Predict Cellular Network Traffic with Markov Logic

Marco Lippi, Marco Mamei, Franco Zambonellli

Dipartimento di Scienze e Metodi dell'Ingegneria
University of Modena and Reggio Emilia, Italy
{marco.lippi,marco.mamei,franco.zambonelli}@unimore.it

Abstract

Forecasting spatio-temporal data is a challenging task in transportation scenarios involving agents. In this paper, we propose a statistical relational learning approach to cellular network traffic forecasting, that exploits spatial relationships between close cells in the network grid. The approach is based on Markov logic networks, a powerful framework that combines first-order logic and graphical models into a hybrid model capable of handling both uncertainty in data, and background knowledge of the problem. Experimental results conducted on a real-world data set show the potential of using such information. The proposed methodology can have a strong impact in mobility demand forecasting and in transportation applications.

1 Introduction

The widespread diffusion of mobile phones and cell networks provides a practical way to collect geo-located information from a large user population. The analysis of such collected data can be a fundamental asset in the development of traffic management, urban planning and transport applications [3; 5; 6; 1; 17]. In this work, we explore the use of anonymized Call Detail Records (CDRs) from a cellular network to estimate and predict the distribution of people across the city. CDRs are generated every time a mobile phone interacts with the cellular network (e.g., to send/receive calls and text messages, or to connect to the Internet). Each CDR contains information about the identity of the mobile phone (typically anonymized with an hashed id), its approximate location (i.e., the network cell where the phone is connected) and a time stamp Accordingly, CDRs can serve as sporadic samples of the approximate locations of the phone's owner [11]. On the basis of such location samples, we estimate the distribution of people across the city and try to predict how they will move.

We focus on *aggregated* CDR data (i.e., data measuring the number of CDRs generated from a region, without any reference to the IDs of the people being involved). This kind of data presents a number of ready-to-market applications as privacy concerns are much reduced (in contrast with CDRs with anonymized IDs). In fact, all major telecommunication companies already have services for the analysis and commercial exploitation of this data. While there are several works analyzing properties of aggregated CDR data and predicting user movements from individual CDRs [16], the task of predicting future density of CDRs from aggregated data is relatively under-explored.

Table 1: Example of aggregated CDR data.

Timestamp	**Cell**	**Type**	**Count**
1/3/2015 13:00	3943	SMS In	4.34
1/3/2015 13:00	3943	Call In	2.34
2/3/2015 13:15	3943	Network	9.34
...	...	...	...
30/4/2015 20:45	3943	Call Out	3.45

Accordingly, the main contribution of this paper is to present a novel approach based on Markov Logic (ML) [15] to predict cellular network traffic across the city (as cellular network traffic is a widely used proxy for people presence, our approach can be applied to predict crowd distribution). The main advantage of ML with respect to other time-series forecasting methods is that ML can easily model the relationship between different areas of the city imposing (probabilistic) constraints on how traffic in an area can influence traffic in another one.

The rest of the paper is structured as follows: Section 2 describes CDR data that are employed in our experiments; Section 3 presents Markov Logic Networks (MLN), the main approach we used for our prediction task; Section 4 describes experiments applying MLN to a set of CDR data and discusses results; Section 5 overviews related work in the area of CDR data analysis and time-series forecasting; finally, Section 6 concludes the paper and indicates some interesting directions for future research.

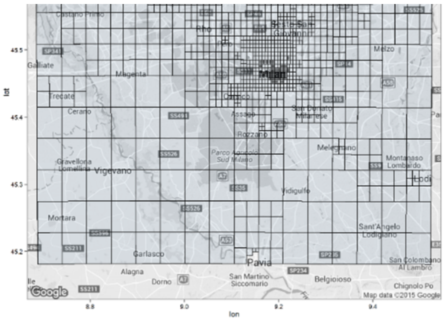

Figure 1: Irregular grid tessellating the area

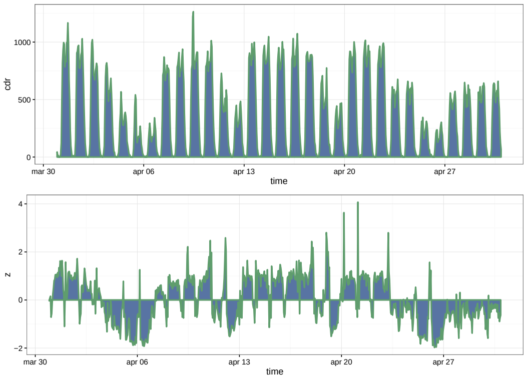

Figure 2: Example of data in a grid cell: (top) Original behaviour extracted from mobile phone data. (bottom) "standardized" score showing the deviation from the mean of that cell at that time, in mean units.

