An Empirical Study on the Regularity of Route Mobility

Farbod Faghihi
University of Helsinki
PO Box 68, FI-00014, Finland
faghihib@cs.helsinki.fi

Petteri Nurmi
University of Helsinki
PO Box 68, FI-00014, Finland
petteri.nurmi@cs.helsinki.fi

Ubicomp/ISWC'16 Adjunct , September 12-16, 2016, Heidelberg, Germany
ACM 978-1-4503-4462-3/16/09.
http://dx.doi.org/10.1145/2968219.2968420

Abstract
Studies on human mobility have demonstrated that transitions between important locations are highly regular, and that in overall, people spend the majority of their time in a small set of important locations. However, what currently is not known is how regular the movements leading from one location to another are, i.e., how regular is the *route mobility* of individuals. The present paper contributes by conducting the first ever empirical study on regularity of route regularity. We demonstrate that routes indeed contain a high degree of regularity and that the uncertainty associated with a route is concentrated along a small set of so-called fork points. Accordingly, our results suggest that routes can be encoded using a combination of sub-segments with high regularity, and a set of fork points that serve as transition points for the sub-segments. We carry out our analysis using the CabSpotting dataset, which contains mobility traces from 538 cabs in San Francisco metropolitan area collected during one month period.

Author Keywords
Trajectory analytic; Path regularity; Trajectory entropy

ACM Classification Keywords
H.m [Information Systems]: Miscellaneous

Introduction

Understanding and characterizing human mobility is a fundamental scientific challenge that has relevance to numerous scientific domains, ranging from urban planning to epidemiology [5], and a wide range of innovative mobile applications and services [6, 12, 15]. Previous research on understanding human mobility has predominantly focused on transitions from one important location (or place) to another, and shown these to be highly regular (see Related Work). However, what is not known is how regular the movements between these locations are, i.e., how regular *human movement trajectories* are.

The present paper contributes by presenting the first ever study on the regularity of route mobility. Our analysis has been carried out using the Cab Spotting dataset [11], which contains the GPS location of 538 cabs from a one month period in San Francisco. Human mobility is necessarily constrained by road network and urban infrastructure, making the movements of (occupied) cabs a feasible proxy for human mobility. As our first technical contribution, we develop a principled methodology for quantifying the entropy of movement trajectories for any given spatial and temporal resolution. The idea is to convert trajectories into Markov chains and quantify their regularity using Markov trajectory entropy [4]. We use our framework to analyze overall regularity of trajectories and demonstrate that, as a whole, trajectories appear irregular and difficult to predict. However, as part of our analysis we examine the uncertainty associated with different locations along the trajectory and demonstrate that most of the uncertainty is concentrated along a small set of so-called *fork points*. Motivated by this result, as our second technical contribution, we develop an algorithm for segmenting trajectories into sub-segments that are highly regular. Using these sub-segments, the overall trajectory can be encoded efficiently and compactly.

Related Work

Previous work on characterizing the regularity of human mobility has mainly focused on (i) characterizing regularity of transitions between frequently visited locations and on (ii) modeling displacements in mobility. In terms of former, Song et al. [14] used cell tower information to characterize transitions and derived an upper bound of 93% for human mobility. Lin et al. [8] repeated the study using GPS data and found 90% potential predictability for building level granularity. Lu et al. [9] studied regularity of mobility patterns in Cote d'Ivoire. Similarly to the other studies, the authors found the theoretical upper bound on predictability to be close to 90%. Lin et al. also demonstrated that mobility predictors can reach accuracy close to the upper bound. Smith et al. [13] refined the limit by considering a more realistic transition model for human mobility, demonstrating that the upper bound of predictability is likely to be $10 - 20\%$ lower than shown by other studies. De Domenico et al. [3] demonstrated a relationship between social interactions and mobility, showing that higher predictability can be achieved when also the social relationships of users are taken into consideration.

Regarding displacements, travel distances have been shown to follow a so-called Lévy flight model [2, 5]. According to this model, human mobility consists of many short flights which are interleaved with some longer distance flights, and the distances of the flights follow power-law distribution. Zhao et al. [16] show how the Lévy flight model can be decomposed into a mixture of log-normal distributions when information about transportation modalities are available.

