Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM 2021)

Web Routineness and Limits of Predictability: Investigating Demographic and Behavioral Differences Using Web Tracking Data

Juhi Kulshrestha,[1,2] **Marcos Oliveira,**[1] **Orkut Karaçalik,**[1,3] **Denis Bonnay,**[4,5] **Claudia Wagner**[1,3]

[1]GESIS - Leibniz Institute for the Social Sciences, Germany, [2]University of Konstanz, Germany,
[3]University of Koblenz-Landau, Germany, [4]Université Paris Nanterre, France, [5]Respondi SAS, France
juhi.kulshrestha@uni-konstanz.de, {marcos.oliveira, orkut.karacalik, claudia.wagner}@gesis.org, dbonnay@u-paris10.fr

Abstract

Understanding human activities and movements on the Web is not only important for computational social scientists but can also offer valuable guidance for the design of online systems for recommendations, caching, advertising, and personalization. In this work, we demonstrate that people tend to follow routines on the Web, and these repetitive patterns of web visits increase their browsing behavior's achievable predictability. We present an information-theoretic framework for measuring the uncertainty and theoretical limits of predictability of human mobility on the Web. We systematically assess the impact of different design decisions on the measurement. We apply the framework to a web tracking dataset of German internet users. Our empirical results highlight that individuals' routines on the Web make their browsing behavior predictable to 85% on average, though the value varies across individuals. We observe that these differences in the users' predictabilities can be explained to some extent by their demographic and behavioral attributes.

Introduction

The World Wide Web is an immense space containing more than six billion websites that keep popping up and dying every second (de Kunder 2020). We spend more than a quarter of our day in this fluid space, performing activities such as shopping, reading the news, or interacting with friends (Kemp 2019). Though the Web is an indispensable part of people's lives today, we still do not fully understand the dynamics of people's web browsing behavior. Such understanding is vital to improve the way we navigate the Web and provides grounds for algorithms to assist users in their exploration of the cyberspace.

In the case of physical space, researchers have already started to understand the fundamental mechanisms governing the dynamics of human mobility (Gonzalez, Hidalgo, and Barabási 2008; Song et al. 2010a). Despite the complexity of our decision-making processes, the way we move in physical space has been shown to present regularities at varying spatiotemporal scales (Pappalardo et al. 2015; Alessandretti et al. 2018; Barbosa et al. 2018). Such regularities emerge from our daily routines and constraints, which in turn make our movements highly predictable (Song et al. 2010b). Understanding these characteristics of human mobility is critical for managing crucial aspects of our society, including public health, transportation, urban planning, to name a few.

With the increasing reliance on the Web, people have switched to performing many everyday activities online rather than offline. With people living more and more of their lives online, the way they traverse the Web affects how they gather information and interact with others, influencing different aspects of society such as education (Ravizza, Uitvlugt, and Fenn 2017), mental health (Culjak, Kowalenko, and Tennant 2016), news consumption (Möller et al. 2019; Scharkow et al. 2020), and political participation (Stier et al. 2020). Uncovering the fundamental properties of web browsing behavior has become imperative to understand society better.

Recently, examining and modeling the dynamics of human movements on the Web have started to receive more attention, with researchers leveraging models developed for mobility in physical space to study them (Zhao et al. 2014; Zhao, Cai, and Lu 2015; Zhao et al. 2016; Barbosa et al. 2016; Hu, Luo, and Liu 2018; Hu, Xia, and Luo 2019; Hazarie et al. 2019). Previous work has uncovered underlying mechanisms in web mobility, such as preferential return and exploration. However, much research has predominantly focused on characterizing distributions and time series. Researchers have often overlooked heterogeneity in users' characteristics and often limited their analysis to aggregated levels and specific platforms. Therefore, the individual-level regularities in web browsing behavior are as yet poorly understood.

In this work, we show that individuals tend to follow routines on the Web—a property we call *web routineness*. These repetitive patterns of online visitation considerably increase the achievable predictability concerning users' browsing behavior. We present an information-theoretic framework for estimating the theoretical limits of predictability of people's web mobility based on previous work (Song et al. 2010b). By examining web tracking data from $2,148$ users, we show that individuals' web routineness makes their browsing behavior predictable to 85% on

average. However, we show that the routineness and predictability vary considerably across users and that users' demographic and behavioral differences can explain part of this variation.

Our work not only has implications for computational social scientists but also computer scientists and industry practitioners. While much prior work (Pirolli and Pitkow 1999; Narvekar and Banu 2015; Deshpande and Karypis 2004; Mabroukeh and Ezeife 2009; Manavoglu, Pavlov, and Giles 2003; Pitkow and Pirolli 1999; Su et al. 2000; Awad and Khan 2007; Khalil, Li, and Wang 2009) has focused on developing specific algorithms for predicting the next visited location on the Web, our work focuses on estimating the theoretical upper limits for correctly predicting an individual's next visited web location based on their browsing trajectories. Our work can inform researchers and practitioners about what performance they can aspire to achieve for which web prediction task by leveraging the routines users follow in their browsing trajectories. Moreover, our results also have implications for privacy researchers since the high predictability for most users indicates that attackers can also predict what location the user would visit next, having observed the routines users follow from their past browsing histories.

Finally, we summarize the contributions of this paper:

- Building upon prior work on mobility in physical space (Song et al. 2010b; Smith et al. 2014; Ikanovic and Mollgaard 2017), we present an adapted information-theoretic framework for estimating the theoretical limits of predictability of humans' mobility on the Web.
- We describe how to construct stationary and non-stationary trajectories from users' raw web tracking data.
- We systematically assess the impact of different design decisions (e.g., temporal and spatial resolution, and stationarity in trajectories) on the measurement and provide practical guidelines for applying the framework to users' web tracking data.
- We investigate users' routineness and predictability on the Web by applying the framework to a web tracking dataset of $2,148$ German internet users. We show that individuals exhibit routineness while browsing the Web and that these routines or repetitive patterns make their browsing behavior predictable to 85% on average. Our results revealed considerable variability in the routineness and predictability across users. We show that users' demographic and behavioral differences can explain part of this variation.
- Finally, we make available our code[1] along with tutorial notebooks and data (Kulshrestha et al. 2021) to the researcher community to encourage further research on mobility on the Web.

Background & Related Work

Human mobility in physical space Research on human mobility has shown that it is characterized by a heterogeneity of travel patterns (Gonzalez, Hidalgo, and Barabási 2008; Pappalardo et al. 2015), a high degree of predictability (Song et al. 2010b), and a strong tendency of humans' to spend most of their time in a few locations (Song et al. 2010a; Alessandretti et al. 2018). We leverage previous work on human mobility (Song et al. 2010b; Ikanovic and Mollgaard 2017; Smith et al. 2014) to present an adapted predictability measurement framework for estimating the upper limits of predictability of people's mobility on the Web.

Adapting physical space mobility models for cyberspace Prior work on investigating mobility in cyberspace, using concepts from physical mobility literature, found several interesting similarities, such as similar scaling properties of return visitation patterns (Barbosa et al. 2016; Hu, Luo, and Liu 2018; Hu, Xia, and Luo 2019), similar superlinear scaling relation between mean frequency of visit and its fluctuation (Zhao et al. 2014), similar focus on a small, stable number of mobile apps (De Nadai et al. 2019) and number of familiar locations in physical space (Alessandretti et al. 2018), and similar memory-based random walk dynamics (Zhao et al. 2016). Several prior studies have applied the framework proposed by Song et al. (Song et al. 2010b) in the online context to study entropy and predictability for mobility in multiplayer online games (Sinatra and Szell 2014), mobile phone users' traffic usage (Tao et al. 2019), and textual conversations (Bagrow, Liu, and Mitchell 2019), information cascades (Kolli, Balakrishnan, and Ramakrishnan 2017), and movement of users across communities (Hu, Luo, and Liu 2018) on specific social media platforms. Instead, we focus on the routines that people follow on the web to estimate the theoretical upper limit of predictability of user's next visited location while traversing the whole Web. We find these cyberspace limits to be lower than the limits computed for physical mobility (Song et al. 2010b; Ikanovic and Mollgaard 2017; Zhao, Cai, and Lu 2015). Furthermore, we observe differences in the predictability of users with different demographic and behavioral characteristics, which were not observed for mobility in the physical space.

