Supplementary Information for “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
[*]Corresponding author. Email: james.bagrow@uvm.edu, Homepage: bagrow.com
[+]Corresponding author. Email: lewis.mitchell@adelaide.edu.au, Homepage: maths.adelaide.edu.au/lewis.mitchell/

Contents

List of Supplementary Figures

List of Supplementary Tables

Supplementary Note 1

Cross-entropy estimator convergence

Supplementary Figure 1 shows the difference between our entropy estimator and traditional Shannon entropy (panel a), as well as the estimator convergence (panels b & c). See the Methods section in the main text for more information.

For the cross-entropy estimator $h_\times(A \mid B)$, we examined the convergence over the lifespan or time window within which the ego has authored tweets. Supplementary Figure 2 shows the convergence of the cross-entropy for the rank-1 alter $h_\times$(ego|alter 1) (left panel), where we truncate both the ego and alter's tweets after some fraction of the ego's lifespan. In general, we found that the cross-entropy estimator saturates within around 50% of the ego's lifespan. The right panel shows a histogram of the slopes of the convergence curves for all users over the final 25% of the ego's lifespan, as a fraction of the final value of $h_\times$(ego | alter 1). These slopes were computed via linear regressions, and many of the slopes are close to zero.

The cross-entropy can also be associated with the predictability by applying Fano's Inequality [1]. Fano's Inequality relies on both the entropy and the cardinality of the random variable; here we take the size of the ego's unique vocabulary as this is the variable we are trying to predict. In Supplementary Figure 3 we present the relationship between cross-entropy and predictability for our data compared with solid lines denoting constant vocabulary-size curves. The predictability of the ego given the alter is lower than the predictability of the ego given the ego because the cross-entropy is greater than the entropy, capturing the increased uncertainty (decreased information) we have by trying to predict the ego given the alter instead of the ego.

Supplementary Note 2

Extrapolating cross-entropy and predictability

We are limited by our data to a window of the top-15 most frequently contacted alters per ego. To address a limit of entropy or predictability as more alters are added, we used a saturating function to extrapolate beyond alter rank $r = 15$.

Specifically, we extrapolated the cross-entropy using the function

$$h(r) = h_\infty + \frac{\beta_0}{\beta_1 + r}, \tag{S1}$$

with the goal of identifying the value of h_∞ and, perhaps more realistically, to estimate $h(r_{\text{dunbar}})$, where $r_{\text{dunbar}} \approx 150$ [2]. Using Levenberg-Marquardt for nonlinear regression, we found best fit parameters of (value ± 95% CI):

$$\begin{aligned} h_\infty &= 5.761978 \pm 0.089699, \\ \beta_0 &= 9.455984 \pm 1.358027, \\ \beta_1 &= 2.553345 \pm 0.444479, \end{aligned}$$

for the cross-entropy of the ego given the alters.

In Supplementary Figure 4 we show the mean cross-entropy as a function of alter rank and compare it with the results of the fitted function. The fit is reasonable. Similarly, fits of the same functional form were applied to the predictability (ego given alter) curves:

$$\Pi(r) = \Pi_\infty + \frac{\beta_0}{\beta_1 + r}, \tag{S2}$$

and here we found best fit parameters

$$\begin{aligned}\Pi_\infty &= 0.608219 \pm 0.006914,\\ \beta_0 &= -0.734410 \pm 0.100195,\\ \beta_1 &= 2.320039 \pm 0.398486.\end{aligned}$$

We also experimented with a second form of extrapolating function:

$$h(r) = h_\infty + \beta_0 r^{-\beta_1}, \qquad \Pi(r) = \Pi_\infty + \beta_0 r^{-\beta_1}. \tag{S3}$$

This function, referred to as **Function 2**, also fits the data well (Supplementary Figure 5) but is a bit less conservative in its extrapolation prediction when extrapolating for $r \to \infty$. To further compare Function 2 and the original function (Function 1), we plotted the residuals between the fits and the data in Supplementary Figure 5.

Taken together, we see that Function 1 (Eqs. (S1) and (S2)), the more conservative estimator, has consistently smaller residuals than Function 2. Both functions' residuals were statistically independent of the exogenous variable r ($p > 0.05$). We concluded that Function 1 is a better choice since it has smaller residuals and is more conservative than Function 2.

Supplementary Note 3

Vocabulary sizes on social media

In Supplementary Figure 6 we present the distributions of the total number of words written per ego and the number of unique words (the vocabulary size) per ego, for the users in our dataset. Egos had (mean ± s.d.) 26802.76 ± 9061.53[1] total numbers of words and 5207.44 ± 1769.48 numbers of unique words. The latter quantity, the vocabulary size, was used in Fano's Inequality to compute the predictability.

Supplementary Note 4

Information content on social media compared with formal written text

To contextualize the entropy rates we estimated for our dataset (most egos had entropies of $5.5 < h < 8$ bits), we compared the entropy rates of formal text with the rates of Twitter users to better understand the information content of social media writings[2]. First, we considered the entropies of some famous example texts (Supplementary Table 1). We considered writers who were famous for being very simple in style (Hemingway) and very complex (Joyce) and found this was reflected in the entropy rates (5.87 bits for Hemingway compared with 7.06 bits for Joyce). The higher entropies reflect that Joyce's word choices are less regular and less predictable than Hemingway's. These formally written and edited texts are very different from social media posts, and yet the range of entropies values we observed was compatible to some extent.