Data Description

We have performed our analysis using the Cab Spotting dataset [11], which contains the GPS locations of 536 cabs sampled during one month (May-June 2008) in San Fran-

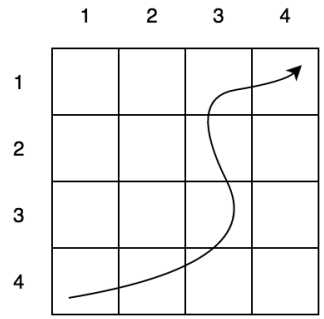

Figure 1: A movement trajectory discretized over a 2-D coordination system

cisco. In total, the data contains over 11 million GPS measurements. The sampling rate of the GPS receiver varies in the data. For our analysis, we restrict to samples containing at most 70 seconds between them. The data contains also an occupancy indicator for each GPS measurement, which we use to identify start and end points of a trajectory.

To quantify the regularity of trajectories, we discretize the data by mapping each GPS measurement into a discrete grid index. We perform the mapping by converting each latitude, longitude pair into a cell index on a $d \times d$ sized grid; see Nurmi et al. [10] for details of the conversion. Discretizing the measurements is essential for computational tractability, and helps to overcome inaccuracies in location measurements, e.g., due to driving on a different lane or due to inaccurate GPS fixes. In our analysis we consider $d = 200$m, which was chosen as a trade-off between location accuracy and computational requirements. Choosing lower values for d results in trajectories with higher resolution, but in the case of the CabSpotting dataset such values will lead in inaccurate trajectories with many inherent gaps. For the Cab Spotting data this corresponds to a model with 120×100 grid cells. Once the points have been converted into grid indices, we represent each trajectory as a string containing the cell indices of the location measurements. Formally, let s and d denote the source and destinations of a trip, the trajectory between s and d is then defined as $T_{s,d} = c_1, ..., c_n$; see Figure 1 for an illustration.

The sparse sampling rate of the GPS receiver can result in trips along the same route generating different trajectories due to gaps in the measurements. To mitigate this issue and to alleviate the effect of gaps in our regularity analysis, we interpolate the measurements to have a constant 10 second sampling interval. We also prune the set of trajectories by removing trips that visit only a small number of cells or that have long update rates. Specifically, we remove (i) all trajectories with a duration less than 5 minutes or longer than 2 hours as these are likely to be erroneous; (ii) contain update rates higher than 70 seconds as these are likely to contain inaccuracies due to GPS unavailability or GPS receiver failure; (iii) have measurements from fewer than four grid cells. In total, this results in $286,629$ distinct trajectories for our analysis.

Quantifying Trajectory Regularity

We quantify regularity of trajectories by modeling them as (first order) Markov chains. Accordingly, each cell along a trajectory corresponds to a state in a Markov chain and movements from one cell to another correspond to transitions between states. Given the collection of all trajectories $\mathcal{T}$, we create a Markov chain $\mathcal{M}(\mathcal{T})$ by constructing a transition probability matrix $P = p_{i,j}$ where $p_{i,j}$ denotes the probability of observing a transition between grid cells i and j. Once the Markov chain has been constructed, we can analyze regularity of trajectories by examining the randomness of the trajectories between each source and destination pair. Any route between an arbitrary source s and destination t can now be seen as a realization of a Markov trajectory between the states corresponding to s and t, and the regularity of the route choices can be examined by analyzing the entropy of the resulting Markov trajectories.

Calculating Entropy of Markov Trajectories

To calculate the entropy of Markov trajectories, we construct a trajectory entropy matrix H that encodes the randomness between each possible source and destination pair. Following Ekroot [4], the following closed form expression can be used to construct H:

$$H = K - \tilde{K} + H_{\Delta}. \quad (1)$$

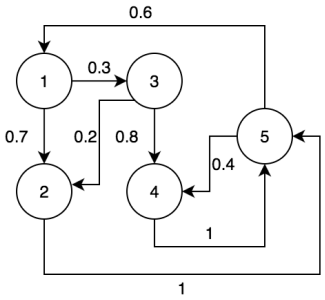

Figure 2: State transition diagram for a 5-state Markov Chain

The matrix H_Δ is a diagonal matrix with entries $(H_\Delta)_{i,i} = \frac{H(\chi)}{\mu_{i\cdot}}$ where $H(\chi)$ is the entropy rate of a state. The entropy rate of a state, in turn, is given by

