# Contrasting social and non-social sources of predictability in human mobility

Zexun Chen,[1] Sean Kelty,[2] Brooke Foucault Welles,[3]
James P. Bagrow,[4] Ronaldo Menezes,[1,*] and Gourab Ghoshal[2,5,†]

[1]*BioComplex Laboratory, Department of Computer Science, University of Exeter, UK*

[2]*Department of Physics and Astronomy,*
*University of Rochester, Rochester, NY, USA*

[3]*Northeastern University, Boston, MA, USA*

[4]*Department of Mathematics & Statistics and Vermont Complex Systems Center,*
*University of Vermont, Burlington, VT, USA*

[5]*Department of Computer Science, University of Rochester, Rochester, NY, USA*

## Abstract

Social structures influence a variety of human behaviors including mobility patterns, but the extent to which one individual's movements can predict another's remains an open question. Further, latent information about an individual's mobility can be present in the mobility patterns of both social and non-social ties, a distinction that has not yet been addressed. Here we develop a "colocation" network to distinguish the mobility patterns of an ego's social ties from those of non-social colocators; individuals not socially connected to the ego but who nevertheless arrive at a location at a similar time as the ego. We apply entropy and predictability measures to analyse and bound the predictive information of an individual's mobility pattern and the flow of that information from their top social ties and from their non-social colocators. While social ties generically provide more information than non-social colocators, we find that significant information is present in the aggregation of non-social colocators: 3–7 colocators can provide as much predictive information as the top social tie, and colocators can replace up to 85% of the predictive information about an ego, compared with social ties that can replace up to 94% of the ego's predictability. The presence of predictive information among non-social colocators raises privacy concerns: given the increasing availability of real-time mobility traces from smartphones, individuals sharing data may be providing actionable information not just about their own movements but the movements of others whose data are absent, both known and unknown individuals.

[*] Correspondence email address: r.menezes@exeter.ac.uk

[†] Correspondence email address: gghoshal@pas.rochester.edu

# 1. INTRODUCTION

The recent availability of extensive geolocated datasets related to human movement, has enabled the quantitative study of human movement at an unprecedented level [1], contributing greatly to insights in estimating migratory flows, traffic forecasting, urban planning, mitigating pollution and epidemic modeling among other applications [2–9]. Several common regularities have been observed across these studies, including bursty activity rates, tendencies to visit a select few locations disproportionately more than others, as well as decreasing likelihood to explore as time goes on [10–15]. A related aspect that can enhance the potential of these findings, particularly for urban planning and the control of epidemics, is the ability to predict the future locations of individuals or groups using their prior history of travel. Indeed, it has been shown that a perfect algorithm can predict, with between 70-90% certainty, an individual's future location given their prior location visits [16], depending upon the spatiotemporal granularity of observations [17].

Human beings are also typically highly social creatures and social structures can influence behavior in a variety of human activities including movement patterns. In fact, it has been shown that social relationships statistically account for between 10% to 30% of all human movement [18]. Social structures inherently encode information flow between parties, such that residual information about an individual can be inferred from their social ties. Such a phenomenon was demonstrated in the context of online interactions, where about 95% of an individual's potential predictive accuracy was contained in their social network, despite no recourse to information about the person in question [19]. Coupled with the observation that movement patterns in the virtual and physical domains are strikingly similar [20], this leads to the intriguing question as to whether one can leverage a person's social network to predict their future mobility patterns, absent any information on their own history. This possibility holds promise for a number of applications, and may be particularly relevant in the context of mitigating future pandemics [21, 22], where a key tool in the arsenal is contact tracing based on mobility patterns [23, 24]. However, accurately mapping human mobility can be challenging due to understandable privacy concerns and people's willingness to disclose or share personal data [25, 26].

Location-Based Social Networks (LBSNs) yield opportunities to examine social relations to human mobility, containing information both about sequences of location visits and (in some cases) information on the underlying social network. At the same time, spatially aggregating these data can reveal individuals in different social circles who visit similar or overlapping locations; for instance, people working in the same building but with different companies, or parents whose children attend the same schools but are unknown to each other. These non-social ties are potential predictors of a person's mobility trajectory. Terming such individuals "non-social colocators," we ask whether and to what extent such colocators yield predictive mobility information, and how this information compares to that of social ties.

Here we apply non-parametric information-theoretic estimators to study human mobility extracted from three Location-Based Social Networks (LBSNs), that contain sequences of location trajectories as well as the (reported) social network of a subset of users. We demonstrate the existence of information transfer in these networks, finding that a given ego's future location visits can be predicted, with between 80–100% of the ego's own accuracy, by studying the historical patterns of just 10 of their alters (ranked by number of common locations visited). Remarkably, non-social colocators, while individually providing less information than social ties, can in the aggregate provide similar levels of predictability. The information flow provided by colocators is also surprisingly robust to temporal-displaced colocations, implying users that never physically colocate can still provide comparable information to social ties. Indeed, the information transfer appears to be driven by the overlap of unique locations visited by the ego and alters, in both social ties and non-social colocators.

The rest of this paper is organized as follows. Section 2 describes the mobility and social datasets we study. We apply information-theoretic tools to these data in Section 3, where we quantify the predictive information of individuals (Sec. 3.1), their social ties and their non-social colocators (Sec. 3.2). We examine the spatial and temporal underpinnings of our results (Sec. 3.3) and end in Section 4 with a discussion of the implications of our findings.

## 2. MATERIALS

Our study uses three publicly available datasets that contain both mobility traces and the social network of a subset of the users of the platforms. The first is BrightKite, a location-based social networking service (LBSN) [18, 27] containing 4,491,143 geo-tagged check-ins by 58,228 users over a period of Apr 2008 - Oct 2010. The second dataset is from Weeplaces, a website that generated visualizations and reports from location-based check-ins in platforms such as Facebook and Foursquare [20]. The considered data contains only Foursquare check-ins, that includes 7,658,368 geotagged check-ins produced by 15,799 users from Nov 2003 to Jun 2011. Finally, we also consider Gowalla [18], another LBSN consisting of 6,442,890 check-ins by 196,591 users over a period of Feb. 2009 - Oct. 2010. (For more details, see Sec. S1.)

