MDPI

Article

Complex contagion features without social reinforcement in a model of social information flow

Tyson Pond [1], Saranzaya Magsarjav [2], Tobin South [2], Lewis Mitchell [2] and James P. Bagrow [1,*]

[1] University of Vermont
[2] University of Adelaide
* Correspondence: james.bagrow@uvm.edu

Version February 1, 2020 submitted to Entropy

Abstract: Contagion models are a primary lens through which we understand the spread of information over social networks. However, simple contagion models cannot reproduce the complex features observed in real world data, leading to research on more complicated complex contagion models. A noted feature of complex contagion is social reinforcement, that individuals require multiple exposures to information before they begin to spread it themselves. Here we show that the quoter model, a model of the social flow of written information over a network, displays features of complex contagion, including the weakness of long ties and that increased density inhibits rather than promotes information flow. Interestingly, the quoter model exhibits these features despite having no explicit social reinforcement mechanism, unlike complex contagion models. Our results highlight the need to complement contagion models with an information-theoretic view of information spreading to better understand how network properties affect information flow and what are the most necessary ingredients when modeling social behavior.

Keywords: online social networks; social media; information spreading; information diffusion; cross-entropy.

1. Introduction

Social networks mediated through online platforms are an increasingly important way in which individuals send and receive information, and their influence is now felt in economics, politics, and the workplace [1–6]. These platforms provide rich opportunities for researchers to collect and study real-world data related to human behaviour and the spread of information. In concert with these datasets, considerable research has worked towards better statistical and information-theoretic tools to quantify information flow [7–9] and towards more accurate mathematical models to understand and even predict information flow [10–12].

A common approach to measuring information flow over a network is to idealize information as a collection of 'packets,' and then track the spread of those packets throughout the network. This approach is especially common when studying social media where keywords such as hashtags or URLs are easily tracked. More complex phenomena, such as the adoption of behaviors can also be monitored and used as a proxy for information flow [13]. Treating information flow in this way brings to mind the spread of infections and the use of epidemiologically-inspired models is popular. In this context, the social "diffusion" of information is often characterized as either a simple contagion or a complex contagion [14]. Simple contagions are those where each exposure can independently lead to an infection. Complex contagions, in contrast, introduce a social reinforcement mechanism where multiple exposures are needed before the contagion can spread.

However, despite its simplicity and popularity, there can be drawbacks to treating information as the contagion of discrete packets. Within social media, for example, there is a wealth of written

information being posted by users that is ignored when focusing only on particular keywords. Likewise, considerable information could be exchanged between individuals without leading to an observable adoption of behavior. Therefore, we argue in this work that a more nuanced approach grounded in information theory can give a better view of information flow in online social networks while more fully utilizing the available data.

The goal of this work is to study how network properties can affect information flow when taking an information-theoretic view on information flow, and how this information-theoretic view compares to contagion. We study the quoter model [12], a simple model for individuals generating text data within social media and apply information-theoretic estimators to the model text. Using both network models and real world network data, we compare the behavior of information flow in this model with traditional simple and complex contagion, to see the similarities and differences we may observe through these contrasting viewpoints. Interestingly, we find that the quoter model exhibits several phenomena characteristic of complex contagion, despite lacking an explicit social reinforcement mechanism, the key feature of complex contagion.

The rest of this work is organized as follows. in Sec. 2 we describe information-theoretic estimators of information flow and mathematical models of information flow and contagion. In Sec. 3 we describe the materials and methods used in this study, including simulation details, measures of information flow, the network properties we investigate, and the network data we use. Section. 4 presents our results comparing contagion models with the information-theoretically-motivated quoter model and exploring how various network properties affect information flow in the quoter model. We conclude with a discussion in Sec. 5.

2. Background

2.1. Measuring information flow

Suppose an individual within a social network generates a stream of text representing posts shared online on Twitter, for example. The entropy rate h of this text captures the information present within it. It can be challenging to estimate h for natural language data as information is present in the ordering of the word, not just the relative frequencies of words [15]. To help address this challenge, Kontoyianni et al. [16] 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. This estimator has been used to study human dynamics including mobility patterns and social media predictability [11,17].

Equation (1) generalizes to an estimator of the **cross-entropy** $h_\times$ between two texts A and B [11,18]:

$$\hat{h}_\times(A \mid B) = \frac{T_A \log_2 T_B}{\sum_{i=t}^{T_A} \Lambda_t(A \mid B)}, \tag{2}$$

where T_A and T_B are the lengths of the two texts, and $\Lambda_t(A|B)$ is the length of the shortest substring starting at position t of text A not previously seen in text B. Previously, in this case, refers to all the words of B written prior to the time when the tth word of A was written. The cross-entropy can be applied directly to the texts of a pair of individuals by choosing B to be the text stream of one individual and A the text stream of the other, and Eq. (2) can be used to measure the information flow between those individuals by asking how much predictive information about one text is contained within the other. This can be a quite powerful and effective measure of information flow, as it satisfies temporal precedence of the text streams and it uses all of the available (text) data for the pair of users.

We focus on the cross-entropy estimated using Eq. (2) as a pairwise measure of information flow, but generalizations can capture information flow from multiple social ties towards a single individual [11,12]. Doing so allows for measures of more complex information flow such as analogs of transfer entropy or causation entropy [7,8,19]. The best extensions of information flow estimators beyond pairwise measures remains an active and fruitful area of research (see also our discussion in Sec. 5).

2.2. Quoter model

To study the effects of network properties on information flow, we use the recently proposed quoter model [12]. The quoter model represents an idealized model of social conversations, meant to capture some of the processes by which individuals in an online social network post text while also being analytically tractable. Nodes in a network generate text streams both by sampling from a given vocabulary distribution and by copying ("quoting") short sub-sequences of text from their neighbors. This model provides a parameter q, the quote probability, that tunes the degree of information flow. (Full details of the model and how we simulate it are given in Sec. 3.1.) After simulating the quoter model for a given number of time steps, a text stream has been generated by each node in the network, and we an estimate the cross-entropies between these texts to study the social flow of written information. See Bagrow and Mitchell [12] for full details on the quoter model.

2.3. Other models of information flow

Contagion approaches are often used to model information flow [14]. A classic simple contagion approach to information flow is compartment models, taken from models of epidemics. Two simple compartment models are Susceptible-Infected (SI) and Susceptible-Infected-Recovered (SIR) models. On a network, a small number of nodes are initially "infected" while the remaining nodes are susceptible. The contagion then spreads from those infected nodes with a constant transmission rate per link so that each node in the "S" compartment has a constant probability to move to the "I" compartment with any given exposure. For SIR models, an additional "R" compartment is used to model a recovery process where infected nodes cease spreading the contagion while also becoming immune to reinfection. Many variants on these models exist.

Complex contagion phenomena are typically captured with threshold models [20,21]. Here nodes are again labeled as susceptible or infected, but the probability for a node i to become "infected" is a function of the number of neighbors of that node already infected. Too few neighbors and there is zero probability that i will be infected. Yet if a sufficient fraction of i's neighbors become infected, then i has a non-zero probability of becoming infected. This *social reinforcement* mechanism is intended to capture the cognitive mechanisms underlying opinion change, knowledge acquisition, and other facets of how individuals respond to and adopt information and ideas [22,23].