We also took the standard ***Brown corpus*** [4], a benchmark text set used in natural language processing and computational linguistics research, as a large-scale baseline of formal text. The corpus consists of approximately 1M words and covers 500 writing samples across 15 fiction and non-fiction categories. Each

[1]For context, this is about the typical length of a novella, defined by the Science Fiction and Fantasy Writers of America as 17500–39999 words [3].

[2]The texts were processed by removing punctuation and casing.

category was broken into 10-thousand-word chunks and the entropies of these chunks were computed. Individual chunks did not span multiple categories and if a chunk at the end of a category was less than 10 thousand words it was discarded to ensure all entropy estimates were computed using the same amount of data. This gave $n = 93$ samples.

We found that formal and social text have the same average value but that the variation across the Twitter sample was significantly greater than across the formal text (Supplementary Figure 7).

Supplementary Note 5

A censoring filter to determine long-range information in the egos and alters

To study the recency of information we applied the (cross-)entropy estimators to censored text, where we removed the recent past of the text and asked how much if any information is lost. If most predictive information is in the recent past, by removing it we should see a significant change in the cross-entropy, although there should always be some loss in information, as the sequences being matched across are always getting shorter.

Specifically, to compute the original cross-entropy (Eq. (2)) between two sequences A and B, we need the cross-parsed match length at position i, $\Lambda_i(A \mid B)$, giving us the shortest subsequence of words in A beginning at position i not seen previously in B. This last part, the past of B, is based on the timestamps of the words: we search all words in B written before the ith word w_i in A: $[w_j \in B \mid t_j(B) < t_i(A)]$, where $t_i(A)$ is the time when the ith word in A was posted (taking all words in a single tweet to be posted at the time the tweet was posted).

The censoring filter simply truncates the past of B at each position i. Instead of searching all of the past of B we instead search the past older than an amount ΔT: $[w_j \in B \mid t_j(B) < t_i(A) - \Delta T]$. By censoring B as we sweep forward in the computation of the cross-entropy, we can estimate how much information is recent versus long-term on average by the change in the cross-entropy rate as a function of ΔT. The same calculation holds for the "self" entropy, simply by setting $B = A$.

We measured the loss of information in the main text out to 24 hours. Here we complement that calculation with Supplementary Figure 8 which presents the information loss out to 1 week. We see in both curves that long-range information is lost by the increasing trend. However, the trend is more shallow for the alters than the ego: taking away more of the ego's past removes more information about the ego than taking away the less recent pasts of the alters.

This censoring filter reduces the amount of data available from which to compute the cross-parsed match lengths $\Lambda_i(A \mid B)$. We investigate the extent of this data loss in Supplementary Figure 9, which shows how the number of censored alter tweets per ego tweet n_{cens} depends on the censoring interval ΔT. The left panel shows distributions of the mean n_{cens} for different lags ΔT from 0.5 to 6 hours. Each distribution is obtained by counting the number of tweets posted by all alters in ΔT hours before each ego tweet, taking the mean of these values, and then plotting the distribution of these mean values. The right panel shows the mean and 95% quantiles of these distributions as a function of ΔT. As expected, the mean n_{cens} increases roughly linearly as a function of ΔT, however the numbers of censored tweets remain relatively low compared to the total amount of data available ($\approx$3200 tweets for each ego).

Supplementary Note 6

Posting frequency and predictability

Main text Fig. 2b demonstrated associations between predictability of the ego given the alter and how frequently either the ego or the alter posted to Twitter, as quantified by the average number of posts per day. Egos who posted more than 8 times per day on average were 16.5% ± 14.9% (difference in mean predictability ± 95% CI on mean) more predictability from their alter than egos who posted less than 1 time per day on average. Likewise, egos where the alter posted more than 8 times per day on average were 23.8% ± 4.46% less predictability from the alter than egos where the alter posted less than 1 time per day on average. These changes in predictability show that egos who post more frequently are more predictable from their alters than less frequent posting egos, while the opposite association holds about egos with frequent and infrequent posting alters. However, these differences in predictability only highlight the extreme ends of the data range, so we also measured the statistical association across the entire posts-per-day range: all measured associations were significant (Supplementary Table 2).

Further, in Supplementary Figures 10 and 11 the association between the posting frequencies of the egos and alters with the predictability of the ego (Supplementary Figure 10: top row), predictability of the alter (Supplementary Figure 10: bottom row), predictability of the ego given the alter (Supplementary Figure 11: top row), and the predictability of the alter given the ego (Supplementary Figure 10: bottom row). We found that the predictabilities of the egos and alters are roughly independent of their posting frequency, except for very infrequently posting alters (Supplementary Figure 11: bottom right), which is likely a result of insufficient data.

The associations between posting frequency and the cross-predictability of the ego given the alter hold even when considering all alters not just the rank-1 alter, as we did in the main text (Supplementary Figure 11: top row). Likewise, the trends also hold (in reverse) when considering the predictability of the alter given the ego (Supplementary Figure 11: bottom row).