$$H(\chi) = -\Sigma_{i,j}\mu_i P_{i,j} log P_{i,j} \quad (2)$$

where μ is the stationary distribution of the Markov chain $\mathcal{M}(\mathcal{T})$ which can be obtained through eigenvalue analysis. The matrix K in Equation 1 is given by

$$(I - P + A)^{-1}(H^* - H_\Delta) \quad (3)$$

where $\tilde{K}$ is a matrix in which the ijth element $\tilde{K}$ equals the diagonal element $K_{j,j}$ of K; A is the matrix of stationary probabilities with entries $A_{i,j} = \mu_j$; $H^*_{i,j} = H(P_{i\cdot}) = -\Sigma_k P_{i,k} log P_{i,k}$ is the matrix of single-step entropies.

To illustrate the concept of entropy of Markov trajectories, consider the Markov chain illustrated in Figure 2, which results in the following probability transition matrix P:

$$P = \begin{pmatrix} 0 & 0.7 & 0.3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0.2 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0.6 & 0 & 0 & 0.4 & 0 \end{pmatrix}$$

By using Equation 1, we can obtain the following matrix of Markov trajectory entropies:

$$H = \begin{pmatrix} 2.7161 & 1.9555 & 6.7135 & 3.3746 & 1.0978 \\ 1.6182 & 3.5738 & 8.3318 & 2.9957 & 0 \\ 2.3401 & 3.5810 & 9.0537 & 1.3210 & 0.72192 \\ 1.6182 & 3.5738 & 8.3318 & 2.9957 & 0 \\ 1.6182 & 3.5738 & 8.3318 & 2.9957 & 1.6296 \end{pmatrix}$$

From the matrix we can observe, e.g., that the entropy of trajectories between cells 1 and 5 is 1.6296 bits, while the entropy of deterministic trajectories, $T_{4,5}$ and $T_{2,5}$, is equal to zero, i.e., they contain no randomness.

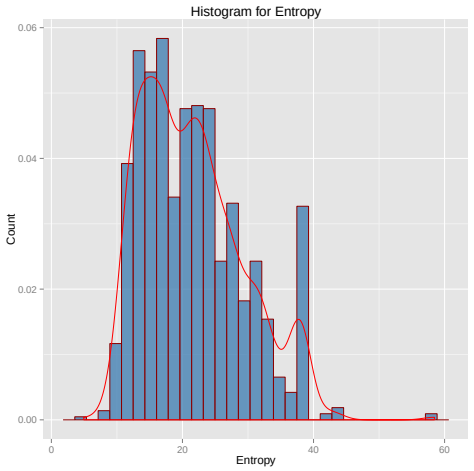

Figure 3: Histogram of trajectory entropies.

Regularity of Trajectories

We first consider the overall regularity of movement trajectories, which is illustrated in Figure 3. From the figure we can observe that generally the entropy of the trajectories contains little variation, with the entropy of most trajectories being between 15 and 35. These values, however, are difficult to interpret or to relate to predictability. To obtain an alternative view on the regularity, Figure 4 contains a heatmap of the entropy rates (see Eq. 2) of the different states / grid cells. From the plot we can observe that the highest uncertainties are associated with the main roads along the downtown areas, however, generally all states tend to have relatively high uncertainty. Thus, state-by-state predictions are likely to result in very poor prediction performance, suggesting that next cell based predictors are not sufficient for predicting human movement trajectories.

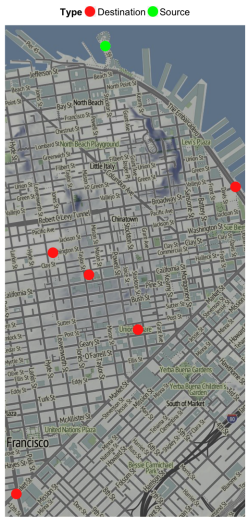

Figure 5: Six popular areas among tourist, each containing several attractions according to Tripadvisor top visited locations.

