REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
14-03-2018	Final Report	15-Mar-2017 - 14-Dec-2017

4. TITLE AND SUBTITLE
Final Report: Predictability Limits in Human Dynamics as a Function of Data Completeness

5a. CONTRACT NUMBER
W911NF-17-1-0127

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Rochester
ORPA
518 Hylan Building, RC Box 270140
Rochester, NY 14627 -0140

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
70343-NS-II.6

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Gourab Ghoshal
UU	UU	UU	UU		19b. TELEPHONE NUMBER 585-276-7748

RPPR Final Report
as of 17-Oct-2022

Agency Code: 21XD

Proposal Number: 70343NSII **Agreement Number: W911NF-17-1-0127**
INVESTIGATOR(S):

Name: Gourab Ghoshal
Email: gghoshal@pas.rochester.edu
Phone Number: 5852767748
Principal: Y

Organization: **University of Rochester**
Address: ORPA, Rochester, NY 146270140
Country: USA
DUNS Number: 041294109 EIN: 160743209

Report Date: 14-Mar-2018 Date Received: 14-Mar-2018
Final Report for Period Beginning 15-Mar-2017 and Ending 14-Dec-2017
Title: Predictability Limits in Human Dynamics as a Function of Data Completeness
Begin Performance Period: 15-Mar-2017 **End Performance Period:** 14-Dec-2017
Report Term: 0-Other

Submitted By: Gourab Ghoshal Email: gghoshal@pas.rochester.edu
Phone: (585) 276-7748

Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 3 **STEM Participants:** 4

Major Goals: The objective of this project was to quantify the predictability of human dynamics from two points of view: under partial information and with the availability of additional information (e.g., using social-network information to improve the prediction of human mobility, and vice-versa). We investigated to which extent the massive availability of data, provided from a multitude of sources, such as phone call registries, social media posts, and web-browser navigation history, can aid developing algorithms such that the maximum predictability (theoretical) is reached. Here we report the following analyses: (1) the characteristics of a candidate data set, (2) the creation of social graphs using human mobility data, (3) the recency effect on social interactions, (4) the predictability in human dynamics, (5) the predictability in human dynamics under partial information, (6) the interplay between the predictability of social ties and human mobility (all previously reported), plus (7) strong evidence of circadian and ultradian rhythms to human mobility an exciting new development not discussed in the original proposal.

The planned timeline for the execution of these tasks is as follows. Completed percentage in parentheses.
Task 1: Data Curation (90%)
--Data Processing on existing CDR datasets 04/17
--Analysis and identification of other relevant candidate datasets (BrightKite, GoWalla, Twitter) 04/17--06/17

Task 2: Decomposing the Activity Data into Social and Mobility Dimensions (100%)
--Generation of Social graphs/temporal networks from CDR data 05/17
---Extraction of social networks from co-location data 05/17-06/17

Task 3: Determining Recency effect in Social Ties (100%)
---Quantification of the recency effect in tie formation 06/17

Task 4: Predictability under scenarios of data limitations (100%)
---Analysis of datasets and their predictability limits 07/17
---Predictability analysis on windows of observation 07/17-08/17
---Analysis of user activity with varying windows of observation 09/17--10/17

Task 5: Interplay between the predictability of social ties and mobility (100%)
---Information of mobility data given social network 10/17
---Information on social network given mobility data 11/17
---Mutual information analysis (spatio-temporal) of the joint distribution 11/17-12/17

RPPR Final Report
as of 17-Oct-2022

Accomplishments: Quantify the effect of the recency bias on social interactions.

One of our objectives was to quantify the effect of the recency bias on social interactions, similar to what has been observed in human mobility. Our empirical analyses based on mobile phone interactions and physical encounters confirmed our hypothesis on the recency bias in social interactions suggesting that approximately 90% of our social interactions are with the last six people we interacted with, even with people we recently first met.

Quantify the predictability limits and their dependence on partial data observation.

The central goal of our STIR proposal was to provide a quantitative, information-theoretic assessment of the predictability limits in human mobility and social interactions under data limitation constraints. Our new findings indicate that 86%-91% of the overall uncertainty about one's visitation patterns can be uncovered in as few as 18 days of observational data.

Quantify the interdependence between mobility and social interactions

Another major goal of our proposal was to measure the level of interdependence between our visitation patterns and our social interactions. More precisely, we quantified the amount of mutual information shared between the users' locations and who they would interact with at that said location. In other words, to what extent knowing who a person is interacting with is a good indicator of where the said person is and vice-versa. Surprisingly enough, when it comes to co-location interactions, knowing where a person is can deliver, on average up to 75% of the information about the friends the said person is interacting with.

Quantify the frequency components of the cyclic patterns of predictability

Following the accomplishment of the initial goals of our proposal, we looked deeper into the temporal regularities we observed in the human mobility predictabilities. Our objective was to gain further insights into the possible roots of the periodic patterns in the predictability. More precisely, we wanted to identify the extent to which some of our daily activity routines (e.g., work schedules and sleeping hours) could account for predictability regularities. Our findings indicate that in addition to our daily routines (i.e., 24h circadian rhythms), our internal biological clock can also shape our mobility-related behaviors at other frequencies such as the 12h circasemidian cycle.

Quantify the role of the temporal components of cyclic patterns of predictability

Finally, we set as the final goal of this phase of our project the analyses of the temporal components of the predictability regularities. In other words, we wanted to understand when the cyclic patterns are more prominent (e. g., weekdays vs. weekends) and what cycles are more prevalent at different time frames. For such we leveraged on the Continuous Wavelet Transform which makes it possible to also extract the temporal components of the time series in addition to the frequencies. Our findings indicated that the circadian cycles (24h) and circasemidian (12h) are present in all periods and time frames whereas 6h routines are limited to the weekdays.

Training Opportunities: One postdoc, two graduate students have been directly funded by this project. In addition, we trained two undergraduate students that assisted with various aspects of the project, but were not directly funded by the grant.

Notable events:

Postdoc Hugo Serrano attended the prestigious Complex Systems Summer School at the Santa Fe institute from 06/11 --07/07.

Undergraduate Alec Kirkley is a first author on a paper in the second round of review with Nature Communications. He recently graduated and is now in the physics graduate program at the University of Michigan, Ann Arbor, supervised by Anatol P. Rappaport University Professor Mark EJ Newman (the PI's former advisor).

RPPR Final Report
as of 17-Oct-2022

Results Dissemination: The following presentations were made on the results generated by this project:

1. A social influence model based on affiliation: A case study from political parties. Presenter: Jose- mar Faustino da Cruz. 9th conference on complex networks (CompleNet), Northeastern University Boston, MA, (04/18)

2. Structural Invariants in Street Networks. Presenter: Gourab Ghoshal. 9th conference on complex networks (CompleNet), Northeastern University Boston, MA, (04/18).

3. The scaling of crime concentration in cities. Presenter: Ronaldo Menezes (Keynote Address) Latin American Conference on Complex Systems (LANET), Puebla, Mexico, (09/17).

4. Urban socioeconomic patterns revealed through morphology of travel routes Presenter: Gourab Ghoshal (Plenary speaker) 8th conference on complex networks (CompleNet), Inter-University Center Dubrovnik, Croatia, (03/17)

Honors and Awards: (i) PI Gourab Ghoshal has been designated a Empire State Division of Science Technology and Innovation (NYSTAR) Distinguished Researcher within the Center of Excellence in Data Science at the University of Rochester.

(ii) He was also presented with the Annual Excellence for teaching Award by the Department of Physics & Astronomy.

(iii) Postdoctoral researcher Hugo Barbosa was selected to attend the highly prestigious (and selective) Complex Systems Summer School at the Santa Fe Institute NM. This is the preeminent program globally for the training of graduate students, postdoctoral researchers and industry professionals, in the tools and techniques of complex systems.

(iv) Co-Pi Ronaldo Menezes gave the Keynote Address in the first ever major conference held on Complex Networks in Latin America (LANET, Puebla, Mexico, 09/17)

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: PD/PI
Participant: Gourab Ghoshal
Person Months Worked: 9.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Co PD/PI
Participant: Ronaldo Menezes
Person Months Worked: 9.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Hugo Barbosa
Person Months Worked: 9.00 **Funding Support:**
Project Contribution:
National Academy Member: N

RPPR Final Report
as of 17-Oct-2022

Participant Type: Graduate Student (research assistant)
Participant: Diogo Pacheco
Person Months Worked: 9.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Marcos Oliveira
Person Months Worked: 9.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Undergraduate Student
Participant: Alec Kirkley
Person Months Worked: 5.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Undergraduate Student
Participant: Adam Kaplan
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

ARTICLES:

RPPR Final Report
as of 17-Oct-2022

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: PLOS ONE
Publication Identifier Type: DOI Publication Identifier: 10.1371/journal.pone.0183110
Volume: 12 Issue: 8 First Page #:
Date Submitted: 8/28/17 12:00AM Date Published: 8/1/17 8:00AM
Publication Location: United States
Article Title: The scaling of crime concentration in cities
Authors: Marcos Oliveira, Carmelo Bastos-Filho, Ronaldo Menezes, Eduardo G. Altmann
Keywords: Mobility, Scaling, power-laws
Abstract: Crime is a major threat to society's well-being but lacks a statistical characterization that could lead to uncovering some of its underlying mechanisms. Evidence of nonlinear scaling of urban indicators in cities, such as wages and serious crime, has motivated the understanding of cities as complex systems—a perspective that offers insights into resources limits and sustainability, but that usually neglects details of the indicators themselves. Here, we developed a framework to characterize crime concentration which divides cities into regions with the same population size. We used disaggregated criminal data from 25 locations in the U.S. and the U.K., spanning from 2 to 15 years of longitudinal data. Our results confirmed that crime concentrates regardless of city and revealed that the level of concentration does not scale with city size.
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Nature Communications
Publication Identifier Type: DOI Publication Identifier: 10.1038/s41467-017-02374-7
Volume: 8 Issue: 1 First Page #: 2229
Date Submitted: 3/14/18 12:00AM Date Published: 12/20/17 5:00AM
Publication Location:
Article Title: Morphology of travel routes and the organization of cities
Authors: Minjin Lee, Hugo Barbosa, Hyejin Youn, Petter Holme and Gourab Ghoshal
Keywords: Urban Systems, Travel Patterns
Abstract: The city is a complex system that evolves through its inherent social and economic interactions. Mediating the movements of people and resources, urban street networks offer a spatial footprint of these activities. Of particular interest is the interplay between street structure and its functional usage. Here, we study the shape of 472,040 spatiotemporally optimized travel routes in the 92 most populated cities in the world, finding that their collective morphology exhibits a directional bias influenced by the attractive (or repulsive) forces resulting from congestion, accessibility, and travel demand. To capture this, we develop a simple geometric measure, inness, that maps this force field. In particular, cities with common inness patterns cluster together in groups that are correlated with their putative stage of urban development as measured by a series of socio-economic and infrastructural indicators, suggesting a strong connection between urban development and increased connecti
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 17-Oct-2022