Supplementary Note 7

Contact volumes and predictability

Here we present in Supplementary Figure 12 the predictability across social ties as a function of how often those social ties contact one another. We ranked the ties of individuals in descending order. Working with ranks helps to account for the variability in contact volumes and overall activity levels across users of social media. Across ranks we found a significant decrease in predictive information, in both directions (predicting the ego given the alter and predicting the alter given the ego).

Supplementary Note 8

Cross-entropy homophily

In the main text we reported a homophily between the entropies of the egos and their alters. Here we explore a similar association with their cross-entropies.

The cross-entropies between the egos and alters are less well correlated, either with the cross-entropy in

the opposite direction, or with the entropies themselves. The correlations (for the rank-1 alters) are:

$$
\begin{aligned}
R\left(\hat{h}(\text{ego}), \hat{h}(\text{alter})\right) &= 0.478, \\
R\left(\hat{h}_{\times}(\text{ego} \mid \text{alter}), \hat{h}_{\times}(\text{alter} \mid \text{ego})\right) &= -0.122, \\
R\left(\hat{h}(\text{ego}), \hat{h}_{\times}(\text{ego} \mid \text{alter})\right) &= 0.240, \\
R\left(\hat{h}(\text{ego}), \hat{h}_{\times}(\text{alter} \mid \text{ego})\right) &= 0.227, \\
R\left(\hat{h}(\text{alter}), \hat{h}_{\times}(\text{ego} \mid \text{alter})\right) &= 0.247, \\
R\left(\hat{h}(\text{alter}), \hat{h}_{\times}(\text{alter} \mid \text{ego})\right) &= 0.300.
\end{aligned}
$$

While significant in all cases, the correlations between the (self) entropies $\hat{h}(\text{ego})$ and $\hat{h}(\text{alter})$ are stronger than between any of the cross-entropies, demonstrating that the effects captured by the cross-entropies over a dyad are different than that captured by the entropies of the individuals in that dyad.

Supplementary Note 9

Reciprocity and information flow

Table 3 provides statistical analyses of the associations reported in main text Fig. 3. Due to potential nonlinear relationships, we report monotonicity coefficients (both Spearman's Rho and Kendall's Tau). We see significant associations between predictability of the ego given the alter and the number of social ties of the ego (cf. main text Fig. 3a). Likewise, we see significant associations between contact volume and predictability of the ego given the alter, with a positive trend for alter-mentions-ego contact volume and a slight negative trend for ego-mentions-alter contact volume. This asymmetry supports information flow capturing directionality in relationships (cf. main text Fig. 3b).

In the main text we reported on the relationship between contact volume and information flow, as measured by the cross-entropy (and mapped into the predictability. A closely related quantity often employed in this context is the Kullback-Leibler divergence, or KL-divergence, $KL(\text{ego} \parallel \text{alter}) \equiv \hat{h}_{\times}(\text{ego} \mid \text{alter}) - \hat{h}(\text{ego})$ [1]. In our data the correlation between $\hat{h}_{\times}(\text{ego} \mid \text{alter})$ and $KL(\text{ego} \parallel \text{alter})$ is quite high (Supplementary Table 4) and so they are effectively the same measure.

We showed that alters who more frequently mention their ego provide more predictive information (lower cross-entropy/KL-divergence) than alters who less frequently mention their ego. Meanwhile, the converse was not true: the ego can mention the alter more or less, and there was not an association with the predictive information possessed by the alter about the ego.

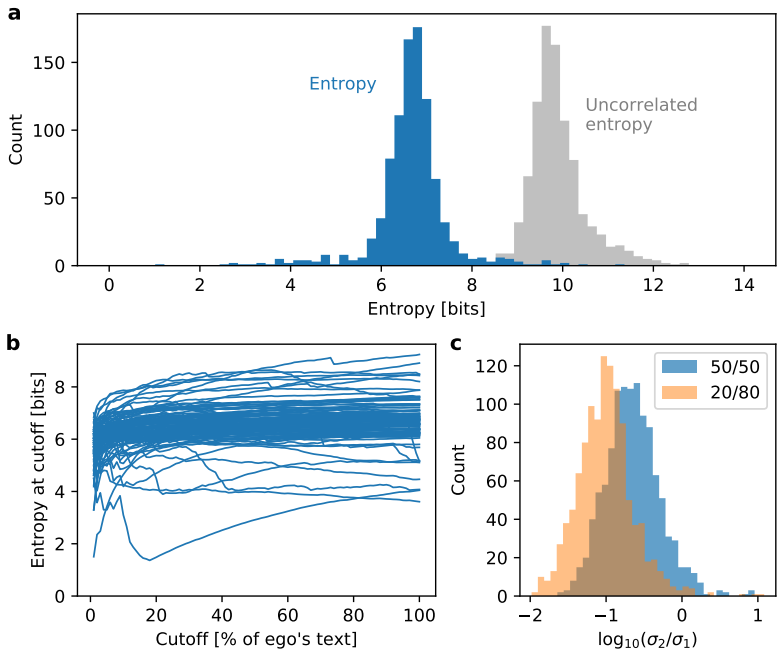
Here we supplement that result by reversing the perspective—instead of asking about the predictive information about the ego possessed by the alter we ask about the predictive information about the alter possessed by the ego. We measure this with the reversed KL-divergence, $KL(\text{alter} \parallel \text{ego}) \equiv \hat{h}_{\times}(\text{alter} \mid \text{ego}) - \hat{h}(\text{alter})$. With this reversal we should expect to also see a reversal in the association of contact volume, and we found this to be the case (Supplementary Figure 13). In Supplementary Figure 13 we compared both KL-divergences and saw that the trends approximately reverse, as expected.

