

Interdependence and Predictability of Human Mobility and Social Interactions

Manlio De Domenico
School of Computer Science
University of Birmingham
United Kingdom
m.dedomenico@cs.bham.ac.uk

Antonio Lima
School of Computer Science
University of Birmingham
United Kingdom
axl162@cs.bham.ac.uk

Mirco Musolesi
School of Computer Science
University of Birmingham
United Kingdom
m.musolesi@cs.bham.ac.uk

ABSTRACT

The study of the interdependence of human movement and social ties of individuals is one of the most interesting research areas in computational social science. Previous studies have shown that human movement is predictable to a certain extent at different geographic scales. One of the open problems is how to improve the prediction exploiting additional available information. In particular, one of the key questions is how to characterise and exploit the correlation between movements of friends and acquaintances to increase the accuracy of the forecasting algorithms.

In this paper we discuss the results of our analysis of the Nokia Mobile Data Challenge dataset showing that by means of multivariate nonlinear predictors it is possible to exploit mobility data of friends in order to improve user movement forecasting. This can be seen as a process of discovering correlation patterns in networks of linked social and geographic data. We also show how mutual information can be used to quantify this correlation. We demonstrate how to use this quantity to select individuals with correlated mobility patterns in order to improve movement prediction. We show that the exploitation of data related to friends improves dramatically the prediction with respect to the case of information of people that do not have social ties with the user. Finally, we discuss how movement correlation is linked to social interactions, in terms of colocation and number of phone calls between individuals.

1. INTRODUCTION

The study of the interdependence of human movement and social ties of individuals is one of the most interesting research areas in computational social science [16]. Previous studies have shown that human movement is predictable to a certain extent at different geographic scales [4, 17, 10]. The applications of these prediction techniques are many, including content dissemination of location-aware information such as advertisements in sponsored applications or search results performed from mobile phones [1].

In this paper we show how it is possible to improve mobility prediction by exploiting the correlation between movements of friends and acquaintances. This can be seen as a process of discovering correlation patterns in networks of linked social information and geographic data. It is possible to exploit such correlations for prediction and inference of aspects related to user behaviour, namely their movements and their social interactions (physical and distant ones through phone calls). In particular, in our analysis we exploit and adapt the concept of *mutual information* [8] in order to quantify correlation and provide a *practical* method for the selection of additional data to improve the accuracy of movement forecasting. We also show how this quantity correlates to different types of social interactions of friends and acquaintances.

More specifically, the contributions of this work can be summarised as follows:

- We first show that by means of a multivariate nonlinear predictor [13] we are able to achieve a very high degree of accuracy in forecasting future user geographic locations in terms of longitude and latitude. We compare it with traditional linear prediction techniques (such as ARMA [7]) and we show that these are not able to capture the dynamics of individuals in the geographic space.
- We discuss how the concept of mutual information can be used to quantify the correlation between two users and we demonstrate that it is possible to exploit movement data of friends and acquaintances in order to improve movement prediction. These social ties are measured using different indicators: presence in the address book of a user, colocation and number of phone calls.
- Finally, we study the correlation between mutual information of mobility traces of two individuals, human movement predictability, and social interactions.

The key findings of our analysis are the following: 1) mobility correlation and the presence of social ties can be used to improve movement forecasting by exploiting mobility data of friends; 2) correlated movement is linked to the existence of physical or distant social interactions and vice versa.