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 0-Other
Journal: Physics Reports
Publication Identifier Type: DOI Publication Identifier: 10.1016/j.physrep.2018.01.001
Volume: Issue: First Page #:
Date Submitted: 3/14/18 12:00AM Date Published: 2/1/18 12:00AM
Publication Location:
Article Title: Human mobility: Models and applications
Authors: Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R. James, Maxime Lenormand, Thomas L
Keywords: Human mobility, Human dynamics, Random walks, Origin–destination matrices
Abstract: Recent years have witnessed an explosion of extensive geolocated datasets related to human movement, enabling scientists to quantitatively study individual and collective mobility patterns, and to generate models that can capture and reproduce the spatiotemporal structures and regularities in human trajectories. The study of human mobility is especially important for applications such as estimating migratory flows, traffic forecasting, urban planning, and epidemic modeling. In this survey, we review the approaches developed to reproduce various mobility patterns, with the main focus on recent developments. This review can be used both as an introduction to the fundamental modeling principles of human mobility, and as a collection of technical methods applicable to specific mobility-related problems. The review organizes the subject by differentiating between individual and population mobility and also between short-range and long-range mobility.
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 4-Under Review
Journal: Nature Communications
Publication Identifier Type: Other Publication Identifier: arXiv preprint arXiv:1709.05718
Volume: Issue: First Page #:
Date Submitted: 3/14/18 12:00AM Date Published:
Publication Location:
Article Title: From the betweenness centrality in street networks to structural invariants in random planar graphs
Authors: Alec Kirkley, Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal
Keywords: Planar graphs, urban systems, street networks
Abstract: We demonstrate that the distribution of betweenness centrality (BC), a global structural metric based on network flow, is an invariant quantity in most planar graphs. We confirm this invariance through an empirical analysis of street networks from 97 of the most populous cities worldwide, at scales significantly larger than previous studies. We also find that the BC distribution is ro- bust to major alterations in the network, including significant changes to its topology and edge weight structure, indicating that the only relevant factors shaping the distribution are the number of nodes and edges as well as the constraint of planarity. Through simulations of random planar graph models and analytical calculations on Cayley trees, this invariance is demonstrated to be a consequence of a bimodal regime consisting of an underlying tree structure for high BC nodes, and a low BC regime arising from the presence of loops providing local path alternatives.
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 3-Accepted
Conference Name: New England Regional Conference on Complex Systems
Date Received: 14-Mar-2018 Conference Date: 12-Apr-2018 Date Published: 12-Apr-2018
Conference Location: SUNY Binghamton, NY
Paper Title: The influence of the circadian and ultradian rhythms to human mobility: empirical evidences from location-based check-ins
Authors: Hugo Barbosa, Marcos Oliveira, Diogo Pacheco, Ronaldo Menezes and Gourab Ghoshal
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 17-Oct-2022

Partners

,

I certify that the information in the report is complete and accurate:
Signature:
Signature Date:

Predicting Human Dynamics under Restrictions of Partial or Incomplete Information

Final Report
W911NF-17-1-0127-P00001

1. Introduction

The objective of this project was to quantify the predictability of human dynamics from two points of view: under partial information and with the availability of additional information (e.g., using social-network information to improve the prediction of human mobility, and vice-versa). We investigated to which extent the massive availability of data, provided from a multitude of sources, such as phone call registries, social media posts, and web-browser navigation history, can aid developing algorithms such that the maximum predictability (theoretical) is reached. Here we report the following analyses: (1) the characteristics of a candidate data set, (2) the creation of social graphs using human mobility data, (3) the recency effect on social interactions, (4) the predictability in human dynamics, (5) the predictability in human dynamics under partial information, (6) the interplay between the predictability of social ties and human mobility (all previously reported), and finally (7) strong evidence of circadian and ultradian rhythms to human mobility, an exciting new development not discussed in the original proposal.

2. Description of the Work

In this section, we detail the results obtained during the 9 months of the project for each task in the proposed schedule. Many of these works have been presented, some results have been published, a couple more are under preparation.

2.1. Data Curation

In our analyses, we used different longitudinal data as proxies of human dynamics, thus our results rely on a proper curation of the data sets. In previous works, we have collected two high-resolution data sets: D_1 that corresponds to 6 months of anonymized mobile-phone traces from a large metropolitan area in Brazil; and D_2 that contains 31 months of check-ins from a social media called BrightKite.

2.1.1 Data Processing of D_1 and D_2

As in any data processing phase, we want to ensure the validity and usefulness of the data sets in the light of the purpose of our work, the partial-information aspect. In general human dynamics research, it is a common practice to remove data taking into account the activity of users. Excessively conservative filters have the potential to bias the data when removing less-active users. In fact, at the individual level, data traces are often sparse (e.g., users with few phone calls or check-ins over a period of time) which turns out to be a source of data incompleteness. Any framework to examine such incompleteness needs to care about users' heterogeneity, thus we decided not to impose a general activity-level threshold to our data. Still, we preprocessed the

data sets to remove *(i)* invalid and/or incorrect data points (e.g., entries with missing coordinates), and *(ii)* duplicate entries (e.g., identical check-ins within less than a second from each other).

2.1.2 Analysis of other candidate data sets

To show the generality of our results, we need different source of data, thus we analyzed other candidate data sets. We used the Gowalla data set that was made available after the social media Gowalla was closed in 2012. The data set is similar to the BrightKite (D_2) in terms of attributes and data characteristics: Gowalla was a location-based social network service, launched in 2007, in which users could check-in at places and were rewarded by their check-ins—as of November 2010 the platform had approximately 600,000 users.

In Section 2.6 we also utilized a data set from Weeplace in addition to Brightkite (D_2), and Gowalla. The Weeplace data set contains more than 7M check-ins produced by more than 15,000 Foursquare users visiting over 1M locations in approximately 50,000 cities worldwide from Nov 2003 to Jun 2011. The data corresponds to the Foursquare users who have provided their data to the Weeplace service (now defunct) in order to create dynamic visualizations of their activities.

2.2. Decomposing activity data into Social and Mobility Dimensions

Data sets that contain both the social and mobility of human activities are rare; even when they exist, they usually offer only coarse-grained information solely about one dimension. For instance, the BrightKite data contain spatial mobility data at the GPS resolution but only present a static aggregated friendship network without temporal information of social interactions. The mobile phone data, frequently used in the literature, provides the social interactions (e.g., the data include identifiers of the sender and the receiver of a text message) with temporal resolution at the level of the seconds; but the mobility dimension consists of only approximations of antenna positions—actual spatial resolution vary largely from place to place.

Notably, social interactions and spatial proximity have been documented as intertwined dimensions in human dynamics[1–3]. With this in mind, we can augment our data using the high-resolution mobility data in order to estimate social interactions which allows us to create social graphs.

2.2.1 Estimation of the social graph from co-locations for D_2 and Gowalla data sets

Here we create co-location networks that represent the social graph of the individuals which emerge from encounters between them over time. To define an encounter, we must have a definition of temporal unit that characterizes a gathering of two individuals. For this, we use two different approaches:

Fixed-length discrete bins (FB) an encounter between two (or more) users consists of a check-in/call at the same location within a given—parametrizable—fixed-length discrete time slot (e.g., between 10:00pm and 11:00pm, Apr 12 2010). The edge weight in a social graph using this approach encodes the number of times users have encountered each other. Note that this method does not capture connections between users who have checked-in at different time slots, even though within less than one hour from each other. For instance, if one user checked-in at 10:49pm and another user at 11:04pm, this method would not consider these events as an encounter since the check-ins have occurred at similar time but at different time slots (i.e., 10–11pm and 11–12pm). It is not unreasonable, however, to believe that if those two users indeed know each other, an encounter between them is likely to occur again in the data.

Fixed-length sliding window (SW) an encounter between two (or more) users consists of a check-in at the same location within a given—parametrizable—fixed-length time window (e.g., one hour). This method is more computationally intensive than the previous one but overcomes the aforementioned drawback; this approach suffers, however, from an overstimation problem since encounters in multiple subsequent windows might inflate the connectivity weight.

2.2.2 Generation of the social graph for D_1

Note that, in the case of mobile phone data, we have the social dimension derived from the phone interactions themselves with high temporal resolution. Nevertheless, we also build social graphs using the FB and SW methods.

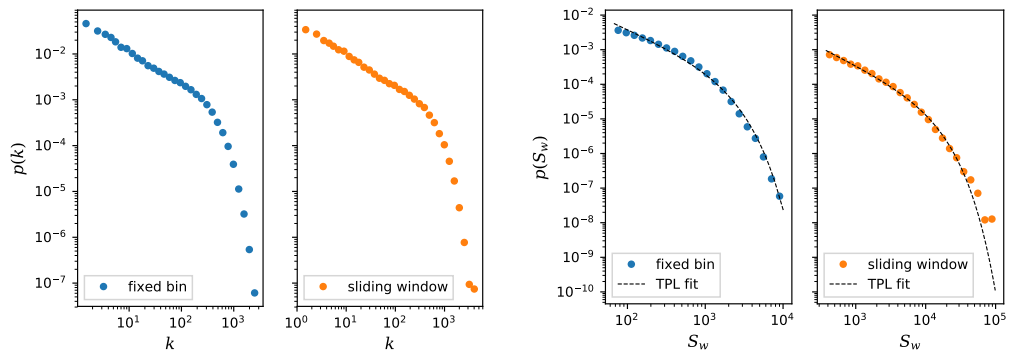
Comparison of the network properties With the social graphs in hands, we compare their basic structural properties, as summarized in Table 1. We found that the networks exhibit similar clustering coefficient $C \approx 0.33$ which means that individuals tend to share friends—a common feature confirmed in the literature[4]. Though the D_1 and D_2 show low degree assortativity, we found that Gowalla exhibit a high assortativity which means that individuals tend to interact with other individuals with similar interaction level. We also analyzed the distributions of degree (i.e., number of friends of each user) and strength (i.e., number of encounters from a user), seen in Figure 1. All the networks exhibit similar features; such properties have been also found in other social networks before which suggests that our approach creates actual social structures. These networks are large graphs and difficult to visualize; still, in Figure 2, we plot the networks of some arbitrary users from the networks—the so–called ego networks: these smaller networks contain only the friends of a certain user (the ego).

Table 1: *Network properties for each of the methods g, number of vertices $|V|$, number of edges $|E|$, clustering coefficient C, edge density d and degree assortativity ρ.*

Data	g	$\|V\|$	$\|E\|$	C	d	ρ
Phone	FB	28088	2210768	0.312918	0.005605	0.044679
	SW	28747	3323624	0.344370	0.008044	0.034015
BrightKite	FB	17307	55531	0.38442	0.000371	0.073291
	SW	20534	107702	0.40205	0.000511	0.070369
Gowalla	FB	57848	453709	0.330514	0.000271	0.410189
	SW	64554	826984	0.310132	0.000397	0.439112

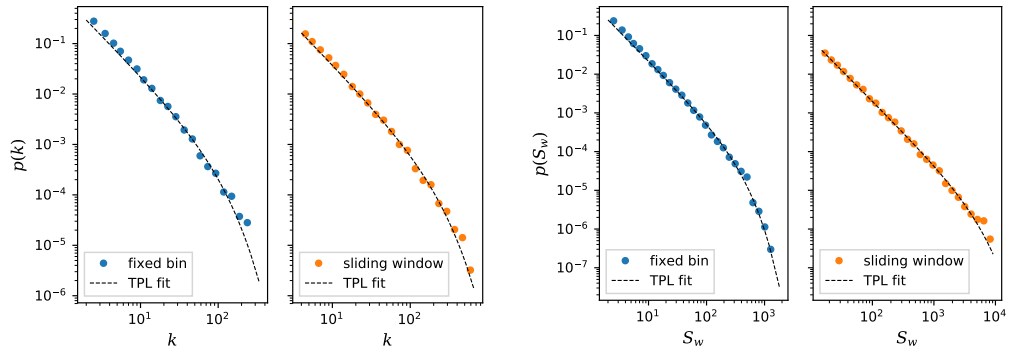
2.3. Recency in social ties

The recency effect refers to the tendency for people to recall the most recent items in a *cognitive list* with far greater accuracy than those that might be in the middle of the list—regardless of the number of times the items appear in the list[5]. The recency effect, or more generally sequential-order biases, has been long known to influence human activity in a variety of socioeconomic contexts[6–9]. In the context of human mobility, we have demonstrated the effect of recency on the visitation patterns: people visit places based not only on frequently-visited locations but also on *recently-visited* locations[10]. To characterize such behavior, we developed a rank-based framework in order to measure the balance between frequency and recency in human dynamics. For this, we