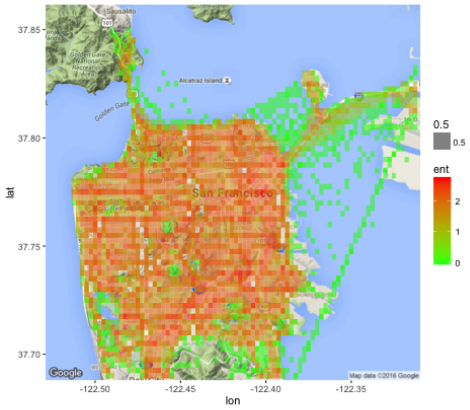

Figure 4: Overall entropy of trajectories

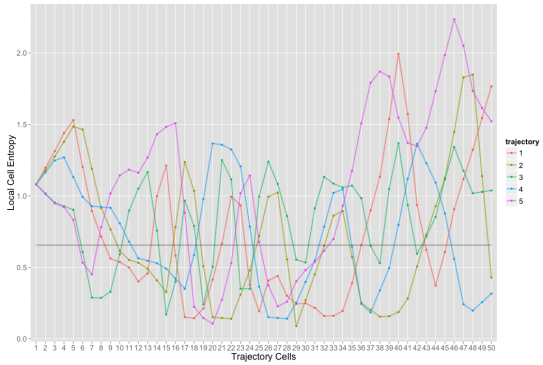

Figure 6: Cell local entropies along five different trajectories chosen between the a specific source and five popular destinations based on the popular tourist attractions. The average value of the state entropies are shown with a black line.

Distribution of Entropy within a Trajectory

To further illustrate the characteristics of route regularity, we consider as an example six popular tourist attractions (according to TripAdvisor.com) in San Francisco; see Figure 5. We select the most popular location as the source of our trajectories and consider the other five as our destinations. To compare the entropies of the trajectories, we interpolate the cell entropies along each trajectory to have the same length and we align the cell trajectories of the different trajectories. This results in a time series representation of the evolution of entropy along each trajectory. The resulting entropies are shown in Figure 6.

From the results we can observe that several cells have very low entropy, suggesting movements within them are highly predictable. Such cells provide little information about movements within the trajectory, and a more compact representation of the trajectory could be obtained by excluding such cells. We can also observe that a small number of grid cells have significantly higher entropy than other cells along the trajectory, appearing as peaks in the resulting plot. These points essentially correspond to decision points along a trajectory. We refer to these points as *fork points*.

Trajectory Segmentation

The regularity analysis revealed that most of the uncertainty associated with route choices is concentrated along a small set of fork points. Each fork point can be effectively understood as a point that segments the overall trajectory into sub-segments that are to a high degree predictable. To identify such point automatically, as the second contribution of the paper we propose a novel online algorithm for detecting fork points from streams of location measurements.

Algorithm 1 ForkPointDetection

1: $stateEntropyArray \leftarrow$ state entropy of the cells based on history
2: $\mu \leftarrow$ mean(stateEntropiesWindow)
3: $\sigma \leftarrow$ sd(stateEntropiesWindow)
4: $z \leftarrow$ thresholdValue
5: $window \leftarrow emptylist$
6: $forkPoints \leftarrow emptylist$
7: **while** *There's a new cell as cell* **do**:
8: $stateEntropyArray \leftarrow stateEtnropyArray.append(entropy(cell))$
9: $\mu \leftarrow$ mean(stateEntropiesWindow)
10: $\sigma \leftarrow$ sd(stateEntropiesWindow)
11: **if** $\frac{entropy(cell)-\mu}{\sigma} \geq z$ **then**
12: $window \leftarrow window.append(entropy(cell)).$
13: **else**
14: **if** window is not empty **then**
15: $forkPoints \leftarrow forkPoints.append(max(window))$
16: $window \leftarrow emptylist$
17: return forkPoints

Our algorithm for detecting fork points is summarized in Algorithm 1. The core idea is to continuously monitor the entropy rate (see Equation 2) of grid cells that are encountered, and to return cells that have a significantly higher entropy rate than other recently encountered cells as the fork points. To accomplish this, two challenges need to be solved: (i) we need to be able to estimate the entropy rate of each grid cell with sufficient accuracy in an online fashion; and (ii) we need to have a mechanism for determining significant deviations in entropy rates in a robust fashion.

