A Refined Limit on the Predictability of Human Mobility

Gavin Smith
Horizon Digital Economy Research
The University of Nottingham
gavin.smith@nottingham.ac.uk

Romain Wieser
Télécom ParisTech, Paris.
ENSG, Marne-la-vallée.
romain.wieser@telecom-paristech.fr

James Goulding and Duncan Barrack
Horizon Digital Economy Research
The University of Nottingham
{first}.{last}@nottingham.ac.uk

***Abstract*—It has been recently claimed that human movement is highly predictable. While an upper bound of 93% predictability was shown, this was based upon human movement trajectories of very high spatiotemporal granularity. Recent studies reduced this spatiotemporal granularity down to the level of GPS data, and under a similar methodology results once again suggested a high predictability upper bound (i.e. 90% when movement was quantized down to a spatial resolution approximately the size of a large building). In this work we reconsider the derivation of the upper bound to movement predictability. By considering real-world topological constraints we are able to achieve a tighter upper bound, representing a more refined limit to the predictability of human movement. Our results show that this upper bound is between 11-24% less than previously claimed at a spatial resolution of approx. 100m×100m, with a greater improvement for finer spatial resolutions. This indicates that human mobility is potentially less predictable than previously thought. We provide an in-depth examination of how varying the spatial and temporal quantization affects predictability, and consider the impact of corresponding limits using a large set of real-world GPS traces. Particularly at fine-grained spatial quantizations, where a significant number of practical applications lie, these new (lower) upper limits raise serious questions about the use of location information alone for prediction, contributing more evidence that such prediction must integrate external variables.**

I. Introduction

Prediction of human movement has received increasing attention from both the research community and industry. Since 2010, when [1] provided evidence that the probability of correctly predicting an individual's next location had an upper bound of 93%, prediction models have become significantly more advanced, focusing on predictions at increasingly fine levels of spatial and temporal granularity. This has led to increased interest in understanding mobility at these levels resulting in [2], [3] and [4] computing the upper bound using the formulation from [1] on different datasets with different properties and spatiotemporal granularities. Knowledge of the upper bounds of mobility predictors not only provides valuable insight into human behaviour, but is also of interest to the vast array of applications that benefit from accurate prediction of an individual's future location: ubiquitous advertising [5]; service provision [6]; and intelligent agents both virtual [7], [8] and physical [9]. Importantly, understanding such upper bounds also enable application designers to work around any constraints such bounds imply, and provides researchers with insight into the design of effective prediction algorithms.

However, knowing an upper bound is only useful if it is relatively tight, i.e. it is close to the maximum predictability actually possible. In this work we show that the approach to deriving the upper bound of predictability in human mobility provided by [1] can be significantly tightened by integrating topological constraints, providing a far lower and hence more accurate upper bound. Moreover, we note that our new approach limits the rate at which the computed upper bound is increasingly over-estimated as the utilized spatial (temporal) quantization becomes more fine(coarse)-grained.

II. Related Work

Consideration of the limits of human mobility predictability and the provision of a mathematical approach to computer an upper bound for a given dataset was first undertaken in [1]. Considering a geo-locational dataset obtained from cell tower logs, heuristically corrected to conform to a 1hr sampling rate, the authors were able to provide an insight into the upper bound on predictability, with a value of 93% reported. Using their formula a number of researches have since interrogated different datasets [2]–[4] in order either extend or address some perceived weaknesses of the approach. [2] replicated the results of [1] using 14 participants via a combination of sensors while also investigating the effect of varying temporal resolution. They demonstrated that the upper limit of predictability increases as the temporal resolution becomes finer-grained. Later, using the Geolife dataset [10] and [3] noted a direct relationship between spatial quantization and the upper bound on predictability [3, p388-389]. Most recently, [4] again considered varying spatiotemporal quantizations but in contrast to [3] used a dataset also including indoor locations. However, compared to all previous studies this produced significantly higher upper bounds (although as their dataset is not publicly available it is not possible to replicate these results). We also note that some works such as [11]–[13] have considered the performance of pre-existing prediction algorithms. However, because we focus on the theoretical bounds of an optimal predictor rather than current algorithms their work is considered complementary rather than directly related.

The work presented here differs substantially from all prior research by integrating topological constraints into the calculation of the upper bound of predictability in [1]. This refinement, which takes into account that some locations are simply unreachable for next-step movement, provides significantly tighter upper bounds and hence more realistic (i.e. lower) upper limits. We focus on the Geolife dataset (also used in [3]) for our empirical tests, also examining the effects of spatiotemporal quantization.

Published in the Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PERCOM 2014).

III. Defining "predictability" and "mobility"

In this section we explicitly define *predictability* and *mobility* as used within our work. Unless explicitly stated otherwise the definitions we adopt correspond to those presented in [1]. While we acknowledge that other definitions of predictability do exist (e.g. [14]), investigating the limits of predictability under such starkly different definitions (and hence end use applications) is beyond the scope of this work.

Mobility Prediction: Prediction of the next-location of an individual at a specific, regular, temporal sampling rate. This is the definition embedded in the approach taken by [1] and requires missing data points to be imnputed. This definition has been used in a range of applications such as [6], [11].

Movement Mobility Prediction: As a sub-problem of *mobility prediction* we define *movement mobility prediction* as the next-location prediction of an individual at a specific, regular, temporal sampling rate but *only* as part of a journey (referred to subsequently as a trajectory). Data points only need to be imputed if they are missing within a trajectory. Examples of work utilising such a definition include [7], [8], [15].

