Understanding User Attributes from Calling Behavior: Exploring Call Detail Records through Field Observations

Ayumi Arai
The University of Tokyo
Tokyo, Japan

Apichon Witayangkurn
Asian Institute of Technology
Pathumthani, Thailand

Hiroshi Kanasugi
The University of Tokyo
Tokyo, Japan

Teerayut Horanont
Thammasat University
Pathumthani, Thailand

Xiaowei Shao
The University of Tokyo
Tokyo, Japan

Ryosuke Shibasaki
The University of Tokyo
Tokyo, Japan

ABSTRACT

Mobile phones are arguably one of the most prolific sources of large-scale human mobility data. The availability of this data has generated a massive body of research focused on understanding the dynamics and patterns of human mobility. However, it is increasingly evident that additional value can be derived from such data. This paper proposes a novel approach for understanding the attributes of mobile users by analyzing calling behavior derived from field survey data, in combination with call detail records (CDRs). Our survey reveals distinctive traits in calling behavior that correspond to user attributes. Analysis results demonstrate that frequent call locations, the variability in call time distributions, and the locations from which calls are made around midday are all keys to distinguishing gender. In addition, the location of calls initiated during the morning hours is a key to analyzing income levels for males.

Categories and Subject Descriptors

I.5.m: Pattern Recognition – *Miscellaneous*;

General Terms

Experimentation, Measurement, Human Factors

Keywords

Call Detail Records, Calling Behavior, Demographic attributes

1. INTRODUCTION

With the rapid spread of mobile phones, analyses of large-scale data such as GPS logs and call detail records (CDRs) have provided detailed descriptions of human mobility. An increasing body of human mobility research is focused on modeling the properties of human mobility patterns, where they are generalized in a quantitative manner. It is also widely agreed that people routinely visit specific locations [16,17], and these are typically home and work places [5]. Various studies have proposed algorithms to estimate the home and work locations of mobile phone users by examining the time and location distribution of call records, which is accomplished by analyzing anonymized

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

MoMM '14, December 08 - 10 2014, Kaohsiung, Taiwan
Copyright 2014 ACM 978-1-4503-3008-4/14/12...\$15.00
<http://dx.doi.org/10.1145/2684103.2684107>

CDRs [6,7]. Research based on GPS log location histories measures users' similarities by analyzing the sequential properties of their trajectories, and the hierarchical properties of their location histories. This body of research argues that users with similar location histories also share similar interests and preferences [10]. Although these studies succeed in mining mobility patterns, the outcomes of large-scale data analyses describe the movement of crowds, because the data is anonymized.

In addition to such large-scale datasets, there are other types of data collected through conventional methods. In the field of urban planning and transportation, many empirical studies have attempted to identify the factors that affect human activities and travel patterns, by analyzing data collected through questionnaire surveys. Such data typically consist of socio-demographic attributes, transportation means and origins, destinations, and purposes of movement, all with associated time stamps. Family and social obligations are often presented as significant factors affecting daily travel and activity behavior [12]. This implies that people's activity patterns are constrained by social ties. This observation is consistent with most previous analysis of human mobility patterns from large-scale datasets, where people routinely visit a limited number of locations [16,19]. Though such conventional approaches may be less efficient in defining human activity and travel patterns in terms of population size and data period length, it can link activity and travel patterns with people's attributes to some extent [15]. Conventional survey data enables common key patterns to be identified from large-scale data, and allows us to understand the hidden properties of anonymized large-scale data. To further investigate the properties of user attributes in anonymized data, it is critical to analyze this large-scale data in combination with such secondary data [11].

In fact, emerging studies are attempting to analyze data derived from mobile phones in combination with secondary data. [3] observed that the ratio of shared phone usage and call type, such as incoming and outbound calls, shows significant differences by gender, on average. In addition, the number of usages during a specified length of time was shown to differ according to income level. This observation is consistent with another research project, which analyzed social connectivity through social networking activities [1]. Utilizing sensor data from volunteer mobile users, [13] proposed prediction models based on user demographic attributes. While these models focus on features derived from smart phones such as acceleration and application usage, they provide features that can be derived from CDRs. For instance, the probability of being at home or work at night is useful for predicting the occupation type. Furthermore, the number of places visited in the evening is useful for predicting marital status.

In this work, we provide a novel approach for extracting features from calling behavior, which can be constructed from anonymized CDRs. This technique can reveal distinctive traits that help identify the demographic attributes of mobile phone users. This study is unique, because we focus on extracting lifestyle traits and routines to identify user attributes. We statistically analyze data collected through a field survey that focused on the demographic attributes, calling behavior, and weekly activity patterns of mobile users. Several key features are empirically derived to analyze user attributes, which are compared with another set of the features generated from CDRs.

Contributions of our work are described below:

- Statistical analysis results are provided for calling behavior based on field survey data. We extract calling behavior traits to differentiate gender and occupation types that correspond to the patterns of routine activities. To do so, we introduce the concept of weekly activity patterns, which specifies whether the day's call records correspond to a day where the user is engaged in their primary routine.
- By comparing the analysis results of the field survey data, which include single-day call records of 922 users, and those of two-month CDRs, we describe the advantages and limitations of analyzing call records from the field survey data.

The remainder of this paper is organized as follows. Section 2 describes the data used for this study. Section 3 explains the mobile user lifestyle and calling behavior prototypes that were used for analyzing the survey data. Section 4 compares the statistical analysis results of the survey data with those based on the CDRs for 58 volunteers, followed by conclusions and discussions for further studies.

2. DATA

In this study, we use two data sources: the aggregated single-day call records of 922 mobile phone users from a field survey, and two-month CDRs for 58 volunteers. A unique characteristic of the survey data is that it includes actual call records from mobile phone users, as single-day records from 922 handsets. The records specify the call's location type and other basic attributes. The data contain neither mobile phone numbers nor any other information explicitly specifying individuals.