Web navigation prediction algorithms Another complementary line of prior work has focused on developing specific algorithms for predicting the next web access by users, given users' web access sequences. For this task, prior work has leveraged the routines in people's browsing trajectories by developing variations of Markov models (Pirolli and Pitkow 1999; Narvekar and Banu 2015; Deshpande and Karypis 2004; Mabroukeh and Ezeife 2009; Manavoglu, Pavlov, and Giles 2003), n-gram sequence models (Pitkow and Pirolli 1999; Su et al. 2000), or hybrid models combining Markov models with ANNs (Awad and Khan 2007) or clustering and association rule mining (Khalil, Li, and Wang 2009). Our goal is not to develop a prediction algorithm, but to estimate the theoretical upper limit of predictability given users' sequence of web accesses. On comparing with prior work, we observe that the previously reported performance matches well with our estimated theoretical limits.

Predictability Measurement Framework

In this section, we describe the framework for estimating the predictability of people's mobility in cyberspace from their

[1] https://tinyurl.com/web-tracking-library

web browsing traces. First, we outline how to construct trajectories from raw browsing data for each user. Then, we detail how we measure uncertainty in users' trajectories using concepts of entropy from the field of Information Theory. These concepts help us to uncover the extent to which individuals follow routines on the Web. Finally, we describe the procedure to estimate the theoretical predictability limits of a user's browsing behavior.

Sharing code In this paper, we attempt to make the framework accessible to diverse research communities, and as such, our code is available as a Python library with corresponding online tutorial notebooks (Karaçalik, Oliveira, and Kulshrestha 2021).

Constructing User Trajectories

To study the mobility of users on the Web, we first need to construct their cyberspace trajectories from the raw web tracking data. Figure 1 illustrates a toy example of raw data: a user visits nine different website URLs, spending different amounts of time on each of them. In this example, each URL belongs to a domain (i.e., A, B, C), and they can be associated with a category, such as 'news', 'social media', or 'search engine'. In our paper, we define a *trajectory* as a discrete sequence of locations visited by a user, where a *location* can be a website URL, a domain, or a category.

We propose two types of web trajectories: stationary and non-stationary. The former accounts for the amount of time spent on locations, while the latter ignores it. These trajectories enable us to study the role of time in measuring users' predictability. In what follows, we describe each trajectory type using the toy data (Figure 1) as the raw browsing trace data. Note that though most of the following examples are trajectories of domains, the definitions are generalizable to other types of locations (i.e., categories or website URLs).

Series of time bins The **stationary trajectories** T^{stat} incorporate the amount of time a user spends on each location. When we analyze users' predictability using this trajectory, we are interested in the task of *predicting the location visited in the next time step*. Practical applications of such a task include real-time recommendations, dynamic caching, and dynamic advertising. With this type of trajectory, we can investigate the impact of visit duration on the emergence of users' routines on the Web.

To generate this trajectory, we use a non-overlapping time window of size Δt to create bins across time. Within each time bin, we choose the most visited location (i.e., the location visited for the longest duration). We break ties randomly and ignore the time bins without any browsing activity. For example, using our toy data, the stationary trajectory of domains would be as follows:

$$T^{\text{stat}} = \{A, A, A, A, A, A, A, A, B, B, B, B, B \\ B, B, B, A, C\}.$$

We can observe that stationarity is captured in the trajectory. This method of generating trajectories is consistent with prior studies of human mobility (Song et al. 2010b; Ikanovic and Mollgaard 2017).

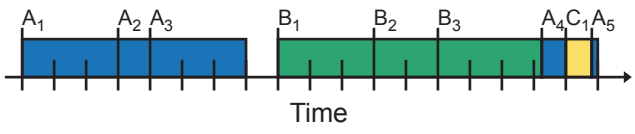


Figure 1: An illustration of raw web tracking data: A user visits nine different website URLs (A_1, A_2, A_3, A_4, A_5, B_1, B_2, B_3, and C_1,) along the timeline. Each URL is from a domain (i.e., A, B, and C) and belongs to a category (e.g., search engine, social media site). Each visit has a time duration, and visits are non-overlapping.

Series of visited locations The non-stationary trajectories ignore the amount of time a user spends on each location, ignoring users' tendency to stay longer at specific locations. When we analyze the predictability of non-stationary trajectories, we are interested in the task of *predicting the next visited location, irrespective of how much later*. Examples of applications include learning visitation patterns to automate users' web activities (e.g., booking a holiday, or researching and buying products or services). These trajectories enable us to investigate users' routines that emerge regardless of the duration of visits.

We build non-stationary trajectories using two different temporal schemes: binned and sequential. The former considers locations for each time bin, whereas the latter considers locations as they appear in the raw browsing trace data. These trajectories are defined as follows:

- **Non-stationary binned trajectory** $T^{\text{binNonStat}}$: First, we create a stationary time series from the raw data using a time window of width Δt. Then, we compress this trajectory by removing repetitive adjacent locations from it. For instance, using our toy example, the non-stationary binned trajectory of domains looks as follows:

  $$T^{\text{binNonStat}} = \{A, B, A, C\}.$$

  By comparing this sequence with T^{stat}, we see that we have removed the stationarity from the trajectory. This method of trajectory construction is consistent with prior work (Ikanovic and Mollgaard 2017).

- **Non-stationary sequential trajectory** $T^{\text{seqNonStat}}$: Instead of compressing the stationary trajectory, we compress the sequence of distinct locations from the raw data. For example, applying this method to our toy data would generate the following trajectory of domains:

  $$T^{\text{seqNonStat}} = \{A, B, A, C, A\}.$$

  This approach creates trajectories that include all distinct locations that were visited independently of the amount of time spent on them. For instance, in our example, domain A appears at the end of trajectory $T^{\text{seqNonStat}}$, but not in $T^{\text{binNonStat}}$.

Note that the two types of non-stationary trajectories converge to each other as Δt gets smaller and smaller, becoming more and more independent of Δt.

Estimating Predictability of Users' Mobility on the Web

Having constructed a user's trajectory, we can now study the dynamics of the user's mobility on the Web. Without loss of generality, we can represent any trajectory as a discrete series

$$T = \{x_1, x_2, \ldots, x_\ell\},$$

where $x_t \in \mathcal{V}$ is a location (i.e., a website URL, a domain, or a category), $\mathcal{V}$ is the set of visited locations, and ℓ is the trajectory length. Here we are interested in the dynamics of this time series; we want to study (1) the location preferences of a user and (2) the emerging visitation patterns in these preferences. Specifically, we aim to investigate the uncertainty embedded in this time series.

In this paper, we use three concepts of uncertainty from the Information Theory field:

(i) Time-uncorrelated entropy (S^{unc}) that accounts for the frequency with which users visit locations.

(ii) Maximum entropy (S^{rand}) that considers only the number of unique locations visited.

(iii) Time-correlated entropy (S) which considers the visitation patterns of the users.

These three measures allow us to quantify the uncertainty in users' browsing behavior, enabling us to estimate the extent to which users exhibit routineness and compute the theoretical limits of predictability of users' trajectories.

Users' location preferences To study a user's location preferences, we can examine the number of unique locations visited by the user and the probability $p(i)$ of visiting a location i (i.e., the normalized frequency of visits).