In this work we focus on the second of these for two reasons. First, the more complex problem of *movement mobility prediction* deserves attention since it often bounds the usefulness of real-world applications, and is hence of practical use in guiding both algorithm developers and implementers. The second reason is purely pragmatic: the largest open movement dataset available at a fine-grained spatiotemporal level is the Geolife dataset [10] (supporting both generalizability and reproducibility) which is a "trajectory dataset" [16]. Additionally, such a definition removes the need to heuristically impute unavailable data (which is prone to significant error and/or bias). We note, however, that the definitions are all sufficiently close that the main contribution of this work is unaffected - assessing the impact of changing the formulation for computing the upper bounds of predictability. Having defined *mobility* we recall the definitions specified in [1] for *predictability* and its maximum theoretical value:

Predictability (Π^{alg}): For a specific prediction algorithm, alg, and a specific temporal sampling rate, predictability is defined as the *mean probability* of correctly predicting a person's next location, given knowledge of all of the possible trajectories that could have led them to that point.

Maximum Predictability (Π^{max}): The highest potential accuracy for predictability (as defined above) is formulated by assuming one possesses the best prediction algorithm that is theoretically possible. Specifically, consider the prediction of the n^{th} location within an individual's movement trajectory and model the probabilities of all possible locations as a random variable, X_n. Additionally model all possible movement trajectories that may have led the individual to that point as the multivariate random variable, h_n (which hence has dimensionality $n-1$). This allows the definition of the conditional probability distribution for the individual's n^{th} location as $P(X_n|h_n)$. Now, consider that the best one can hope from any predictor is to return the value from this distribution that corresponded to the *highest probability* next location. [1] define this *most likely* next location as x_{ML}, and denote its probability of occurrence as $\pi(h_n)$. The expected correct-prediction rate from this optimal algorithm can therefore be written as $\sum_{h_n} P(h_n)\pi(h_n)$. However, this just represents expected success for one specific step, so to calculate overall predictability, Π^{max}, the average over all values of n is taken:

$$\Pi^{max} \equiv \lim_{t\to\infty} \frac{1}{t} \sum_{n=1}^{t} \left[\sum_{h_n} P(h_n)\pi(h_n)\right] \quad (1)$$

Unfortunately, it is not possible to directly compute Π^{max}, so instead work is devoted to finding as tight an overestimate of its value as possible (an upper bound), based on empirical data. An upper bound of this form is denoted $\Pi^{\overline{max}}$.

IV. Measuring the limits of predictability

One method for calculating an upper bound, $\Pi^{\overline{max}}$, on the maximum predictability of human movement was given in [1]. Their primary intuition was that, having modelled an individual's movement pattern as a stochastic process, one can compute a corresponding entropy rate and therefore quantify the amount of uncertainty inherent to their mobility behaviour (intuitively, a process with low uncertainty will be highly predictable; a process with high uncertainty will not). Song et al. managed to establish an explicit formula which connected that entropy rate, $\mathcal{H}(\mathcal{X})$, with the upper bound on predictability, $\Pi^{\overline{max}}$, and by applying a numerical solver to this formula they were then able to calculate a value for $\Pi^{\overline{max}}$. We now summarize how this derivation was performed[1], before going on to show in section V how to improve upon it and produce a greatly refined version of their upper bound.

A. Connecting Entropy to an Upper bound for Predictability

Consider the *entropy rate*, $\mathcal{H}(\mathcal{X})$, of a stochastic process. This can be defined in terms of the joint entropy of its t random variables, but if the process is stationary it can also be formulated as:

$$\mathcal{H}(\mathcal{X}) = \lim_{t\to\infty} \frac{1}{t} H(X_t|X_{t-1}, X_{t-2}, \ldots, X_1) \quad (2)$$

Recall that $P(X_n|h_n)$ represents the distribution for a single timestep in the process, conditional on the values of the timesteps preceding it, h_n. Given the definition of conditional probability one can rewrite Eq. 2 as the weighted average of the conditional entropy of each of the t time indexed variables:

$$\mathcal{H}(\mathcal{X}) = \lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} \left[\sum_{h_n} P(h_n) H(X_n|h_n)\right] \quad (3)$$

It is this equation which is ultimately linked in [1] to the maximum predictability, Π^{max}. First note that:

$$H(X_n|h_n) = -\sum_{x\in X} P(x|h_n) \log_2 P(x|h_n) \quad (4)$$

By taking the slightly unusual move of isolating the probability, π (for notational convenience we drop the parametrisation here since h_n is fixed in this and subsequent contexts), that an individual will undertake their *most likely* next-location step, x_{ML}, then this equation can be rewritten as:

$$H(X_n|h_n) = -\pi \log_2 \pi - \sum_{x\neq x_{ML}, x\in X} P(x|h_n) \log_2 P(x|h_n) \quad (5)$$

In this form it becomes clearer as to how one might define an upper bound for the equation. [1] achieves this by replacing the *actual* next-step distribution, $P(X_n|h_n)$, with a new distribution, $P(X_n^{'}|h_n)$, that preserves the probability that an

[1]Most of the derivation comes from the supporting material for [1], [17].

individual makes the most likely move, π, but which assumes that all of the other remaining possibilities are *equally likely* with a probability $\frac{1-\pi}{N-1}$ (where N is the number of possible location values). This ensures that the entropy function (denoted here as $H_F(\pi)$) is now solely dependent on π:

$$\begin{aligned} H_F(\pi) &\equiv H(X_n^{'}|h_n) \\ &= -\pi \log_2 \pi - (1-\pi) \log_2 \left(\tfrac{1-\pi}{N-1}\right) \end{aligned} \quad (6)$$