To estimate the entropy rate of grid cells, we maintain an online estimate of the probability transition matrix P of the underlying Markov chain. This can be simply accomplished by keeping a running count of the number of transitions between grid cells. Note that the interpolation of the measurements ensures that subsequent measurements are in neighboring cells and consequently we only need to store eight values per cell. These values can even be maintained on mobile devices as the number of cells encountered by an individual is typically relatively small. To calculate the entropy rate, we need to solve the stationary distribution μ of P, which can be done either on demand or periodically when the matrix P changes significantly.

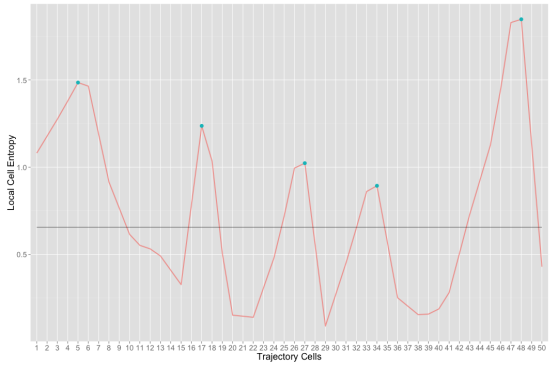

Figure 7: Detected fork-points of an example trajectory where the threshold is shown with the black line.

We detect significant deviations in the entropy rates using a statistical significance test. Specifically, we calculate running estimates of the mean and standard deviation of the overall entropy rate of a trajectory and derive a z-score for each cell that is encountered. Whenever the z-score of a cell exceeds a threshold of statistical significance, we initiate peak detection and buffer measurements until the score of the cell falls below the initial threshold. The cell with the

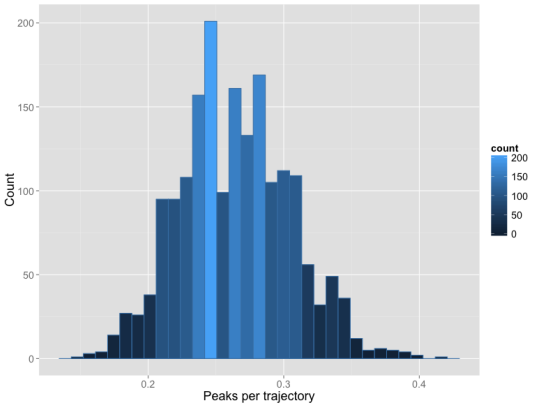

Figure 8: The number of fork-points per trajectory based on the San Francisco cab trajectory traces

maximal z-score is then selected as the fork point. Figure 7 demonstrates our fork-point detection algorithm over an example trajectories. The fork-points are shown as blue dots and the black line is is our threshold.

After detecting the fork-points in each trajectory it is possible to compress and encode the trajectory with its fork-points. To illustrate the benefits of this approach, we selected the most repeated location in the cab spotting dataset as the source of the trips and extracted all trajectories. A histogram of the compression rate of the trajectories is shown in Figure 8, demonstrating that most of the time even 70% savings can be achieved in the trajectory representation. This can be used also for other purposes, e.g., trajectory tracking systems such as EnTracked [7, 1] can use fork points to schedule location updates in order to minimize overall energy consumption of the tracking.

Discussion and Summary

The fact that the majority of a trajectory is regular is interesting for several reasons. On a system level, location and trajectory tracking solutions, such as EnTracked [7, 1], can reduce the sampling rate of GPS and other energy-heavy sensors during segments with high predictability. As we have shown, these are typically the longest segments in a trajectory, suggesting that significant reductions in energy consumption can be achieved.

We investigated regularity using a uniform spatial resolution. A limitation with this approach is that factors such as multiple lanes or large traffic junctions can affect the regularity results as the locations of the users fall to neighboring cells, decreasing the overall regularity. In addition, we only considered trajectories from cabs. While cab traces are a feasible proxy for human mobility, cab mobility may have certain properties that can differ from, e.g., journeys by private cars. We plan to further analyze this in our future work. We also plan to investigate how information about the traffic network can be used to refine our regularity results.

