Information flow reveals prediction limits in online social activity

James P. Bagrow[1,2,*], Xipei Liu[1,2], and Lewis Mitchell[1,2,3,†]

[1]Department of Mathematics & Statistics, University of Vermont, Burlington, VT, United States
[2]Vermont Complex Systems Center, University of Vermont, Burlington, VT, United States
[3]School of Mathematical Sciences, North Terrace Campus, The University of Adelaide, SA 5005, Australia
[*]E-mail: james.bagrow@uvm.edu
[†]E-mail: lewis.mitchell@adelaide.edu.au

January 21, 2019

Abstract

Modern society depends on the flow of information over online social networks, and users of popular platforms generate significant behavioral data about themselves and their social ties [1, 2, 3, 4, 5]. However, it remains unclear what fundamental limits exist when using these data to predict the activities and interests of individuals, and to what accuracy such predictions can be made using an individual's social ties. Here we show that 95% of the potential predictive accuracy for an individual is achievable using their social ties only, without requiring that individual's data. We use information theoretic tools to estimate the predictive information within the writings of Twitter users, providing an upper bound on the available predictive information that holds for any predictive or machine learning methods. As few as 8-9 of an individual's contacts are sufficient to obtain predictability comparable to that of the individual alone. Distinct temporal and social effects are visible by measuring information flow along social ties, allowing us to better study the dynamics of online activity. Our results have distinct privacy implications: information is so strongly embedded in a social network that in principle one can profile an individual from their available social ties even when the individual forgoes the platform completely.

The flow of information in online social platforms is now a significant factor in protest movements, national elections, and rumor and misinformation campaigns [6, 7, 8]. The study of social contagion [9], for example, is predicated on the flow of information over social ties, and has benefited greatly from the availability of massive online social datasets and platforms on which to perform observational and experimental studies [10, 11]. Data collected from online social platforms are a boon for researchers [2] but also a source of concern for privacy, as the social flow of predictive information can reveal details on both users and non-users of the platform [5, 12, 13]. Measuring information flow is challenging, in part due to the complexity of natural language and in part due to the difficulty in defining a quantitative and objective measure of information. Owing to these challenges, proxies are often studied instead: *structural* proxies focus on network characteristics such as the movements of keywords [4, 7, 14, 15] or adoptions of behaviors [16, 17, 18]. *Temporal* proxies attempt to quantify the information contained in the timings of user activity, as temporal relationships between user activity are known to reflect underlying coordination patterns [19, 20].

However, neither of the above approaches consider the full extent of information available: both the complete language data provided by individuals and their temporal activity patterns. Although, for example, temporal proxies are necessary in social networks where time series data are available but message content is not, for privacy or other reasons (for example, in mobile phone datasets), public postings to online social platforms present a unique opportunity to explore the textual content of messages in conjunction with their timings, giving a richer understanding of social ties.

Information theory allows us to mathematically quantify the information contained within data, and is well suited to data in the form of online written communication. Although the mathematical definition of information is somewhat distinct from our commonly held notions of information and meaning, or semantics, information-theoretic measures are crucial for understanding how algorithms can learn from data. Nowadays, with such large volumes of data generated by online social platforms, both researchers and platform providers are often forced to interact with a platform's data only computationally, using algorithms to quantify and make inferences about users, and the accuracy of these inferences is predicated on the mathematical information contained within a user's data.

In this work, we apply information-theoretic estimators to study information and information flow within a collection of Twitter user activities. These estimators fully incorporate language data while also accounting for the temporal ordering of user activities. We find that meaningful predictive information about individuals is encoded in their social ties, allowing us to determine fundamental limits of social predictability, independent of actual predictive or machine learning methods. We explore the roles of information recency and social activity patterns, as well as structural network properties such as information homophily between individuals.

We gathered a dataset of $N = 13,905$ users, comprising egocentric networks from the Twitter social media platform, and a total of $m = 30,852,700$ public postings from these users. Each of the $n = 927$ ego-networks consisted of one user (the ego) and their 15 most frequently mentioned Twitter contacts (the alters), providing us with ego-alter pairs on which to measure information flow. See 'Data collection and filtering' in the Methods section for full details on the data processing.

The ability of a machine learning method to accurately profile individuals from their online traces is reflected in the predictability of their written text. Indeed, with a language model trained to predict the words a user will post online, in principle, one can construct a profile of the user by evaluating the likelihoods of various words to be posted, such as terms related to politics. Thus, quantifying the predictive information contained within a user's text allows us to understand the potential accuracy such methods can potentially achieve given a user's data.

A text's predictive information can be characterized by three related quantities, the entropy rate h, the perplexity 2^h, and the predictability Π. The entropy rate quantifies the average uncertainty one has about future words given the text one has already observed (Fig. 1a). Higher entropies correspond to less predictable text and reflect individuals whose interests are more difficult to predict. In the context of language models, it is also common to consider the perplexity. Whereas the entropy rate specifies how many bits h are needed on average to express subsequent unseen words given the preceding text, the perplexity tells us that our remaining uncertainty about those unseen words is equivalent to that of choosing uniformly at random from among 2^h possibilities. For example, if $h = 6$ bits (typical of individuals in our dataset), the perplexity is 64 words, which is a significant reduction from choosing randomly over the entire vocabulary (social media users have ≈5000-word vocabularies on average; see Supplementary Note 1.3 for full distributions). Finally, the predictability Π, given via Fano's inequality [21], is the probability that an *ideal* predictive algorithm will correctly predict the subsequent word given the preceding text. Repeated, accurate predictions of future words indicate that the available information can be used to build profiles and predictive models of a user's writing (see also below for subsequent discussion), and estimating Π allows us to fundamentally bound the usefulness of the information present in a user's writing without depending on the results of specific predictive algorithms. We emphasize that the information-theoretic *predictability* as defined here is distinct from *prediction*, in that it does not actually make predictions about future text. Instead, this predictability provides a method-independent upper bound on prediction accuracy.