Equally importantly, in altering the distribution in this way its entropy can only have *increased*. This seems intuitive because the distribution $P(X_n'|h_n)$ is *at least* as random as the original and leads to Theorem 1 (replicated from [17, p12], full proof of claim from [17] in appendix A):

Theorem 1. $H(X_n|h_n) \leq H_F(\pi)$

Applying this inequality to the definition for the overall entropy rate (Eq. 3) provides the connection to the maximum predictability, Π^{max}. This is replicated from [17, p 13-14] herein as Theorem 2, which notes that a connection between the entropy rate of the empirical process and Π^{max} can be achieved by the re-parametrisation of the H_F function with the notion of overall predictability.

Theorem 2. $\mathcal{H}(\mathcal{X}) \leq H_F(\Pi^{max})$

Finally, [17, p14] show that an upper bound for Π^{max}, $\Pi^{\overline{max}}$, will occur at the boundary case of Theorem 2, when $\mathcal{H}(\mathcal{X}) = H_F(\Pi^{max})$. Substituting Eq. 6, yields:

$$\begin{aligned} \mathcal{H}(\mathcal{X}) &= H_F(\Pi^{\overline{max}}) \\ &= -\Pi^{\overline{max}} \log_2 \Pi^{\overline{max}} - (1-\Pi^{\overline{max}}) \log_2 \tfrac{1-\Pi^{\overline{max}}}{N-1} \end{aligned} \quad (7)$$

This is the key equation - if $\mathcal{H}(\mathcal{X})$ is known it is now possible to calculate $\Pi^{\overline{max}}$ by sending Eq. 7 through a numeric solver. One problem remains: finding the value of $\mathcal{H}(\mathcal{X})$. Unfortunately, the generating function behind the stochastic process, $\mathcal{X}$, is often unknown (and even when it is specified a closed form for its entropy rate can often not be defined) so the direct calculation of $\mathcal{H}(\mathcal{X})$ is impractical. The solution to this issue is to estimate the entropy rate empirically [18].

V. A Refined Predictability Limit

It is important to recognize that $\Pi^{\overline{max}}$ is only an upper bound - and as such, is potentially an over-estimate. In this section we adapt Song et al.'s formulation in order to derive a much tighter, and consequently more realistic, upper bound for predictability.

Recall from the definition of $H_F(\pi)$ (Eq. 6) that an individual's *unknown* next-step probability distribution $P(X_n|h_n)$ was replaced by the new distribution $P(X_n'|h_n)$, the entropy of which was an upper bound to the original because it preserved the probability of the most likely next step, π, but distributed the remaining $1-\pi$ probability uniformly over all other locations. We propose that it is possible to significantly improve on this upper bound. The reason for this is that in reality we *do* have knowledge about the structure of $P(X_n|h_n)$ that we can integrate into our upper bound replacement - real-world spatial constraints and physical limitations on movement speeds that mean a number of the possible outcomes of the random variable, X_n, *must* be zero, given the history h_n.

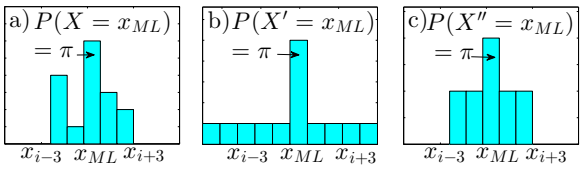

Fig. 1. (a) An illustrative example of an actual "true" distribution (unknown in practice) over next step locations with unreachable locations denoted by zero probabilities due to real world topological constraints. (b) The higher entropy approximation, given we know π, proposed in [1]. (c) The approximation proposed in this work, guaranteed to have a entropy higher than (a) but lower than (b) by taking into account that both π and the number of zero-valued locations is known (note: exact placement of the zero/non-zero elements is not, nor required to be, known since it does not change the distribution entropy).

Specifically, if we know that only a limited number (N_r) of distinct next-step locations are reachable from a given location, then exactly $N - N_r$ probabilities in our upper bound distribution must be zero in the actual (unknown) $P(X_n|h_n)$ distribution. Therefore, rather than distributing the remaining $1-\pi$ probability uniformly over all locations other than x_{ML}, which would incorrectly assign probability to unreachable locations, we distribute the remaining $1-\pi$ probability uniformly over $N_r - 1$ locations, those known to be neither zero nor x_{ML}. In similar fashion to [1], when assuming π is known, this results in a distribution with a corresponding entropy greater than or equal to that of the true distribution. Importantly, however, our approach additionally ensures a corresponding entropy *lower* than that achieved in Song et al.'s formulation as we are applying a uniform distribution to a smaller or equal number of locations. Hence our model of this distribution is closer to that of the true distribution (see Fig. 1).

