Mobility-aware COVID-19 Case Prediction using Cellular Network Logs

***Abstract*—In this paper, our goal is to model the aggregate mobility of individuals in a city by analyzing cellular network connections, and then leverage the designed mobility model to model and predict the number of COVID-19 infections in future. We collaborate with one of the main cellular network providers in Brazil, and collect and analyze cellular network connections from 973 antennas for all users in the city of Rio de Janeiro and its suburbs from April 5, 2020 to July 2, 2020. To aid implementation of our mobility-aware prediction model by government officials, we use the geographical municipal categorization of Rio and design a Markovian model that captures the mobility across different municipalities. We determine the transition probabilities of the Markov chain by analyzing mobility events at the user-level between antennas from the cellular network connectivity logs. We then combine the aggregate mobility characteristics across municipalities as obtained from the transition probabilities with the number of reported COVID-19 cases in a municipality during a particular week in the design of our mobility-aware COVID-19 case prediction model to predict the number of cases for the following week. We conduct experiments and observe that the steady state and transient performance of the Markovian model matches closely with those observed from the actual logs. Having empirically validated the performance of the Markovian model, we then compare the performance of linear and polynomial versions of our mobility-aware model with a mobility-agnostic linear regression model and demonstrate that our models significantly outperform the baseline model in terms of metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Overall, our experiments demonstrate that incorporating mobility can help in superior prediction of future COVID-19 infections and can be used by authorities to enforce targeted regional lockdowns to mitigate the spread of the disease instead of widely unpopular blanket lockdown measures.**

I. Introduction

COVID-19 is a global pandemic that has infected human beings in all countries of the world. At the time of writing this paper, COVID-19 is the leading cause of death in the United States and multiple countries in the world have been placed in a lockdown to combat the spread of the virus. While there is research on how the virus spreads, there is limited work on understanding the impact of mobility on the rate of infections, which is one of the reasons blanket lockdowns/closures are still being widely used as the primary measure to combat the spread of the virus.

In this paper, our goal is to present a simple, yet effective approach to understand the relationship between mobility and infection rates at the region level in a city. Our methodology follows a data-driven approach to investigate and model the mobility of individuals in a city by analyzing cellular network connections, and then leverages the designed mobility model to model and predict the number of COVID-19 infections in the future. To this end, we partner with TIM Brazil, one of the largest cellular network providers in Brazil to collect anonymized cellular network connection logs (i.e., 3G/4G connections, text messages, calls) for all users in the city of Rio de Janeiro for every 5 minute interval from 973 antennas between April 2020 and July 2020 to discern human mobility patterns. We also use publicly available COVID-19 infection data for the different municipal administrative regions (27 in all) of Rio de Janeiro. By designing a mobility-aware COVID-19 prediction model and demonstrating its efficacy via a large scale study consisting of a total of over 10 billion connections spanning over 2 million users per day in the second most populous city of Brazil, our goal is to develop a robust tool that can be easily used by government authorities around the world to curb the spread of the disease.

We first identify mobility events (i.e., movement of an individual from one antenna to the next) by perusing through the connection logs. We then use these mobility events to design a Markovian model that succinctly captures the aggregate movement of users. To avoid state space explosion of the Markovian model and to also aid the implementation of our mobility-aware COVID-19 prediction model, we consider the municipalities of Rio and associate the states of our Markov chain to the various municipalities. Mobility events between two antennas in the same municipality correspond to a same state transition while mobility events between two antennas in different municipalities correspond to transition between different states in the Markov chain.

We analyze the mobility events for each week to determine the probabilities of the transition matrix for that week. We then design mobility-aware prediction models that uses the transition probabilities encoding the mobility between a source and destination regions and combines it with the corresponding number of infections in the source to predict the number of infections in the destination for the next week. We design two versions of our predictive model—a simple linear model, and a more complex model that considers higher order polynomials of the past variables to predict the number of future infections.

We conduct extensive experiments to *i)* first validate the efficacy of the Markovian model, and *ii)* demonstrate the capability of our mobility-aware COVID-19 prediction model that uses the transition probabilities of the Markovian model. We compare the steady state and transient state probabilities of our model with the observed distributions obtained from the data and observe that the results match closely, which validates the effectiveness of our Markovian mobility model. We com-

pare the linear and polynomial versions of our mobility-aware COVID-19 case prediction model with a baseline mobility-agnostic linear regression model that just uses the number of past cases in a particular region to predict the number of future cases and observe that our models significantly outperform the baseline model in terms of metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Relative Error (RE). We also observe that though the mobility-aware linear model provides better interpretability, the polynomial models provide better prediction performance as they can better capture the underlying relationships in the data.

Our analysis, mobility, and infection prediction models arms governments and administrative authorities with the right set of tools to design effective policies to contain the pandemic. Our approach can be easily scaled and extended to other cities, states, and countries of the world, is non-intrusive to user privacy, and can be implemented with ease by government agencies to determine which sub-regions of any geographical unit need attention. Instead of adopting widely unpopular blanket lockdown measures, our mobility-aware models can help officials investigate the correlation between human mobility and rising number of infections and design lockdown policies at the regional level. Our models also lay the foundation for designing advanced traffic control patterns at the city level so that the underlying human mobility pattern can be altered to further mitigate the spread of the disease.

II. Related Work

In this section, we outline research in the areas of mobility modeling and COVID-19 that is closely related to our work. As the primary goal of this paper is to understand the connection between human mobility and the spread of COVID-19 by analyzing cellular network connectivity data and to use it to predict the number of future cases, we discuss research related to mobility prediction and modeling, and mining cellular network traffic in the context of human mobility.

Analyzing user mobility in cellular networks [15], inferring mobility-traffic interactions based on WLAN traces [3], and modeling semantic-rich human mobility using hidden Markov models [19] are a few examples of related research focused on mobility modeling and analysis. Cao et. al. investigate mobility predictability from WLAN logs [6], while a multi-dimensional approach to mobility prediction is adopted in [4]. Additionally, in recent years various machine learning approaches have been developed for mobility prediction [18], [11], including deep learning models such as attention-based recurrent neural networks [9].