Our dataset that has been provided for the Nokia Mobility Data Challenge is composed of information related to 39

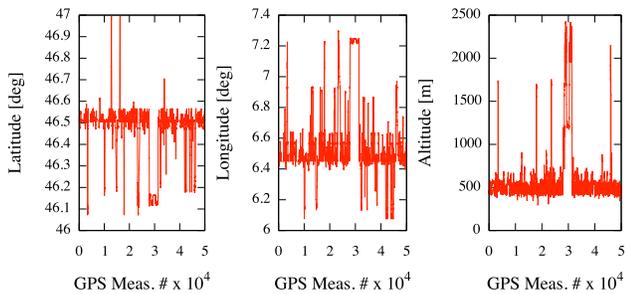


Figure 1: Time series corresponding to the movements of user 179. No periodic behaviour is apparent in the movement traces of the user.

users [15], including the following: GPS traces, telephone numbers, call and SMS history, Bluetooth and WLAN history. We use the information of 25 of them, since the dataset does not include phone numbers for 14 of them; therefore, it is not possible to detect if and when phone calls occur between them. We use GPS traces to analyse the movement of the users.

2. MULTIVARIATE NONLINEAR TIME SERIES PREDICTION

We now present how we apply nonlinear time series prediction methods to the problem of forecasting the future GPS coordinates of the users, given in input initially the past movement history. We will then extend this model by also considering the movement of other users (in particular friends) as input of the nonlinear predictor. In Fig. 1 we show 5×10^4 time-ordered GPS measurements corresponding to the position of user 179 on the Earth. We firstly apply linear prediction models to this time series. The time series appears rather noisy with alternating spikes, nearly flat values, corresponding to stationary points, and fluctuation around an average value. We try to model such movements in the space with a simple multivariate $AR(p) + \text{noise}$ process.

As for the order p of the multivariate autoregressive model that best approximates the original time series, we have chosen the one that minimises an information criterion, according to Akaike [2] and Schwarz [18]. We have found that $p = 24$ provides the best approximation. Hence, we use such a model to perform a multivariate linear forecasting of 1000 GPS measurements for user 179. We validate the model by comparing the latest 1000 real GPS measurements against the forecasted ones¹. The results are shown in Fig. 2, where the real movements are indicated with dots and the forecasting with the linear model is indicated by the solid line. It is evident that the forecasting is not in agreement with observations. In fact, the prediction error on the position (latitude and longitude) is of the order of 3° , whereas the error on the altitude is generally larger than 600 m.

However, although the time series are not regularly sampled,

¹The latest 1000 real GPS measurements have not been included in the procedure adopted to estimate the best order p .

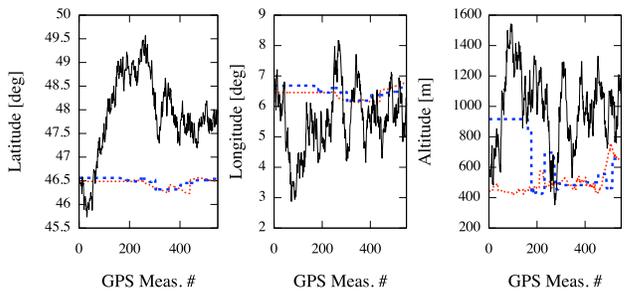


Figure 2: Multivariate nonlinear prediction of user 179 mobility in the geographic space: the first 600 predictions, corresponding to about 60 hours, are shown. The dotted line represents the true movements, the solid line indicates the prediction within an ARMA model, while the dashed line indicates the data obtained by means of a multivariate nonlinear predictor.

we find that they show some features typical of deterministic dynamics contaminated by noise. In fact, preliminary inspection of phase space reconstruction by means of Takens' embedding theorem shows an underlying structure, typical of deterministic dynamical systems.

We model the position of a user on the Earth with a time-varying four-dimensional state vector \mathbf{s}_n with the following dimensions: hour of the day h_n , latitude ϕ_n , longitude λ_n and altitude ξ_n . The prediction of the future states of vector \mathbf{s}_n can be performed using different approaches [13]. We choose the method based on the reconstruction of the phase space of \mathbf{s}_n by means of the delay embedding theorem, since this is considered the best state-of-the-art solution to this problem. While the scalar sequence of coordinates may appear completely non deterministic, it is possible to uncover the characteristics of its dynamic evolution by analysing sub-sequences of the time series itself. In order to investigate the structure of the original system, the time series values must be transformed in a sequence of vectors with a technique called delay embedding. For a univariate time series measurement x_n of a d -dimensional dynamical system, the Takens' embedding theorem [20] allows to reconstruct a m -dimensional space ($m \geq 2d + 1$) with the same dynamical characteristics of the original phase space. The key idea is to build a delay vector \mathbf{x}_n by using delayed measurement defined as follows:

$$\mathbf{x}_n \equiv (x_{n-(m-1)\tau}, x_{n-(m-2)\tau}, \dots, x_{n-\tau}, x_n), \quad (1)$$

where τ is a time delay. Hence, the reconstruction depends on the two parameters m and τ , which have to be estimated. This technique can be extended to the case of the embedding of a multivariate time series² [5].

Under the hypotheses of Takens' theorem, i.e., not noisy time series of infinite length, the underlying dynamics can be fully reconstructed by using only univariate measurements of the dynamical system of interest. Unfortunately, real-world measurement are noisy and with finite length: hence, the phase space reconstruction is more precise if multivariate

²We refer to [21] (and references therein) for an overview of practical applications of multivariate embedding.

measurements of the dynamical system under investigation are performed.

Let us indicate with N the number of measurements corresponding to an M -dimensional time series $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N$, with $\mathbf{y}_i \equiv (y_{1,i}, y_{2,i}, \dots, y_{M,i})$ and $i = 1, 2, \dots, N$. The resulting delay vector is

$$\begin{aligned} \mathbf{v}_n \equiv & (y_{1,n-(m_1-1)\tau_1}, y_{1,n-(m_1-2)\tau_1}, \dots, y_{1,n}, \\ & y_{2,n-(m_2-1)\tau_2}, y_{2,n-(m_2-2)\tau_2}, \dots, y_{2,n}, \\ & \dots \\ & y_{M,n-(m_M-1)\tau_M}, y_{M,n-(m_M-2)\tau_M}, \dots, y_{M,n}), \end{aligned} \quad (2)$$

where m_j and τ_j , $j = 1, 2, \dots, M$ are respectively the embedding and time delays corresponding to each component of the multivariate time series.

Intuitively, this method searches the past history to find and extract sequences of values that are very similar to the recent history. Assuming a certain degree of determinism in the system, the assumption is that, given a certain state (in our case geographic coordinates), there is a strong probability that this will be followed by the same next state.

Although several methods have been proposed to estimate the values of embedding and time delays, in our analysis we consider the same time delay τ for all the series. In fact, for a given user, we have found that the first local minimum of the average mutual information [9], generally adopted to estimate τ in the univariate case, is of the same order of magnitude for any component. This fact has also practical implications, since it simplifies the application of this methodology for the analysis of our data. The optimal embedding dimension is estimated by exploiting the method of false nearest neighbours [14, 13, 11] in the case of multivariate embedding [3]. For any point in the data, an m^* -dimensional phase space is considered and the number of false nearest neighbours, i.e., points which are neighbours in the m^* -dimensional space but not in the $(m^* + 1)$ -dimensional one, is estimated. The desirable embedding dimension m is such that the percentage of false nearest neighbours is small, e.g., below 10%. Any efficient algorithm for counting nearest neighbours is allowed: in particular, we adopt the method implemented in the TISEAN software [12].

Finally, the multivariate nonlinear prediction (MNP) is performed by approximating the dynamics locally in the phase space by a constant (see [6] for further information). In the delay embedding space, all the points in the neighbourhood \mathcal{U}_n of the state \mathbf{v}_n are taken into account in order to predict the coordinates at time $n + k$. Hence, the forecast $\hat{\mathbf{v}}_{n+k}$ for \mathbf{v}_{n+k} is given by

$$\hat{\mathbf{v}}_{n+k} = \frac{1}{|\mathcal{U}_n|} \sum_{\mathbf{v}_j \in \mathcal{U}_n} \mathbf{v}_{j+k}, \quad (3)$$

i.e., the average over the states which correspond to measurements k steps ahead of the neighbours \mathbf{v}_j .