Within these data, each *event*, i.e., an instance of a location visit, is timestamped and tagged with a unique location ID. In all datasets, a location visit $v$ is represented by a tuple $v = (u, \ell, t)$, meaning a user $u$ visited a location $\ell$ at time $t$. At a user level, a trajectory composed of $N_u$ discrete observations is characterized by a sequence of $N_u$ location-time pairs $(\ell_i, t_i)_{i \in 1 \ldots N_u}$ where $\ell_i$ stands for the location visited at step $i$ and time $t_i$. We assume a user who visits $N_u$ locations in total, visits $n_u \leq N_u$ *distinct* locations, with equality holding only if the user never visits a location more than once. To filter out spurious activities, we exclude inactive users and discard records with missing attributes. Furthermore, for purposes of statistical significance, we also discard users who have logged $N_u < 150$ check-ins (our results are robust to these filtering criteria; see Sec. S1.3 and Figs. S3 and S4). After filtering, we are left with 510,308 events across 6,132 users in Brightkite, 924,666 events across 11,533 users in Weeplaces, and finally 850,094 events across 9,937 users for Gowalla (cf. Table S1 for further details). In Fig. S1 we show the check-in maps for each of the datasets indicating global coverage with the highest concentrations in North America and Western Europe. In Fig. S2 we plot the corresponding total distinct locations visited by all users, jump-length and radius of gyration distributions finding statistical trends in the LBSN data consistent with other sources of mobility data [1].

Each of the datasets have social networks collected by their respective API's (details in Sec. S1.1), however, not all of the users in the network log check-ins. Given that our goal is to examine the information transfer in these social networks as it relates to location visits, we focus on users with logged location-trajectories. To quantify the information provided by colocated non-social ties, we construct colocation networks where a tie is included between an ego and alter if they checked in at the same location within a particular time window (See Sec. S2 for details on egocentric network construction). We assume that individuals who colocate more often contain more predictive information about one other's whereabouts, so the ranking criteria is based on the frequencies of colocations (see Sec. S2.4, Figs. S5 and S6). All results presented in the main manuscript correspond to a one hour temporal bin, but our results are robust to varying temporal frames (Sec. S2.5 and Fig. S7).

## 3. RESULTS

### 3.1. Information contained in egos

We begin our analysis by examining the information contained in the location trajectories of all ego's in each of the datasets; this serves as a baseline when comparing information flow with social ties and non-social colocators. The degree of uncertainty in capturing the future locations of a trajectory $A$, given past observations, is encoded in the entropy rate $S_A$ of the trajectory. Accounting for both frequency of location visits, as well as temporal ordering (specific ordered sequences in the data), we make use of a non-parametric estimator [28, 29] given by the expression

$$\hat{S}_A = \frac{N \log_2 N}{\sum\limits_{i=1}^{N} \Lambda_i}, \tag{1}$$

where for a trajectory $A$ of $N$ moves of an individual, $\Lambda_i$ is the length of the shortest trajectory sub-sequence beginning at position $i$ not seen previously, and the entropy is measured in bits. This estimator has been applied to mobility patterns and online social activities [16, 19]. In the absence of any structure in the sequence, the expression converges to the standard Shannon entropy [19]. In Fig. 1A, we plot $\hat{S}_A$ for the three datasets finding peaks between 4–5 bits with varying degrees of spread. (One dataset, BrightKite, looks distinct from the others for reasons we discuss shortly.)

An intuitive measure to interpret these results is the *perplexity* $2^S$: we are as uncertain about future visits for a trajectory with entropy rate 3 bits, for example, as we would be when choosing uniformly at random from $2^3 = 8$ possible locations. Using $\hat{S}_A$ from Fig. 1A, this implies that on average knowing the past history of the typical ego allows us to reduce the possible number of future location visits to between 16–32 sites. Given the

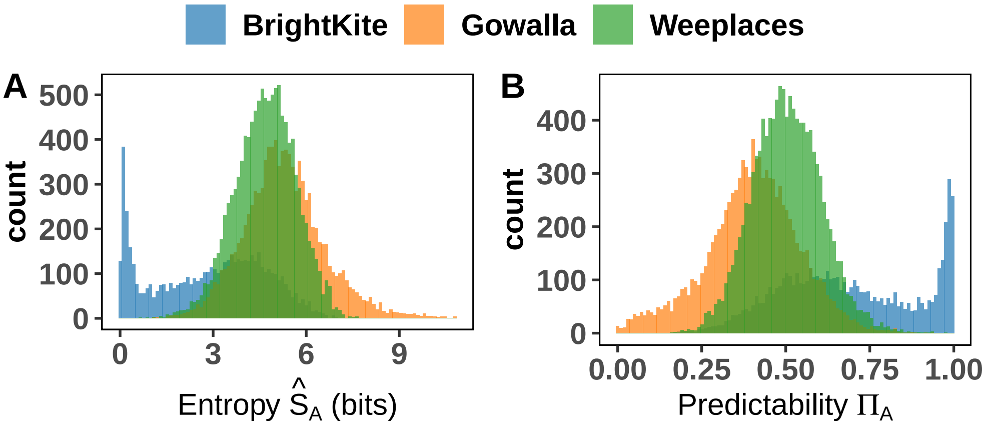

FIG. 1. **Entropy and Predictability in three mobility datasets.** **A**, The distribution of the entropy $\hat{S}_A$ (Eq. (1)) for each of the three datasets. **B**, The corresponding distribution of predictability $\Pi_A$, calculated by inverting Eq. (2), tells us how well an ideal algorithm can predict an individual's future location given their mobility history.

average number of distinct locations (Fig. S2) that a typical user visits across the datasets (107 total distinct locations per user on average in BrightKite, 166 in Weeplaces, and 198 in Gowalla), information due to the spatiotemporal regularities of ego trajectories represent an order of magnitude reduction from choosing across all locations uniformly at random, a result consistent with that found in other mobility studies [16, 17].

The entropy rate can also be interpreted using Fano's inequality [30] to define the *predictability* $\Pi_A$, the upper bound of how often an *ideal predictive algorithm* can correctly guess the next location visit, given prior history. This predictability is calculated by inverting Fano's inequality:

$$\hat{S}_A \leq H_A(\Pi_A) + (1 - \Pi_A)\log_2(n-1), \tag{2}$$

where $n$ is the number of distinct locations visited and $H(x)$ is the binary entropy function capturing the entropy of a simple Bernoulli trial (in this case achieving maximal predictability or not). Utilizing $\Pi_A$ allows us to leverage information theory to mathematically bound the performance of all real predictive methods given an information sources inferred uncertainty.

