A Framework to Model Human Behavior at Large Scale during Natural Disasters

Jay Ghurye
Department of
Computer Science,
University of Maryland,
College Park, 20740
Email: jayg@cs.umd.edu

Gautier Krings
Real Impact Analytics
Brussels, Belgium
Email: gautier.krings@realimpactanalytics.com

Vanessa Frias-Martinez
College of
Information Studies,
University of Maryland,
College Park, 20740
Email: vfrias@umd.edu

***Abstract*—Natural disasters affect millions of people every year. Understanding human behavior is critical to improve both emergency planning and prevention. However, emergency responders typically struggle to gain access to timely, fine-grained models of human behavior during disasters. In this paper, we propose a novel framework to analyze behavioral changes during disasters using Call Detail Records (CDRs) from a telecommunications company. CDR datasets are collections of spatio-temporal traces that can characterize individual mobility and social network behaviors at very fine scales. The proposed framework exploits the granular behavioral models to evaluate the similarities and differences between the normal and the disaster patterns.**

The framework consists of three steps: data pre-processing, behavioral baseline computation and disaster analytics. The data pre-processing step uses data mining techniques to extract individual mobility traces and individual social network features from the CDR data. The behavioral baseline computation step computes the baselines that characterize normal mobility and social network behaviors in non-disaster scenarios. This step uses n-th order Markov Chain models to approximate mobility patterns from the CDR spatio-temporal data. Finally, the disaster analytics step allows for the statistical analysis of behavioral changes during disasters by comparing the real behaviors observed during a disaster with the behaviors that would have been expected under normal circumstances (baselines). We use the framework to analyze Rwanda's 2012 floods and show that disasters tend to disrupt both mobility patterns and communication behaviors while recovery times can take several weeks.

I. Introduction

Natural disasters such as hurricanes, floods or tornadoes affect millions of individuals every year. As a result, governments spend millions of dollars in emergency response allocating resources to mitigate the damages. Effective resource allocation requires a deep understanding of how humans react when a disaster takes place. However, gathering human behaviors at large scale during a disaster is not trivial. For example, some natural disasters like floods might generate temporary displacements or permanent relocations. Drawing a complete picture of such population mobility patterns is extremely difficult. Generally, emergency responders in the field gather data by interviewing affected individuals, but the coverage of these interviews can be pretty limited.

The widespread use of cell phones worldwide has allowed to model human behaviors at large scale through the use of Call Detail Records (CDR) [1]. CDRs are collected by cell phone companies for billing purposes every time a phone call is made or received. Each CDR contains information regarding the phone numbers involved in the communication, date, time and the location (as a pair of latitude and longitude) of the cellular towers that gave coverage to the service. As previous research has shown, CDRs can offer a detailed picture of how humans move and interact with each other [2]–[5]. In this paper, we propose a novel framework to automatically extract large-scale models of human behavior during disasters using CDRs. The main objective is to allow emergency responders understand how humans react to a disaster. The resulting behavioral models will provide valuable information not only to critically allocate resources once a disaster happens, but also to enhance emergency planning and prevention.

The proposed framework uses a combination of data mining, n-th order Markov Chain models and statistical analyses to infer normal mobility patterns and social network behaviors from CDR data and to automatically quantify behavioral changes regarding displacements and communication patterns when a disaster happens. Unlike previous work [6], our approach uses CDR data which is sparser (both temporally and spatially) than GPS data and thus more challenging in terms of accurate mobility inference. More importantly, it offers the advantage that the framework will be useful in emerging regions with very limited resources where GPS cell phones, let alone GPS collection systems, are a rarity; and where the high penetration rates of cell phones offer the opportunity of modeling mobility at large scale. The main contributions of the proposed framework are:

- A framework that uses a mixture of data mining and machine learning techniques to extract mobility and social network features from CDR data which are then used to automatically build baselines that characterize normal behaviors under non-disaster scenarios.
- A framework to automatically evaluate the statistically significant differences between the normal mobility and social behaviors of a population (pre-disaster) and the reactions observed during and after a natural disaster takes place.

- An evaluation of a real flood scenario in Rwanda using CDR data from the major telecommunications carrier. The evaluation provides insight information regarding significant changes in mobility and social network patterns. These insights will prove useful in understanding displacements qualitatively and quantitatively so as to improve emergency planning as well as in evaluating social network changes that could prove critical to design communication plans for the affected populations.

The rest of the paper is organized as follows: we first describe the general framework with its three steps. Next, we explain each step in depth: the data pre-processing to extract individual mobility and social network features from CDR data; the behavioral baseline computation step to compute baselines that characterize mobility and social network behaviors under normal circumstances exclusively using geolocation information from CDR data; and the disaster analytics that allows to automatically extract statistically significant behavioral changes in terms of displacements and communication patterns of a population during a disaster. Next, we describe our evaluation results using CDR data during the flood season in Rwanda, in 2012 and we finalize with a description of the related work and a discussion of our main conclusions.

II. Framework Overview

This paper presents a framework that automatically analyzes the behavioral response to a disaster using as input CDR data. The main objective is to help emergency responders understand human behaviors during disasters so as to improve their mitigation plans and resource allocation. Figure 1 shows the proposed framework. It receives as input a CDR dataset *i.e.,* a large-scale spatio-temporal series with millions of calls from individuals that live in a geographical area hit by a disaster. Initially, the dataset is divided into two groups: the pre-disaster dataset, which contains all CDRs until the day of the disaster and is used to compute the mobility and social network baselines, and the post-disaster dataset, with all the CDRs from the day of the disaster onwards which is used to quantify the behavioral differences with respect to a normal period of time. Given that input, the framework follows three main steps. Here, we present a brief overview. The next three sections describe each step in detail.

The *data pre-processing* step uses the CDR data to extract three pre- and post-disaster features that will be used throughout the framework, namely: mobility traces, ego-social network features and home location. The mobility traces are computed for each individual in the sample as the set of transitions between cellular towers and are used as an approximation of the mobility patterns for any given individual. The ego-social network features model various individual social network aspects including degree or volume of incoming and outgoing communications, reciprocity, transitivity and friendship. Pre-disaster mobility traces and ego-social network features are used in the second step of the framework to compute the behavioral baselines that characterize general mobility and social network patterns of the population under study. The third variable, home location, is critical to model behavioral changes with respect to the geographic distribution of the population. Since the CDR data does not contain any information regarding the home location of the individuals, we present a data mining approach to approximate it given a large-scale spatio-temporal series for a given individual.