In reality however, π is not a known, set value and it is the entropy (rate) itself that is fixed (being computed empirically from an individual's historic data). This is the reason why our formulation results in a lower score for predictablity, because in order to maintain the same entropy despite the fact it possesses fewer non-zero symbols, $P(X_n|h_n)$ must be distributed in a more even fashion. Given the uniform distribution of all other locations, this is only possible if π, the probability of the most likely location, is lower than it would otherwise have been considered in [1], resulting in a reduced bound for overall predictability. We define our alternative entropy function, $H_F'(\pi)$, as follows:

$$H_F'(\pi) \equiv H(X^{''}|h_n) \quad (8)$$

$$= -\pi \log_2 \pi - (1-\pi) \log_2 \left(\tfrac{1-\pi}{N_r-1}\right) \quad (9)$$

Having defined such a function we now must prove it is both concave (required for an equivalent Theorem 2) and satisfies the inequality: $H(X_n|h_n) \leq H_F'(\pi)$ (our equivalent to Theorem 1), after which the re-parameterisation of $H_F'(\pi)$ by Π can, as before, be equated with the empirical entropy rate (such that $\Pi = \Pi^{\overline{max}}$). This allows the new equation:

$$\begin{aligned} \mathcal{H}(\mathcal{X}) &= H_F'(\Pi^{\overline{max}}) \\ &= -\Pi^{\overline{max}} \log_2 \Pi^{\overline{max}} - (1-\Pi^{\overline{max}}) \log_2 \tfrac{1-\Pi^{\overline{max}}}{N_r-1} \end{aligned} \quad (10)$$

to be put through a numerical solver in order to obtain a refined $\Pi^{\overline{max}}$. To prove that $H(X_n|h_n) \leq H_F'(\pi)$ we first provide a lemma regarding the maximum value of the entropy of a distribution with N_r non-zero outcomes.

Lemma 1. *The maximum entropy of a distribution with N_r non-zero outcomes is $log_2 N_r$. Proof: See Appendix B.*

Theorem 3. $H(X_n|h_n) \leq H'_F(\pi)$

Proof: Consider that the optimal prediction algorithm defined in [17] always predicts x_{ML}. A prediction error will occur when this prediction differs from an individuals actual move (i.e. $x \neq x_{ML}$). Let us define the probability of this event as $P(e)$, and denote its associated binary entropy as $H(E)$. Correspondingly:

$$P(x_{ML}|h_n) = \pi = 1 - P(e|h_n) \tag{11}$$

As shown in the derivation of Fano's inequality [19], this allows us to rewrite the entropy, $H(X_n|h_n)$, as:

$$H(X_n|h_n) \quad = H(E|h_n) + P(e|h_n)H(X_{x \neq x_{ML}}|h_n) \tag{12}$$

By noting that $H(X_{x \neq x_{ML}}|h_n)$ is the entropy of an ensemble of $N-1$ elements and since we know $P(X_n|h_n)$ has at least $N - N_r$ zero elements, lemma 1 indicates that the entropy of such an ensemble cannot exceed $\log_2(N_r - 1)$. Eq. 12 can therefore be re-written as:

$$H(X_n|h_n) \leq H(E|h_n) + P(e|h_n)\log_2(N_r - 1) \tag{13}$$

Substituting 11 into 13:

$$H(X_n|h_n) \leq -\pi \log_2 \pi - (1-\pi)\log_2\left(\tfrac{1-\pi}{N_r - 1}\right) \tag{14}$$

Since RHS of Eq. 14 equals Eq. 9: $H(X_n|h_n) \leq H'_F(\pi)$ ■

To prove $H'_F(\pi)$ is concave, note that if one additionally parameterizes $H_F(\pi)$ by N one has $H_F(\pi) = H_F(\pi, N)$ and $H'_F(\pi) = H_F(\pi, N_r)$. Therefore, since $H_F(\pi, N = N_r)$ is concave and monotonically decreasing in π [17], so is $H'_F(\pi)$.

Finally we provide the proof that the refined approach does correspond with a tighter upper bound on the maximum predictability. Specifically, define $\Pi_1^{\overline{max}}$ as the solution to $\mathcal{H}(\mathcal{X}) = H_F(\Pi^{\overline{max}})$ (as per the original method in [1]) and $\Pi_2^{\overline{max}}$ as the solution to $\mathcal{H}(\mathcal{X}) = H'_F(\Pi^{\overline{max}})$ (the refined approach), then it follows that $\Pi_2^{\overline{max}} \leq \Pi_1^{\overline{max}}$:

Proof: $\Pi_2^{\overline{max}} \leq \Pi_1^{\overline{max}}$

Since: $\mathcal{H}(\mathcal{X}) = H_F(\Pi_1^{\overline{max}}) = H'_F(\Pi_2^{\overline{max}})$,

and: $H'_F(\Pi)$, $H_F(\Pi)$ are concave, monotonically decreasing in Π and log monotonically increasing in N,

and: $H'_F(\Pi) \leq H_F(\Pi)$ (since $N_r \leq N$)

Then: $\Pi_2^{\overline{max}} \leq \Pi_1^{\overline{max}}$ ■

A. Defining Reachability

Recall that the main correction we have made to the approach in [1] is to address its assignment of a positive "next step" probabilities to all locations - even those impossible to reach. The magnitude of the corrective effect we implement is therefore dependent on the difference between the number of all possible (distinct) locations, N, and the number of locations that are reachable in reality, N_r since Π is monotonically increasing in N_x via $H_F(\Pi, N_x)$ (see Fig. 2). Since the number of reachable locations differs depending on the current location, the maximum number of reachable (i.e. non-zero probability) locations across all possible choices of "current location" is used so as to not underestimate the number of possible reachable locations, and maintain the upper bound. Formally N_r is calculated from an empirical symbolic time series $\mathcal{T} = \{s_1, s_2, \ldots, s_m\}$, with the set of all possible spatial locations being Ω, as $N_r = \max_{x \in \Omega} |\{s_{i+1} : s_i = x\}|$. This is a per-individual, data-driven over-estimation of reachability. Importantly, this data-driven approach prevents the arbitrary over-estimation of possible "next-step" locations, which occurs when techniques such as grid based reachability, unpersonalised "ground cover" maps[2] (e.g. [20]) or arbitrarily sized sets of data-driven possible locations[3] (as previously used in [1] and subsequent work) are used.

