Framework to Study Migration Decisions Using Call Detail Record (CDR) Data

Viren Dias, *Member, IEEE*, Lasantha Fernando, *Member, IEEE*, Yusen Lin, *Member, IEEE*, Vanessa Frias-Martinez, *Member, IEEE*, and Louiqa Raschid, *Fellow, IEEE*

***Abstract*— This article addresses the challenges of using call detail record (CDR) data to study migration. Repurposing CDR data for this task have many advantages, including the lower costs of data collection and the potential for contemporaneous analysis. We present a framework for the repurposing and analysis of CDR data. We identify the home location of a subscriber, with corresponding confidence measures, and determine if the subscriber is a definite migrant, likely migrant, likely nonmigrant, or definite nonmigrant. A predictive model then uses mobility and social network features, extracted from the CDR data, to predict the individual decision to migrate. We are the first to address the challenging task of predicting the migration decision at the individual level. We also provide insight into features that can have an impact on the decision to migrate. An in-depth evaluation using CDR data from two provinces in Sri Lanka provides a granular map of migrant inflow and outflow. The success of our prediction model and the insights gained from the evaluation prepare the way for the repurposing of CDR data for social good with a focus on migration.**

***Index Terms*— Call detail record (CDR) data, migration patterns, migration prediction.**

I. Introduction

THE worldwide adoption of mobile devices, and human activity within the cyber–physical space, has provided a valuable and ubiquitous stream of digital traces, known as call detail record (CDR) data. In particular, this is a log of the location of a mobile device as recorded by contacts with base transceiver stations (BTSs). There have been many successful applications of repurposing CDR data, for example, in the areas of transportation and urban computing [1]–[3]. More recently, there has been awareness of these resources in the wider research community, in particular among computational social scientists who are interested in repurposing CDR data to study human behavior. Agencies such as the World Bank and UNICEF have also been actively engaged in utilizing CDR data through collaborations with mobile service providers.

Manuscript received October 16, 2021; revised January 20, 2022; accepted April 18, 2022. This work was supported in part by the USA National Science Foundation under Grant CNS 1750102 and in part by the grants from the International Development Research Centre (IDRC), Canada. *(Corresponding author: Louiqa Raschid.)*

Viren Dias and Lasantha Fernando are with LIRNEasia, Colombo 08, Sri Lanka (e-mail: viren@lirneasia.net; lasantha@lirneasia.net).

Yusen Lin is with the UMIACS, University of Maryland, College Park, MD 20742 USA (e-mail: yusenlin@umd.edu).

Vanessa Frias-Martinez is with the UMIACS, University of Maryland, College Park, MD 20742 USA, and also with the School of Information, University of Maryland, College Park, MD 20742 USA (e-mail: vfrias@umd.edu).

Louiqa Raschid is with the UMIACS, University of Maryland, College Park, MD 20742 USA, and also with the Smith School of Business, University of Maryland, College Park, MD 20742 USA (e-mail: lraschid@umd.edu).

Digital Object Identifier 10.1109/TCSS.2022.3177727

Internal migration refers to the migration of individuals from one region to another within the same geopolitical entity, typically within the same country [4]–[6]. In recent years, there has been an increase in the volume, types, and complexity of human internal migration in many countries, mostly due to economic crises, political instability, and various types of natural disasters [6], [7]. Economists have developed econometric models to predict individual internal migration decisions [8], [9]. These models typically rely on census data, which is often unavailable, out of date, or difficult and costly to generate. Recognizing these limitations, there has been research on repurposing ubiquitous data generated in a passive manner, e.g., email, Web, and social media data [10]–[12]. A limitation of these new approaches is that they may suffer from a bias problem. To explain, the demographic and economic backgrounds of email, Web, and social media users may not be representative of the population at large [13].

In an attempt to overcome this limitation of biased data, there has been more recent research to repurpose digital mobility trace data. As mentioned, CDR data are a log of the location of a device as recorded by BTSs within a cellular network. An alternate source of location data is harvested by location intelligence companies (via the mobile application developer SDK) when smartphone-based mobile applications share their location with the GPS network. Of the two sources, CDR data are known to have much higher penetration rates across diverse population groups, and research has shown that CDR data can be representative of the population at large [14], [15]. CDR data have been used to model behaviors during pandemics, such as the H1N1 flu outbreak and natural disasters [16], [17]. Research has also successfully demonstrated the use of CDR data to identify migrants and measure the volume and direction of flow of internal migration, for example, in Rwanda and Namibia, respectively [18], [19]. Much of this research has been limited to aggregate analysis (volume and flow) but has not studied an individual's migration decision.

To summarize, census data provide accurate models for migration, but the data may be outdated and are expensive to generate. Email, Web, and social media data, as well as smartphone location data, may be biased. CDR data, in contrast, have been shown to be representative. We present a holistic framework for the repurposing and analysis of CDR data to better understand migration and determine if the subscriber is a migrant. We are the first to develop a prediction model that uses mobility and social network features to predict the individual decision to migrate.

To identify the home locations of a subscriber, we extend the heuristic approach in [20]. Our extension includes metrics to consider the confidence of a BTS being the home location of a subscriber. Using the home location and corresponding confidence measures, we label a subscriber as a definite migrant, likely migrant, likely nonmigrant, or definite nonmigrant. Our work is the first to determine confidence in migration and provide nuanced labeling of migration status. The confidence feature is important since migration maps need to provide an accurate migrant status that will be trusted by the social science community that studies migration. The current gold standard to study migration are the (more expensive to produce) census data or self-reported survey data; both have a reputation for providing accurate migration and demographic information [8], [21]. We validate the correctness of our predicted CDR-based results against the gold standard census data. used toward the identification of a or climate change, under extreme situations CDR-based migration maps could also be computed continuously and contemporaneously, producing frequent statistics that would be expensive to collect using surveys. To illustrate this, we produce a map of migrant inflow and outflow for two administrative areas in Sri Lanka circa 2013. One, the Western Province, is the most populous and most developed in the country. In contrast, the second area, the Northern Province, emerged from a 30-year civil war in 2009. We identify the areas with high levels of inflow and/or outflow, i.e., churn, and areas with high levels of net gain or loss of migrants, net gain, or loss of migrants.

We are the first to develop a prediction model for the individual's decision to migrate. Our objective is twofold: 1) to predict whether a subscriber will migrate and 2) to understand the role that behavioral features, including mobility and/or social network relationships, may play in that decision. We build upon previous work exploring the role of spatial dynamics [23] and social networks [24] on migration [23]. Further research showed that migrants rely on social support to deal with their migration experiences [24]. Our model will also extensive research on features extracted from CDR data [25], [26]. Of special note in our model will be identifying a social relationship feature reflecting the presence of migrants as close contacts in an individual's network as relevant to the migration decision.

The main contributions of this article are given as follows:

1) an end-to-end framework for CDR-based analysis of migration that can be contemporaneous;
2) an enhanced algorithm to identify the home location of a subscriber, confidence measures, and nuanced labeling of migration status;
3) a novel model to predict the individual decision to migrate;
4) an in-depth evaluation using CDR data from the Western and Northern Provinces of Sri Lanka.

II. Related Work

A. Mobility Trace Data and Human Mobility

The ubiquitous presence of mobile devices has generated a valuable stream of digital traces containing location information that has been used to model human activity within the cyber–physical space [57], [60], [61]. For example, researchers have explored the use of location information extracted from social media, e.g., geotagged tweets or foursquare visits, to predict the next location visited by a person [58]. Researchers have used GPS data collected from mobile applications to predict trip purpose and route choice [64]. CDRs have been used to approximate origin–destination (OD) matrices that characterize aggregate flows between two locations, so as to model travel behaviors, such as commuting patterns [59]. CDR data have also been used to predict evacuation patterns during natural disasters [62].

Some of these tasks, such as identifying the next location visited and inferring trajectories, require real-time processing, which is often done on a mobile device. As a result, novel architectural approaches need to be designed to ensure the efficient functioning of the mobile network [63]. We build on these works to model migration behaviors using location information extracted from CDR data. However, our specific tasks—to predict the home location and to predict migration—do not require real-time data or processing. These tasks can be executed offline, using archival data, and they do not impose real-time performance restrictions on the mobile device or the network.

A range of computational methods has been used for human mobility predictive tasks, as reported in the literature. This includes logistic regression (LogRed); random forest or XGBoost [64]; radiation models [65]; Markov models [62]; or deep learning approaches [66]. Typically, the deep learning approaches may show some (often limited) performance advantage over other methods, and this depends on the specific task. We use LogRed in our research for the following reasons: 1) the lack of spatiotemporal data complexity, i.e., predicting the migration decision only requires aggregate features, and this task will not benefit from deep learning and 2) the importance of model transparency. Deep learning approaches do not provide clear insights into the specific features that play a role in a person making a decision to migrate [67]. However, such insights are of utmost importance in the domain of migration research. LogRed is a popular modeling approach that can, indeed, provide such insights.

B. Mobility Trace Data and Migration

There exist two distinct types of internal migration: 1) circular, repetitive, and nonpermanent moves, such as migrant workers who move periodically between cities and rural areas [4] and 2) permanent migration, when individuals remain in their final destinations [5]. Researchers have extensively studied internal migration movements, so as to develop policies that try to prevent migrants from being left behind and allow them to adapt to their new settings. These policies can help enhance work opportunities, lifestyle, and family and social relationships [29]. A comprehensive review of internal migration patterns across 15 countries in Asia, including Sri Lanka, is given in [6] and [30].