The size of $\mathcal{V}$, denoted by $N = |\mathcal{V}|$, relates to the preference breadth of a user. It tells us how broadly this user explores the Web throughout a particular time frame. This quantity, however, neglects the frequency with which the users visit locations, missing the likely existence of favorite websites, categories, or domains. To account for frequency, we need to examine the spread of the probability distribution $p(i)$. For instance, if a user visits all locations at the same frequency, then $p(i) \approx 1/N$, whereas if a user mostly visits a specific location j, then $p(i)$ peaks at $i = j$. We want to quantify the peakiness of the distribution $p(i)$.

From an information-theoretic perspective, we want to measure the uncertainty of a random variable (in our case, the random variable is the time series T). For analyzing this scenario without accounting for temporal correlations, we can use Shannon entropy, denoted here as the **time-uncorrelated entropy** S^{unc}, defined as

$$S^{\text{unc}}(p) = -\sum_{i \in \mathcal{V}} p(i) \log_2 p(i) \tag{1}$$

and expressed in bits (Cover and Thomas 2006). $S^{\text{unc}}(p)$ measures the average amount of memory needed to store the outcome of the random variable associated with $p(\cdot)$. That is, the more certain we are about the outcome of a random variable, the less memory we need to store the variable, thus the lower $S^{\text{unc}}(p)$.

In our case, Eq. (1) quantifies the uncertainty concerning the locations a specific user visits over time. For instance, when a user keeps visiting only one location, the entropy $S^{\text{unc}}(p)$ is zero because of the low uncertainty of this user's behavior. In contrast, a user without a favorite location exhibits the highest uncertainty and the maximum entropy. The **maximum entropy** S^{rand} occurs when $p(i) = 1/N$ for $\forall i$ and can be defined as the following:

$$S^{\text{rand}}(p) = \log_2 N. \tag{2}$$

This quantity represents the uncertainty in a user's location when the user's visitation behavior follows a uniform distribution (i.e., without any preference towards specific locations). $S^{\text{rand}}(p)$ can be seen as the worst-case scenario of a variable with a sample space of size N, such that this uncertainty equals to guessing a user's location *randomly*.

Users' visitation patterns Though $S^{\text{unc}}(p)$ provides us with a simple uncertainty measure, it neglects long-range temporal correlations in the trajectories—a potential feature in web mobility. A temporal correlation might emerge because of repetitive patterns of visitation on the Web (e.g., always accessing a social media website after reading the news). Such patterns in a user's trajectory decrease the uncertainty about this user, regardless of the Shannon entropy. Indeed, a user might lack favorite locations but have favorite patterns of visitation (see Figure 2). To capture this uncertainty in web mobility, we need to examine trajectories as the result of processes that generate sequences.

First, we denote a block of L consecutive random variables as $X^L = X_1 \ldots X_L$, and $x^L = x_1 \ldots x_L$ is an instance of this variable. In our case, x^L can be any sequence

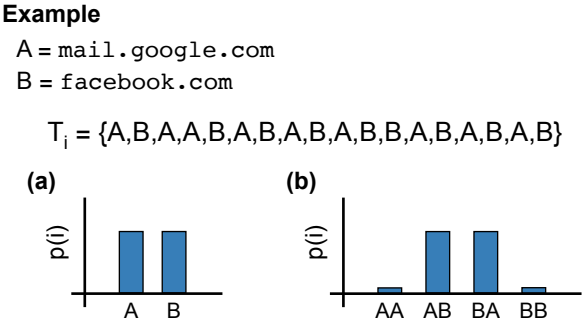


Figure 2: We need to account for long-range temporal correlations to study the predictability of users' browsing behavior. For example, consider a hypothetical user trajectory of accessing an e-mail account then a social media website (top). When a user follows such a sequence over time, the absence of favorite locations makes the Shannon entropy to approach its maximum value, $S^{\text{unc}} \approx S^{\text{rand}}$ **(a)**. This trajectory, however, is quite predictable; we can see a clear pattern repeating itself. Indeed, when we examine the probability of the blocks of length $L = 2$, we find some favorite patterns in the trajectory **(b)**. The rate at which the uncertainty increases with L is called the source entropy rate. This quantity measures the intrinsic uncertainty of a user trajectory as we learn about the sequences that this user follows over time.

of length L in a trajectory of a user. If we denote $p^L(\cdot)$ as the probability function of X^L, we can define the entropy of p^L as the following:

$$S'(p^L) = -\sum_{s^L \in \mathcal{V}^L} p^L(s^L) \log_2 p^L(s^L), \quad (3)$$

where $\mathcal{V}^L$ is the sample space containing all the visited sequences of length L, and we establish $S'(p^0) = 0$ (Crutchfield and Feldman 2003). This quantity is the so-called block entropy or total Shannon entropy of length-L, and it is a convenient measure of uncertainty of the sequences in a trajectory (Crutchfield 2011).

Instead of focusing on a specific length L, we can examine the relationship between $S'(p^L)$ and L to investigate the uncertainty in the trajectories. Specifically, we would like to analyze how $S'(p^L)$ changes as we have longer and longer L-blocks. In this way, we learn about the uncertainty of a trajectory while discounting for the recurrent patterns in it. For this purpose, we use the source entropy rate of a stochastic process, defined as

$$h_\mu = \lim_{L \to \infty} S'(p^L)/L, \quad (4)$$

when the limit exists (Crutchfield and Feldman 2003; Cover and Thomas 2006). The entropy rate tells us the irreducible uncertainty of a process that persists even when we account for the temporal correlations. In our case, h_μ quantifies the intrinsic uncertainty of a trajectory even when we learn about the sequences (i.e., patterns of visitation) that this person follows throughout time.

In our paper, we follow previous work (Song et al. 2010b; Smith et al. 2014; Ikanovic and Mollgaard 2017) and estimate the entropy rate of a trajectory using the Lempel–Ziv compression algorithm; we call this estimate the **time-correlated entropy** and denote it as S. The Lempel–Ziv algorithm attempts to find the optimal dictionary to compress a given sequence, and it has been shown to quickly converge to the entropy rate as the sequence length approaches infinity (Kontoyiannis et al. 1998).

By comparing time-correlated entropy with time-uncorrelated entropy, we can investigate the role of routines in users' web browsing behavior. The higher the difference between the entropies, the more individuals repeat visitation patterns. These recurrent visitation patterns occur because of preferred sequences in visiting websites, categories, or domains. We call this preference **web routineness**. This property of web browsing behavior emerges from individuals' choices under the structural constraints of the Web and can affect the extent to which we can predict their trajectories.

Estimating predictability These entropy quantities enable us to assess the uncertainty embedded in the browsing behavior of individuals; however, they fail to explicitly tell us how well we could predict a user's trajectory despite its intrinsic uncertainty. We want to learn about the probability Π of correctly predicting future locations, given a past series of observations. It is possible to show that Π is subject to Fano's inequality[2], and has an upper bound, denoted as Π^{max}.

In our case, this upper bound reveals the theoretical upper limit to predict an individual's future location correctly if we restrict ourselves to only the browsing trajectories from the web tracking data. For instance, a user having an upper limit of $\Pi^{\text{max}} = 0.6$ exhibits an intrinsic uncertainty that makes their behavior indistinguishable from random 40% of the time. Because of this randomness, the best accuracy level that a predictive algorithm can achieve for this user is 60%. That is, Π^{max} tells us how much predictive power we can theoretically have by possessing just the browsing trajectories from the web tracking data of an individual.

It is impractical to compute Π^{max} directly, but the quantity has an explicit relationship with the entropy rate:

$$S = H_b(\Pi^{\text{max}}) + (1 - \Pi^{\text{max}}) \log_2(N - 1), \quad (5)$$

where $H_b(\cdot)$ is the binary entropy function, defined as $H_b(p) = -p \log_2 p + (1-p) \log_2 (1-p)$, and N is the alphabet size (i.e., the number of unique locations visited) (Song et al. 2010b). To find the **predictability upper-bound** Π^{max} of a user, we just have to solve Eq. (5) using a numerical solver, provided that we know S and N for this user. This approach also allows us to find the hypothetical predictabilities of users if their trajectories were random or absent of temporal correlations by substituting S in Eq. (5) with the corresponding entropy value. More precisely, we replace S with S^{rand} and S^{unc} to find Π^{rand} and Π^{unc}, respectively.