2 Dataset

We focus the analysis on aggregated CDR data that count SMSs, calls and Internet traffic over specific areas of the city and at time intervals. Specifically, the geographic area under analysis is tessellated with an irregularly shaped grid, similar to a Voronoi tessellation. Thus, the more cell network antennas present, the denser the grid (see Figure 1.

The resulting dataset (illustrated in Table 1 is a set of counters that estimate, for each cell of the grid and each 15-minutes-interval, the number of SMSs, calls and Internet traffic. Counters can also be fractional to take into account CDR interactions originating in a cell and ending in another one. The original data comprises about 12 million records like the ones depicted. For each cell and CDR type, the typical plot resembles the one in Figure 2 (top) where it is possible to observe daily and weekly patterns in city dynamics.

To better highlight variability in our data, for each cell, we computed the mean (μ_t) of the time series in that 15 minutes interval (i.e., the mean among all days at that time) and obtain a "standardized" score by computing $\hat{x}_t = (x_t - \mu_t)/\mu_t$. The resulting time series shows the deviation from the mean of that cell at that time, in mean units (e.g., $\hat{x}_t = 1$ means that there is twice – 100% more – as many people than usual), see Figure 2 (bottom). Finally, we discretized $\hat{x}_t$ into a set of classes. Working with discrete values notably simplify computations, without compromising the actual significance and interpretability of the results.

To predict cellular network traffic, we apply Markov logic [15], a statistical relational learning method, to perform *collective* classification on a grid of cells. While traditional machine learning classifiers typically treat the examples as independent and identically distributed, statistical reltional learning approaches are capable of taking into account relations and inter-dependencies between the examples to be classified, so that a joint classification spanning multiple examples can be performed. In particular, we aim to exploit the spatial relationships between cells, as the nature of CDR data is inherently relational along this dimension: at time t, the traffic at two cells c_1 and c_2 spatially close in the network will typically be strongly inter-related.

3 Methodology

Inspired by the work in [10] for road traffic flow forecasting, we modeled our domain with a set of logic predicates that describe the dynamics of CDR traffic data during time and across different cells. Supposing to discretize the amount of traffic in C classes, predicates `Class0(cell,time)`, ..., `ClassC(cell,time)` can be used to indicate the fact that, at a certain cell and at a given time, the traffic quantity falls in one of such classes. A simple predicate `Neighbors(cell,cell)` indicates that two cells are spatially close in the grid. Time is modeled with predicate `Next(time,time)`. Additional information about the day of the week and the part of the day can be also easily modeled with logic predicates. Given a domain described in terms of logic facts, a Markov Logic Network (MLN) consists in a set of weighted rules that model relationships among the existing predicates. For example, a simple predictor that forecasts, for cell c and time t, the same traffic class observed at cell c at time $t-1$, is obtained with the rules:

```
Class0(c,t1) ∧ Next(t1,t2) => Class0(c,t2)
                     ...
ClassC(c,t1) ∧ Next(t1,t2) => ClassC(c,t2)
```

Clearly, such rules are not *always* true, but they are true with a certain probability. Given a collection of observations of past events, the Markov logic framework allows to *learn* the weights of such rules directly from data. The higher the weight, the higher is a probability that the rule will be true. Once the weights of the MLN have been learned, one can use the model to compute the truth value of some query predicates. In the case of this work, the aim is to forecast the dynamics of CDR traffic in the future. Given the current state of the CDR traffic network, by performing inference over the MLN it is possible to retrieve the cell configuration that maximizes the probability of the rules in the model. In order to exploit spatial relationships, rules like the following one can be used:

```
ClassC(c1,t1) ∧ Next(t1,t2) ∧ Neighbors(c1,c2)
                => ¬ Class0(c2,t2)
```

Such rule means that, if there is a high traffic (class C) in a certain cell $c1$ at time $t1$, then at the next time step $t2$ it is unlikely that a neighbor cell $c2$ will have very low traffic (class 0). Complex relationships and dependencies can by modeled with such rules.

