The quoter model: A paradigmatic model of the social flow of written information

James P. Bagrow[1,a)] and Lewis Mitchell[2,b)]
[1]Department of Mathematics and Statistics, University of Vermont, Burlington, Vermont 05405, USA
[2]School of Mathematical Sciences, North Terrace Campus, The University of Adelaide, Adelaide, South Australia 5005, Australia

(Received 31 October 2017; accepted 23 February 2018; published online 11 July 2018)

We propose a model for the social flow of information in the form of text data, which simulates the posting and sharing of short social media posts. Nodes in a graph representing a social network take turns generating words, leading to a symbolic time series associated with each node. Information propagates over the graph via a quoting mechanism, where nodes randomly copy short segments of text from each other. We characterize information flows from these text via information-theoretic estimators, and we derive analytic relationships between model parameters and the values of these estimators. We explore and validate the model with simulations on small network motifs and larger random graphs. Tractable models such as ours that generate symbolic data while controlling the information flow allow us to test and compare measures of information flow applicable to real social media data. In particular, by choosing different network structures, we can develop test scenarios to determine whether or not measures of information flow can distinguish between true and spurious interactions, and how topological network properties relate to information flow. *Published by AIP Publishing.* https://doi.org/10.1063/1.5011403

Rich datasets on human activity and behavior are now available, thanks to the widespread adoption of online platforms such as social media. The primary artifact generated by users of these platforms is text in the form of written communication. These symbolic data are invaluable for research on information flow between individuals and across large-scale social networks, but working with and modeling natural language data is challenging. While most models of social information flow focus on compartment models, contagion models, or cascades, the richness of the text data available to researchers underscores the importance of incorporating the full information present in text into modeling efforts. In this paper, we propose a model for how groups of individuals embedded in a social network can generate streams of text data based on their own interests and the interests of their neighbors in the network. The goal is to more explicitly capture the dynamics inherent to human discourse. We show how to relate parameters in the model to quantities underlying information-theoretic estimators specifically aimed at understanding information flow between sources of text. By controlling the graph topology and model parameters, we can benchmark how information flow measures applied to text deal with spurious interactions and confounds.

Recently, considerable effort has taken place to better understand information flow in dynamical systems and real datasets.[1] On one hand, new measures and algorithms have been developed to better understand information flow interactions and related phenomena, including transfer entropy,[2] symbolic transfer entropy,[3] convergent cross-mapping,[4] and causation entropy.[5,6] On the other hand, new large-scale datasets have allowed researchers to better understand at scale the spread of information in a complex system, especially those involving online social networks and social media such as Twitter.[7,8] Especially interesting are studies applying information-theoretic tools to large-scale social media data, such as Ver Steeg and Galstyan, who consider the shared information present in the timings of tweets posted by social ties on Twitter,[9] and Borge-Holthoefer *et al.*, who use symbolic transfer entropy to investigate predictive signals of collective action such as protests in the time series of the numbers of tweets posted in different geographic regions.[10] These recent studies show that tools developed from information theory and dynamical systems theory can successfully be applied to human dynamics data captured from online platforms such as Twitter.

Most research on information flow within online media either considers proxies of information flow, such as tracking the spread of particular keywords, or uses information-theoretic tools focused on the timing of social media posts.[9,10] Yet the posts themselves are packed with potentially useful data: the text generated by users of online platforms is their primary artifact and, when available for study, should be the focus of research. Fortunately for the study of information flow, information theory has a rich history of working with symbolic data such as text.

Given the importance of focusing on the text data, there is currently a lack of models for the problem of studying information flow as measured from the text generated by users in a social network. Most work focuses on modeling information flow as a type of contagion, cascade, or diffusion process.[7,11–13] These works are invaluable for studying information flow but by compartmentalizing nodes into groups that have or have not adopted an innovation, been "infected," etc.

[a)] Electronic mail: james.bagrow@uvm.edu
[b)] Electronic mail: lewis.mitchell@adelaide.edu.au

they generally neglect the full richness of the text generated by users in this setting.

Our goal here is to propose and analyze a simple model of the discourse underlying the text generation process online. Nodes within a given graph (representing individuals within a social network) generate symbolic time series (the time-ordered text) based on what they and their neighbors in the network say, and we relate this to information-theoretic estimators of information flow between the texts of different individuals. Doing so provides insights into how well these estimators can distinguish true versus spurious interactions, detect confounding effects, and help us relate network topological properties to the features of information flow.

The rest of this paper is organized as follows. In Sec. I, we discuss background material on entropy estimators for written text and how they may be used to measure information flow. In Sec. II, we introduce the quoter model and discuss its different components. In Sec. III, we analyze the quoter model between two individuals and compare our analytic predictions with simulations. Section IV extends these simulations to a number of network structures and investigates the interplay between network topology and information flow. We conclude with a discussion of our results and potential future directions in Sec. V.

I. BACKGROUND

A. Entropy and information flow in text

The information content in a written text can be quantified with its entropy rate h, the number of additional bits (or other unit of information) needed on average to determine the next word[14] of the text given past words.[15] The entropy rate is maximized for a text that is completely random such that preceding words will not give useful information for determining a subsequent word. Conversely, the entropy rate is zero for a deterministic sequence of words such that knowledge of previous words only gives all the information necessary to specify the subsequent word.

There is a rich history of practical entropy estimators for text.[16–18] The challenge when working with real text is that there is information in the ordering of words, not just their relative frequencies—shuffling a text preserves the (unigram) Shannon entropy but destroys much of the information in the text. To account for the ordering of words, one needs to evaluate the complete joint (or conditional) distribution of word occurrences, and estimating these probabilities requires enormous amounts of data.

Kontoyianni *et al.*[19] proved that the estimator

$$\hat{h} = \frac{T \log_2 T}{\sum_{t=1}^{T} \Lambda_t} = \frac{\log_2 T}{\bar{\Lambda}}, \tag{1}$$

converges to the true entropy rate h of a text, where T is the length of the sequence of words and Λ_t is the *match length* of the prefix at position t: it is the length of the shortest substring (of words) starting at t that has not previously appeared in the text. (For simplicity, we now omit the ˆ symbol distinguishing the estimator from the true quantity.) Theorems underlying nonparametric estimators such as Eq. (1) play an important role in the mathematics of data compression. Indeed, some authors have even used compression software to estimate the entropy of text. However, using compression software risks introducing bias, as specific compression code (such as gzip) trades off optimal compression rates in order to run much more efficiently. Due to these trade offs, one should instead work directly with the theoretical estimator [Eq. (1)] to more accurately estimate h.

