15] Renaud Lambiotte, Vincent D Blondel, Cristobald De Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, 2008.

[16] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1):10, 2015.

[17] Marton Karsai, Kimmo Kaski, Albert-László Barabási, and János Kertész. Universal features of correlated bursty behaviour. *Scientific Reports*, 2, 2012.

[18] Ming-Xia Li, Wen-Jie Xie, Zhi-Qiang Jiang, and Wei-Xing Zhou. Com-

munication cliques in mobile phone calling networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(11):P11007, 2015.

[19] Wei Wei, Xiaojun Zhu, and Qun Li. Lbsnsim: Analyzing and modeling location-based social networks. In *INFOCOM, 2014 Proceedings IEEE*, pages 1680–1688. IEEE, 2014.

[20] Ana Nika, Asad Ismail, Ben Y Zhao, Sabrina Gaito, Gian Paolo Rossi, and Haitao Zheng. Understanding and predicting data hotspots in cellular networks. *Mobile Networks and Applications*, pages 1–12, 2016.

[21] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[22] Hao Wang, Manolis Terrovitis, and Nikos Mamoulis. Location recommendation in location-based social networks using user check-in data. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL'13, pages 374–383, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2521-9. doi: 10.1145/2525314.2525357. URL `http://doi.acm.org/10.1145/2525314.2525357`.

[23] Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. Time-aware point-of-interest recommendation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 363–372, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2034-4. doi: 10.1145/2484028.2484030. URL `http://doi.acm.org/10.1145/2484028.2484030`.

[24] Jinhee Kim, Soora Rasouli, and Harry J.P. Timmermans. Social networks, social influence and activity-travel behaviour: a review of models and empirical evidence. *Transport Reviews*, 38 (4):499–523, 2018. doi: 10.1080/01441647.2017.1351500. URL `https://doi.org/10.1080/01441647.2017.1351500`.

[25] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1082–1090, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0813-7. doi: 10.1145/2020408.2020579. URL `http://doi.acm.org/10.1145/2020408.2020579`.

[26] Steffen Rendle. Factorization machines with libFM. *ACM Trans. Intell. Syst. Technol.*, 3(3):57:1–57:22, May 2012. ISSN 2157-6904.

[27] Defu Lian, Cong Zhao, Xing Xie, Guangzhong Sun, Enhong Chen, and Yong Rui. Geomf: Joint geographical modeling and matrix factorization for point-of-interest recommendation. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 831–840, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2956-9. doi: 10.1145/2623330.2623638. URL

http://doi.acm.org/10.1145/2623330.2623638.

[28] Jie Bao, Yu Zheng, David Wilkie, and Mohamed F Mokbel. A survey on recommendations in location-based social networks. *ACM Transaction on Intelligent Systems and Technology*, 2013.

[29] Jia-Dong Zhang, Chi-Yin Chow, and Yanhua Li. Lore: Exploiting sequential influence for location recommendations. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 103–112. ACM, 2014.

[30] Quan Yuan, Gao Cong, and Aixin Sun. Graph-based point-of-interest recommendation with geographical and temporal influences. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 659–668, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2598-1. doi: 10.1145/2661829.2661983. URL http://doi.acm.org/10.1145/2661829.2661983.

[31] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. Recommender systems survey. *Knowledge-based systems*, 46: 109–132, 2013.

[32] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, Aug 2009. ISSN 0018-9162. doi: 10.1109/MC.2009.263.

[33] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, pages 173–182. International World Wide Web Conferences Steering Committee, 2017.

[34] Jeroen BP Vuurens, Martha Larson, and Arjen P de Vries. Exploring deep space: Learning personalized ranking in a semantic space. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, pages 23–28. ACM, 2016.

[35] Steven Mudda, Defu Lian, Silvia Giordano, Danyang Liu, and Xing Xie. Spatial-aware deep recommender system. In *Proceedings of Ubiquitous Intelligence and Computing*. IEEE SmartWorld, 2018.