Hence, we use MNP to forecast the same 1000 GPS measurements discussed above. Again, we validate the model by comparing the latest 1000 real GPS measurements against the forecasted ones. The results for user 179 are shown in

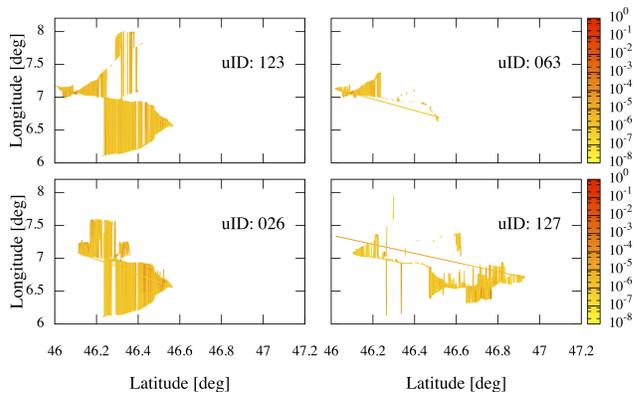


Figure 3: PDF of positions occupied by four different users. Top: users are friends or acquaintances. We say that two individuals are friends or acquaintances if one of them is in the other’s address book. Bottom: users are not friends or acquaintances. The colour indicates the frequency of occupation.

Fig. 2, where the real movements are indicated with triangles and the forecasting with the nonlinear method is indicated by the dashed line. The number of nearest neighbours used to build the neighbourhood \mathcal{U}_n has been kept fixed to 10. Intriguingly, the nonlinear forecasting is in excellent agreement with observations of latitude and longitude, with a global position prediction error equal to 0.19° , and in good agreement with the altitude coordinate, with a global altitude forecasting error equal to 219.43 m.

The global error on the time series prediction is estimated separately for each component using the following formula:

$$e_j = \sqrt{\frac{1}{N} \sum_{n=1}^N (\hat{s}_{j,n} - s_{j,n})^2}, \quad (4)$$

with $j = 1, 2, \dots, M$ with $M = 4$, $N = 1000$. The overall error between the prediction position and the real one is given by the geodesic distance.

3. MUTUAL INFORMATION

In this section, we will briefly introduce the concept of mutual information and we will show how this quantity can be exploited in our analysis to measure the correlation between the movement of different individuals. In the following section, we will then discuss how mutual information can be used to select mobility data of other users that can be exploited as inputs of the nonlinear predictors in order to improve the prediction accuracy.

Let us assume that \mathbf{X} and \mathbf{Y} are two multivariate stochastic variables, and let us indicate with $P_{\mathbf{X}}(\mathbf{x})$ and $P_{\mathbf{Y}}(\mathbf{y})$, respectively, the corresponding Probability Density Functions (PDF). The joint probability is indicated by $P_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})$. The mutual information $\mathcal{I}(\mathbf{X}, \mathbf{Y})$ between such two variables is defined as follows:

$$\mathcal{I}(\mathbf{X}, \mathbf{Y}) = \sum_{\mathbf{x} \in \mathbf{X}} \sum_{\mathbf{y} \in \mathbf{Y}} P_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) \log \frac{P_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})}{P_{\mathbf{X}}(\mathbf{x})P_{\mathbf{Y}}(\mathbf{y})}. \quad (5)$$