With COVID-19 spreading rapidly in countries around the world, researchers have also investigated the impact of COVID-19 on human behavior and mobility. Gao et al. [10] map county-level mobility pattern changes in the US in response to COVID-19. The impact of the COVID-19 pandemic on mobile network operator traffic is outlined by Lutu et al. [14]. Huang et al. [13] also analyze the impact of the COVID-19 pandemic on transportation-related behaviors using human mobility data. Recent research has also investigated the impact of government measures such as lockdowns and decreased human mobility on COVID-19 related deaths in the UK [12]. Arimura et al. [5] evaluate the changes in urban mobility in Sapporo city, Japan due to COVID-19 emergency declarations, while Feldman et al. demonstrate the impact of lockdown measures on internet traffic [8]. In comparison, Chang et al. explain higher COVID-19 infection rates among disadvantaged groups using mobility network modeling [7]. Researchers have also developed a variety of machine learning models for predicting the number of COVID-19 cases [2], [17]. Additionally, Achterberg et al. [1] and Pizzuti et al. [16] develop and compare network-based prediction models for understanding the COVID-19 outbreak.

III. Background on Rio de Janerio

In this section, we first discuss the geography and socio-economic distribution of people in the city of Rio de Janeiro. We use the municipal administrative characterization of Rio in the design of our mobility-aware COVID-19 prediction models. Rio de Janeiro is the second most populous city of Brazil and it has 33 administrative regions or municipalities. In our analysis, however, we consider a total of 27 municipalities, we merge some regions with their closest neighbors, due to absence of sufficient number of antennas in some regions.

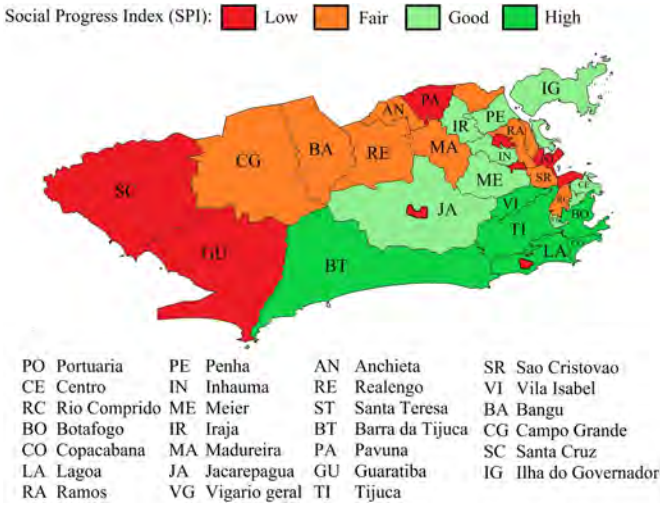


Fig. 1: The Municipal Administrative Regions of Rio de Janeiro and their SPI

Figure 1 shows the municipalities of Rio, color-coded according to their Social Progress Index (SPI), a metric that captures the socio-economic situation of people. Like most cities around the world, Rio has neighborhoods that are more prosperous than others and have higher SPI. The central zone of Rio is the main business district of the city and consists of administrative regions such as CE (Centro), ST (Santa Terasa), RC (Rio Comprido), SR (Sao Cristovao), and PO (Portuaria).

The northern zone consists of regions such as VI (Vila Isabel), ME (Meier), AN (Anchieta), and TI (Tijuca). The north zone contains roughly 40% of the population, a vast majority of whom can be categorized as middle-class residents.

The south zone is the wealthiest and most expensive area of the city. It consists of coastline regions such as LA (Lagoa), CO (Copacabana), and BO (Botafogo). The residents in this zone account for 10% of the entire population of the city. The western zone is the largest and most populous zone of the city. It houses more than 40% of the entire population of the city. Some regions in the west zone have low SPI such as SC (Santa Cruz), CG (Campo Grande), while other regions such as BT (Barra da Tijuca), JA (Jacarepagua) have higher SPI.

IV. Data and Methods

In this section, we provide an overview of the two datasets we analyze in this study. We collaborate with one of the largest cellular network providers of Brazil, TIM Brazil, and collect cellular network connection logs for all users in the city of Rio. We also use publicly available COVID-19 infection data for the different municipalities of Rio. Our objective is to first understand the aggregate mobility of people during COVID-19 using the cellular network connectivity dataset and to then effectively leverage the two datasets to design mobility-aware COVID-19 prediction models.

A. Cellular Network Connectivity Data

This data consists of anonymized cellular network logs (i.e., phone calls, text messages, 3G/4G data connections) of users along with the information of the specific antennas through which the connections are established. Table I shows a couple of examples of connection logs. Each log represents a connection made by a user (denoted by a hash) along with the location of the antenna (latitude and longitude) at a specific time. We collect data from April 5^{th} to July 2^{nd} and our dataset consists of more than 10 billion logs for this entire time period. As Brazil imposed strict lockdowns from March 15 to June 1, our data starts from the 4^{th} week of the lockdown and includes data for 5 weeks after the lockdown ends. Our dataset consists of approximately 2 million users per day and as we collect data from the entire city of Rio, we record data from 973 antennas.

Because our aim is to utilize human mobility to accurately predict the number of COVID-19 cases, we first identify mobility events from the network logs. If a user moves from one antenna to another antenna with different timestamps, we consider this to be a mobility event. For example, if a user connects to antenna 1 on April 05 at time 21.45.00 and then connects again to antenna 2 on the same day at the time 22.30.00, we consider this movement to be a single mobility event. We repeat this process for each user to obtain all the mobility events.

TABLE I: Cellular Network Connectivity Dataset

Timestamp	User ID	Latitude	Longitude
timestamp-1	hash-1	-23.003431	-43.342206
timestamp-2	hash-2	-22.8415	-43.278389

Figure 2 represents the variation in the number of total connections and the number of mobility events over weeks. The number of connections (i.e., the blue line), is shown on the left y-axis and the number of mobility events (i.e., the red line) is shown on the right y-axis. The vertical dotted line in the figure is the day when Brazil eased its lockdown (June 1), which corresponds to beginning of week 9 in our analysis. We observe that the number of connections and the number of mobility events per week increase significantly even before the lockdown is eased. The average number of connections per week increases from ∼341 million in the during-lockdown period to ∼365 million in the after-lockdown period. We observe a similar trend in the number of mobility events. The average number of mobility events per week rises from ∼85 million in the during-lockdown period to ∼99 million in the after-lockdown period.