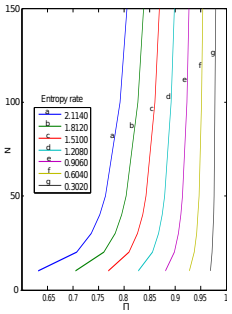

Fig. 2. Plot showing Π & N in $H_F(\Pi, N)$ inc. monotonically.

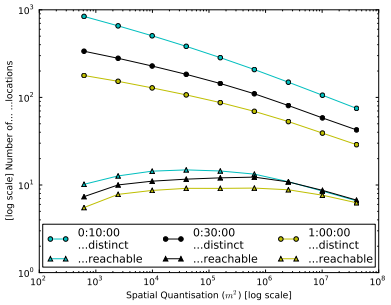

Fig. 3. Graph showing the number of distinct locations vs reachable locations

An important empirical observation is that in using more fine-grained spatial quantizations one significantly increases the difference between N and N_r. This difference, based on examining all trajectories from the Geolife dataset [10], is shown as a log-log plot in Fig. 3 highlighting not only a significant difference between N and N_r in general, but also that this difference is particularly pronounced for more fine-grained spatial quantization levels. As such it is expected (and validated in section VI-C) that the difference between the upper bound calculated by the original formula from [1] and the refined upper bound calculated via this work will be more pronounced for higher spatial quantizations.

VI. An Empirical Evaluation

In this section we implement both the original [1] and our new method for calculating an upper limit for predictability and test it on real world data. The sensitivity of these limits to both temporal and spatial resolution is also investigated. In order to aid replicability a supporting website is available[4].

A. Dataset

The Geolife dataset [16] contains 182 individual's GPS trajectories (longitude/latitude points, points omitted when no GPS was available) of varying lengths and sampling rates over a period of five years. Because many individuals have an insufficient number of data points to obtain an accurate estimate of their entropy, we are only able to consider data for a subset of individuals in our analysis. To determine individuals with sufficient data we calculated the entropy rate estimate

[2]Often people will never visit places they *could*, e.g. commercial buildings.
[3]which varies significantly due to how far each individual has ever travelled
[4]Supporting website: www.cs.nott.ac.uk/~gss/mobility.php

for every individual. If this value fell above the theoretical maximum entropy rate[5], then the corresponding individual was not selected. Of the remaining individuals, plots of the estimated entropy rate as a function of time were studied. Those exhibiting obvious signs of non-convergence were not selected[6]. This left a group of 42 individuals[7]. However, it was not possible to use the whole set of 42 individuals at all spatiotemporal quantizations considered. In particular, far more data is required to even approximately estimate the entropy rate of an individual when combining a fine-grained spatial quantization with a coarse grained temporal quantization (because coarse grained temporal quantization results in a reduced number of points and fine-grained spatial quantization results in significantly more unique locations and hence lower repeating patterns). Instances where this occurred are clearly marked in the results and are discussed accordingly. In this way, we are able to quantitatively examine the majority of spatiotemporal quantizations, and treat any quantization levels which are restricted to a small sample of individuals in a more qualitative fashion.

B. Methodology

As detailed in section IV, an upper bound on predictability can be estimated from symbolised versions of individuals' historic movement logs. To symbolise the GPS data a hierarchical, equal area quantization of the Earth is used [21]. Each individual's upper bound on the maximum predictability is calculated using both the original method proposed by [1] and our new method. The mean across all individuals is then reported, before iterating for each combination of spatial and temporal quantization level. Spatial quantization levels of 618m^2, 2474m^2, 9896m^2, 39586m^2, 158347m^2, 633388m^2, 2.53km^2, 10.13km^2, 40.54km^2 along with temporal quantization levels of $5, 10, 15, 30, 45$ and 60 minutes were investigated. These spatial and temporal quantizations reflect a similar range of quantizations to that previously considered in [3], [4]. Note that more fined-grained spatial quantization is likely to lead to inaccuracies due to GPS error and higher temporal quantizations are unlikely to provide enough data.

When calculating the empirical entropy rate we use the following estimator[8] based on Lempel-Ziv data compression [18], where $\mathcal{T}$ is an individual's observed time-series(that we model as having been generated by the stochastic process $\mathcal{X}$) and Λ_i is the length of the longest pattern that starts at position i, but which has not been seen prior to that point. It has been proven that the estimate $\hat{\mathcal{H}}(\mathcal{X})$ converges to the actual entropy rate, $\mathcal{H}(\mathcal{X})$, when t approaches infinity [18].

$$\hat{H}(\mathcal{X}) = H(\mathcal{T}) = \left(\frac{1}{t} \sum_{i=2}^{t} \frac{\Lambda_i(\mathcal{T})}{\log_2 i} \right)^{-1} \tag{15}$$

C. Results and Discussion

Our results are shown via the three heatmaps in Fig. 4, and empirically confirm that our method not only achieves a tighter upper bound than the original approach in [1] in all circumstances, but that this improvement is sizeable especially at finer grained levels of quantization. The heatmaps compare: *(left)* the mean upper bound on the maximum predictability as computed by the original method from [1]; *(middle)* the mean upper bound as computed by the refined method presented in this work; and *(right)* the mean difference between the two approaches. In all cases these values were computed over the six temporal quantizations and nine spatial quantizations.