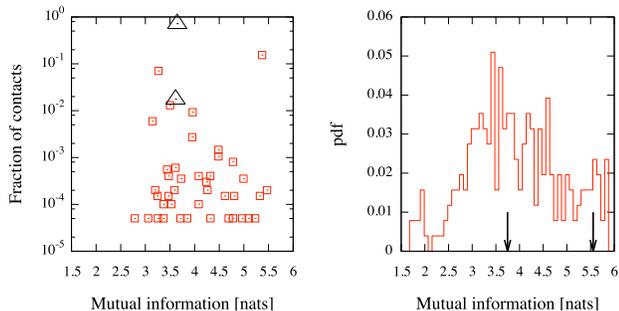


Figure 4: Left panel: scatter plot of the fraction of contacts vs the mutual information estimated for pairs of users with at least one contact, where triangles indicate the two pairs of users connected by social ties in the dataset. Right panel: pdf of mutual information estimated for pairs of users with no contacts at all, where arrows indicate the value of mutual information for the only two pairs of user with social ties, and no contacts, in the dataset.

The mutual information³ quantifies how much information the variable \mathbf{Y} provides about the variable \mathbf{X} . For this reason, it can be used as an estimator of the amount of correlation between \mathbf{X} and \mathbf{Y} . In fact, if the two variables are totally uncorrelated then $P_{\mathbf{X}\mathbf{Y}}(\mathbf{x}, \mathbf{y}) = P_{\mathbf{X}}(\mathbf{x})P_{\mathbf{Y}}(\mathbf{y})$ and $\mathcal{I}(\mathbf{X}, \mathbf{Y}) = 0$.

In our analysis \mathbf{X} represents the motion of a user on the Earth, the random samples \mathbf{x} drawn from \mathbf{X} correspond to geographic coordinates, whereas the PDF of \mathbf{x} quantifies the fraction of time spent by the user in a particular position. In Fig. 3 the two-dimensional PDF of positions occupied by four different users is shown. Users 063 and 123 are friends or acquaintances, while users 026 and 127 are not. We say that two individuals are friends or acquaintances if one of them is in the other’s address book.

We use the mutual information to quantify the amount of correlation between the motion of different users, or, equivalently, how much information the motion \mathbf{Y} provides about the motion \mathbf{X} . However, it is worth noting that friendship or acquaintanceship are only sufficient conditions to have a high mutual information, but they are not necessary. In fact, the mutual information between two users who are not friends or acquaintances can be still high if the two users behave similarly in space and time, i.e., if their motion is similar for some reasons (e.g., a pair of students of the same university department or a pair of colleagues of the same company).

4. EXPLOITING MOVEMENT CORRELATION AND SOCIAL TIES

We now discuss how mobility traces of individuals that have correlated geographic patterns and social ties can be used to improve the accuracy of movement forecasting.

Our approach can be summarised as follows: assuming that we want to predict the movement of a user A , instead of

³The units of mutual information are nats when the natural logarithm is used.

Nodes	Social link	Position Error	Altitude Error
026 127	None	0.167°	66.33 m
063 123	Present	0.011°	20.95 m
094 009	Present	0.003°	5.57 m

Table 1: Global error, defined by Eq. (4), on the prediction of position and altitude for pairs of users connected through social links (defined as presence in the address book of the user).

having only the vector describing the location of user A as input, we will also consider the movement history of another user B , characterised by mobility patterns that are strongly correlated to those of the user we would like to predict. This measure is given by the mutual information introduced in the previous section.

From a mathematical point of view, the idea is to use a 8-dimensional vector that is given by the juxtaposition of the two time-varying state vector representing the states (time-stamped GPS coordinates) of user A and another user B with highly correlated mobility patterns, which we indicate with \mathbf{s}_{n_A} and \mathbf{s}_{n_B} , as inputs of the multivariate nonlinear predictor.

We find that by using additional traces from another user with high correlation, the accuracy of the prediction improves at least by one order of magnitude (and often of two orders of magnitude) with respect to the case of using only his/her own mobility data. Moreover, it is interesting to note that social ties can also be used to select the user for the additional traces as input. In fact, we find that if we select mobility patterns of individuals that are in the address book of the user, the performance of the predictor improves dramatically. At the same time, we would like to stress the fact that the dataset contains a small number of users, so it is difficult to make claims about the general validity of this finding.