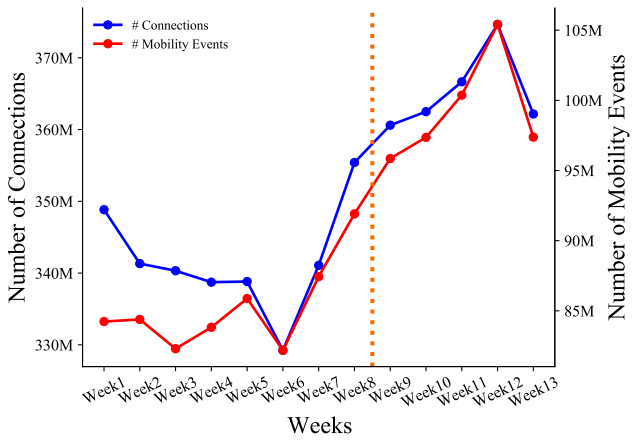


Fig. 2: The number of connections and mobility events per week from April 5^{th}, 2020 to July 2^{nd}, 2020

B. COVID-19 Case Data

The second dataset we use in this study provides us the number of positive COVID-19 cases for each municipality on a daily basis. Table II provides some illustrative examples for this dataset. Each row in the table indicates a single positive COVID-19 case. The latitude and longitude represents a location in the region where the positive case is detected. The region codes in the table correspond to regions shown in Figure 1. For instance, the first row corresponds to a positive COVID-19 case in the Campo Grande region (region code 144) on a specific date (as denoted by the timestamp). Though we have COVID-19 cases data from March, we primarily consider data from April 5^{th} to July 2^{nd} as this is aligned with our cellular network connections data.

TABLE II: COVID-19 Case Dataset

Timestamp	Latitude	Longitude	Region Code
timestamp-1	-22.888272	-43.552508	144
timestamp-2	-22.898441	-43.223156	10

Figure 3 shows the number of confirmed COVID-19 cases per day from March 12 to October 31. Orange dashed lines on March 16 and June 1 represent the beginning and end of the lockdown, respectively. We observe from the figure that the average number of cases per day in the during-lockdown

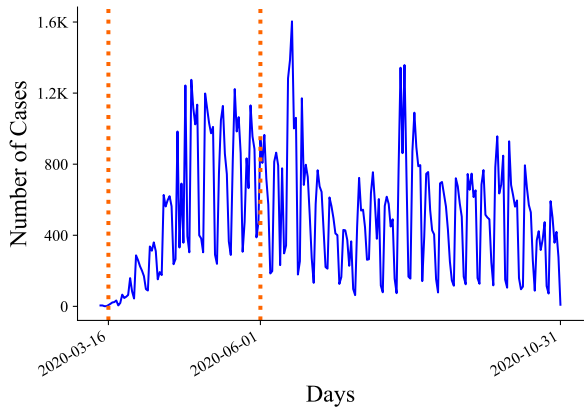


Fig. 3: The number of cases per day from March 12^{th}, 2020 to October 31^{st}, 2020

period is 516.2 and it increases to 686.8 in the after-lockdown period. The average number of cases in a day increases 33% after lockdown restrictions are eased.

V. Mobility-Aware COVID-19 Case Prediction Models

In this paper, our goal is to design mobility-aware COVID-19 case prediction models to predict the number of future infections and to better understand the relationship between human mobility and the spread of the disease. To this end, we first utilize the cellular network connection logs to discern mobility patterns between the different municipalities of Rio and then combine this aggregate mobility information with the case information in the various regions to make informed predictions for the future. By developing these models, our aim is to help local authorities implement targeted lockdown measures instead of widely unpopular blanket lockdowns to mitigate the spread of the disease. We first describe our Markovian models for modeling the mobility across municipalities and then discuss the linear and polynomial versions of our mobility-aware COVID-19 case prediction model.

A. Markov Models for Human Mobility

To elegantly capture and model aggregate human mobility patterns in Rio, we design a Markov model at the municipality level. Each state in our Markov model corresponds to a municipality and the transitions between states encode movement between municipalities. Our rationale behind mapping states of the model to the municipalities in Rio is multi-fold. First, municipalities are local jurisdictions and are the finest granularity at which it is possible to implement government policies (e.g., lockdowns). Secondly, as described in Section IV, we discern mobility events from connection logs based on transitions of a user between antennas. As our datasets spans the entire city of Rio, it contains 973 antennas. If we map states of the Markov chain to the antennas themselves, it would lead to state space explosion and fail to provide meaningful transition probabilities. Finally, the finest granularity at which one can obtain meaningful COVID-19 case information is also at the municipality level. These factors necessitate that the Markov model be designed to capture mobility at the municipality level. Having provided an overview of the state space of the Markov model, we next discuss how we determine the transition matrix of the Markov model.

1) Empirical Transition Matrix: Recall that we peruse the cellular network connectivity logs to identify mobility events between the different antennas for all the users. We use these mobility events to determine the transition probabilities. We note that there are multiple antennas in the same municipality. Therefore, if a mobility event occurs between two antennas in the same municipality, it corresponds to a same state transition and if the mobility events occur between two antennas in different municipalities, then it corresponds to transitions between the corresponding states of the Markov chain. For example, if a user's trace is of the form {CE, CE, BT, MA}, we obtain the following transitions from the Markov chain: CE to CE, CE to BT, and BT to MA.

To enable authorities design effective policies, our aim is to determine the number of COVID-19 cases for the next week using the data for the past week. Therefore, we determine the transition probabilities of the Markovian models on a weekly basis. We consider all the mobility events for the past week to determine the total number of transitions between the various states. We then normalize the values to empirically obtain the transition probabilities from our data.

B. COVID-19 Case Prediction Model

Having developed the Markovian model and determined its transition matrix in the previous subsection, here, we discuss how we combine the mobility model with the current active COVID-19 cases to predict the number of future cases. Our objective is to design parsimonious mobility-aware prediction models that provide good prediction performance, but yet can be easily implemented in practice. We design two versions of our mobility-aware prediction model — *i)* a linear model, and *ii)* a polynomial model.