Supplementary Note 10

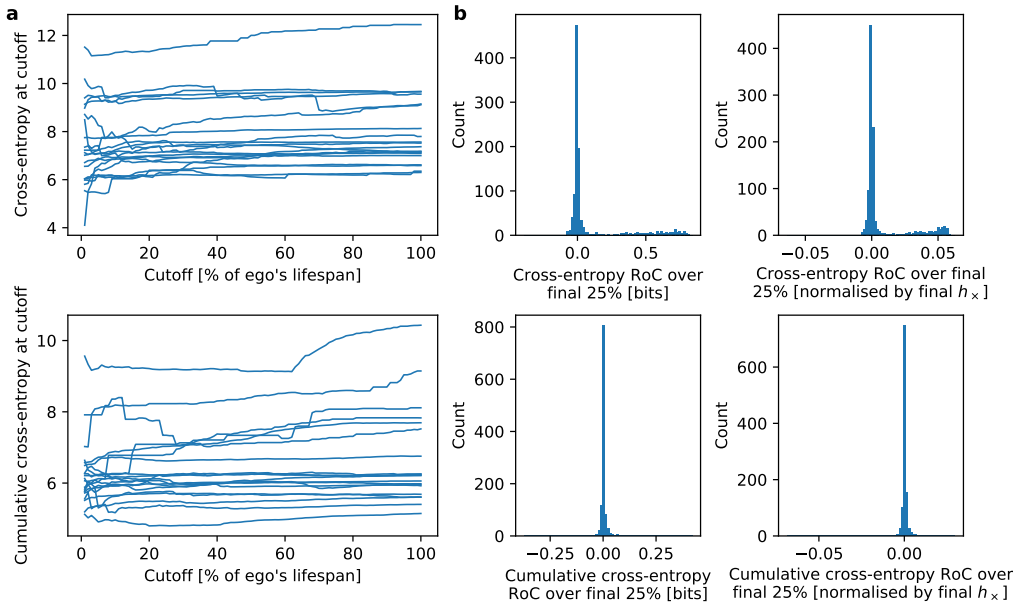
Interrelations between information-theoretic quantities

In Supplementary Table 4 we present the Spearman rank correlation coefficients between the primary information-theoretic quantities we computed, including the KL-divergence: $KL(\text{ego} \parallel \text{ater}) \equiv \hat{h}_{\times}(\text{ego} \mid \text{alter}) - \hat{h}(\text{ego})$. The cross-entropy and KL-divergence are strongly correlated.

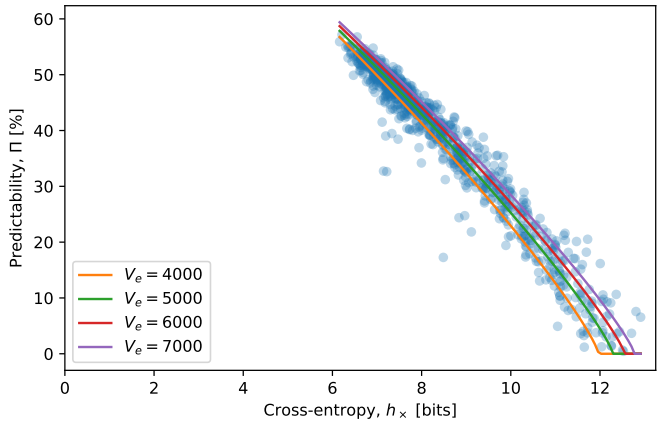
Supplementary Figures



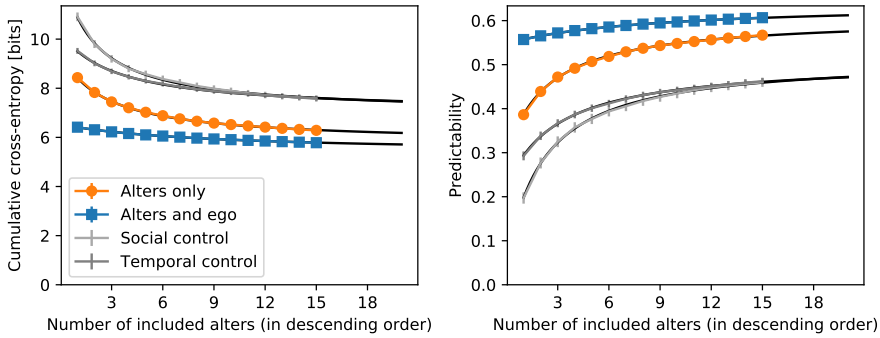
Supplementary Figure 1: **a**, Correlations in the text account for ≈3 additional bits of information. The uncorrelated entropy (considering only the relative frequencies of words posted by Twitter users) is approximately 3 bits higher than the correlated entropy as estimated from Eq. 1. **b & c**, Entropy estimator convergence. **b**, The estimator generally saturates well within our data window, as evidenced by the flattening of the entropy estimate as we examine more of the ego's text. **c**, Here we compute the variance of each ego's entropy over two portions of the curves at left. One distribution compares the variance of the final 50% of the data to the initial 50%, while the other compares the variance of the final 20% of the data to the initial 80%. The latter shows significantly smaller variability, underscoring how entropy estimates have converged within our data window. The left plot shows a random selection of egos, while the right covers all dyads in our dataset.