Complex contagion leads to several phenomena that differ from simple contagion. For one, there is an interesting *cascade window* where network density leads to a non-monotonic relationship with the spread of the contagion. Often denser networks lead to less spread, unlike simple contagion where a contagion will spread more easily as denser networks afford more opportunities (links) for spreading. Another feature of complex contagion is the complicated role of clustering where clustering can appear to either promote or inhibit contagion [24–26]. Complex contagion also exhibits a "weakness of long ties" effect, where long ties impede the flow of contagion [27], in contrast with the seminal "strength of weak ties" result [28] that implies long-range ties have an out-sized role in promoting information flow. The goal of our work here is to study the information-theoretic view of information flow we adopt here with the quoter model and compare to the effects of complex contagion that is commonly used as a *non*-information-theoretic view to study information flow.

3. Materials and methods

In this study, we use the quoter model on networks to elucidate the role of network structure on information flow. Here we describe the procedures to simulate the quoter model, measure information flow between nodes in networks, we describe the network features we study in relation to information flow, and we provide the details on the network models (random graphs) and real world network datasets we study.

3.1. The quoter model

We use the following process to simulate the quoter model on a given network. The quoter model requires a directed graph $G = (V, E)$ and, in the most general case, quote probabilities q_{uv} on each directed edge (we say node v (ego) may quote u (alter) if the edge $u \to v$ exists and has $q_{uv} > 0$). We simplify this for our simulations: when an ego generates new text, with probability q (bidirectional quoting) we pick an alter (predecessor) uniformly at random to quote from; otherwise, with probability $1 - q$ the user generates new content. If an ego quotes an alter, then they copy a random segment of the alter's past text and append this onto their growing text stream. We take the "quote length" (number of words) being copied to be Poisson-distributed with mean λ) for all users; otherwise, the ego generates new content randomly by sampling with replacement from a vocabulary distribution $W(w)$ and appending those samples onto their growing text stream, where the number of samples is again Poisson-distributed with mean λ. We assume a common, fixed vocabulary distribution $W(w)$ that follows a Zipf law of word use, as in prior studies and motivated by real world language usage patterns [12]. Specifically, a Zipf law defines the probability of using word w to be a power law based on the rank r_w of w: $W(w) = H_{z,\alpha}^{-1} r_w^{-\alpha}$, where z is the vocabulary size and $H_{z,\alpha} = \sum_{r=1}^{z} r^{-\alpha}$. Here we take $z = 1000$ as in [12] and, unless otherwise stated, focus on the exponent $\alpha = 1.5$, an exponent often-observed in social media data. We focus in this work on $q = 1/2$ and $\lambda = 3$ but we explore the robustness of our results to other parameter choices in App. A. This process repeats for $T = 1000N$ time steps so that each user has generated approximately $1000\lambda = 3000$ words when complete.

3.2. Measuring information flow over the network

After generating text streams for all nodes in G by iterating the quoter model, the cross-entropy estimator (Eq. 2) is then applied as needed to compute $h_\times$. We compute the cross-entropy over all edges, $\{h_\times\} = \{h_\times(u \mid v) \mid (u, v) \in E\}$, and report the mean $\langle\{h_\times\}\rangle$ and variance $\mathrm{Var}(\{h_\times\})$ of these values. Likewise, the predictability Π, given by Fano's Inequality [29], is a functionally-equivalent measure of information flow (as we assume the same vocabulary sizes for nodes in the quoter model). We focus on link-based cross-entropies although the cross-entropy estimator can be applied to non-neighboring nodes. Indeed, when studying the role of community structure in modular networks (see Sec. 3.4), we also consider cross-entropies between nodes in different modules, to assess information flow between and within said modules.

3.3. Simulating contagion models

To compare and contrast information flow in the quoter model, we also simulate traditional models of information flow, specifically simple and complex contagion. For simple contagion we simulate a stochastic SIR model on different networks (1000-node Erdős-Rényi and Barabási-Albert networks, as well as a sample of real world networks) using [30]. For the simulations here we set the transmission rate 20 and recovery rate 1. We initialize with a random 5% of the nodes infected, and run 10 outbreaks on 100 realisations of the network for each choice of average degree $\langle k \rangle$. For complex contagion we use exactly the same parameters, except we introduce a threshold function for transmission as in [21], where the transmission rate is set to zero if the proportion of infected neighbors is below some threshold ϕ (and we set $\phi = 0.18$ following [21]). For all simple and complex contagion

simulations we measure the peak outbreak size, noting that larger outbreak sizes conventionally correspond to greater information flow.

3.4. Assessing the impact of structure on dynamics

In this work we use a number of network models (random graphs) tailored to control for various network properties such as density, clustering, and modular structure. Here we describe the models and properties we study in relation to information flow in the quoter model.

Density and average degree

To explore how network density relates to information flow, we create Erdős-Rényi and Barabási-Albert networks of N nodes with varying average degree, $\langle k \rangle$, allowing us to the tune their densities. For the Erdős-Rényi networks we add edges independently with probability $p = \langle k \rangle / (N-1)$. For the Barabási-Albert model we start with $m = \langle k \rangle / 2$ nodes with no edges and add nodes which each form m links with previous nodes according to preferential attachment. Here we measure how cross-entropies varies with the densities of the networks using their average degree $\langle k \rangle$ and edge density $M/\binom{N}{2}$ where M is the total number of edges in the network. To complement the Erdős-Rényi and Barabási-Albert results, we also compare the densities of real networks with their average cross-entropy.

Degree heterogeneity

To assess the role of degree heterogeneity on information flow, we study the simplest random graph model with tunable degree heterogeneity, termed "dichotomous networks" in [31]. Dichotomous networks are generated via the configuration model. They have only two types of nodes – those with degree k_1 and those with degree k_2. We assume there are $N/2$ nodes of each degree and fix $k_1 + k_2$ so that the average degree is fixed. The mean and variance of the degree distribution, respectively, are given by $\mu = \frac{1}{2}(k_1 + k_2)$ and $\sigma^2 = (k_1 - k_2)^2/4$. We are interested in how the cross-entropy varies with k_1/k_2. When $k_1/k_2 = 1$ the network reduces to a random k-regular graph ($\sigma^2 = 0$), while $\sigma^2 \to \infty$ as $k_1/k_2 \to 0$.

Clustering

Clustering or triadic closure, the tendency towards forming triangles, is a key feature of social networks. We studied clustering using a network model with tunable numbers of triangles and with a randomization procedure that can lower the number of triangles in an existing network. We quantify a network's clustering using *transitivity* $T(G)$, the fraction of possible triangles in the network which actually exist: $T(G) = 3N_{\text{triangles}}/N_{\text{triads}}$, where $N_{\text{triangles}}$ counts the number of triangles in the network and N_{triads} is the number of triads or paths of length 2.

We constructed "small-world" networks using the Watts-Strogatz (WS) model [32] to tune their clustering. We generated a one-dimensional periodic lattice of N nodes with k nearest-neighbor connections, and randomly rewired lattice edges with a rewiring probability p. Varying the rewiring probability p allows us to tune the network diameter and clustering.

While the Watts-Strogatz model lets us generate networks with different clustering values, a generic challenge when assessing the impact of clustering (and other network properties) on dynamics is generating networks with tunable clustering, but for which other structural properties, such as density or diameter, can be controlled for. To study the relationship between transitivity and information flow, we apply the established degree-preserving stochastic rewiring or "x-swap" method [33–35], in which we repeatedly choose two links at random and two randomly selected endpoints of those links are swapped as long as the number of links does not change by swapping and the network does not become disconnected. These swaps lower transitivity while fixing the number of links and degrees of all nodes in the network. We performed $5M$ swaps for each real network.