1) Linear Model: Our linear mobility-aware prediction model determines the number of COVID-19 cases for the next week in a municipality (i.e., destination) by considering the linear weighted sum of the current COVID-19 infections in the different municipalities (sources) multiplied by the one step transition probability from the different sources to the destination (Eqn. (1)).

$$c_j(t+1) = m_{0j} + m_{1j}(\sum_i p_{ij}(t)c_i(t)) \quad (1)$$

where $c_j(t+1)$ denotes the number of cases in week $(t+1)$ in state j, $c_i(t)$ denotes the cases in week t in state i, $p_{ij}(t)$ is the probability of moving from state i to state j (1-step transition probability from the Markov chain). m_{0j} and m_{1j} are the intercept and slope of the line. We then fit the best line that minimizes the mean squared error to determine m_{0j} and m_{1j}. Our rationale behind this approach is to account

for the number of active infections at a source and to then use the mobility metric (i.e., transition probability from the source to the destination) as a measure to capture the spread of the infection from source to destination.

2) Polynomial Model: While linear models are easy to understand and interpret, we also explore higher order polynomial models to make better predictions. One of the main advantages of higher order polynomial models over linear models is that they can capture nuances in the underlying data better and thus often fit the data better than linear models. Unlike the linear model, the higher-order polynomial model in Eqn (2) fits the best curve. We experiment with primarily polynomials of orders 2 and 3 to keep the number of parameters to a minimum and to avoid overfitting the model to the data.

$$\begin{aligned} c_j(t+1) = m_{0j} + m_{1j}(\sum_i p_{ij}(t)c_i(t)) + \\ m_{2j}(\sum_i p_{ij}(t)c_i(t))^2 + \\ .... \\ + m_{nj}(\sum_i p_{ij}(t)c_i(t))^n \end{aligned} \tag{2}$$

where $m = (m_{0j}, m_{1j}, ..., m_{nj})$ are the coefficients of the polynomial terms.

VI. Experiments

In this section, we first present experimental results to validate the efficacy of the Markovian mobility model. We then conduct experiments to demonstrate the superior prediction performance of our mobility-aware COVID-19 case prediction models when compared with the baseline mobility-agnostic linear prediction model. Overall, our experiments show that the Markovian model accurately captures the aggregate mobility across municipalities and that mobility information can be effectively incorporated in the design of COVID-19 case prediction models. Thus, our research arms local authorities with important tools to design policies aimed at mitigating the spread of the disease.

A. Markovian Model Validation Results

In this subsection, we compare the steady state and transient state performance of the Markov model with those observed directly from the data to demonstrate its efficacy. Recall that in our empirical approach, we determine the one-step transition matrix on a weekly basis directly from the data. Figure 4 shows the one-step transition probabilities for week 1 in the form of a heatmap. In Figure 4, the y-axis represents source regions while the x-axis depicts destination regions. From the heat map we observe that the transition probability (i.e., outgoing mobility probability) to other states for some western zone municipalities, such as SC (Santa Cruz) and CG (Campo Grande), is high for neighboring municipalities. The main reason is that these municipalities cover a large geographical area and are located considerably far from the central and northern zones

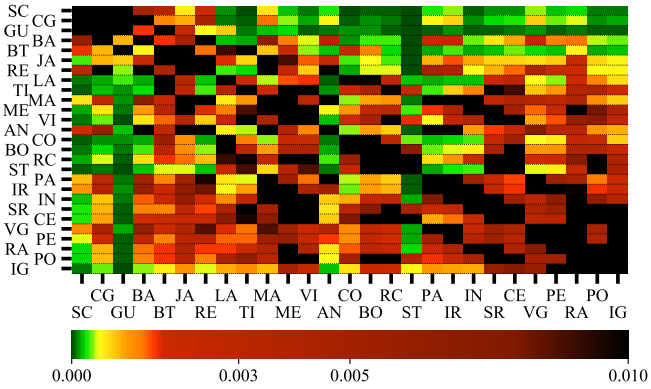
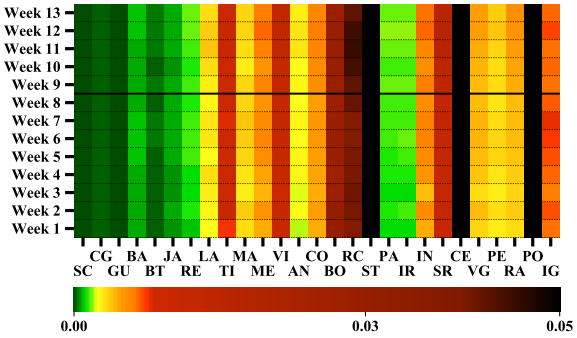
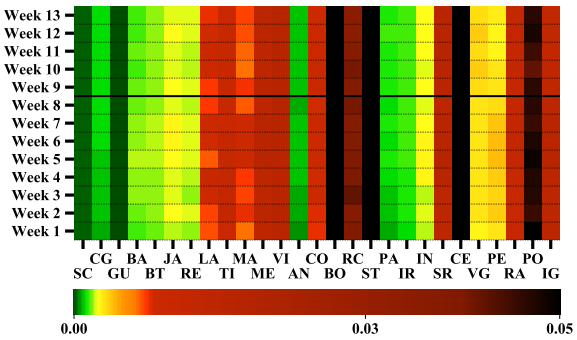


Fig. 4: One-step transition probabilities of week 1



(a) One-step incoming transition probabilities into CE

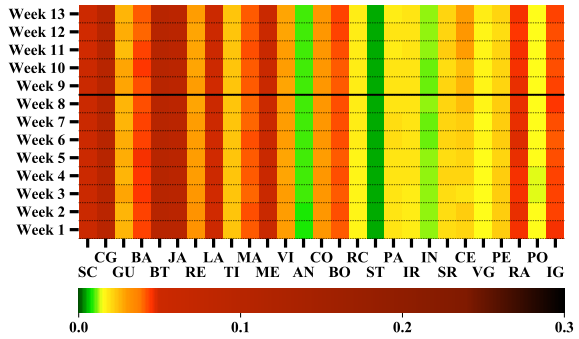


(b) One-step outgoing transition probabilities from CE

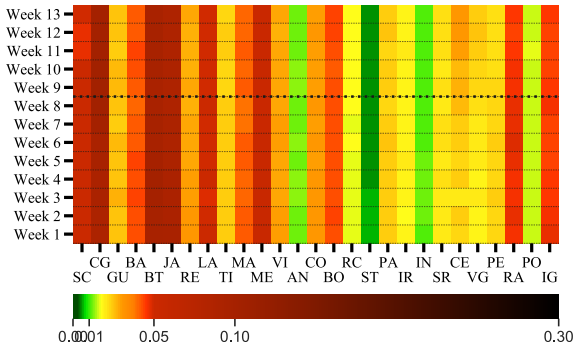
Fig. 5: Heat maps for incoming and outgoing one-step transition probabilities

and therefore, the transition probabilities to municipalities in those zones is low. In contrast, the central zone municipalities, such as CE (Centro) and PO (Portuaria), have high transition probability for the northern zone municipalities. We also see a similar trend for some northern zone municipalities such as IR (Iraja), MA (Madureira) that have high transition probabilities for central zone municipalities.