2.1 Field Survey Data

To understand the hidden properties of CDRs, we conducted a field survey, named the Survey on Patterns of Activity for Comprehensive Explorations of Mobile Phone Users in Dhaka (SPACE). The purpose of SPACE is to understand the calling behavior, characteristics, and lifestyles of mobile phone users. Our survey site was Greater Dhaka, which is composed of Dhaka City Cooperation (DCC), surrounding municipal areas (Savar and Karaniganj Upazila), and regions outside of the urban area (OUA, including portions of the Narayanganj, Gazipur, and Narsingdi Districts). SPACE was conducted from November 27, 2013 through January 4, 2014. The survey interviewed 810 households, and included 922 mobile phone users.

To capture calling behavior along with the diverse characteristics of mobile phone users, we employed two-staged stratified sampling. We first segmented our survey site by Primary Sampling Units (PSUs) and classified them into three subgroups based on the type of dominant land use such as residential, commercial, and industrial. Among them, 15 PSUs were selected by considering population distributions. Subsequently, 18 households each were sampled from high, middle, and lower

income groups for each PSU. Because it is difficult to obtain household income data for sampling, we set the criteria for the three income groups based on building types, ownership, facility, and durable consumer goods.

Table 1. Basic attribute distribution of the mobile phone users in SPACE data

Gender	Male	Female
Users of a specified operator	60%	40%
Roles in the household	Male	Female
Head of Household	85%	9%
Spouse	0%	76%
Daughter/Son of head/spouse	12%	10%
Others	2%	5%
Age group	Male	Female
Under 20	3%	5%
20 – 29	16%	28%
30 – 39	31%	33%
40 – 49	26%	24%
50 – 59	16%	7%
60 and more	7%	0%
Primary activity = Occupation type	Male	Female
Working (not household tasks) = <i>Worker</i>	89%	18%
Household tasks = <i>Household</i>	1%	77%
Schooling = <i>Student</i>	4%	4%
Others = <i>Others</i>	6%	1%
Individual annual income	Male	Female
0 BDT ~	8%	80%
~ 48,000 BDT (620 USD)	3%	1%
~ 144,000 BDT (1,800USD)	23%	7%
~ 312,000 BDT (4,000 USD)	31%	7%
~ 576,000 BDT (7,400 USD)	14%	2%
~ 996,000 BDT (12,800 USD)	10%	2%
More than 996,000 BDT	10%	1%

Table 1 describes the basic features of mobile phone users for the leading telecommunications operators in the SPACE data. Eighty-five percent of males are categorized as the head of the household, and are generally considered to be financially responsible for the people living with them. Further, 76% of females are categorized as a spouse and 77% of them consider household tasks to be their primary activity. This implies that most female users are married and care for their family members at home. We assume the females are able to allocate some money at their disposal, even if they are not engaged in income-earning activities. Individual income levels among mobile phone users in the SPACE appear to be higher than average for Dhaka, where annual income per capita is 1,350 USD [4]. It partially infers that this operator's mobile phone tariff is relatively high. The primary activity is specified as *Worker*, *Household*, *Student*, or *Others*; hereinafter, this is referred to as the occupation type.

2.2 Call Detail Records

We used CDRs for 58 mobile users, who receive service from one of leading telecommunications operators in Bangladesh. Their basic attribute information is also collected. It included gender, age, weekly activity patterns, annual income, and call location types such as home and work places. Table 2 describes their basic features. CDRs consist of time, antenna locations, and duration of voice calls from November and December 2013.

Table 2. Basic attribute distribution for 58 mobile users

Gender	Male	Female
Users of a specified operator	47%	53%
Roles in the household	Male	Female
Head of Household	85%	13%
Spouse	0%	81%
Daughter/Son of head/spouse	7%	0%
Others	8%	6%
Age group	Male	Female
Under 20	11%	23%
20 – 29	19%	35%

30 – 39	33%	23%
40 – 49	15%	13%
50 – 59	15%	6%
60 and more	7%	0%
Primary activity = Occupation type	Male	Female
Working (not household tasks) = <i>Worker</i>	70%	13%
Household tasks = <i>Household</i>	0%	84%
Schooling = <i>Student</i>	4%	0%
Others = <i>Others</i>	26%	3%
Individual annual income	Male	Female
0 BDT ~	14%	83%
~ 48,000 BDT (620 USD)	0%	0%
~ 144,000 BDT (1,800 USD)	41%	7%
~ 312,000 BDT (4,000 USD)	22%	7%
~ 576,000 BDT (7,400 USD)	15%	3%
~ 996,000 BDT (12,800 USD)	4%	0%
More than 996,000 BDT	4%	0%

3. LIFESTYLE AND CALLING BEHAVIOR

To understand activity patterns, it is vital to identify a certain set of locations where people spend the majority of their time. Many studies identify two dominant locations for people as their home and work or school (hereinafter referred to as “*Home*” and “*Primary out-of-home location*”), which can explain a significant portion of their activity [16]. Identifying the time and location distribution of the users’ whereabouts enabled us to define their activity patterns and portions of their lifestyles. Therefore, in this section we examine the time distribution of calls initiated from *Home*, *Primary out-of-home location*, or *Others* by analyzing the calling behavior of SPACES’s mobile phone users. Because sequential location histories can infer similarities between people [10], we expect the distributions to vary according to differences in user attributes. We extracted calling behavior prototypes according to gender and occupation type, and examine how their traits differ.

For the analysis provided this section we introduce the concept of weekly activity patterns, which specifies whether mobile users are engaged in their primary routine on the day the call was made. Routines can be any activities users spend the majority of their time on during the day. The routine that the user follows on the highest number of days during the week is considered as their primary routine.