Information theory has a long history of estimating the mathematical information content of text [22, 23, 24, 25]. Crucially, information is present not just in the words of the text but also in their order of appearance. Thus, we applied a nonparametric entropy estimator that incorporates the full sequence structure of the text [25]. This estimator has been proved to converge asymptotically to the true entropy rate for stationary processes and has been applied to human mobility data [26]. See 'Measuring the flow of predictive information' and 'Estimator convergence on our data' in the Methods section for further details on the entropy estimators and their convergence rates on these data.

We focus on four aspects of information flow over social networks, exploring both content and timing of messages: (i) the extent to which information is encoded through language into an individual's social ties, (ii) the importance of recency to information flow between individuals, (iii) the role of tie strength between individuals in the flow of information, and (iv) the relationship between structural network properties such as homophily and information. We first examined the information content of the egos themselves. Their text streams were relatively well clustered around $h \approx 6.6$ bits, with most falling between 5.5–8 bits (Fig. 1b). Equivalently, this corresponds to a perplexity range of

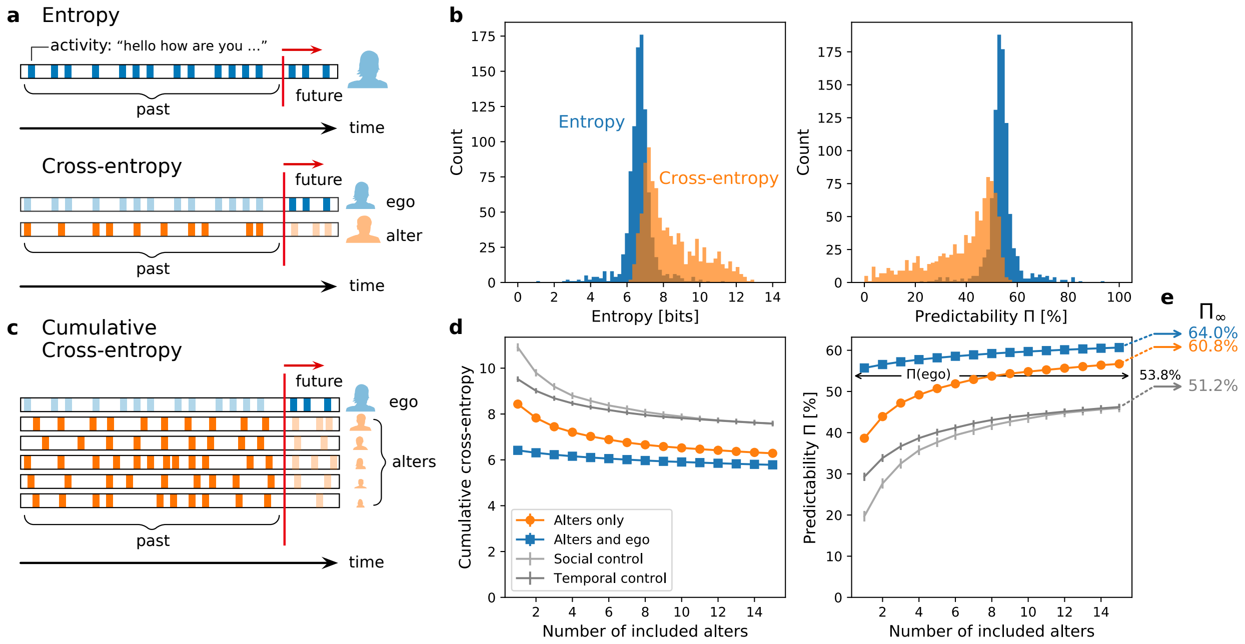

Figure 1: Information and predictability in online social activity. **a**, A user posts written text over time and we would predict their subsequent words given their past writing. Treating each user's posts as a contiguous text stream, the entropy rate tells us how uncertain we are about a user's future writing given their past. To study information flow, the cross-entropy rate tells us how much information about the future text of one user (the ego, blue) is present in the past text of another user (the alter, orange). **b**, Most users have entropies and predictabilities (blue) in a well-defined range, whereas the cross-entropies and associated predictabilities (orange) indicate a broad variety of social information flow levels. **c**, Predictive information may be available in the pasts of multiple alters, so we computed the cumulative cross-entropy as we included more alters in order of most to least frequently contacted. **d**, As the past activities of more alters are used to predict the ego, more information is available and the entropy drops and predictability rises (orange). Including the ego's past with the alters (blue) shows that the alters provided non-redundant predictive information. **e**, Extrapolating beyond our data window estimates the prediction limit Π_∞ of online activity. Error bars denote mean ± 95% CI.

≈45–256 words, far smaller than the typical user's ≈5000-word vocabulary, and a mean predictability of ≈53%, quite high for predicting a given word out of ≈5000 possible words on average (for example, choosing words uniformly at random corresponds to a predictability of 0.02%). We found this typical value of information comparable to other sources of written text, but social media texts were more broadly distributed around the mean—individuals were more likely to be either highly predictable or highly unpredictable compared with formally written text (see Supplementary Note 1.4).

Next, instead of using the entropy rate to ask how much information is present in what the ego has previously written regarding what the ego will write in the future, we now ask how much information is present on average in what the ***alter*** has previously written regarding what the ego will write in the future (Fig. 1a). If there is consistent, predictive information in the alter's past about the ego's future, especially beyond the information available in the ego's own past, then there is evidence of predictive information flow.