Equation (1) generalizes naturally to a **cross-entropy** between two sequences A and B.[20,21] To do so, define the *cross-parsed match length* $\Lambda_t(A|B)$ as the length of the shortest substring starting at position t of sequence A not previously seen in sequence B. If sequences A and B are *time-aligned*, as in a written conversation unfolding over time, then "previously" refers to all the words of B written prior to the time when the tth word of A was written. The estimator for the cross-entropy rate is then

$$h_\times(A \mid B) = \frac{T_A \log_2 T_B}{\sum_{i=1}^{T_A} \Lambda_i(A \mid B)}, \tag{2}$$

where T_A and T_B are the lengths of A and B, respectively. The log term in Eq. (2) has changed to $\log_2 T_B$ because now B is the "database" we are searching over to compute the match lengths and the T_A factor is due to the average of the Λ_t'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. Further, $h_\times(A \mid A) = h$. Despite a similarity in notation, the cross-entropy is distinct from the conditional entropy (which requires estimating a joint probability distribution of A and B, something that is not easy to estimate from social media text data, for example). The cross-entropy can be applied directly to text of a pair of individuals by choosing B to be the text stream of one individual and A the text stream of the other.

While our focus in this work is on the cross-entropy between pairs of individuals, $h_\times$ can be generalized further to $h_\times(\mathrm{A} \mid \mathcal{B})$, quantifying the predictive information regarding the text in string A contained within a *set* of strings $\mathcal{B}$.[21] This lets us understand the information flow from multiple social ties to a single individual. It also allows us to construct transfer entropy-*like* measures: $h(A) - h_\times(\mathrm{A} \mid \{\mathrm{A}, \mathrm{B}\})$ measures how much if any extra information is present on average in the past text of B about the future future text of A, beyond the information already present in the past text of A. Doing so is important when inferring information flow from data, as it is important to determine whether or not the information in B is redundant if one already has the information in A.[2,5,6]

B. Social information flow

In a previous work, we showed how to use the cross-entropy [Eq. (2)] as a measure of information flow between individuals posting to the Twitter.com social media platform.[21] We concatenated the texts of all public tweets for a given Twitter user into a long stream of text and then applied the aforementioned entropy and cross-entropy measures to users, pairs of users, and ego-centric networks consisting of users and their most frequent contacts. Measuring information flow with the cross-entropy naturally incorporates the temporal ordering of the tweet text and uses all the available information in the texts of the individuals, whereas other measurement methods limit themselves to proxies of information flow, such as tracking the spread of keywords like hashtags or URLs.

The focus of that work was on measuring information flow from text data. When developing and applying estimators in such scenarios, it is useful to have plausible models with which to build examples and test cases. However, most work modeling information flow has focused on the study of information as "packets" spreading between individuals, typically represented in Twitter's case by the hashtags or URLs. This allows researchers to apply contagion models, such as Susceptible-Infected or other compartmental models, complex contagion models, and more.[11,22–24] Contagion models are very well studied on network topologies, but in this case they neglect the dynamical processes governing written communication. The back-and-forth nature of discussions, for example, may generate far more information flow within the text than would be measurable from the spread of keywords alone.

II. THE QUOTER MODEL

We propose the "quoter model" as a simplified way to capture the dynamics governing the written conversations taking place between individuals in a social network. The model consists of N individuals embedded as the nodes $\mathcal{V}$ of a social network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $|\mathcal{V}| = N$ and there are $|\mathcal{E}| = M$ edges connecting those nodes. For generality we take the graph to be directed such that an edge $(i,j) \in \mathcal{E}$ represents communication from node j to node i via the quoting process described below.

Each member of the graph generates written text over time, represented as a symbolic time series or "word stream." At timestep t, individual i generates a number of new words according to one of the two mechanisms, growing his or her word stream. The number of new words at timestep t is $\lambda_i(t) \sim L_i(t)$, where this number is drawn from an integer-valued length distribution $L_i(t)$. This probability distribution may be time-independent or evolve as a function of time, and this distribution may vary across users ($L_i \neq L_j, j \neq i$) or not ($L_i = L_j \equiv L$). After choosing the number of words to generate, the actual words are generated according to one of the two mechanisms:

1. $\lambda_i(t)$ draws with replacement from a vocabulary distribution W_i (with probability $1 - q_{ij}$).
2. A contiguous sequence of $\lambda_i(t)$ words are copied from a random position within the previously written text of a neighbor j of node i (with probability q_{ij}).

This process is then repeated for all individuals in the network until their text streams have reached a desired length or a desired number of timesteps have elapsed. The first mechanism is intended to represent the creation of new content while the second mechanism is the quoter action of the model. The quote probabilities q_{ij} tune the relative strengths of the two mechanisms by how often i quotes from the past text of j. We illustrate one step of the model for a pair of individuals in Fig. 1.

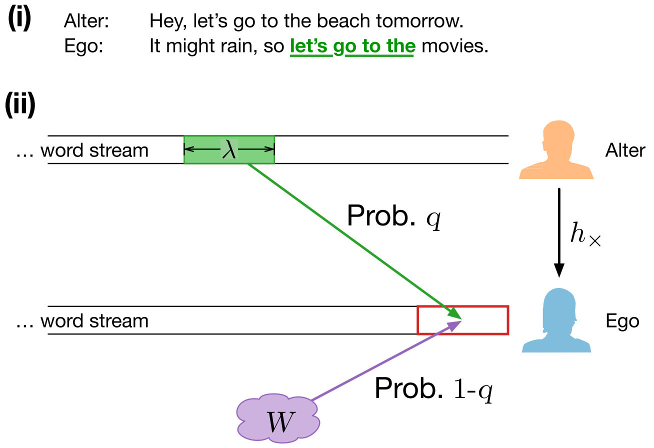
(i)
Alter: Hey, let's go to the beach tomorrow.
Ego: It might rain, so let's go to the movies.
(ii)
... word stream
λ
Alter
Prob. q
h×
... word stream
Ego
Prob. 1-q
W

FIG. 1. The quoter model for the social flow of information. **(i)** The repeated occurrences of short quoted passages such as this one throughout a written conversation indicate information flow. **(ii)** In the model, words are generated by individuals at each time step, forming word streams. To model information flow we use two mechanisms: at each timestep, with probability $1 - q$ the ego draws λ new words randomly from a specified vocabulary distribution W; otherwise, with probability q the ego copies a passage of length λ taken from a random position in the past words of the alter.

The idea underlying the second mechanism is that when two individuals are discussing a topic verbally or in writing, and they are listening to one another, then there will be a back and forth of small sequences of common words. The quotes generated by the second mechanism are not meant to capture full-length, long form quotations such as retweets, but instead short shared sequences of text. Alice: "That's the right way to go"; Bob: "No, this is the right way." In this example, the exchange between Alice and Bob leads to a short quotation of Alice by Bob ("the right way") and from this exchange only we can at least surmise that Bob is probably receiving and reacting to Alice's text. Of course, Bob could have responded in an equivalent way without that short quote. However, over the course of very long conversations we expect more such quotations to occur on average, and they will likely occur more often in conversations when there is more information flow than in conversations where there is little information flow.