Examining information flow on the randomized network compared with information flow on the original network can then illustrate what effect, if any, transitivity had on information flow.

Community structure and modularity

Community structure is another inherent property of social networks. It is commonly quantified using modularity [36]:

$$Q = \frac{1}{2M} \sum_{i,j} \left(a_{ij} - \frac{k_i k_j}{2M} \right) \delta(c_i, c_j),$$

where M is the total number of links, the sum runs over all pairs of nodes in the network, $\mathbf{A} = [a_{ij}]$ is the adjacency matrix of the network, k_i is the degree of node i, δ is the Kronecker delta, and c_i denotes the community containing i. The community structure encoded in the $\{c_i\}$ can be found using a community detection algorithm or it may be planted within a network model. To investigate community structure within a network model, we examined instances of the stochastic block model (SBM) [37,38] with N nodes and two planted blocks, or groups of nodes, denoted A and B, of equal size $m = N/2$. Here there are two connection probabilities: p_0 (the within-block connection probability) and p_1 (the between-block connection probability) governing the probability for a link to form between nodes in the same block and in different blocks, respectively. The expected modularity in this two-block stochastic block model is

$$Q = \frac{1}{2} \left(\frac{p_0 - p_0 m + p_1 m}{p_0 - p_0 m - p_1 m} \right).$$

Our main quantities of interest are the average cross-entropy on within-block edges, $\langle h_\times(\text{within}) \rangle$, the average cross-entropy on between-block edges $\langle h_\times(\text{between}) \rangle$ and their difference, $\Delta h_\times \equiv \langle h_\times(\text{between}) \rangle - \langle h_\times(\text{within}) \rangle$. These quantities describe to what extent information flows within and between communities.

We also computed modularity for real networks using the Louvain method [39]. The Louvain method is a hierarchical community detection algorithm that finds a partition of nodes that maximizes modularity Q. As commonly done, we initialize each node in its own community.

Multiple vocabulary distributions

A recent study [40] showed that heterogeneity in the dynamical parameters can be as important as structural heterogeneity. Communities offer an obvious way to implement such heterogeneity: We also investigate a two-block SBM where we distinguish the two groups A and B by giving them different Zipf exponents α_A, α_B, respectively, for their vocabulary distributions.

3.5. Network datasets

To supplement the above graph models, we also studied contagion and quoter model dynamics on real-world networks. We developed a corpus of 10 social networks spanning a range of sizes and densities that were used as the basis for simulation. See App. B for details on network sources and processing. Table 1 shows several descriptive statistics for the networks we analyzed.

Table 1. Descriptive statistics for real world networks used in this study. ASPL: Average Shortest Path Length. Modularity computed using the Louvain method [39].

Network	$\|V\|$	$\|E\|$	$\langle k \rangle$	Density	Transitivity	ASPL	Modularity	Assortativity
Sampson's monastery	18	71	7.9	0.464	0.53	1.54	0.29	−0.07
Freeman's EIES	34	415	24.4	0.740	0.82	1.26	0.07	−0.15
Kapferer tailor	39	158	8.1	0.213	0.39	2.04	0.32	−0.18
Hollywood music	39	219	11.2	0.296	0.56	1.86	0.20	−0.08
Golden Age	55	564	20.5	0.380	0.53	1.64	0.45	−0.13
Dolphins	62	159	5.1	0.084	0.31	3.36	0.52	−0.04
Terrorist	62	152	4.9	0.080	0.36	2.95	0.52	−0.08
Les Miserables	77	254	6.6	0.087	0.50	2.64	0.56	−0.17
CKM physicians	110	193	3.5	0.032	0.16	4.24	0.61	0.11
Email Spain	1133	5452	9.6	0.009	0.17	3.61	0.57	0.08

4. Results

Here we compare information flow in the quoter model with traditional simple and complex contagion (Sec. 4.1), then investigate how degree heterogeneity (Sec. 4.1), clustering (Sec. 4.2) and network modularity (Sec. 4.3) affect information flow. We also study how heterogeneity in the parameters affects information flow compared to the effects of network structure (Sec. 4.4).

4.1. Information flow and models of contagion

A distinguishing feature of simple and complex contagion is that denser networks lead to higher spreading for simple contagion and lower spreading (mostly) for complex contagion. We illustrate this difference using simulations in Figs. 1A,B. The decrease in spreading in complex contagion is due to its social reinforcement mechanism: it is more difficult for a contagion to spread when egos have many alters as more alters must adopt the contagion before the ego does. Yet we see in Fig. 1C that the quoter model, which lacks an explicit social reinforcement mechanism, also exhibits lower information flow at higher density. Here we measure information flow using predictability on links (Sec. 3.2), which is functionally equivalent in our simulations to the cross-entropy $h_\times$ (Fig. 1C inset). These results also hold on our corpus of real world networks (Fig. 2).

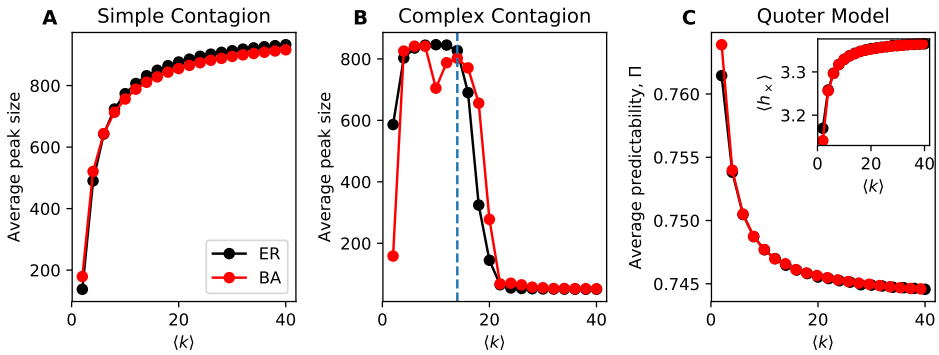

Figure 1. Denser networks are associated with higher information flow for simple contagion but lower information flow for both complex contagion and the quoter model. Here density is measured by average degree $\langle k \rangle$ for Erdős-Rényi (ER) & Barabási-Albert (BA) model networks. (**A**) Simple contagion. (**B**) Complex contagion (**C**) Quoter model. (Panel C, inset) Average cross-entropy on links; higher cross-entropies correspond to lower predictabilities and lower information flow. Networks consisted of $N = 1000$ nodes and each point constitutes 200 simulations.