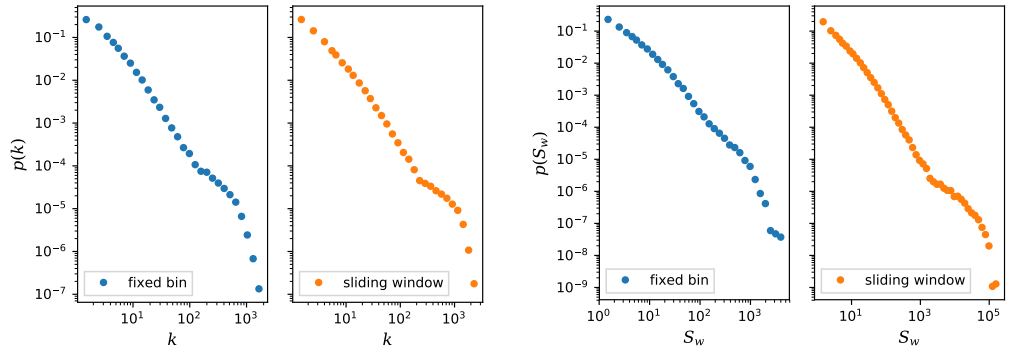
Degree distribution *Strength distribution*

(a) *Phone*



Degree distribution *Strength distribution*

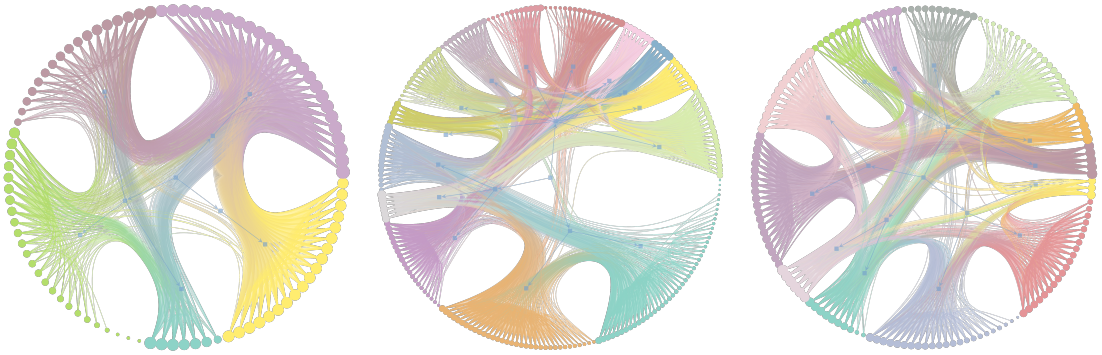
(b) *BrightKite*



Degree distribution *Strength distribution*

(c) *Gowalla*

Figure 1: *Comparing the degree (left) and strength (right) distributions of the* ***(a)*** *Phone,* ***(b)*** *BrightKite, and* ***(c)*** *Gowalla encounters network as produced from both the fixed bins and sliding window methods. In all data sets and for both methods used to build the networks, we have found fat tail distributions; suggesting both methods are able to generate actual social structures.*



(a) (b) (c)

Figure 2: *Co-location networks -representative examples of three ego encounter networks of selected nodes - colors represent the nodes' corresponding stochastic block model partitions. Even for ego-networks of different size, the clustering effect (friends sharing friends) is clear.*

defined two variables: K_f, the rank of frequently visited locations, and K_s, the rank of recently visited locations. At any given time, locations recently visited exhibit low K_s, while locations frequently visited present low K_f. By analyzing the relationship between K_s and K_f, we found that individuals have a stronger preference for returning to recently-visited locations, *even* if these were not visited frequently in the past[10].

The ubiquity of the recency effect on human dynamics suggests the existence of such phenomenon in social interactions. To examine this, we analyze the balance of the recency and frequency in the interactions among people.

2.3.1 Quantification of the recency effect in tie formation

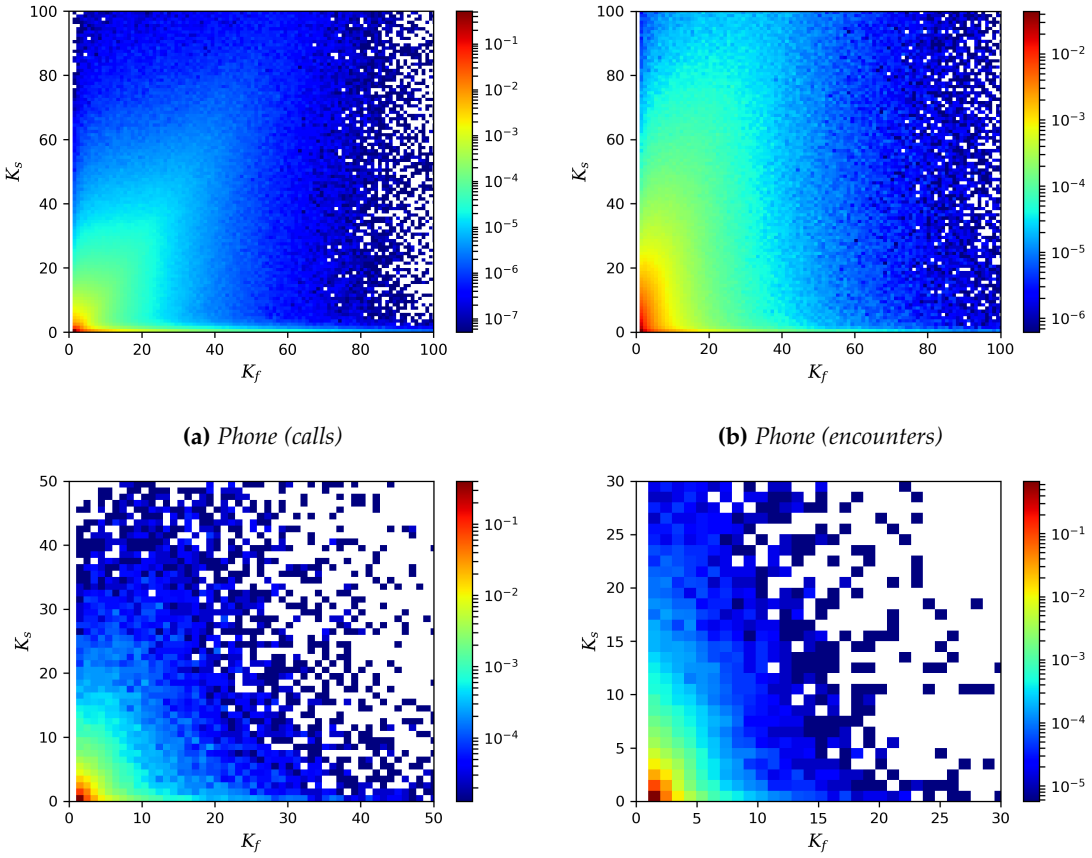
To quantify the recency effect, we created the trajectories of encounters for each individual in each data set then built the ranks K_s and K_f for the users which allow us to analyze the returns of the user to recent–frequent interactions. For this, we constructed the heat maps, seen in Figure 3, that show the number of encounters that already happened X times (frequency) and that happened Y units of time ago (recency). The bottom of each heatmap (i.e., recent interactions) exhibit high number of encounters which means that people tend to interact with people they interacted with recently. Our results confirm that social interactions also exhibit the recency effect.

2.4. Predictability under scenarios of data limitation

We want to examine the predictability of human dynamics from two points of view: under partial information and with the availability of additional information (e.g., using social-network information to improve the prediction of human mobility, and vice-versa). Hence, we analyzed the predictability of users in the data sets regarding the *mobility* and *social* dimensions:

Mobility predictability We use the visitation history of a user and the sequence of visited locations. For this, we transformed our data into time series of locations.

Interactions predictability To evaluate the predictability of social interactions, we create time series representing the sequence of people each user interacted with over time. In the phone data, the sequence represent the true social interaction (i.e., calls or texts). In the other two



(a) *Phone (calls)*

(b) *Phone (encounters)*

(c) *BrightKite*

(d) *Gowalla*

Figure 3: *Recency bias in social interactions. Similarly to human mobility, social interactions also exhibit the recency effect: people tend to interact with people who they interacted with recently. We analyzed (a) the calls between people and used the sliding-window approach to examine the encounters in three data sets: (b) phone, (c) BrighKite, and (d) Gowalla. The increased interaction probability at the bottom for low K_s indicates the recency effect in all cases.*

data sets, we define an interaction as an encounter (defined in Section 2.2.1) between two people.

2.4.1 Analysis of the data sets and their predictability limits

Before delving into the interplay between *mobility* and *social* dimensions, we need to measure their individual capacity of predictability. With representations of mobility and interactions as time series, we can examine users to find out their favorite location or preferred friends which give us an approximation of their predictability: people usually go to their favorite place and usually interact with their favorite person. Still, to characterize individuals, we must also take into account the second favorite one, the third favorite, and so on; in fact, we need to quantify the uncertainty that people have about their favorite ones. For this, we use the Shannon entropy which measures

the uncertainty of a random variable (i.e., the time series here), defined as:

$$S(X) = -\sum_{i}^{N} p(x_i) \log_2 p(x_i). \quad (1)$$

where $p(x_i)$ represents the probability of an individual to visit a specific location i (or to interact to a person i); and N is the number of unique locations (or peers). For instance, if the person k only interacts with the person m, then $p(X = m) = 1$ and $p(x_i) = 0$ for the rest of the world, which leads to $S(X) = 0$ uncertainty: we are sure about the preferences of this person. When a person lacks a preferred location/person, we have the maximum uncertainty (and entropy). Such case occurs when $p(x_i)$ presents the same value for all x_i. Note that, since $\sum_i^N p(x_i) = 1$, this case leads to $p(x_i) = 1/N$, then the maximum entropy is $S_r = \log_2 N$ (the subscript r refers to random). From a practical perspective, entropy quantifies the degree of *uncertainty* about future events; for example, with $S(X) = 3$, the average number of next possible locations/interactions to $2^{S(X)} = 8$. Entropy-related measures have proven themselves useful as a powerful estimator for the limits of predictability of human trajectories at different levels such as shoppers visitation patterns[11] and wireless cell usage[12].

Note that S_r regards to the number of unique locations/interactions while S incorporates the frequency. They fail to consider, however, the temporal sequence that locations are visited or that a person interacts with people. A person may exhibit high unpredictability with respect to frequency, but the sequences that this person follow might present patterns. For example, consider a sequence of going to work, then to the gym, and going home. If a person follows such sequence every day, the absence of preferred location makes the uncertainty to approach $S \approx S_r$, its maximum; yet, we can see a clear pattern repeating itself—this person is actually quite predictable. To evaluate the uncertainty of the flow of information, we have to analyze the entropy rate of the time series which gives us the intrinsic uncertainty of the random variable including any temporal pattern. In our analysis, we estimate the entropy rate using the Lempel–Ziv algorithm[13,14].

Nevertheless, the aforementioned entropy quantities only capture an indirect aspect of the predictability of human dynamics. We want a precise characterization so we measure the probability to correctly predict future locations, given a past series of observations[11] with respect to a particular definition of the entropy. Let us denote this probability Π, and its maximal value Π^{max}. For instance, a person with $\Pi^{\max} = 0.7$ means that we can correctly predict future locations 70% of the time using a predictive algorithm, whereas 30% of the time the whereabouts of this person are indistinguishable from a random trajectory. In other words, the quantity $\Pi^{\max}$ is the upper limit for any predictive algorithm to correctly predict the future whereabouts of an individual.