Hence, we perform the same analysis described in Section 2, but including in the multivariate nonlinear prediction the time series of movements corresponding to other users. The global prediction error, defined by Eq. (4), of position and altitude is reported in Tab.1 for the three cases. As shown in this table, we observe that the additional information provided by the movement of a user socially linked to that taken into consideration improves the prediction by more than one order of magnitude with respect to the case of users who are not socially linked to each other.

For each pair of users in the dataset, we count the total number of Bluetooth contacts and calls. Then, we estimate the mutual information defined by Eq. (5) for each pair. In order to quantify the amount of correlation between the fraction of contacts and the mutual information, we build a scatter plot between these two observables. The result is shown in the left panel of Fig. 4, by considering only pair of users with at least one contact. The points corresponding to pairs of users with social ties are also shown (triangles). In the right panel of Fig. 4, we show the PDF of mutual information obtained by considering only pairs of users with no contacts at all. The mutual information corresponding to pairs of users

with social ties is shown (arrows). Even if these plots show interesting correlations for this specific dataset, we believe no generalisations can be drawn from these plots, because of the lack of sufficient statistics.

5. DISCUSSION

In the context of mobile applications, the prediction of mobility patterns of users is of great interest for several reasons. For instance, mobility forecasting could be used to determine where the person will be and who he/she will meet. Such an information can enable location-based mobile applications to provide personalised services relating to the context the user is in.

However, we are aware that the method we propose presents some scalability issues for the implementation and the deployment of the proposed technique. In particular, it is well known that calculating mutual information in a multidimensional environment (in this case, for a number of users larger than two) is computationally expensive and does not scale efficiently. In fact, in this case the computational complexity scales as $\mathcal{O}(N^n)$, where N is the subset of users and n is the cardinality of the tuple taken into account. However, the problem we are dealing with usually involves no more than 100 users (e.g., the size of the circle of most significant friends for an individual). For this reason, we can still evaluate mutual information values for any pair of users, which scales as $\mathcal{O}(N^2)$. Nonetheless, the multivariate embedding reconstruction is not feasible for a phase-space larger than 40-dimensional. Even for a 2-coordinate signal representing mobility traces of a single user, it is not unusual to have a large embedding reconstruction due to noisy data. Hence, the mobility traces of no more than three users should be considered simultaneously.

It is worth noting that many factors could be considered as signals of social ties, some being better than others, depending on the scenario taken into consideration. As a consequence, the quality of predictions might be deeply affected, either positively or negatively, by the criteria used to detect social ties. In the Nokia MDC dataset, we had no information about social ties between individuals, neither of real nor virtual nature. In the provided dataset, the presence of an individual in the address book of another one actually represents the strongest definition of a social tie. Moreover, two individuals with no social ties might show similar mobility patterns, resulting in a high value of mutual information. However, it is worth remarking that the presence of a social tie is only a sufficient but not necessary condition for having a significant correlation between the mobility patterns of two users. In fact, it could happen that two individuals are not socially linked (i.e., they are not friends, co-workers and so on), but their mobility can be highly correlated, e.g., in the case of an individual whose work depends on the actions of another one. If this is the case, the multivariate nonlinear predictor will greatly benefit of such a correlation. On the other hand, it could happen that two individuals with social ties show very different mobility patterns, degrading the accuracy of the predictor. It is likely that individuals with strong social ties (students, friends, co-workers and so on) behave similarly and their mobility traces are characterised by patterns with a high value of mutual information. Hence, the accuracy of the predictor will be improved in the case

the dynamics of users is highly correlated, even if a social tie does not exist.

Finally, we would like to point out two possible refinements of this work, which we intend to investigate in larger datasets. In this paper we have considered only the mutual information of the mobility patterns between two individuals at time zero, i.e., we estimate the mutual information for traces that are not separated by any lag time. Actually, the movements of two individuals could be correlated but not synchronised: a time-delayed mutual information should be able to capture such a feature. Hence, we plan to investigate the ranking of the users with respect to the maximum value of the time-delayed mutual information between their mobility patterns. A possible second refinement is the use of multivariate nonlinear prediction with non-uniform embedding (different delays) and local polynomial fitting [19] in order to increase the accuracy of the prediction.