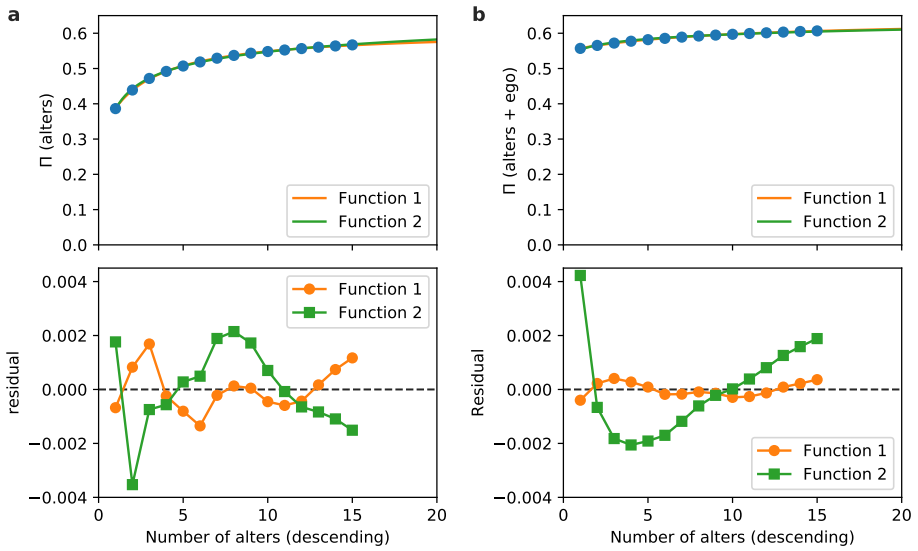
Supplementary Figure 2: Convergence of the cross-entropy estimator. **a**, The estimator saturates well within the lifespan of the ego's tweets, generally within 50% of the lifespan. **b**, The distributions of the slope (RoC: rate-of-change) over the final 25% of the curves. The majority of egos have very flat RoCs at the end of their data windows. In the left plots we show a randomly selection of egos, while the distributions on the right curve all dyads in our dataset.



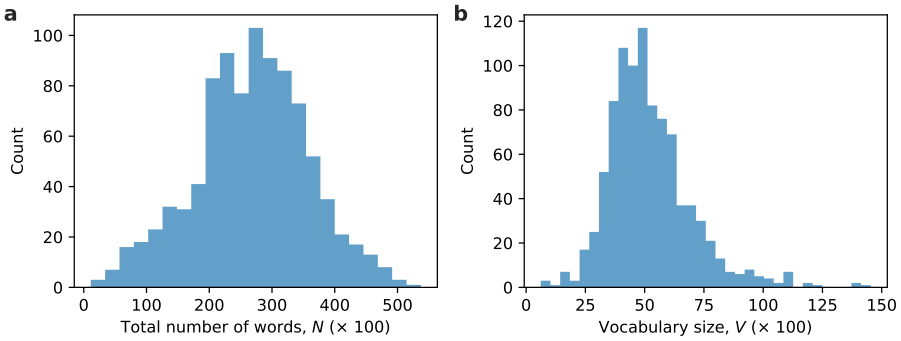
Supplementary Figure 3: Cross-entropy $\hat{h}_\times$(ego | alter) and predictability Π across different ego vocabulary sizes v_e indicated by color.



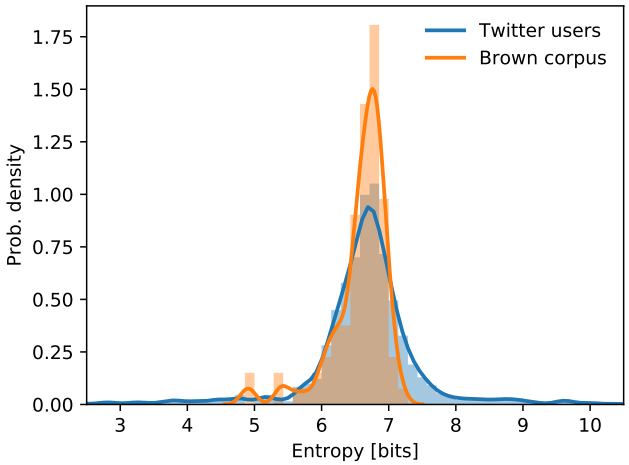
Supplementary Figure 4: Extrapolating cross-entropy and predictability. The fitted functions (Eqs. (S1) and (S2), solid lines) compared with the original cross-entropy data (averaged for each alter rank). Note that the function was fitted to the original and not averaged cross-entropy values. Points and error bars denote means and 95% CIs, respectively, on the data, while error bars without points denote the same quantities for the social and temporal controls. Colors distinguish the controls (light gray: social control; dark gray: temporal control), and whether the ego's information was included alongside the alters' (blue: alter and ego; orange: alter only). After fitting to the original data, the extrapolating function was plotted out to a value of 20 (beyond our data window of 15 alters) to highlight the extrapolation.