3.1 Behavioral Norms and Diverse Lifestyles

Dhaka is the capital city of Bangladesh; it is one of the emerging economies in South Asia, enjoying an average of 4.8% annual GDP growth per capita [22]. Its strong growth has accelerated the urbanization of Bangladesh, leading to a high concentration of economic and demographic growth in metropolitan areas. Bangladesh currently has very vibrant mobile phone markets, which include 112 million active subscribers in a total population of 155 million [2]. Considering the substantial and increasing subscription rate, the mobile phone is a useful platform for acquiring a general picture of human mobility in Dhaka.

Bangladeshi society has long promoted divisions in social space, and fostered differences in behavioral norms between males and females in various aspects of social, cultural, and religious traditions [18]. As a result, females have often assumed the role of family caregiver [20]. However, attitudes regarding their social roles have been changing, owing to advancements in educational opportunities that have occurred along with economic development. Consequently, females, particularly those in middle or upper-middle income households, have started to transform their traditional roles significantly [21]. As described in Table 1, mobile users in the SPACE data include a variety of population groups. Therefore, the SPACE data enable us to capture such

transformations while traditional behavioral norms continue to exist.

3.2 Weekly Activity Patterns

In this subsection, we describe the weekly activity patterns of the 922 mobile users recorded in the SPACE data. Weekly activity patterns are determined based on the number of weekly routines (primary and non-primary), and the number of days on which the primary routine is followed. By classifying the activity according to a pattern, we examine how activity patterns are linked to calling behavior in the following section. We assume that these patterns can also be partially extracted from anonymized CDRs, which allows us to estimate significant locations such as home and work places. Based on this technique, we consider it possible to reconstruct weekly activity patterns from CDRs, to which we can apply this concept.

Table 3 classifies the mobile users into four patterns, according to the following criteria. Pattern (1) captures those who have only one type of routine per week. That is, their weekly activities do not fundamentally change. Fifty-six percent of the users are classified into this pattern. As shown in Table 3, the number of routines is one, and the number of primary routine days is seven. The majority of users classified into this pattern tend to have activities outside of home such as commuting to an office, shopping at a market, or sending their children to school. We assume some of those locations will appear in call records as the *Primary out-of-home location*. The population in Pattern (1) consists of those who are primarily engaged in income-earning activity on a daily basis, and those who primarily perform household tasks and care for family members at home. All seven days of the week are considered to be their primary routine days, unless they have other routines such as social activities in their community on specific days, or small side businesses that occupy their spare time. In addition, approximately 2% of those in Pattern (1) do not have specific tasks, and therefore remain at home. Most of them are elderly, retired, or young children not enrolled in school. Although they primarily remain at home all day and are not engaged in any specific tasks, all days of the week are counted as their primary routine days because their daily activities do not change.

Table 3. Weekly activity patterns of 922 mobile users in SPACE data

Pattern	Description of people classified into the pattern	Number of routines	Number of days for the primary routine	%
(1)	Those who repeat their primary routine activity every day. Mainly composed of income earners and those who do household tasks.	1	7	56
(2)	Those who repeat their primary routine activity every day, except for Friday. Mainly composed of income earners or are students.	2	6	30
(3)	Those who repeat their primary routine activity every day except for Friday and Saturday. Mostly composed of public sector employees or students.	2	5	9
(4)	Mostly income earners and their activity patterns do not follow Pattern (1), (2), or (3).	1 or more	1 to 6	5

The population in Pattern (2) consists of users who have two routines per week, where six days (except for Friday) are devoted to their primary routine. The majority of users in this pattern are income earners or students, because it is common to have only

one non-working or non-school day in Dhaka, which is typically a Friday. As described in Table 3, the number of routines in Pattern (2) is two, and the number of primary routine days is six. Similarly, the population in Pattern (3) primarily consists of users who are income earners or students. However, those who are classified into this pattern devote five days, except for Friday and Saturday, to their primary routine. As listed in Table 3, only a small percentage of users are classified into Pattern (3). Few people in Dhaka have two non-working or non-schooling days. It is likely that many users in Pattern (3) work in the public sector, because the government sets Friday and Saturday as official holidays. The remaining users are classified into Pattern (4). Most of them are engaged in income earning activity and have both working and non-working days. Their primary routine days do not essentially follow Patterns (1), (2), or (3). For instance, some of them may be engaged in income-earning activity for five or six days per week and their non-working days are neither Friday nor Saturday.

Then, we examine how the four patterns are distributed within major occupation types. Figure 1 shows the distribution of the weekly activity patterns across the occupation types. We note that we do not provide the distribution of the occupation types for weekly activity patterns. This is because we sampled the same number of households from each income group, but the income group distribution among general mobile users was not confirmed. That is, the composition of the occupation type for each weekly activity pattern can be affected by the distribution of the income level, if the population composition differs by income level. As explained in Table 2, we categorize mobile users into four occupation types: *Worker* is composed of users whose primary activity is earning an income; *Household* is composed of users who perform household tasks; *Student* includes users who are enrolled in school; and *Other* is composed of users who do not fit into any of the aforementioned categories. As Figure 1 shows, there are certain patterns that are specific to occupation types. For instance, 40% of *Worker* is composed of Pattern (1) and an additional 40% is composed of Pattern (2). *Household* is primarily represented by Pattern (1); similarly, *Student* is primarily represented by Pattern (3). Assuming that the distribution of call locations can partially reflect the times and locations where people spend the majority of their day, the weekly activity patterns, derived from calling behavior, can be keys to understanding the activity of mobile users.

Figure 1. Distribution of weekly activity patterns of mobile users in SPACE data across occupation types

3.3 Calling Behavior by Gender

This subsection compares the calling behavior of males and females by analyzing the call records of the 922 mobile users. First, we examine calling behavior by analyzing call location trends. Figure 2 shows percentages for the number of calls at (a) *Home* and (b) *Primary out-of-home location* among the total number of calls on the primary routine day. As the figure shows, males and females exhibit distinctly different call location trends. Across all patterns, females tend to call predominantly from *Home* while males call from both the *Home* and *Primary out-of-home location*. This indicates that identifying dominant call locations on

the primary routine day is important for determining gender. This feature is particularly distinctive among those classified into Pattern (1), where the primary routine is repeated every day.