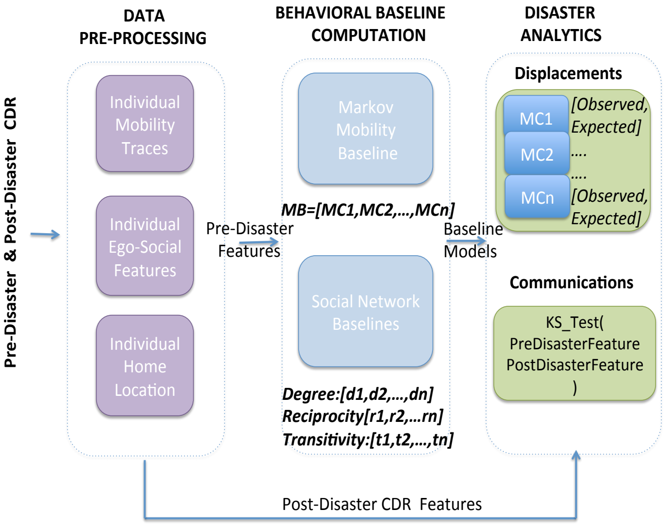

Fig. 1. Overview of the proposed framework with its three steps: CDR pre-processing to extract main mobility and social features; behavioral baseline computation to extract normal behaviors during non-disaster periods; and disaster analytics to evaluate statistical differences between normal, pre-disaster behavior and the behaviors observed during and after the natural disaster.

The *behavioral baseline computation* step computes the baselines that accurately represent normal mobility and social network behaviors during non-disaster periods. These baselines will be used in the disaster analytics step to infer the expected mobility and social behaviors under normal circumstances and compare them against the actual behavior observed during a disaster so as to measure behavioral changes. While individual social network behaviors can be easily represented as averages characterizing the degree, reciprocity or transitivity of the communications for each individual; mobility behaviors are much more complex to represent. For that reason, the framework computes one social network baseline per social feature (degrees, reciprocity and transitivity), where each baseline is a distribution containing the individual average feature values for all the population. For example, given a population of p individuals, the reciprocity baseline will be a distribution with p elements where each element is the individual average reciprocity computed using all the pre-disaster CDR data. On the other hand, the mobility baseline is represented as a set of n-th order Markov Chain models (one per individual) that allow to probabilistically infer, for each individual, the next location based on a set of previously visited locations. The framework automatically explores various n-th order Markov Chain models using both visited locations and social network information and selects the best n value based on the accuracy of the Markov Chain models in representing the population's normal behavior.

Finally, the *disaster analytics* step takes as input the behavioral baselines and the post-disaster features, and extracts statistically significant differences between the normal, expected behaviors and the actual behaviors observed after the disaster. The objective is to quantify the difference between expected and observed behaviors so as to provide critical information to emergency responders regarding displacements, communications and their spatial and temporal characteristics. For the displacements, the framework takes each individual n-th order Markov Chain model in the mobility baseline and uses them to infer what the expected location would be given a set of n previously visited locations by an individual. Displacements are measured by comparing the expected location against the observed one in the post-disaster period, and by repeating the process across all individuals and all their visited locations. For the communications, the framework runs statistical analyses to compare the baseline of each social feature against its actual distribution in the post-disaster period so as to understand the role that communications could play for the deployment and dissemination of mitigation services. Next, we explain each step in detail.

III. Data Pre-processing

In this section, we present the approaches used to extract the mobility traces, ego-social network features and home location for each individual both during the pre- and the post-disaster periods. The input to this step is a large-scale spatio-temporal CDR series where each element is of the type (I_i, I_j, T_i, T_j, t): I_i is an individual at location T_i calling individual I_j at location T_j at time t. The location of a cellular tower T_z is expressed as a pair $(latitude, longitude)$. There exist only a set of such possible locations determined by the number of cellular towers $T_1, ..., T_n$ that give coverage to the geographical area under study. Given such time series, the mobility traces MT_{I_z} of an individual I_z are computed as the set of observed transitions between cellular towers at different time stamps t:

$$MT_{I_z} = \bigcup_{T,t} (T_i, T_j, t)$$

On the other hand, given a set of individuals $I = \{I_1, ..., I_n\}$ in the large-scale spatio-temporal CDR series, we compute the ego-social network for a given individual I_i as the set of individuals with whom a communication has been established in the past. The framework computes the following features for each ego-network: input and output network degrees, reciprocity, number of friends and transitivity. Input (ID) and output (OD) network degrees represent the number of incoming and outgoing communications for a given individual i in her ego-network and are computed as:

$$ID_i = \sum_{I_j} V(i \leftarrow j)$$

where $V(i \leftarrow j)$ measures the number of calls from j to i, and the output degree as the total number of calls from i to others

$$OD_i = \sum_{I_j} V(i \rightarrow j)$$

We compute the friends for a given individual I_i as the set of individuals with whom a *strong, reciprocated* communication has been established in the past. Reciprocity measures the relationship between how much an individual calls other people and how much the other people reciprocate that behavior. The framework computes this variable as $R_i = \sum_{j=1}^{N} R_{i,j} \Big/ N$ where N is the total number of communicated individuals and

$$R_{i,j} = 1 - \left| 1 - \frac{2V_{i \rightarrow j}}{V_{i \rightarrow j} + V_{j \rightarrow i}} \right|$$

where $V_{x \rightarrow y}$ is the volume of calls made from individual x to individual y and $R_{i,j} \in [0, 1]$ [7] . However, reciprocity can be high between two individuals that have shared very few communications. Thus, we define friends as individuals that not only have a high reciprocity, but also whose communications represent an important percentage th of the overall communication graph. Formally, we define the set of friends of an individual I_i as:

$$F(I_i) = \bigcup_{j} I_j \ s.t. \ (R_i \geq r) \wedge (V_{i \leftrightarrow j} \geq V_i * th)$$

where $V_{i \leftrightarrow j}$ is the number of calls between I_i and I_j, V_i is the total volume of calls for I_i and where r and th are calibrated empirically. More details are presented in the evaluation section. The last network feature, transitivity, measures the number of connections that an individual has with her friends' friends [8] and is computed as:

$$T(i) = V_{i \leftrightarrow j} \ s.t. \ j \in F(F(I_i))$$

Finally, to approximate the home location of an individual, we use the centroid of the radius of gyration modeled using all the locations visited by an individual. The radius of gyration approximates the geographical area typically covered by an individual and we use its centroid as a proxy for the location of the home. Formally, given the set of towers $T_1, ..., T_n$ with coordinates $(lat_1, long_1), ..., (lat_n, long_n)$ and being n_i the number of times tower T_i is visited by a given individual, we compute the radius of gyration r for I_i as the deviation from the center of masses for each recorded position weighted by the number of times each location is visited:

$$r_{I_i} = \sqrt{\sum_{i=1}^{N} dist(C, T_i)^2}$$

where $dist(C, T_i)$ is the Euclidean distance between tower T_i, n_i is the number of times a location is visited and C is the centroid computed as:

$$c_{lat} = (\sum_{i=1}^{N} n_i * lat_i) \Big/ \sum_{i=1}^{N} n_i$$

and

$$c_{long} = (\sum_{i=1}^{N} n_i * long_i) \Big/ \sum_{i=1}^{N} n_i$$

IV. Behavioral Baseline Computation

This section presents the methods used to compute the mobility and social network baselines that characterize a population's normal, expected behavior in non-disaster scenarios.

A. Mobility Baseline

The mobility baseline represents the normal behavior of a population as a set of n-th order Markov Chain (MC) models [9], one per individual. Each n-th order MC is used to approximate the locations that an individual visits under normal circumstances, and constitutes a baseline for the mobility of a given individual. The framework automatically computes the best n value as the one that best approximates the mobility patterns for the population. The underlying assumption is that the current location of an individual depends on the previous n locations visited in the past. Formally, n-th order Markov Chain models are designed such that given a set of states S the probability of being at a given state depends on the previously n visited locations *i.e.,* $P(X_n = S_n) = P(X_{n-1} = S_{n-1}, X_{n-2} = S_{n-2}, ..., X_{n-m} = S_{n-m})$ for $n > m$. Given the mobility traces MT_{I_z} for each individual I_z the framework explores different n values for MC(n) to infer the next visited location (cellular tower T_{n+1}) based on the n previously visited locations (cellular towers $T_1, ...T_n$). For that purpose, the framework divides the individual mobility traces into randomly selected training and testing sets. The training set is used to compute each individual n-th order MC model while the testing set will be used to assess the accuracy of the models to infer the next visited location. The process is repeated multiple times for different randomly selected training and testing sets and the average accuracy across runs is used. The n value that most accurately approximates all the locations observed across all individuals will be used to represent the mobility baseline as the set of individual MC(n) models:

$$MB = \bigcup_{i=1,...,p} MC(n)_{I_i}$$

Each individual MC model is represented as a transition matrix. For a given n-th order, the matrix contains the frequencies at which each location is visited given a set of previous n locations. This transition matrix has as many rows as possible combinations with repetition (CR) of n previous towers T and as many columns as potential next locations (towers T):

$$MC(n)_{size} = [CR(T, n), T]$$

As such, the transition matrices tend to be sparse, suggesting that a large number of the potential transitions between locations do not happen *i.e.,* $P_{T_{n+1}, \{T_n, ..., T_1\}} = 0$, which is not necessarily true but rather a limitation of the model. To account for this, the framework applies the Laplacian smoothing technique to the transition matrices with $\alpha = 1$ [10]. To assess the accuracy of each individual transition matrix, the framework compares the inferred location, given a set of previously visited n locations, with the actual visited location observed in the testing set. The inferred location is the one with the highest frequency in the transition matrix for a given set of previous locations. This procedure is repeated for all transitions observed in the testing set while the MC(n) matrix is also updated in the process.

To select the best value for n, the framework measures the accuracy for each order across all users in the disaster area. We consider two scenarios: (i) *exact location*, where the next location prediction given a set of n previous locations, is correct if the inferred position is exactly the observed position in the mobility trace; and (ii) *top locations*, where the next location prediction, given a set of n previous locations, is correct if the observed position in the testing set is in the set of top 30% most frequently visited locations in the transition matrix. For a given n, the accuracy of the mobility baseline for the *exact location* is measured as the average accuracy across all individual MC models when the *exact location* is used:

$$ACC(MB) = \sum_{i=1,p} ACC(MC(n)_{I_i})/p$$

where $ACC(\cdot)$ represents the accuracy of an individual MC model inferring the next location and p is the total number of individuals in the sample. The accuracy of the MC(n) model for each individual is computed as the number of times the observed location in the testing set is equal to the inferred location using the individual $MC(n)_{I_i}$ model and the previous n locations *i.e.,*

$$ACC(MC(n)_{I_i}) = |T_z == MC(n)_{I_i}|$$

where $||$ is the cardinality of the set, T_z the observed location in the testing set and $MC(n)_{I_i}$ is the inferred next location using the n-th order transition matrix (which is the location with the highest frequency). Similarly, the accuracy for mobility baseline with the *top locations* approach given an order n is computed as the average accuracy across all individual MC(n) models where each individual accuracy is obtained as the number of times the observed location is in the set of towers at the top 30% based on the frequencies in the transition matrix:

$$ACC(MC(n)_{I_i}) = |T_z \in \bigcup_j MC(n)_{I_i}[, j]|$$

where $MC(n)[, j]$ represents all the locations j in the transition matrix whose frequencies are in the top 30%.

To explore enhancements to the accuracy of the individual MC models and thus the mobility baseline, we will incorporate the hypothesis that mobility patterns are also influenced by one's social network [11]. For that purpose, we propose to modify the next location inference by using not only one's own transition matrix, but also the transition matrices from all the friends. As a result, the prediction will be considered correct if the observed location, given a set of previous n locations, is the same as the inferred one using the MC model from the individual or any of her friends. Formally, the accuracy of a given $MC(n)_{I_i}$ with the *exact location* approach will be computed as

$$ACC(MC(n)_{I_i}) = |T_z \in \bigcup_{F_i}(MC(n)_{I_i}, MC(n)_{I_{F_i}})|$$

where T_z is the observed location in the testing set and $MC(n)_{I_{F_i}}$ are the inferred locations using the transitions matrices from i's friends F_i (where friends are defined as described in the previous section). For the *top locations* approach, the accuracy is computed as

$$ACC(MC(n)_{I_i}) = |T_z \in \bigcup_{F_i}(MC(n)_{I_i}, MC(n)_{I_{F_i}}[, j])|$$

where $MC(n)_{I_{F_i}}[, j]$ are the top 30% locations j across all of i's friends transition matrices. Once the best value for n is computed, the framework builds the mobility baseline as the set of MC(n) models, one per individual in the population under study.