Acknowledgments

The work was supported in part by the Academy of Finland, under the project Sampling in Pervasive Sensing Systems, grant 296139. The work only reflects the authors' views.

REFERENCES

1. Sourav Bhattacharya, Henrik Blunck, Mikkel Kjærgaard, and Petteri Nurmi. 2015. Robust and Energy-Efficient Trajectory Tracking for Mobile Devices. *IEEE Transactions on Mobile Computing* 14 (2015), 2. DOI:http://dx.doi.org/10.1109/TMC.2014.2318712
2. D. Brockmann, L. Hufnagel, and T. Geisel. 2006. The scaling laws of human travel. *Nature letters* 439 (2006), 462–465.

3. Manlio De Domenico, Antonio Lima, and Mirco Musolesi. 2013. Interdependence and Predictability of Human Mobility and Social Interactions. *Pervasive and Mobile Computing* 9 (2013), 798–807. http://arxiv.org/abs/1210.2376
4. Laura Ekroot and Thomas M. Cover. 1993. The entropy of Markov trajectories. *IEEE Transactions on Information Theory* 39 (1993), 1418–1421.
5. Marta C. González, César A. Hidalgo, and Albert-László Barabási. 2008. Understanding individual human mobility patterns. *Nature* 453 (2008), 779–782. DOI:http://dx.doi.org/10.1038/nature06958
6. Samuli Hemminki, Petteri Nurmi, and Sasu Tarkoma. 2013. Accelerometer-Based Transportation Mode Detection on Smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems (SenSys)*. ACM.
7. Mikkel Baun Kjærgaard, Sourav Bhattacharya, Henrik Blunck, and Petteri Nurmi. 2011. Energy-efficient Trajectory Tracking for Mobile Devices. In *Proceedings of the 9th International Conference on Mobile Systems, Applications and Services (MobiSys)*.
8. Miao Lin, Wen-Jing Hsu, and Zhuo Qi Lee. 2012. Predictability of Individuals' Mobility with High-Resolution Positioning Data. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM Press, 381–390.
9. Xin Lu, Erik Wetter, Nita Bharti, Andrew J. Tatem, and Linus Bengtsson. 2013. Approaching the Limit of Predictability in Human Mobility. *Scientific Reports* 3 (2013).
10. Petteri Nurmi, Sourav Bhattacharya, and Joonas Kukkonen. 2010. A grid-based algorithm for on-device GSM positioning. In *Proceedings of the 12th International Conference on Ubiquitous Computing (UbiComp)*. 227–236. DOI: http://dx.doi.org/10.1145/1864349.1864385
11. Michal Piorkowski, Natasa Sarafijanovic-Djukic, and Matthias Grossglauser. 2009. CRAWDAD data set epfl/mobility (v. 2009-02-24). Downloaded from http://crawdad.cs.dartmouth.edu/epfl/mobility. (February 2009).
12. James Scott, A. J. Bernheim Brush, John Krumm, Brian Meyers, Michael Hazas, Stephen Hodges, and Nicolas Villar. 2011. PreHeat: controlling home heating using occupancy prediction. In *Proceedings of the 13th international conference on Ubiquitous computing (Ubicomp)*. ACM, 281–290.
13. Gavin Smith, Romain Wieser, James Goulding, and Duncan Barrack. 2014. A refined limit on the predictability of human mobility. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications*.
14. Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of Predictability in Human Mobility. *Science* 19, 5968 (2010), 1018–1021.
15. Libo Song, D. Kotz, Ravi Jain, and Xiaoning He. 2006. Evaluating Next-Cell Predictors with Extensive Wi-Fi Mobility Data. *IEEE Transactions on Mobile Computing* 5, 12 (2006), 1633–1649.
16. Kai Zhao, Mirco Musolesi, Pan Hui, Weixiong Rao, and Sasu Tarkoma. 2015. Explaining the power-law distribution of human mobility through transportation modality decomposition. *Scientific Reports* 5 (2015).