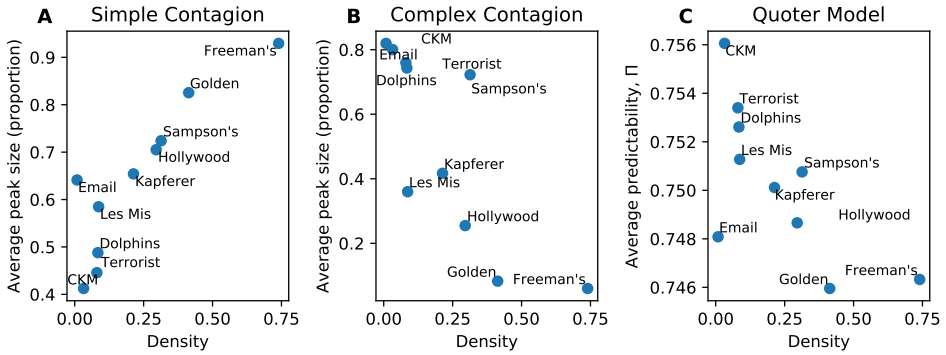

Figure 2. Information flow on real world networks. (**A**) Simple contagion. (**B**) Complex contagion. (**C**) Quoter model. Here information flow measures (average peak size, average text predictability) are compared to network density $M/\binom{N}{2}$. The association between information flow and density, either positive (simple contagion) or negative (complex contagion, quoter model), is significant (Wald test on non-zero regression slope, $p < 0.05$).

Somewhat surprisingly, in Fig. 1C we see that Erdős-Rényi (ER) and Barabási-Albert (BA) networks are qualitatively indistinguishable in terms of information flow, despite the preponderance of hubs in the latter that we expect would play an out-sized role in information flow. To better understand this observation, we investigated the variance of $h_\times$ over links in Fig. 3A. We see that the cross-entropy varies more from link to link in the BA networks than for ER networks, indicating that hubs do not move the average information flow but do create fluctuations in the flow, especially for sparser networks.

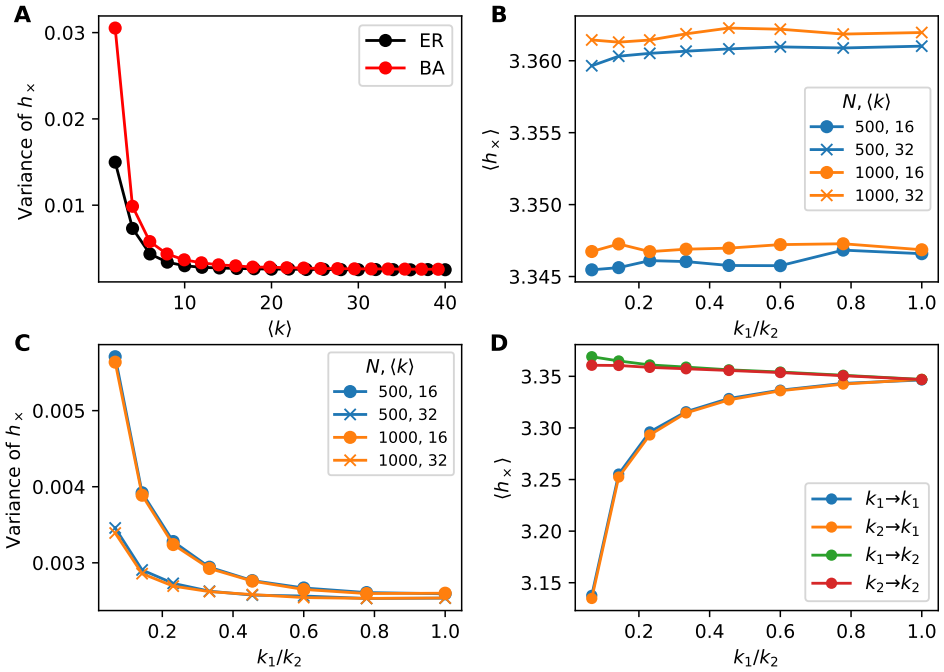

Figure 3. Exploring the variance of information flow. (**A**) Variance of cross-entropy is higher at low densities for BA than ER networks despite the average $h_\times$ being similar (Fig. 1C). (**B-D**) Information flow on dichotomous networks (random networks where all nodes have degree k_1 or degree k_2, allowing tunable degree heterogeneity) of size $N \in \{500, 1000\}$ with $\langle k \rangle \in \{16, 32\}$. Each point constitutes 500 trials. **(B)** Average cross-entropy versus k_1/k_2. Degree heterogeneity does not affect average cross-entropy, supporting Fig. 1C. Network size has a smaller affect on $h_\times$ compared to the average degree. **(C)** Variance of cross-entropy versus k_1/k_2. Higher degree heterogeneity (lower k_1/k_2) leads to higher variation in $h_\times$ over links, indicating the existence of highly predictive nodes and nodes that contribute little predictive information within heterogeneous networks. **(D)** Dichotomous networks of size $N = 1000$ and $\langle k \rangle = 16$. Average cross-entropy over links conditioned on degrees of endpoints (predicting ego from alter). Only the degree of the ego matters, approximately, not the degree of the alter.

To further explore the role of network structure heterogeneity, we investigate dichotomous networks (Sec. 3.4). Here half the nodes have degree k_1 and the other half have degree k_2. Varying the degree ratio k_1/k_2 allows us to tune the degree variance within this simplified network model. In Fig. 3B we see that the total number of nodes and average degree change the average information flow while the degree heterogeneity (k_1/k_2) has little effect. Yet degree heterogeneity does affect the variance of information flow (Fig. 3C). These simpler dichotomous networks show the same effects as observed previously in BA networks.

The simplified bimodal degree distribution of dichotomous networks also lets us explore the effects of ego and alter degrees by computing conditional expectations of $h_\times$ conditioned on degree. We see from the grouping of curves in Fig. 3D that the degree of the ego (the node being predicted) but not the alter (the node predicting) plays a role in the information flow: degree-k_1 egos have more information flow than degree-k_2 egos regardless of the degree of the alter.

4.2. Interplay of clustering and information flow

Next, we study how clustering (transitivity) affects information flow. Clustering plays a complicated role in both simple and complex contagion [24,26] and we report interesting, if mixed, results in Fig. 4 with the quoter model's information flow.