Figure 1B shows the distributions of predictability, finding key difference across the three platforms. One dataset in particular, BrightKite, shows a distinct spike of highly predictable ($\Pi_A \approx 1$) users. In BrightKite, a large fraction of users visit only between 1–3 unique locations (Fig. S2A), leading to a low entropy rate and thus a disproportionately high degree of predictability. Conversely, in Gowalla, while $\Pi_A$ is peaked at $\approx 40\%$, we find a wide spread around the peak, given that some users visit many locations (indeed, 23 users never return to a previously visited location), leading to a high entropy rate, and low predictability. This likely stems from Gowalla incentivizing its users to discover new locations (see Sec. S1). Finally, in Weeplaces, $\Pi_A$ is peaked at $\approx 50\%$ with a tighter bound around the peak as compared to the other datasets. This observed diversity in mobility behavior across the three platforms provides us with a robustness check on the results to follow.

### 3.2. Information contained in social ties and non-social colocators

Next we examine information flow, how much mobility information about the ego is contained in the sequence of location visits of their alter(s), absent any information about the ego's own location history. This can be measured by the *cross-entropy* [19, 31], which is greater than the entropy when the alter contains less information on the ego than the ego itself, and quantifies information loss when we have access to only the alter's past. To estimate the cross-entropy between two sequences $A$ and $B$, Eq. (1) can be modified to account for two sequences $A$ and $B$ (representing the mobilities of the ego and alter, respectively) according to

$$\hat{S}_{A|B} = \frac{N_A \log_2(N_B)}{\sum_{i=1}^{N_A} \Lambda_i(A|B)}, \tag{3}$$

where $N_A$ and $N_B$ are the lengths of the sequences $A, B$, and the cross-parsed match length $\Lambda_i(A|B)$, is the length of the shortest location sub-sequence starting at position $i$ of sequence $A$ not previously seen in sequence $B$. Here, 'previously' refers to those locations $\ell_j$ in sequence $B$ with $t_j < t_i$, the timestamp of the check-in location $\ell_i$ in sequence $A$. As with the cross-entropy, one can generalize the predictability $\Pi_A$ to the cross-predictability $\Pi_{A|B}$ by applying Eq. (2) to Eq. (3). For the remainder of this paper, both social ties and non-social colocators have been processed by retaining alters that provide better-than-random information about their ego, as well as removing from the colocation network any spurious colocators (see Secs. S2.1 and S2.2 and Table S2).

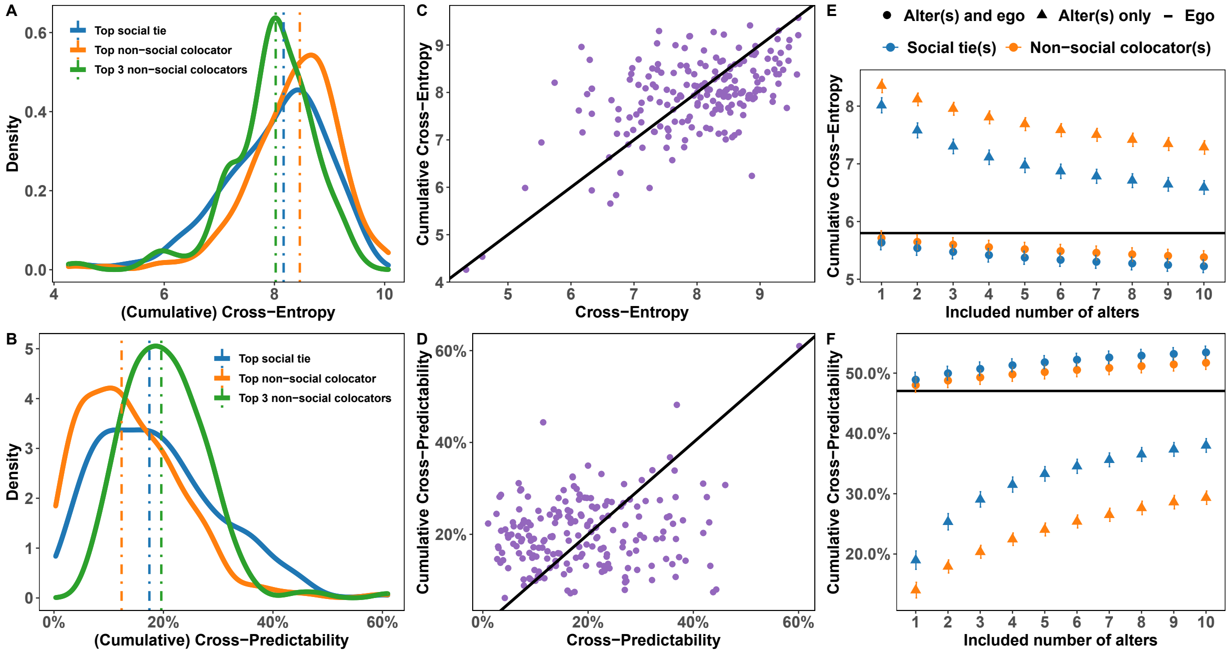

FIG. 2. **Cross-entropy and predictability in social ties and non-social colocators**. **A** Distributions of $\hat{S}_{A|B}$ for the rank-1 social tie (median 8.17 bits), non-social colocator (median 8.46 bits), and $\hat{S}_{A|\mathcal{B}}$ for the top-3 non-social colocators (median 8.02 bits) in Weeplaces. **B** The corresponding $\Pi_{A|\mathcal{B}}$ for the social (median17.43%), and non-social colocators (median 12.35%), and $\Pi_{A|\mathcal{B}}$ for the top-3 non-social colocators (median 19.60%). **C** $\hat{S}_{A|B}$ encoded in the top-social tie as a function of $\hat{S}_{A|\mathcal{B}}$ for the top-3 non-social colocators. Each point corresponds to a single ego and the solid line denotes $y = x$. **D** As in panel **C** but with predictability instead of cross-entropy. **E, F** $\hat{S}_{A|\mathcal{B}}$ and $\Pi_{A|\mathcal{B}}$ after accumulating the top-ten social alters and non-social colocators. Horizontal lines denote the average entropy (5.80 bits) of egos and their self-predictability (47.05%). Shapes indicate whether the self-predicatbility of the the ego was included in the sequence (circles) or excluded (triangles). Error bars denote 95% CI.