Figure 2. Percentage of calls from (a) *Home* and (b) *Primary out-of-home location* against total calls on the primary routine day by gender (a) Upper (b) Lower

Similarly, Figure 3 shows percentages of the number of calls from (a) *Home* and (b) *Primary out-of-home location* compared to the total number of calls on the non-primary routine day. In contrast to primary routine day trends, non-primary routine day trends for males and females are relatively similar. However, the figure shows large differences in the call location distributions of Patterns (2) and (3). This corroborates our assumption that the distribution of call locations, which partially reflects the time spent at the locations, accurately reflects differences in the activity patterns of different population groups. There is no observation for Pattern (1), because those classified into this pattern only have primary routine days. Further, we do not have an observation for Pattern (4) on non-primary routine days in SPACE data. As a result, single-day call records from users classified into Pattern (4) are utilized to represent all of their primary routine days.

Figure 3. Percentage of calls from (a) *Home* and (b) *Primary out-of-home location* against total calls on non-primary routine day by gender (a) Upper (b) Lower

We then compared the calling behavior of males and females in terms of time distribution and call length. Using five features, we attempted to capture calling behavior trends based on the frequency, time, and duration of calls for the primary and non-primary routine days. Figure 4 compares the calling behavior of the primary routine day and that of the non-primary routine day by gender. Each ratio r in the table of Figure 4 is obtained by solving the following equation for each feature:

$$r = \frac{\text{Value of a feature, which is calculated from calling records of non-primary routine days}}{\text{Value of a feature, which is calculated from calling records of primary routine days}}$$

Features (A), (B), and (C) capture calling behavior based on the call frequency and time of day, while (D) and (E) capture calling

behavior based on the duration of calls. For example, the value of feature (A) is the ratio of the number of calls on the primary routine day to the number of calls on the non-primary routine day. The greater the value deviates from unity, the more the users call on the non-primary routine day. The value of (A) for males, which is unity, indicates that the frequency of calls for males did not change between their primary and non-primary routine day. That is, the frequency of calls for males fundamentally did not change according to the type of day. In contrast, the value for females, 1.23, indicates females tended to call more frequently on their non-primary routine days. Based on the trend observed from feature (A), we conclude that the frequency of calls for males was consistent throughout a week, whereas the call frequency for females fluctuated, in that they tended to make a greater number of calls on the non-primary routine day.

Figure 4. Comparison of calling behavior on the primary routine day and non-primary routine day

Features (B) and (C) explain trends in the time distribution of calls for the primary and non-primary routine days. Here, the time of calls denotes the time when the user initiated communication with the call's recipient (Missed calls, which are not accepted by the recipient, are not included in CDRs.) Feature (B) is the average deviation of the call times calculated for each user and (C) is the average variance of those calls. The values of both features for males, which are close to unity, imply that calling behavior on the primary routine days and non-primary routine days were similar. That is, on average, if the user tended to call at a specific time on primary routine days, the user exhibited similar calling behavior on non-primary routine days, and vice versa. Conversely, the value of both features for females, 1.25 and 1.23 respectively, significantly deviate from unity, indicating that the call time distribution of the primary routine days differed from that of the non-primary routine days. Similar to the trend in the frequency of calls, the call time distribution trend indicates that calling behavior for males was similar throughout the week; but different for females, where greater variability between the primary routine days and non-primary routine days is observed in terms of call times.

The values for features (D) and (E) show trends in the duration of calls for the primary routine days and non-primary routine days. Feature (D) is the average duration per call per person. For example, the value of feature (D) for males is slightly smaller than unity. This indicates that the average call duration for males was slightly longer on primary routine days compared to non-primary routine days. As for females, average call durations for the primary and non-primary routine days were essentially equal. Contrary to the trends in the frequency and time distribution of calls, the average duration per call for females was similar throughout the week but different for males, because males tended to make longer calls on primary routine days. Feature (E) is the average of the total duration of calls per day per person. Similar to feature (D), the value of feature (E) for males is slightly smaller than unity. This implies that the total duration of calls per day for

males was slightly longer on primary routine days compared to non-primary routine days. Conversely, the value of (E) for females is greater than unity. This indicates that females tended to spend a greater amount of time on the phone during their non-primary routine days. Thus, regarding the duration of calls, females tended to be on the phone longer during their non-primary routine days, while males tended to be on the phone longer during their primary routine days.

To summarize, we have identified gender-specific traits in calling behavior. By analyzing the five features, we conclude that considering the type of day (primary or non-primary routine day) while examining call records is crucial for extracting gender-wise traits. Assuming that weekly activity patterns can also be reconstructed from the CDRs, we conclude that our findings can be utilized to estimate the gender of the user who produced an anonymized CDR.

3.4 Calling Behavior by the Occupation type

Similarly, we compared calling behavior across occupation types. Figure 5 compares the calling behavior recorded on primary routine days and that of non-primary routine days across occupation types. The features used are the same as those in Figure 4. We can observe distinctive differences in calling behavior according to occupation type. For instance, the value of feature (A) for *Household*, 1.64, indicates that users who were primarily engaged in household tasks tended to call more frequently on their non-primary routine days, while those in the *Worker* and *Student* categories appeared to exhibit similar call frequency trends on their primary and non-primary routine days. This indicates that differences in calling frequency between the primary and non-primary routine days are keys to identifying users who are primarily engaged in household tasks, where greater variability is observed.