Replacing the ego's past writing with the alter's past converts the entropy to the *cross-entropy* (see 'Measuring the flow of predictive information' and 'Estimator convergence on our data' in the Methods section). The cross-entropy is always greater than the entropy when the alter provides less information about the ego than the ego, and so an increase in cross-entropy tells us how much information we lose by only having access to the alter's information instead of the ego's. Indeed, estimating the cross-entropy between each ego and their most frequently contacted alter (Fig. 1b), we saw higher cross-entropies than using the ego's own text, spanning from 6–12 bits compared with 5–9 bits (equivalently, perplexities from 64–4096 words compared with 32–512 words, or predictabilities spread from 0–60% compared with 40–70%). Whereas less frequently contacted alters provided less predictive information than alters in close contact (see Supplementary Notes 1.6 and 1.7), even for the closest alters there was a broader range of cross-entropies than the entropies of the egos themselves. This implies a diversity of social relationships: sometimes the ego is well informed by the alter, leading to a cross-entropy closer to the ego's entropy, whereas other times the ego and alter exhibit little information flow.

Thus far we have examined the information flow between the ego and individual alters, but actionable information regarding the future of the ego may be embedded in the combined pasts of multiple alters (Fig. 1c). To address this, we generalized the cross-entropy estimator to multiple text streams (see 'Measuring the flow of predictive information' and 'Estimator convergence on our data' in the Methods section). We then computed the cross-entropies and predictabilities as we successively accumulated alters in order of decreasing contact volume (Fig. 1d). As more alters were considered, cross-entropy decreased and predictability increased (Spearman's $\rho = -0.505$ 95% CI [-0.517, -0.492], $p < 0.001$ and $\rho = 0.527$ [0.515, 0.540], $p < 0.001$, respectively), which is sensible as more potential information is available. Interestingly, with 8–9 alters, we observed a predictability of the ego given the alters at or above the original predictability of the ego alone—with 10 alters, the predictability was significantly greater than that of the ego alone (two-tailed test, $t(1852) = -3.32$, $p < 0.001$). As more alters were added, up to our data limit of 15 alters, this increase continued. Paradoxically, this indicated that there is potentially more information about the ego within the total set of alters than within the ego itself.

To understand this apparent paradox, we need to address a limitation with the above analysis: it does not incorporate the ego's own past information. It may be that the information provided by the alters is simply redundant when compared to that of the ego. To see whether this is the case, we simply included the ego's past alongside the alters, generalizing the estimator to an entropy akin to a transfer entropy [27,28], a common approach to studying information flow. This entropy is computed in the 'Alters and ego' curves in Fig. 1d. A single alter provided a small amount of extra information beyond that of the ego, 1.9% more predictability. This value provided us a quantitative measure of the extent of information flow between individual users of social media. Beyond the most frequently contacted alter, as more alters were added, this extra predictability grew: at 15 alters and the ego there was 6.9% more predictability than via the ego alone. Furthermore, the information provided by the alters without the ego is strictly less than the information provided by the ego and alters together, resolving the apparent paradox.

However, this extra predictability also appeared to saturate, and if so then eventually adding more alters will not provide extra information (see Supplementary Note 1.2). This observation is compatible with *Dunbar's number* which uses cognitive limits to argue for an upper bound on the number of meaningful ties that an ego can maintain (≈150 alters for humans) [29]. Thus, the question becomes: given enough ties, what is the upper bound for predictability?

To extrapolate beyond our data window, we fitted a nonlinear saturating function to the curves in Fig. 1d, (see Supplementary Note 1.2 for details and validation of our extrapolation procedure). From fits to the raw data extrapolated to infinity, we found a limiting predictability given the alters of $\Pi_\infty = 60.8\% \pm 0.691\%$ (95% CIs) (Fig. 1e). Of course, egos will not have an infinite number of alters, so a more plausible extrapolation point may be to Dunbar's number: $\Pi_{150} = 60.3\%$, within the margin of error for Π_∞, indicating that saturation of predictive information has been reached. Similarly, extrapolating the predictability including the ego's past gives $\Pi_\infty = 64.0\% \pm 1.54\%$ ($\Pi_{150} = 63.5\%$).

These extrapolations showed that significant predictive information was available in the combined social ties of individual users of social media. In fact, there is so much social information that an entity with access to all social media data will have only slightly more potential predictive accuracy (≈64% in our case) than an entity that has access to the activities of an ego's alters but not to those of that ego (≈61%). This may have distinct implications for privacy: if an individual forgoes using a social media platform or deletes their account, yet their social ties remain, then, potentially, that platform owner still possesses $95.1\% \pm 3.36\%$ of the achievable predictive accuracy of the future activities of that individual.

Two issues can affect the cross-entropy as a measure of information flow. The first is that the predictive information may be due simply to the structure of English: commonly repeated words and phrases will represent a portion of the information flow. The second is that of a common cause: egos and alters may be independently discussing the same concepts. This is particularly important on social media with its emphasis on current events [3].

To study these issues, we constructed two types of controls. The first randomly pairs users together by shuffling alters between egos. The second constructed pseudo-alters by assembling, for each real alter, a random set of posts made at approximately the same times as the real alter's posts, thus controlling for temporal confounds. See 'Control procedures' in the Methods section for more information. Both controls used real posted text and only varied the sources of the text. As shown in Fig. 1d, the real alters provided more social information than either control. Although there was a decrease in entropy as more control alters were added, the control cross-entropy remained above the real cross-entropy (two-tailed test, $t(23293) = -103.8$, $p < 0.001$) and the control predictability remained below the real predictability ($t(21103) = 119.0$, $p < 0.001$). We also observed that, for a single alter, the temporal control had a lower cross-entropy than the social control ($t(23293) = -117.5$, $p < 0.001$) and therefore temporal effects explain more information than social effects (underscoring the role of social media as a news platform [3]), although both controls eventually converge to a limiting predictability of 51.2%. This demonstrates that useful predictive information is encoded in real social ties, beyond that expected from the structure of language alone.