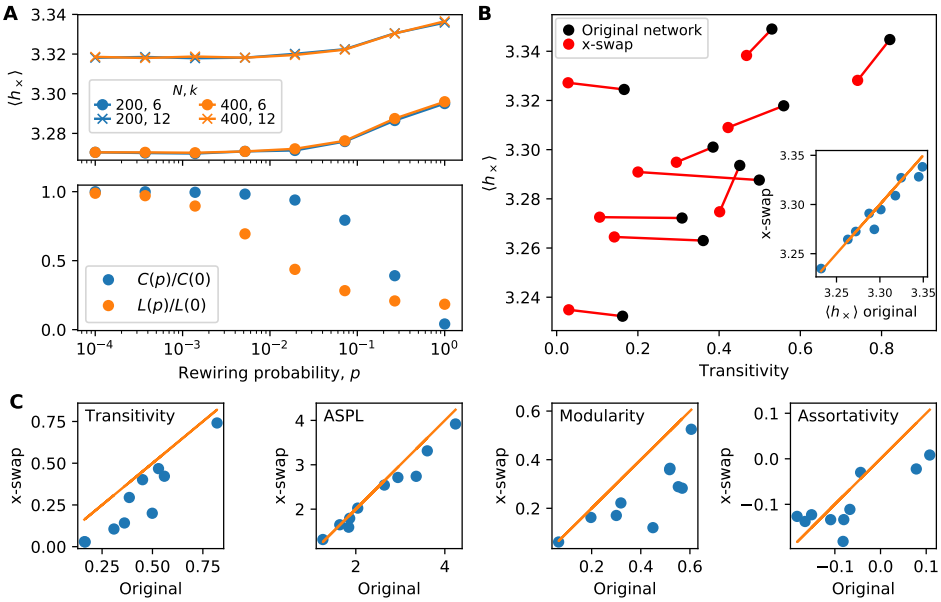

Figure 4. Mixed effects of clustering on information flow. **(A)** Information flow on small-world networks of size $N \in \{200, 400\}$ and average degree $k \in \{6, 12\}$. As network rewiring increases (and clustering decreases) $h_\times$ increases. This suggests that clustered networks promote information flow. Rewiring a small-world network changes the diameter (L) as well the clustering (panel A, bottom); however, $h_\times$ begins to increase primarily when the clustering begins to drop, not when diameter begins to drop. Each point constitutes 300 trials. **(B)** Average cross-entropy versus transitivity for real-world networks. By randomizing networks using the standard "x-swap" method (Sec. 3.4), we can lower the transitivity and investigate how $h_\times$ changes. Some networks show little change in $h_\times$ on randomized networks compared with the original networks, while others show a slight decrease in $h_\times$. This is especially visible in the inset comparing $h_\times$ directly. **(C)** Several network properties before and after the x-swap method. While the x-swap method lowers transitivity, it also alters other important network properties, making it challenging to isolate the role of clustering from other properties.

First, in Fig. 4A we study information flow for small world networks that are randomly rewired to remove clustering [32]. Regardless of network size or average degree, information flow decreases (higher $h_\times$ in top panel of Fig 4A) as clustering decreases (Fig 4A bottom panel). Note that rewiring also changes the diameter of the small-world network, but we see that the main increase in $h_\times$ occurs when clustering begins to drop. In small-world networks, clustering tends to promote information flow.

Next, in Fig. 4B we investigate transitivity in the corpus of real world networks. For each network, we compute information flow on the original network and on a replicate of the network that is randomized by the "x-swap" method. The x-swap method lowers transitivity for all networks but for half of the networks it also lowers $h_\times$, contradicting the previous results on small world networks by indicating that transitivity *inhibits* information. However, it is challenging to draw a sharp conclusion from this x-swap procedure as it also affects other network properties simultaneously. We illustrate this in Fig. 4C where we compare four network properties in the original and x-swapped networks. X-swapping affects transitivity but also average shortest path length (ASPL), modularity and assortativity (degree correlations). This means the changes in information flow seen in Fig. 4B may be due to changes in a combination of these (and possibly other) network properties. Unfortunately, it remains an open research problem how to systematically control for network properties to uncover their affects on dynamics.

4.3. Community structure and the weakness of long ties

The effects of long-range links on information flow have been investigated for some time, from the seminal "strength of weak ties" [28] and the contrasting "weakness of long ties" in complex contagion [27]. Here we investigate long ties in the context of community structure: In networks with densely connected groups of nodes, long ties act to bridge nodes in different groups. How does information flow differ between groups compared to flow within groups?

Using the stochastic block model (Sec. 3.4) with two groups of equal size as a model for networks with dense modules, we study in Fig. 5 information flow between and within groups. The two-group SBM is parameterized by two connection probabilities, the probability for a link within each group (p_0) and the probability for a link between the two groups (p_1). In Figure 5A we see that information flow decreases as p_0 increases and the network becomes more dense. Likewise, the difference in information flow $\Delta h_\times$ increases due to between-block links containing less predictive information (Fig. 5B). This supports the well-known "weakness of long ties" feature of complex contagion. For larger values of p_1, when there are more links connecting the groups making them less distinct, this difference decreases. The collapse of curves in Fig 5C indicates $\Delta h_\times$ is entirely predicated on the network modularity Q.

Interestingly, we also remark that $\Delta h_\times$ is always positive—even when $p_0 < p_1$ (equivalently, $Q < 0$). We would expect more information flow between groups than within when within this "anti-community" regime of the SBM, when there are more links between groups than within groups, yet we observe a weak effect otherwise.

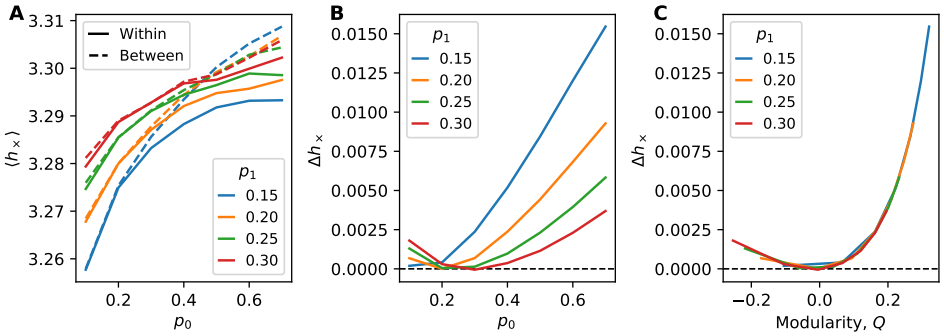

Figure 5. Information flow within the stochastic block model (SBM) of $N = 100$ (two blocks of size $N = 50$). Each point constitutes 10k trials. **(A)** Average cross-entropy on within-block edges and between-block edges as a function of the within-block connection probability p_0 for different between-block connection probabilities p_1. (**B**, **C**) Examining the cross-entropy difference $\Delta h_\times \equiv \langle h_\times(\text{between})\rangle - \langle h_\times(\text{within})\rangle$ across (**B**) connection probabilities and (**C**) modularity Q. Examining $\Delta h_\times$ as a function of modularity Q shows a clear collapse across values of SBM probabilities. Interestingly, anti-community structure ($Q < 0$) still leads to positive $\Delta h_\times$, indicating that information flow is still more prevalent within blocks.

4.4. The role of dynamic heterogeneity

In our results so far, we have treated nodes as identical within the quoter model and focused only on their topological differences within the network. Yet recent studies have underlined the importance of comparing dynamic heterogeneity with structural heterogeneity [40]. Here we taken an exploratory step in this direction by considering a generalization of the quoter model where nodes have different vocabulary distributions.

We explored how information flow changes in the stochastic block model when the nodes in the two blocks have different vocabulary distributions. This is intended to model a difference in the nodes between the two groups, capturing in the quoter model a social homophily in how egos write. Specifically, we assume they have the same vocabularies and follow Zipf distributions, but the exponent of the Zipf distribution is different: nodes in block A have exponent α_A and nodes in block B

have exponent α_B. A larger α (steeper distribution) corresponds to a less diverse vocabulary, and could capture a group of nodes that is more consistent and repetitive in their dialog. In contrast, a lower α (shallower distribution) may describe a group of nodes that uses more diverse words.

Figure 6 shows how information flow changes when the two blocks have different vocabulary distributions (Fig. 6A,C) compared with the same distribution (Fig. 6B). For illustration, we show the Zipfian vocabulary distributions for the two groups as insets in Fig. 6. We observe a much larger trend in how cross-entropy changes with modularity when the exponents are not equal compared to when they are equal. This underscores how structural features (the degree of modularity) greatly magnifies the effects of intrinsic dynamic heterogeneity (different vocabulary distributions). While modularity plays a role even when the two groups have identical vocabulary distributions (Fig. 5), this difference is challenging to detect in Fig. 6B when viewed on the scale of groups with different vocabulary distributions (Fig. 6A,C).