Finally, we compare the accuracy of the mobility baseline against two models already proposed in the literature: (i) a *memoryless baseline* where each individual Markov Chain is of type MC(0) *i.e.,* the current state of an individual is independent of the previously visited states [12]; and (ii) a *time-based memoryless baseline* where each individual Markov Chain, TMC(0), considers the current state to be independent of the previously visited states but dependent of the date and time [13]. The MC(0) model infers next location always as the most frequent location visited by that individual across all training data and independently of previous visited locations. On the other hand, the TMC(0) model infers the next location as the most frequent one visited by that individual at a given day and time *i.e.,* for each day of the week and time of the day a most frequent location is given. As such, it represents a more granular memoryless model than the MC(0).

B. Social Network Baseline Models

To characterize the normal social network behavior of the population under study, the framework computes four different baselines, one per social network feature: input degree, output degree, reciprocity and transitivity. Specifically, each baseline is defined as a distribution where each element represents an individuals' average value for a given social network feature. Given the set of social measures per individual i: ID_i, OD_i, R_i, and T_i, the baselines are computed as follows:

$$\begin{aligned} IDB &= \{ID_1, ID_2, ..., ID_p\} \\ ODB &= \{OD_1, OD_2, ..., OD_p\} \\ RB &= \{R_1, R_2, ..., R_p\} \\ TB &= \{T_1, T_2, ..., T_p\} \end{aligned}$$

where p is the total population under study, ID_i is the average input degree for individual I_i in the population; OD_i is the average output degree for individual I_i; R_i is the average reciprocity for individual I_i; and finally, T_i is the average transitivity for individual I_i. The framework computes these distributions and compares them against the actual distributions observed in the post-disaster period to measure the behavioral differences as explained in the next section.

V. Disaster Analytics

In this section, we describe the automatic analyses that the framework performs to gain a better understanding of the behavioral changes that citizens undergo when a disaster happens. This step requires as input the mobility and social network baselines to infer normal behavior and the post-disaster spatio-temporal CDR to compare against and quantify the behavioral changes.

A. Displacements

The framework measures displacements using two variables: the distances between inferred (normal) and observed locations after the disaster (we will refer to it as $D1$) and the changes in the distribution of transition lengths ($D2$). While the first variable quantifies the general impact of the disaster in the mobility patterns *i.e.,* overall changes with respect to normal (inferred) behavior; the second variable characterizes specific types of changes in the mobility patterns with respect to distances travelled. Additionally, both variables are also used by the framework to provide a measure of disaster recovery *i.e.,* the amount of time it takes to recover the normal behavior that individuals had before the disaster happened.

The framework computes $D1$ as the average weekly distance between inferred and observed locations across all individuals. Observed locations are extracted from the post-disaster CDR features while the individual MC models in the mobility baseline are used to infer what the next visited location would be for an individual under normal circumstances given that she has visited n previous locations during the post-disaster period. Formally,

$$D1 = \sum_i \sum_{j=1}^{M} d(T_j, MC(n)_{I_i})/W$$

where I_i is an individual in the sample, T_j are the towers used by that individual each week after the disaster, $d()$ is a function that computes the Haversine distance for two pairs of coordinates, $MC(n)_{I_i}$ is the MC model for individual I_i and W is the number of weeks in the evaluation. The MC model infers what the normal location would be, given the set of previous visited locations, had the disaster not happened.

To compute $D2$, the framework extracts the distribution of the probabilities that transitions happen at different length ranges (R in miles) for a given period of time. Formally, it computes $\{P(d(T_i, T_j, t, t') \in R)\}$ where $R \in \{[0, 1\ mi), [1\ mi, 2\ mi), ..[n\ mi, (n+1)\ mi]\}$ and (T_i, T_j) represents a normal (inferred) or an observed transition during time period (t, t') and across all users I_i. After computing the probabilistic distribution for both observed and inferred transitions during a given period of time, the framework runs Kolmogorov-Smirnov (KS) statistical tests to evaluate the behavioral differences between each pair of distributions.

Finally, the framework also provides information regarding the destination of the displacements. Specifically, it evaluates three types of destinations: (i) friends' home locations, (ii) urban versus rural locations and (iii) new versus already visited locations. To evaluate the number of displacements that have a friends' home location as final destination, we use the home locations computed in the data pre-processing step. However, since destinations are extracted as cellular tower locations T_i, and home location is expressed as the centroid of the radius of gyration, the framework considers as friends' home the tower whose cellular coverage contains the home centroid. To model urban versus rural locations, the framework uses manually-provided boundaries surrounding the main cities in the disaster area under study and measures the number of displacements that fall within the boundaries (urban) or not. Finally, new locations are defined as those that have not been visited before the disaster by the individual.

B. *Communications*

Another relevant information for emergency planners is the ability to understand how communication patterns change among individuals when a disaster happens. Such information might shed some light into information diffusion techniques to reach out to affected individuals. In fact, understanding how people communicate could help authorities devise better communication plans to reach those who are in most need.

In this step, the framework takes as input the post-disaster CDR features and the social network baselines that represent the normal communication behaviors. Next, it builds one post-disaster distribution for each social network feature and runs KS statistical tests to evaluate whether there exist statistically significant differences between the pre- and post-disaster distributions for any the four social features: input and output degrees; reciprocity and transitivity. As discussed earlier, input and output degrees measure the volume of communications individuals have, reciprocity measures the relationship between how much an individual calls other people and how much the other people reciprocate that behavior, and transitivity measures the number of connections that an individual has with her friends' friends. We expect these variables will provide insights into how communications might change to reach out to others (higher input or output degrees or reciprocity) or to connect with friends' of friends (transitivity) when a disaster happens.

VI. Evaluation

In this section, we present an evaluation of the proposed framework using spatio-temporal CDR data from a telecommunications company in Rwanda. Rwanda suffers every year from heavy rains and floods specially in the northern province of Musanze. Heavy rains typically affect crops creating food insecurity; and can damage roads, schools and hospitals, generating a lot of disruptions in the lives of citizens. On April 12, 2012 heavy rains led to floods in Musanze provoking damages in thousands of households [14]. The results discussed here might help emergency responders gain a deep understanding of the behavioral patterns during the disaster and prepare for the floods to come in following years.