Given the importance of temporal information in online activity, to what extent is this reflected in the information flow? Do recent activities contain most of the predictive information or are there long-term sources of information? To estimate recency effects, we applied a censoring filter to the ego's text stream, removing at each time period the text written in the previous ΔT hours and measuring how much the mean predictability decreased compared with the mean predictability including the recent text. Increasing ΔT decreased Π, especially evident when removing the first 3–4 h worth of text (these intervals correspond to 6.2–7.8 tweets ignored per word on average; Fig. 2a): we found an average decrease in predictability of 1.4% at 4 h. This 1.4% loss in predictability relative to the uncensored baseline is comparable to the 1.9% gain from the rank-1 alter that we observed in Fig. 1d. In other words, close alters tended to contain a quantity of information about the ego comparable to the information within just a few hours of the ego's own recent past. Beyond 24 h the predictability loss continued approximately linearly (visually; see Supplementary Note 1.5 and Supplementary Fig. 8). We also applied this censoring procedure to the alters alone and the alters combined with the ego, excluding their recent text and measuring how the cross predictability changed on average from their respective baselines. We found a similar drop in predictability during the first few hours (0.8% and 1.3% in the first 4 h given alters and alters plus ego, respectively), but then a more level trend than when censoring the ego alone (a further decrease of 0.1% and 0.3%, respectively, between 4 and 24 h, compared with 0.4% for the ego alone over the same interval). This leveling off showed that less long-term information was present in the alters' pasts than within the ego's.

Next, we studied recency by the activity frequencies of alters and egos. Individuals who post frequently to social media, keeping up on current events, may provide more predictive information about either themselves or their social ties than other, infrequent posters. We found that the self-predictability of users was actually independent of activity frequency (Supplementary Note 1.4), but there were strong associations between activity frequency and social information flow: egos who posted 8 times per day on average were 17% ± 14.9% (95% CI) more predictable given their alters than egos who posted once per day on average (Fig. 2b). Interestingly, this trend reversed itself when considering the activity frequencies of the alters: alters who posted 8 times per day on average were 23% ± 4.46% less predictive of their egos than alters who posted once per day on average. Both trends in Fig. 2b were significant (Spearman's $\rho = 0.276$ [0.216, 0.335], and $\rho = -0.437$ [-0.487, -0.383], respectively, $p < 0.001$; see Supplementary Note 1.6).

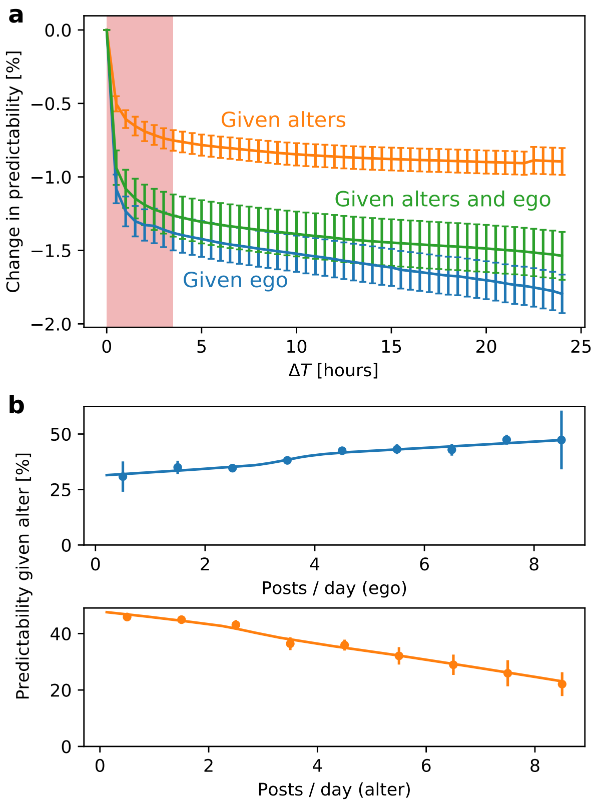

Figure 2: Recency of information. a, Removing the most recent ΔT hours of activity, most predictive information about the ego is contained in the most recent 3–4 h (shaded region, 1.4% drop). In all cases, information extends backwards beyond these time intervals, but the ego (blue) contains more long-range past information (0.6% more predictability) than the combined alters alone (orange, alters and ego green). **b**, Egos who post more frequently are 17% ± 14.9% more predictable from their alter than egos who post less frequently, whereas frequently posting alters provide 23% ± 4.46% less information about their egos than alters who post less often. Lines in panel b denote a LOWESS fit. Error bars denote mean ± 95% CI.

Highly active alters tended to inhibit information flow, perhaps due to covering too many topics of low relevance to the ego.

Information flow reflects the social network and social interaction patterns (Fig. 3). We measured information flow for egos with more popular alters compared with egos with less popular alters. Alters with more social ties provided less predictive information about their egos than alters with fewer ties (Fig. 3a). This trend was significant (Spearman's $\rho = -0.199$ [-0.224, -0.175], $p < 0.001$; see Supplementary Note 1.9). Qualitatively, the decrease in predictability of the ego was especially strong up to alters with ~400 ties, where the bulk of our data lies, but the trend continued beyond this as well. This decreasing trend belies the power of hubs in many ways: although hubs strongly connect a social network topologically [30], limited time and divided attention across their many social ties bound the hub alter's ability to participate in information dynamics mediated by the social network and this is reflected in the predictability.

Reciprocated contact is an important indicator of social relationships [31], especially in online social activity