Figure 2 shows the results of our information metrics on the Weeplaces dataset. Panels **A** and **B** show the distribution of the cross-entropy and predictability for the rank-1 social tie and non-social colocated alter. We see that the top social tie provides slightly more information than the top colocator, with predictability slightly right-skewed (Fig. 2B). While social ties provide more predictive information, the distribution also shows the existence of some non-social colocators that provide mobility information comparable to that provided by social ties. Furthermore, the predictability of egos are positively correlated with the predictability of their top alter (Fig. S10), meaning highly predictable egos tend to have highly predictable top alters, and similarly more unpredictable egos tend to have less predictable alters. These findings are consistent with that seen for Brightkite (Fig. S8**A,B**) and Gowalla (Fig. S9**A,B**).

We've thus far looked at the individual and pair-wise information in ego-alter pairs, a limited analysis, given that these are being considered as information sources in isolation. Next, we examine the information content of a multiplicity of an ego's alters, or in other words examine the information content of a subset of the ego's colocators by adapting the cross-entropy to a set of alters. To estimate the amount of information needed to encode the next location of sequence $A$ given the location information in *a set of sequences* $\mathcal{B}$, we generalize the pairwise cross-entropy to the *cumulative cross-entropy* according to [19], thus,

$$\hat{S}_{A|\mathcal{B}} = \frac{N_A \log_2(N_{A\mathcal{B}})}{\sum_{i=1}^{N_A} \Lambda_i(A|\mathcal{B})}, \tag{4}$$

where $\Lambda_i(A|\mathcal{B}) = \max\{\Lambda_i(A|B), B \in \mathcal{B}\}$ is the longest cross-parsed match length over any of the sequences in the set of sequences $\mathcal{B}$, $N_{A\mathcal{B}} = \sum_{B\in\mathcal{B}} w_B N_B / \sum_{B\in\mathcal{B}} w_B$ is the average of the lengths $N_B$, and $w_b$ is the number of times that matches from sequence $A$ are found in each sequence $B \in \mathcal{B}$. Note that if there is only one sequence in $\mathcal{B}$, Eq. (4) reduces to Eq. (3). By applying Fano's inequality (Eq. (2)), we denote the corresponding cumulative cross-predictability as $\Pi_{A|\mathcal{B}}$. In Fig. S6 we see that on average, alters associated with more frequent colocations contain more information content, as a consequence, we rank alters according to frequency of colocations. Plotting the average number of colocations between ego-alter pairs saturate at around 10 alters in all three datasets (Fig. S5), and therefore we examine the information content of the top-10 most frequently colocated social alters and non-social colocated alters. For a fair comparison between these two different sources of mobility information, we focus on egos in both the social and colocation network with at least ten alters in each network, leading to 33 (Brightkite), 97 (Gowalla), and 199 (Weeplaces) egos (cf. Table S2 for details).

By moving from one non-social colocator to three, we see in Fig. 2A and B that considerably more predictive information is present, with the peak of $\Pi_{A|B}$ shifted significantly rightward. Further, the peak is now at a higher value than the peak for the top social tie (Fig. 2B), indicating that many egos are better predicted by three non-social colocators than they are by their top social tie. We further emphasize this relationship in Fig. 2C,D with scatter plots comparing the cross-entropy of the top social tie to the cumulative cross-entropy of the top-3 non-social colocators; any points above the line $y = x$ in panel **D** demonstrate more information flow from the colocators about the ego than from the top social tie. Individually, non-social colocators are less informative than social ties, but collectively they can meet or exceed the information content of individual social ties.

Expanding on the comparison between social ties and non-social colocators, in Fig. 2**E**,**F**, we plot the cumulative cross-entropy and cross-predictability $\Pi_{A|\mathcal{B}}$, finding a progressive increase in predictability as we accumulate more alters (positive Spearman's $\rho$ across 88.94% all users, $p < 0.05$). A given number of social ties provides more information on average than the same number of non-social colocators, as demonstrated by the lower curve in entropy in panel **E** and higher curve in predictability in panel **F**. Specifically, 94.47% of egos in Weeplaces show significantly higher social tie predictability than non-social colocator predictability (paired one-sided $t$-test, $p < 0.01$). However, while the colocator curve in Fig. 2**F** sits below the social curve for a given number of alters, we do see that on average a greater number of colocators can exceed the information content of a small number of social ties. For instance, the top-3 non-social colocators provide higher predictability than the top social tie, and the top-7 colocators provide higher predictability than the top-2 social ties Consistent results hold for the BrightKite and Gowalla datasets (cf. Sec. S3, and Fig. S8F and Fig. S9F) and for different definitions of colocation (cf. Fig. S7).

While alter information about the ego is important to understand, especially for matters of privacy (see Discussion), we also wish to understand whether that information is redundant when given the ego's past. We therefore show in Fig. 2 **E**,**F** curves including the ego's past alongside that of the alters. We find that non-redundant information exists in both types of alters, with a gain of $\approx$ 10% predictability for the top-10 social ties and $\approx$ 14% for the top-10 non-social colocators. Figure 2**F** also demonstrates that $\Pi_{A|\mathcal{B}}$ appears to saturate as more alters are included, an effect observed in other studies [19]. This saturation effect is examined in Sec. S4, where $\Pi_{A|\mathcal{B}}$ is extrapolated beyond our data window of 10 alters by fitting a nonlinear saturating function and estimating the extrapolation predictability $\Pi_\infty$. For Weeplaces we find $\Pi_\infty = 44.32\%$ and 39.79% for social ties and colocators, respectively. Compared to the average $\Pi_{\text{ego}} = 47.05\%$ all egos in the network, this means that 94% and 85%, respectively, of the *potential* predictability of an ego is in principle available in that ego's alters. The closeness of these values underscore the high degree of predictive information available in the non-social colocators. The corresponding findings for the other two datasets are shown in Table S3.