A. Model components

The main components of the quoter model are (i) the graph topology, which may be as simple as a single directed link between two individuals, (ii) the quote probabilities q_{ij}, (iii) the length distributions L_i, and (iv) the vocabulary distributions W_i. We study several graph topologies in this work. The quote probabilities q_{ij} can be considered as edge weights on the social network, and there is considerable flexibility in assigning those weights.

The length distributions L_i govern the amount of text generated per timestep and the total length of the text: the expected length after t timesteps will be $\langle L \rangle \times t$. We primarily focus on two cases here, the constant length distribution $L(\lambda_t) = \delta_{\lambda\lambda_t}$, where δ_{ij} is the Kronecker delta; and a Poisson distribution $L(\lambda_t) = e^{-\lambda}\lambda^{\lambda_t}/\lambda_t!$ with mean λ.

The vocabulary distribution W_i gives the relative frequencies of words for individual i. In this work we consider two example W's. The first is a uniform distribution over a fixed number of z unique words: $W(w) = 1/z, w = 1, \ldots, z$. The binary case corresponds to $z = 2$. The second vocabulary distribution is a basic Zipf's law that incorporates the skewed distributions typically observed in real text corpora.[25] Here, the probability of a word w depends on its rank r_w, with the most probable word having rank $r_w = 1$. Zipf's law then defines word probabilities that obey a power law form with r: $W(w) \propto r_w^{-\alpha}$, where α is a power law exponent. This distribution is normalized by $H_{z,\alpha} = \sum_{r=1}^{z} r^{-\alpha}$, the generalized harmonic number. This distribution also holds for infinite vocabularies ($z \to \infty$) so long as $\alpha > 1$, in which case the normalization constant converges to the Reimann zeta $\zeta(\alpha)$.

III. MODEL ANALYSIS

Here, we study the basic quoter model between two individuals (referred to as the "ego" and the "alter") where the ego copies the alter but the alter does not copy the ego. We focus on the case of uniform vocabulary distribution $W(w) = 1/z, w = 1, \ldots, z$, and we assume both individuals draw from the same W, although our analysis is not specific to these assumptions.

To quantify the flow information from the alter to the ego via the cross-entropy $h_\times(\text{ego} \mid \text{alter})$, we need to compute the mean $\Lambda = T^{-1} \sum_{t=1}^{T} \Lambda_t$, where Λ_t is the length of the shortest substring of words beginning at position t in the ego's text which has not been observed in the text of the alter prior to "time" t (Sec. I), and T is the total length of the text. To model Λ_t, we assume that (i) two terms contribute to Λ_t: the mean Λ when a quote occurs (call it Λ_{quote}) and the mean Λ when no quote occurs (call it Λ_{random}) and (ii) the quote probability q weights these two possibilities:

$$\Lambda_t(\text{ego} \mid \text{alter}) = (1 - q)\Lambda_{\text{random}} + q\Lambda_{\text{quote}}, \tag{3}$$

where we have suppressed the dependence on position t in Λ_{random} and Λ_{quote}. We need to determine both Λ_{random} and Λ_{quote} as functions of the vocabulary distribution and the current amounts of text generated.

A. Prefix matches when not quoting

It is possible as the ego is drawing words from the vocabulary distribution that due to chance a string of words will be generated that previously appeared in the past text of the alter. This will depend on the vocabulary distribution and the length of the alter's past text.

Suppose the alter has posted a total of t words so far and the ego has just posted m new words. The probability that one of the new words posted by the ego matches a random word previously posted by the alter is $\sum_w W(w)^2 \equiv d$. This is the probability that two draws from the vocabulary distribution give the same word, irrespective of the particular word, and is the Simpson index (also known as the Herfindahl–Hirschman index) of the vocabulary distribution.[26,27] The probability of at least m new ego words matching with m prior alter words at a particular location in the alter's past text is d^m. Since there are approximately t locations in the alter's text at which a match may occur (assuming $t \gg m$), the expected number of matches of length m or more is $td^m \equiv C(m)$. Then, the expected length of the longest match m_* occurs at the value of $m = m_*$ for which $C(m_*) \geq 1$ and $C(m_* + 1) < 1$. Solving $C(m_*) = 1$ for m_* gives an expected longest match length of $m_* = \ln(t)/\ln(1/d)$, or

$$\Lambda_{\text{random}} = \frac{\ln(t)}{\ln(1/d)} + 1, \tag{4}$$

since Λ is always one more than the match length.

B. Prefix matches when quoting

If a quote of length λ occurs at position t, then $\Lambda_t = \lambda + 1$ only if any words of the ego subsequent to the λ quoted words do not happen to match the words of the alter subsequent to the original quoted passage. In other words, even if deterministically a match of length λ occurs, Λ_t may be longer due to chance. Specifically, the probability that $\Lambda_t = \lambda + 1 + m$, $m \geq 0$, is $d^m(1 - d)$, as a value of m requires that the next m words will match and the $(m + 1)$-th word will not match. Note that, unlike the previous calculations, this probability does not involve the total text length of the alter t because these post-quote matches cannot occur anywhere in the alter's text except in the positions following the quoted passage (neglecting duplicate passages). From this probability, the expected Λ_t is

$$\sum_{m=0}^{\infty} (\lambda + 1 + m)\, d^m(1 - d) = \lambda + 1 + \frac{d}{1 - d}, \tag{5}$$

meaning that, on average, random chance increases Λ_t by an amount $\frac{d}{(1-d)}$.

However, it is not necessarily reasonable to neglect duplicate passages. Indeed, the number of duplicate passages may be significant for certain combinations of parameters: the probability that a different location of the alter's past is the start of a passage of length λ equal to the randomly chosen quoted passage is d^λ, and the expected number of such duplicate passages within the alter's text (including the original passage) is $\approx td^\lambda + 1$. For $t = 10^4$, $d = 1/5$, and $\lambda = 3$, for example, the expected number of duplicates is 17.

The probability for at least m words of the ego's text subsequent to the newly quoted passage to also match m words following the original passage in the alter is d^m, so the expected number of times matches of length m or longer will occur following any of the duplicate passages in the alter is $\approx (td^\lambda + 1)d^m$. The longest match length m_* occurs at the value of m for which the number of these matches is ≈ 1, or $m_* = \ln(td^\lambda + 1)/\ln(1/d)$. Lastly, the expected total match length when quoting is then $\lambda + \ln(td^\lambda + 1)/\ln(1/d)$.