Comparing predictabilities Having the three predictability values (Π^{rand}, Π^{unc}, and Π^{max}) for the users provides us with a framework for comparing users and investigating the impact of emerging preferences and routines in their trajectories. First, predictability values lie within $[0, 1]$, making it convenient to compare users with a varying number of visited locations—different from absolute entropy numbers. Second, the distinct types of predictability help us interpret the results by yielding two null models. Indeed, finding the hypothetical values of Π^{rand} and Π^{unc} helps us to contextualize Π^{max}. Each of these predictability values represents a version of a user's trajectory when we remove a specific feature of it. Π^{unc} represents the predictability of a trajectory if we randomize the original trajectory, removing all the temporal correlations. Π^{rand} represents the predictability of a trajectory if we remove all repeated visits to locations from the original trajectory, removing all user preferences. With these values as a baseline, we can now compare them with Π^{max} to understand the impact of web routineness on the predictability of a user's web browsing behavior.

Dataset of Web Browsing Traces

Our anonymized dataset consists of one month's (October 2018) web tracking data of $2,148$ German users. For each user, the data contains the URL[3] of the webpage the user visited, the domain of the webpage, category of the domain[4],

[2] We omit details of the analytical derivation and refer the interested reader to the literature (Song et al. 2010b; Smith et al. 2014).

[3] URLs are anonymized, and personally identifiable information such as user names or passwords have been removed.

[4] Categories were inferred using https://docs.webshrinker.com/v3/website-category-api.html, which provides 41 distinct cate-

Gender	Population	Sample	Age	Population	Sample	Age	Population	Sample
Female	50.6	51.2 (48.5)	18-24	9.7	12.7 (11.3)	45-54	19.3	22.8 (24.9)
Male	49.4	48.8 (51.5)	25-34	16.3	20.1 (18.6)	55-64	18.6	21.6 (24.2)
			35-44	15.6	18.4 (16.1)	65-80	20.5	4.4 (4.9)

Table 1: Comparison of German population margins of gender (male, female) and age (18-80 years) with our sample's composition (percentage values). Filtered sample in parentheses. Unfiltered sample refers to all $2,148$ users, while the filtered sample depicts the $1,455$ users who have long enough trajectories to estimate their predictability reliably (see Section "Ensuring Data is Sufficient to Estimate the Predictability").

time of visit, and active seconds spent by the user on the page. In total, these 2,148 users made 9,151,243 URL visits, spanning 49,918 unique domains.

Evaluating web tracking data We used web traffic data from Alexa[5] to evaluate whether the most visited domains in Germany also feature prominently in our dataset. We compared the 5000 most visited domains from Alexa with the aggregated number of visits to these domains in our web tracking data. We found them to be highly correlated (Pearson's correlation coefficient $r = 0.78$).

Comparing the demographics of our panelists and the German population For each user in our dataset, we have self-reported information (collected via a survey) about their gender and age. We compare our sample's distributions with German population margins for gender and age (Table **??**).[6] We observe that our sample's gender distribution matches closely with the German population's. However, the older population (65 years and above) is under-sampled in our panel. This difference could potentially be due to lower internet usage by older people, or due to the opt-in, non-probability recruitment of the panelists.

Ethics The web tracking data for the users in our dataset was collected by a GDPR compliant digital panel company in Europe. All the participants of the company's online panel have agreed to anonymously share their survey (e.g., the gender and age information we use) and behavioral data for business or scientific purposes. The participants willing to be tracked install an executable file on their desktop. In return, they receive a monthly reward for their data.

Sharing data We share our anonymized dataset to spur further research in this area (Kulshrestha et al. 2021).

gories. We manually added two more categories 'email' and 'productivity'. Moreover, for popular domains with many sub-domains with distinctly different functions (such as *google.com*), we manually re-coded the sub-domains to correct categories. See https://github.com/gesiscss/webtracking/ for more details.

[5]While Alexa is also not representative of German online population, we use it since it is the most established source for web traffic data: https://docs.aws.amazon.com/AlexaTopSites/latest/ApiReference_TopSitesAction.html.

[6]In the absence of high-quality data about German internet users' demographics, we compare with offline German population margins for the year 2018: https://www.destatis.de/EN/Themes/Society-Environment/Population/Current-Population/_node.html.

Guidelines for Applying the Predictability Measurement Framework

We now provide some practical guidelines for applying the framework. As the first step, we present a principled method for ensuring that the data is sufficient for measuring predictability. Then, we discuss the impact of some design decisions on the measured predictability and offer guidelines that could help researchers to reflect on these decisions systematically.

Ensuring Data is Sufficient to Estimate the Predictability

To ensure that we have enough data to estimate users' predictability, we examine the convergence of $\Pi^{\max}$ as we add more and more data. The rationale here is that the quality of S (i.e., the estimate of the entropy rate, needed to find $\Pi^{\max}$) depends on the trajectory length. As this length approaches infinity, S converges to the entropy rate. We study this convergence by analyzing the predictability of trimmed versions of users' trajectories. For this, we first denote $\Pi^{\max}(\ell')$ as the predictability of a user estimated using only the first ℓ' locations of this user's trajectory; then, we study the absolute consecutive differences of $\Pi^{\max}(\ell')$, defined as:

$$\Delta\Pi^{\max}(\ell') = |\Pi^{\max}(\ell') - \Pi^{\max}(\ell' - 1)|,$$

where $1 < \ell' \leq \ell$.

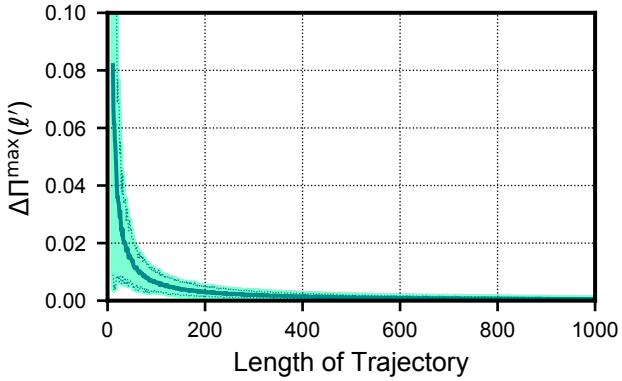


Figure 3: Mean absolute change in maximum predictability with increasing trajectory length for random 100 users. The shaded region depicts the 5^{th} to 95^{th} quantile range.

To examine the predictability convergence, we selected 100 random users and computed $\Delta\Pi^{\max}$ of their non-stationary $T^{\text{binNonStat}}$ domain trajectories (since these are the shortest trajectories). We found that the predictability values stabilize for a trajectory length of $\ell' = 100$, such that the mean difference is small ($\Delta\Pi^{\max}(100) = 0.006$), as shown in Figure 3. Therefore, we decided that the length of users' non-stationary trajectories must be at least 100 locations long. In the rest of the paper, we perform our analysis with the $1,455$ users who satisfy this criterion, excluding 693 users from our study. These $1,455$ users made a total of $8,910,779$ URL visits spanning $48,721$ unique domains, and their gender and age distributions are depicted in parentheses in Table **??**.

Impact of Design Decisions on Predictability

Next, we discuss the impact of three critical design decisions on the estimated predictability: (i) stationarity of trajectories, (ii) temporal resolution, and (iii) spatial resolution.