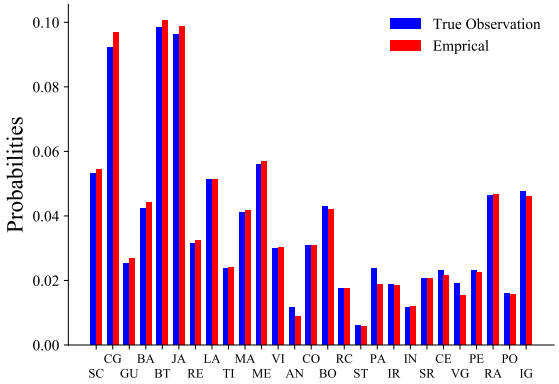
We next investigate the one-step incoming and outgoing



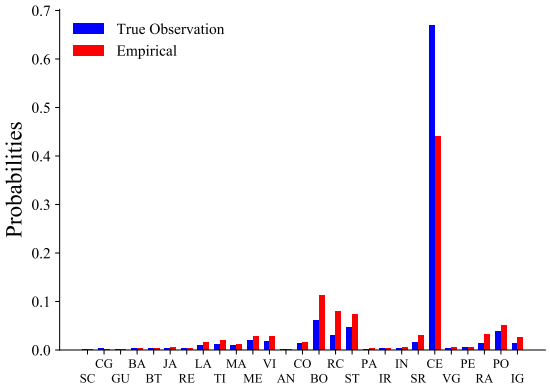
(a) Empirical Approach



(b) True Observations



(c) Steady State Probabilities



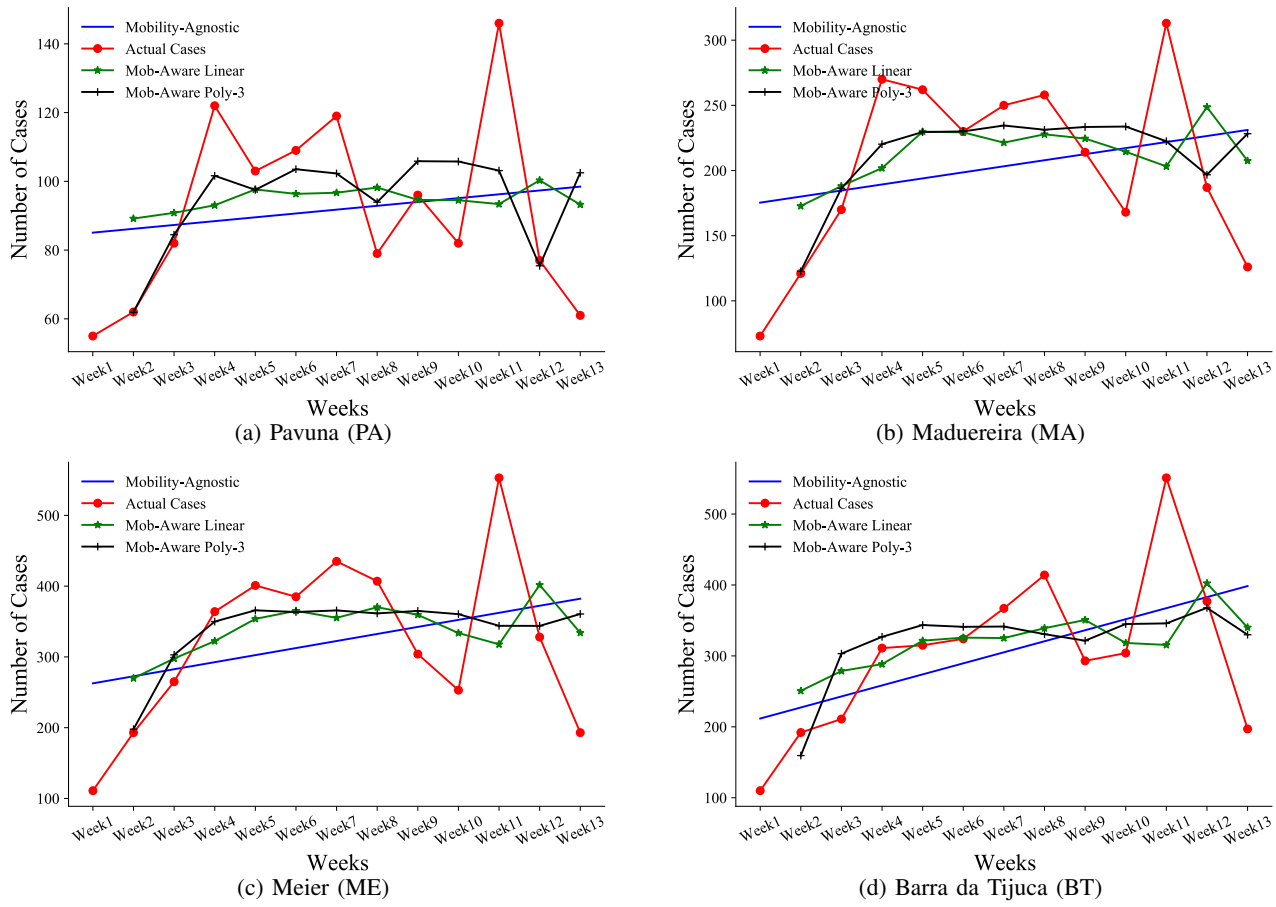
(d) 2-step Outgoing Mobility Probabilities

Fig. 6: Validation of Markovian Model

transition probabilities for the downtown area of Rio (i.e., CE (Centro)). Figures 5a and 5b show the one-step incoming and outgoing transition probabilities for CE for all the weeks, respectively. The horizontal lines between week 8 and week 9 represent the time that the lockdown restrictions are eased. We observe that the probability of getting into CE from its neighboring regions such as ST, RC, and PO is high both during and after lockdown, with the probability increasing for region RC for the after-lockdown period (Figure 5a). We also observe that the likelihood of getting into CE from some north zone municipalities such as TI, VI, and IG, and some south zone municipalities such as BO and CO is reasonably high during and after lockdown. In comparison, the incoming probability to CE from the western municipalities (i.e., SC, CG, and GU) is rather low. One logical reason could be the combination of large geographical distance between the western municipalities and CE and lockdown restrictions, making long distance commutes considerably harder. For other regions such as BT, JA, LA, and ME, we observe that the probability of getting into CE increases after the Brazilian government eased the lockdown restrictions. One explanation could be that people living in these regions started commuting to the downtown regions (i.e., CE) for work after the lockdown is eased.