A majority of the research uses census data. Recognizing the limitations of relying on census data, there has been research on repurposing ubiquitous data generated in a passive manner, e.g., email, Web, and social media data [10]–[12].

Email service logs have been used to identify international migration rates [10]. Anonymized log data from Yahoo! users have been used to generate short- and medium-term migration flows across countries [11]. Research using Twitter data studied internal and international migrations to determine flow direction and volume [12]. Much of this research has been limited to aggregate analysis, i.e., volume and flow, but has not studied individual migration decisions. A more serious limitation of all these approaches is that they suffer from a bias problem. To explain, the demographic and economic backgrounds of email, Web, and social media users are typically not representative of the population at large [13].

In an attempt to overcome this limitation of biased data, there has been more recent research to repurpose digital mobility trace data. CDR data are a log of the location of a device as recorded by BTSs within a cellular network; it is collected by providers and used for billing purposes. Location data are also harvested by location intelligence companies via the mobile application developer SDK; it is collected when the smartphone-based mobile applications share locations with the GPS network. These data have been widely used for digital marketing. GPS data collection requires smartphone ownership, which is not as common as cell (mobile) phone ownership, especially among lower income groups. GPS data can potentially incur a higher selection bias in comparison to CDR data [14], [15]. More relevant to our research, CDR data can also be used to reconstruct each subscriber's social network. This is typically not possible using mobile application-based location data from the GPS network. Our research will show the benefits of both mobility and social network features in predicting the decision to migrate.

C. Individual Migration Decision Prediction

There exists extensive work on the use of econometric models to predict whether an individual will migrate to a different region and to understand the reasons behind that decision. Related work has studied the role that expected earnings, housing prices, crime, weather, and employment might play on individual migration decisions [8], [9], [21]. A majority of this research relies on the existence of census data or migration-focused surveys. Discrete choice models assess whether an individual would migrate to a given region. For example, LogRed-based classifiers were used to predict whether an individual would migrate to one of the 324 metropolitan areas in the U.S. [27]. They used a set of individual demographic and socioeconomic variables extracted from individual U.S.-based Public Use Microdata Survey, as well as housing costs, climate, crime, and topography variables.

When census data are available, these models are indeed excellent. However, collecting census data, in general, and high granular individual data, in particular, are expensive and not accessible to many resource-constrained countries. Furthermore, available census data may be outdated since more detailed granular data are collected decennially.

D. Home Location for Migration Identification

There exists an important body of work that uses CDR data to determine migration flows, i.e., to approximate the number of people that migrate to a region based on the number of home location changes computed using CDR data. For example, a large dataset of 72 billion CDR data records collected from October 2010 to April 2014 in Namibia was used to determine internal migrant flows within 13 regions [19]. The estimated flows were then compared with census-derived migration statistics to assess the accuracy of using CDR-based home location changes as a proxy for migration flows. CDR data from the 2005-2008 period were used to approximate internal migration flows in Rwanda and assess the role that population characteristics might play in migration [18].

Most of the current studies define the subscriber's home location as the BTS where the subscriber was observed most frequently during a given interval [20], [26], [35], [36]. While such methods have the advantage of simplicity, they can also be prone to error. Our research considers the confidence of a BTS (or a group representing a BTS entity) being the home location of a subscriber. We label a subscriber as a definite migrant, likely migrant, likely nonmigrant, or definite nonmigrant. Our work is the first to determine confidence in migration using CDR data. We posit that confidence in the home location identification is necessary for CDR-based predictive models to be adopted widely in the relevant research community. That community typically uses census data with accurate migration statistics and demographic information that is self-declared via surveys [8], [21].

In addition, our work is the first to show preliminary results toward the identification of a *migration window* that characterizes the time interval during which the migration has taken place. These insights will be important when assessing migrations during natural disasters, where collecting specific migration dates via surveys is a much harder task [43].

E. CDR-Based Migration Features

CDR data have been widely used to model mobility and social network behaviors that could be indicative of migration intentions [2], [16], [44], [45]. For example, spatial dynamic features, such as entropy or radius of gyration, and social ties features, such as the number of contacts or communication entropy, have been used to characterize postmigration behaviors [25], [26]. CDR-based social network data were used to evaluate the evolution of social ties during the migration processes [34].

We extend prior work with novel spatial and social diversity measures to incorporate more nuanced modeling of the relationships. Of special note is that we consider the presence of migrants as contacts in an individual's network.

We note that the features used in this article were at the granularity of individual subscribers. We also computed features that required an aggregation over the complex features of each subscriber in each ego network. This level of granularity and complexity was typically much higher compared to prior research, in particular transportation analysis, where features may be aggregated over each BTS.

TABLE I
CDR DATA SUMMARY STATISTICS

Statistic	Province	
	Western	Northern
Number of subscribers	4,946,819	1,177,251
Number of BTSs	672	246
Average calls per subscriber		
Working days	3.50	4.08
Non-working days	3.02	3.97
Average subscribers per BTS		
Working days	1,238	386
Non-working days	1,400	556
Average calls per BTS		
Working days	24,446	18,247
Non-working days	19,153	16,994

III. DATASETS AND METHODOLOGY FOR MIGRATION IDENTIFICATION

We first describe the CDR dataset. We then describe an enhanced algorithm to identify the home location of a subscriber and corresponding confidence measures. We also highlight how the confidence measure can help identify a migration window.

A. Call Detail Records

We use CDR data collected by a telecommunications company from January to September 2013 in the Western and Northern Provinces of Sri Lanka. The first level of administrative decentralization in Sri Lanka is into nine *provinces*; our data are from two of these provinces, namely, the Western and Northern Provinces. Each province is further subdivided into *districts*. There are six districts in the Northern Province, three districts in the (most populous) Western Province, and 25 districts overall. A district is further subdivided into *divisional secretariat divisions* (commonly known as DSDs). There are a total of 331 DSDs in Sri Lanka; the Northern Province contains 34, and the Western Province has 40. We define the migrants within the dataset as those subscribers who have different home locations and different DSDs when considering the first time period of January to March 2013 and the second time period of July to September 2013.

As shown in Table I, the dataset contains approximately five million subscribers in the Western Province and over one million in the Northern Province. We note that the CDRs are granular records at the level of individual subscribers. However, the data are pseudonymized, and we do not have access to any information about the subscriber. This lack of ground truth clearly complicates our tasks of prediction and validation. The CDR data only include voice calls; while the same provider and device are used for a range of other applications, these data were not included. Data cleaning included the removal of duplicate records, subscribers associated with BTSs that were not in the dataset, and so on.

The Spark Graph API was not optimal to handle some computations. Spark API-based computations had to be tuned to use memory efficiently; a similar computation using Hadoop was not memory bound, but there was often a tradeoff with the speed of execution using Hadoop. Computing the extensive set of features used for both home location identification and migration prediction, using the Spark dataset API, took up to two days on a small cluster of five nodes, each with four cores.

The two provinces that are used for the evaluation are very different. The Western Province is the most populous and most developed. In contrast, the Northern Province had emerged from a long civil war that ended in 2009. In the Western Province, the average count of calls per subscriber is 3.50 on working days; this goes down to 3.02 on nonworking days. For subscribers in the Northern Province, the average count of calls, for both working and nonworking days, is higher at 4.08 and 3.97, respectively. Subscribers in the Western Province may be accessing apps more frequently instead of making voice calls. We note that, while we may not be capturing all the activity for these subscribers in the Western Province, we do not believe that this data gap will have a negative impact on the accuracy of the home location algorithm or on the features that impact the decision to migrate. This is, indeed, validated by comparing our results of migrant counts with census-based results for Sri Lanka [30].

B. Base Transceiver Stations

There are 1557 individual BTSs in the Western and Northern Provinces of Sri Lanka. The distribution of BTS in some parts of the Western Province is very dense. Furthermore, calls are often switched between BTS in close proximity for load-sharing purposes. For these reasons, we made the decision to associate each BTS with a map display segment. This enabled us to consider a group of closely located BTS as a *single BTS entity* when generating features for the purposes of home location identification and migration prediction.

We define a map display segment as a 1 km $\times$ 1 km grid. The initial placement of the grid was random, and a total of 100 possible grid placements were considered by shifting the initial grid laterally and longitudinally by 100-m increments. The final placement of the grid was determined by minimizing the error term, i.e., the cumulative Euclidean distance between each BTS and the center of the grid cell containing the BTS.

All BTSs contained within the same grid cell were considered to be a *single BTS entity*. The location of the BTS entity was defined as the center of the grid cell. The coverage area of the grid cell was the merge of areas defined by the Voronoi tessellation of each individual BTS within the grid cell. As a result of this process, we consolidated the data into 918 BTS entities for the two provinces.