However, unlike with Λ_{random}, adding 1 to this expected total match length is not an accurate expression for the average Λ_{quote}. When $\lambda + \ln(td^\lambda + 1)/\ln(1/d)$ is much larger than Λ_{random}, then the match length Λ_t at that text position t will almost certainly be due only to the single quoted passage. This means that the subsequent Λ_{t+1} will likely be 1 fewer than Λ_t, because a random match that would extend Λ_{t+1} is unlikely. Likewise, $\Lambda_{t+2} = \Lambda_t - 2$, and so forth, until the match lengths are short enough that random matching is again probable. Accounting for this, we expect the average

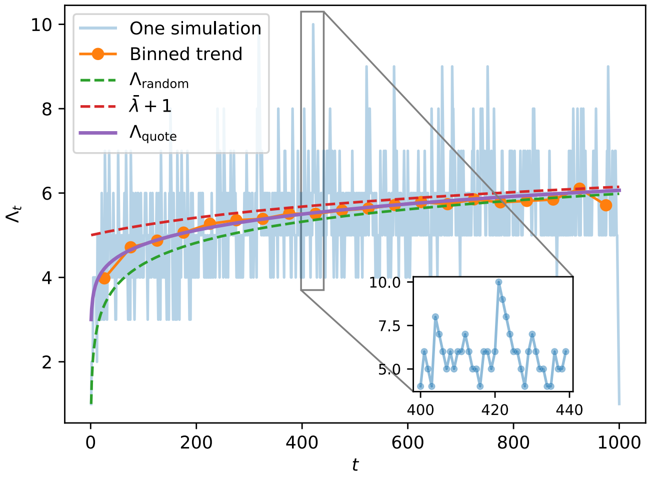

FIG. 2. Illustration of Λ_t when quoting to demonstrate the relationship between Λ_{random}, $\bar\lambda$, and Λ_{quote}. Here, we show a single realization of the model (parameters: $q = 1, z = 4, \lambda = 4$). Individual realizations show considerable variability so we also include a binned trend averaged over $n = 10$ realizations. This trend agrees well with Λ_{quote} for these parameters. The inset plot highlights a spike in Λ_t (at $t = 421$) and how it decays linearly back to the approximate level of Λ_{random}.

Λ_{quote} to be roughly equal to

$$\frac{1}{\bar\lambda - \Lambda_{\text{random}} + 2} \sum_{j=0}^{\bar\lambda+1-\Lambda_{\text{random}}} \left(\bar\lambda + 1 - j\right), \tag{6}$$

where $\bar\lambda = \lambda + \frac{\ln(td^\lambda+1)}{\ln(1/d)}$. Equivalently, this is the average of the two endpoints, $\bar\lambda + 1$ and Λ_{random}, and therefore:

$$\Lambda_{\text{quote}} = \frac{1}{2}\left[\lambda + \frac{\ln(td^\lambda + 1)}{\ln(1/d)} + \frac{\ln(t)}{\ln(1/d)} + 2\right]. \tag{7}$$

We illustrate the relationship between Λ_{random}, $\bar\lambda$, and Λ_{quote} in Fig. 2, showing a single simulation of the model and highlighting a spike in Λ_t above Λ_{random} and how it decays back down to Λ_{random}.

With these expressions for Λ_{random} and Λ_{quote}, we can now compute Λ and from it the cross-entropy.

C. Cross-entropy

To compute the cross-entropy $h_\times$ between the ego and alter requires computing the total Λ summed over all positions in the ego's text where matches can occur then dividing $T \ln T$ by that Λ: $h_\times = T \ln T/\Lambda$, where $\Lambda = \sum_{t=1}^{T} \Lambda_t$. Using the previously derived expected contributions to Λ for the two mechanisms and approximating the sum over the text positions with an integral give the following expression for Λ:

$$\begin{aligned}\Lambda &\approx \int_0^T \left[(1-q)\Lambda_{\text{random}} + q\Lambda_{\text{quote}}\right] dt \\ &= \frac{T}{\ln(1/d)} \left\{(1-q)\left(\ln\frac{T}{d} - 1\right)\right. \\ &\quad \left. + \frac{q}{2}\left[\ln\frac{T}{d^{\lambda+2}} + \left(\frac{1}{Td^\lambda} + 1\right)\ln(Td^\lambda + 1) - 2\right]\right\}, \end{aligned} \tag{8}$$

which can be substituted into $T \ln T/\Lambda$ to compute the cross-entropy as a function of q, λ, d, and T.

The limit of large text using Eq. (8) gives

$$\lim_{T\to\infty} h_\times(\text{ego} \mid \text{alter}) = \lim_{T\to\infty} \frac{T \ln T}{\Lambda} = \ln(1/d), \tag{9}$$

which is the Rényi entropy of the vocabulary distribution:

$$h_\alpha = \frac{1}{1-\alpha} \ln\left(\sum_w W(w)^\alpha\right), \tag{10}$$

with $\alpha = 2$. Note also that q has dropped out of this limit, implying that, given sufficient text, the entropy of the model will be that of the underlying vocabulary distribution only. However, as we shall see, for finite T, even quite large, q still plays an important role in the overall cross-entropy.

D. Comparison with simulations

To test our theoretical predictions, we simulate the quoter model and compared our predicted cross-entropy [substituting Eq. (8) into $T \ln T/\Lambda$ and converting to bits] with that computed directly from the simulations [Eq. (2) on the simulated text sequences]. We simulate the one-link, two-node model for 10^3 and 10^4 timesteps, giving expected text lengths of $T = 10^3\lambda$ and $T = 10^4\lambda$, respectively. Here, we choose for both nodes $W(w) = 1/z, w = 1, \ldots, z, L(t) = \text{Pois}(\lambda), q_{ij} = q$ and $q_{ji} = 0$ (denoting the ego as i and the alter as j). Overall, we find reasonable qualitative agreement between our predictions and the simulations, as shown in Fig. 3. However, there are some systematic discrepancies. While the absolute difference in entropies between predictions and simulations is small, often less than 0.1-0.2 bits, this means that the treatment above does not capture everything present in the model.

Beyond Fig. 3, which explores the cross-entropy as a function of q for different λ and $d = 1/z$ parameters, it is also useful to inspect the two limiting cases of no quotes ($q = 0$) and all quotes ($q = 1$). Figure 4 explores how the cross-entropy depends on d when $q = 0$. Since there are no quotes, we expect no dependence on λ and we indeed see strong collapse across the simulations and the theory (there is a slight difference between the curves only because the total length of the generated text depends on λ). Further, there is good agreement with predictions (solid lines) except at values of very low d (equivalently, high z). Agreement improves considerably at higher T although predicted values are still below those of the simulations. In this case, $h_\times$ depends entirely on Λ_{random}, and the expression for Λ_{random} [Eq. (4)] primarily gives only the *scaling* of Λ_{random} with accuracy.