From Figure 5b, we observe that the outgoing probability from CE to certain municipalities (e.g., RC, ST, and PO) that have high incoming probability into CE is also high. This is interesting as it can help authorities aggregate municipalities that exchange traffic with one another and impose partial lockdowns (instead of blanket ones) by identifying mobility clusters. We also observe that the outgoing probability from CE to most regions such as LA, MA, ME, BO, and CO is higher than the incoming probability to CE from these regions both during and after lockdown. In comparison, we notice that the outgoing probability from CE to some municipalities in the north zone regions such as AN, PA, IN, and VG is lower than incoming probabilities to CE from these regions both during and after lockdown.

We next evaluate the validity of the Markov chain model by comparing the steady state and transient state performance of the empirical transition matrix with those observed directly from the traces. Figures 6a and 6b show the empirical and observed steady state performance of the Markov chain for all the weeks. We see from the figures that the steady state performance of the Markov chain matches the observed probabilities very closely. To quantitatively demonstrate the



(a) Pavuna (PA)
(b) Maduereira (MA)
(c) Meier (ME)
(d) Barra da Tijuca (BT)

Fig. 7: Comparison of mobility-aware linear and polynomial (order 3) models with mobility-agnostic baseline

performance similarity between the empirical approach and the observed values, we present a bar graph that shows the steady-state probabilities for week 1 for both approaches (Figure 6c). We again observe that the empirical steady-state probabilities closely match the observed values for each municipality. The figure also shows that some municipalities have higher steady-state probabilities than others. The main reason for this is that some municipalities such as BT, JA, and LA are quite populous, cover a larger geographical area, and have a significantly larger number of antennas when compared to others. To further quantitatively evaluate the proximity of both approaches, we calculate the variance of the distributions for each approach. We first obtain the variances for each week and then calculate the mean of these variances. This provides us the average variance across all the weeks. The average variance of the steady-state probabilities of the observed probabilities across all weeks is $\sim$0.000633 and the average variance of the empirical steady-state probabilities across all weeks is $\sim$0.000686. These numbers further demonstrate the closeness of the empirical and observed values.

To evaluate the transient state performance, we analyze the two-step outgoing transition probability for both approaches. Figure 6d shows the 2-step transition probability for region CE for week 1. We see that the outgoing 2-step empirical transition probability results match the true observations reasonably. One important reason for the difference is that a significant number of user mobility traces end with unused CE as a source, which impacts the observed values from the traces. For example, if a user trace is of the form {ME ,..., BT, CE, BT, CE, CE}, only the first CE is used for calculating the 2-step transition probability. The last two CEs are ignored in this case. The number of ignored CEs corresponds to 1.66% of total CEs in the traces (i.e., 23,090 out of 1,389,321). Overall, our analysis demonstrates that our Markov model is able to model human mobility patterns across municipalities quite accurately.

B. Mobility-Aware Model Prediction Results

In this subsection, we present results demonstrating the superior performance of the linear and polynomial mobility-aware COVID-19 case prediction models by comparing them with a baseline mobility-agnostic linear regression model. The baseline linear regression model considers only the past COVID-19 cases in a region and produces the best fit straight line for the data. We evaluate the models with respect to 3 different error metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Relative Error (RE).

$$\text{MAE}(y, \hat{y}) = \frac{1}{N} \sum_{i=0}^{N-1} |y_i - \hat{y}_i| \tag{3}$$

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2} \quad (4)$$

$$\text{RE}(y, \hat{y}) = \frac{1}{N}\sum_{i=0}^{N-1}\frac{|y_i - \hat{y}_i|}{y_i} \quad (5)$$

where y_i and $\hat{y}_i$ are the i^{th} actual and predicted values, and N denotes the number of samples (in our case there are 12 samples, one corresponding to each week).

Figure 7 shows the prediction results for four selected regions. We pick 4 representative regions (PA, MA, ME and BT) that are more populous and have four different SPI values: low, fair, good, and high, respectively. We observe that among the regions presented in Figure 7, regions with high SPI have higher number of cases when compared to regions with lower SPI. This is counter-intuitive because people living in regions with lower SPI may not have luxury of working remotely and have to venture out of their homes for their work. A possible explanation is the disparity in testing among the different regions, with access to testing being harder in regions with lower SPI.

From Figure 7, we observe that the mobility-agnostic linear regression fits a straight line (i.e., blue line) with respect to the actual case numbers (i.e., red points). The green and black lines correspond to the mobility-aware linear and polynomial (order 3) models. Note that our mobility-aware linear and polynomial models fit linear and higher order polynomials between past and future cases while taking the one-step transition probabilities into account (Eqns 1 and 2). As a result, even our linear model shows a zig-zag pattern in Figure 7 as we are plotting the number of cases versus the weeks in this figure. We observe qualitatively from the figures that our mobility-aware linear and polynomial models provide better performance than the baseline mobility-agnostic model.

We present the RMSE, MAE, and RE in Tables III and IV. In Table III, we present error rates for regions that have low or fair SPI and in Table IV, we present the error rates for regions that have good or high SPI. Colors show the SPI of the region as classified in Figure 1. We observe from the tables that the mobility-aware linear and polynomial models overall outperform the mobility-agnostic linear regression model for all the error metrics. We observe that the mobility-aware linear and 2^{nd} order polynomial models perform better than the mobility-agnostic model with respect to all 3 error metrics for all except 4 regions (SC, ST, CE, JA). For JA, the mobility-agnostic model predicts better than the linear mobility-aware model with respect to RMSE. Note that JA contains sub-regions with contrasting SPIs, which can pose a challenge when predicting the cases for the region as a whole. In comparison, the mobility-aware 3^{rd} order polynomial model outperforms the mobility-agnostic model baseline for 26 regions (except SC). As expected higher-order polynomial models provide superior prediction performance as they can better model the underlying patterns in the data.