C. Home Location and Migration Identification

To estimate the home location of a subscriber, we used an extended version of the algorithm described in [20] with additional metrics calculated to determine the confidence of our estimation. In particular, for each BTS associated with each subscriber, we computed the following metrics: day

TABLE II
FEATURES USED FOR HOME LOCATION IDENTIFICATION, CALCULATED FOR EACH BTS VIA WHICH CALLS WERE MADE BY EACH SUBSCRIBER

Label	Description
BTS ID	The 1 km by 1 km grid cell ID.
Day count	The count of days in which calls were made at any time of the day.
Night count	The count of days in which calls were made between 9 p.m. and 5 a.m.
Neighborhood night count	The count of days in which calls were made, including calls made via neighbouring BTS within that BTS entity, between 9 p.m. and 5 a.m.
Day span	The number of days between the first and final calls made.
BTS confidence	The ratio of night count, to the cumulative sum of night counts for all BTS via which calls were made by the subscriber.
Neighborhood confidence	The ratio of neighborhood night count, to the cumulative sum of night counts for all BTS (within a BTS entity) via which calls were made by the subscriber.

count, night count, neighborhood night count, and day span. Then, we computed the following two confidence measures.

1) *BTS Confidence:* The ratio of the count of nights during which calls were made via a given BTS to the cumulative night count over all BTS.
2) *Neighborhood Confidence:* The ratio of the count of nights during which calls were made via a given BTS, or any of its neighbors within the BTS entity, to the cumulative night count over all BTSs.

Definitions are given in Table II.

We filtered the data to exclude BTSes as a potential home location for each subscriber based on the filtering criteria in Table II. The lower bound thresholds for BTS exclusion were chosen after a careful examination of each of the distributions to eliminate the long tail of BTSs that were not frequented by some subscribers. We verified that this exclusion utilized over 90%–95% of the data over all subscribers. A partial sensitivity analysis was also conducted to ensure that the final BTS selected as the home location was not impacted by the choice of the threshold.

The heuristic for home location identification in [20] only considered the *top-one ranked BTS*. Our enhancement is to consider the *top-three BTS, ranked by confidence*, in determining if a migration (nonmigration) is definite or likely. We ranked all candidate BTSs by neighborhood confidence and then by BTS confidence. We used a standard ranking with ties, where we skipped the relevant count of tied positions in the ranking; this is informally referred to as a "1-2-2-4" ranking. In the event of a tie for the top-one rank, we first checked if the tied BTS were neighbors within a BTS entity. If true, a BTS from the BTS entity was selected as the home location. If false, the subscriber was not further included in our experiments since we were unable to determine a home location for that subscriber.

TABLE III
HOME LOCATION CHANGE CLASSIFICATION DEFINITIONS

Label	Description
Definitely changed	All home locations ranked 1 to 3, disregarding their corresponding ranks, are different between the two time periods.
Likely changed	All home locations ranked 1, disregarding their corresponding ranks, are different between the two time periods.
Likely unchanged	At least one home location ranked 1 to 3, and its corresponding rank, is identical in both time periods.
Definitely unchanged	All home locations ranked 1 to 3, and their corresponding ranks, are identical in both time periods.

TABLE IV
MIGRATION CLASSIFICATION DEFINITIONS

Label	Description
Definite migrant	Home location definitely changed, and Divisional Secretariat Division (DSD) changed.
Likely migrant	Home location likely changed, and DSD changed.
Likely non-migrant	Home location definitely changed, likely changed, or likely unchanged, and DSD unchanged.
Definite non-migrant	Home location definitely unchanged, and DSD unchanged.

This next step, *also unique to our approach*, compared the top-three BTS across the two time periods and labeled a subscriber's home location as follows using the decision criteria in Table III: 1) definitely changed; 2) likely changed; 3) likely unchanged; and 4) definitely unchanged. The definitions are in Table III. For example, all of the top-three BTS have to be identical and, in the same order, in both time periods, to label the home location as definitely unchanged and the subscriber as a definite nonmigrant.

We note that a home location change might represent a move of a short distance within a given city; this is considered to be a relocation but not a migration. Thus, to disentangle relocation behavior from actual migration, we only considered migrants whose home location has changed from one DSD to another between the two time periods. To determine the DSD for a subscriber, we assign a BTS to a DSD such that the BTS location (center of the grid cell) lies within a DSD boundary.

We then considered whether this change in home location resulted in a change in DSD as well and labeled each subscriber as follows: 1) definite migrant; 2) likely migrant; 3) likely nonmigrant; and 4) definite nonmigrant. The descriptions are outlined in Table IV.

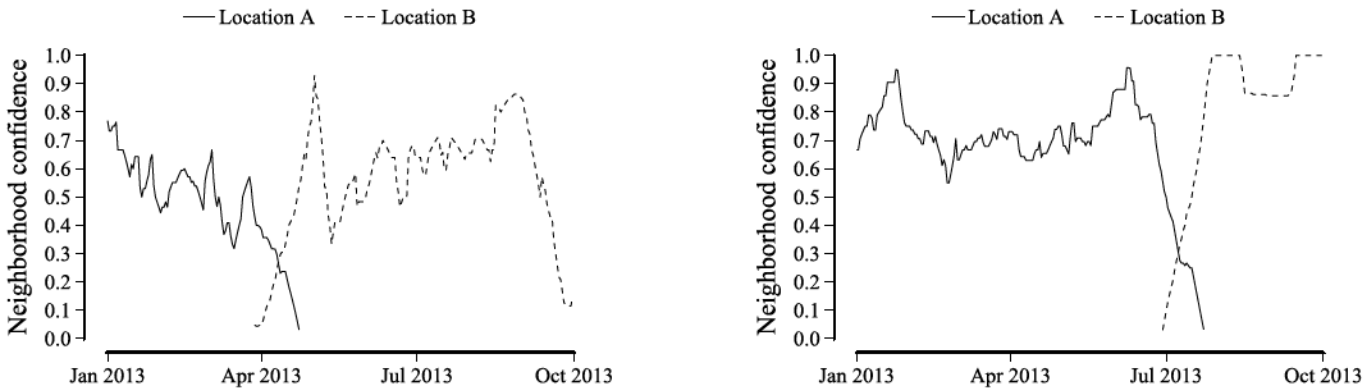

Fig. 1. Neighborhood confidence distribution with a clear home location shift.

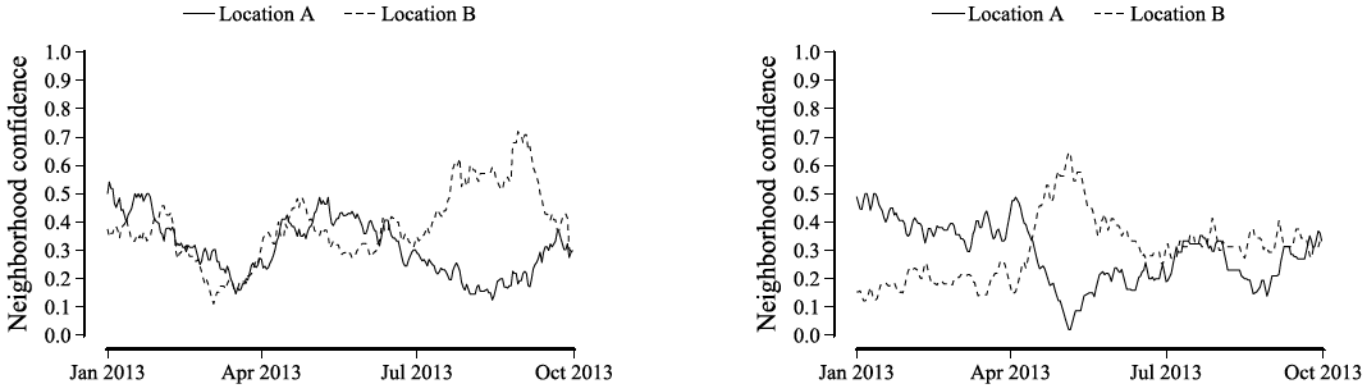

Fig. 2. Neighborhood confidence distribution with an unclear home location shift.

D. Results and Validation

We computed the home locations for all subscribers in the Western and Northern Provinces of Sri Lanka for the two time periods: January to March 2013 and July to September 2013. Figs. 1 and 2 provide plots of the neighborhood confidence of the two most prominent BTSs for four sample subscribers. As we can observe, the neighborhood confidence plots in Fig. 1 clearly identify that the home location has shifted—from solid to dashed—for both subscribers. More importantly, such distributions also allow us to pinpoint the actual window when the home location has shifted, i.e., the actual window of migration. In contrast, the confidence distributions in Fig. 2 do not clearly identify if there is a shift in the home location for the corresponding two subscribers.

Tables V and VI show the counts for the number of subscribers in the Western and Northern Provinces broken down by a change in home location and DSD, and migration class, respectively. The percentages of definite migrants identified in the Western and Northern Provinces are 1.62% and 2.84%, respectively. When considering all of the CDR data (not just the two provinces reported in this article), the definite migrant rate is 1.25%, and the likely migrant rate is 1.99%.