Impact of stationarity in trajectories When applying the predictability framework, we first need to decide which trajectory to construct. This decision is primarily driven by which prediction task we want to study. While T^{stat} is suitable for predicting the next visited location in the next time bin, $T^{\text{binNonStat}}$ or $T^{\text{seqNonStat}}$ are suitable for predicting the next visited location independent of when it is visited. To investigate how predictability changes depending on these tasks, we computed the predictability of the users using these three types of trajectory. We found that these tasks satisfy the following inequality:

$$\Pi^{\max}_{\text{binNonStat}} < \Pi^{\max}_{\text{seqNonStat}} < \Pi^{\max}_{\text{stat}},$$

where the predictability values correspond to the trajectories $T^{\text{binNonStat}}$, $T^{\text{seqNonStat}}$, and T^{stat}, respectively. We observe that as the stationarity in the trajectories increases (i.e., the user stays in the same location for more consecutive time bins), the maximum predictability also increases, as shown in Figure 4. Our results indicate that it is easier to predict the location visited by the user in the next time bin rather than the next (distinct) location visited irrespective of the time of visit.

Impact of temporal resolution The next decision is the choice of temporal resolution Δt. We expect that when Δt is close to zero, the uncertainty in the next visited location (i.e., entropy) is minimal, and the predictability approaches 100%. To investigate this conjecture, we computed the entropy and predictability of the users' T^{stat} domain trajectories using varying time bins $\Delta t \in [0.25, 0.5, 0.75, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]$ minutes.[7]

As expected, we found that, as the temporal resolution gets more coarse-grained, the entropy increases while the predictability decreases (see Figure 5). Note that different values of Δt represent slightly different prediction tasks.

In our case, we would like to set a reasonable value to Δt based on our data. For this, we measure the average amount of time that users spend on each domain visit, finding an average of around 30 seconds (M=36.91, 95% CI [35.53, 38.29]). Note that the larger the size of Δt, the more short-duration visits would get removed from trajectories. Therefore, we use $\Delta t = 1$ minute for all other results in the paper.[8]

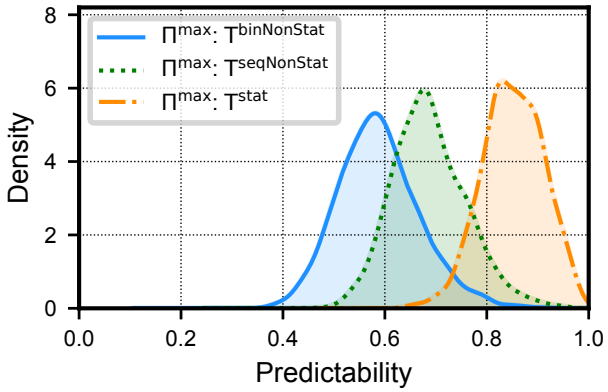


Figure 4: Impact of stationarity in the trajectories on the predictability.

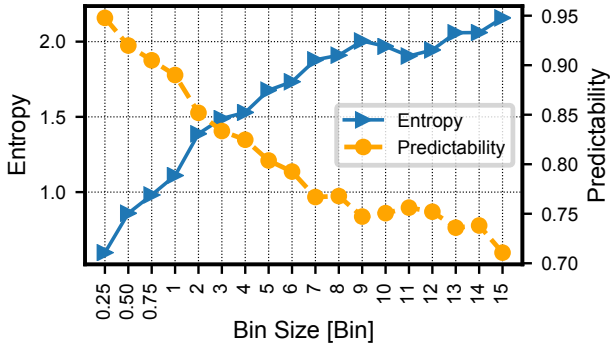
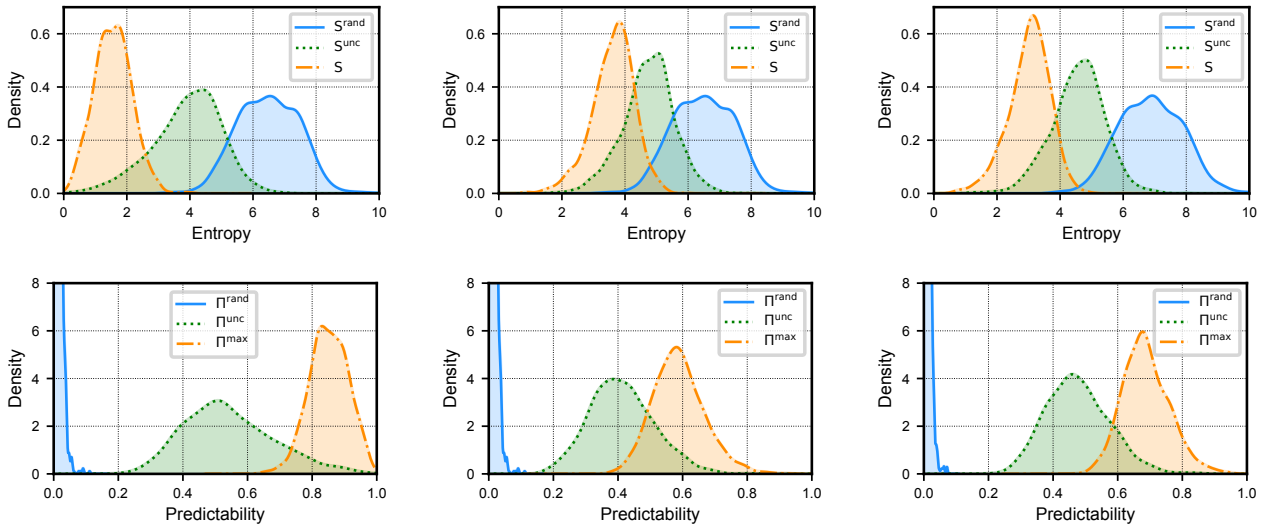


Figure 5: Impact of temporal resolution on the predictability.

Impact of spatial resolution The final design decision pertains to the spatial resolution of the locations in the analysis. In the cyberspace, this spatial resolution relates to hierarchical features of locations, such as URLs, sub-domains, domains, and categories. Deciding the resolution of the analysis typically depends on the problem at hand. In the limiting case, when the spatial resolution is too coarse-grained, and every web location is virtually the same, the predictability of the next visited location is perfect (i.e., the next location is always the same location). To investigate the role of spatial resolution, we measure the predictability of users' $T^{\text{seqNonStat}}$ trajectories of locations with decreasing spatial resolution, from URLs ($2,910,824$ unique URLs) to domains ($48,721$ unique domains) to categories (409 unique categories). We find that, indeed, the maximum predictability increases with more coarse-grained spatial granularity, as shown in Figure 7. These results imply that we should only compare the

[7] We found the average session length to be 883 seconds, which is just under 15 minutes. Hence, we vary Δt up to 15 minutes only.

[8] We also generated our results with Δt equal to 3, 5, and 8 minutes, finding that the trends in our results still hold.



(a) Stationary trajectory T^{stat}

(b) Non-stationary binned trajectory $T^{binNonStat}$

(c) Non-stationary sequential trajectory $T^{seqNonStat}$

Figure 6: Distribution of the three types of entropies (S^{rand} in blue, S^{unc} in green, and S in orange in the upper panel), and predictabilities (Π^{rand} in blue, Π^{unc} in green, and Π^{max} in orange in the lower panel) of all users in our dataset computed using: **(a)** stationary trajectories T^{stat}, **(b)** non-stationary binned trajectories $T^{binNonStat}$, and **(c)** non-stationary sequential trajectories $T^{seqNonStat}$.

predictability of users' movements if the spatial resolution is similar. In the rest of the paper, we present results at the resolution of domains, unless specified otherwise.

Predictability of Users' Mobility on the Web

For each user in our dataset, we computed the three entropy (S^{rand}, S^{unc}, and S) and predictability (Π^{rand}, Π^{unc}, and Π^{max}) values defined in Section "Estimating Predictability of Users' Mobility on the Web" for the stationary (T^{stat}) and non-stationary ($T^{binNonStat}$, $T^{seqNonStat}$) trajectories of visited domains; their distribution are presented in Figure 6.