For all data sets, we calculated each entropy and its $\Pi^{\max}$ predictability. Figure 4 shows the distribution of mobility entropies (S_r, S, and S_t) and the respective maximum prediction probabilities (Π_r, Π, and Π_t) for our three data sets. In the left panels, we measure the entropies (uncertainties) while in the right panels we estimate the maximum mobility predictability using these entropies. Entropy S_r (in blue) measures the uncertainty while considering only the number of visited locations; S (in green) incorporates the frequency distribution, and S_t (in orange) incorporates the sequence of location visits; progressively, showing reduced uncertainty in the human mobility, i.e., providing lower search spaces for the next location. The maximum predictability (right panels) presents dramatic increase as we move from knowledge of just the number of locations to the temporal sequence of location visits. Maximum human mobility predictability varies significantly among data sets; the phone data not only has higher predictability, but also presents clearly distinct progressive values for the entropies.

Similarly, Figure 5 presents the distributions of social interaction entropies (S_r, S, and S_t) and the respective maximum prediction probabilities (Π_r, Π, and Π_t). The social interaction results

are quite consistent when compared to the human mobility ones. However, the predictability patterns exhibited for the data based on *potential* encounters are similar to each other, but clearly different from the phone data where the real social interactions were captured. Finally, it is not clear whether human mobility or social interaction are intrinsically more predictable than the other.

2.4.2 Predictability analysis on windows of observations

Different sources of data allowed us to show that people are predictable in their mobility and their social interactions. Researchers still fail to quantify, however, the amount of data needed to achieve such predictability. To tackle this question, we manipulate our data in order to examine the impact of data loss on the predictability. Specifically, we apply different types of data–loss filters (e.g., noise, limited observational window) on the complete data to generate synthetic partial data. Here we declare the complete data as the ground truth of human predictability—the intrinsic uncertainty present in human mobility and social interactions. With this assumption, we can estimate the amount of uncertainty explained when we use partial data compared to the inherent uncertainty of human behavior—we can quantify the predictability gain as a function of the amount of information available.

To evaluate this gain, we need a tool to measure how different the partial data are from the complete data regarding their predictability. As a first approach, we analyze the partial trajectories of individuals which are created using a limited observational time window. A trajectory of a person is represented by the mass function $p(x)$ that gives the probability that this individual is at the place x. To describe an incomplete trajectory, we represent the trajectory of an individual using t units of time using the mass function $q_t(x)$. Thus, $q_t(x)$ gives the probability of an individual be at place x given that we have data that contain t units of time. Here we evaluate the distance between the complete trajectory (i.e., ground truth) and a given incomplete trajectory (i.e., partial information). Hence, we employ the cross-entropy λ_{Π}^{t} that measures the distance between q_t and p, defined as the following:

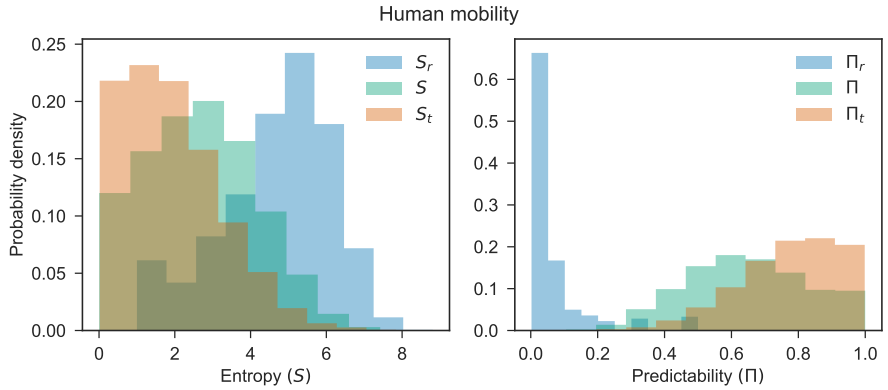
$$\lambda_{\Pi}^{t} = -\sum_{x} p(x) \log q_t(x).$$

In practical terms, the cross-entropy between p and q_t gives us the amount of information that q_t explains of p. As our synthetic partial data becomes more complete, the partial data approaches the complete data, thus $q_t(x)$ explains well $p(x)$ and λ_{Π}^{t} reaches its maximum value. To examine limits of predictability of human dynamics under restriction of partial information, we measured λ_{Π}^{t} for each user in the BrightKite, Phone, and Gowalla data sets, and calculated the average λ_{Π}^{t} in each data set, as depicted in Figure 6.

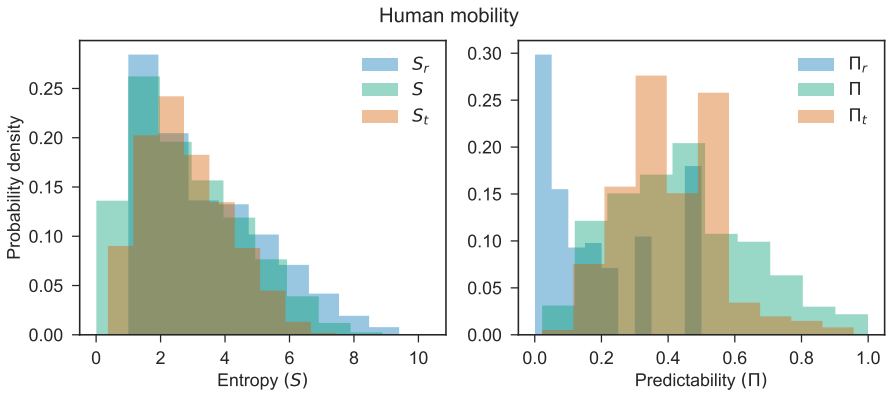
In all cases, we found that the gain of information for observing individuals for one more day decreases with the window length. As shown in Figure 6, λ_{Π}^{t} quickly increases with the size of the observational window t until it reaches a point of slow increasing pace. Note that both data sets share similar points where the pace starts to decrease. The cross-entropy λ_{Π}^{t} enables us to assess the amount of data necessary to reach similar predictability of a full data set. Still, we also need to take into account the differences between users.

2.4.3 Description of user distribution based on activity

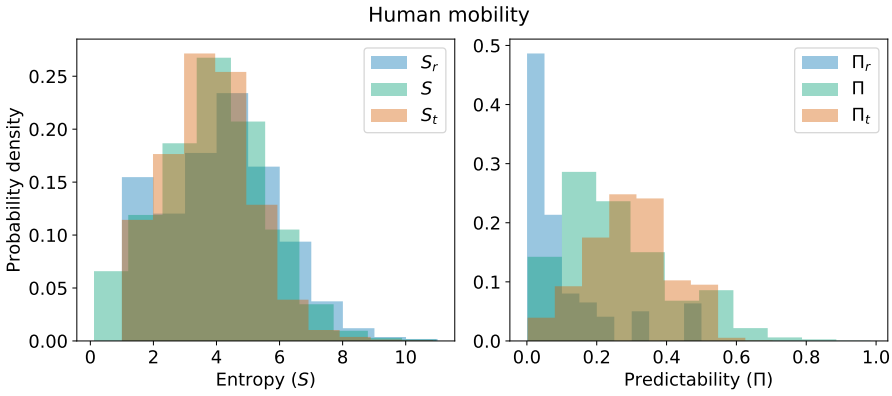
As in other human facets, human mobility and social interactions tend to exhibit a high degree of heterogeneity. If on the one hand most people use their mobile phones a few times a day, some users, on the other hand, make hundreds of calls a day. Hence the characterization of the individual-level dynamics (and their predictability) needs to consider the very different activity



(a) *Phone*

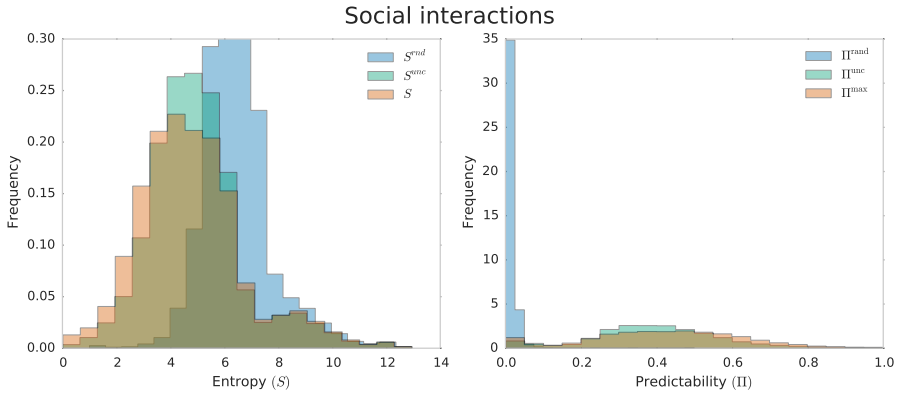


(b) *BrightKite*

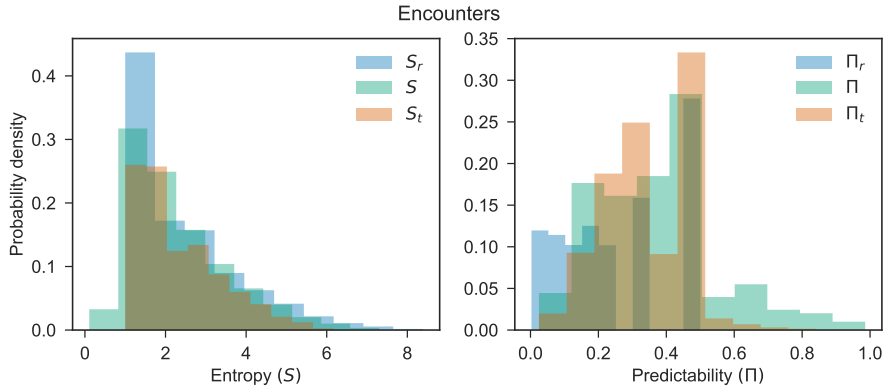


(c) *Gowalla*

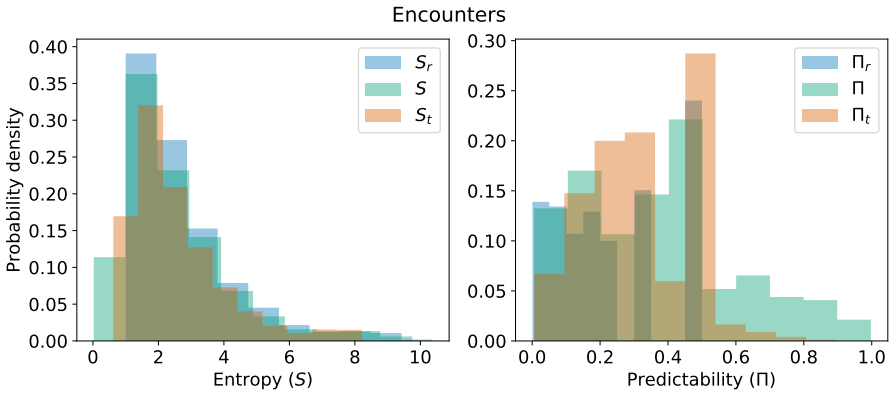
Figure 4: *Distribution of mobility entropies and prediction probabilities for three distinct measurements of entropy, in three data sets. In the left panels, we measure the entropies (uncertainties) while in the right panels we estimate the maximum mobility predictability using these entropies. Entropy S_r (in blue) measures the uncertainty while considering only the number of visited locations; S (in green) incorporates the frequency distribution, and S_t (in orange) incorporates the sequence of location visits; respectively, showing often reduced uncertainty in the human mobility, i.e. providing lower search spaces for the next location. The maximum predictability (right panels) presents dramatic increase as we move from knowledge of just the number of locations to the temporal sequence of location visits. Maximum human mobility predictability varies significantly among data sets.*



(a) *Phone*



(b) *BrightKite*



(c) *Gowalla*

Figure 5: *Distribution of social interaction entropies and prediction probabilities for three distinct measurements of entropy, in three data sets. In the left panels, we measure the entropies (uncertainties) while in the right panels we estimate the maximum social interaction predictability using these entropies. Entropy S_r (in blue) measures the uncertainty while considering only the size of the social network; S (in green) incorporates the frequency distribution of the interactions, and S_t (in orange) incorporates the sequence of social interactions; respectively, showing often reduced uncertainty in the human mobility, i.e. providing lower search spaces for determining the next interaction (who). The maximum predictability (right panels) presents increase as we move from knowledge of just the size of the social circle to the temporal sequence of interaction.*