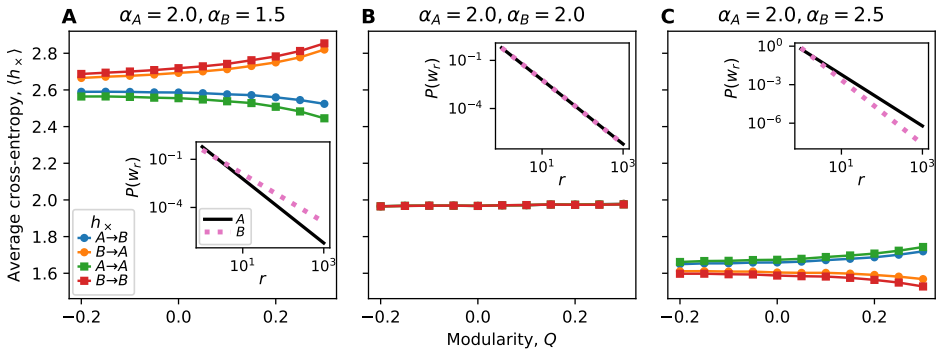

Figure 6. Effects of dynamic heterogeneity on information flow in the stochastic block model. Nodes in group A have Zipfian vocabulary distribution with exponent α_A while nodes in B have exponent α_B. The between-block connection probability is fixed ($p_1 = 0.15$) and the within-block connection probability p_0 is varied to generate a range of modularities. Since the structure is symmetric (subgraphs A and B have the same size and expected density), we only show the result of fixing $\alpha_A = 2$ and varying α_B. Each point constitutes 150 trials. **(A)** The vocabulary distribution of group A has a lower Shannon entropy than of B, and this difference is visible from examining links $A \to A$ and $B \to B$. When examining links $A \to B$ and $B \to A$, the cross-entropy is mainly dependent on the vocabulary distribution of the alter. As modularity increases, differences between the predictabilities of various nodes are exaggerated. **(B)** In homogeneous communities, the cross-entropy does not vary with modularity at such a scale. **(C)** The vocabulary distribution of group A has a higher Shannon entropy than of B. Similar mirror results are seen as in panel A.

5. Discussion

In this work, we have studied how the social flow of written information can be affected by network properties such as the density of links, preponderance of triangles, and modular or community structure. We focused on the quoter model, a toy model for a network of individuals to communicate by generating text sequences and applied information-theoretic estimators of the information flow to these text. We compared results of information flow in the quoter model with traditional simple and complex contagion models.

A particularly intriguing facet of the interplay between quoter model dynamics and network topology is how the quoter model exhibits both the density-driven inhibition of information flow and the weakness of long ties that are signatures of complex contagion, despite lacking an explicit mechanism of social reinforcement. Social reinforcement, the idea that individual's adopt a piece of information only after receiving repeat exposure from social ties, is considered one of the characteristics that distinguishes complex contagion from epidemic spreading. Social reinforcement mechanisms better model how people perceive and react to information. Yet we found here that social reinforcement is not strictly necessary when modeling a more nuanced view of information flow. In particular,

considering text streams (as generated by the quoter model) and predictive measures of information flow (as quantified using cross-entropy estimators) allows us to capture how information can be "drowned out" by the increased "cross-talk" that occurs in denser networks, showing how increased density can inhibit information flow. Further pursuing this line of investigation may give more insight into information flow and even human behavior within social networks.

We also found a mixed combination of results relating clustering to information flow. For small-world (Watts-Strogatz) networks, increasing the clustering leads to a significant increase in information flow (decrease in cross-entropy). At the same time, however, experiments on real world networks showed the opposite effect: randomizing networks to lower transitivity while preserving connectedness and the degree distribution leads to a *decrease* in information flow. However, this well established randomization procedure does not control for other network properties such as modularity or average shortest path length, so it remains an open question if the interplay of multiple effects may resolve the discrepancy between these results.

Another interesting result related information flow to community structure, with the modularity Q used to measure the strength of the modular divide. When $Q > 0$, meaning there were fewer links between modules than expected, we found in Fig. 5 an increase in cross-entropy between modules compared with the cross-entropy between nodes that share a module, as expected by the "weakness of long ties". However, we found the same increase in cross-entropy when $Q < 0$, where there were more links between modules than expected. We would initially expect this regime of "anti-community" structure to have more information flow between modules as there exist more links to facilitate this flow. One possible reason for this anti-community result is that nodes in the same group, while having fewer direct links to one another, may have many links to common nodes in the other group, leading to more similar inputs to their texts. This nonlocal interplay of information flow and network structure is an intriguing avenue for future work.

There are some important limitations to discuss regarding this work. We only considered undirected, unweighted networks. In the context of social networks, this implies all relationships are reciprocal and equal in strength. Future work should extend to directed, weighted networks. Further, a more exhaustive study of the robustness of results to parameter choices is necessary (we take a first step towards this in App. A). Likewise, cross-entropy (Eq. (2)) is a somewhat simplistic information-theoretic measure of information flow, and it is important to consider more advanced measures. Measures such as transfer or causation entropy can offer more insight, quantifying non-redundant information and allowing us to identify indirect influences [7,8]. However, in the context of time-ordered social text data, it is challenging to estimate conditional entropies, making it non-obvious how to implement such measures [12]. Finally, while we observed a number of features that are signatures of complex contagion, not all features of complex contagion are exhibited by the quoter model. For example, there is an optimal modularity that maximizes spreading of complex contagions within the stochastic block model: if Q is either too small or too large then the contagion will not spread [41]. We were unable to observe a corresponding feature within the quoter model. This warrants further investigation, in particular to understand if this is due to how the quoter model differs from complex contagion models, or if it is due to the information-theoretic measure of information, or a combination of the two.

In general, contagion models are a successful way to study information flow in social networks, but to gain more insight it is necessary to adopt more nuanced views of information flow. We argue here that information theory can provide a pathway towards these insights, especially when combined with models such as the quoter model that capture features of human behavior while also modeling key aspects of the data being generated by social network platforms.

Author Contributions: Conceptualization, Tyson Pond, Lewis Mitchell and James Bagrow; Funding acquisition, Lewis Mitchell and James Bagrow; Investigation, Tyson Pond, Saranzaya Magsarjav and Lewis Mitchell; Methodology, Tyson Pond, Tobin South and James Bagrow; Project administration, James Bagrow; Software, Tyson Pond, Tobin South and Lewis Mitchell; Supervision, Lewis Mitchell and James Bagrow; Validation, Tyson Pond, Saranzaya Magsarjav and Lewis Mitchell; Visualization, Tyson Pond; Writing – original draft, Tyson Pond and James Bagrow; Writing – review & editing, Tyson Pond and James Bagrow. All authors have read and agreed to the published version of the manuscript.