Extrapolation analysis demonstrates the overall relative value of non-social colocators,

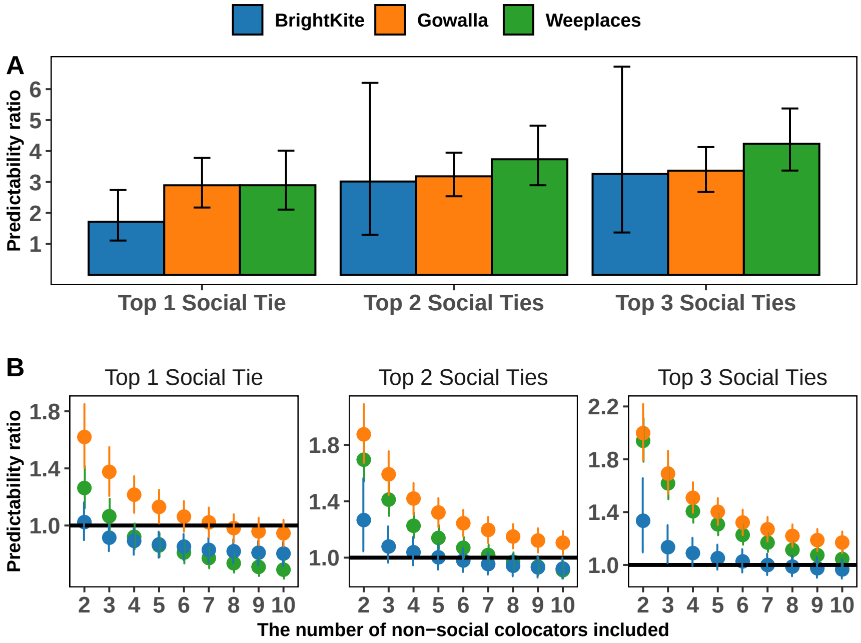

FIG. 3. **Quantifying the predictive information aggregated from non-social colocator(s) with respect to social ties.** **A** The predictability ratio $\Pi_{\text{ego}\,|\,\text{social tie(s)}}/\Pi_{\text{ego}\,|\,\text{non-social colocator(s)}}$ between the top non-social colocator and (left-to-right) the top, top-2, and top-3 social tie(s). **B** The predictability ratios between the top 2–10 non-social colocators and the top, top-2, and top-3 social tie(s). Error bars denote 95% CI.

but it does not allow us to determine more precisely how many non-social colocators equal the information content for a given number of social ties. Therefore, to better quantify the relative information content provided by social ties compared with non-social colocators, we examine the predictability ratio $\Pi_{\text{ego}\,|\,\text{social tie(s)}}/\Pi_{\text{ego}\,|\,\text{non-social colocator(s)}}$ across all three mobility datasets. In Fig. 3**A** we present the distributions of predictability ratio comparing the top non-social colocator to the top-$k$ social ties ($k = 1, 2, 3$). For the top social ties ($k = 1$) we see that BrightKite colocators provide the closest information with a ratio just below 2, meaning the social tie provides approximately twice the predictability of the colocator. In Gowalla and Weeplaces, the difference is even stronger, with the top social ties providing approximately three times the predictability of the top colocator. Moving from the top colocator to multiple colocators, in Fig. 3**B** we plot the predictability ratio for increasing numbers of nonsocial colocators; when this curve crosses the horizontal line at a ratio of 1, we have equal amounts of information. For example, examining the first panel in **B**, we see that curve for BrightKite intersects between 1 and 2 colocators, between 7 and 8 colocators, and between 3 and 4 for Weeplaces. This suggests that three Weeplaces colocators are equivalent to the top social tie, while 7–8 Gowalla colocators are equivalent to the top social tie. In Brightkite, a dataset characterized by high predictability and low entropy, one added social tie provides the same degree of information as two non-social colocators. Across all datasets, we see that an aggregate of fewer than 10 non-social colocators can equal the information of the top social tie. While more colocators are needed to equal the aggregate of the top-2 or top-3 social ties, the observed decreasing trends suggest a convergence in the amount of information contained in either flavor of tie.

### 3.3. Underlying Spatial and Temporal Mobility Characteristics

We next determine the key factors that determine the near identical types of information transfer in both types of ties, despite having no overlap in the pair-wise connections. One of the possibilities driving the quantity of information on the ego provided by alters is the information inherent in the locations themselves. That is, it is reasonable to surmise that information about the ego is derived from shared visits to common locations, given that predictability of the ego itself depends on the patterns of location visits in their trajectory. While alters do not necessarily visit all the locations that their egos do, nor would they necessarily visit at the same time (see below), one can hypothesize that higher-ranked alters share more distinct locations with the ego than lower-ranked ones. If the trend is similar across both social ties and non-social colocators, then this would be a plausible mechanism for the similarity in the observed cross-predictability.

To measure this, we compute the proportion of unique locations visited by the ego and its alters. For an ego $A$ we define the *Overlapped Distinct Location Ratio (ODLR)* $\eta$ as the fraction of $A$'s visited locations also visited by an alter $B$. Formally,

$$\eta_{A|B} = \frac{|Y_A \cap Y_B|}{|Y_A|} \tag{5}$$

where $Y_A$ and $Y_B$ are the sets of locations visited by $A$ and $B$, respectively, and $|\cdot|$ denotes set cardinality.

In Fig. 4**A** we plot the ODLR as a function of alter rank. As alters are ranked according to the frequency of overlap of any location visit of the ego, as opposed to distinct location visits, there is no reason to *a priori* expect that a rank-1 alter will share the most number