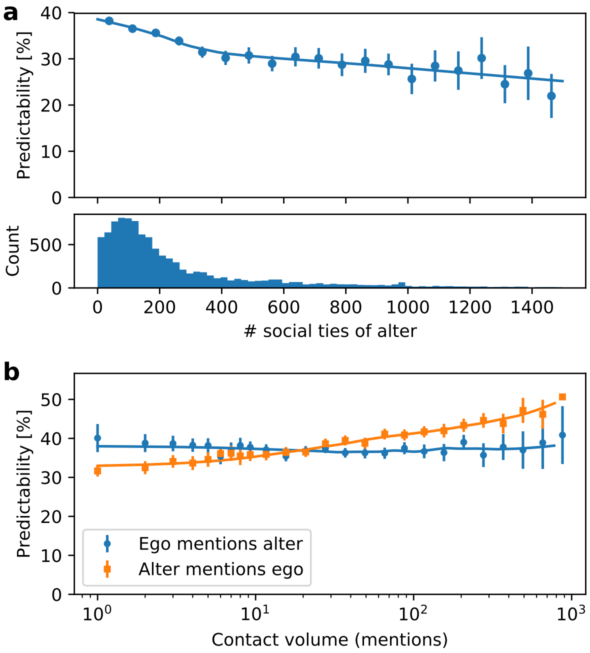

Figure 3: Social interactions are visible in information flow. a, Alters with more social ties of their own provided less information about the ego than less popular alters (Spearman's $\rho = -0.199$ [-0.224, -0.175], $p < 0.001$). **b**, Information flow captures directionality in relationships, which is a key factor in social dynamics. Alters who often contact the ego provide more predictive information about the ego than alters who rarely mention the ego (Spearman's $\rho = 0.226$ [0.202, 0.250], $p < 0.001$). Yet, if the ego frequently mentions the alter, it does not necessarily mean that the alter will provide more predictive information about the ego (Spearman's $\rho = -0.0185$ [-0.0440, 0.00704], $p = 0.156$). Lines denote a LOWESS fit. Error bars denote mean ± 95% CI.

where so much communication is potentially one-sided [3]. In Fig. 3b, we investigated how directionality in contact volume, how often the ego mentions the alter and vice versa, related to information flow. We found that the ego was more predictable given the alter for those dyads in which the alter more frequently contacted the ego (Spearman's $\rho = 0.226$ [0.202, 0.250], $p < 0.001$; see Supplementary Note 1.9), but there was little change across dyads when the ego mentioned the alter more or less frequently (Spearman's $\rho = -0.0185$ [-0.0440, 0.00704], $p = 0.156$; see Supplementary Note 1.9). We also observed a similar trend for information flow but in reverse, when predicting the alter given the ego (see Supplementary Note 1.9. These trends captured the reciprocity of information flow: an alter frequently contacting an ego will tend to give predictive information about the ego, but the converse is not true: an ego can frequently contact her alter but that does not necessarily mean that the alter will be any more predictive, as evidenced by the relatively flat trend in Fig. 3b.

Finally, comparing the entropy of an ego with the entropy of their alters reveals a strong homophily effect in terms of their (self) information (Fig. 4). The entropy rates of the ego and alter on a given dyad were correlated (Fig. 4a). Figure 4a covers the correlation between the ego and the rank-1 alter (Spearman's $\rho = 0.440$ [0.386, 0.490],

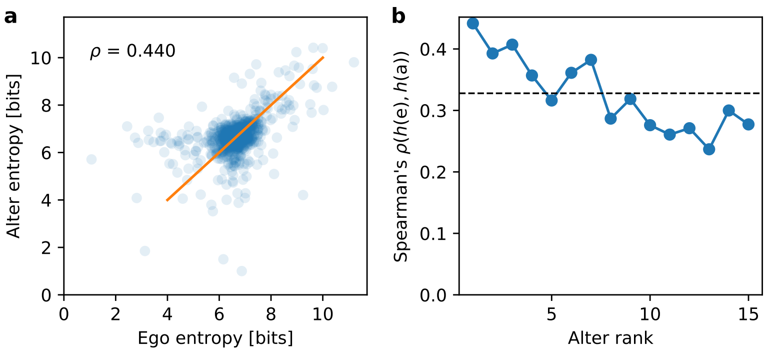

Figure 4: An 'information homophily' between egos and alters. The entropies of egos and their alters are strongly correlated (Spearman's $\rho = 0.440$ [0.386, 0.490, $p < 0.001$), indicating a ***homophily effect***. **a**, The entropy rate h of egos compared to the entropy rate of their rank-1 (most frequently contacted) alter. The straight line $y = x$ provides a guide for the eye. **b**, The Spearman's ρ between ego entropies and alter entropies as a function of alter rank; all plotted ρ are significant ($p < 0.001$). Correlation generally decreases with alter rank. The dashed line indicates ρ across all ranks.

$p < 0.001$). In Fig. 4b, we plot the Spearman's ρ between $\hat{h}$(ego) and $\hat{h}$(alter) as a function of alter rank. These correlations were significant for all ranks ($p < 0.001$). The correlation drops consistently over the first five or so alters, implying that the homophily effect is connected with contact volume. Interestingly, we see weaker associations in cross-entropy (see Supplementary Note 1.8); further investigation of these and other information homophily effects has the potential to improve our ability to control for homophily in order to explore social contagion.

The ability to repeatedly and accurately predict the text of individuals delivers considerable value to the providers of social media, allowing providers to develop profiles to identify and track individuals [32, 33] and even manipulate information exposure [34]. For example, a language model may be trained on available data to generate new text in the "voice" of a user [35, 36, 37] and with such a language model one could derive a profile for the user by querying it for the likelihood that the user will make certain kinds of statements (for example, how likely are certain statements about one political party or another). Language models derived in this way can have important consequences: combining predictions from a language model with an algorithm for recommending new social ties, for example, has the potential to create or exacerbate filter bubbles [34]. The optimal accuracy a trained model can achieve when making these text predictions is mathematically bounded by the predictive information that we estimate here. That information is so strongly embedded socially underscores the power of the social network: by knowing who the social ties of an individual are and what the activities of those ties are, our results show that one can, in principle, accurately profile even those individuals who are not present in the data [5].

Experimental studies are crucial for improving on our results. For example, we have shown how a platform provider can use information from a user's social ties as a substitute for missing information from that user. Yet, in reality, this substituted predictive information can become outdated as the social system and its members are not static entities but evolve over time. This evolution challenges prediction, as a user forgoing or deleting their account can change a social tie's future behavior, even if only through the fact of no longer interacting with that user, affecting the future predictability of that user. Experiments can help understand the effects of this evolution on prediction. Likewise, any research involving observational data, such as ours, will have difficulty distinguishing social contagion from homophily [16, 38]. Our focus on predictive information flow is one of the strongest measures possible given the text data that we study, in that we explicitly utilize the time ordering of information when calculating cross-entropy. However, information flow alone is not sufficient evidence of contagion. To establish contagion would require controlling for the tendency of similar alters to share ties, which the present study provides a first step towards. Of course, the gold standard for causal influence remains experimental interventions. Such experimental studies are ideal for studying both dynamic social effects and contagion phenomena.