The all-quote case is explored in Fig. 5. In this case, we expect a strong dependence on λ and indeed we see a change of more than two bits of cross-entropy at the lower diversity values when moving from $\lambda = 3$ to $\lambda = 9$. We also see good agreement between predictions and simulations except at low d, although in this case agreement improves considerably at low d for the longer text length.

Overall, we find that our treatment of the model captures the basic qualitative links between q, d, λ, and the total text length. Agreement is not perfect, indicating that more behavior is going on than is being modeled, particularly at low d, or entropy estimators based on Λ are biased for finite text, or some combination thereof. A more rigorous treatment

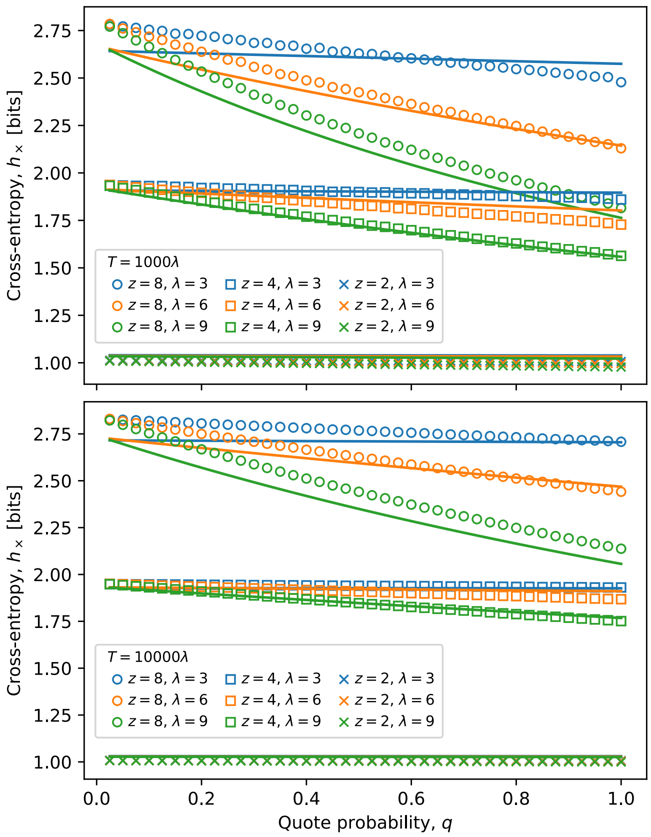

FIG. 3. The theoretical predictions (lines) give qualitative agreement with simulations (symbols), although there are systematic discrepancies, especially at lower vocabulary diversities $d = 1/z$.

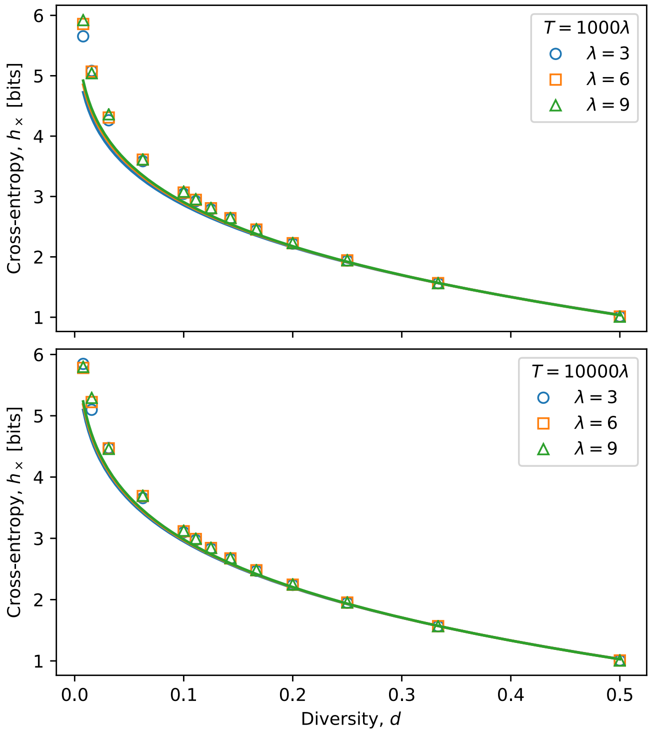

FIG. 4. The limiting case of $q = 0$ for different levels of vocabulary diversity $d = \sum_w W(w)^2$. There is reasonable agreement between cross-entropy using Eq. (8) and simulations except at low values of d. Agreement improves with larger T.

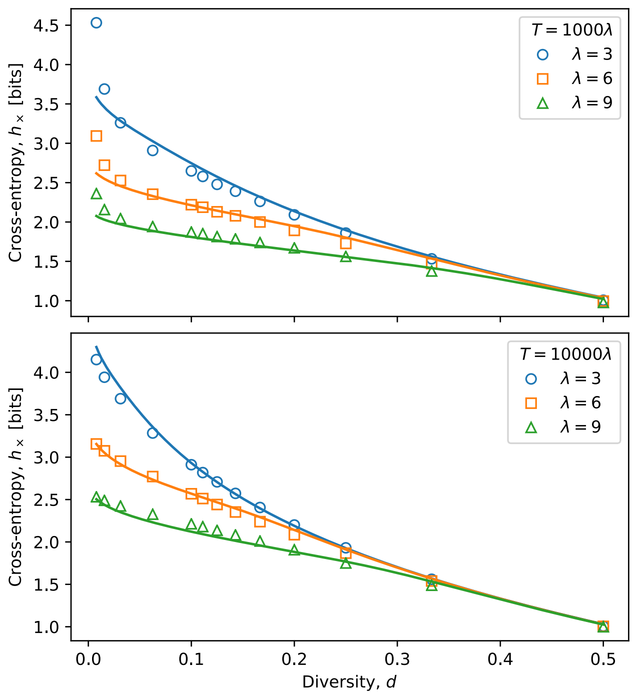

FIG. 5. The limiting case of $q = 1$ for different levels of vocabulary diversity d. Symbols denote simulations and lines denote predicted cross-entropy using Eq. (8). Agreement is reasonable in this case, and agreement improves for larger T.

of the model may be able to distinguish between these two possibilities and can extend the analysis to more complex arrangements than a single link between a pair of individuals.