Figure 5. Comparison of calling behavior on the primary and non-primary routine days across occupation types

Further, the feature (B) values for *Household* and *Student*, 1.78 and 1.68, respectively, and the feature (C) values for *Household* and *Student*, 2.18 and 1.78, respectively, largely deviate from unity, while (B) and (C)'s values for *Worker* are close to unity. This indicates that the call times for *Household* and *Student* varied more on non-primary routine days than they did on primary routine days. In contrast, trends for *Worker* are again similar for the primary and non-primary routine days. That is, *Household* and *Student* exhibit more flexibility in the timing of calls on the primary routine days. This indicates that a call time distribution that is consistent throughout the week can be a key to identifying users who are engaged in some type of income-earning activities.

Regarding the duration per call and total duration of calls per day, we observe that users in the *Household* category have particular trends compared to users in *Worker* and *Student*. As shown by features (D) and (E) for *Household*, those who were primarily

engaged in household tasks made shorter but more frequent calls on their primary routine days, whereas the duration per call and total duration of calls per day for *Worker* and *Student* did not vary according to the type of day. This indicates that variability in the call duration and the total duration of calls between primary and non-primary routine days can be a key to identifying users who are engaged in household tasks. Furthermore, we show that combining analysis results for features (A), (B), and (C) enables us to identify users who are primarily engaged in income-earning activities, household tasks, or school, based on calling behavior.

PROTOTYPES OF CALLING BEHAVIOR

In this section, we narrowed the time window from a weekly basis to an hourly basis, to understand trends in calling behavior across gender and occupation types. As a result, we aggregated the call times by hour in this section. We extracted parts-based representations of calling behavior by conducting vector quantization against the time and location distribution of call records. This enabled us to cluster the data into mutually exclusive prototypes [8].

3.5 Method for Extracting Prototypes

We employed non-negative matrix factorization (NMF) for vector quantization. NMF was applied to the call records of the 922 mobile users, where the distribution of call records for a single day is expressed as a 72×1 matrix. The first set of 24 elements out of 72 consists of hourly counts of call records for *Home*. The first element is the total number of calls from *Hour 0*, occurring between 0:00 and 0:59, and the 24th element denotes *Hour 23*, occurring between 23:00 and 23:59. The next set of 24 elements is structured similarly to the first set, and it accounts for the number of calls from the *Primary out-of-home location*. In a similar manner, hourly counts of calls for other locations, which are any locations except for home and work/school, are captured by an additional set of 24 elements. As a result, we obtained 922 sets of 72×1 column vectors. To employ NMF, we solved the equation below by following an algorithm, which allows only additive combinations [9]:

$$V \approx WH$$

where we obtained non-negative matrix factors W and H given a non-negative matrix V . Given 922 sets of 72×1 column vectors, the vectors were placed in the columns of a 72×922 matrix V . This matrix was approximately factorized into a $72 \times r$ matrix W , and an $r \times 922$ matrix H . Here we selected $r = 3$ to analyze differences in the major features of calling behavior. We then defined the cost function that evaluates the quality of approximation for iterative updates of W and H . We calculated the distance between two non-negative matrices and measured the square of the Euclidian distance [14]. As a result of repeated iterations, we obtained an optimal matrix factorization.

To understand the calling behavior prototypes for males and females, we split the call records obtained from SPACE by gender. Then, we separated the records by primary and non-primary routine days. We assumed that the calling behavior of the primary routine days was different from that of the non-primary routine days, based on the analysis results described in the previous section. We must note that the call records we obtained from SPACE were the aggregation of single-day records for 922 users. That is, the users that generated the records for the primary routine day were not the same as those that generated the non-primary routine day records. To maintain the population composition in terms of income level, we set our survey schedule to randomize the distribution of the days of the week for each

income level. Thus, we assume that this condition did not bias our sampling, and did not affect the analysis results.

3.6 Prototypes of the Primary Routine Day

Figures 6(a) and 6(b) show three calling behavior prototypes for males and females, whose call records in the SPACE data fall on their primary routine days. Each prototype is expressed in a 24×3 matrix, showing time distributions between zero and 23 hours in the rows, and call locations of *Home*, *Primary out-of-home location*, and *Others* in each column from left to right. The intensity of the color indicates the intensity of calls for the specified time band and location. For instance, in each matrix, the cell on the top of the left column represents the intensity of calls for *Hour 0* from home. The percentage indicates how significant each component is for explaining the population's call record trends. For example, Pattern m1, whose percentage is 46% and the greatest among the three, is the most dominant pattern for the calling behavior of males on the primary routine day.

Figure 6. Time distributions of call locations for (a) male and (b) female on the primary routine day (a) Upper (b) Lower

As shown in Figure 6, Pattern m1, which is the most dominant pattern for males, shows that males tended to call from *Primary out-of-home location* around midday, peaking in *Hour 12*. In contrast, the most dominant pattern for females, Pattern f2, shows a peak of *Hour 21* at *Home*. Furthermore, all extracted patterns for females show a higher intensity of calls from *Home*. This trend follows the one exhibited in Figure 2, which indicates that call locations could be a key to distinguishing the gender of users. Although the trend in call locations partially captures gender differences, frequently recorded hours for the remainder of males, Pattern m2 and m3, are apparently difficult to distinguish from those for females, Patterns f1, f2, and f3. For both males and females, intensive calling hours are *Hour 11*, *Hour 12*, *Hour 20*, and *Hour 21* from *Home*. It is still fair to conclude that the location of calls around midday could be a key to identifying the user's gender. That is, males mostly call from *Primary out-of-home location* and females call from *Home* around midday on their primary routine days.

We then investigated what occupation types and individual income levels represent each prototype described in Figures 6(a) and 6(b). Figures 7(a) and 7(b) describe the distribution of the three patterns illustrated in Figures 6(a) and 6(b) across occupation types *Worker*, *Household*, *Student*, and *Others*. Interestingly, we can observe that the dominant pattern for males differs according to the occupation type. Regarding *Worker*, Pattern m1 is the most common pattern, in which users tend to call around midday from *Primary out-of home location*. Among males, users in the *Household* and *Student* categories tended to call at *Home* around midday as described in Pattern m2. However, the common pattern for females categorized as *Worker*, *Household*, and *Student* are all the same, Pattern f2, where they tended to call from *Home* early in the morning and late evening. This indicates that calls from *Primary out-of home location* around midday can be a key to identifying males who are primarily engaged in income-earning activity.