For both types of trajectories, we found that the 'real' uncertainty (S) in people's movement in cyberspace is much lower than the expected entropy values when we disregard the order (S^{unc}) or users' preferences (S^{rand}) for the visited locations, as shown in the upper panel of Figure 6. These results indicate that people follow repetitive visitation patterns on the Web, that is, people exhibit web routineness in their browsing trajectories.

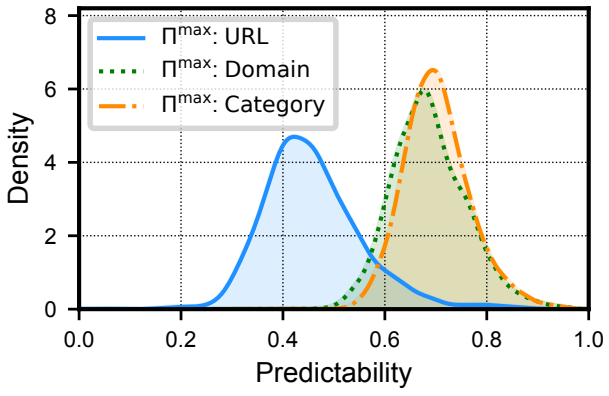
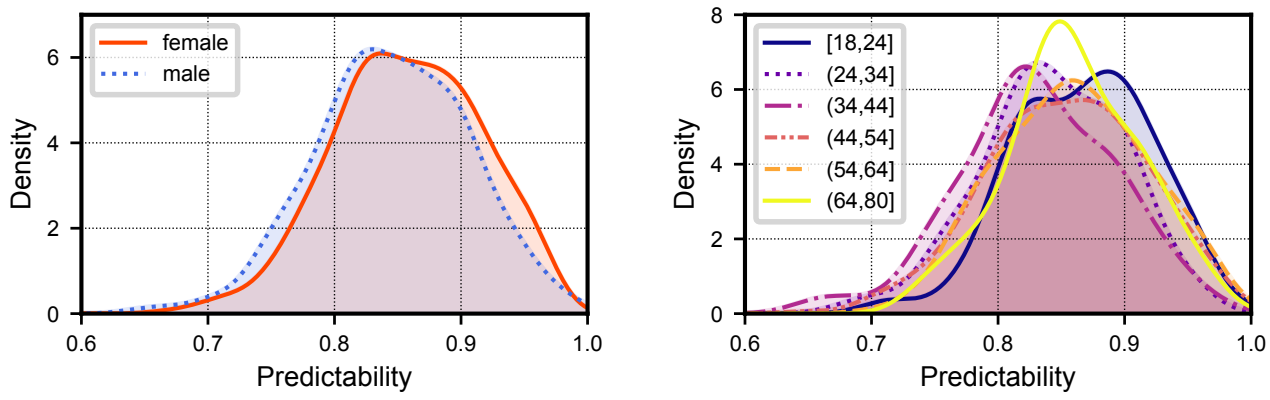


Figure 7: Impact of spatial resolution on the predictability.

In the case of the theoretical limits of predictability, we found that the more information we have about the users' trajectories—from the number of locations (Π^{rand}) to preference for locations (Π^{unc}) to order of location visited (Π^{max})—the higher the predictability. Our results show that the mean predictability increases from less than 5% to more than 85% (95% CI $[84.79, 85.41]$) for T^{stat}, around 60% (95% CI $[58.38, 59.20]$) for $T^{binNonStat}$, and around 70% (95% CI $[68.40, 69.14]$) for $T^{seqNonStat}$ (shown in the lower panel of Figure 6). Such an increase in predictability indicates that the routines users follow increase the predictive power we can theoretically achieve based on their web browsing trajectories.

When we compare our estimated theoretical limit of predictability with the performance of previously developed web access prediction algorithms that only utilize users' web access sequences, (Awad and Khan 2007; Narvekar and Banu 2015; Deshpande and Karypis 2004; Mabroukeh and Ezeife 2009; Khalil, Li, and Wang 2009; Manavoglu, Pavlov, and Giles 2003; Pitkow and Pirolli 1999), we observe that prior results are in line with our estimated theoret-



(a) Gender

(b) Age

Figure 8: Distribution of predictability $\Pi^{\max}$, computed using stationary trajectories T^{stat}, for users in our dataset grouped by: **(a)** gender and **(b)** age.

ical limits.[9]

Lastly, note that even though there are definite peaks of the maximum predictability distributions in Figure 6, there is a considerable spread across different users. This spread suggests that it is easier to predict the next visited location for some users than for the others. In the next section, we will explore which behavioral and demographic factors explain this variation in users' predictability.

What Explains the Predictability?

Demographics We now turn our attention towards examining whether users with different demographic attributes demonstrate different degrees of maximum predictability for their web browsing behavior. We focus our analysis on two demographic features: gender and age. Although in this section, we present the results computed using T^{stat} trajectories, our results also hold for non stationary trajectories.

- **Gender:** Our dataset of $1,455$ users consists of 51.5% men and 48.5% women (see Table **??**). We evaluated the differences between the distributions of the predictability for the men and women (depicted in Figure 8(a)) using a two-sample Kolmogorov-Smirnov (KS) test (Berger and Zhou 2014). We obtained a KS score of 0.086 at the p-value of 0.008 for the predictability. Since the p-value is below 0.05, we can reject the null hypothesis and conclude that the distributions of predictability of men and women are significantly different. Additionally, we also computed the effect size using Cliff's delta (Cliff 1993). The value of $d = 0.11$ indicates that there is an 11% chance that a randomly chosen woman has higher predictability than a randomly chosen man.
- **Age:** For studying whether users of different ages demonstrate different browsing behavior and hence are predictable to differing extents, we divided our set of $1,455$ users into six age groups: [18-24, 25-34, 35-44, 45-54, 55-64, 64-80] which contain [161, 267, 231, 357, 347, 70] users respectively (see Table **??**). Following the same procedure as above, we applied the two-sample KS test to the probability distribution of users in every pair of age groups. As before, we also computed Cliff's delta. We show the pairs of age groups for which the differences were significant in Table **??**.[10] We note that the "middle-aged" users (35-44) are typically lesser predictable than users of most other ages, as shown in the table. Interestingly, even among the "young" users (18-24 and 25-34), $d = -0.24$ indicates a 24% chance that a randomly chosen user below the age of 24 has higher predictability than a randomly chosen user above 24 years of age. We show the distribution of the predictability of users in different age groups in Figure 8(b).

Browsing Behavior Likely, the differences in predictability between men and women and young and older people can be explained by differences in their browsing behavior. We, therefore, analyze the impact of users' browsing characteristics on the predictability of the domain visited in the next time bin (using T^{stat} trajectories) and analyze differences between the behavior of different groups. We observe similar trends for the next visited categories and other trajectories but omit the results for brevity. For this analysis, we constructed some augmented features for each user in our

[9] We compare the prediction accuracy from prior work with our estimated limits for $T^{\text{seqNonStat}}$ trajectories since the web access sequences used in prior work match closest to them. However, we must note that without knowing the differences between the trajectories used in previous works and our trajectories as well as the differences in their server log datasets and our web tracking dataset, this is a less than optimal comparison.

[10] To adjust for multiple pairwise comparisons ($n = 15$), we apply the Bonferroni correction (Shaffer 1995). Doing so, we set alpha value for the entire set of age-based comparisons to $\alpha = 0.05$ by setting alpha value of each comparison to $\alpha/n = 3.3 \times 10^{-3}$.