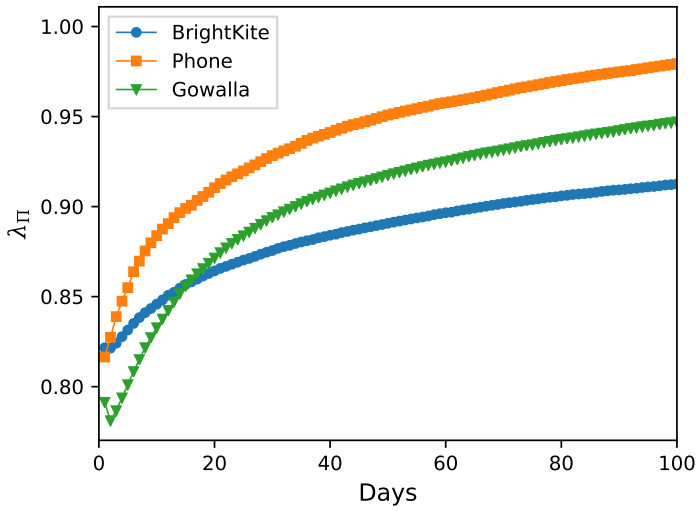


Figure 6: *The impact of partial information on the predictability of human dynamics. As we increase the observational window, the partial data increases its similarity to the complete data. This gain of information, however, stabilizes at a certain point; from this point, the gain for observing individuals for more one day decreases with the window length. In the plot, each data point represents the average λ_Π^t of the users in a data set.*

patterns from people. For such, here we analyze the basic statistics of the user in the data sets. Figure 7a depicts the probability distribution of the number of records per day. Note that all data sets present distributions that exhibit similar skewness which allows the existence of user with more than 100 records a day. If we want to evaluate individual-level predictability, we need to take into account such disparities (as we show in Section 2.4.4).

We also need to examine the regularity that users create records in the data sets. So we evaluated the fraction of the days that a user creates a record from the complete observational time frame of this user (seen in Figure 7b). From this analysis, we found that most of the users in Brighkite use the service sporadically; Gowalla exhibits similar characteristic but also presents quite active users. The data set from mobile phone, however, contains users with a wide range of usage patterns with almost no bias towards a specific type of user.

2.4.4 Analysis of Temporal Dependence of Predictability for Groups of Users

In this section we explore the diversity of activity patterns of users by analyzing their predictability in separate groups. In addition, we analyze the temporal dependence of the limits of predictability by aggregating activities by weekday and hour. Our proposed approach here was to dis-aggregate information by first looking at the distribution of users based on their frequency and exploratory levels of activity. We propose three classes of activity to analyze groups of users based on the following criteria:

Singular Location Visited (S) – group users based on the number of different locations where their activity took place, i.e. the number of different cell towers in the phone data set or the number of different places checked in the BrightKite and Gowalla data sets. See results in Figure 8.

Geographical Coverage (r_g) – users are grouped based on their geographical exploration pattern,

measured by the *radius of gyration* in kilometers. Lower r_g means a user tend to perform their activities (calls/check-ins) in locations close to each other. See results in Figure 9.

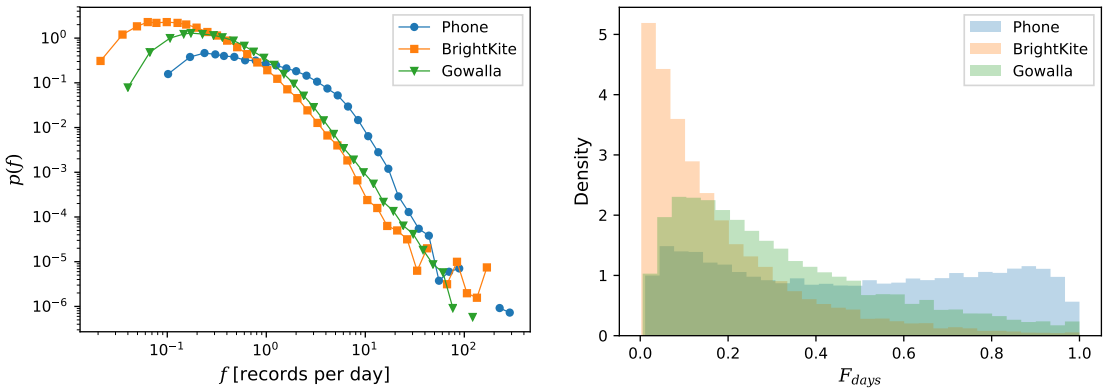
Monthly Frequency of Activity (C_M) – users are grouped by the average number of calls/check-ins per month, regardless the uniqueness or the distance of the location where the activity is performed. See results in Figure 10.

With the defined classes, we conduct a multivariate analysis between the level of activity in each class, the temporal dynamic pattern observed weekly, and the generalization obtained using three data sets.

First, by analyzing the activity levels for the proposed classes of users (unique locations S, geo coverage r_g, and frequency C_M), we notice for all data sets, the human mobility predictability decreases as S, r_g, and C_M increase. That is, it is harder to predict the mobility of users with more *diverse* routes; regardless if diversity is given by the number of unique locations visited, the geographical area explored, or the amount of monthly data (traces). This effect is less pronounced when comparing the geographical coverage in the BrightKite (Figure 9b) and Gowalla (Figure 9c) data sets.

We can split the temporal analysis to compare *weekdays vs. weekends* and *day vs. night*. According to the phone data set (figures 8a, 9a, and 10a), users are less predictable during the week, while from the location-based social network data sets, users are less predictable on the weekends. We also observe this dichotomy when comparing day and night predictability; that is, user mobility is more predictable at night for the phone data. The distinction between the predictability on weekdays or weekends is again less pronounced in the geographical coverage groups in the social network data sets (figures 9b and 9c).

Despite being unanimous in the relation of predictability and activity diversity, the data sets could not provide a general interpretation of the limits of human predictability under the temporal aspects. Although from the phone data the predictability is higher during nights or weekends, regardless the criteria of classes (S, r_g, and C_M) analyzed, the results from the social network data



(a) **(b)**

Figure 7: *People present distinct activity patterns. The probability distribution of the number of records per day (a) shows the possibility of the existence of users with more than* 100 *records a day. Also, (b) the fractions of the days in each users' observational time frame in which a user has created a record (i.e., phone call or check-in) shows that the regularity which users create a record in the data sets present a wide range of usage patterns in mobile phone, while most of users of Brightkite are sporadic users.*

sets are consistently the opposite. Probably, this difference is due to the nature of these data sets where the phone data captures work habits while the social data encompass more leisure time. Moreover, the quasi-collapsing of groups of users for all classes at night in the phone data (figures 8a, 9a, and 10a) suggests a lower level of activity (records) during this period.

2.5. Interplay Between the Predictability of Social Ties and Mobility

2.5.1 Locality and social interactions

In order to better understanding this interplay between location and social interactions, a first natural step is to investigate to what extent knowing the location of a person can help us to know with whom they are going to interact (i.e., meet or call). In other words, we want to quantify how much information is shared between our visitation patterns and social interactions, or, how much could one learn about individuals' whereabouts just by knowing with whom they are interacting with (and vice-versa). A natural quantity of choice in this case would be the Mutual Information (MI) between location and interactions. We can define the Mutual Information between a user's visitation and interaction sequences as

$$MI(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

where X represent a vector with the visitation sequence of a user and Y who they interacted with at each of these locations such that $p(x)$ is the fraction of the visits to a location x, $p(y)$ is the fraction of the interactions with the contact y and $p(x,y)$ is the probability of said user meeting the contact y at location x. Since the mutual information is given in bits, we used the normalized mutual information, which we termed as NMI defined as

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}}$$

where $H(X)$ corresponds to the Shannon entropy of X as defined in Eq. (1). The choice for the normalized version of the MI stems from the fact that instead of giving a value in bits, it is a dimensionless quantity in the interval $[0,1]$, such that two completely independent random variables have an expected NMI of 0 whereas two variables that are completely dependent one another have an $NMI = 1$.

Figure 12 shows the NMI as measured for each user in the data suggesting a high degree of shared information between location and interactions. In the context of geo-tagged check-ins (orange, green and red areas), it means that encounters with particular contacts tend to take place at specific locations, such as workplaces and school, which is not exactly surprising. Moreover, our results show that not only face-to-face interactions tend to occur at particular locations but also mobile phone contacts (blue area), meaning that there is also an interdependence (to a lesser degree) between where we are and who we call. Although it might be an intuitive and even unsurprising derivation, the quantitative result is of high relevance in the context of this work, especially in the scenarios where one dimension (e.g., social or spatial) is unknown.

Furthermore, similarly to what we did for other quantities, we investigated the temporal regularities in the NMI by grouping the activities by day of the week and hour of the day (Figure 12) and averaging across all users. Once again, we can clearly see the daily regularities in the data indicating that the amount of information one can potentially gain by inspecting solely one of the activity dimensions about the other has has a marked temporal regularity. A closer look at each of the individual curves reveals an interesting difference between the location-based check-ins

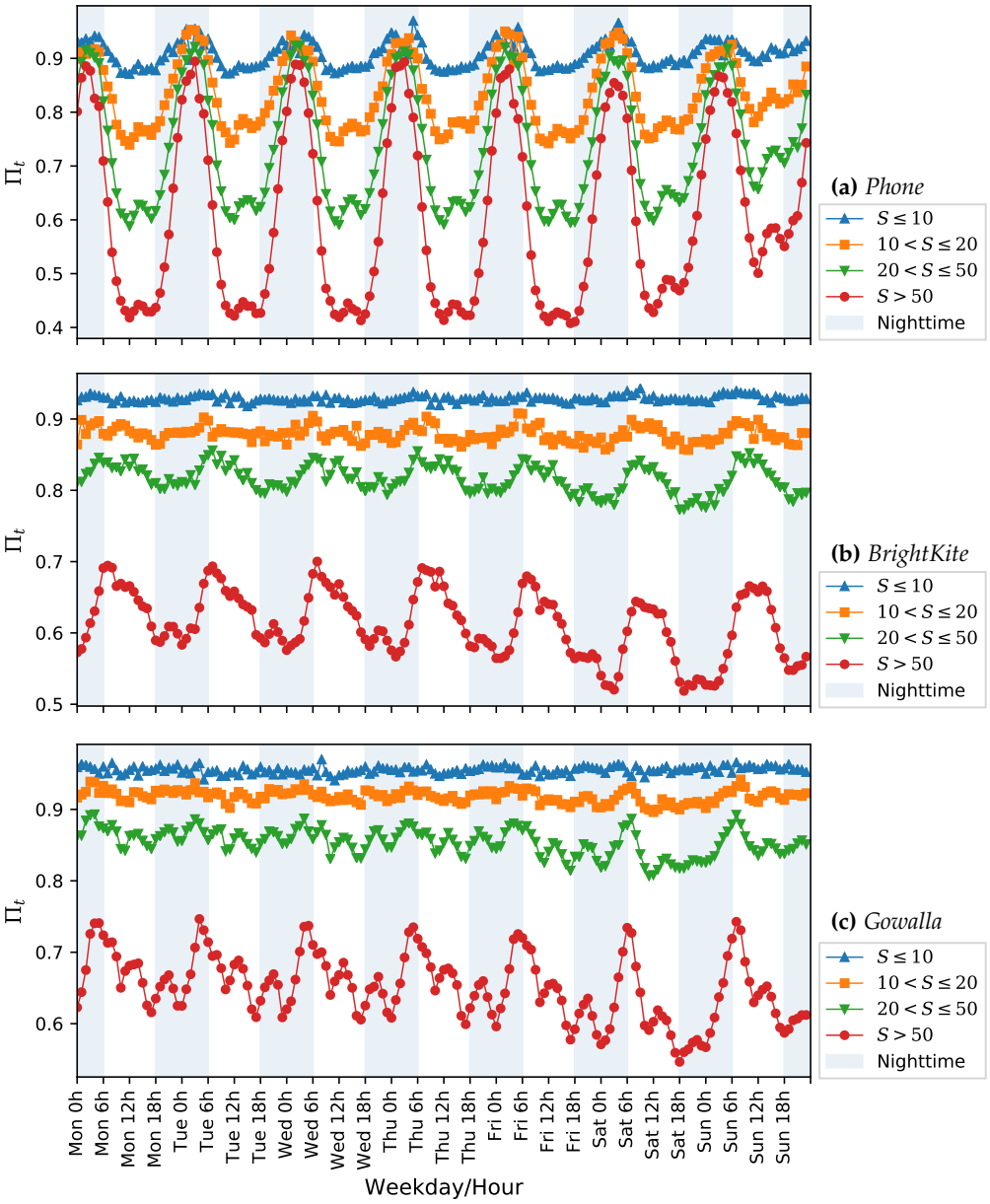


Figure 8: *Singular-locations-visited (S) groups predictability. Temporal dependence of the limits of predictability by the hour and the day of the week. For all data sets, predictability decreases for groups visiting more number of unique locations. Human trajectory is more predictable at night and on weekends when using the phone data (**a**), but at day and on weekdays for the social network data sets (**b** and **c**).*