[36] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[37] Xin Lu, Erik Wetter, Nita Bharti, Andrew J Tatem, and Linus Bengtsson. Approaching the limit of predictability in human mobility. *Scientific reports*, 3:srep02923, 2013.

[38] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Next place prediction using mobility markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, page 3. ACM, 2012.

[39] Wesley Mathew, Ruben Raposo, and Bruno Martins. Predicting future locations with hidden markov models. In *Proceedings of the 2012 ACM*

Conference on Ubiquitous Computing, pages 911–918. ACM, 2012.

[40] Francesco Calabrese, Giusy Di Lorenzo, and Carlo Ratti. Human mobility prediction based on individual and collective geographical preferences. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 312–317. IEEE, 2010.

[41] Haoyi Xiong, Daqing Zhang, Daqiang Zhang, and Vincent Gauthier. Predicting mobile phone user locations by exploiting collective behavioral patterns. In *Ubiquitous Intelligence & Computing and 9th international Conference on Autonomic & Trusted Computing (UIC/ATC), 2012 9th international conference on*, pages 164–171. IEEE, 2012.

[42] Fahad Alhasoun, May Alhazzani, Faisal Aleissa, Riyadh Alnasser, and Marta González. City scale next place prediction from sparse data through similar strangers. In *Proceedings of ACM KDD Workshop, Halifax, Canada*, 2017.

[43] Balázs Cs. Csáji, Arnaud Browet, V.A. Traag, Jean-Charles Delvenne, Etienne Huens, Paul Van Dooren, Zbigniew Smoreda, and Vincent D. Blondel. Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, 392(6):1459 – 1473, 2013. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2012.11.040. URL `http://www.sciencedirect.com/science/article/pii/S0378437112010059`.

[44] James P Bagrow and Yu-Ru Lin. Mesoscopic structure and social aspects of human mobility. *PloS one*, 7(5):e37676, 2012.

[45] Santi Phithakkitnukoon, Zbigniew Smoreda, and Patrick Olivier. Socio-geography of human mobility: A study using longitudinal mobile phone data. *PloS one*, 7(6):e39253, 2012.

[46] Francesco Calabrese, Zbigniew Smoreda, Vincent D Blondel, and Carlo Ratti. Interplay between telecommunications and face-to-face interactions: A study using mobile phone data. *PloS one*, 6(7):e20814, 2011.

[47] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1100–1108, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0813-7. doi: 10.1145/2020408.2020581. URL `http://doi.acm.org/10.1145/2020408.2020581`.

Appendix

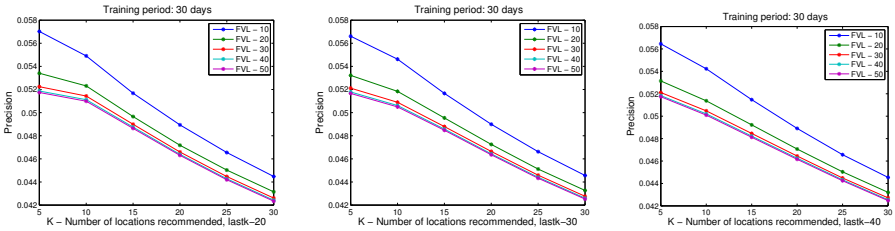


Figure 21: Performance of REGULA on CDR-based LBSN for different FVLs, α=10, Box=1km, and 30 days for training.

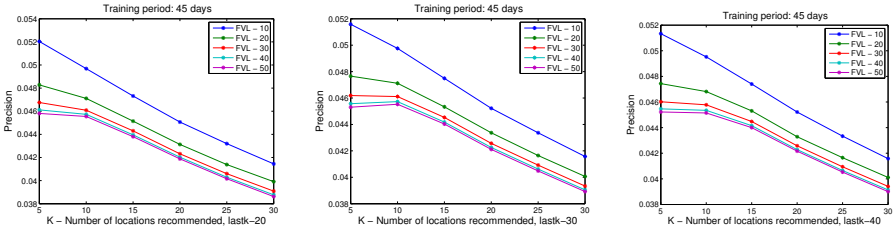


Figure 22: Performance of REGULA on CDR-based LBSN for different FVLs, α=10, Box=1km, and 45 days for training.