Age groups	KS test		Cliff's delta
	KS score	p-value	
18-24 & 25-34	0.19	8.3×10^{-4}	-0.24
18-24 & 35-44	0.26	1.8×10^{-6}	-0.34
35-44 & 45-54	0.16	6.2×10^{-4}	0.19
35-44 & 55-64	0.21	3.1×10^{-6}	0.25
35-44 & 65-80	0.28	2.1×10^{-4}	0.27

Table 2: Differences in the predictability of users of different age groups. (Only significantly different pairs included.)

dataset based on their web browsing behavior. We then examined how the users' maximum predictability varies with these browsing features using Pearson's correlation coefficient r. Next, we describe our constructed browsing features and the variation of predictability with them.

- **User activity:** We quantify how active a user is via the total active seconds that the user is browsing and the count of the user's total domain visits during our one-month observation period. We find that predictability is positively correlated with total active seconds that the user spends browsing ($r = 0.4001$); however, it is not correlated with the total domain visits ($r = 0.0074$).
- **Diversity of user interests:** We capture how varied user's interests are, as the number of distinct domains and categories that the user visits. We observe that the more varied a user's interests are, the lower is the user's predictability as indicated by a negative correlation between predictability and number of distinct domains ($r = -0.2282$) and categories ($r = -0.1616$) visited.
- **User stationarity:** When a user spends more amount of time on average on each visit to a domain, we consider the user to have more stationarity in their browsing behavior since they are stationary on a domain for a long time before moving to another. Therefore, we consider the degree of stationarity in a user's browsing trajectory to be captured by the mean and median amount of time spent on each domain visit. In both cases, we observe a strong positive correlation between predictability and stationarity in the user's browsing, with r equal to 0.6502 and 0.3320, respectively. This increase in predictability with the increase in stationarity in users' browsing behavior also complements our results in Section "Impact of Design Decisions on Predictability".

Gender-based differences in Browsing Behavior While it is interesting that women are more predictable than men, the higher predictability for women in our dataset can potentially be explained by the lower diversity of interests (given by the number of distinct domains visited) for women (M=140.22, 95% CI $[132.16, 148.29]$) than men (M=161.46, 95% CI $[151.81, 171.13]$), and higher stationarity (captured as mean seconds spent per domain visit) for women (M=37.33, 95% CI $[35.54, 39.14]$) than men (M=33.14, 95% CI $[31.47, 34.83]$).

Age-based differences in Browsing Behavior In the analysis presented earlier, we found that the "middle-aged" users (35-44) are typically lesser predictable than users of most other ages. We observe that the users in the age group 35-44 have the least amount of stationarity in their browsing with the lowest mean amount of time spent per domain visit (M=29.32, 95% CI $[27.21, 31.42]$). This low stationarity could indicate that users in this age group perform a large variety of tasks on the Web. Even among "young" users (18-24 and 25-34), higher predictability of users below 24 can partially be explained by their higher stationarity (M=34.44, 95% CI $[31.17, 37.72]$) as compared to users above 24 (M=29.36, 95% CI $[27.74, 30.97]$).

Comparing Predictability in Physical- and Cyber-space

Finally, we briefly compare our computed predictability values for mobility in cyberspace with those presented in prior work for mobility in the physical space. We observe that the mean predictability of the next visited location is lower in cyberspace (85% for stationary and 59% for non-stationary trajectories) than in physical space (93% for stationary (Song et al. 2010b) and 71% for non-stationary (Ikanovic and Mollgaard 2017) trajectories).[11] Also, while we do observe differences in the distribution of predictability in the cyberspace for users with different demographic attributes, such gender- and age-based differences were not observed for the predictability in the physical space (Song et al. 2010b).

Discussion & Future Work

In this work, we adapted prior work on mobility in physical space to present an information-theoretic framework for estimating the theoretical limits of the predictability of people's mobility on the Web and provided some practical guidelines for its application. We applied the framework to a dataset of German users to demonstrate that people exhibit web routineness. Users' repetitive visitation patterns increase the theoretically achievable predictability of their web browsing behavior to up to 85%. We also observed gender- and age-based differences in the predictability of users' web browsing behavior. We highlighted that behavioral differences between demographic groups could partly explain these differences.

Framework premises In our paper, we analyzed the uncertainty embedded in time series (i.e., web trajectories), which enabled us to estimate the predictability upper bounds of users' trajectories on the Web. We highlight two relevant aspects of this framework. First, this upper bound assumes that individuals are somewhat stationary and unlikely to change their web browsing behavior drastically. In our analyses, we failed to recognize drastic changes during the

[11] It is worth noting that while we do compare the predictability results in physical- and cyber-space, it is not clear what the comparable values of Δt should be. Since we observe similarly reduced predictability for non-stationary trajectories also, we believe the reduction in predictability in cyberspace is a robust result.

observation period of one month. Future work could examine whether our results hold if one analyzes data that spans a more extended period (e.g., several years). Second, this upper bound tells us the theoretical upper limit to correctly predict an individual's future location by restricting ourselves to web tracking data only (i.e., location sequences). We show that having just this type of data grants one with a high potential predictive power. However, this upper bound is likely to increase by using different kinds of data, such as users' characteristics (e.g., demographic and behavioral features), visited locations' characteristics (e.g., topical categories), or temporal features (e.g., time of the day when web location was accessed, duration of access); these are also directions for future research.

Data limitations We focused on a dataset limited to one country and desktop browsing only. We are aware that a significant amount of browsing happens over mobile devices. In the future, we would like to compare our results for desktop browsing with those for mobile browsing. Moreover, there is always a chance that some of the web browsing activity that we analyze was not generated by humans but by bots. However, we believe this is unlikely for our dataset since the panelists' behavior is closely monitored by the panel company, and the panelists also take part in frequent surveys. In this regard, this type of data is of high quality and allows ethically studying the dynamics of human behavior on the Web.

Future work In the future, we would like to expand our current study to model the impact of more demographic and behavioral attributes of users and to include users from multiple countries to observe cross-cultural differences in the routineness of people on the Web. We would also like to explore the definition of distance measures or cost functions of moving from one web location to another to broaden our analysis to use distance-based or cost-aware models of mobility (e.g., radius of gyration). Another line of further research concerns the privacy implications of our work. By progressively removing available browsing history data for a user and studying the change in predictability for the next visited location, we can estimate the privacy risks against an attacker who may get access to a user's partial or full browsing history. Finally, in this work, we estimate the theoretical limits of predictability achievable on account of the routineness that people demonstrate in their web browsing behavior. Designing specific prediction algorithms to attain this theoretical maximum performance could be pursued in the future.

Acknowledgments

We acknowledge the support of Respondi AG, which provided the web tracking and survey data used in the present paper free of charge for research purposes, with special thanks to François Erner and Luc Kalaora at Respondi for their insights and help with data extraction.

References

Alessandretti, L.; Sapiezynski, P.; Sekara, V.; Lehmann, S.; and Baronchelli, A. 2018. Evidence for a conserved quantity in human mobility. *Nature Human Behaviour* 2(485).

Awad, M. A.; and Khan, L. R. 2007. Web navigation prediction using multiple evidence combination and domain knowledge. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 37(6): 1054–1062.

Bagrow, J. P.; Liu, X.; and Mitchell, L. 2019. Information flow reveals prediction limits in online social activity. *Nature human behaviour* 3(2): 122–128.

Barbosa, H.; Oliveira, M.; Pacheco, D.; Menezes, R.; and Goshal, G. 2018. The Influence of the Circadian and Ultradian Rhythms to Human Mobility: Empirical Evidences from Location-Based Check-Ins. In *First Northeast Regional Conference on Complex Systems (NERCCS)*.

Barbosa, H. S.; de Lima Neto, F. B.; Evsukoff, A.; and Menezes, R. 2016. Returners and Explorers Dichotomy in Web Browsing Behavior?A Human Mobility Approach. In *Complex Networks VII*, 173–184. Springer.

Berger, V. W.; and Zhou, Y. 2014. Kolmogorov–smirnov test: Overview. *Wiley statsref: Statistics reference online* .

Cliff, N. 1993. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological bulletin* 114(3): 494.