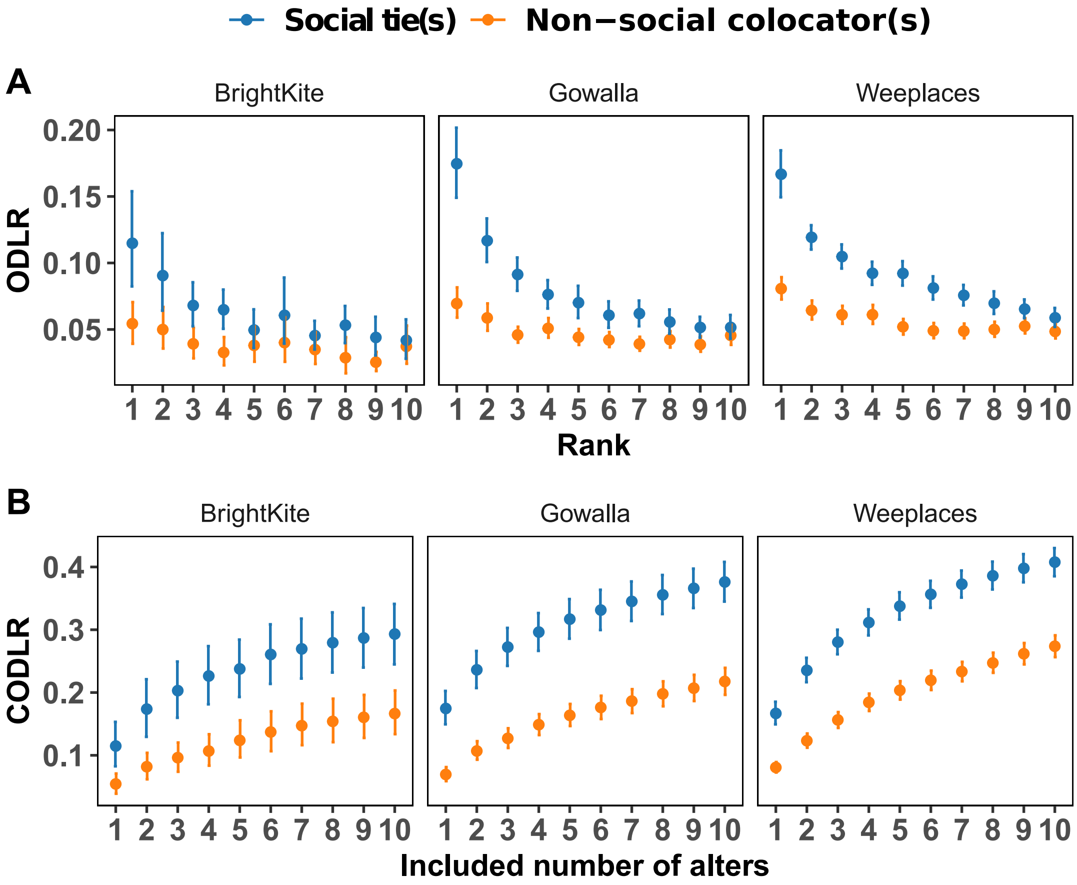

FIG. 4. **The degree of ego–alter distinct location overlap**. **A**, The Overlapped Distinct Location Ratio (Eq. (5)) indicates that higher ranked alters share more unique location visits than lower-ranked ones, with the top (rank-1) alter showing the most shared location. The trend is stronger for social ties than non-social colocators. **B**, The Cumulative Overlapped Distinct Location Ratio (Eq. (6)) shows increasing discovery of unique locations in the the ego's trajectory as alters are added in order of decreasing rank, but that the rate of discovery slows. Error bars denote 95% CI.

of distinct locations in their trajectory with the ego. Nevertheless, that is indeed what is observed across all datasets, with a monotonically decreasing trend of ODLR as one considers lower-ranked alters. This monotonic trend is considerably stronger for social ties than for non-social colocators, although the difference diminishes with the number of alters added.

The ODLR fails to consider locations shared across multiple alters, instead focusing on one alter at a time. Yet we previously saw (Fig. 2) the importance of examining the aggregate information, particularly when comparing non-social colocators to social ties. Therefore, we generalize the ODLR to a *Cumulative Overlapped Distinct Location Ratio*

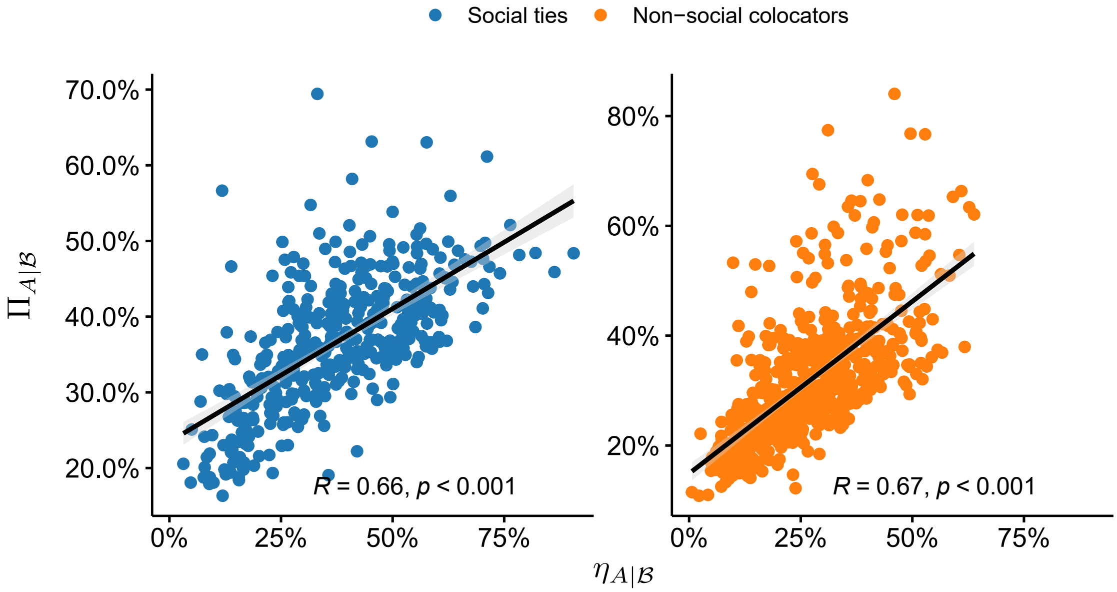

FIG. 5. **Connecting location overlap to information transfer.** Regression analysis of cumulative cross-predictability $\Pi_{A|\mathcal{B}}$ and CODLR $\eta_{A|\mathcal{B}}$ for the Weeplaces dataset with the top-10 alters. Here $R$ is Pearson's correlation coefficient. The solid lines are linear regressions.

(CODLR) by taking the union of the location sets of multiple alters according to,

$$\eta_{A|\mathcal{B}} = \frac{|\cup_{B\in\mathcal{B}}(Y_A \cap Y_B)|}{|Y_A|}, \tag{6}$$

where $A$ is the ego and $\mathcal{B}$ is the set of all alters. We plot the results in Fig. 4**B** finding similar increasing monotonic trends across datasets and networks. As alters are added, more information on the ego's unique locations are discovered, saturating at between 30-40% after 10 social ties, and between 15-30% for non-social colocators. Once again we observe that larger numbers of colocators provide comparable location overlap as a smaller number of social ties, emphasizing both the relative importance of social ties and the extent of useful information present in the aggregation of non-social colocators.