To validate our results, we compare our statistics with the closest census data statistics for Sri Lanka. First, we summarize some known limitations: We note that the cell phone penetration rate in Sri Lanka is reported as ≈ 60% circa 2019 [38]. We, therefore, expect that some fraction of migrations will be undetected due to the lack of cell phone ownership. There is also potential bias in our sample; cell phone subscribers that use the device consistently to provide sufficient history

TABLE V

BREAKDOWN OF SUBSCRIBER COUNTS WITH RESPECT TO CHANGE OF HOME LOCATION AND DSD IN EACH PROVINCE. (a) WESTERN PROVINCE (TOTAL POPULATION: 5 851 130). (b) NORTHERN PROVINCE (TOTAL POPULATION: 1 061 315)

(a)

	DSD		
Home location	Changed	Unchanged	Total
Definitely changed	95,027	31,665	126,692
Likely changed	48,943	71,284	120,227
Likely unchanged	—	155,244	155,244
Definitely unchanged	—	611,118	611,118
Total	143,970	869,311	1,013,281

(b)

	DSD		
Home location	Changed	Unchanged	Total
Definitely changed	30,177	9,203	39,380
Likely changed	14,024	16,877	30.901
Likely unchanged	—	41,964	41,964
Definitely unchanged	—	143,976	143,976
Total	44,201	212,020	256,221

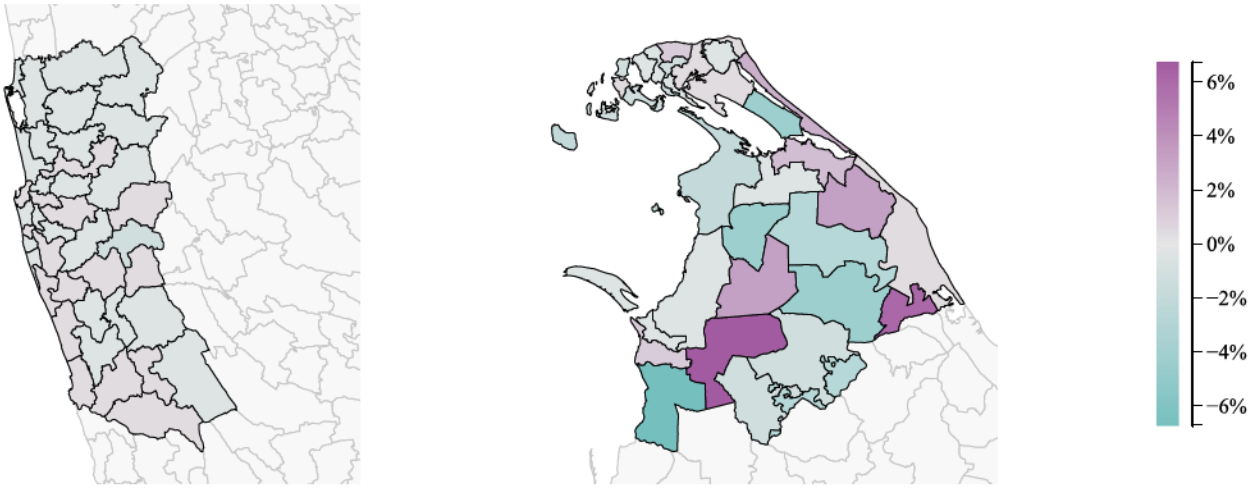

Fig. 3. Net migration (inflow − outflow) as a proportion of the population per DSD for the Western Province on the left and the Northern Province on the right.

TABLE VI
BREAKDOWN OF SUBSCRIBER COUNTS WITH RESPECT TO MIGRATION CLASSIFICATION. (a) WESTERN PROVINCE (TOTAL POPULATION: 5 851 130). (b) NORTHERN PROVINCE (TOTAL POPULATION: 1 061 315)

(a)

Migration class	Count	Percentage
Definite migrant	95,027	1.62
Likely migrant	48,943	0.84
Likely non-migrant	258,193	4.41
Definite non-migrant	611,118	10.44
Total	1,013,281	17.32

(b)

Migration class	Count	Percentage
Definite migrant	30,177	2.84
Likely migrant	14,024	1.32
Likely non-migrant	68,044	6.41
Definite non-migrant	143,976	13.57
Total	256,221	24.14

TABLE VII
BREAKDOWN OF DSDS WITH RESPECT TO CHURN (INFLOW + OUTFLOW) AND NET MIGRATION (INFLOW − OUTFLOW) AS A PROPORTION OF THE POPULATION. THE LABELS WERE DETERMINED BASED ON THE ABSOLUTE VALUES OF THEIR DISTRIBUTIONS FOR BOTH PROVINCES COMBINED: "HIGH" REPRESENTS THE TOP QUARTER, "MED" REPRESENTS THE INTER-QUARTILE RANGE, AND "LOW" REPRESENTS THE BOTTOM QUARTER. (a) WESTERN PROVINCE (TOTAL 40 DSDS). (b) NORTHERN PROVINCE (TOTAL 34 DSDS)

(a)

	Net Change					
Churn	− High	− Med	Low	+ Med	+ High	Total
Low	—	4	12	2	—	18
Med	1	9	4	5	—	19
High	—	2	—	1	—	3
Total	1	15	16	8	—	40

(b)

	Net Change					
Churn	− High	− Med	Low	+ Med	+ High	Total
Low	—	—	1	—	—	1
Med	3	9	2	2	1	17
High	8	—	—	2	6	16
Total	11	9	3	4	7	34

may also be associated with a higher socioeconomic status. We further note that we are not differentiating long- and short-term migrations.

Our validation is with respect to census data from 2008, which reports internal migration rates of $\approx$ 1.8% at the country level [37]. The aggregate crude migration intensity (ACMI), the percentage of permanent changes of address, for the five years prior to 2012, is reported as 8% across all of Sri Lanka [6], [30]. This translates to an average yearly value of $\approx$1.6%. Overall, our estimates from the CDR data (of 1.99%) and the statistics reported from the census data counts (between 1.6% and 1.8%) appear to be reasonably consistent.

E. Migration Observations in Two Provinces

Fig. 3 shows a map for net migration, as a proportion of the population at the DSD level of granularity, for: 1) Western Province and 2) Northern Province, respectively.

Fig. 3 and Table VII highlight different migration patterns in the Western and Northern Provinces. The Western Province has 40 DSDs; approximately half experience low churn and half medium churn. In contrast, for the 34 DSDs of the Northern Province, approximately half experience medium churn and half high churn. The net migration loss and gain show similar differences. Seven DSDs in the Northern Province experience high net gain, while 11 experience high net loss. In contrast, the net gain or net loss in the Western Province is limited to medium or low values. To summarize, the Northern Province is characterized by high or medium churn, as well as high and medium net loss or gain in migration. In contrast,

the Western Province is characterized by medium or low churn and medium or low net loss or gain in migration.

To provide some context for these observations, we note that these two provinces experience the highest levels of both churn and net migration in Sri Lanka. The Western Province is the most populous and developed in Sri Lanka. While this can attract migrants, the high cost of living and other issues can dampen enthusiasm. The Northern Province represents a very different scenario. In 2013, at the time of data collection, the Northern Province was five years out of a three-decade-long civil war and was experiencing significant socioeconomic changes. These contextual observations could provide some intuition into the drivers for the different migration patterns.

To summarize, our evaluation was specific to a single country, but it spanned two very different provinces based on demographic and socioeconomic metrics. We also validated our evaluation against the closest census data statistics for Sri Lanka. We expect that our methodology to determine the confidence in the home location and label the migration status will have similar performance accuracy across other datasets; however, the actual mobility behavior will vary based on the region, country, and local conditions. Some limitations of the approach are summarized under conclusions and future work.

IV. Individual Migration Prediction

Our objective is twofold: 1) to predict whether a subscriber will migrate and 2) to understand the role that behavioral features including mobility and/or social network relationships may play in predicting that migration decision.

We frame the migration identification problem as a classification task. Recall that we define the internal migrants within the dataset as those subscribers who have changed home locations and changed DSDs when considering the first time period of January–March 2013 and the second time period of July–September 2013. We use the set of features from the first time period to predict whether a subscriber will migrate between the first and second time periods.

While classification tasks are ubiquitous, the prediction of individual human decisions, such as migration, is both novel and challenging. The difficulty increases as we consider feature extraction from noisy big data, such as CDR data. There has been prior research on migration prediction, but it has relied on an accurate survey and demographic features [27]. The closest prediction research using CDR data is to predict BTS locations that a subscriber may have visited to uncover trips made by the subscriber who is hidden in the CDR data [3]; we note that such trip completion predictions are simpler since they consider a limited subset of the data. There is comparable research on predicting human decisions using noisy social media data. For example, migration across social media sites is estimated using features extracted from these sites [39]. This research does not predict at the individual level. Individual-level prediction models are presented in [40] and [41]. User behavior in creating a new social media post is studied in [40], and linking to a post is studied in [41].

To summarize, predicting individual human decisions, such as migration, is challenging. The prior prediction has relied on an accurate survey and demographic data. Our research is the first to predict these decisions using noisy CDR data. We note that the advantages of an accurate, continuous, and contemporaneous prediction are the ability for planners to stage interventions, offer customized services, and so on to mitigate the costs and stresses associated with migration.

A variety of methods have been explored in the research in [3], [40], and [41], including random forest, ranked support vector machines, LogRed and deep learning using CNNs and RNNs. A deep learning approach slightly outperformed other methods for the task of uncovering hidden trips [3]. However, a deep learning approach only has a performance advantage when the CDR features/dataset is disaggregated at the granularity of a single trip or a single day, and when the data are noisy and sparse, as in the case of uncovering hidden trips. For our task of migration prediction, however, we aggregate features over the entire training period and provide a single count or value of that feature for each subscriber. In this setting, there is a marginal performance advantage from using a deep learning approach.

On the other hand, a significant shortcoming of deep learning is the difficulty of interpreting the models. For our migration problem, social scientists have a strict requirement that they must understand the models driving human migration patterns. Thus, in our setting, the marginal potential performance advantage of prediction accuracy from deep learning is clearly offset by the difficulty of interpreting such models. With that in mind, we chose to address our task using a simpler LogRed model that is straightforward to interpret. We compare the results with XGBoost [33], which has been shown to be accurate and efficient across many diverse datasets.