A. *Dataset*

We use a spatio-temporal CDR series for a temporal range from December 1st, 2011 to June 30th, 2012 covering all the communications in or out of the Musanze province. The dataset is fully anonymized *i.e.,* real cell phone numbers have been transformed into keys to preserve privacy. It contains for each recorded cell phone call the key for the caller and callee, the location of the cellular towers where the caller and callee were when the call happened (as pairs latitude, longitude) and the date and time at which the communication took place. The whole dataset contains approximately 1.5 billion records.

B. *Data Pre-processing*

First, we compute the mobility traces for each individual as the set of existing transitions between two continuous visited locations in the spatio-temporal CDR series. Figure 2 shows the Probability Density Function (PDF) for the transitions' lengths using the pre-disaster CDR data. We observe that, in general, the most probable transitions between any two given points are under 20 miles of length, which covers almost any trip distances in and between major cities in the north of the province: Ruhengeri, Butare or Mutura among others. However, there exists another peak with relatively high probability at around 40 miles which probably represents trips from Ruhengeri to the southern rural parts of the province which are the farthest away from the urban hubs in the north. Longer distances, possibly representing trips outside the province, are much less probable.

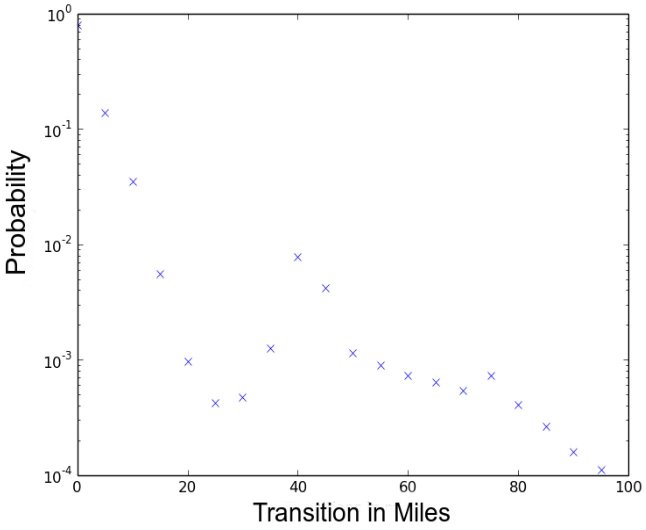

Fig. 2. Probability Density Function (PDF) for the pre-disaster transitions' lengths using CDR data. Two prominent peaks appear at $20mi$ and $40mi$. The first probably represents trips in and between major cities in the Musanze province ($20mi$) while the second is probably associated to trips between the capital city and the southern rural areas which are the farthest away from the capital city.

The framework also models the ego-social network for each individual extracting the input and output degrees, the reciprocity, number of friends and transitivity of the ego-social network. Further details are given in section VI-C2. Finally, the home location for each individual in the dataset is obtained as the centroid of the radius of gyration. Figure 3 shows an example of the home location computation for an individual who has used a total of 11 cellular towers throughout the seven months of the data. The size of the towers in the plot is proportional to the volume of activity that the user has had in each tower. We observe that this specific user has visited several towers in the northwestern part of the province with higher frequencies than other areas. As a result, the center of masses, and thus her home location, is approximated to be where the red dot lies.

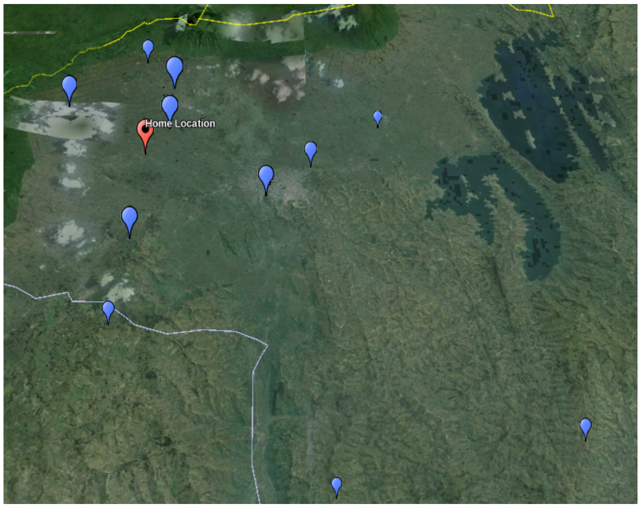

Fig. 3. Home location for one individual in our dataset. This individual visits a total of 11 distinct towers in the time period under study. However, the northern towers are visited the most (notice larger size in the plot) while the southern towers are visited with less frequency. Thus, the home location is located in the north, where the center of masses lies.

C. Behavioral Baselines

1) Mobility Baseline: Table I shows the accuracy of the mobility baseline for different n-th order MC models. The accuracy is measured as the average accuracy across all individual MC models using the pre-disaster spatio-temporal CDR features. All values are computed as the average results of multiple runs resulting from dividing the pre-disaster series into randomly selected training and testing sets. We only show results for orders $n = 1$ and $n = 2$ since larger orders gave mobility baseline accuracies lower than 10%. We discuss both mobility baseline accuracy (ACC) and error (ERR) for each MC order, where accuracy measures the average number of correct predictions and error the average distance between the inferred and the real locations across all predictions in the testing set. Results are discussed for both the Exact and the Top Locations (30%) approaches. For the latter, we also report the average distance ($Dist$) between all the towers at the top 30% as an approximation of the area covered by the set of towers that are considered correct inferences. The table also shows results for the other two existing mobility baselines: MC(0) and TMC(0).

TABLE I
ACCURACIES AND ERRORS FOR THE MOBILITY BASELINES USING MC(N) MODELS WITH N=1 AND N=2 AND THE TWO EXISTING BASELINES: MEMORYLESS AND TIME-BASED MEMORYLESS. RESULTS ARE SHOWN FOR BOTH EXACT-LOCATION INFERENCE AND TOP30% LOCATION INFERENCE.