We connect ODLR and information flow directly in Fig. 5, by plotting $\eta_{A|\mathcal{B}}$ against $\Pi_{A|\mathcal{B}}$ for the top-10 alters in both types of networks, observing a strong, approximately linear trend (Pearson's $R \approx 0.66$ for social ties; $R \approx 0.67$ for non-social colocators; both significant, $p < 0.001$). Disentangling the plots by progressively adding alters from rank-1 to rank-10 shows a monotonically increasing trend for the correlations (Figs. S11 and S12). The corresponding results for BrightKite (Figs. S13 and S14) and Gowalla (Figs. S15 and S16)

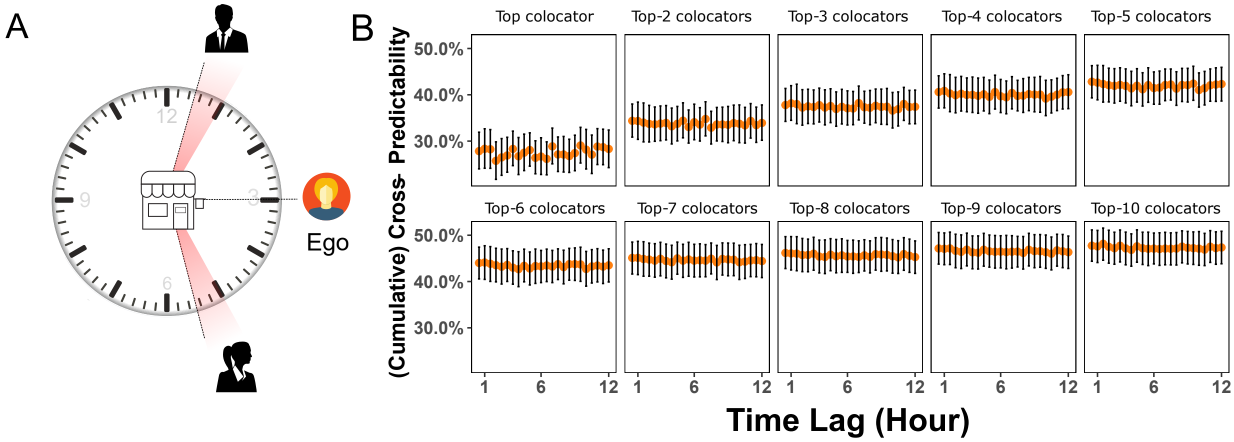

FIG. 6. **Temporal stability of non-social colocator information**. **A** An example of two time-displaced colocators who visit the same location as the ego on a $T = 2.5$ hour time lag. **B** The (cumulative) cross-predictability influence of temporal-lag for non-social colocator(s) in Weeplaces. Each point corresponds to a co-location network resulting from the amount of temporal offset between an ego and alters visit to a common location. Error bars denote 95% CI.

reveal the same trends.

The observed connection between information transfer and location overlap, behooves one to to ask whether temporal effects are a key factor. In other words, our choice of colocation is based on the simple idea that individuals in the same place at the same time contain information about mobility patterns of each other (see also Sec. S2.5). We can relax this condition and also consider individuals that visit the same locations as the ego, but displaced in time. For example, residents of a neighborhood can stop at their local corner store at different times of day and never run into each other, but their visits are always five hours apart because of their respective work schedules. We can investigate this by creating networks of time-displaced colocators, where now an ego and alter colocate if an alter visits a location in the time windows $[T, T - 1/2]$ hours prior or $[T - 1/2, T]$ hours following an ego visitation at the same location (see the illustration in Fig. 6**A**). Note that $T = 1/2$ h yields the fully connected window $[-1/2, 1/2]$.

We examine the trend in $\Pi_{A|B}$ as a function of the temporal-lag $T$ in Fig. 5**B**. Each network resulting from the different temporal lags will generally have different ego-alter pairs, and the set of egos with at least 10 alters may change. We consider then the common alters who have at least 10 alters for all the networks between 1/2 h and 12 h with 30-minute intervals. The ego-alter overlap between these networks shown in Figure S17 indicates that the networks are quite different, meaning co-located alters are not necessarily the same as time-displaced co-located alters. As before, there is an increase in predictability as alters are added, yet for a given set of alters, remarkably there is little-to-no difference between any of the networks in terms of their information content, at least within the investigated temporal ranges. Furthermore, this suggests potentially unexpected sources of mobility information, as the non-social co-locators selected here do not necessarily have to be visitors of the same location at the same time to provide predictive mobility information about the ego.

## 4. DISCUSSION

Using information-theoretic measures, we analyzed the spatiotemporal structure of the mobility trajectories of a set of users in three publicly available location-based social networks. Entropy measures were used to quantify the sequential information contained in a user's physical trajectory which revealed differences in our datasets based on the context of how users used the apps. Using these measures, we then compared the information present in the mobility patterns of an individual's (the ego's) social ties compared with non-social colocators, other users who frequently visited the same locations as the ego. Across datasets we found the importance of social ties: consistently more information about the ego's future location was present in the past locations of the social ties than in the past locations of the non-social ties, and this held when aggregating information from multiple number of alters. Interestingly however, this implies something important: that groups of many non-social colocators can in principle provide as much information as a smaller set of social ties, meaning that non-social sources of mobility information are in principle available. If access to social data are limited, these non-social data may, in the aggregate, be used as a replacement. A future study on the mobility information of an individual carried by non-social colocators should consider the possibility that social-economics or demographics could play a role. For instance, people working in the same building but for different companies are likely to share similar social-economic status, likewise, parents taking their children to schools may share similar household responsibilities. The extent in which these shared factors affect the predictability of a ego mobility can be investigated in richer datasets.

Our study relied on observational data taken from three location-based social networks.

This introduces crucial caveats. In particular, the social ties reported in the datasets are incomplete reflections of a person's full social circle, and the nature of such ties may differ in the online and offline domains. Likewise, not all locations visited by an individual are recorded in these social networks, which rely primarily on user check-ins, so we expect mobility trajectories to be under-sampled as well. Followup work, including richer, more detailed data and even experimental studies, are needed to address these concerns, yet our robustness checks, including observing consistent trends across datasets and across our sampling criteria, already provide evidence that our results appear to be robust.