A. Prediction Model Features

We describe the set of features that will be used as predictors of whether a subscriber will migrate. We build upon features from the literature that characterizes human behavior. Table VIII shows the set of features grouped into three classes.

1) Calling Patterns: These features are straightforward summary statistics, including the count of calls, duration of calls, count of distinct social contacts, and so on. For a given ego subscriber, we defined a social contact as having had at least one incoming call and one outgoing call with another subscriber.

2) Social and Spatial Diversities: We extend previous research and capture the behavioral diversity of a subscriber. To do so, we consider social and spatial diversity measures based on the Shannon entropy as defined by [32]. The measures are defined as follows.

1) *Social diversity (social_entropy):* This is defined by considering the diversity in the total call volume occurring between a subscriber and members of their social networks. For a subscriber i with K contacts, the value is given as follows:

$$\text{social\_entropy}(i) = \frac{-\sum_{k=1}^{K} p_{ik} \log(p_{ik})}{\log(K)}$$

TABLE VIII
FEATURES USED TO PREDICT MIGRATION

Label	Description
Calling patterns	
out_call_count	The count of outgoing calls.
inc_call_count	The count of incoming calls
all_call_count	The count of total calls.
out_call_duration	The total duration of outgoing calls.
inc_call_duration	The total duration of incoming calls.
all_call_duration	The total duration of all calls.
unique_bts_count	The distinct count of contact BTSs.
radius_of_gyration	The distance between home and visit BTS, weighted by visit frequency.
travel_distance_total	The total distance travelled.
travel_distance_max	The maximum distance travelled.
unique_contact_count	The count of distinct contacts.
contact_rate	The average count of calls initiated/received per contact.
contact_distance	The average distance between a home BTS contact's home BTS.
Social and spatial diversity	
hl_contact_bts_count	The distinct count of BTSs of a subscriber's contacts.
long_contact_bts_count	The count of contact BTSs outside of a subscriber's home province.
prov_contact_bts_count	The count of contact BTSs within a subscriber's home province.
contact_bts_count	The distinct count of BTSs contacted.
total_call_volume	The count of calls initiated / received.
contact_count	The distinct contact count (at least 1 call).
social_entropy	See definition in the text.
spatial_entropy	See definition in the text.
hl_spatial_entropy	See definition in the text.
prov_spatial_entropy	See definition in the text.
long_spatial_entropy	See definition in the text.
Social relationships	
network_1_migrant_count	The count of migrant contacts with 1 reciprocal call.
network_2_migrant_count	The count of migrant contacts with 2 reciprocal calls.
network_5_migrant_count	The count of migrant contacts with 5 reciprocal calls.

where

$$p_{ik} = \frac{V_{ik}}{\sum_{k=1}^{K} V_{ik}}$$

and V_{ik} is the total call volume between subscribers i and k.

2) *Spatial Diversity (spatial_entropy):* This is defined by considering the diversity in the total call duration to each BTS within the ego social network of a subscriber. For a subscriber i who has contacts in M distinct BTS areas, the value is given as follows:

$$\text{spatial\_entropy}(i) = \frac{-\sum_{m=1}^{M} p_{im} \log(p_{im})}{\log(M)}$$

where

$$p_{im} = \frac{D_{im}}{\sum_{m=1}^{M} D_{im}}$$

and D_{im} is the total call duration for subscriber i to BTS area m.

The preceding features capture diversity across contacts and their associated BTSs. However, this does not consider spatial diversity. For example, an individual subscriber could have a higher diversity with local contacts who share the same home location, or they may have higher diversity with contacts who live in other DSDs or other provinces. To account for this, we propose the following three novel diversity measures:

1) *Home Location-Based Spatial Diversity (hl_spatial_entropy):* We determine the home location of the altered contacts of the ego node. We consider the time spent by an ego node communicating with a contact who lives in a given BTS area to measure hl_spatial_entropy. For a subscriber i who has contacts with home locations in A distinct BTS, the value is computed as follows:

$$\text{hl\_spatial\_entropy}(i) = \frac{-\sum_{a=1}^{A} p_{ia} \log(p_{ia})}{\log(A)}$$

where

$$p_{ia} = \frac{D_{ia}}{\sum_{a=1}^{A} D_{ia}}$$

and D_{ia} is the total call duration for subscriber i to BTS area a.

2) *Province-Based Spatial Diversity (prov_spatial_entropy):* The definition is the same as for spatial diversity; however, only the BTS areas within the home location province of the ego subscriber are considered.

3) *Long Distance Spatial Diversity (long_spatial_entropy):* The definition is the same as for spatial diversity; however, only the BTS areas outside the home location province of the ego subscriber is considered.

3) Social Relationships: Prior research on predicting individual migration used census and survey data but did not consider social characteristics [8], [9], [21], [27]. Social relationship features, such as the count of contacts, were used to characterize postmigration behavior [25], [26]. CDR-based social network features were used to study the evolution of social ties during the migration process [34]. Our work is the first to use social relationship features for migration prediction.

In particular, we consider the presence of migrants as contacts in a subscriber's network.

B. Prediction Model Results

We provide some details about the experiment settings as follows.

1) We characterize each definite migrant, likely migrant, likely nonmigrant, and definite nonmigrant, with all the behavioral features described above.
2) We use collinearity tests to eliminate highly correlated features (with a correlation estimate > 0.7).
3) We use tenfold cross-validation to train and test the migration classification model.
4) We report on the area under the ROC curve (AUC) metric—averaged across tenfold—for the following subsets.
   a) The *entire dataset* of definite migrants, likely migrants, likely nonmigrants, and definite nonmigrants.
   b) A *restricted dataset* of only definite migrants and definite nonmigrants.
   c) A *balanced dataset* of equal numbers of definite migrants and definite nonmigrants generated by randomly downsampling the larger number of nonmigrants and repeated 50 times. Note that the AUC scores (averaged across all 50 iterations) did not show an improvement over the *restricted* dataset.
5) We also report on the AUC scores at different migration distance thresholds, i.e., models trained considering only migrants that migrate a minimum distance of 5, 10, and 15 km.

Table IX shows the performance of the LogRed model and XGBoost as measured by the AUC. The values in parentheses following the AUC are the population counts. Considering the entire dataset and LogReg, the AUCs are 0.70 and 0.73 for the Western and Northern Provinces, respectively; the AUCs for XGBoost are 0.73 and 0.76, respectively. These values improve when we consider larger migration distance thresholds. For distances greater than 15 km, the AUC for LogReg can go up to 0.82 and 0.80 for the Western and Northern Provinces, respectively, and 0.86 and 0.83 for XGBoost. The trend is that the AUC increases with longer migration distances, i.e., the predictive algorithm appears to be more robust at detecting long-distance home location changes.

The prediction model for the restricted dataset with only definite migrants and definite nonmigrants shows better results as expected. We observe that the LogReg AUC increases to 0.77 and 0.80 (from 0.70 and 0.73) for the Western and Northern Provinces, respectively. The XGBoost AUC increases to 0.80 and 0.83 (from 0.73 and 0.76). The downsampling experiments to create a *balanced* dataset for the count of definite migrants and definite nonmigrants did not produce any significant improvement for AUC; we do not report on AUC for the *balanced* dataset in Table IX. We note that the downsampling does have an impact on the significance of features in prediction, as discussed in Section IV-C.

TABLE IX
PERFORMANCE AS MEASURED USING THE AUC FOR A LOGRED MODEL AND XGBOOST FOR MIGRATION CLASSIFICATION FOR TWO PROVINCES. THE VALUES IN PARENTHESES ARE THE POPULATION COUNT IN THOUSANDS IN EACH PROVINCE. (a) WESTERN PROVINCE. (b) NORTHERN PROVINCE

(a)

Model	Migration distance threshold (km)			
Dataset	< 5	≥ 5	≥ 10	≥ 15
LogReg				
Entire	0.70 (107.7)	0.77 (75.2)	0.80 (61.7)	0.82 (54.0)
Restricted	0.77 (71.7)	0.79 (59.3)	0.82 (49.9)	0.83 (44.0)
XGBoost				
Entire	0.73 (107.7)	0.80 (75.2)	0.84 (61.7)	0.86 (54.0)
Restricted	0.80 (71.7)	0.83 (59.3)	0.85 (49.9)	0.86 (44.0)

(b)

Model	Migration distance threshold (km)			
Dataset	< 5	≥ 5	≥ 10	≥ 15
LogReg				
Entire	0.73 (32.7)	0.77 (28.7)	0.79 (24.2)	0.80 (22.4)
Restricted	0.80 (22.5)	0.81 (21.4)	0.82 (19.5)	0.83 (18.4)
XGBoost				
Entire	0.76 (32.7)	0.79 (28.7)	0.82 (24.2)	0.83 (22.4)
Restricted	0.83 (22.5)	0.84 (21.4)	0.85 (19.5)	0.85 (18.4)

To summarize, both LogRed and XGBoost perform well with a slight advantage for XGBoost. All trends are consistent across both methods. The accuracy is slightly higher for the following across both methods: 1) the restricted dataset of definite migrants and nonmigrants; 2) the Northern Province; and 3) long-distance home location changes. There is no performance improvement for the *balanced* dataset.