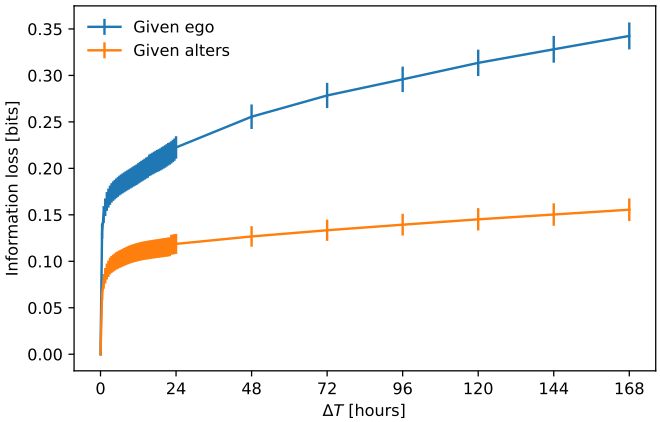
Supplementary Figure 5: Extrapolations and residuals for the predictability functions (Function 1: Eq. (S2); Function 2: Eq. (S3)). **a**, Comparison of the measured cross-entropies (for the top-15 alters) with the extrapolation functions and mean residuals between function fit and original data. **b**, Same as panel A but for fits of the same form as Eqs. (S2) and (S3) to $\Pi(r)$ including the past of the ego along with the alters. Overall, Function 1 was slightly more conservative than Function 2, extrapolating to a slightly smaller value of Π, and had lower residuals. Colors denote the data (blue), Function 1's fit and residual (orange), and Function 2's fit and residual (green).



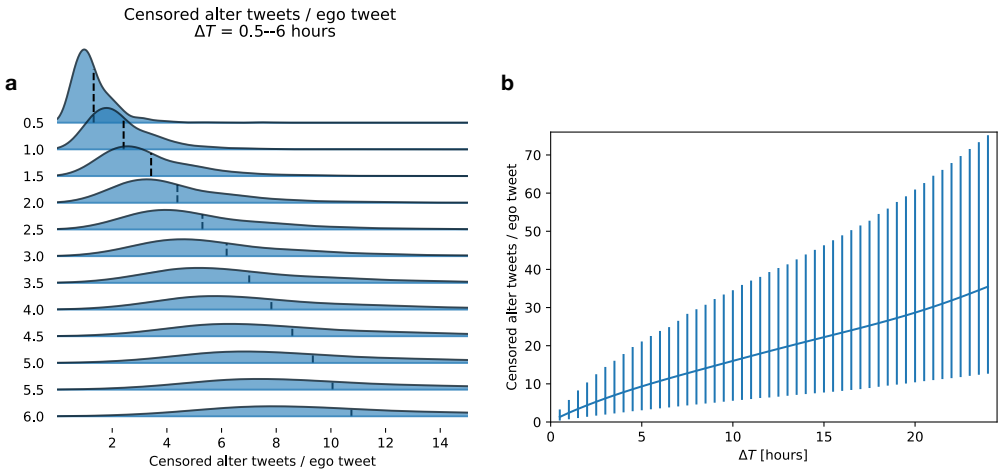
Supplementary Figure 6: Distributions of Twitter ego vocabulary size. **a**, The total number of words written. **b**, The vocabulary size or number of unique words written.



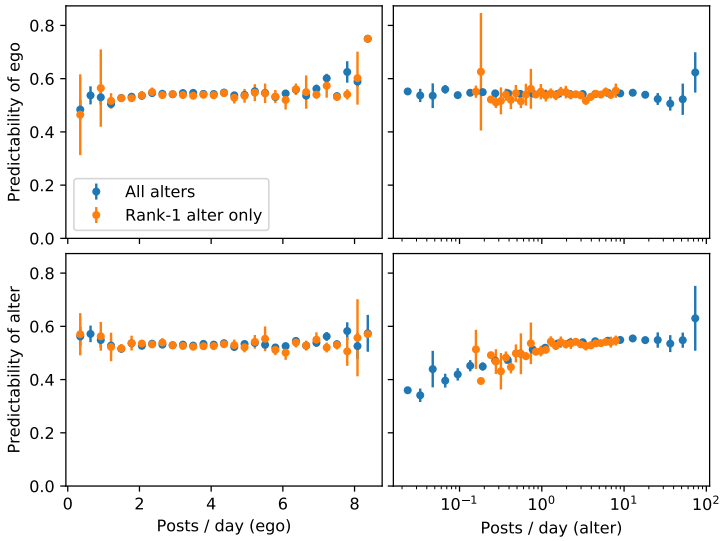
Supplementary Figure 7: Entropy distributions for social and formal written text corpora. We found that the distributions have the same central tendency (Mann-Whitney U test: $U = 39797$, $p > 0.1$) but different dispersions (Fligner-Killeen test on variances, $\chi^2 = 15.580$, $p < 10^{-4}$) Brown corpus was taken from NLTK v3.2.1 corpora [5].