The time-ordered cross-entropy (Fig. 1a) applied here to online social activity is a natural, principled information-theoretic measure that incorporates all of the available textual and temporal information. Although weaker than full causal entailment, by incorporating time ordering, we identify social information flow as the presence of useful, predictive information in the past of one's social tie beyond that of the information in one's own past. Doing so closely connects this measure with Granger causality and other strong approaches to information flow [27, 39].

Methods

Data collection and filtering We selected a random sample of individuals for study from the Twitter 10% Gardenhose feed collected during the first week of April 2014. From this, we uniformly sampled individuals who had tweeted in English (as reported by Twitter in the metadata for each tweet) during this time period and had 50–500 followers, as reported in the feed metadata. The lower follower cutoff is to avoid inactive and bot accounts, whereas the higher cutoff is to ensure that individuals in our sample have comparably sized ego-networks and to avoid studying unusually popular outlier accounts, such as celebrity accounts. We remark that generating a sample from the 10% feed necessarily introduces a small bias towards more active individuals, those who have tweeted at least once into that feed. For each user, we then collected their complete public tweet history excluding retweets (up to 3200 most recent public messages, as allowed by Twitter's Public REST API limit [40]). As discussed later in this section, we then applied to these users a filtering procedure, including both computational tools and human raters to help ensure sufficient data on individual activities and to limit bots and non-individual accounts from our sample. When finished, we retained a final sample of $n = 927$ individual egos and their top-15 alters ($n = 13,905$ total users). For each initially sampled ego, we collected the user IDs of the account whom the ego 'at-mentioned' most frequently in their public tweets, forming the rank-1 alter. Using such mentions is of course not the only way to define a social tie on Twitter; follower relationships, numbers of retweets, or shared textual features (such as hashtags or keywords) could all be reasonably employed to define a social network. However, defining social ties using mentions gives a stronger signal than simply Twitter following, as it demonstrates active communication on the behalf of at least one of the individuals of a social tie. Defining social ties as related to the number of mentions also captures a degree of social closeness, whereas follower or following has no strength associated with it. As was done with the egos, the REST API was then used to retrieve the complete public tweet history of this alter. Examining the messages of the (ego, rank-1 alter) dyad, we retained egos where the ego's tweets covered at least a 1-year period, the alter's tweets covered at least a 1-year period, and the ego at-mentioned at least 15 unique Twitter users (including the rank-1 alter). For dyads who satisfied these criteria, we collected the full public messages of the remaining 14 most at-mentioned alters, giving us the full public activities of the ego and their top-15 most mentioned alters.

To limit the effects of bots and non-personal accounts, we moved beyond the basic filtering criteria listed above and employed both computational tools and human raters to examine the accounts of the egos in our dataset. These tools were applied in April

2017. A small number of accounts in our sample were suspended or deleted after our data collection period and were not available online to be examined, so we simply retained these unrated accounts in our sample. We used the botometerAPI [41, 42, 43, 44, 45] to score the probability that an ego account was a bot, and eliminated $n = 46$ accounts that scored above 50%. This tool examines Twitter accounts along several dimensions to estimate the likelihood that the account belongs to a bot. Next, we asked human raters to examine the accounts and report whether the account appeared to belong to a real person or a non-personal entity, such as a corporation or a bot. Two independent raters examined each account's Twitter homepage if available. We removed $n = 84$ accounts for which both raters agreed that the account did not belong to an individual, beyond those already flagged by the botometer scores. Raters were recruited on Amazon Mechanical Turk and compensated at a rate of US$0.10 per three Twitter accounts. Finally, we also removed a small number of accounts ($n = 31$) showing convergence issues with our entropy estimators, as inferred by negative Kullback–Leibler divergences from the ego to the alter or vice versa. This gave our final sample size of $n = 927$ egos and their top-15 associated alters, comprising $m = 30,852,700$ total tweets.

Control procedures We performed two controls for the cross-entropy experiments: random tweets or 'temporal control' and random alters or 'social control'. For the temporal control, we constructed proxy tweet streams for the alters that preserved the approximate times at which alters had written messages. To do this, we substituted for each real alter tweet a randomly sampled English-language tweet posted during the same hour as the real alter tweet. The randomly sampled replacement tweets were taken from the 10% Gardenhose feed. In the social control, we randomized the ego networks, swapping the tweet text streams of true alters with those of randomly chosen alters. This control does not preserve the times at which the original alters had authored tweets, hence the use of the previous temporal control.

Text processing To apply the entropy estimators discussed below, we first need to process and tokenize the texts of users. The UTF-8 encoded text of each user was processed by removing casing, punctuation (except for twitter specific "@" and "#" symbols), and URLs (identified as words beginning with "http://" or "https://"). All tweet texts were concatenated into a single text string in time order (based on the tweet timestamps), except for "retweets" which were all excluded in order to focus on the effect of shared language and avoid artificially inflating predictability scores. The text was then tokenized into words by segmenting on whitespace.

Measuring information in written text The entropy (rate) h of a sequence of words is the number of bits needed to encode the next word, on average, given past words. Kontoyianni *et al.* [25] proved convergence for a nonparametric estimator $\hat{h}$ for h:

$$\hat{h} = \frac{N \log N}{\sum_{i=1}^{N} \Lambda_i}, \tag{1}$$

where N is the length of the sequence of words and Λ_i is the match length of the prefix at position i, that is, it is the length of the shortest subsequence (of words) starting at i that has not previously appeared. (All logarithms are base 2.) If the sequence of words were randomly shuffled, breaking any long-range structure, this estimator converges to the traditional Shannon entropy on unigrams (see Supplementary Fig. 1).