Figure 7. Distribution of occupation types for three principle patterns for (a) male and (b) female on the primary routine day (a) Upper (b) Lower

Figures 8(a) and 8(b) describe the distribution of the calling patterns illustrated in Figures 6(a) and 6(b) for individual income levels. We split call records for males into four groups based on their individual income level. Call records for females were split into three income groups and a separate non-income group, because the majority of the female users did not earn an income.

Figure 8. Distribution of individual income levels for three principle patterns for (a) male and (b) female on the non-primary routine day (a) Upper (b) Lower

Among males, we observed that their common patterns varied according to their income level. As described in Figure 6(a), the higher income levels contained larger ratios of Pattern m1. Conversely, the lower individual income levels contained larger ratios of Pattern m3. This result is consistent with our field observations where males in lower-income groups tend to be engaged in self-employed jobs whose work locations tend to be their own home. However, no particular trends were observed among females across the income levels. Pattern f2 was the most

common calling pattern for all income levels. These trends imply that calling behavior could be a key to understanding the individual income level for males, but not for females.

It is worth noting that the income level used here is not a household income but individual income, which means we assess how much each person earns annually. Incidentally, we do not observe distinctive differences in calling patterns due to different household income levels when we use the household income level for analysis. We assume this implies that calling behavior strongly reflects the characteristics of individuals, rather than those of households.

3.7 Prototypes of the Non-primary Routine Day

Figures 9(a) and 9(b) contain three prototypes composed of time and location distributions of call records. These call records were generated by calls made on non-primary routine days. All patterns for males show that they tended to call from *Home* on non-primary routine days. By contrast, we observed that some females tended to call from *Primary out-of home location* on non-primary routine days. As described in Figure 6(b), *Home* was the most common call location for females on their primary routine days. This is probably because working outside of the home was unlikely to be the primary routine for the female users. Because females have such opportunities only on limited days (significantly fewer days than males), working outside the home tends to be counted as an activity on the non-primary routine day. This result strongly implies that, among those categorized into *Worker*, male and female users exhibited opposite call location trends on their primary and non-primary routine days. That is, analyzing the distribution of call locations and the type of day can reveal gender differences among users in the *Worker* category. We note that the factorization may not be the optimal approach for understanding non-primary routine day trends for our data, because the number of their records is limited for the type of days.

Figure 9. Time distributions of call locations for (a) male and (b) female on the non-primary routine day (a) Upper (b) Lower

Figures 10(a) and 10(b) describe the distribution of the calling patterns illustrated in Figures 9(a) and 9(b) for occupation types

Worker, Household, and Student. Pattern m1 was the most common pattern for males overall. In contrast, we note that the dominant pattern varied according to the occupation type among females. This indicates that time and location distribution trends on the non-primary routine day varied according to the occupation type among females. Incidentally, as shown in Figure 10(a), we provide no observation for males who are engaged in household tasks on the non-primary routine day in the SPACE data. This reflects the predominant societal norms in Bangladesh, where most males are engaged in income-earning activity while most females perform household tasks.

Figure 10. Distribution of occupation types for three principle patterns for (a) male and (b) female on the non-primary routine day (a) Upper (b) Lower

Last, we examined the distribution of calling patterns for individual income levels. The income levels were classified following the same rules used in Figures 8(a) and 8(b). As described in Figure 11(a), the dominant pattern for all income levels among males was Pattern m1, where the intensity of calls was highest at *Hour 18*. This trend was opposite from that of primary routine days, where time of day and call location distribution could indicate income levels for males. Conversely for females, dominant patterns differed across income levels, as shown in Figure 11(b). Again, the trend was opposite from that of primary routine days. Pattern f3 was common in the middle level, where calls originated on late evenings from *Home*. However, Pattern f1 was dominant for no-income females where calls tended to be made in the late morning and early afternoon from *Home*.

Figure 11. Distribution of individual income levels for three principle patterns for (a) male and (b) female on non-primary routine days (a) Upper (b) Lower

These trends followed those discussed in Section 3, where males were more flexible on their primary routine days and females were more flexible on their non-primary routine days. Our analysis results indicate that time of day and call location distribution trends are only useful for determining the income levels of female users on non-primary routine days.

4. COMPARING SINGLE-DAY CALL RECORDS WITH TWO-MONTH CDRs

In this section, we compare basic calling behavior statistics generated from single-day call records from SPACE data and CDRs for 58 volunteers. In the comparison, we employ a set of statistics under two different time frames, referred to as daily and weekly basis time frames. In the daily basis time frame, we generate one set of statistics by calculating averages per day. In the weekly basis time frame, we generate two sets of statistics by separately calculating averages for primary and non-primary routine days. As a result, we can examine the advantages and limitations of capturing calling behavior from survey data.

4.1 Calling Behavior by Gender

First, we compared SPACE data and CDRs by gender. Table 4(a) and 4(b) present overall calling behavior trends for SPACE data and CDRs in the daily basis time frame. The F/M values represent the ratio of Female (F) to Male (M), which were calculated to compare the trends of (a) SPACE and (b) CDRs. For instance, the F/M value for feature (C) was 0.95 for (a) SPACE and 0.99 for (b) CDRs. That is, both ratios are slightly smaller than unity. This indicates that the feature (C) values for males were slightly higher than those for females for both (a) SPACE and (b) CDRs. By comparing the trends for the remaining features in the two tables, we concluded that their overall trends were similar across the features.