C. Feature Analysis

We report on the significance of features across the various experimental settings for the LogRed. To do so, we report on the odds ratio (OR) for each feature and setting. In LogRed, OR represents the effect of a predictor variable on the likelihood that the outcome will occur. In our migration prediction setting, the OR value allows us to explore the effect of calling patterns, social and spatial Diversities, and social relationships on the individual migration decision. Features with OR values greater than one will reveal a positive impact on the migration decision, whereas OR values less than one will reveal a negative impact on the migration decision, and OR values close to 1 reveal features with little to no impact on the migration decision. OR values can also be interpreted as unit increases, whereby a one-unit increase in a given feature increases the odds of a migration decision, as reflected by the change in the OR value. We note that interpreting the results is not straightforward for XGBoost since the coefficients have to be normalized across a random forest.

OR values for selected features across experiment subsets are reported in Table X. Recall that this includes the entire dataset, a restricted dataset of only definite migrants and

TABLE X
ORs. (a) WESTERN PROVINCE. (b) NORTHERN PROVINCE

(a)

Feature	Migration distance threshold (km)			
Dataset	< 5	≥ 5	≥ 10	≥ 15
avg_contact_distance				
Entire	1.280	1.240	1.226	1.225
Restricted	1.313	1.265	1.243	1.241
Balanced	1.572	1.543	1.580	1.660
cv_contact_rate				
Entire	1.109	1.138	1.151	1.158
Restricted	1.153	1.165	1.174	1.179
Balanced	1.172	1.187	1.200	1.204
stddev_unique_bts_count				
Entire	1.147	1.142	1.146	1.146
Restricted	1.242	1.230	1.229	1.228
Balanced	1.260	1.239	1.227	1.217
stddev_radius_of_gyration				
Entire	1.142	1.202	1.230	1.247
Restricted	1.192	1.222	1.244	1.259
Balanced	1.232	1.296	1.354	1.408
stddev_hl_spatial_entropy				
Entire	1.071	1.109	1.130	1.139
Restricted	1.095	1.114	1.131	1.138
Balanced	1.077	1.094	1.113	1.119
cv_prov_spatial_entropy				
Entire	1.081	1.080	1.076	1.075
Restricted	1.101	1.097	1.090	1.089
Balanced	1.121	1.122	1.119	1.122
ego_network_5_migrations				
Entire	1.199	1.242	1.248	1.252
Restricted	1.315	1.293	1.293	1.296
Balanced	1.347	1.316	1.317	1.320

(b)

Feature	Migration distance threshold (km)			
Dataset	< 5	≥ 5	≥ 10	≥ 15
avg_contact_distance				
Entire	1.420	1.352	1.362	1.367
Restricted	1.530	1.461	1.435	1.428
Balanced	1.782	1.681	1.674	1.684
cv_contact_rate				
Entire	1.136	1.158	1.175	1.182
Restricted	1.196	1.200	1.208	1.210
Balanced	1.237	1.244	1.256	1.262
stddev_unique_bts_count				
Entire	1.171	1.164	1.157	1.155
Restricted	1.289	1.268	1.248	1.243
Balanced	1.340	1.309	1.285	1.284
stddev_radius_of_gyration				
Entire	1.173	1.183	1.189	1.197
Restricted	1.224	1.222	1.226	1.228
Balanced	1.294	1.297	1.310	1.318
ego_network_5_migrations				
Entire	1.303	1.387	1.376	1.389
Restricted	1.534	1.555	1.501	1.507
Balanced	1.601	1.634	1.562	1.575

definite nonmigrants, and a balanced subsampled dataset of equal counts of definite migrants and definite nonmigrants. In addition, we use the migration distance thresholds of 5, 10, and 15 km.

1) Calling Patterns: Several calling patterns showed a significant positive impact on the likelihood of being a migrant, i.e., the OR values were greater than one. The *contact_distance* is the distance between the home BTSs of a subscriber and their contacts. Table X shows that, for a unit increase in the average contact distance, the OR value across settings takes values in a range from 1.225 to 1.782. Thus, subscribers with contacts living further away have an increased probability of migration. The OR values for this feature have values greater than one for the entire, restricted, and balanced datasets; the positive impact also holds across all of the distance thresholds.

The contact rate is the count of calls exchanged with a contact. The OR value for the *contact_rate* is positive and takes values in the range of 1.109–1.262 across the settings, i.e., those who have more intense interactions with contacts in their social network may be better positioned to migrate. We also observe that the value of the OR increases as the distance threshold increases and when moving from the entire dataset to a balanced dataset. This confirms that the more intense interactions with contacts have a significant positive impact, in particular when considering longer migration distances.

The radius of gyration is the distance between a home BTS and any visited BTS, weighted by visit frequency. Table X shows that a unit increase in the value leads to OR values in the range of 1.142–1.408. These numbers reflect that subscribers exhibiting greater mobility might be more likely to make a migration decision.

2) Social and Spatial Diversities: Two of the three novel entropy measures proposed have a small significant effect on the migration outcome, i.e., the OR values for the entire dataset are slightly above one. Small increases in the OR value are observed for the Western Province for distance thresholds of greater than 5 km. Specifically, for each unit increase in the home-location-based *hl_spatial_entropy*, the odds of being a migrant are in the range of 1.071–1.119. Hence, communications with subscribers who have diverse home locations contribute to an increase in migration. A similar outcome is observed for the other novel feature—the province-based *prov_spatial_entropy*; a unit increase results in OR values in the range of 1.081–1.122.

3) Social Relationships: A very significant feature is *network_5_migrant_count*. As explained in Table VIII, this feature measures the count of contacts who have already migrated and who remain in close contact. Close contact was defined as exceeding five weekly calls between the subscriber and the contact. Fig. 4 provides the distribution of the count for this feature, for migrants and nonmigrants, for the two provinces. As can be observed, for both provinces, the median count of this feature for migrants exceeds that for nonmigrants. Furthermore, there is less variance in the count for migrants. The median count for the Northern Province is higher, both for migrants and nonmigrants, in comparison with the Western Province. The high OR values for this feature reflect its strong positive impact on a subscriber's migration decision; the impact holds across migration distance thresholds and

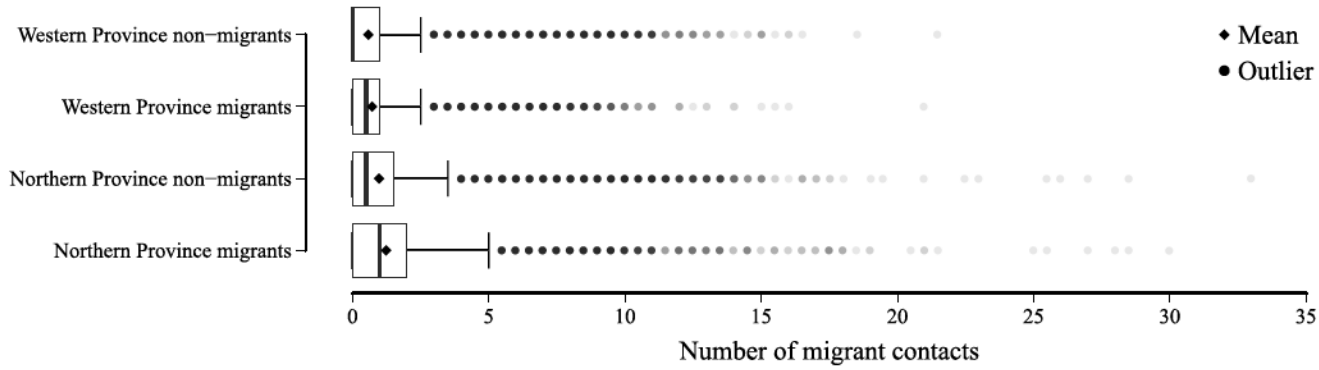

Fig. 4. Histogram of five call migrant counts for migrants and nonmigrants in the Western and Northern Provinces.

provinces. In the Northern Province, for each unit increase in close contact, the odds of migration are in the range of 1.303–1.601. For the Western Province, this feature has OR values in the range of 1.199–1.347. This result reflects that subscribers who maintain strong connections with other migrants are more likely to make the individual decision to migrate.

V. Conclusion, Limitations, and Future Work

A. Conclusion

This research is the first holistic end-to-end approach to repurpose CDR data to study migration. Our research identified the home locations of subscribers together with corresponding confidence measures. We identified definite migrants, likely migrants, definite nonmigrants, and likely nonmigrants. We created detailed maps and identified migrations patterns at the DSD level for two provinces in Sri Lanka. Our research is the first to predict the individual migration decision using CDR-based features. Of note is that the social relationship of staying in close contact with migrants is a strong indicator of future migration.

B. Limitations

Our dataset is from a single carrier circa 2013; at the time, the count of mobile cellular subscriptions was approximately 2.36 million with a total population of 20.32 million [42]. This may limit our coverage of the population, and it may be biased toward urban and mid- to high-income subscribers because of the barriers to cell phone adoption in low-income communities. Nevertheless, a GSMA report from 2013 revealed that competition between operators in Sri Lanka lowered prices, potentially giving access to higher numbers of low-income individuals [56]. Our approach to determine confidence in the home location filtered out subscribers that did not have significant call activity at night, or whose activity levels are low overall. We are, thus, likely to miss the behavior of migrants with low cellular usage patterns. While some providers collect data at higher frequencies (a.k.a. network data) that may cover subscribers with low activity, such data are rarely made available for research purposes. Our evaluation was specific to a single country, but it spanned two very different provinces based on demographic and socioeconomic metrics. We expect that our methodology to determine the confidence in the home location and to label the migration status will have similar performance accuracy across other datasets.