The ideas underlying estimators such as Eq. (1) play an important role in the mathematics of data compression algorithms. Indeed, some authors have used practical compression software to estimate the information content of a text. However, such estimates tend to be biased, as specific compression implementations (such as gzip) tend to sacrifice small amounts of extra compression to run much more efficiently. Owing to these approximations, it is important to work directly with the theoretical estimator to more accurately estimate h, as we have when we applied Eq. (1).

Measuring the flow of predictive information To generalize Eq. (1) to a cross-entropy between two sequences A and B, we define the **cross-parsed match length** $\Lambda_i(A|B)$ as the length of the shortest subsequence starting at position i of sequence A not previously seen in sequence B. If sequences A and B are ***time-aligned***, as in timestamped social media posts, then 'previously' refers to all of the words of B written prior to $t_i(A)$, the time when the ith word of A was posted, according to the timestamp of the respective tweet. The estimator for the cross-entropy rate is then

$$\hat{h}_\times(A \mid B) = \frac{N_A \log N_B}{\sum_{i=1}^{N_A} \Lambda_i(A \mid B)}, \tag{2}$$

where N_A and N_B are the lengths of A and B, respectively. An estimator of the relative entropy (or KL-divergence), similar to Eq. (2), was introduced by Ziv and Merhav [46]. The log term in Eq. (2) has changed to $\log N_B$ because now B is the 'database' (or window, in Lempel-Ziv terms) we are searching over when we compute the match lengths; the N_A factor is due to the average of the Λ_i's taking place over A. The cross-entropy tells us how many bits on average we need to encode the next word of A given the information previously seen in B. Furthermore, $\hat{h}_\times(A \mid A) = \hat{h}$. The cross-entropy can be applied directly to an ego-alter pair by choosing B to be the text stream of the alter and A the text stream of the ego.

We now wish to generalize the cross-entropy to $\hat{h}_\times(A \mid \mathcal{B})$, estimating the average amount of information needed to encode the next word of sequence A given the information in a *set* of sequences $\mathcal{B}$. Take the cross-parsed match length for a set of databases to be $\Lambda_i(A \mid \mathcal{B}) = \max\{\Lambda_i(A \mid B), B \in \mathcal{B}\}$, that is, the longest match length over any of the sequences in $\mathcal{B}$. This cross-parsing implies a new $\log N_{A\mathcal{B}}$ factor in the estimator, where $N_{A\mathcal{B}}$ is the average of the lengths N_B ($B \in \mathcal{B}$), weighted by the number of times matches were found in each sequence $B \in \mathcal{B}$. (If the same match length occurs for more than one sequence $B \in \mathcal{B}$ then each such sequence receives a weight in the average.) The estimator is

$$\hat{h}_\times(A \mid \mathcal{B}) = \frac{N_A \log N_{A\mathcal{B}}}{\sum_{i=1}^{N_A} \Lambda_i(A \mid \mathcal{B})}, \tag{3}$$

where $N_{A\mathcal{B}} = \sum_{B\in\mathcal{B}} w_B N_B / \sum_{B\in\mathcal{B}} w_B$ and w_B is the number of times that matches from A are found in $B \in \mathcal{B}$. Note that $\sum_B w_B \geq N_A$ due to possible ties, with equality holding if no ties occur. Note that Eq. (3) reduces to Eq. (2) when $|\mathcal{B}| = 1$.

Equation (3) lets us build the cumulative cross-entropy by appropriate choices of $\mathcal{B}$. Here, we sequentially added alters to the set $\mathcal{B}$ in order of decreasing contact volume (i.e., $\mathcal{B} = \{\text{alters}\}$), to understand how information grows as more alters are made available. Likewise, Eq. (3) lets us build the transfer entropy-like measures by additionally including the ego within the set $\mathcal{B}$ (i.e., $\mathcal{B} = \{\text{ego}\} \cup \{\text{alters}\}$).

We implemented Eqs. (1)-(3) in Python. See code availability statement.

Estimator convergence on our data The estimator given by Eq. (1) has been proven to converge asymptotically under stationarity assumptions [25]. However, our data are finite, and so we investigated the convergence properties of the estimator empirically (see Supplementary Figure 1b,c). In general, we observed that the entropy (1) saturates after around 1000 tweets (approximately 10,000 words). Likewise, the cross-entropy estimator $h_\times(A \mid B)$ tends to converge within around 50% of the ego's observed lifespan (see Supplementary Note 1.1).

Code Availability

The code used to generate the results of this paper is available from the corresponding authors upon request.

Data Availability

Data that support the findings of this study are available at Figshare.

References

1. Kossinets, G. & Watts, D. J. Empirical Analysis of an Evolving Social Network. *Science* **311**, 88–90 (2006).
2. Lazer, D. *et al.* Computational social science. *Science* **323**, 721 (2009).
3. Kwak, H., Lee, C., Park, H. & Moon, S. What is Twitter, a Social Network or a News Media? Categories and Subject Descriptors. In *19th International Conference on the World Wide Web (WWW '10)*, 591–600 (2010).
4. Bakshy, E., Messing, S. & Adamic, L. A. Exposure to ideologically diverse news and opinion on Facebook. *Science* **348**, 1130–1132 (2015).
5. Garcia, D. Leaking privacy and shadow profiles in online social networks. *Science Advances* **3** (2017).
6. Shirky, C. The political power of social media: Technology, the public sphere, and political change. *Foreign Affairs* **90**, 28–41 (2011).
7. Lotan, G. *et al.* The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *Int. J. Comm* **5**, 31 (2011).
8. Del Vicario, M. *et al.* The spreading of misinformation online. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 554–559 (2016).
9. Castellano, C., Fortunato, S. & Loreto, V. Statistical physics of social dynamics. *Reviews of Modern Physics* **81**, 591–646 (2009).
10. Kramer, A. D., Guillory, J. E. & Hancock, J. T. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 8788–9790 (2014).