Formally, a Markov logic network (MLN) is a set of first-order logic formulae $\mathcal{F} = \{F_1, \dots, F_n\}$, each with an attached real-valued weight $w = \{w_1, \dots, w_n\}$. Given a finite set of constants $C = \{c_1, \dots c_k\}$ (the objects in the domain – in our case described above, cells and timestamps), an MLN induces a Markov network (or Markov random field) where nodes are ground atoms,[1] and edges are present between nodes that appear together in at least one grounding of some formula. An MLN defines a probability distribution over possible variable configurations (i.e., *worlds*):

$$P(X = x) = \frac{1}{Z} \exp\left(\sum_{j}^{|\mathcal{F}|} w_j n_j(x)\right) \tag{1}$$

where $n_j(x)$ is the number of groundings that satisfy formula F_j in x. As stated above, the higher is the weight of a rule, the higher the probability that formula is satisfied.

In most of the application scenarios, some of the predicates (i.e., of random variables) are always observed, which means they are given as *evidence*, whereas others are to be guessed at prediction time, thus being observed only during the training phase. This is called a *discriminative* setting. The truth value of query variables is obtained from the evidence variables, and from the weighted MLN, through a process of *inference*, that computes the maximum a posteriori (i.e., the most probable) configuration of such variables.

The set of parameters w_j associated to the MLN rules can be learned with several different algorithms. In a discriminative setting, MLN weights are learned by maximizing the *conditional log-likelihood* (CLL) of query atoms Y, given the evidence X. The conditional probability $P(Y = y|X = x)$ is defined as:

$$P(Y = y|X = x) = \frac{1}{Z_x} \exp\left(\sum_{i \in F_Y} w_i n_i(x, y)\right) \tag{2}$$

where $n_i(x, y)$ is the number of groundings of formula i in the configuration (x, y) and F_Y is the set of first-order clauses containing query predicates Y. The gradient of the CLL can be computed as:

$$\begin{aligned} \frac{\partial}{\partial w_i} \log P_w(y|x) &= n_i(x,y) - \textstyle\sum_{y'} P_w(y'|x) n_i(x,y') && (3) \\ &= n_i(x,y) - E_w[n_i(x,y)] && (4) \end{aligned}$$

Exactly computing the expected number of true groundings $E_w[n_i(x, y)]$ is an intractable problem, thus approximate algorithms are typically employed. One possible solution is to use the counts in the MAP state y_w^*: in this way, MAP inference is called as a subroutine for each step of the learning algorithm.

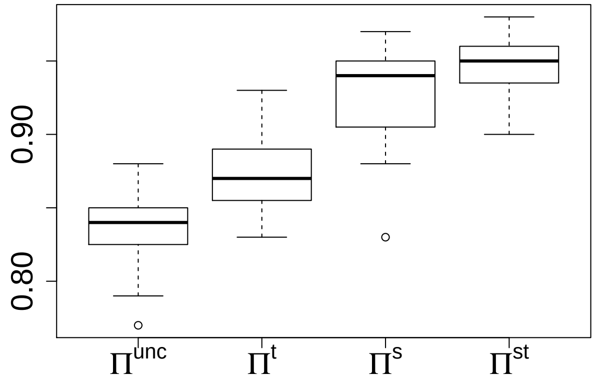

Figure 3: Predictability according to Fano's inequality associated with uncorrelated, time, spatial and spatiotemporal correlated entropies.

Table 2: Comparison of the classifiers employed in the experimental study. We report accuracy and average F_1 over the five classes.

Predictor	Accuracy	Avg F_1
Random	65.7	19.8
Majority	80.3	16.1
RW	74.3	38.8
MLN	77.4	40.6

4 Experiments

We focused the analysis on the province of Milan and we used aggregated CDR data collected over a period of two months: from March 1, 2015 to April 30, 2015 and including calls and SMS sent and received and network traffic. In our experiments we sum together all SMS and calls sent/received, while we do not consider network traffic (as it is expressed in a different format in our data). For each cell, at a given time t this sum is our x_t. The area under analysis is tessellated in 1,419 cells. Cell areas can range from 0.04 Km^2 in the city center to 40 Km^2 in the suburbs. Temporal resolution is 15 minutes. For each cell, we compute the standardized score $\hat{x}_t = (x_t - \mu_t)/\mu_t$ and we discretized those values in 5 classes associated to intervals: $[-\infty, 0.25]$, $[0, 25, 0.50]$, $[0.50, 0.75]$, $[0.75, 1]$, $[1, \infty]$ (i.e., the class $[0.75, 1]$ contains those values having from 75% to 100% more traffic than the mean at that time). Overall, 79% of data fall in the first class, 11% in the second one, 4% both in the third and fifth one, 2% in the fourth one.