6. CONCLUSIONS

Through the analysis of the Nokia Mobile Data Challenge traces, we have shown that it is possible to exploit the correlation between movement data and social interactions in order to improve the accuracy of forecasting of the future geographic position of a user. In particular, mobility correlation, measured by means of mutual information, and the presence of social ties can be used to improve movement forecasting by exploiting mobility data of friends. Moreover, this correlation can be used as an indicator of potential existence of physical or distant social interactions and vice versa.

Acknowledgements

This work was supported through the EPSRC Grant ‘‘The Uncertainty of Identity: Linking Spatiotemporal Information Between Virtual and Real Worlds’’ (EP/J005266/1).

7. REFERENCES

- [1] L. Aalto, N. G othlin, J. Korhonen, and T. Ojala. Bluetooth and WAP Push Based Location-aware Mobile Advertising System. In *Proceedings of MobiSys '04*, pages 49–58, 2004.
- [2] H. Akaike. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, C-19(6):716–723, 1974.
- [3] S. Boccaletti, D. Valladares, L. Pecora, H. Geffert, and T. Carroll. Reconstructing embedding spaces of coupled dynamical systems from multivariate data. *Physical Review E*, 65(3):35204–35204, 2002.
- [4] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439:462–465, 2006.
- [5] L. Cao, A. Mees, and K. Judd. Dynamics from multivariate time series. *Physica D*, 121(1-2):75–88, 1998.
- [6] M. Casdagli. Nonlinear prediction of chaotic time series. *Physica D*, 35(3):335–356, 1989.
- [7] C. Chatfield. *The Analysis of Time Series An Introduction*. Chapman and Hall CRC, 2004.
- [8] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2006.
- [9] A. Fraser and H. Swinney. Independent coordinates

- for strange attractors from mutual information.
Physical Review A, 33(2):1134, 1986.
- [10] M. C. González, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453:779–782, June 2008.
- [11] R. Hegger and H. Kantz. Improved false nearest neighbor method to detect determinism in time series data. *Physical Review E*, 60(4):4970, 1999.
- [12] R. Hegger, H. Kantz, and T. Schreiber. Practical implementation of nonlinear time series methods: The TISEAN package. *Chaos*, 9(2):413–435, 1999.
- [13] H. Kantz and T. Schreiber. *Nonlinear time series analysis*. Cambridge University Press, 1997.
- [14] M. Kennel, R. Brown, and H. Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review A*, 45(6):3403, 1992.
- [15] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, and M. Miettinen. The Mobile Data Challenge: Big Data for Mobile Computing Research. In *Proceedings of the Mobile Data Challenge by Nokia Workshop. Colocated with Pervasive'12*, 2012.
- [16] D. Lazer et al. Computational Social Science. *Science*, 323:721–723, February 2009.
- [17] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell. NextPlace: A Spatio-temporal Prediction Framework for Pervasive Systems. In *Proceedings of Pervasive'11*, San Francisco, California, USA, June 2011.
- [18] G. Schwarz. Estimating the Dimension of a Model. *Ann. Stat.*, 6(2):461–464, 1978.
- [19] L. Su. Prediction of multivariate chaotic time series with local polynomial fitting. *Computers & Mathematics with Applications*, 59(2):737–744, 2010.
- [20] F. Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Lecture Notes in Mathematics*, volume 898, pages 366–381. Springer-Verlag, 1981.
- [21] I. Vlachos and D. Kugiumtzis. State space reconstruction from multiple time series. In *Topics on Chaotic Systems: Selected Papers from Chaos 2008 International Conference*, page 378. World Scientific Publishing, 2009.