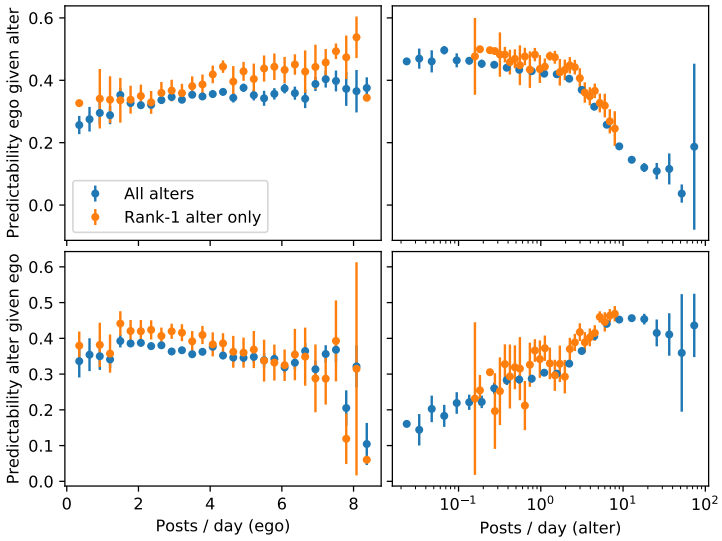
Supplementary Figure 8: Alters provided less long-range information about the ego than the ego itself. This plot complements the loss in predictability shown in the main text and extends ΔT beyond the 24-hour window to a one-week period. Error bars show 95% CIs. Up to 24 hours we show information loss every 30 minutes; every 24 hours thereafter.



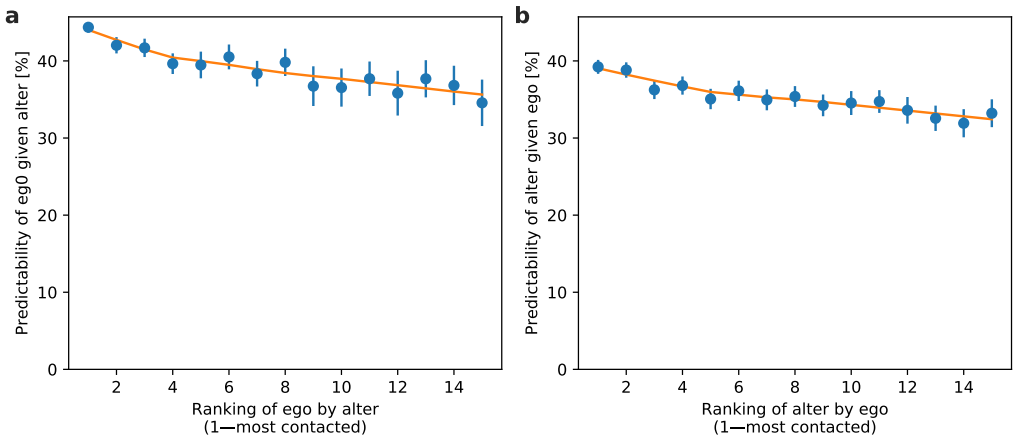
Supplementary Figure 9: Number of alter tweets censored per ego tweet increases with ΔT. **a**, Distributions of mean number of censored alter tweets per ego tweet for lags ΔT from 0.5–6 hours. Vertical lines show the mean of each distribution. **b**, Mean number of censored alter tweets/ego tweet as a function of ΔT. Error bars show 95% quantiles.



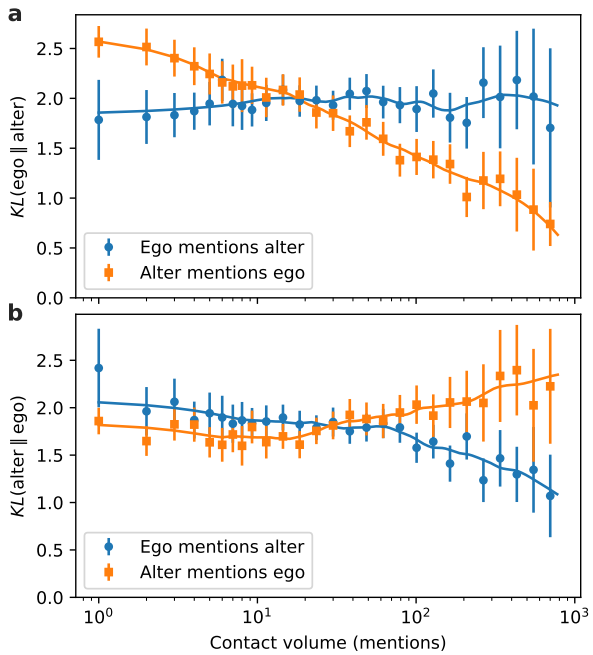
Supplementary Figure 10: Self-predictabilities are approximately independent of activity frequency, with the exception of predictability of the alter as a function of the alter's activity frequency (lower right). This is primarily due to insufficient data: alters who post very infrequently have low predictability, but qualitatively the trend levels off for alters who post more than ≈1 time per day. Error bars show 95% CIs.



Supplementary Figure 11: Association between cross-predictability and posting frequency holds for all alters. Here we repeated the trends shown in the main text where we considered the rank-1 alter only, but now we included all alters as well. Due to alters who post very frequently and very infrequently, we used a logarithmic scale on the right column. Error bars show 95% CIs.