Following the work in [16], we try to establish upper bounds for the predictability of aggregate (discretized) CDR behavior across cells. We compute different entropy measures for each cell: *(i)* The random entropy

$$s^{rand} = log_2 N = 2.3$$

[1] In a ground atom all variables have been substituted by constants.

	0	1	2	3	4
0	48,015	3,738	931	437	976
1	4,257	1,894	381	110	122
2	951	418	568	103	145
3	470	134	164	111	191
4	1,163	163	194	208	1,546

Table 3: Confusion matrix of the MLN predictor. Rows: true labels, columns: predictions.

N is the number of values exhibited by the cell – all cells have 5 values.

(ii) The uncorrelated entropy

$$s^{unc} = -\sum_{j=1}^{N} p_j * log_2 p_j$$

(iii) The time (Markov) correlated entropy

$$s^{t} = -\sum_{\hat{x}_t}\sum_{\hat{x}_{t-1}} p(\hat{x}_t, \hat{x}_{t-1}) log_2 p(\hat{x}_t|\hat{x}_{t-1})$$

(iv) The spatial correlated entropy

$$s^{s} = -\sum_{\hat{x}_t} p(\hat{x}_t, \hat{x}_t^1 ... \hat{x}_t^k) log_2 p(\hat{x}_t|\hat{x}_t^1 ... \hat{x}_t^k)$$

Where $\hat{x}_t^1 ... \hat{x}_t^k$ are values in neighbor cells.

(v) Finally, the spatio-temporal correlated entropy

$$s^{st} = -\sum_{\hat{x}_t}\sum_{\hat{x}_{t-1}} p(\hat{x}_t, \hat{x}_{t-1}, \hat{x}_t^1 ... \hat{x}_t^k) log_2 p(\hat{x}_t|\hat{x}_{t-1}, \hat{x}_t^1 ... \hat{x}_t^k)$$

Naturally, for each cell, we will have $s^{rand} \geq s^{unc} \geq s^t, s^s \geq s^{st}$. We then compute the predictability Π associated to each entropy according to Fano's inequality [16]. This is an upper bound for any algorithm predicting $\hat{x}_t$. $\Pi^{rand} = 20\%$ for all cells. Results for other predictabilities are in Figure 3. For example, from these results, we can infer that the upper bound for a classifier using only temporal information (1 step – 15 minutes back) is about 85% (median value). Despite these encouraging predictability results, it is worth noting that the class distribution is highly skewed (e.g., 79% of all the $\hat{x}_t$ are in one class), and thus a simple majority classifier would get very good results in terms of accuracy, according to Π^{unc}. Therefore, more careful analysis is needed.

We run experiments to test the Markov Logic predictor focusing on a subset of 23 cells[2], and we employed the first half of the data (March) for training our system, while the remaining part (April) was used as test set. For Markov logic, we used the Alchemy software,[3] training our model for 1,000 epochs with the voted perceptron algorithm. All the other software parameters were left to their default values. We compared four different predictors (see Table 2). As a first baseline, we measured the performance of a classifier that randomly predicts one of the four classes, by drawing from a probability distribution that knows the true proportions between the classes (called Random in Table 2). As a second baseline, we employ a classifier that simply always predicts the most frequent class (named Majority in Table 2), that is class 0 in our case, corresponding to low traffic. As a third predictor, we use a Random Walk (RW in Table 2) that produces as a forecast at time t the same class that was observed at time $t-1$ (for each cell independently). Finally, we employ an MLN exploiting spatial relationships between the cells. The task is to predict the status of the grid 15 minutes ahead in the future. We report both the accuracy and average F_1 over the five classes (being F_1 of a single class the harmonic mean between its precision and recall). As already introduced when discussing Π^{unc} predictability, the accuracy is clearly dominated by the most frequent class, that is present 80.3% of the times in the test set (which is, in fact, the accuracy of the Majority predictor), while the average F_1 gives the same importance to each of the five classes, and it is thus more significant in this setting. For example, the Majority predictor achieves the best accuracy, but actually it is a completely useless system, as it never predicts something different from the low-level traffic. It is interesting to see that the RW predictor already achieves a significant improvement over Random, thus proving to be a very strong competitor. This behavior suggests that CDR traffic has a dynamic which changes smoothly through time, and 15 minutes ahead in the future is a short horizon to observe big changes in the network configuration. Nevertheless, the MLN approach achieves better results than RW, both in terms of accuracy, and of average F_1. Table 3 reports the confusion matrix for the MLN model: rows/columns represent the true/predicted values, respectively (position i, j in the matrix indicates the number of examples of class i that are predicted to belong to class j).