C. Open Challenges

We would first like to study migration patterns in depth. This includes differentiating long- and short-term migrants, circular migration patterns, and so on. We would like to extend our novel work on determining confidence in the home location to more precisely determine the actual migration window. This will enhance the granularity, and utility, of the migration maps. More important, it may be a valuable tool in predicting migration, as an earlier migration decision may have a later cascading impact through social relationships.

A holistic end-to-end framework for the CDR-based analysis must be based upon the architecture and infrastructure for contemporaneous processing that can provide close to real-time migration maps to social scientists and policymakers. Due to privacy reasons, CDR data are not freely available. Nevertheless, cell phone companies have developed numerous collaborations with academic and nonprofit partners, such as the World Bank and UNICEF to use CDR data in high-stake settings, including poverty [48] and health [49]. In addition, data challenges have also been proposed to give broader access to aggregated CDR datasets to the larger research community, e.g., the Syrian refugees challenge [31] or the D4D in Senegal [51]. Finally, we note that the COVID19 pandemic has spurred significant interest in the potential use of CDR data to monitor and predict community spread [52], [53].

Finally, the end-to-end framework will have to support the shipment of sensitive CDR data from providers and must satisfy multiple requirements around privacy. Although the CDR data are pseudonymized, under some (limited) circumstances, CDR data were reverse engineered to deidentify individuals [54]. GDPR-compliant approaches have been proposed recently to implement ethically founded CDR-based frameworks [55]. Recommendations from these studies could be incorporated into the proposed framework to transform it into a CDR-based GDPR-compliant framework. We will explore this transformation in future work.

REFERENCES

[1] S. Jiang, J. Ferreira, and M. C. González, "Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore," *IEEE Trans. Big Data*, vol. 3, no. 2, pp. 208–219, Jun. 2017.

[2] Y. Xu, S.-L. Shaw, Z. Zhao, L. Yin, Z. Fang, and Q. Li, "Understanding aggregate human mobility patterns using passive mobile phone location data: A home-based approach," *Transportation*, vol. 42, no. 4, pp. 625–646, Jul. 2015.

[3] Z. Zhao, J. Zhao, and H. N. Koutsopoulos, "Individual-level trip detection using sparse call detail record data based on supervised statistical learning," in *Proc. Transp. Res. Board Annu. Meeting*, 2016, pp. 1–18.

[4] G. Hugo, *What we Know About Circular Migration and Enhanced Mobility*, vol. 7. Washington, DC, USA: Migration Policy Institute, 2013.

[5] D. Akeju, "Africa, internal migration," in *The Encyclopedia Human Migration*. Hoboken, NJ, USA: Wiley, 2013.

[6] M. Bell, A. Bernard, E. Charles-Edwards, and Y. Zhu, *Internal Migration Countries Asia: Across-National Comparison*. Cham, Switzerland: Springer, 2020.

[7] R. Skeldon, *Migration Development: A Global Perspective*. Evanston, IL, USA: Routledge, 2014.

[8] S. Bertoli, J. Fernández-Huertas Moraga, and F. Ortega, "Crossing the border: Self-selection, earnings and individual migration decisions," *J. Develop. Econ.*, vol. 101, pp. 75–91, Mar. 2013.

[9] R. J. Nawrotzki, F. Riosmena, and L. M. Hunter, "Do rainfall deficits predict U.S.-bound migration from rural Mexico? Evidence from the Mexican census," *Population Res. Policy Rev.*, vol. 32, no. 1, pp. 129–158, Feb. 2013.

[10] E. Zagheni and I. Weber, "You are where you e-mail: Using e-mail data to estimate international migration rates," in *Proc. 3rd Annu. ACM Web Sci. Conf.*, 2012, pp. 348–351.

[11] B. State, I. Weber, and E. Zagheni, "Studying inter-national mobility through IP geolocation," in *Proc. 6th ACM Int. Conf. Web Data Mining*, 2013, pp. 265–274.

[12] E. Zagheni, V. R. K. Garimella, I. Weber, and B. State, "Inferring international and internal migration patterns from Twitter data," in *Proc. 23rd Int. Conf. World Wide Web*, Apr. 2014, pp. 439–444.

[13] Z. Tufekci, "Big questions for social media big data: Representativeness, validity and other methodological pitfalls," in *Proc. ICWSM*, vol. 14, 2014, pp. 505–514.

[14] V. Frias-Martinez and J. Virseda, "Cell phone analytics: Scaling human behavior studies into the millions," *Inf. Technol. Int. Develop.*, vol. 9, no. 2, pp. 1–35, 2013.

[15] V. Frias-Martinez, C. Soguero-Ruiz, E. Frias-Martinez, and M. Josephidou, "Forecasting socioeconomic trends with cell phone records," in *Proc. 3rd ACM Symp. Comput. Develop.*, 2013, pp. 1–10.

[16] E. Frias-Martinez, G. Williamson, and V. Frias-Martinez, "An agent-based model of epidemic spread using human mobility and social network information," in *Proc. IEEE 3rd Int. Conf. Privacy, Secur., Risk Trust, IEEE 3rd Int. Conf. Social Comput.*, Oct. 2011, pp. 57–64.

[17] S. Isaacman, V. Frias-Martinez, and E. Frias-Martinez, "Modeling human migration patterns during drought conditions in La Guajira, Colombia," in *Proc. 1st ACM SIGCAS Conf. Comput. Sustain. Societies*, Jun. 2018, pp. 1–9.

[18] J. E. Blumenstock, "Inferring patterns of internal migration from mobile phone call records: Evidence from Rwanda," *Inf. Technol. Develop.*, vol. 18, no. 2, pp. 107–125, 2012.

[19] S. Lai *et al.*, "Exploring the use of mobile phone data for national migration statistics," *Palgrave Commun.*, vol. 5, no. 1, p. 34, Dec. 2019.

[20] S. Isaacman *et al.*, *Identifying Important Places in Peoples Lives from Cellular Network Data* (Lecture Notes in Computer Science), vol. 6696, no. 6. Cham, Switzerland: Springer, 2011, pp. 133–151.

[21] J. Kennan and J. R. Walker, "The effect of expected income on individual migration decisions," *Econometrica*, vol. 79, no. 1, pp. 211–251, 2011.

[22] P. Bohra-Mishra, M. Oppenheimer, and S. M. Hsiang, "Nonlinear permanent migration response to climatic variations but minimal response to disasters," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 27, p. 9780, 2014.

[23] S. Li and P. Zhao, "Restrained mobility in a high-accessible and migrant-rich area in downtown Beijing," *Eur. Transp. Res. Rev.*, vol. 10, no. 1, Mar. 2018.

[24] C. Arul, H. A. Wilkin, A. Holley, and S. R. M. Hua, "International migrant workers' use of mobile phones to seek social support in Singapore," *Inf. Technol. Int. Develop.*, vol. 9, no. 4, pp. 1–19, 2013.

[25] L. Hong, J. Wu, E. Frias-Martinez, A. Villarreal, and V. Frias-Martinez, "Accuracy and bias in the identification of internal migrants using cell phone data," in *Proc. Int. Conf. Web Social Media (ICWSM) Workshop Making Sense Online Data Population Res.*, 2018.

[26] L. Hong, J. Wu, E. Frias-Martinez, A. Villarreal, and V. Frias-Martinez, "Characterization of internal migrant behavior in the immediate post-migration period using cell phone traces," in *Proc. 10th Int. Conf. Inf. Commun. Technol. Develop.*, Jan. 2019, pp. 1–12.

[27] A. J. Plantinga, C. Détang-Dessendre, G. L. Hunt, and V. Piguet, "Housing prices and inter-urban migration," *Regional Sci. Urban Econ.*, vol. 43, no. 2, pp. 296–306, Mar. 2013.

[28] M. J. Greenwood, "Research on internal migration in the United States: A survey," *J. Econ. Literature*, vol. 1975, pp. 397–433, Jun. 1975.

[29] E. Hopkins, F. Bastagli, and J. Hagen-Zanker, "Internal migrants and social protection: A review of eligibility and take-up," in *Proc. ODI Work. Paper*, 2016.

[30] S. Perera, "Internal migration in Sri Lanka," in *Internal Migration Countries Asia: A Cross-National Comparison*, M. Bell, A. Bernard, E. Charles-Edwards, and Y. Zhu, Eds. Cham, Switzerland: Springer, 2020.

[31] A. Salah *et al.*, "Introduction to the data for refugees (D4R) challenge on mobility of Syrian refugees in Turkey," in *Guide to Mobile Data Analytics Refugee Scenarios*. Cham, Switzerland: Springer, 2019, pp. 3–27.

[32] N. Eagle, M. Macy, and R. Claxton, "Network diversity and economic development," *Science*, vol. 328, no. 5981, pp. 1029–1031, 2010.

[33] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[34] S. Phithakkitnukoon, F. Calabrese, Z. Smoreda, and C. Ratti, "Out of sight out of mind-how our mobile social network changes during migration," in *Proc. IEEE 3rd Int. Conf. Privacy, Secur., IEEE 3rd Int. Conf. Social Comput.*, Oct. 2011, pp. 515–520.