Model	Exact		Top Locations (30%)			Friends
	ACC	ERR	ACC	ERR	Dist	ACC
MC(1)	40.21%	5.01mi	64.55%	3.50mi	6.69mi	66.44%
MC(2)	44.32%	2.96mi	72.06%	2.55mi	6.55mi	73.98%
MC(0)	44.47%	6.24mi	59.32%	1.70mi	7.41mi	60.4%
TMC(0)	22.91%	5.68mi	29.88%	2.45mi	7.18mi	30.13%

The table shows that the best mobility baseline accuracies are obtained using MC(2) models with values of 44.32% and 72.06% when measured via Exact position and Top 30%, respectively. These results show that considering the two previously visited locations considerably improves the next-location inference with respect to considering only one previous location. In fact, the accuracy for the mobility baseline using $MC(1)$ models where the next-location inference is solely based on the previously visited location decreases to 40.21% and 64.55% for Exact and Top 30%, respectively. As expected, the accuracy results for the Exact prediction are considerably lower than the Top 30% approach: from 44.32% to 72.06% in the best case. However, the errors for the Exact predictions are larger meaning that the Top 30% approach reduces the distance between the inferred and the real locations since it considers a set of potential candidates. Although one could argue that considering multiple locations always reduces the error since the chance of finding a closer location is higher, the average distance between any two locations of the sets considered in the top 30% experiments is $6.55mi$ which makes the errors reported considerably small given the location distribution.

We also observe that adding information from friends' MC models, increments the best mobility baseline accuracies by $\approx 2\%$ for any order n. In an attempt to better understand why friends' MC models appear to improve the next-location inference, we divided the users into those for whom the friendship approach increased the accuracy and those for whom the friendship approach did not. For each group, we computed average values for the following variables: number of locations (DT), number of contacts (DC), distance from friends (FD), number of friends (NF), radius of gyration (RG), total calls made (TC) and average distance (TD) from home to the locations visited (TD). Figure 4 shows that, in general, individuals for whom the next-location prediction improved using friends' information are users that tend to have higher volumes of calling interactions and larger mobility patterns, which probably enhances the amount of information available in the mobility models and thus the final accuracy in the prediction.

A comparison with the existing mobility baselines in the literature shows that the proposed framework enhances the mobility inference accuracy. In fact, the memoryless baseline were previous locations are not taken into account for the inference shows accuracies approximately 14% worse than the mobility baseline with MC(2) models and the Friends approach, which confirms the fact that previous locations help in predicting next visited locations. As for the time-based memoryless approach, the results are even worse which implies that inferring next location at specific times without taking into account previous visited locations is not a feasible solution. All the accuracies reported in the table were the best across different friendship threshold values $th \in \{1\%, ..., 20\%\}$. The value that gave the best accuracy results was $th = 10\%$ *i.e.,* a friend is someone with whom your reciprocal communication represents at least 10% of the total volume of your communication graph.

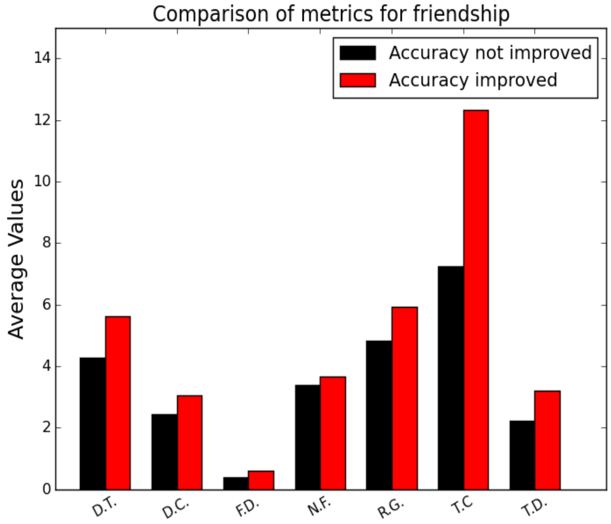

Fig. 4. Average variable values for individuals for whom the friends' baseline model improved prediction accuracies. The variables explored are: number of locations (DT), number of contacts (DC), distance from friends (FD), number of friends (NF), radius of gyration (RG), total calls made (TC) and average distance (TD) from home to the locations visited (TD).

2) Social Network Baselines: The four distributions IDB, ODB, RB, TB are computed across all individuals. The average input degree for the input degree baseline IDB is $\mu = 2.54$; $\sigma = 0.43$, while the average output degree in the ODB baseline is $\mu = 2.39$; $\sigma = 0.36$. The average reciprocity for the reciprocity baseline is $r = 0.8$; $\sigma = 0.008$. Finally, the average transitivity across all individuals in the transitivity baseline is $\mu = 0.26$; $\sigma = 0.67$ which indicates a low level of friendship between friends of friends.

D. Disaster Analytics

1) Displacements: Figure 5 shows the average weekly distances between inferred (normal) and observed locations after the floods (variable $D1$). We used the best mobility baseline: individual MC(2) models with friends' information, to infer the expected locations during the post-disaster period. The Figure shows that the largest displacements happened during the first three weeks after the disaster, reaching a maximum difference of $10mi$. This reveals that the floods had the greatest impact on changing the mobility patterns with respect to normal behavior mostly during the first three weeks. Weeks 3 and 4 also show average changes of up to $9mi$ after which the impact decreases down to $5mi$, which shows that the floods still generated mobility disruptions from the typical visited locations, although at a smaller scale. It is important to note that after 10 weeks we do not observe a trend towards zero. This indicates that for this flood, the recovery time for people to go back to their normal location visit patterns was longer than two months; which could be indicative of long-term or permanent displacements.

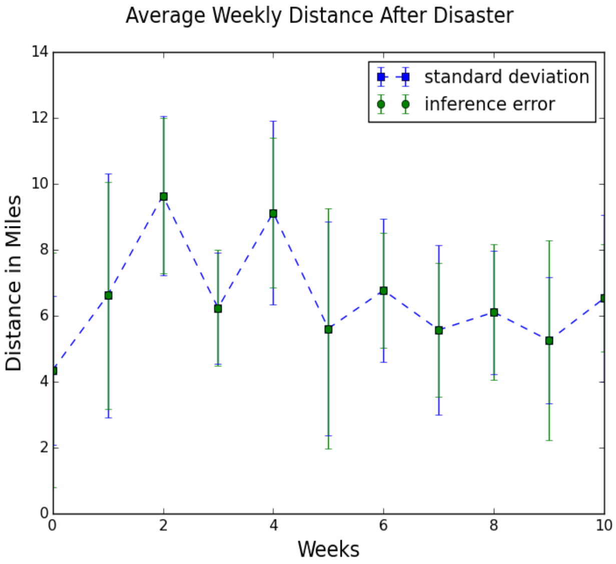

Fig. 5. Average weekly distances between inferred and observed locations after the disaster happens (variable $D1$). The bars show the average inference errors and the standard deviation of the variable D1.