TABLE III: Error rates for low and fair SPI regions

Regions	Models	Errors		
		MAE	RMSE	RE
PA	Mobility-Agnostic	21.557	25.270	0.236
	Mob-Aware Linear	20.543	24.526	0.228
	Mob-Aware Poly-2	15.474	20.851	0.153
	Mob-Aware Poly-3	15.424	20.832	0.151
SC	Mobility-Agnostic	42.608	53.395	0.204
	Mob-Aware Linear	58.809	69.069	0.302
	Mob-Aware Poly-2	57.570	68.657	0.263
	Mob-Aware Poly-3	56.800	66.095	0.260
GU	Mobility-Agnostic	11.286	13.221	0.246
	Mob-Aware Linear	9.087	11.080	0.220
	Mob-Aware Poly-2	8.831	10.550	0.190
	Mob-Aware Poly-3	8.746	10.488	0.187
PO	Mobility-Agnostic	7.427	9.400	0.341
	Mob-Aware Linear	5.298	6.899	0.248
	Mob-Aware Poly-2	5.283	6.899	0.195
	Mob-Aware Poly-3	5.236	6.896	0.191
MA	Mobility-Agnostic	53.094	60.463	0.273
	Mob-Aware Linear	44.968	54.146	0.233
	Mob-Aware Poly-2	35.755	48.488	0.160
	Mob-Aware Poly-3	35.832	48.472	0.159
RE	Mobility-Agnostic	35.073	42.382	0.267
	Mob-Aware Linear	29.400	36.488	0.223
	Mob-Aware Poly-2	26.430	32.699	0.176
	Mob-Aware Poly-3	26.074	32.392	0.171
RA	Mobility-Agnostic	42.097	55.151	0.205
	Mob-Aware Linear	39.983	49.410	0.199
	Mob-Aware Poly-2	39.721	49.386	0.188
	Mob-Aware Poly-3	34.390	47.128	0.161
CG	Mobility-Agnostic	69.791	75.474	0.248
	Mob-Aware Linear	56.576	65.890	0.212
	Mob-Aware Poly-2	57.419	65.852	0.192
	Mob-Aware Poly-3	56.966	64.940	0.189
BA	Mobility-Agnostic	62.891	73.394	0.335
	Mob-Aware Linear	55.799	65.626	0.305
	Mob-Aware Poly-2	55.827	65.577	0.253
	Mob-Aware Poly-3	50.091	63.916	0.224
AN	Mobility-Agnostic	28.652	32.968	0.343
	Mob-Aware Linear	22.467	28.131	0.275
	Mob-Aware Poly-2	21.015	26.469	0.206
	Mob-Aware Poly-3	19.233	24.652	0.194
SR	Mobility-Agnostic	22.199	25.412	0.317
	Mob-Aware Linear	18.162	22.346	0.265
	Mob-Aware Poly-2	17.334	21.900	0.212
	Mob-Aware Poly-3	16.894	21.714	0.203
RC	Mobility-Agnostic	16.532	18.840	0.256
	Mob-Aware Linear	13.827	16.693	0.215
	Mob-Aware Poly-2	11.544	14.439	0.158
	Mob-Aware Poly-3	11.565	14.438	0.159
VG	Mobility-Agnostic	16.369	20.362	0.310
	Mob-Aware Linear	15.390	19.932	0.294
	Mob-Aware Poly-2	12.156	18.685	0.187
	Mob-Aware Poly-3	12.328	18.592	0.185

VII. Conclusion

In this paper, we designed a Markovian model to model the aggregate mobility of people between different regions of a city during the COVID-19 pandemic using cellular network connectivity logs. We then combined the transition probabilities between various regions with the number of active infections in them to develop a mobility-aware COVID-19 case prediction model that predicts the number of future infections. We conducted experiments on billions of cellular network connectivity logs during the pandemic and demon-

TABLE IV: Error rates for good and high SPI regions

Regions	Models	Errors		
		MAE	RMSE	RE
ME	Mobility-Agnostic	90.728	104.250	0.294
	Mob-Aware Linear	76.774	95.450	0.246
	Mob-Aware Poly-2	64.302	90.427	0.185
	Mob-Aware Poly-3	65.776	90.159	0.187
JA	Mobility-Agnostic	92.647	121.985	0.243
	Mob-Aware Linear	80.686	125.314	0.216
	Mob-Aware Poly-2	72.799	117.345	0.187
	Mob-Aware Poly-3	69.433	105.358	0.185
CE	Mobility-Agnostic	89.179	115.988	0.320
	Mob-Aware Linear	118.915	140.280	0.465
	Mob-Aware Poly-2	117.819	140.059	0.358
	Mob-Aware Poly-3	84.348	100.305	0.270
ST	Mobility-Agnostic	4.333	5.696	0.212
	Mob-Aware Linear	5.024	5.894	0.266
	Mob-Aware Poly-2	4.735	5.778	0.222
	Mob-Aware Poly-3	3.655	4.754	0.177
IN	Mobility-Agnostic	33.341	39.068	0.310
	Mob-Aware Linear	30.161	36.391	0.286
	Mob-Aware Poly-2	25.153	30.823	0.200
	Mob-Aware Poly-3	23.265	29.515	0.186
IR	Mobility-Agnostic	28.599	33.493	0.218
	Mob-Aware Linear	23.483	31.492	0.176
	Mob-Aware Poly-2	20.214	28.499	0.149
	Mob-Aware Poly-3	20.289	28.497	0.149
PE	Mobility-Agnostic	23.829	27.520	0.229
	Mob-Aware Linear	19.094	23.464	0.184
	Mob-Aware Poly-2	17.918	22.204	0.164
	Mob-Aware Poly-3	17.759	22.200	0.163
IG	Mobility-Agnostic	48.271	56.677	0.269
	Mob-Aware Linear	41.116	51.923	0.231
	Mob-Aware Poly-2	36.517	49.510	0.186
	Mob-Aware Poly-3	33.918	49.275	0.173
BT	Mobility-Agnostic	69.373	90.805	0.234
	Mob-Aware Linear	62.498	89.477	0.207
	Mob-Aware Poly-2	60.125	85.854	0.181
	Mob-Aware Poly-3	59.307	82.071	0.187
LA	Mobility-Agnostic	47.067	63.438	0.187
	Mob-Aware Linear	45.516	57.453	0.177
	Mob-Aware Poly-2	45.441	57.451	0.166
	Mob-Aware Poly-3	43.952	56.771	0.162
TI	Mobility-Agnostic	42.929	50.826	0.234
	Mob-Aware Linear	35.825	46.619	0.194
	Mob-Aware Poly-2	30.699	42.471	0.149
	Mob-Aware Poly-3	31.351	40.425	0.154
VI	Mobility-Agnostic	25.904	34.381	0.192
	Mob-Aware Linear	23.916	32.039	0.174
	Mob-Aware Poly-2	23.710	31.874	0.164
	Mob-Aware Poly-3	23.139	31.849	0.160
BO	Mobility-Agnostic	38.715	51.309	0.161
	Mob-Aware Linear	32.229	47.975	0.131
	Mob-Aware Poly-2	31.691	46.199	0.121
	Mob-Aware Poly-3	31.593	43.892	0.121
CO	Mobility-Agnostic	28.413	37.820	0.162
	Mob-Aware Linear	25.197	32.826	0.144
	Mob-Aware Poly-2	23.978	32.143	0.124
	Mob-Aware Poly-3	21.999	31.209	0.111