IV. THE QUOTER MODEL ON NETWORKS

Moving beyond our treatment of a single pair of individuals (Sec. II), here we numerically investigate the quoter model on four simple network topologies (see Fig. 6): A *chain* of N nodes where each node copies from the previous node (i), a *fork* where one node influences two nodes (ii), a *collider* where a node is influenced by two nodes simultaneously (iii), and larger Erdős-Rényi and Barabási-Albert random graphs (iv) (not shown in Fig. 6). These topologies allow us to better understand, in a simplified context, the interplay between network topology and the dynamics of information flow as measured via the cross-entropy. The chain allows us to understand the attenuation of information flow with distance, the fork and the collider provide simple motifs to investigate confounds and spurious links, and the larger graph models can shed light on how global network properties such as density can affect information flow.

A. (i) Chain of quoters

We investigate the attenuation of information by simulating the quoter model over a unidirectional chain of nodes $v_0, v_1, \ldots, v_{N-1}$, where each node has probability q of quoting the node directly before them in the chain, except for the first node in the chain which only draws from W:

$$q_{ij} = \begin{cases} q & \text{if i > 0, i = j + 1;} \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

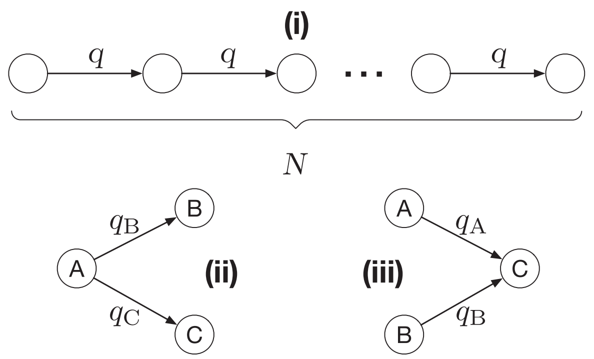

FIG. 6. Model network topologies. (i) Chain of N quoters, each with unidirectional quote probability q; (ii) Fork, with quote probabilities q_{B} and q_{C} for nodes B and C to copy A respectively; (iii) Collider, with quote probabilities q_{A} and q_{B} for C to copies A and B, respectively.

At each timestep, each node in the chain writes or quotes $\lambda_t \sim \mathrm{Pois}(\lambda = 3)$ words, each of which is then drawn from a 1000-word truncated Zipf distribution with exponent $\alpha = 3/2$. (Results were found to be very similar when using a uniform distribution with the same number of words.) We simulate the model on $N = 10$ nodes for 10,000 timesteps, so $T \approx 10,000\lambda$.

Figure 7 shows the cross-entropy of node i from the first node 0 in the chain, which generates original text. For reasonable values of the quote probability $q < 0.5$ information attenuates quickly, with $h_\times$ having saturated by approximately the third link in the chain. Only at very high quoting probabilities ($q = 0.95$) do we observe greater information flow (lower cross-entropy) for nodes further along the chain.

B. (ii) Fork and (iii) Collider

To investigate how cross-entropy distinguishes between information flow from different sources, we simulate the quoter model on the three-node "fork" and "collider" graph shown in Fig. 6. First, for the fork graph [Fig. 6(ii)], using the same parameters as above [$\lambda_t \sim \mathrm{Pois}(\lambda = 3)$, $w \sim \mathrm{Zipf}(z = 1000, \alpha = 3/2)$], we vary the probabilities q_{B} and q_{C} with which nodes B and C, respectively, copy the source node A, which generates original content (drawing words from W only). The top and bottom panels of Fig. 8 show the cross-entropy of C from A and of C from B, respectively, averaged over 1000 realizations of the model. As expected, $h_\times(\mathrm{C} \mid \mathrm{A})$ shows no dependence on q_{B} and decreases approximately linearly as the quote probability q_{C} grows (Fig. 8; top).

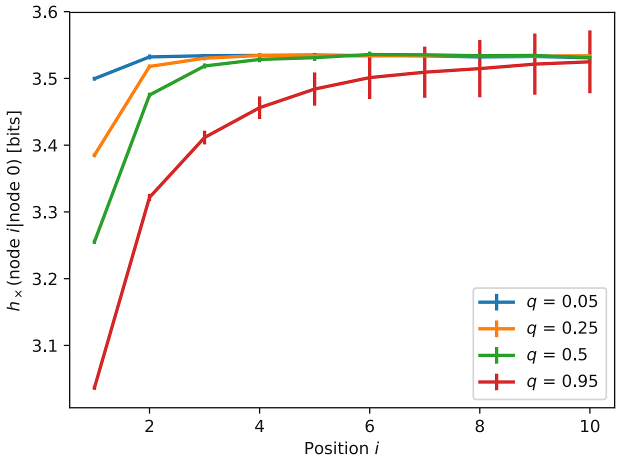

FIG. 7. Attenuation of information in a chain of quoters. Cross-entropy increases (information flow decreases) with both distance from the source node and decreasing quote probability q, generally saturating for $q \leq 0.5$ by a separation of no more than four steps.

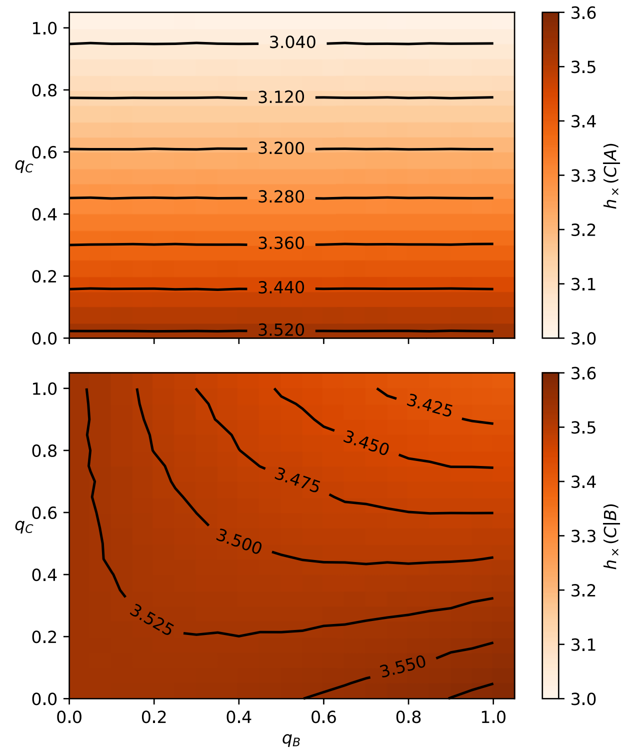

FIG. 8. Information flow on the fork graph as a function of the quote probabilities q_{B} and q_{C}. (Top) We find that the information flow from the source A to a target (either B or C; C shown only) depends only on the quote probability for the source-target link. (Bottom) The target-target link (B to C or C to B; B to C shown only) shows a mixed dependence on q_{A} and q_{B}. However, the cross-entropy values are higher than those observed for the source-target links, for most regions of the $(q_{\mathrm{C}}, q_{\mathrm{B}})$-space. We discretized q_{B} and q_{C} into steps of 0.05 and interpolated to obtain the level curves in the figures.