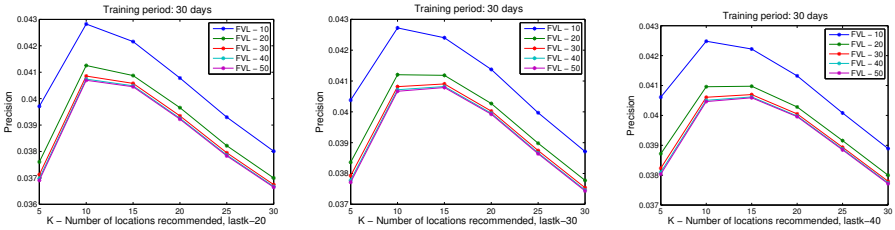


Figure 23: Performance of REGULA on CDR-based LBSN for different FVLs, α=10, Box=2.5km, and 30 days for training.

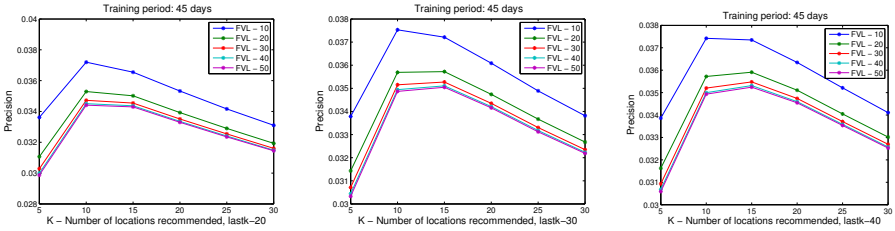


Figure 24: Performance of REGULA on CDR-based LBSN for different FVLs, α=10, Box=2.5km, and 45 days for training.

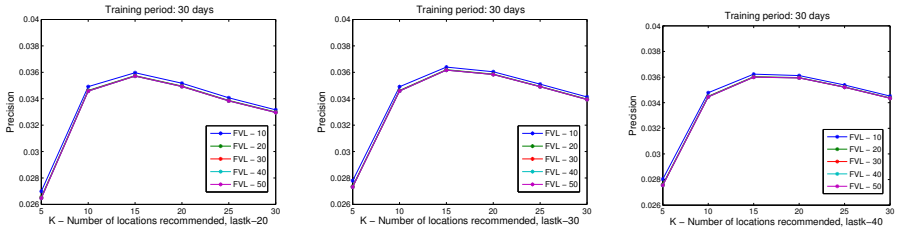


Figure 25: Performance of REGULA on CDR-based LBSN for different FVLs, α=10, Box=5km, and 30 days for training.

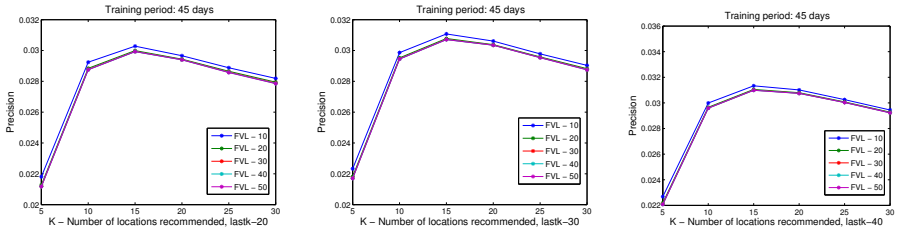


Figure 26: Performance of REGULA on CDR-based LBSN for different FVLs, α=10, Box=5km, and 45 days for training.