Funding: This material is based upon work supported by the National Science Foundation under Grant No. IIS-1447634.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASPL	Average Shortest Path Length
BA	Barabási-Albert
ER	Erdős-Rényi
SBM	Stochastic Block Model
SI	Susceptible-Infected
SIR	Susceptible-Infected-Recovered
WS	Watts-Strogatz

Appendix A. Further exploring quoter model parameters

To support our results, here we explore other choices of quoter model parameters (q and λ). The simulations are done on smaller networks to make it less computationally expensive to do a wide sweep of the parameter space. We first simulate the quoter model on ER, BA, and small-world networks for $q \in \{0.1, 0.5, 0.9\}$ and vary $\langle k \rangle$ or the rewiring probability, p, to support results from Sec. 4.1 and Sec. 4.2. We then simulate the ER, BA, and small-world experiments again for various combinations of the quote probability q and mean quote length λ. We evaluate the robustness of results for ER networks as follows. For each combination of (q, λ), we calculate the difference $\langle h_\times \rangle_{k=20} - \langle h_\times \rangle_{k=6}$, whereby $\langle h_\times \rangle_{k=20}$ we mean the average cross-entropy on ER networks of average degree $k = 20$. The quantity will be positive if density inhibits information flow. This allows us to assess the how the magnitude of our results vary with (q, λ), although it does not confirm a monotonic trend holds. We repeat these calculations with the BA networks and extend them to the small-world networks by replacing $\langle k \rangle$ with $p \in \{0, 1\}$. In general, we find in Figs. A1 and A2 that our results are qualitatively robust to parameter choices, with the exception of very small values of q, as we expect.

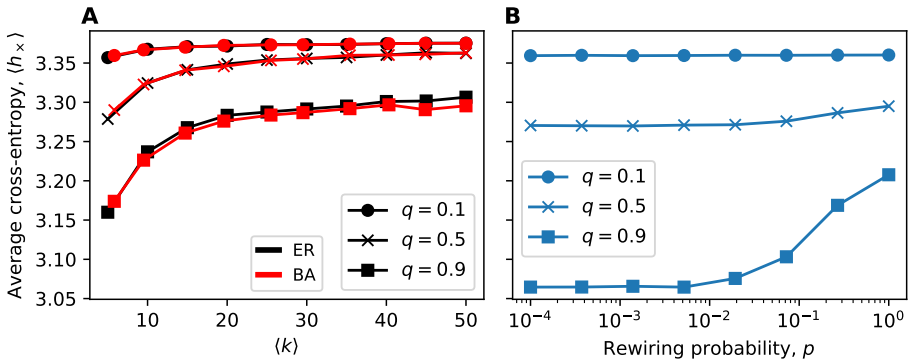

Figure A1. Trends in information flow in ER, BA, and small-world networks for $q \in \{0.1, 0.5, 0.9\}$. With the exception of very low quote probabilities, we see qualitatively similar trends. **(A)** ER & BA networks of size $N = 100$ with varying average degree. Each point constitutes 200 simulations. **(B)** Small-world networks of size $N = 200$ with $k = 6$ with varying rewiring probability. Each point constitutes 500 simulations.

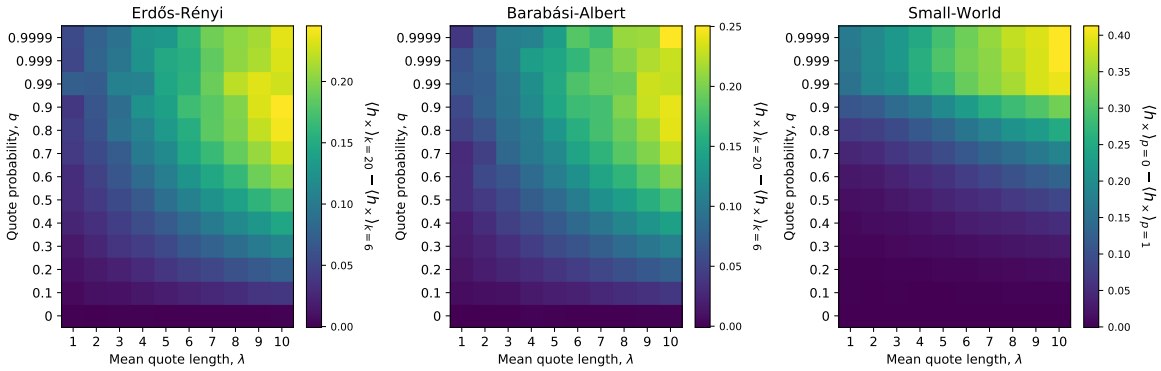

Figure A2. Effects of quoter model parameter choices on observed trends. Information flow is lower for denser ER and BA networks across a range of q and λ with the effect being more pronounced at higher values of q and λ. Likewise, for small-world networks, more clustering (lower p) exhibits higher $h_\times$ than less clustering (higher p), with the effect being most pronounced at $q > 0.5$ regardless of λ. Here, ER & BA networks had $N = 100$ and small-world networks had $N = 200$ and $k = 6$. Each cell constitutes 100 simulations.

Appendix B. Network corpus

All networks studied here can be found through the Index of Complex Networks (ICON) [42]. We converted any directed or weighted networks to undirected (bi-directional) and unweighted. Details for each of the ten networks:

1. Les Miserables co-appearances [43] [Undirected, Weighted].
2. Hollywood film music [44] [Undirected, Weighted]. This is a bipartite network; we converted it to a one-mode projection (nodes are composers and two composers are linked if they worked with the same producer).
3. Freeman's EIES dataset [45] [Directed, Weighted]. We used the "personal relationships (time 1)" network.
4. Sampson's monastery [46] [Directed, Weighted]. We used the Pajek dataset. The weight of a directed link represents how an individual rates the other. The rating can be positive (1,2,3 = top 3 ranked) or negative (-1,-2,-3 = worst 3 ranked). We chose to only keep links which were positive.
5. Golden Age of Hollywood [47] [Directed, Weighted]. We used the aggregated network over 1909-2009.
6. 9-11 terrorist network [48] [Undirected, Unweighted].
7. CKM physicians social network [49] (1966) [Directed, Unweighted]. We used "CKM physicians Freeman" networks hosted by Linton Freeman, and chose the "friend" network (i.e., the third adjacency matrix). We took only the giant component.
8. Kapferer tailor shop [50] (1972) [Undirected, Unweighted]. We used the "Kapferer tailor shop 1" Pajek dataset (kapfts1.dat).
9. Dolphin social network [51] (1994-2001) [Undirected, Unweighted].
10. Email network (Uni. R-V, Spain, 2003) [52] [Directed, Unweighted]. We used the "email-uni-rv-spain-arenas" network.

References

1. Lazer, D.; Pentland, A.; Adamic, L.; Aral, S.; Barabasi, A.L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; Jebara, T.; King, G.; Macy, M.; Roy, D.; Van Alstyne, M. SOCIAL SCIENCE: Computational Social Science. *Science* **2009**, *323*, 721–723.