The dependence of C upon B in the fork is more complex, however, with the cross-entropy $h_\times(\mathrm{C} \mid \mathrm{B})$ of the non-existent link between B and C decreasing with both increasing q_{B} and q_{C} (Fig. 8; bottom). However, there exists a clear separation in the values of cross-entropy between the two cases, with $h_\times(\mathrm{C} \mid \mathrm{B})$ being significantly larger than $h_\times(\mathrm{C} \mid \mathrm{A})$ for most quote probabilities except the region where both q_{B} and q_{C} are close to 1. Cross-entropy therefore effectively identifies the direction of real information flow for this model graph.

Due to the fork's symmetry, the results for $h_\times(\mathrm{B} \mid \mathrm{A})$ and $h_\times(\mathrm{B} \mid \mathrm{C})$ are identical to those shown in Fig. 8. Likewise, the analogous $h_\times(\mathrm{C} \mid \mathrm{A})$ and $h_\times(\mathrm{C} \mid \mathrm{B})$ for the collider network topology [Fig. 6(iii)] appear similar to the top panel of Fig. 8: with no dependence between A and B in the collider, $h_\times(\mathrm{C} \mid \mathrm{A})$ decreases linearly with q_{A} and shows no dependence on q_{B} (not shown).

C. (iv) Random networks

Finally, we investigate the quoter model on larger networks, modeled as random graphs. We simulate the quoter model on Erdős-Rényi (ER)[28,29] and Barabási-Albert (BA)[30] random graphs. ER graphs are simple models that capture only the overall density of a network but are a useful starting point. BA graphs capture the "scale-free" property observed in real-life social networks. Using graphs of $N = 100$, we create directed, weighted networks of varying average node degree.[31] To create directed ER networks, we chose pairs of nodes i and j and created an edge from i to j with probability p. For the BA networks, we used the standard preferential attachment method with edges pointing in both directions. This construction means that quoting is always bidirectional in the BA networks, but not necessarily in the ER networks. Other options are possible for the BA network, e.g., creating directed links from newer nodes to older nodes through the preferential attachment process; however, this would have rendered these networks a directed tree rather than graph, as was desired here.

Quote probabilities q_{ij} are chosen from $U(0,1)$, with $q_{ii} \sim U(0,1)$ representing the probability of a node generating new content (after normalizing such that $q_{ii} + \sum_j q_{ij}A_{ij} = 1$, where A is the adjacency matrix of the graph). The quoter model is then run for $5000N$ timesteps over the network, updating a randomly chosen node at each timestep, and using the same vocabulary (W) and quote-length (L) distributions as above. At the end of the simulation each node has generated text of length $T \approx 5000\lambda = 15,000$ words. We simulate 100 realizations of the network and quoter model dynamics on both the ER and BA networks.

Information flow on these graphs as a function of the graph's average node degree $\langle k\rangle$ is shown in Fig. 9. As average degree increases in the network, the average cross-entropy of a node i from its neighbors j also increases, meaning that i becomes less predictable from its neighbors with increasing density. The BA graphs show slightly lower median cross-entropy, however, with larger variation across realizations. The presence of high-degree hubs in BA graphs means that cross-entropy can exhibit a larger range of variation, with the self probability q_{ii} at hub nodes i to generate new content driving much of the information flow on the network. The increasing trend of cross-entropy with average node degree indicates that information "sources" and "sinks" become increasingly difficult to identify in a network, as the density of connections increases.

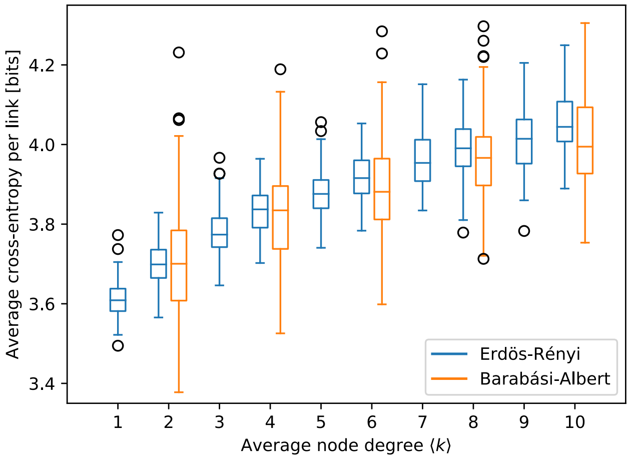

FIG. 9. Average information carried by edges in a network decreases as network density increases, as evidenced by the increase in cross-entropy. Both types of networks contain 100 nodes, and boxes represent the distribution of cross-entropy over 100 realizations each. (Boxplots have been shifted left and right where they would otherwise overlap, for clarity.)

V. DISCUSSION

In this paper, we introduced the quoter model as a simple, paradigmatic model of the flow of information. Considerable effort has been put into measuring information flow in online social media, both from proxies such as tracking keywords and from information-theoretic tools. Models of the dynamics underlying these processes are invaluable for better understanding information flow, and the goal of our work is to introduce a model that more directly relates to information flow in text data than traditional contagion-style models, but without being overly complicated. Our model mimics at a basic level the overall dynamics of text streams posted online, and here we showed that one can derive expressions for the information flow between written texts as measured via the cross-entropy.

The analysis we performed here showed good qualitative agreement with simulations in general, but there remains room for improvement. Nevertheless, the ability to find tractable expressions for information-theoretic quantities highlights how the basic quoter model can provide better insights into information flow over social networks. Indeed, we proposed this model because empirical benchmarks for information flow over social networks are difficult to find. However, as many dynamic processes can be represented by symbolic time series, models like the quoter model may even be useful when studying information flow in more general contexts.

The language generator we studied here is a relatively simplistic bag-of-words model: individuals simply draw words from a given vocabulary distribution W. More realistic models should be explored. One possibility would be a time-dependent W. For example, one could endow W with a latent context C: $W(w \mid C)$ and allow the context to vary (slowly) over a space of contexts. A Markov chain over this context space would be one way to introduce dynamic context shifts. Such a context dependence can then be used to model topical shifts over the length of a discourse. If two users exhibit the same context shifts, their vocabulary distributions will tend to "sync up" "with each other", and this should lead to a lower cross-entropy than if contexts were not shared.

This dynamic context shift in quoted discourse suggests a natural time-based generalization to the model as well. With quoting behavior likely to occur within a short "attention span" of the time of the original message, it makes sense to incorporate a probability of quoting into the model which decays over time. While the form of this probability likely introduces an extra parameter, it is plausible that this parameter could be estimated from real data. Future work will explore the possibility of fitting the quoter model to real datasets.