Table 4. Daily basis calling behavior by gender for (a) SPACE data and (b) CDRs (a) Upper (b) Lower

	(A) Total number of calls per day (times/day)	(B) Call time deviation per day (hours)	(C) Call time variance per day (hours)	(D) Average duration of calls per day (seconds)	(E) Total duration of calls per day (seconds)
Male(M)	4.70	2.69	10.65	110.98	458.65
Female(F)	3.85	2.41	10.08	135.79	491.47
F/M	0.82	0.90	0.95	1.22	1.07

	(A) Total number of calls per day (times/day)	(B) Call time deviation per day (hours)	(C) Call time variance per day (hours)	(D) Average duration of calls per day (seconds)	(E) Total duration of calls per day (seconds)
Male(M)	4.07	1.90	8.70	98.98	371.72
Female(F)	3.70	1.85	8.64	113.52	401.11
F/M	0.91	0.97	0.99	1.15	1.08

We then split the CDRs into primary and non-primary routine days. We used the same set of features for r that were employed for Figure 4 in 3.3. Using CDRs, five features were calculated for males and females in Figure 12. Interestingly, there were two main differences between the trends in the SPACE data and CDRs.

Figure 12. Comparison of calling behavior on the primary and non-primary routine day of CDRs by gender

First, gender differences were more noticeable in terms of call duration, i.e. feature (D) and (E), in the CDRs; conversely, gender differences were more noticeable in terms of the frequency and time distribution of calls, i.e. features (A), (B), and (C), in the SPACE data. Second, differences between the primary and non-primary routine days were more distinctive in the calling behavior of males in the CDRs, while those differences were more

significant in the calling behavior of females in the SPACE data. At this time, we are not able to identify the reason we obtained opposite trends when we changed the length of the time frame.

4.2 Calling Behavior by Occupation type

Similarly, we compared SPACE data and CDRs by occupation type. Table 5(a) and 5(b) present overall calling behavior trends for SPACE data and CDRs in the daily basis time frame. H/W values represent the ratio of *Household* (H) to *Worker* (W) and S/W values represent the ratio of *Student* (S) to *Worker* (W), which help compare the trends in the two tables. For example, the ratio of H/W to S/W for feature (C) was 0.97 to 0.71 for (a) SPACE, and 0.95 to 0.68 for (b) CDRs. That is, the ratio of H/W to S/W for (a) SPACE was 1.366, which was calculated dividing 0.97 by 0.71. Similarly, the ratio of H/W to S/W for (b) CDRs was 1.397, which was calculated by dividing 0.95 by 0.68. By comparing the two ratio values 1.366 and 1.397, we observed that the relative trends of three values, W, H, and S, were similar for (a) SPACE and (b) CDRs. By comparing the trends for the remaining features in the two tables, we again concluded that their overall trends were similar across the features, as they were for gender.

Table 5. Daily basis calling behavior by occupation type for (a) SPACE data (b) CDRs (a) Upper (b) Lower

	(A) Total number of calls per day (times/day)	(B) Call time deviation per day (hours)	(C) Call time variance per day (seconds)	(D) Average duration of calls per day (seconds)	(E) Total duration of calls per day (seconds)
Worker(W)	4.73	2.70	10.69	112.34	476.81
Household(H)	3.67	2.40	10.37	137.91	474.40
Student(S)	3.78	2.11	7.61	119.06	383.02
H/W:S/W	0.78 : 0.79	0.89 : 0.78	0.97 : 0.71	1.23 : 1.06	0.99 : 0.80

	(A) Total number of calls per day (times/day)	(B) Call time deviation per day (hours)	(C) Call time variance per day (seconds)	(D) Average duration of calls per day (seconds)	(E) Total duration of calls per day (seconds)
Worker(W)	4.12	1.98	9.17	93.65	355.18
Household(H)	3.73	1.84	8.70	126.21	448.38
Student(S)	2.36	1.38	6.22	93.04	201.46
H/W:S/W	0.91 : 0.57	0.93 : 0.70	0.95 : 0.68	1.35 : 0.99	1.26 : 0.57

We then split the CDRs into primary and non-primary routine day records, as we did for Figure 5 in Section 3.4. Statistics generated under the weekly basis time frame are shown in Figure 13. Apart from the trends observed when analyzing gender in the previous subsection, we note completely different trends between the CDRs in Figure 5 and the SPACE data in Figure 13. Regarding the different trends derived from the lengthening of the time frame, we will review additional literature and attempt to identify clues for further research in the following subsection.

Figure 13. Comparison of calling behavior on the primary and non-primary routine day for CDRs by occupation type

In sum, our analysis of daily trends showed that similar trends could be observed between the single-day call records, which were gathered from various individuals for primary and non-

primary routine days (with the specification of the day type), and the CDRs. This implies that the SPACE field survey data were sufficient to understand the daily traits obtainable from longer-term CDRs. However, when we expanded the time frame from daily to weekly, the single-day records were insufficient to obtain trends that we would expect to extract from longer-term CDRs.

4.3 Necessity of Capturing Bounded Nature in Calling Behavior

As described in the previous sections, there are inconsistent trends between SPACE data and CDRs, which appear to be caused by changing time frames. We consider traits that cannot be captured by analyzing collections of single-day call records to be associated with the bounded nature of human activity patterns. This probably results from differences in the sizes of the individual time windows in which people repeat their routines in different recurrence rates. In fact, different time frames, e.g. 24, 48, and 72 hours, are used to describe the recurrence and temporal periodicity of the bounded nature of human trajectories [5]. Because calling records can trace human mobility patterns to some extent, repeatedly visited locations such as home, work places, and other significant locations are keys to extracting such bounded nature in calling behavior.

By comparing the statistical analysis results from SPACE data and CDRs, we described the advantages and limitations of analyzing calling behavior using collections of single-day call records. We assume that it will be necessary to further examine the routines, which are expressed by returning to a few limited locations on a regular basis within different time frames. One possible method to capture the various time frame lengths would be to expand the field survey's interview time frame from a single-day to three days. However, from a practical standpoint, it may be difficult to obtain meaningful responses by asking respondents to recall activity that occurred during the previous three days. Thus, further efforts are necessary to improve our approach, by developing a method that enables us to capture the bounded nature from calling records.