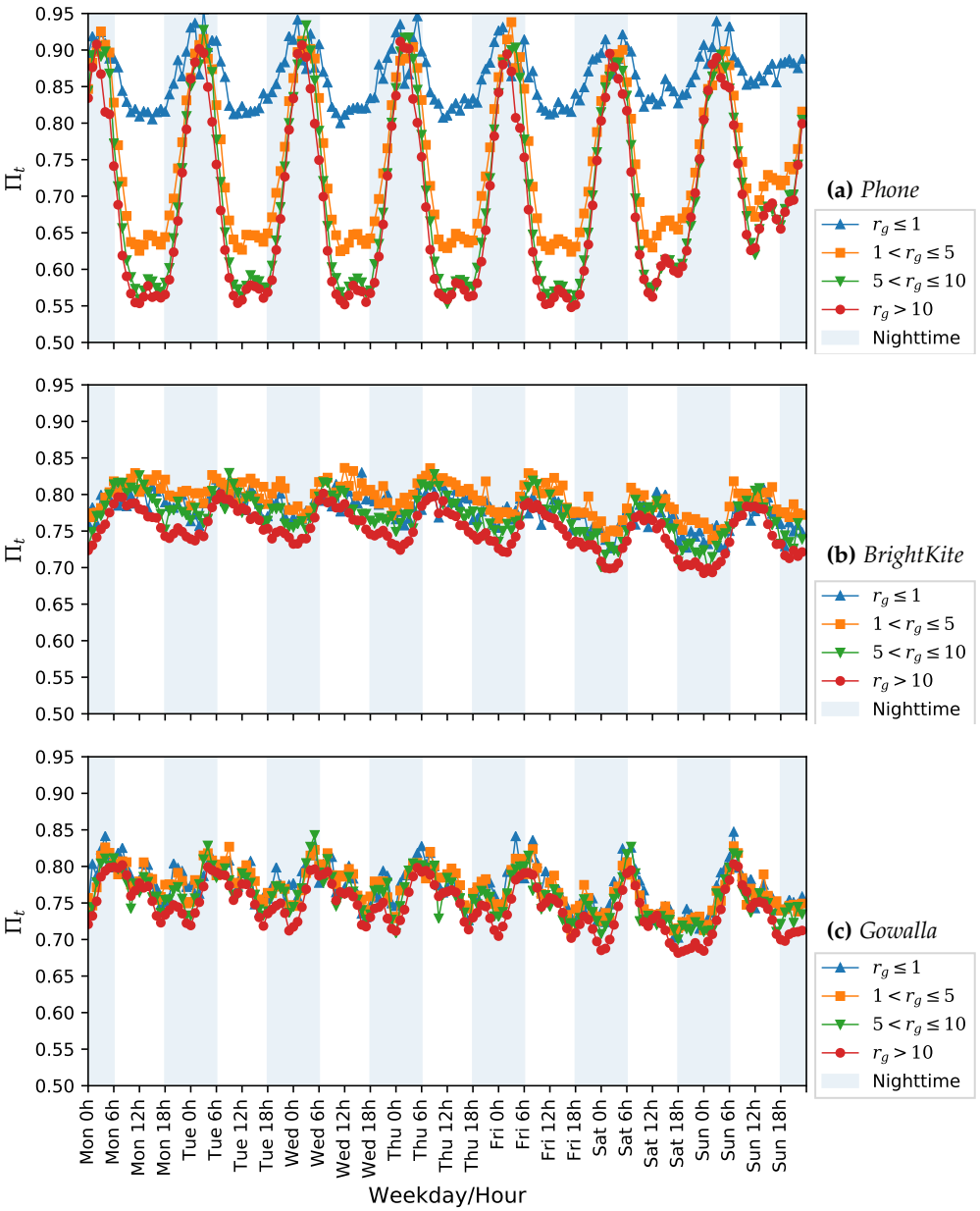


Figure 9: *Geographical-coverage (r_g) groups predictability. Temporal dependence of the limits of predictability by the hour and the day of the week. For all data sets, predictability decreases for groups with wider radius of gyration. Human trajectory is more predictable at night when using the phone data (**a**), but at day for the social network data sets (**b** and **c**).*

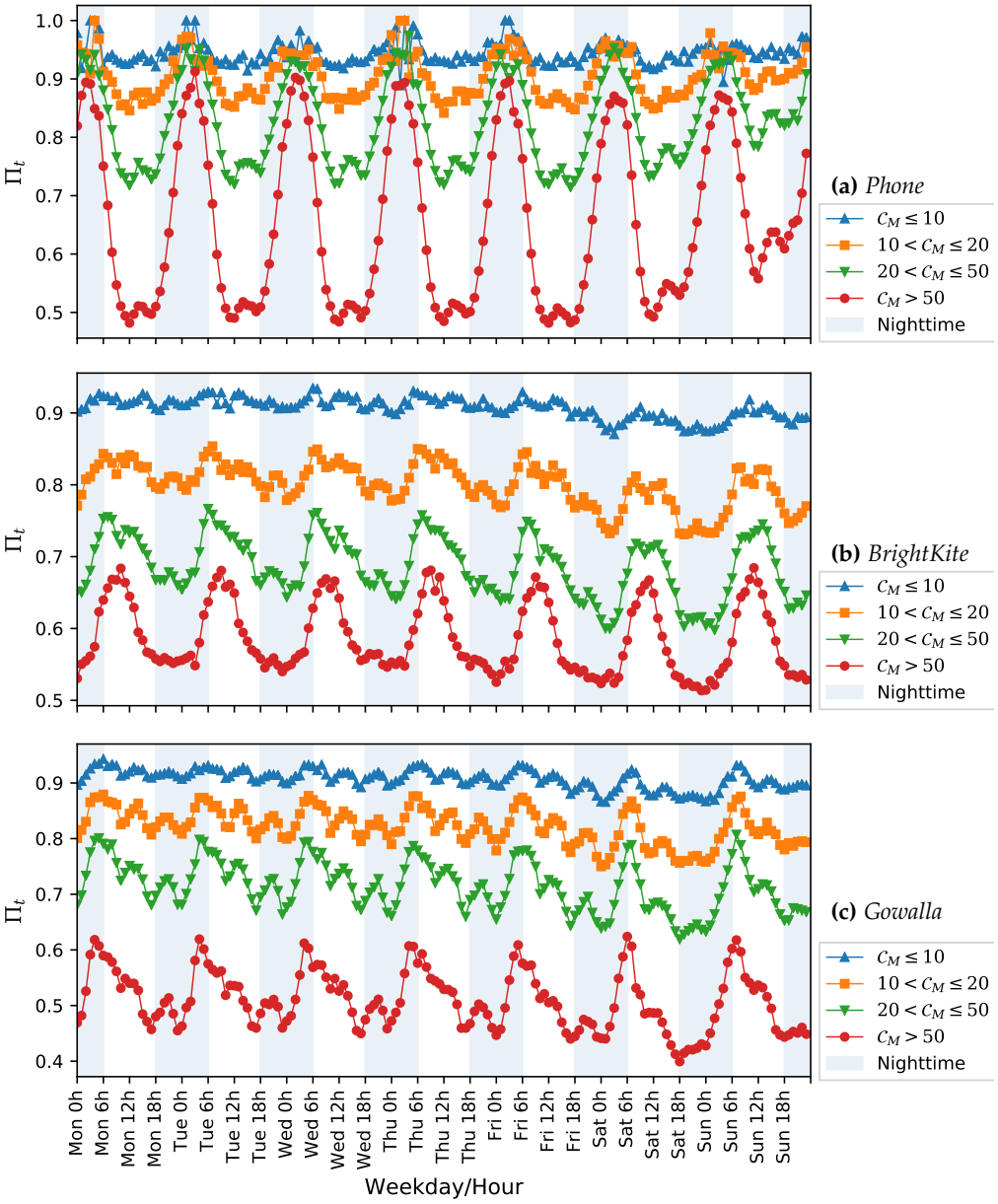


Figure 10: *Monthly-Activity-Frequency (C_M) groups predictability. Temporal dependence of the limits of predictability by the hour and the day of the week. For all data sets, predictability decreases for groups with higher monthly activity frequency (i.e. more average records). Human trajectory is more predictable at night and on weekends when using the phone data (**a**), but at day and on weekdays for the social network data sets (**b** and **c**).*

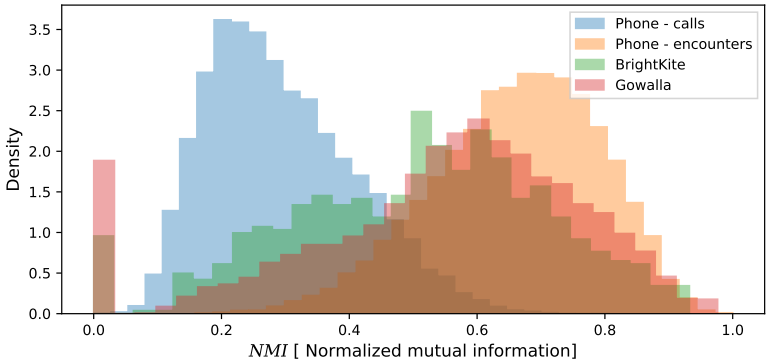


Figure 11: *Overall normalized Mutual Information between locations and their interactions (encounters and phone contacts), computed for each individual user in the datasets. Indeed, in all three datasets we can observe a marked interdependence between locality and social interactions, evidencing that encounters with particular contacts tend to take place at specific locations, such as workplaces and school. Moreover, our results show that not only face-to-face interactions tend to occur at particular locations but also mobile phone contacts (blue area), meaning that there is also a interdependence (to a lesser degree) between where we are and who we call.*

and the mobile phone usage. As we can see, most of the mutual information in the phone data is concentrated in the nighttime, whereas in the location-based check-ins most the NMI is peaked during the daytime. It can suggest that location-based social platforms and phone calls are used in different ways. For instance, phone calls late in the night during weekdays are rare and tend to occur at very few locations (mostly home) and to be targeted to very specific contacts such as a partner or close friend. Social encounters, on the other hand, tend to occur mostly after the regular working hours, hence the offset in the NMI between the two classes of data.