[35] V. Frias-Martinez, J. Virseda-Jerez, and E. Frias-Martinez, "Socio-economic levels and human mobility," in *Proc. Qual. Meets Quantum Workshop (QMQ)*, 2010, pp. 1–6.

[36] V. Frias-Martinez, J. Virseda-Jerez, and E. Frias-Martinez, "On the relation between socio-economic status and physical mobility," *Inf. Technol. Develop.*, vol. 18, no. 2, pp. 91–106, Apr. 2012.

[37] E Unit, "Weekly epidemiological report," in *Report of the Ministry of Health, Nutrition and Indigenous Medicine*. Colombo, Sri Lanka: Ministry of Health, Government of Sri Lanka, 2017. [Online]. Available: http://www.epid.gov.lk/web/images/pdf/wer/2017/vol_44_no_05-english.pdf

[38] LIRNEAsia. (2018). *AfterAccess: ICT Access and Use in Asia and the Global South*. [Online]. Available: https://lirneasia.net/wp-content/uploads/2018/10/LIRNEasia-AfterAccess-Asia-Report.pdf

[39] S. Kumar, R. Zafarani, and H. Liu, "Understanding user migration patterns in social media," in *Proc. Conf. Artif. Intell. (AAAI)*, 2011, pp. 1204–1209.

[40] S. Wu, T. Elsayed, W. Rand, and L. Raschid, "Predicting author blog channels with high value future posts for monitoring," in *Proc. Conf. Artif. Intell. (AAAI)*, 2011, pp. 1261–1266.

[41] S. Wu and L. Raschid, "Prediction in a microblog hybrid network using bonacich potential," in *Proc. 7th ACM Int. Conf. Web Data Mining*, Feb. 2014, pp. 383–392.

[42] *GSMA Association*. Accessed: Jan. 2022. [Online]. Available: https://www.gsma.com/mobileeconomy/asiapacific/

[43] J. A. Groen and A. E. Polivka, "Going home after Hurricane Katrina: Determinants of return migration and changes in affected areas," in *Demography*, vol. 47, no. 4. Cham, Switzerland: Springer, 2010, p. 821.

[44] C. Kang, X. Ma, D. Tong, and Y. Liu, "Intra-urban human mobility patterns: An urban morphology perspective," *Phys. A, Stat. Mech. Appl.*, vol. 391, no. 4, pp. 1702–1717, Feb. 2012.

[45] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proc. 17th Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1082–1090.

[46] L. Hong and V. Frias-Martinez, "Modeling and predicting evacuation flows during hurricane irma," *EPJ Data Sci.*, vol. 9, no. 1, Dec. 2020.

[47] M. Vanhoof, W. Schoors, A. Van Rompaey, T. Ploetz, and Z. Smoreda, "Comparing regional patterns of individual movement using corrected mobility entropy," *J. Urban Technol.*, vol. 25, no. 2, pp. 27–61, Apr. 2018.

[48] M. Hernandez, L. Hong, V. Frias-Martinez, and E. Frias-Martinez, "Estimating poverty using cell phone data: Evidence from Guatemala," World Bank Policy Res. Working Paper, Tech. Rep., 2017.

[49] K. H. Jones, H. Daniels, S. Heys, and D. V. Ford, "Challenges and potential opportunities of mobile phone call detail records in health research: Review," *JMIR mHealth uHealth*, vol. 6, no. 7, p. e161, Jul. 2018.

[50] A. A. Salah *et al.*, "Introduction to the data for refugees challenge on mobility of Syrian refugees in Turkey," in *Guide to Mobile Data Analytics Refugee Scenarios*. Cham, Switzerland: Springer, 2019, pp. 3–27.
[51] Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D. Blondel, "D4D-senegal: The second mobile phone data for development challenge," 2014, *arXiv:1407.4885*.
[52] K. H. Grantz *et al.*, "The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology," *Nature Commun.*, vol. 11, no. 1, p. 4961, Dec. 2020.
[53] E. Knippenberg and M. Meyer, "The hidden potential of mobile phone data: Insights on COVID-19 in the Gambia," in *The World Bank Data Blog*. Washington, DC, USA: World Bank, 2020. [Online]. Available: https://blogs.worldbank.org/opendata/hidden-potential-mobile-phone-data-insights-covid-19-gambia
[54] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Sci. Rep.*, vol. 3, p. 1376, Mar. 2013.
[55] K. H. Jones, H. Daniels, S. Heys, and D. V. Ford, "Toward an ethically founded framework for the use of mobile phone call detail records in health research," *JMIR mHealth uHealth*, vol. 7, no. 3, Mar. 2019, Art. no. e11969.
[56] GSMA. (2013). *GSMA Intelligence. Country Overview: Sri Lanka*. [Online]. Available: https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2016/02/Country_Overview_Sri_Lanka.pdf
[57] E. Cesario, C. Comito, and D. Talia, "An approach for the discovery and validation of urban mobility patterns," *Pervas. Mobile Comput.*, vol. 42, pp. 77–92, Dec. 2017.
[58] C. Comito, "NexT: A framework for next-place prediction on location based social networks," *Knowl.-Based Syst.*, vol. 204, Sep. 2020, Art. no. 106205.
[59] M. S. Iqbal, C. F. Choudhury, P. Wang, and C. M. González, "Development of origin-destination matrices using mobile phone call data," *Transp. Res. C, Emerg. Technol.*, vol. 40, no. 1, pp. 63–74, Mar. 2014.
[60] C. Fu, G. McKenzie, V. Frias-Martinez, and K. Stewart, "Identifying spatiotemporal urban activities through linguistic signatures," *Comput., Environ. Urban Syst.*, vol. 72, pp. 25–37, Nov. 2018.
[61] V. Frias-Martinez and J. Virseda, "Cell phone analytics: Scaling human behavior studies into the millions," *Inf. Technol. Int. Develop.*, vol. 9, no. 2, 2013.
[62] J. Ghurye, G. Krings, and V. Frias-Martinez, "A framework to model human behavior at large scale during natural disasters," in *Proc. 17th IEEE Int. Conf. Mobile Data Manage. (MDM)*, Jun. 2016, pp. 18–27.
[63] C. Comito, D. Talia, and P. Trunfio, "An energy-aware clustering scheme for mobile applications," in *Proc. IEEE 11th Int. Conf. Comput. Inf. Technol.*, Aug. 2011, pp. 15–22.
[64] S. Nair, K. Javkar, J. Wu, and V. Frias-Martinez, "Understanding cycling trip purpose and route choice using GPS traces and open data," in *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, 2019, vol. 3, no. 1, pp. 1–26.
[65] S. Isaacman, V. Frias-Martinez, and E. Frias-Martinez, "Modeling human migration patterns during drought conditions in La Guajira, Colombia," in *Proc. 1st ACM SIGCAS Conf. Comput. Sustain. Societies*, Jun. 2018, pp. 1–9.
[66] J. Wu, L. Hong, and V. Frias-Martinez, "Predicting perceived level of cycling safety for cycling trips," in *Proc. 27th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2019, pp. 456–459.
[67] D. Castelvecchi, "Can we open the black box of AI?" *Nature News*, vol. 538, no. 7623, p. 20, 2016.

Viren Dias (Member, IEEE) received the M.Eng. degree in engineering science from the University of Oxford, Oxford, U.K., in 2014.

He is currently a Senior Researcher with LIRNEasia, Colombo, Sri Lanka, a nonprofit think tank.

Lasantha Fernando (Member, IEEE) received the bachelors' and master's degrees in computer science and engineering from the University of Moratuwa, Moratuwa, Sri Lanka, in 2013 and 2019, respectively. He is currently pursuing the Ph.D. degree in computer science with the University of Waterloo, Waterloo, ON, Canada.

His primary research interests include stream data processing systems and large-scale data management.

Yusen Lin (Member, IEEE) received the bachelors' degree in computer science from Xiamen University, Xiamen, China, in 2016, and the master's degree in electrical engineering from the University of Maryland, College Park, MD, USA, in 2021.

He is currently an Algorithm Engineer with Mesoor, China. His research expertise is in natural language processing and data science in finance, with a particular focus on neural machine translation, text to SQL, and pretraining models.

Vanessa Frias-Martinez (Member, IEEE) received the bachelor's degree in computer science from the University of Valladolid, Valladolid, Spain, in 1999, and the Ph.D. degree in computer science from Columbia University, New York, NY, USA, in 2008.

She is currently an Associate Professor with the University of Maryland, College Park, MD, USA. Her research focuses on the analysis of large-scale spatiotemporal data to model the interplay between human mobility patterns, social networks, and the physical environment, with the main objective of informing decision-making in areas such as transportation, natural disasters, or socioeconomic development.

Dr. Frias-Martinez was a recipient of the National Science Foundation (NSF) CAREER Award.

Louiqa Raschid (Fellow, IEEE) received the B.Tech. degree in electrical engineering from IIT Madras, Chennai, India, in 1980, and the Ph.D. degree in electrical engineering from the University of Florida, Gainesville, FL, USA, in 1987.

She is currently the Dean's Professor with the University of Maryland, College Park, MD, USA. Her research in data management and data sciences crosses multiple application domains, including biomedical applications to disaster information management to social media models to economic and financial ecosystems.

Dr. Raschid is also a fellow of the Association of Computing Machinery (ACM).