All three heatmaps include a black line, above which cells denote spatiotemporal quantizations which a person could not walk across within a single timestep[9]. This is of interest since quantizations above this line have a clear advantage with respect to "next step" prediction. The reason for this is that, while the dataset contains a variety of transportation types, a non-trivial number of trajectories were undertaken on foot[10]. In these trajectories, many "next-steps" will be to the same region as the one they are currently in. Hence, a prediction of "no movement" will *automatically* be correct a predominant amount of the time, solely as a consequence of the granularity chosen. As such spatiotemporal quantizations below this line are often of greater interest. Fig. 4, *(left)*, shows the mean upper bound of predictability over all individuals as computed by the original approach. This acts as a baseline to consider the magnitude of the refinement provided by our new formulation. Results are, as expected, slightly lower than prior work. This is to be expected as we focus on next-location predictability only within trajectories (see section III) - and this means we do not include a comparatively large amount of non-movement, which is easier to predict. At the same quantization levels used in the original work of [1], we report a predictability limit of between 81.45-85.57%, compared to the 93% they reported.

We also compare our results with [3]. At the finest spatial resolution used in that study of 350,000m^2, at a five minute temporal resolution, an upper bound of $\approx 98.5\%$ was reported compared to our 93.05-94.7%. While this is relatively close, as the temporal quantization becomes more coarse the difference increases substantially - where [3] reports $\approx 91.7\%$, at an hour spatial quantization, we report 74.23-78.20%. Such differences, however, are well within expectations - when an individual is sampled at a coarse temporal granularity the uncertainty of their location increases significantly, and this is only emphasized when non-movement is not considered. We note also that these differences have no effect on our subsequent analysis, as this result set remains a cogent baseline against which we can compare the impact of our new method. However, it does emphasize (even before the application of a refined approach to calculating the upper bound of predictability), that previously reported upper bounds are inflated due to the inclusion of significant amounts of non-movement.

Our main contribution is illustrated in Fig. 4, *middle* and *right*, which highlight the importance of incorporating topological constraints in calculating upper bounds on maximum predictability. Across all quantization levels an average refinement of 12.9% was achieved (with a refinement of 17.69% achieved for quantizations below the black line). Excluding the combinations which could not include all 42 individuals

[5]Due to the entropy rate definition, $\mathcal{H}(\mathcal{X})$, and lemma 1 this is $\log_2(N_r)$.

[6]Typically due to a low number of data points.

[7]Specifically, we used individuals with ids: 0-5,7,9,12-17,22,24,28,30,35,36, 38-40,43,44,50,52,55,68,71,82,84,85,92,96,101,104,119,126,153,167,179.

[8]We note this estimator, used in [3], is slightly different to that used by [1]. While both estimators converge to the same value in the limit preliminary experiments on synthetic data suggested the former converges faster.

[9]Based on the *preferred walking speed* of 5km per hour from [22].

[10]$\approx 41\%$ of the labelled Geolife data was denoted "walking" [16].

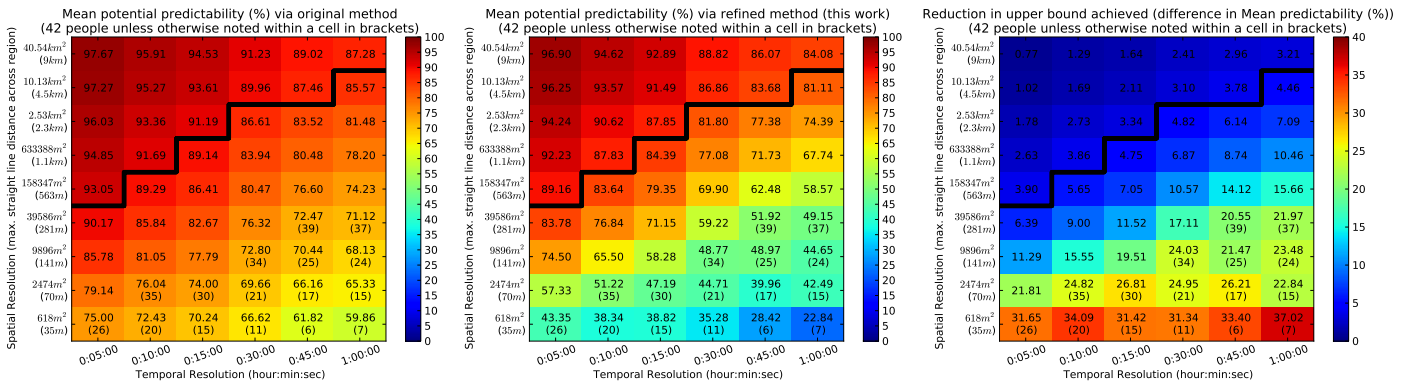

Fig. 4. Predictability upper bound results for varying spatiotemporal quantizations. Dark black line separates the spatiotemporal quantizations one can cross within one time period walking (those below the line) at the preferred walking speed of 5km/hour [22]. LEFT: Predictability upper bounds as defined in [1]. MIDDLE: Predictability upper bounds as defined in this work. RIGHT: Difference in predictability upper bounds between the two methods.

(due to the extra data requirements at fine-grained levels) the average refinement was still 10.4% for the combinations below the line and 6.86% including those above.