2.6. Circadian and Ultradian Rhythms to Human Mobility

In the last step of the project, we looked deeper into the temporal regularities in human mobility predictability. We used the Continuous Wavelet Transform (CWT) to examine the time series in the frequency domain. The wavelet transform is frequently used to extract from a time series both their time and frequency components. The method has a long history of successful applications to a variety of domains such as climate prediction[15], digital image processing[16], and medical imaging[17] to name a few.

The wavelet transform of a discrete sequence $Y = \{y(1), y(2), \ldots, y(N)\}$ having observations with a uniform step δt can be defined as the following:

$$W_Y(s,n) = \sqrt{\frac{\delta t}{s}} \sum_{t=1}^{N} y(t) \psi^* \left[\frac{(t-n)\delta t}{s} \right], \tag{2}$$

where the '$*$' denotes the complex conjugate and s is the wavelet scale. The wavelet transform can be seen as a convolution of X with a translated and scaled version of a wavelet function $\psi(\cdot)$. By varying the scale s and translating over time (i.e., varying n), we construct a representation of the amplitude of the different periodic features of X and how they vary with time. In our analyses, we used the Morlet wavelet due to its improved frequency resolution in comparison with other candidate functions[15,18].

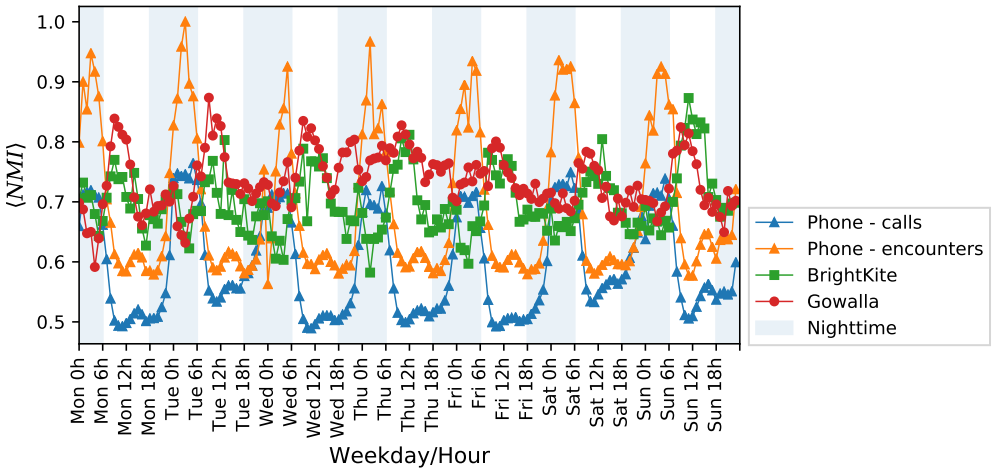


Figure 12: *Temporal regularities in the Normalized Mutual Information . between locations and interactions. Here we show the average of the amount of information that is shared by location and interaction sequences.*

2.7. Wavelet Analyses

Recall that the entropy of a user can be computed using Eq. (1). Then, given a particular definition of a mobility entropy S and the number of visited locations N, this quantity can be calculated by inverting the relation

$$S = H(\Pi) + (1 - \Pi)\log_2(N - 1), \tag{3}$$

from which we see that this measure is a function of the number of locations N and the functional form of the entropy S. The quantity $H(x)$ is the well-known binary entropy function with general form $H(x) = -x\log_2 x + (1 - x)\log_2(1 - x)$.

Our activity routines are characterized by temporal regularities with time and frequency components. Here we measured the entropy and predictability values for individual discrete-time bins, treating the visits occurring within each bin as an independent sequence. We define the bins to be one-hour-long windows, since the majority of our activities or visits tend to be at least one hour. As we are also interested on the daily predictability changes, we include the weekdays as a second-level in this binning scheme. Each bin t represents the activities occurring within each hour of the week (e.g., Monday/9h-10h, Thursday/15h-16h), for which we compute the entropy and estimate the predictabilities values Π_t^u for each person u, using Eq. (1) and Eq. (3).

We first analyzed the aggregated patterns of the entire population for each data set. For this, we calculated the sample mean of the predictabilities Π_t across all users within each time bin t. In Fig. 13, we can see the average Π_t with a remarkable 24h periodic patterns in all three data sets, having high predictability periods during evenings and nighttime hours (peaking around 4-5am), and secondary peaks between noon and 5pm.

Though all data sets are from location-based social network platforms, the predictability amplitude in the Weeplace data is much larger than what we observed in the other data sets. A possible cause is the fact that the Weeplace data was provided by Foursquare users who were interested in visualizing their check-in history. It is reasonable to believe that these users were, on average, more active (in terms of their check-in history) than a regular LBSN user. In fact, the median number of check-ins of BrightKite and Gowalla users were 11 and 25 respectively, whereas for the Weeplace this figure was 329 check-ins. Also, a median Weeplace user has visited 131

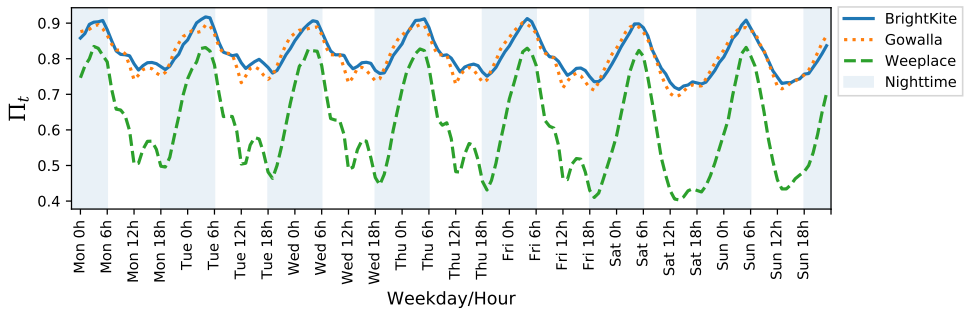


Figure 13: ***Time-dependent predictabilities*** *The average predictability Π_t across all users exhibits daily peaks (4-5am) and secondary peaks (12-5pm) of predictability throughout the time series.*

single locations while for BrightKite and Gowalla these numbers were 8 and 19 unique locations respectively.

With an information-theoretic measure instead of a simpler quantity such as the relative location frequencies or the pure location diversity, we have encapsulated in a single quantity both the location diversity and their relative frequencies. To decompose the mobility regularities into their different time-frequency components, now we perform different instances of time-frequency analyses, leveraging on on the continuous wavelet transform of population and individual-level predictability data. For simplicity, we examine here the rescaled and centered version of the time series $\hat{\Pi}_t = (\Pi_t - \mu_{\Pi_t})/\sigma_{\Pi_t}$, where μ_{Π_t} and σ_{Π_t} denote respectively the mean and the standard deviation of Π_t.

As shown in Fig. 14a, $\hat{\Pi}_t$ exhibits remarkably similar curves in all three datasets, including most of their main and secondary components. Though the data sets are from different sources, the three data sets seem to capture the same regularities. Such finding suggests that the temporal variation of the mobility diversity is activity-independent and therefore is likely to be a characteristic manifestation of the underlying human dynamics.

To analyze these regularities, we evaluated the wavelet transform of $\hat{\Pi}_t$ (Fig. 14b,d,e) and found a strong spectral agreement in all data sets. Our finding revealed major peaks of high predictability during the nighttime and valleys around noon. Minor peaks are also observed in the afternoon during the weekdays. The wavelet transform represents the power level of both the periods of the regularities (y axis) and the temporal localization when they happened throughout the time series (x axis). Moreover, a 95% significance region is represented as the areas delimited by the solid black lines whereas the uncertainty region due to boundary effects is under the hashed area. For details on the computation of the significance and uncertainty regions we refer the interested reader to Ref.[19].

For the Weeplace data, the wavelet analysis (Fig. 14b) reveals that the circadian period (approximately 24h) is the most prominent component of the predictability regularity, depicted as the yellowish contoured area, although for Monday and Sunday the area is under the uncertainty region (hashed area) accounting for boundary effects. More surprising, however, is the fact that the second strongest component is not the 8h period, as we would expect—given the working schedules and the sleeping cycles—but rather a 12h period (i.e., the circasemidian period) corresponding to the second harmonic of the circadian rhythm. This period manifests in the plot as a reddish contoured area. The third strongest component is centered approximately around the 6h regime during the day. To describe these components, we also estimated the global wavelet spectrum, which gives us the true power spectrum of the time series. A closer inspection of the

global power spectrum (Fig. 14c) shows that these three main components can also be observed in the Fourier spectrum.

In the BrightKite data (Fig. 14d), the 12h period is even more pronounced than what we observed in the Weeplace data, whereas the 6h period, although present in the spectrum, is not statistically significant. Conversely, in the Gowalla data (Fig. 14e), the 12h frequency is not sufficiently pronounced whereas the fourth harmonic (6h) is indeed statistically significant. These results suggest that human traveling behavior could be influenced by internal biological processes way beyond the sleeping and feeding necessities, evidenced by the presence of the 12h rhythm.

However, an alternative hypothesis to the influence of the 12h rhythm is that these periods are

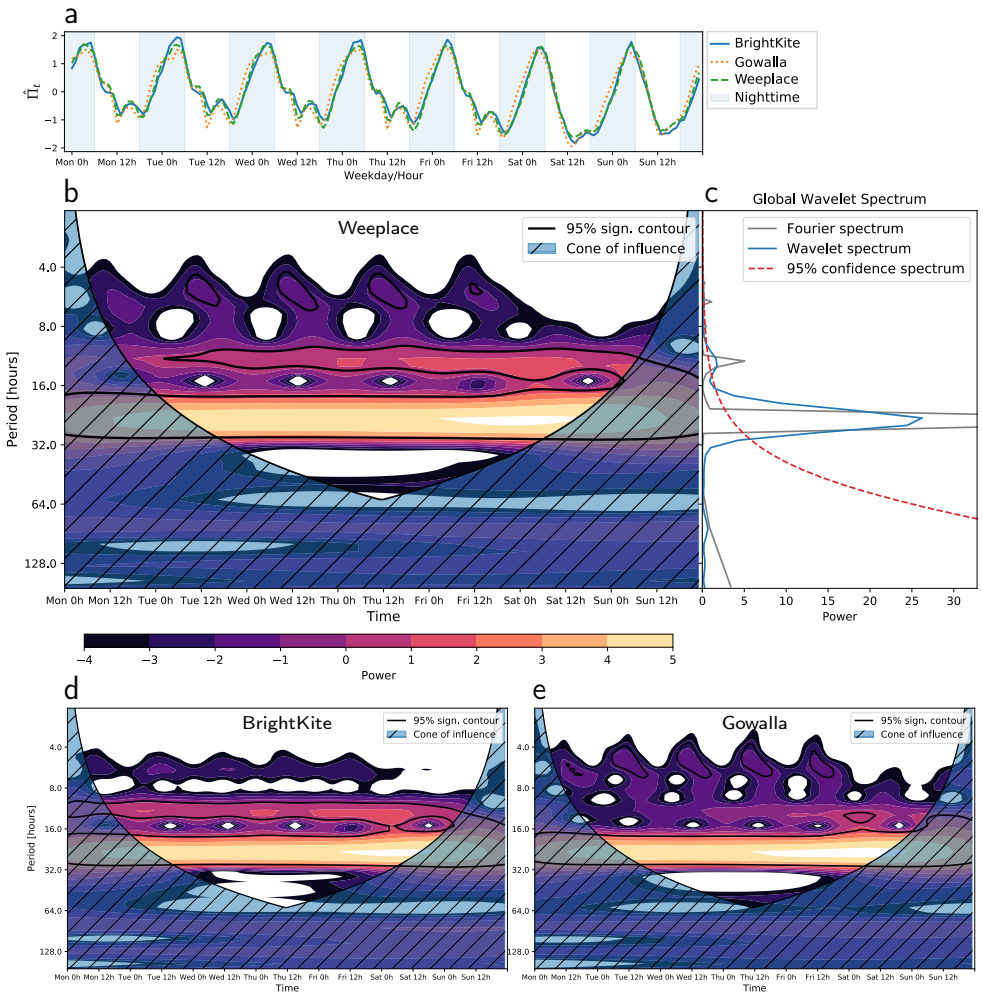


Figure 14: ***Wavelet analysis*** *– (**a**) The rescaled predictabilities $\hat{\Pi}_t$ reveal a remarkably strong agreement of all three dataset both in the time and frequency domains. (**b**) The wavelet spectrum reveals that the circadian period (approximately 24h) is the most prominent component of the predictability regularity while the second most-pronounced frequency is the 12h period. Additionally, the third strongest component is a period of approximately 6h. A closer inspection of the power spectra (**c**) shows that these three main components can also be observed in the Fourier spectrum. (**d**) In the BrightKite data, the 12h period is even more pronounced than what we observed in the Weeplace data, whereas the 6h period, although present in the spectrum, is not statistically significant. (**e**) Conversely, the 6h frequency is statistically significant in the Gowalla data.*

in fact rooted on population-level inhomogeneities on the activity routines. For instance, the 12h component could be explained by the same 24h rhythms if a large proportion of the users had their 24h schedules offset by 12h. To test for this hypothesis we performed the wavelet analysis on individual-level data. Since the 95% significance region is a function of the lag-1 autocorrelation α [19], we computed the α correlation for each individual user. Our analyses indicated that the α correlation distribution exhibits a bell-shaped curve with mean $\mu_\alpha \approx 0.58 \pm 0.1$. We then selected users in each dataset with $\alpha \approx 0.58$ and performed the wavelet analyses on their predictability data. Fig. 15a depicts the normalized predictabilities $\hat{\Pi}_t^u$ of one typical user from each dataset.