Lastly, there is much room for future exploration of network topology and its relationship to information flow. As the quoter model allows us to design "planted" interactions, we can implement the quoter dynamics on constructed networks and then test whether algorithms can successfully infer true interactions and reject spurious interactions. We did this here with the fork and collider graphs. Moving beyond those small motifs, one area of network structure worth exploring in future work is that of network topologies exhibiting clustering, to investigate the effect of community structure[32] on information flow.

ACKNOWLEDGMENTS

We gratefully acknowledge the resources provided by the Vermont Advanced Computing Core at the University of Vermont and the Phoenix HPC service at the University of Adelaide. This material is based upon work supported by the National Science Foundation (Grant No. IIS-1447634). 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).

[1] E. M. Bollt and J. Sun, “Editorial comment on the special issue of ‘Information in dynamical systems and complex systems’,” Entropy **16**, 5068–5077 (2014).

[2] T. Schreiber, “Measuring information transfer,” Phys. Rev. Lett. **85**, 461 (2000).

[3] M. Staniek and K. Lehnertz, “Symbolic transfer entropy,” Phys. Rev. Lett. **100**, 158101 (2008).

[4] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch, “Detecting causality in complex ecosystems,” Science **338**, 496–500 (2012).

[5] J. Sun and E. M. Bollt, “Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings,” Physica D **267**, 49–57 (2014).

[6] J. Sun, D. Taylor, and E. M. Bollt, “Causal network inference by optimal causation entropy,” SIAM J. Appl. Dyn. Syst. **14**, 73–106 (2015).

[7] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, Information diffusion through blogspace, in *WWW New York, NY, May 17–20, 2004* (ACM, New York, NY, 2004), pp. 491–501.

[8] J. L. Iribarren and E. Moro, “Impact of human activity patterns on the dynamics of information diffusion,” Phys. Rev. Lett. **103**, 038702 (2009).

[9] G. Ver Steeg and A. Galstyan, Information transfer in social media, in *WWW, Lyon, France, April 16–20, 2012* (ACM, New York, NY, 2012), pp. 509–518.

[10] J. Borge-Holthoefer, N. Perra, B. Gonçalves, S. González-Bailón, A. Arenas, Y. Moreno, and A. Vespignani, “The dynamics of information-driven coordination phenomena: A transfer entropy analysis,” Sci. Adv. **2**(4), e1501158 (2016).

[11] R. S. Burt, “Social contagion and innovation: Cohesion versus structural equivalence,” Am. J. Sociol. **92**, 1287–1335 (1987).

[12] D. J. Watts, “A simple model of global cascades on random networks,” Proc. Natl. Acad. Sci. U.S.A. **99**, 5766–5771 (2002).

[13] A. Vespignani, “Modelling dynamical processes in complex sociotechnical systems,” Nat. Phys. **8**, 32–39 (2012).

[14] In this work we consider the word-level entropy rate, but it is also common to work on a per-character basis.

[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, New York, 2012).

[16] C. E. Shannon, “Prediction and entropy of printed English,” Bell Syst. Tech. J **30**, 50–64 (1951).

[17] W. Ebeling and T. Pöschel, “Entropy and long-range correlations in literary English,” EPL **26**, 241 (1994).

[18] T. Schürmann and P. Grassberger, “Entropy estimation of symbol sequences,” Chaos **6**, 414–427 (1996).

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

[20] J. Ziv and N. Merhav, “A measure of relative entropy between individual sequences with application to universal classification,” IEEE Trans. Inf. Theory **39**, 1270–1279 (1993).

[21] J. P. Bagrow, X. Liu, and L. Mitchell, Information flow reveals prediction limits in online social activity, (2017), arXiv:1708.04575.

[22] L. Weng, A. Flammini, A. Vespignani, and F. Menczer, “Competition among memes in a world with limited attention,” Sci. Rep. **2**, 335 (2012).

[23] J. L. Toole, M. Cha, and M. C. González, “Modeling the adoption of innovations in the presence of geographic and media influences,” PLoS ONE **7**, e29528 (2012).

[24] S. Melnik, J. A. Ward, J. P. Gleeson, and M. A. Porter, “Multi-stage complex contagions,” Chaos **23**, 013124 (2013).

[25] G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (Addison-Wesley, Reading, MA, 1949).

[26] The vocabulary distribution $W(w)$ enters into our analysis through its diversity d, the probability of drawing the same word twice irrespective of the particular word w. In the case of text obeying Zipf’s law, the diversity is $d = \zeta(2\alpha)/\zeta(\alpha)^2$, where ζ is the Reimann zeta function. For $\alpha = 3/2$, which gives a reasonable approximation for the vocabulary distribution of real text[33] although there is evidence that Zipf’s law is too simplistic for real text corpora,[33,34] we find that $d \approx 0.176$. Note that when Zipf’s law is defined over a finite vocabulary of z words, the diversity becomes $d = H_{z,2\alpha}/(H_{z,\alpha})^2$. Again for the plausible exponent $\alpha = 3/2$, and a vocabulary size of $z = 1000$ total words, this gives $d \approx 0.185$, slightly higher than the infinite case. These two examples show that real vocabulary can tend to relatively high values of d despite the large number of words, because the probabilities for those words are so heavily skewed.

[27] The diversity can also account for individuals with different vocabulary distributions. In that case, $d = \sum_w W(w)^2$ becomes $d = \sum_w W_i(w)W_j(w)$. This only requires that both distributions are defined over the same set of words, which can always be achieved simply by taking any words missing in W_i (W_j) but present in W_j (W_i) and including them in W_i (W_j) with a probability of zero.

[28] P. Erdős and A. Rényi, “On random graphs. I,” Publ. Math. **6**, 290–297 (1959).

[29] P. Erdős and A. Rényi, “On the evolution of random graphs,” Publ. Math. Inst. Hung. Acad. Sci. **5**, 17–60 (1960).

[30] A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” Science **286**, 509–512 (1999).

[31] This was done for ER graphs by varying the probability p for each link to exist. For BA graphs, we varied the attachment parameter m, which leads to only even-valued average degrees (see Fig. 9).

[32] M. Girvan and M. E. Newman, “Community structure in social and biological networks,” Proc. Natl. Acad. Sci. U.S.A. **99**, 7821–7826 (2002).

[33] J. R. Williams, J. P. Bagrow, C. M. Danforth, and P. S. Dodds, “Text mixing shapes the anatomy of rank-frequency distributions,” Phys. Rev. E **91**, 052811 (2015).

[34] J. R. Williams, P. R. Lessard, S. Desu, E. M. Clark, J. P. Bagrow, C. M. Danforth, and P. S. Dodds, “Zipf’s law holds for phrases, not words,” Sci. Rep. **5**, 12209 (2015).