Figure 4 shows a case study in which spatial relationships help to improve the accuracy of the predictions. The traffic classes for the cells in the network are represented for some timestamp t (left), and for the subsequent timestamp $t+1$ (right). In this scenario, the traffic in cells A and B increases (from green to yellow), which is a case where a Random Walk predictor would fail. The MLN model, on the other hand, correctly forecasts the traffic classes for cells A and B by exploiting spatial relationships, as most of the neighbors at time t belong to a high (yellow, orange or red) traffic class.

5 Related Work

Fueled by the "recent" availability of telecoms' CDR data, a number of researchers try to estimating and predicting the distribution of people across the city on this basis. The works in [7; 2; 14; 12] present approaches to estimate the attractiveness of areas in the city from the combination of cellular network activity and other information sources. They try to estimate the location

[2] We chose the cells whose id contains the prefix 3943.

[3] http://alchemy.cs.washington.edu

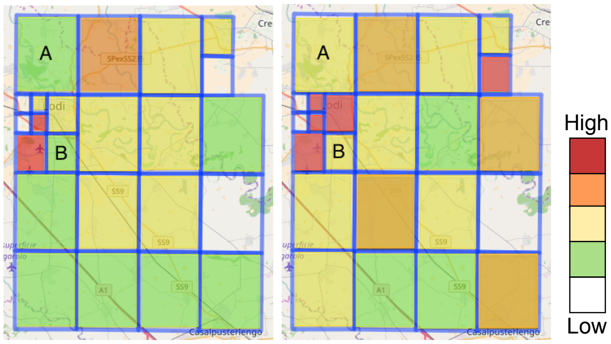

Figure 4: CDR traffic at time t (left) and $t+1$ (right). Traffic at cells A and B increases, following the trend in the neighboring cells.

of cellular network traffic and to use it as a proxy of the number of people in that area.

A similar approach is also adopted by Telefonica's *Smart Steps*, Telecom Italia (TIM) *City Forecast*, and Vodafone *mAnalytics* platforms. All these approaches estimate the present people distribution, and could fruitfully be enriched by our forecasting module.

The work in [10] presents a comparison of multiple algorithms for forecasting vehicular traffic. The data used is based on ad-hoc road traffic sensors, but provides a representation of vehicle counts that is roughly similar to our CDR counters. The algorithm presented in this work extends the algorithms discussed in [10] to the case of multi-class prediction.

The work in [13] predicts people density on the basis of individual CDR data. To predict a user's position, they use a simple model based on previous most frequent locations (in that day at that time). Since humans tend to have very predictable mobility patterns [1, 4, 5], this simple model turns out to give a good predictability baseline. Our work deals with aggregated data that can be processed in a much more privacy-compliant way. Nevertheless, we think that our Markov Logic approach could be fruitfully extended to this other setting and could improve performance by modeling relations between individuals.

The work in [9] and [18] deal with the problem of people density prediction in urban areas on the basis of CDR and GPS traces respectively. Both approaches are based on a recurrent (deep) neural network. Neural networks for different regions are trained separately, but both approaches introduce mechanisms (e.g., using a shared layer between networks of neighbor areas) to take into account spatial dependencies. In our future work we plan to better investigate and compare Markov Logic and deep neural networks. A first comment we can make is that Markov Logic encoding relations via first-order predicate is typically much more interpretable than neural networks.

[8] and [4] presents a set of mechanisms to compute people density on the basis of advanced models of individual people mobility. Individual predictions are aggregated to predict density. [8] deals with individuals CDR data and build people profile in terms of home and work locations. [4] deal with finer-grained GPS data and precisely model trajectories on a shorter time frame. We think that both these aspects could take advantage of the explicit spatial (inter-personal) relations that can be expressed via Markov Logic.