11. Mønsted, B., Sapieżyński, P., Ferrara, E. & Lehmann, S. Evidence of Complex Contagion of Information in Social Media: An Experiment Using Twitter Bots. *PLoS ONE* **12**, e0184148 (2017).

12. Jurgens, D., Tsvetkov, Y. & Jurafsky, D. Writer profiling without the writer's text. In *Social Informatics. SocInfo 2017. Lecture Notes in Computer Science*, vol. 10540, 537–558 (2017).

13. Garcia, D., Goel, M., Agrawal, A. K. & Kumaraguru, P. Collective aspects of privacy in the Twitter social network. *EPJ Data Science* **7** (2018).

14. Gruhl, D., Guha, R., Liben-Nowell, D. & Tomkins, A. Information diffusion through blogspace. In *WWW*, 491–501 (ACM, 2004).

15. Bakshy, E., Rosenn, I., Marlow, C. & Adamic, L. The role of social networks in information diffusion. In *WWW*, 519–528 (ACM, 2012).

16. Aral, S., Muchnik, L. & Sundararajan, A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 21544–21549 (2009).

17. Centola, D. The spread of behavior in an online social network experiment. *Science* **329**, 1194–1197 (2010).

18. Aral, S. & Walker, D. Identifying influential and susceptible members of social networks. *Science* **337**, 337–341 (2012).

19. Ver Steeg, G. & Galstyan, A. Information transfer in social media. In *WWW*, 509–518 (ACM, 2012).

20. Borge-Holthoefer, J. *et al.* The dynamics of information-driven coordination phenomena: A transfer entropy analysis. *Science Advances* **2** (2016).

21. Cover, T. M. & Thomas, J. A. *Elements of Information Theory* (John Wiley & Sons, 2012).

22. Shannon, C. E. Prediction and entropy of printed english. *Bell Syst. Tech. J* **30**, 50–64 (1951).

23. Brown, P. F., Pietra, V. J. D., Mercer, R. L., Pietra, S. A. D. & Lai, J. C. An estimate of an upper bound for the entropy of english. *Comput. Ling.* **18**, 31–40 (1992).

24. Schürmann, T. & Grassberger, P. Entropy estimation of symbol sequences. *Chaos* **6**, 414–427 (1996).

25. Kontoyiannis, I., Algoet, P., Suhov, Y. M. & Wyner, A. Nonparametric entropy estimation for stationary processes and random fields, with applications to english text. *IEEE Trans. Inf. Theory* **44**, 1319–1327 (1998).

26. Song, C., Qu, Z., Blumm, N. & Barabási, A.-L. Limits of predictability in human mobility. *Science* **327**, 1018–1021 (2010).

27. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **85**, 461 (2000).

28. Staniek, M. & Lehnertz, K. Symbolic transfer entropy. *Phys. Rev. Lett.* **100**, 158101 (2008).

29. Dunbar, R. I. Coevolution of neocortical size, group size and language in humans. *Behav. Brain. Sci.* **16**, 681–694 (1993).

30. Albert, R., Jeong, H. & Barabasi, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).

31. Wasserman, S. & Faust, K. *Social network analysis: Methods and applications* (Cambridge university press, 1994).

32. De Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M. & Blondel, V. D. Unique in the crowd: The privacy bounds of human mobility. *Sci. Rep.* **3**, 1376 (2013).

33. de Montjoye, Y.-A., Radaelli, L., Singh, V. K. & Pentland, A. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science* **347**, 536–539 (2015).

34. Pariser, E. *The filter bubble: What the Internet is hiding from you* (Penguin UK, 2011).

35. Mosteller, F. & Wallace, D. L. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association* **58**, 275–309 (1963).

36. Katz, S. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE transactions on acoustics, speech, and signal processing* **35**, 400–401 (1987).

37. Bengio, Y., Ducharme, R., Vincent, P. & Jauvin, C. A neural probabilistic language model. *Journal of machine learning research* **3**, 1137–1155 (2003).

38. Shalizi, C. R. & Thomas, A. C. Homophily and contagion are generically confounded in observational social network studies. *Sociological methods & research* **40**, 211–239 (2011).

39. Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424–438 (1969).

40. Twitter REST APIs. Available from: `https://dev.twitter.com/rest/public` (2016). Accessed: 2016-07-07.

41. Botometer API. Available from: `https://botometer.iuni.iu.edu/` (2016). Accessed: 2016-07-07.

42. Varol, O., Ferrara, E., Davis, C. A., Menczer, F. & Flammini, A. Online human-bot interactions: Detection, estimation, and characterization. In *ICWSM* (2017).

43. Davis, C. A., Varol, O., Ferrara, E., Flammini, A. & Menczer, F. BotOrNot: A system to evaluate social bots. In *WWW Developers Day* (2016).

44. Ferrara, E., Varol, O., Davis, C. A., Menczer, F. & Flammini, A. The rise of social bots. *Communications of the ACM* **59** (2016).

45. Subrahmanian, V. S. *et al.* The DARPA Twitter Bot Challenge. *Computer* **49**, 38–46 (2016).

46. Ziv, J. & Merhav, N. A measure of relative entropy between individual sequences with application to universal classification. *IEEE Trans. Inf. Theory* **39**, 1270–1279 (1993).

Acknowledgements

We gratefully acknowledge the resources provided by the Vermont Advanced Computing Core. This material is based upon work supported by the National Science Foundation under Grant No. IIS-1447634 (J.P.B.). L.M. acknowledges support from the Data To Decisions Cooperative Research Centre (D2D CRC), and the ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author Contributions

J.P.B. and L.M. designed the research. L.M. oversaw data collection and processing. X.L. collected and analyzed human rater data. J.P.B. and L.M. analysed the data and wrote the manuscript.

Competing Interests

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.P.B. (email: james.bagrow@uvm.edu) or L.M. (email: lewis.mitchell@adelaide.edu.au).