Figure 6 shows the distributions of transition lengths between consecutive locations for the inferred (normal) and observed post-disaster locations as a probability density function (PDF) for approximately one, two, three and ten weeks after the floods. Since the normal PDFs for each time range were very similar, and for clarity purposes, we only plot one normal PDF (70 days later), although the statistical analyses were carried out comparing one-on-one inferred versus observed distributions. We observe that while for short transitions (up to $\approx 20mi$) there is not much change between the normal and the actual observed post-disaster behaviors; there exist significant changes for larger transitions ($\geq 20mi$). This result might indicate that despite the floods, people still moved locally at similar lengths, although to different places given the differences observed in variable $D1$. However, people reduced their large-scale trips significantly as can be observed 5, 10, 20 and 70 days after the disaster. Interestingly, we also observe a small recovery pattern in the long transitions

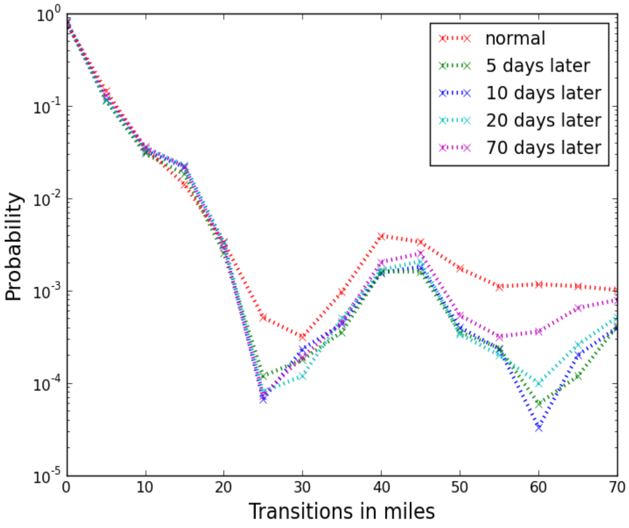

Fig. 6. Distribution of transition lengths between consecutive locations for the inferred (70 days later) and observed locations 5, 10, 20 and 70 days after the floods happened.

($\geq$ $20mi$) as days pass after the floods. In fact, 10 weeks (70 days) after the disaster, the longer transitions are more probable than before, although still not at normal levels. At 10 weeks, $D1$ still showed that individuals were not visiting the usual locations which means that although the long transitions might be recovering, some of these still represent trips to locations different than usual. All the differences discussed between inferred and observed transitions at different weeks were statistically significantly different with a KS test at $p < 0.01$ [15].

Regarding the types of destinations citizens went to when the flood happened, we observed that 26.67% of the displacements during the disaster had as destination an area where the home location of at least one friend was located. Additionally, approximately 13.98% of the displacements were to urban locations, and approximately 35.62% of the displacements observed during the post-disaster period were to new places that citizens had not visited in the past. Overall, the main behavioral trends reveal that individuals stayed more in rural areas, mostly going to known places (presumably not affected by the floods) and where they had few friends.

2) Communications: To investigate the effect of the floods on the communication patterns between users, we explore the statistically significant differences between the degrees, reciprocity and transitivity distributions before and after the floods took place. Regarding the input and output degrees, a KS test between the baselines and the post-disaster series gave a statistical significant difference between the two with $p < 0.01$. Overall, the communication behaviors changed after the floods happened with an increase in the average incoming calls (input degrees) from 2.38 to 2.46; and a decrease in the average outgoing calls (output degrees) from 2.42 to 2.34.

To look more in depth into these changes, Figure 7 shows the weekly input degree and output degree for 20 weeks before and 10 weeks after the floods. It can be seen that both incoming and outgoing volumes of calls are higher than normal mostly for the first two weeks after the floods happened. This increase reflects that right after the floods individuals are both reaching out and being reached out by their contacts either to seek help or to let others know they are fine. After that, there is a decrease in both values specially in the output degree which goes lower than average values before the floods. The affected individuals appear to cut down their outgoing communications at least for several weeks after the floods. This finding reveals that although emergency responders could use cell phone communication to reach out to affected individuals, it might be difficult to spread the word given that individuals appear to be reaching out to their contacts less that usual (decrease in the average output degrees). On the other hand, we did not observe any significant differences in terms of reciprocity or transitivity. As a result, emergency responders will also have difficulties in spreading information across friends' networks, although individuals appear to be responsive to incoming communications.

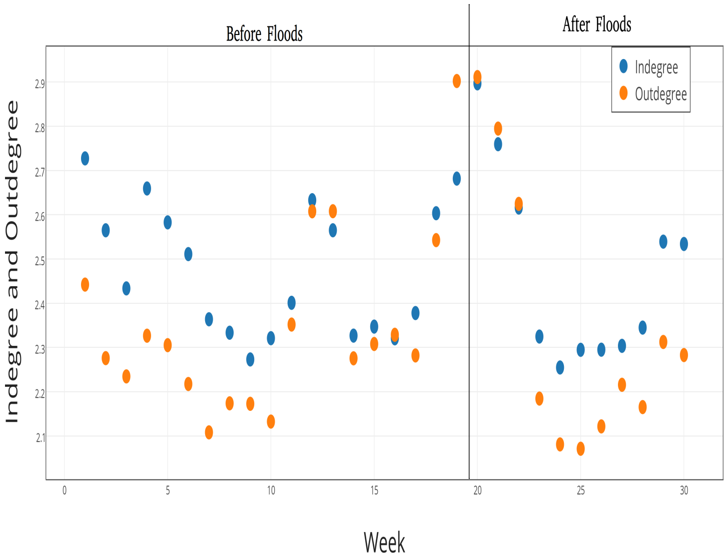

Fig. 7. Average weekly input degree and output degree before and after the floods in Musanze province. We observe a statistically significant increase in the average input degrees and a decrease in the average output degrees right after the floods.

VII. Related Work

From a disaster analytics perspective, there exists an important body of work that uses Call Detail Records to analyze disasters. Moumny *et al.* explored communication patterns during an earthquake [16] using communication data from cellphones; Bengsston *et al.* analyzed aggregated displacements upon an earthquake in Haiti [17]; and Song *et al.* provide an intelligent system to infer mobility patterns during disasters. Pastor *et al.* and Morales *et al.* carried out an aggregated analysis studying general mobility trends upon floods [18], [19]. A more general analysis by Bagrow *et al.* compared the collective response to large-scale emergencies such as bombings, blackouts, earthquakes and big festivals, among others [20]. Our paper expands the state of the art by moving beyond mere behavioral analysis contributing with a general framework that allows for the extraction of behavioral trends upon different types of disasters based on inferred behaviors rather than simply using the observed data.