2. Tumasjan, A.; Sprenger, T.O.; Sandner, P.G.; Welpe, I.M. Predicting elections with twitter: What 140 characters reveal about political sentiment. Fourth international AAAI conference on weblogs and social media, 2010.
3. Conover, M.D.; Ferrara, E.; Menczer, F.; Flammini, A. The Digital Evolution of Occupy Wall Street. *PLoS One* **2013**, *8*, e64679.
4. Castells, M. *Networks of outrage and hope: Social movements in the Internet age*; John Wiley & Sons, 2015.
5. de Montjoye, Y.A.; Radaelli, L.; Singh, V.; Pentland, A. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science* **2015**, *347*, 536–539.
6. Garcia, D. Leaking privacy and shadow profiles in online social networks. *Sci. Adv.* **2017**, *3*, e1701172.
7. Schreiber, T. Measuring Information Transfer. *Phys. Rev. Lett.* **2000**, *85*, 461–464.
8. Sun, J.; Bollt, E.M. Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. *Physica D* **2014**, *267*, 49–57.
9. Borge-Holthoefer, J.; Perra, N.; Gonçalves, B.; González-Bailón, S.; Arenas, A.; Moreno, Y.; Vespignani, A. The dynamics of information-driven coordination phenomena: A transfer entropy analysis. *Sci. Adv.* **2016**, *2*, e1501158.
10. Wang, D.; Wen, Z.; Tong, H.; Lin, C.Y.; Song, C.; Barabási, A.L. Information spreading in context. Proceedings of the 20th international conference on World wide web - WWW '11. ACM Press, 2011, pp. 735–744.
11. Bagrow, J.P.; Liu, X.; Mitchell, L. Information flow reveals prediction limits in online social activity. *Nat Hum Behav* **2019**, *3*, 122–128.
12. Bagrow, J.P.; Mitchell, L. The quoter model: A paradigmatic model of the social flow of written information. *Chaos* **2018**, *28*, 075304.
13. Centola, D. The Spread of Behavior in an Online Social Network Experiment. *Science* **2010**, *329*, 1194–1197.
14. Borge-Holthoefer, J.; Banos, R.; Gonzalez-Bailon, S.; Moreno, Y. Cascading behaviour in complex socio-technical networks. *J. Complex Networks* **2013**, *1*, 3–24, [https://academic.oup.com/comnet/article-pdf/1/1/3/1370358/cnt006.pdf].
15. Shannon, C. Prediction and Entropy of Printed English. *Bell System Technical Journal* **1951**, *30*, 50–64.
16. Kontoyiannis, I.; Algoet, P.; Suhov, Y.; Wyner, A. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Trans. Inform. Theory* **1998**, *44*, 1319–1327.
17. Song, C.; Qu, Z.; Blumm, N.; Barabasi, A.L. Limits of Predictability in Human Mobility. *Science* **2010**, *327*, 1018–1021.
18. Ziv, J.; Merhav, N. A Measure of Relative Entropy between Individual Sequences with Application to Universal Classification. Proceedings. IEEE International Symposium on Information Theory. IEEE, IEEE, 1993, pp. 352–352.
19. Sun, J.; Taylor, D.; Bollt, E.M. Causal Network Inference by Optimal Causation Entropy. *SIAM J. Appl. Dyn. Syst.* **2015**, *14*, 73–106.
20. Granovetter, M. Threshold Models of Collective Behavior. *American Journal of Sociology* **1978**, *83*, 1420–1443.
21. Watts, D. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences* **2002**, *99*, 5766–5771.
22. Centola, D.; Eguíluz, V.M.; Macy, M.W. Cascade dynamics of complex propagation. *Physica A* **2007**, *374*, 449–456.
23. Ugander, J.; Backstrom, L.; Marlow, C.; Kleinberg, J. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences* **2012**, *109*, 5962–5966, [https://www.pnas.org/content/109/16/5962.full.pdf].
24. Miller, J.C. Percolation and epidemics in random clustered networks. *Phys. Rev. E* **2009**, *80*, 020901.
25. Pastor-Satorras, R.; Castellano, C.; Van Mieghem, P.; Vespignani, A. Epidemic processes in complex networks. *Rev. Mod. Phys.* **2015**, *87*, 925–979.
26. O'Sullivan, D.J.; O'Keeffe, G.J.; Fennell, P.G.; Gleeson, J.P. Mathematical modeling of complex contagion on clustered networks. *Front. Phys.* **2015**, *3*, 71.
27. Centola, D.; Macy, M. Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology* **2007**, *113*, 702–734.
28. Granovetter, M.S. The Strength of Weak Ties. In *Social Networks*; Elsevier, 1977; pp. 347–367.
29. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons, Inc., 1991.
30. Miller, J.; Ting, T. EoN (Epidemics on Networks): A fast, flexible Python package for simulation, analytic approximation, and analysis of epidemics on networks. *JOSS* **2019**, *4*, 1731.

31. Lambiotte, R. How does degree heterogeneity affect an order-disorder transition? *Europhys. Lett.* **2007**, *78*, 68002.
32. Watts, D.J.; Strogatz, S.H. Collective dynamics of 'small-world' networks. *Nature* **1998**, *393*, 440–442.
33. Singh, P.; Sreenivasan, S.; Szymanski, B.; Korniss, G. Threshold-limited spreading in social networks with multiple initiators. *Sci Rep* **2013**, *3*, 2330 EP –.
34. Milo, R.; Kashtan, N.; Itzkovitz, S.; Newman, M.E.; Alon, U. On the uniform generation of random graphs with prescribed degree sequences. *arXiv preprint cond-mat/0312028* **2003**.
35. Blitzstein, J.; Diaconis, P. A Sequential Importance Sampling Algorithm for Generating Random Graphs with Prescribed Degrees. *Internet Mathematics* **2011**, *6*, 489–522.
36. Newman, M.; Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **2004**, *69*, 026113.
37. Danon, L.; Díaz-Guilera, A.; Duch, J.; Arenas, A. Comparing community structure identification. *J. Stat. Mech.* **2005**, *2005*, P09008–P09008.
38. Karrer, B.; Newman, M. Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **2011**, *83*, 016107.
39. Blondel, V.D.; Guillaume, J.L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, *2008*, P10008.
40. de Arruda, G.F.; Petri, G.; Rodrigues, F.A.; Moreno, Y. Impact of the distribution of recovery rates on disease spreading in complex networks. *Phys. Rev. Research* **2020**, *2*, 013046.
41. Nematzadeh, A.; Ferrara, E.; Flammini, A.; Ahn, Y.Y. Erratum: Optimal Network Modularity for Information Diffusion [Phys. Rev. Lett. 113, 088701 (2014)]. *Phys. Rev. Lett.* **2014**, *113*, 088701.
42. Clauset, A.; Tucker, E.; Sainz, M. The Colorado index of complex networks. *Retrieved July* **2016**, *20*, 2018.
43. Knuth, D.E. *Stanford GraphBase: A platform for combinatorial computing*; Addison-Wesley, 1993.
44. Faulkner, R.R. *Music on demand: composers and careers in the Hollywood film industry*; Transaction Books, 1983.
45. Freeman, S.C.; Freeman, L.C. *The networkers network: A study of the impact of a new communications medium on sociometric structure*; School of Social Sciences University of Calif., 1979.
46. Sampson, S.F. A novitiate in a period of change: An experimental and case study of social relationships. PhD thesis, Cornell University, 1968.
47. Taylor, D.; Myers, S.A.; Clauset, A.; Porter, M.A.; Mucha, P.J. Eigenvector-Based Centrality Measures for Temporal Networks. *Multiscale Model. Simul.* **2017**, *15*, 537–574.
48. Krebs, V. Uncloaking Terrorist Networks. *First Monday* **2002**, *7*, 43–52.
49. Burt, R.S. Social Contagion and Innovation: Cohesion versus Structural Equivalence. *American Journal of Sociology* **1987**, *92*, 1287–1335.
50. Kapferer, B. *Strategy and transaction in an African factory: African workers and Indian management in a Zambian town*; Manchester University Press, 1972.
51. Lusseau, D.; Schneider, K.; Boisseau, O.J.; Haase, P.; Slooten, E.; Dawson, S.M. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behav. Ecol. Sociobiol.* **2003**, *54*, 396–405.
52. Guimerà, R.; Danon, L.; Díaz-Guilera, A.; Giralt, F.; Arenas, A. Self-similar community structure in a network of human interactions. *Phys. Rev. E* **2003**, *68*, 065103.