Cover, T. M.; and Thomas, J. A. 2006. *Elements of Information Theory*. Dordrecht: John Wiley & Sons, 2 edition. ISBN 978-0-471-24195-9.

Crutchfield, J. P. 2011. Between order and chaos. *Nature Physics* 8(1): 17–24. ISSN 1745-2473. doi:10.1038/nphys2190.

Crutchfield, J. P.; and Feldman, D. P. 2003. Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos* 13(1): 25–54. ISSN 10541500. doi:10.1063/1.1530990.

Culjak, G.; Kowalenko, N.; and Tennant, C. 2016. Awareness, Access and Use of Internet Self-Help Websites for Depression by University Students. *JMIR Mental Health* 3(4): e48. ISSN 2368-7959. doi:10.2196/mental.5311.

de Kunder, M. 2020. Daily Estimated Size of the World Wide Web. https://www.worldwidewebsize.com/. Accessed: 2020-11-12.

De Nadai, M.; Cardoso, A.; Lima, A.; Lepri, B.; and Oliver, N. 2019. Apps, Places and People: strategies, limitations and trade-offs in the physical and digital worlds. *arXiv preprint arXiv:1904.09350* .

Deshpande, M.; and Karypis, G. 2004. Selective Markov models for predicting Web page accesses. *ACM transactions on internet technology (TOIT)* 4(2): 163–184.

Gonzalez, M. C.; Hidalgo, C. A.; and Barabási, A.-L. 2008. Understanding individual human mobility patterns. *Nature* 453: 779–782.

Hazarie, S.; Barbosa, H.; Frank, A.; Menezes, R.; and Ghoshal, G. 2019. Uncovering the role of spatial constraints in the differences and similarities between physical and virtual mobility. *arXiv preprint arXiv:1907.04122* .

Hu, T.; Luo, J.; and Liu, W. 2018. Life in the "Matrix": Human Mobility Patterns in the Cyber Space. In *Twelfth International AAAI Conference on Web and Social Media*.

Hu, T.; Xia, Y.; and Luo, J. 2019. To Return or to Explore: Modelling Human Mobility and Dynamics in Cyberspace. In *The World Wide Web Conference*, 705–716. ACM.

Ikanovic, E. L.; and Mollgaard, A. 2017. An alternative approach to the limits of predictability in human mobility. *EPJ Data Science* 6(1): 12.

Karaçalik, O.; Oliveira, M.; and Kulshrestha, J. 2021. webtracking: A Library to Analyze Web Browsing Behavior. https://github.com/gesiscss/web_tracking/.

Kemp, S. 2019. Digital trends 2019: Every single stat you need to know about the internet. https://tnw.to/Bk8Ba. Accessed: 2020-11-12.

Khalil, F.; Li, J.; and Wang, H. 2009. An integrated model for next page access prediction. *IJ Knowledge and Web Intelligence* 1(1/2): 48–80.

Kolli, N.; Balakrishnan, N.; and Ramakrishnan, K. 2017. On Quantifying Predictability in Online Social Media Cascades Using Entropy. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 109–114.

Kontoyiannis, I.; Algoet, P.; Suhov, Y.; and Wyner, A. 1998. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Transactions on Information Theory* 44(3): 1319–1327. ISSN 00189448. doi:10.1109/18.669425. URL http://ieeexplore.ieee.org/document/669425/.

Kulshrestha, J.; Oliveira, M.; Karacalik, O.; Bonnay, D.; and Wagner, C. 2021. A web tracking data set of online browsing behavior of 2,148 users. doi:10.5281/zenodo.4383164. URL https://doi.org/10.5281/zenodo.4383164. https://doi.org/10.5281/zenodo.4383164.

Mabroukeh, N. R.; and Ezeife, C. I. 2009. Semantic-rich markov models for web prefetching. In *2009 IEEE International Conference on Data Mining Workshops*, 465–470. IEEE.

Manavoglu, E.; Pavlov, D.; and Giles, C. L. 2003. Probabilistic user behavior models. In *Third IEEE International Conference on Data Mining*, 203–210. IEEE.

Möller, J.; van de Velde, R. N.; Merten, L.; and Puschmann, C. 2019. Explaining Online News Engagement Based on Browsing Behavior: Creatures of Habit? *Social Science Computer Review* 0894439319828012.

Narvekar, M.; and Banu, S. S. 2015. Predicting user's Web navigation behavior using hybrid approach. *Procedia Computer Science* 45: 3–12.

Pappalardo, L.; Simini, F.; Rinzivillo, S.; Pedreschi, D.; Giannotti, F.; and Barabási, A.-L. 2015. Returners and explorers dichotomy in human mobility. *Nature communications* 6: 8166.

Pirolli, P. L.; and Pitkow, J. E. 1999. Distributions of surfers' paths through the World Wide Web: Empirical characterizations. *World Wide Web* 2(1-2): 29–45.

Pitkow, J.; and Pirolli, P. 1999. Mining longest repeating subsequences to predict world wide web surfing. In *Proc. UsENIX symp. on Internet Technologies and systems*, 1.

Ravizza, S. M.; Uitvlugt, M. G.; and Fenn, K. M. 2017. Logged In and Zoned Out: How Laptop Internet Use Relates to Classroom Learning. *Psychological Science* 28(2): 171–180. ISSN 0956-7976. doi:10.1177/0956797616677314.

Scharkow, M.; Mangold, F.; Stier, S.; and Breuer, J. 2020. How social network sites and other online intermediaries increase exposure to news. *Proceedings of the National Academy of Sciences* 117(6): 2761–2763.

Shaffer, J. P. 1995. Multiple hypothesis testing. *Annual review of psychology* 46(1): 561–584.

Sinatra, R.; and Szell, M. 2014. Entropy and the predictability of online life. *Entropy* 16(1): 543–556.

Smith, G.; Wieser, R.; Goulding, J.; and Barrack, D. 2014. A refined limit on the predictability of human mobility. In *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 88–94. IEEE. ISBN 978-1-4799-3445-4. doi:10.1109/PerCom.2014.6813948.

Song, C.; Koren, T.; Wang, P.; and Barabási, A.-L. 2010a. Modelling the scaling properties of human mobility. *Nature Physics* 6(10): 818–823. ISSN 1745-2473. doi:10.1038/nphys1760.

Song, C.; Qu, Z.; Blumm, N.; and Barabási, A.-L. 2010b. Limits of predictability in human mobility. *Science* 327(5968): 1018–1021.

Stier, S.; Kirkizh, N.; Froio, C.; and Schroeder, R. 2020. Populist Attitudes and Selective Exposure to Online News: A Cross-Country Analysis Combining Web Tracking and Surveys. *The International Journal of Press/Politics* doi:10.1177/1940161220907018.

Su, Z.; Yang, Q.; Lu, Y.; and Zhang, H. 2000. Whatnext: A prediction system for web requests using n-gram sequence models. In *Proceedings of the First International Conference on Web Information Systems Engineering*, volume 1, 214–221. IEEE.

Tao, Y.; Cui, H.; Nian, Y.; Hong, Y.; Wang, J.; and Zhang, H. 2019. Behavior predictability and consistency of mobile users' traffic usage between different years based on entropy theory. *International Journal of Communication Systems* 32(16): e4052.

Zhao, Y.-M.; Zeng, A.; Yan, X.-Y.; Wang, W.-X.; and Lai, Y.-C. 2016. Unified underpinning of human mobility in the real world and cyberspace. *New Journal of Physics* 18(5): 053025.

Zhao, Z.-D.; Cai, S.-M.; and Lu, Y. 2015. Non-Markovian character in human mobility: Online and offline. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 25(6): 063106.

Zhao, Z.-D.; Huang, Z.-G.; Huang, L.; Liu, H.; and Lai, Y.-C. 2014. Scaling and correlation of human movements in cyberspace and physical space. *Physical Review E* 90(5): 050802.