However, the corrections to upper bounds of human predictability are particularly of note when the spatial quantization is fine-grained - the level relevant to a large number of interesting pervasive applications (e.g. location-sensitive advertising [5] or detecting route errors [8]). Specifically, the average refinement of the upper bound achieved in this work is between 10-24% for a spatial quantization of 9896m^2, 21-26% for 2474m^2 and 31-37% for 618m^2. While these figures are not always supported by a sample size as large as 42 individuals they do help to show a strong trend in results: as temporal quantization becomes coarser, and spatial quantization becomes finer our approach has a greater impact on the upper bounds of maximum predictability. These results also match the expectations from the theory discussed in section V.

Considering the absolute values in Fig. 4, *(centre)*, it is of note that the refined upper bound varies quite substantially. Of particular interest is when the upper bound is low. This is because, as an upper bound, the value only denotes that one cannot create a predictor that performs *better* - not that it is possible in reality to create one that reaches that level of performance. As such high bounds are of limited value, and should serve only as encouragement to researchers by hinting at the possibility of better predictive algorithms. In contrast low values provide an explicit indication that movement data alone will likely not achieve research and/or industry goals with regard to prediction.

The lowest upper bound of 22.84% occurs under the finest spatial quantization level (618m^2) and at the coarsest temporal quantization level. While the lower number of individuals contributing to the average means that this particular result must be taken with caution, the general trend indicates that attempting prediction at this, or any finer (coarser), spatial (temporal) granularity is unlikely to provide useful performance based on locational trails alone. For a significant number of applications which require high levels of predictability to function (such as ones where predictions will directly result in user interaction) this statement applies to *all* considered temporal quantization levels at the spatial granularity of 618m^2 (and likely 2474m^2) because at those levels even the best possible prediction algorithm will only be correct 57.33% of the time (which seems insufficient to support a robust service). In fact, out of all the spatiotemporal quantization levels considered that individuals could cross on foot, and hence not introducing systematic bias due to the analysis of next step predictability, only 35% (14/37) had an upper bound of more than 70%.

While the low predictability limits highlighted in this work have potential to cast doubts on the use of predictive algorithms based on human movement traces, at least for prediction with spatial granularity less than 9896m^2, it is important to realise that it does not preclude advances beyond these limits through the incorporation of other information sources into prediction. Indeed, such systems have been advocated for numerous years, empirically outperforming systems relying on singular information sources. As such these results contribute solid evidence that such systems *will be required* if performance targets are desired beyond the values reported in Fig. 4. A final point to remember is that these are mean upper bounds. Some individuals and/or groups will have higher predictability, particularly in some datasets (e.g. as discussed in [2]).

VII. Conclusion

In this work we reconsidered the derivation of the upper bound of the maximum (upper limit) predictability of human mobility. Previous results based on a method for computing the upper bound by [1] have suggested high predictability at or above 90% across a range of spatiotemporal quantization levels. Noting an overestimation of the bound in the formulation in [1], due to a lack of topological constraints, an alternate formulation is provided which is shown both theoretically and empirically to provide a lower, and therefore closer to the true, upper bound providing the opportunity to revisit the studies of [1]–[4].

Under a definition of mobility that focuses on human movement, which is of greater use to many applications, we show starkly different results to prior work. Our empirical results show that predictability can decrease almost exponentially as the spatial quantization levels used are refined and that, even at modestly fine spatial quantization levels, predictability

can be lower than required for many real world applications (e.g. 57.33% under a spatial quantization of 2474m^2 and 5 minute sampling). Furthermore, a strong trend is evident that this gets worse as temporal sampling rates are reduced and the predictions are made over a longer time window. Such results provide an important, sobering, look at the role of movement traces for predicting users locations, indicating that on their own (and without integration of further external knowledge) they will not provide sufficient foundation. These results therefore provide solid evidence that in these cases prediction algorithms should either: use a different approach completely (i.e. approaches following different definitions of prediction or mobility); focus on delineating sub-populations with exceptionally regular behaviour; or integrate external variables in order to achieve the desired prediction accuracy to meet the applications performance needs.

ACKNOWLEDGEMENTS

The authors would like to thank our shepherd Silvia Santini and the anonymous reviewers for their valuable comments. This work was funded by the RCUK Horizon Digital Economy Research Hub grant, EP/G065802/1.

APPENDIX A
PROOF OF THEOREM 1: $H(X_n|h_n) \leq H_F(\pi)$

This proof makes explicit the claim in [17, p12] that $H(X_n|h_n) \leq H_F(\pi)$ "represents an appropriate rewritting of Fano's inequality".

Proof: Consider the optimal prediction algorithm, which will always predict x_{ML}. A prediction error occurs when this prediction differs from the individuals actual move (i.e). Let us define the probability of this event as $P(e)$, and denote its associated binary entropy as $H(E)$. Correspondingly:

$$P(x_{ML}|h_n) = \pi = 1 - P(e|h_n) \tag{16}$$

By Fano's inequality [19]:

$$H(X_n|h_n) \leq H(E|h_n) + P(e|h_n)\log_2(N-1) \tag{17}$$

Substituting 16 into 17:

$$H(X_n|h_n) \leq -\pi \log_2 \pi - (1-\pi)\log_2\left(\tfrac{1-\pi}{N-1}\right) \tag{18}$$

Since RHS of Eq. 18 equals Eq. 6: $H(X_n|h_n) \leq H_F(\pi)$ ■

APPENDIX B
PROOF OF LEMMA 1

Proof: Consider a distribution $P(X)$ with a set V of N_r non-zero outcomes and a set Z of N_z zero outcomes.