strated the efficacy of the Markovian mobility model. We then showed that our case prediction models far outperformed baseline mobility-agnostic models. Our work can be utilized by government authorities to better understand the spread of the disease and implement targeted region-wise lockdowns instead of widely unpopular blanket lockdown measures.

REFERENCES

[1] ACHTERBERG, M. A., PRASSE, B., MA, L., TRAJANOVSKI, S., KITSAK, M., AND VAN MIEGHEM, P. Comparing the accuracy of several network-based covid-19 prediction algorithms. *International journal of forecasting* (2020).

[2] AHMAD, A., GARHWAL, S., RAY, S. K., KUMAR, G., MALEBARY, S. J., AND BARUKAB, O. M. The number of confirmed cases of covid-19 by using machine learning: Methods and challenges. *Archives of Computational Methods in Engineering* (2020), 1–9.

[3] ALIPOUR, B., TONETTO, L., DING, A. Y., KETABI, R., OTT, J., AND HELMY, A. Flutes vs. cellos: Analyzing mobility-traffic correlations in large wlan traces. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications* (2018), pp. 1637–1645.

[4] ALIPOUR, B., TONETTO, L., KETABI, R., YI DING, A., OTT, J., AND HELMY, A. Where are you going next? a practical multi-dimensional look at mobility prediction. In *Proceedings of the 22nd International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems* (2019), pp. 5–12.

[5] ARIMURA, M., HA, T. V., OKUMURA, K., AND ASADA, T. Changes in urban mobility in sapporo city, japan due to the covid-19 emergency declarations. *Transportation Research Interdisciplinary Perspectives 7* (2020), 100212.

[6] CAO, P. Y., LI, G., CHAMPION, A. C., XUAN, D., ROMIG, S., AND ZHAO, W. On human mobility predictability via wlan logs. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications* (2017), IEEE, pp. 1–9.

[7] CHANG, S. Y., PIERSON, E., KOH, P. W., GERARDIN, J., REDBIRD, B., GRUSKY, D., AND LESKOVEC, J. Mobility network modeling explains higher sars-cov-2 infection rates among disadvantaged groups and informs reopening strategies. *medRxiv* (2020).

[8] FELDMANN, A., GASSER, O., LICHTBLAU, F., PUJOL, E., POESE, I., DIETZEL, C., WAGNER, D., WICHTLHUBER, M., TAPIDOR, J., VALLINA-RODRIGUEZ, N., ET AL. The lockdown effect: Implications of the covid-19 pandemic on internet traffic. In *Proceedings of Internet Measurement Conference (IMC)* (2020).

[9] FENG, J., LI, Y., ZHANG, C., SUN, F., MENG, F., GUO, A., AND JIN, D. Deepmove: Predicting human mobility with attentional recurrent networks. In *Proceedings of the 2018 world wide web conference* (2018), pp. 1459–1468.

[10] GAO, S., RAO, J., KANG, Y., LIANG, Y., AND KRUSE, J. Mapping county-level mobility pattern changes in the united states in response to covid-19. *SIGSPATIAL Special 12*, 1 (2020), 16–26.

[11] GEBRIE, H., FAROOQ, H., AND IMRAN, A. What machine learning predictor performs best for mobility prediction in cellular networks? In *2019 IEEE International Conference on Communications Workshops (ICC Workshops)* (2019), IEEE, pp. 1–6.

[12] HADJIDEMETRIOU, G. M., SASIDHARAN, M., KOUYIALIS, G., AND PARLIKAD, A. K. The impact of government measures and human mobility trend on covid-19 related deaths in the uk. *Transportation research interdisciplinary perspectives 6* (2020), 100167.

[13] HUANG, J., WANG, H., FAN, M., ZHUO, A., SUN, Y., AND LI, Y. Understanding the impact of the covid-19 pandemic on transportation-related behaviors with human mobility data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2020), pp. 3443–3450.

[14] LUTU, A., PERINO, D., BAGNULO, M., FRIAS-MARTINEZ, E., AND KHANGOSSTAR, J. A characterization of the covid-19 pandemic impact on a mobile network operator traffic. In *Proceedings of the ACM Internet Measurement Conference* (2020), pp. 19–33.

[15] NIKHAT, S., AND MEHMET-ALI, M. An analysis of user mobility in cellular networks. In *Proceedings of the International Symposium on Mobility Management and Wireless Access* (2018).

[16] PIZZUTI, C., SOCIEVOLE, A., PRASSE, B., AND VAN MIEGHEM, P. Network-based prediction of covid-19 epidemic spreading in italy. *Applied Network Science 5*, 1 (2020), 1–22.

[17] RAJ, R., SEETHARAM, A., AND RAMESH, A. Ensemble regression models for short-term prediction of confirmed covid-19 cases. *AI for Social Good Workshop* (2020).

[18] ZHANG, H., AND DAI, L. Mobility prediction: A survey on state-of-the-art schemes and future applications. *IEEE Access 7* (2018), 802–822.

[19] ZHU, W., ZHANG, C., YAO, S., GAO, X., AND HAN, J. A spherical hidden Markov model for semantics-rich human mobility modeling. In *AAAI Conference on Artificial Intelligence* (2020).