6 Conclusion

We presented a novel approach based on Markov Logic [15] to predict cellular network traffic across the city. As this traffic is a widely used proxy for people presence, our approach can be applied to predict crowd distribution. The distribution of people across the city and the prediction of their aggregated mobility can find application in a number of scenarios from transport and mobility to urban planning. Result show the effectiveness of the MLN framework in this prediction task. MLN is able to take advantage of spatial relationships among cells to perform a collective forecasting of people distribution across the city. While in the present work we modeled only relations between neighboring cells, in our future work we will take into account also relations between distant cells (e.g., a large crowd in a stadium can impact the number of people of a train station in the future event if the station is far away form the staduim). Moreover, we will conduct a comprehensive analysis trying to compare different algorithms for this task.

Acknowledgment

Work supported by the POR-FESR 2014-2020 Project: LUME PlannER - Tools per la pianificazione di viaggi sostenibili presso LUoghi storici, Musei, Eventi artistici e culturali dell'Emilia Romagna
Source of the Dataset: TIM Big Data Challenge 2015, *www.telecomitalia.com/bigdatachallenge*

References

[1] R. Becker, R. Caceres, 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*, 56(1):74–82, 2013.

[2] F. Calabrese, F. Pereira, G. Lorenzo, L. Liang, and C. Ratti. The geography of taste: Analyzing cellphone mobility and social events. In *International Conference on Pervasive Computing*, Helsinki, Finland, 2010.

[3] F. Calabrese, C. Ratti, M. Colonna, P. Lovisolo, and D. Parata. Real-time urban monitoring using cell phones: a case study in rome. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):141–151, 2011.

[4] Z. Fan, X. Song, R. Shibasaki, and R. Adachi. Citymomentum: An online approach for crowd behavior

prediction at a citywide level. In *Ubicomp*, Osaka, Japan, 2015.

[5] L. Ferrari and M. Mamei. Discovering city dynamics through sports tracking applications. *IEEE Computer*, 44(12):61–66, 2011.

[6] L. Ferrari, M. Mamei, and M. Colonna. Discovering events in the city via mobile network analysis. *Journal of Ambient Intelligence and Humanized Computing*, 5(3):265–277, 2014.

[7] F. Girardin, A. Vaccari, A. Gerber, A. Biderman, and C. Ratti. Quantifying urban attractiveness from the distribution and density of digital footprints. *International Journal of Spatial Data Infrastructure Research*, 4:175–200, 2009.

[8] S. Isaacman, R. Becker, R. Cceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger. Human mobility modeling at metropolitan scales. In *ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*, Low Wood Bay, Lake District, UK, 2013.

[9] V. Liang, R. Ma, W. S. Ng, L. Wang, M. Winslett, H. Wu, S. Ying, and Z. Zhang. Mercury: Metro density prediction with recurrent neural network on streaming cdr data. In *IEEE International Conference on Data Engineering*, Helsinki, Finland, 2016.

[10] M. Lippi, M. Bertini, and P. Frasconi. Collective traffic forecasting. In *ECML/PKDD Proceedings, Barcelona, Spain*, 2010.

[11] M. Mamei and M. Colonna. Estimating attendance from cellular network data. *Internaional Journal of Geographic Information Science*, 30(7):1281–1301, 2016.

[12] J. Neumann, M. Zao, A. Karatzoglou, and N. Oliver. Event detection in communication and transportation data. *Pattern Recognition and Image Analysis*, 7887:827–838, 2013.

[13] N. Ponieman, A. Salles, and C. Sarraute. Human mobility and predictability enriched by social phenomena information. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Niagara, Ontario, Canada, 2013.

[14] D. Quercia, G. D. Lorenzo, F. Calabrese, and C. Ratti. Mobile phones and outdoor advertising: Measurable advertising. *IEEE Pervasive Computing*, 10(2):28–36, 2011.

[15] M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62(1):107–136, 2006.

[16] C. Song, Z. Qu, N. Blumm, and A. Barabsi. Limits of predictability in human mobility. *Science*, 327(5968), 2010.

[17] F. Zambonelli. Toward sociotechnical urban superorganisms. *IEEE Computer*, 45(8):76–78, 2012.

[18] J. Zhang, Y. Zheng, D. Qi, R. Li, X. Yi, and T. Li. Predicting citywide crowd flows using deep spatio-temporal residual networks. *Artificial Intelligence*, 259:147 – 166, 2018.