From a baseline computation and inference perspective, various mobility and social network data such as check-in locations, GPS traces or CDR data have been use to analyze mobility patterns. Liu *et al.* or Noulas *et al.* have used check-in data from social networking to model user behavior in urban areas [21]–[23]; while Becker *et al.* use cellular network data to analyze daily range of travel, carbon footprints and traffic volumes [24]. Cellular data has also been in determining important places in users lives based on large population [25]. The effect of users' ego networks on their own behavior has also been studied in detail. Sadilek *et al.* proposed a probabilistic model to predict location of a user based on the locations of his friends [26]; while Cho *et al.* studied location prediction based on friends' users movements [27].

VIII. Conclusions and Future Work

Understanding human behavior upon disasters is critical to improve both emergency planning and prevention. However, emergency responders typically struggle to gain access to individual models of human mobility and communication. In this paper, we have proposed a novel framework to analyze human behavior during disasters using large-scale spatio-temporal CDR series. The framework extracts behavioral features from the spatio-temporal data to then infer general behavioral baselines that are used to study behavioral changes during disasters. As such, the framework can be used for any type of disaster given that the spatio-temporal datasets are provided for the duration of the disaster. Our evaluation has shown that the framework can generate valuable information both in terms of understanding displacements as well as for evaluating communication changes.

References

[1] S. Isaacman, R. Becker, R. Caceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger, "Human Mobility Modeling at Metropolitan Scales," *10th ACM International Conference on Mobile Systems, Applications and Services (MobiSys 2012)*, 2012.

[2] R. Becker, R. Caceres, K. Hanson, J. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, "A Tale of One City: Using Cellular Network Data for Urban Planning," *IEEE Pervasive Computing* , 2010.

[3] V. Frias-Martinez, C. Soguero-Ruiz, E. Frias-Martinez, and M. Josephidou, "Forecasting socioeconomic trends with cell phone records," in *Proceedings of the 3rd ACM Symposium on Computing for Development*, ser. ACM DEV '13, 2013.

[4] R. Lambiotte, V. Blondel, K. C., H. E., P. C., S. Z., and P. Dooren, "Geographical dispersal of mobile communications networks," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 21, 2008.

[5] R. Ahas, A. Aasa, Y. Yuan, M. Raubal, Z. Smoreda, Y. Liu, C. Ziemlicki, M. Tiru, and M. Zook, "Everyday spacetime geographies: using mobile phone-based sensor data to monitor urban activity in harbin, paris, and tallinn," *International Journal of Geographical Information Science*, vol. 29, no. 11, pp. 2017–2039, 2015.

[6] X. Song, Q. Zhang, Y. Sekimoto, and R. Shibasaki, "Intelligent system for urban emergency management during large-scale disaster," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[7] L. Gao, C. Song, Z. Gao, A.-L. Barabási, J. P. Bagrow, and D. Wang, "Quantifying information flow during emergencies," *Scientific reports*, vol. 4, 2014.

[8] T. Q. Phan and E. M. Airoldi, "A natural experiment of social network formation and dynamics," *Proceedings of the National Academy of Sciences*, vol. 112, pp. 6595–6600, 2015.

[9] W. R. Gilks, *Markov chain monte carlo*. Wiley Online Library, 2005.

[10] D. A. Field, "Laplacian smoothing and delaunay triangulations," *Communications in applied numerical methods*, vol. 4, no. 6, 1988.

[11] M. Musolesi and C. Mascolo, "Designing mobility models based on social network theory," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 11, no. 3, pp. 59–70, 2007.

[12] X. Lu, L. Bengsston, and P. Holme, "Predictability of population displacement after the 2010 haiti earthquake," *Proceedings of the National Academy of Sciences*, vol. 109, 2012.

[13] E. Frias-Martinez, G. Williamson, and V. Frias-Martinez, "An agent-based model of epidemic spread using human mobility and social network information," *IEEE International Conference on Social Computing*, 2011.

[14] ReliefWeb, "Thousands affected by heavy rains and floods in Rwanda," http://reliefweb.int/report/rwanda/thousands-affected-heavy-rains\newline-and-floods-rwanda, 2015, online;Accessed:2015-08-20.

[15] D. Eadie, W.T.and Drijard, F. E. James, M. Roos, and B. Sadoulet, "Statistical methods in experimental physics," p. 269271, 1971.

[16] B. Moumni, V. Frias-Martinez, and E. Frias-Martinez, "Characterizing social response to urban earthquakes using cell-phone network data: The 2012 oaxaca earthquake," in *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, ser. UbiComp '13 Adjunct. New York, NY, USA: ACM, 2013, pp. 1199–1208. [Online]. Available: http://doi.acm.org/10.1145/2494091.2497350

[17] L. Bengtsson, X. Lu, A. Thorson, R. Garfield, and J. von Schreeb, "Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in haiti," *PLoS medicine*, vol. 8, no. 8, p. e1001083, 2011.

[18] D. Pastor, A. Morales, and et al., "Flooding through the lens of mobile phone activity," *IEEE Global Humanitarian Technology Conference (GTHC)*, 2014.

[19] A. Morales, D. Pastor, and et al.., "Studying human behavior through the lens of mobile phones during floods," *Netmob*, 2015.

[20] Bagrow, J.P. and Wang, D. and Barabasi, A.L., "Collective Response of Human Populations to Large-Scale Emergencies," *PLoS ONE*, 2010.

[21] Y. Liu, Z. Sui, C. Kang, and Y. Gao, "Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data," *PloS one*, vol. 9, no. 1, p. e86026, 2014.

[22] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "An empirical study of geographic user activity patterns in foursquare." *International Conference on Web and Social Media*, vol. 11, pp. 70–573, 2011.

[23] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo, "A tale of many cities: universal patterns in human urban mobility," *PloS one*, vol. 7, no. 5, p. e37027, 2012.

[24] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky, "Human mobility characterization from cellular network data," *Communications of the ACM*, vol. 56, no. 1, pp. 74–82, 2013.

[25] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in peoples lives from cellular network data," *Pervasive computing*, 2011.

[26] A. Sadilek, H. Kautz, and J. P. Bigham, "Finding your friends and following them to where you are," in *ACM WSDM*. ACM, 2012.

[27] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1082–1090.