$$H(X) = \textstyle\sum_{v\in V} P(v)\log_2 \frac{1}{P(v)} + \sum_{z\in Z} P(z)\log_2 \frac{1}{P(z)} \tag{19}$$

$$\leq \log_2 \textstyle\sum_{v\in V} P(v)\frac{1}{P(v)} \tag{20}$$

$$= \log_2 |V| = \log_2 N_r \tag{21}$$

Eq. 19 is by the definition of entropy, Eq. 20 by the fact that when calculating entropy $0\log_2(0)$ is taken to be zero and by definition all $P(z) = 0$ with the inequality due to Jensen's inequality since $\log_2 x$ is strictly convex [23]. ■

REFERENCES

[1] C. Song, Z. Qu, N. Blumm, and A.-L. Barabsi, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[2] B. Jensen, J. Larsen, K. Jensen, J. Larsen, and L. Hansen, "Estimating human predictability from mobile sensor data," in *IEEE Int'l Workshop Machine Learning for Signal Processing (MLSP)*, 2010, pp. 196–201.

[3] M. Lin, W.-J. Hsu, and Z. Q. Lee, "Predictability of individuals' mobility with high-resolution positioning data," in *Proc. ACM Conf. on Ubiquitous Computing (UBICOMP)*. ACM, 2012, pp. 381–390.

[4] W. Qian, K. Stanley, and N. Osgood, "The impact of spatial resolution and representation on human mobility predictability," in *Int'l Workshop Web & Wireless Geographical Information Systems*. Springer, 2013.

[5] J. Krumm, "Ubiquitous advertising: The killer application for the 21st century," *Pervasive Computing, IEEE*, vol. 10, no. 1, pp. 66–73, 2011.

[6] B. Jung, M. Choi, H. Youn, and O. Song, "Vertical handover based on the prediction of mobility of mobile node," in *IEEE Int'l Conf. Pervasive Computing and Communications Workshops*, 2010, pp. 534–539.

[7] J. Froehlich and J. Krumm, "Route prediction from trip observations," *SAE SP*, vol. 2193, p. 53, 2008.

[8] D. Patterson, L. Liao, K. Gajos, M. Collier, N. Livic, K. Olson, S. Wang, D. Fox, and H. Kautz, "Opportunity knocks: A system to provide cognitive assistance with transportation services," in *Proc. ACM Conf. on Ubiquitous Computing (UBICOMP)*, 2004, pp. 433–450.

[9] B. D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. Srinivasa, "Planning-based prediction for pedestrians," in *IEEE/RSJ Int'l Conf. Intelligent Robots and Systems (IROS)*. IEEE, 2009, pp. 3931–3936.

[10] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *Proc. Int'l Conf. World Wide Web (WWW)*. ACM, 2009, pp. 791–800.

[11] Y. Chon, H. Shin, E. Talipov, and H. Cha, "Evaluating mobility models for temporal prediction with high-granularity mobility data," in *Int'l Conf. Pervasive Computing and Communications*, 2012, pp. 206–212.

[12] W. Mathew, R. Raposo, and B. Martins, "Predicting future locations with hidden markov models," in *Proc. ACM Conf. on Ubiquitous Computing (UBICOMP)*. ACM, 2012, pp. 911–918.

[13] P. Baumann, W. Kleiminger, and S. Santini, "The influence of temporal and spatial features on the performance of next-place prediction algorithms," in *Proc. ACM Int'l Conf. on Pervasive and Ubiquitous Computing (UbiComp)*. New York, NY, USA: ACM, 2013, pp. 449–458.

[14] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. Campbell, "Nextplace: A spatio-temporal prediction framework for pervasive systems," in *Int'l Conf Pervasive Computing*. Springer, 2011, pp. 152–169.

[15] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti, "Wherenext: a location predictor on trajectory pattern mining," in *Proc Int'l Conf Knowledge discovery & data mining (KDD)*. ACM, 2009, pp. 637–646.

[16] Microsoft Research. Geolife dataset and user guide: 1.3. [Online]. Available: http://research.microsoft.com/en-us/downloads/

[17] C. Song, Z. Qu, N. Blumm, and A.-L. Barabsi. (2010) Limits of predictability in human mobility (supporting material). [Online]. Available: www.sciencemag.org/content/327/5968/1018/suppl/DC1

[18] I. Kontoyiannis, P. Algoet, Y. Suhov, and A. Wyner, "Nonparametric entropy estimation for stationary processes and random fields, with applications to english text," *Information Theory, IEEE Transactions on*, vol. 44, no. 3, pp. 1319–1327, 1998.

[19] R. Fano, "Fano inequality," *Scholarpedia*, vol. 3, no. 10, p. 6648, 2008.

[20] J. Krumm and E. Horvitz, "Predestination: Where do you want to go today?" *Computer*, vol. 40, no. 4, pp. 105–107, 2007.

[21] K. Górski, E. Hivon, A. Banday, B. Wandelt, F. Hansen, M. Reinecke, and M. Bartelmann, "Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere," *The Astrophysical Journal*, vol. 622, no. 2, p. 759, 2005.

[22] J. Rose, H. Ralson, and J. Gamble, "Energetics of walking," in *Human walking*. Williams & Wilkins, 1994, pp. 45–72.

[23] R. McEliece, *The theory of information and coding*. Cambridge University Press, 2002.