The presence of predictive information, both socially and otherwise, has crucial implications. Privacy protections regarding social data are important to protect sensitive information about a user and their social ties. Social information flow suggests that an individual's future movements can be predicted by studying the mobility patterns of a few acquaintances. On the other hand, our study also demonstrates that social ties are not the only source of predictive mobility information, and measures of colocation are enough to uncover novel sources of mobility information. This means that locations monitoring individual visits, for example, a grocery store tracking the smartphones of shoppers [32], may in principle be collecting the building blocks of mobility profiles, and individuals providing access to their mobility data may also be providing information about both social and non-social ties [33–35]. While these data can inform important applications such as contact tracing in the early stages of a disease outbreak, significant ethical concerns surrounding such information sources make it critical to place strong access constraints on mobility information. Indeed, the results presented here provide further impetus to the ongoing debate on best practices for privacy protection, both in terms of legislation and ethical algorithmic development.

[1] Barbosa, H. *et al.* Human mobility: Models and applications. *Physics Reports* **734**, 1–74 (2018).

[2] Batty, M. *The new science of cities* (MIT press, 2013).

[3] Simini, F., Gonzalez, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484**, 96–100 (2012).

[4] Uherek, E. *et al.* Transport impacts on atmosphere and climate: Land transport. *Atmospheric*

*Environment* **44**, 4772–4816 (2010).

[5] Lee, M., Barbosa, H., Youn, H., Holme, P. & Ghoshal, G. Morphology of travel routes and the organization of cities. *Nature Communications* **8** (2017). 1701.02973.

[6] Kirkley, A., Barbosa, H., Barthelemy, M. & Ghoshal, G. From the betweenness centrality in street networks to structural invariants in random planar graphs. *Nature Communications* **9**, 2501 (2018).

[7] Pan, W., Ghoshal, G., Krumme, C., Cebrian, M. & Pentland, A. S. Urban characteristics attributable to density-driven tie formation. *Nature Communications* **4**, 1961 (2013).

[8] Tizzoni, M. *et al.* Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. *BMC Medicine* **10**, 165 (2012).

[9] Toole, J. L. *et al.* The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies* **58, Part B**, 162–177 (2015).

[10] Vázquez, A. *et al.* Modeling bursts and heavy tails in human dynamics. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* **73**, 1–19 (2006). 0510117.

[11] Brockmann, D., Hufnagel, L. & Geisel, T. The scaling laws of human travel. *Nature* **439**, 462–465 (2006). 0605511.

[12] Song, C., Koren, T., Wang, P. & Barabási, A. L. Modelling the scaling properties of human mobility-supp. *Nature Physics* **6**, 818–823 (2010). 1010.0436.

[13] Rhee, I. *et al.* On the levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking* **19**, 630–643 (2011).

[14] Boyer, D., Crofoot, M. C. & Walsh, P. D. Non-random walks in monkeys and humans. *Journal of the Royal Society Interface* **9**, 842–847 (2012). 1110.0763.

[15] Hasan, S., Schneider, C. M., Ukkusuri, S. V. & González, M. C. Spatiotemporal Patterns of Urban Human Mobility. *Journal of Statistical Physics* **151**, 304–318 (2013).

[16] Song, C., Qu, Z., Blumm, N. & Barabasi, A.-L. Limits of Predictability in Human Mobility. *Science* **327**, 1018–1021 (2010).

[17] Ikanovic, E. L. & Mollgaard, A. An alternative approach to the limits of predictability in human mobility. *EPJ Data Science* **6**, 12 (2017).

[18] Cho, E., Myers, S. A. & Leskovec, J. Friendship and mobility: User movement in location-based social networks. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, KDD '11, 1082–1090 (Association for Computing Machinery, New York, NY,

USA, 2011).

[19] Bagrow, J. P., Liu, X. & Mitchell, L. Information flow reveals prediction limits in online social activity. *Nature Human Behaviour* **3**, 122–128 (2019).

[20] Hazarie, S., Barbosa, H., Frank, A., Menezes, R. & Ghoshal, G. Uncovering the differences and similarities between physical and virtual mobility. *Journal of the Royal Society Interface* **17**, 20200250 (2020).

[21] Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).

[22] Wu, F. *et al.* A new coronavirus associated with human respiratory disease in china. *Nature* **579**, 265–269 (2020).

[23] Flaxman, S. *et al.* Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261 (2020).

[24] Davies, N. G. *et al.* Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study. *The Lancet Public Health* **5**, e375–e385 (2020).

[25] Fahey, R. A. & Hino, A. COVID-19, digital privacy, and the social limits on data-focused public health responses. *International journal of information management* **55**, 102181–102181 (2020).

[26] Bengio, Y. *et al.* The need for privacy with public digital contact tracing during the COVID-19 pandemic. *The Lancet Digital Health* **2**, e342–e344 (2020).

[27] Grabowicz, P., Ramasco, J., Gonçalves, B. & Eguíluz, V. Entangling mobility and interactions in social media. *PloS one* 1–16 (2014).

[28] Lempel, A. & Ziv, J. On the Complexity of Finite Sequences. *IEEE Transactions on Information Theory* **22**, 75–81 (1976).

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

[30] Cover, T. & Thomas, J. A. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing (Wiley-Interscience, New York, New York, USA, 2006).

[31] Ziv, J. & Merhav, N. A measure of relative entropy between individual sequences with application to universal classification. *IEEE Transactions on Information Theory* **39**, 1270–1279

(1993).

[32] Enck, W. *et al.* Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones. *ACM Transactions on Computer Systems (TOCS)* **32**, 1–29 (2014).

[33] Horvát, E.-A., Hanselmann, M., Hamprecht, F. A. & Zweig, K. A. One plus one makes three (for social networks). *PLOS ONE* **7**, 1–8 (2012).

[34] Sarigol, E., Garcia, D. & Schweitzer, F. Online privacy as a collective phenomenon. In *Proceedings of the second ACM conference on Online social networks*, 95–106 (2014).

[35] Garcia, D. Leaking privacy and shadow profiles in online social networks. *Science advances* **3**, e1701172 (2017).