5. CONCLUSIONS

In this work, we identified calling behavior traits that can distinguish gender and occupation types by comparing the analysis results of SPACE data with those of CDRs. Analysis results suggest that a higher ratio of calls from home can be a key to distinguishing females from males. Females tended to call from home around midday on their primary routine day. In addition, the variability in the frequency and time distribution of calls from female users was more distinctive, according to the type of day. That is, constant frequency and time distribution of calls throughout the week were keys to identifying male users. Regarding the average duration per call and total duration of calls per day, females tended to use the phone more often on the non-primary routine day, whereas trends for males did not exhibit such variability, according to the type of day. Specifically for males on the primary routine day, the higher the individual income level, the higher the probability that they would initiate calls from their primary out-of-home locations around midday. The lower the income level of the user, the higher the probability that they would initiate calls from home in the morning. Conversely, there were no distinctive differences in primary routine day calling behavior for females related to their individual income level.

Our findings suggested that time of day and call location distribution were keys to extracting differences in calling behavior by gender and occupation type. As the basis for this analysis, we

introduced the concept of weekly activity patterns, which specifies whether the day of the call corresponds to a day on which the user is engaged in their primary routine. Analysis results suggest that the type of day could be also a key for extracting traits from call records. By comparing the statistical analysis results of SPACE data and CDRs, we also concluded that the current approach is not sufficient for capturing the regularity of individual calling behavior within different time frames.

With our work, we exploited the potential of deriving demographic attributes from anonymized CDRs. Although experiments were performed with a limited number of CDRs, the techniques developed in this study are capable of extracting gender and occupational traits from large-scale CDRs. Experimental results infer that our approach is capable of constructing a demographic attribute prediction model based on anonymized CDRs. Considering the importance of specifying call locations to extract demographic attributes, further studies are necessary to improve existing CDR location labeling methods. Further investigation is required to develop a method to extract the bounded nature of human mobility from calling behavior by advancing the concept of weekly activity patterns. This would enhance the extraction of useful traits from call records to distinguish occupation types; these are closely related to individual regularity that varies according to the length of time frames.

6. ACKNOWLEDGMENTS

We appreciate the telecommunications operator and volunteers who provided data for the research. Part of this work was supported by GRENE-ei, funded by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan.

7. REFERENCES

- [1] Aarthi, S., Bharanidharan, S., Saravanan, M., and Anand, V. 2011. Predicting customer demographics in a mobile social network. In *2011 IEEE International Conference on Advances in Social Networks Analysis and Mining*, 553-554. IEEE.
- [2] Bangladesh Telecommunication Regulatory Commission (BTRC). Mobile phone subscribers in Bangladesh 2014 April. Retrieved 23 May, 2014, from <http://www.btrc.gov.bd/content/mobile-phone-subscribers-bangladesh-april-2014>
- [3] Blumenstock, J., and Eagle, N. 2010. Mobile divides: gender, socioeconomic status, and mobile phone use in Rwanda. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, 6. ACM.
- [4] Financial Express. Where Dhaka stands today. Retrieved 30 May, 2014, from <http://www.thefinancialexpress-bd.com/old/index.php?ref=MjBfMDJfMjZfMTNfMV85MI8xNjEyOTE=>
- [5] González, M. C., Hidalgo, C. A., and Barabási, A. L. 2008. Understanding individual human mobility patterns. *Nature*. 453, 7196, 779-782.
- [6] Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M. J., Rowland, J., and Varshavsky, A. 2011. Identifying important places in people's lives from cellular network data. In *Pervasive Computing*, 133-151. Springer Berlin Heidelberg.
- [7] Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., & Willinger, W. 2012. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, 239-252. ACM.
- [8] Lee, D. D. and Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*. 401, 788-791.
- [9] Lee, D. D. and Seung, H. S. 2000. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing 13*. Cambridge, MIT Press.
- [10] Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., and Ma, W. Y. 2008. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 298-307. ACM. New York, NY.
- [11] Lu, X., Bengtsson, L., and Holme, P. 2012. Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences*, 109, 29, 11576-11581.
- [12] Mo, K., Tan, Ben., Zhong, Erheng., and Yang, Q. 2012. Report of task 3: your phone understands you. Paper presented at Nokia mobile data challenge 2012 workshop, Newcastle, UK, 18-19 June 2012.
- [13] Lu, X., and Pas, E. I. 1999. Socio-demographics, activity participation and travel behavior. *Transportation Research Part A: Policy and Practice*, 33(1), 1-18.
- [14] Paatero, P. and Tapper, U. 1997. Least squares formulation of robust non-negative factor analysis. *Chemometr. Intell. Lab.* 37, 23-35.
- [15] Pas, E. I. 1984. The effect of selected sociodemographic characteristics on daily travel-activity behavior. *Environ. Plann. A*, 16, 5, 571-581.
- [16] Song, C., Koren, T. K., Wang, P., and Barabási, A. L. 2010. Modelling the scaling properties of human mobility. *Nat. Phys.* 6, 10, 818-823.
- [17] Song, C., Qu, Z., Blumm, N., and Barabási, A. L. 2010. Limits of predictability in human mobility. *Science*, 327.
- [18] Sultana, N. 2004. Polygamy and Divorce in Rural Bangladesh. *Empowerment*, 11, 75-96.
- [19] Szell, M., Sinatra, R., Petri, G., Thurner, S., and Latora, V. 2012. Understanding mobility in a social petri dish. *Scientific reports*, 2.
- [20] White, S. 1992. Arguing with the Crocodile: Gender and Class in Bangladesh. London: Zed Books Ltd.
- [21] World Bank. 2012. Information and communications for development 2012: maximizing mobile. Washington, D. C., World Bank.
- [22] World Bank. 2013. World development indicators. 2013. Washington, D. C., World Bank.