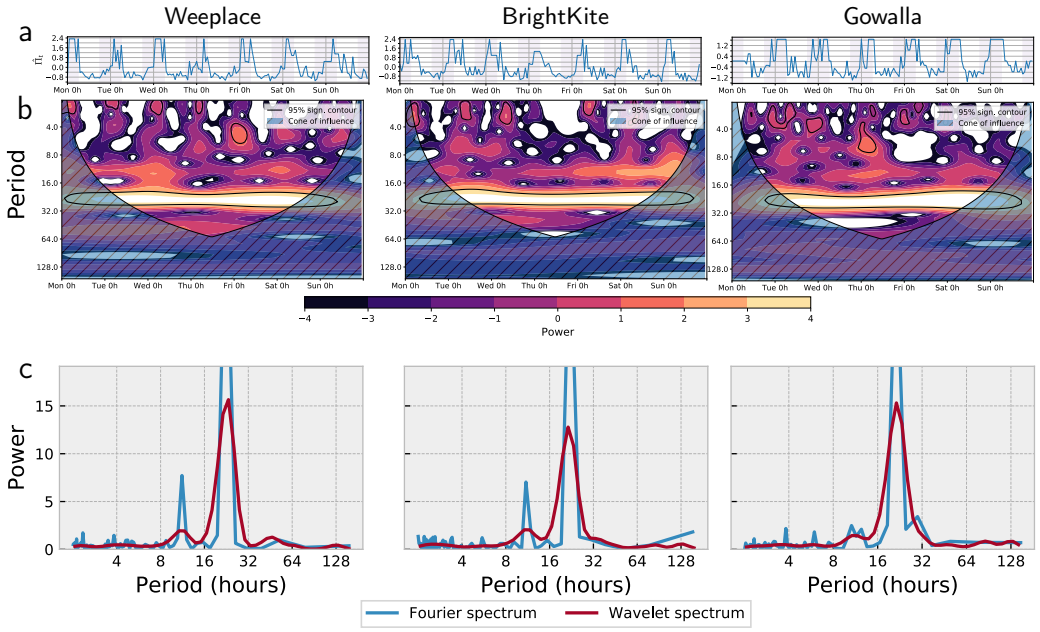


Figure 15: ***The wavelet power spectra of typical users from each dataset*** *– (**a**) The rescaled predictabilities $\hat{\Pi}_t$ of individual users are not as smooth as what we observed in the population data. (**b**) Nevertheless, the wavelet spectra suggests the presence of the 12h rhythms. (**c**) This evidence is further corroborated by analyses of the Fourier and wavelet power spectra.*

At an individual-level, the predictability curves are not as smooth as in the population data and neither are the frequency components. The wavelet analyses (Fig. 15b), however, reveal the 12h and 24h cycles, even though the 12h one is not sufficiently strong to satisfy the 95% significance threshold, likely due to data sparsity. Nevertheless, a closer loot at the global wavelet spectrum (Fig. 15c) makes it evident that the 12h cycle is manifested in the data of all three users, being observed in both the wavelet and Fourier spectra.

In summary, in this phase of the project, we explored the time-frequency components of the cyclic variations observed in the uncertainty and predictability-levels of human visitation patterns, which are signature of the synchronized nature of many of our activities (e.g., work schedules) [20], convoluted with socially-instituted divisions of the time (e.g., hours and calendars). However, our results indicate that in addition to the daily routines, mobility diversity is also marked by periods of approximately 12h and 6h. Moreover, the absence of a marked 8h cycle – the duration of typical working and sleeping hours – in detriment of a prominent 12h period indicates that the decision-making processes responsible for our visitation patterns are being influenced by our internal biological cycles beyond our sleeping and feeding necessities as well as other

socially-motivated routines. Finally, our findings might represent an important advancement to our understanding of the fundamental processes governing human mobility behavior which is intrinsically liked to the predictability of such behavior.

3. Work supported by the grant

This list below consists of all work supported by the grant.

3.1. Publications

1. *Human mobility: Models and applications*. Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, Marcello Tomasini. Physics Reports (in press), (2018).

2. *The scaling of crime concentration in cities* Marcos Oliveira, Carmelo Bastos-Filho, Ronaldo Menezes. PLoS One. **12(8)**, e0183110, (2017)

3. *The influence of the circadian and ultradian rhythms to human mobility: empirical evidences from location-based check-ins*. Hugo Barbosa, Marcos Oliveira, Diogo Pacheco, Ronaldo Menezes, Gourab Ghoshal. Proceedings of New England Regional Conference on Complex Systems (NERCCS) (in press) (2018) .

4. *Morphology of travel routes and the organization of cities*. Minjin Lee, Hugo Barbosa, Hyejin Youn, Petter Holme, Gourab Ghoshal. Nature communications **8(1)**, 2229, (2017).

5. *From the betweenness centrality in streets to structural invariants in street networks*. A Kirkley, H Barbosa, M Barthelemy, G Ghoshal. in review with Nature Communications

6. *Structural motifs in human mobility*. Hugo Barbosa, Marcos Oliveira, Diogo Pacheco, Ronaldo Menezes and Gourab Ghoshal. in review with Journal of Royal Society Interface.

3.2. Presentations

1. *A social influence model based on affiliation: A case study from political parties*. Presenter: Josemar Faustino da Cruz. 9th conference on complex networks (CompleNet), Northeastern University Boston, MA, (04/18)

2. *Structural Invariants in Street Networks*. Presenter: Gourab Ghoshal. 9th conference on complex networks (CompleNet), Northeastern University Boston, MA, (04/18).

3. *The scaling of crime concentration in cities*. Presenter: Ronaldo Menezes (Keynote Address) Latin American Conference on Complex Systems (LANET), Puebla, Mexico, (09/17).

4. *Urban socioeconomic patterns revealed through morphology of travel routes* Presenter: Gourab Ghoshal (Plenary speaker) 8th conference on complex networks (CompleNet), Inter University Center Dubrovnik, Croatia, (03/17)

References

[1] L. Backstrom, E. Sun, and C. Marlow, "Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity," in *Proceedings of the 19th international conference on World wide web - WWW '10*, (New York, New York, USA), p. 61, ACM Press, 2010.

[2] W. Pan, G. Ghoshal, C. Krumme, M. Cebrian, and A. S. Pentland, "Urban characteristics attributable to density-driven tie formation," *Nature Communications*, vol. 4, p. 1961, 2013.

[3] P. Deville, C. Song, N. Eagle, V. D. Blondel, A.-L. Barabási, and D. Wang, "Scaling identity connects human mobility and social interactions," *Proceedings of the National Academy of Sciences*, vol. 113, pp. 7047–7052, jun 2016.

[4] M. Schlapfer, L. M. a. Bettencourt, S. Grauwin, M. Raschke, R. Claxton, Z. Smoreda, G. B. West, and C. Ratti, "The scaling of human interactions with city size," *Journal of The Royal Society Interface*, vol. 11, pp. 20130789–20130789, jul 2014.

[5] A. Wingfield and D. L. Byrnes, *The Psychology of Human Memory*. Elsevier Science, 2013.

[6] Y. Li and N. Epley, "When the best appears to be saved for last: Serial position effects on choice," *Journal of Behavioral Decision Making*, vol. 22, pp. 378–389, oct 2009.

[7] W. B. de Bruin and G. Keren, "Order effects in sequentially judged options due to the direction of comparison," *Organizational Behavior and Human Decision Processes*, vol. 92, no. 1-2, pp. 91–101, 2003.

[8] W. Bruine de Bruin, "Save the last dance for me: Unwanted serial position effects injury evaluations," *Acta Psychologica*, vol. 118, no. 3, pp. 245–260, 2005.

[9] W. Bruine de Bruin, "Save the last dance II: Unwanted serial position effects in figure skating judgments," *Acta Psychologica*, vol. 123, no. 3, pp. 299–311, 2006.

[10] H. Barbosa, F. B. de Lima Neto, A. Evsukoff, and R. Menezes, "The effect of recency to human mobility," *European Physical Journal Data Science*, vol. 4, p. 21, 2015.

[11] C. Krumme, A. Llorente, M. Cebrian, A. S. Pentland, and E. Moro, "The predictability of consumer visitation patterns.," *Scientific reports*, vol. 3, p. 1645, jan 2013.

[12] L. Song, D. Kotz, R. Jain, and X. He, "Evaluating next-cell predictors with extensive Wi-Fi mobility data," *IEEE Transactions on Mobile Computing*, vol. 5, no. 12, pp. 1633–1648, 2006.

[13] A. Lempel and J. Ziv, "On the Complexity of Finite Sequences," *IEEE Transactions on Information Theory*, vol. 22, pp. 75–81, jan 1976.

[14] I. Kontoyiannis, P. Algoet, Y. Suhov, and A. Wyner, "Nonparametric entropy estimation for stationary processes and random fields, with applications to English text," *IEEE Transactions on Information Theory*, vol. 44, pp. 1319–1327, may 1998.

[15] D. Nalley, J. Adamowski, B. Khalil, and A. Biswas, "Inter-annual to inter-decadal streamflow variability in Quebec and Ontario in relation to dominant large-scale climate indices," *Journal of Hydrology*, vol. 536, pp. 426–446, 2016.

[16] J.-P. Antoine, P. Carrette, R. Murenzi, and B. Piette, "Image analysis with two-dimensional continuous wavelet transform," *Signal processing*, vol. 31, no. 3, pp. 241–272, 1993.

[17] P. S. Addison, *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance*. CRC press, 2017.

[18] X. Mi, H. Ren, Z. Ouyang, W. Wei, and K. Ma, "The use of the Mexican Hat and the Morlet wavelets for detection of ecological patterns," *Plant Ecology*, vol. 179, pp. 1–19, jul 2005.

[19] C. Torrence and G. P. Compo, "A Practical Guide to Wavelet Analysis," *Bulletin of the American Meteorological Society*, vol. 79, pp. 61–78, jan 1998.

[20] A. J. Morales, V. Vavilala, R. M. Benito, and Y. Bar-Yam, "Global patterns of synchronization in human communications," *Journal of The Royal Society Interface*, vol. 14, p. 20161048, mar 2017.