Supplementary Figure 12: Less frequently contacted ties provide less predictive information. Here we plot the mean predictability of the ego given the alter averaged over ego-alter pairs conditioned on the rank of the ego by the alter (panel **a**) or rank of the alter by the ego (panel **b**), with rank-1 being the most frequently contacted social tie. Error bars show 95% CIs and the solid line denotes a LOWESS fit that provides a guide for the eye.



Supplementary Figure 13: Reciprocated information flows are captured in both directions. **a**, In the main text we reported on the trend between contact volume and the cross-entropy from the alter to the ego. We repeat that figure here but with the KL-divergence. **b**, In comparison, if we consider the opposite divergence, from the ego to the alter, we see a similar trend but reversed: egos which more frequently mention their alter give more predictive information (lower divergence) than egos which mention their alter less often. In both panels, error bars show 95% CIs, solid lines denote LOWESS fits that provide a guide for the eye, and colors indicate direction of contact (either ego mentions alter or alter mentions ego).

Supplementary Tables

Supplementary Table 1: Entropy rates of some example texts. Samples 1 and 2 were the first and second 38,000 words of each text, respectively (a bit longer than the typical Twitter user's text stream). Hemingway is known for his simple writing style while Joyce is famous for the opposite; this is well reflected in their respective entropy rates.

Text	Author	$\hat{h}$ (sample 1) [bits]	$\hat{h}$ (sample 2) [bits]
For Whom the Bell Tolls	Ernest Hemingway	5.870953	5.910003
Gravity's Rainbow	Thomas Pynchon	5.881336	5.881336
The Fellowship of the Ring	J.R.R. Tolkien	6.439215	6.340354
Ulysses	James Joyce	7.067339	7.227677

Supplementary Table 2: Predictibility vs. posting frequency (cf. main text Fig. 2b). Both trends shown in Fig. 2b had statistically significant correlations, reported here using the Spearman's rho and Kendall's tau correlation measures. Confidence intervals on τ were computed as per Sec. 8.3 of Hollander *et al.* [6].

	Spearman's ρ [95% CI]	p-value	Kendall's τ [95% CI]	p-value
Posts / day (ego)	0.276 [0.216, 0.335]	$< 10^{-16}$	0.183 [0.145, 0.222]	$< 10^{-16}$
Posts / day (alter)	−0.437 [-0.487, -0.383]	$< 10^{-43}$	−0.291 [-0.327, -0.256]	$< 10^{-39}$

Supplementary Table 3: Predictibility vs. social ties (cf. main text Fig. 3a) and contact volume (cf. main text Fig. 3b). The asymmetry in the associations of predictability with the two directions of contact volume demonstrate how information flow captures the directionality of relationships. Confidence intervals on τ were computed as per Sec. 8.3 of Hollander *et al.* [6].

	Spearman's ρ [95% CI]	p-value	Kendall's τ [95% CI]	p-value
Num. social ties of alter	−0.199 [-0.224, -0.175]	$< 10^{-53}$	−0.133 [-0.150, -0.117]	$< 10^{-52}$
Contact vol. (Ego → alter)	−0.0185 [-0.0440, 0.00704]	0.156	−0.0124 [-0.0290, 0.00500]	0.156
Contact vol. (Alter → ego)	0.226 [0.202, 0.250]	$< 10^{-68}$	0.154 [0.137, 0.170]	$< 10^{-67}$

Supplementary Table 4: Cross-entropy and KL-divergence are strongly correlated (bold). Here we present the nonparametric Spearman correlation between information-theoretic quantities computed over the $n = 927$ ego-(rank-1 alter) dyads.

	$\hat{h}$(ego)	$\hat{h}$(alter)	$\hat{h}_\times$(e | a)	$\hat{h}_\times$(a | e)	KL(e ‖ a)	KL(a ‖ e)
$\hat{h}$(ego)	1.000000	0.439867	0.302718	0.201535	-0.142122	0.000338
$\hat{h}$(alter)	0.439867	1.000000	0.257699	0.236107	0.048474	-0.176895
$\hat{h}_\times$(e | a)	0.302718	0.257699	1.000000	-0.281961	**0.858274**	-0.395016
$\hat{h}_\times$(a | e)	0.201535	0.236107	-0.281961	1.000000	-0.384678	**0.881891**
KL(e ‖ a)	-0.142122	0.048474	**0.858274**	-0.384678	1.000000	-0.409757
KL(a ‖ e)	0.000338	-0.176895	-0.395016	**0.881891**	-0.409757	1.000000

Supplementary References

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012. 2, 6

[2] R. I. M. Dunbar. The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews*, 6(5):178–190, 1998. 2

[3] Science Fiction and Fantasy Writers of America, Inc. Nebula rules. http://nebulas.sfwa.org/about-the-nebulas/nebula-rules/. Accessed: 2017-08-05. 3

[4] W. N. Francis and H. Kucera. Brown corpus manual. *Brown University*, 2, 1979. 3

[5] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python*. O'Reilly Media, Inc., 2009. 13

[6] M. Hollander, D. A. Wolfe, and E. Chicken. *Nonparametric statistical methods*, volume 751. John Wiley & Sons, 3rd edition, 2013. 20