Social Network Analysis and Mining manuscript No.
(will be inserted by the editor)

On the Computation of Entropy Production in Stationary Social Networks

**Tobias Hoßfeld, Valentin Burger, Haye Hinrichsen,
Matthias Hirth, Phuoc Tran-Gia**

Received: February 5, 2014/ Accepted: n.a.

Abstract Completing their initial phase of rapid growth social networks are expected to reach a plateau from where on they are in a statistically stationary state. Such stationary conditions may have different dynamical properties. For example, if each message in a network is followed by a reply in opposite direction, the dynamics is locally balanced. Otherwise, if messages are ignored or forwarded to a different user, one may reach a stationary state with a directed flow of information. To distinguish between the two situations, we propose a quantity called *entropy production* that was introduced in statistical physics as a measure for non-vanishing probability currents in nonequilibrium stationary states. The proposed quantity closes a gap for characterizing online social networks. As major contribution, we show the relation and difference between entropy production and existing metrics. The comparison shows that computational intensive metrics like centrality can be approximated by entropy production for typical online social networks. To compute the entropy production from real-world measurements, the need for Bayesian inference and the limits of naïve estimates for those probability currents are shown. As further contribution, a general scheme is presented to measure the entropy production in small world networks using Bayesian inference. The scheme is then applied for a specific example of the R mailing list.

Keywords Social Network Characterization, Entropy Production, Bayesian Inference, Practical Computation

Tobias Hoßfeld, Valentin Burger, Matthias Hirth, Phuoc Tran-Gia
University of Würzburg, Institute of Computer Science, Chair of Communication Networks, Am Hubland, 97074 Würzburg, Germany. E-mail: tobias.hossfeld@uni-wuerzburg.de

Haye Hinrichsen
University of Würzburg, Department of Physics and Astronomy, Am Hubland, 97074 Würzburg, Germany. E-mail: hinrichsen@physik.uni-wuerzburg.de

1 Introduction

Due to the rapid growth of social media in the last decade, many theoretical studies have been focused on the growth dynamics of social networks (Jin et al. 2001). In such a social network, individuals are connected to each other (e.g. friends in facebook, sender and receiver in mail networks) and there is an interaction between the individuals (e.g. exchanging messages). Hence, beside the network topology, the interaction among individuals characterize the social network. However, recent observations by Barnes and Andonian (2011) indicate that many social networks are approaching a plateau of constant size, e.g. due to logistic growth models and resulting upper population bounds like in Hoßfeld et al. (2011a). In such a matured state, the dynamics of the network are approximately stationary in a statistical sense, meaning that the network topology as well as the probability for receiving and sending messages do not change in the long-term limit.

The dynamics of a stationary state of a network is not uniquely given, rather there is a large variety of possible realizations. For example, the three individuals shown in Fig. 1 may send messages (a) randomly in both directions or (b) in clockwise direction. In both situations the individuals send and receive messages at constant rate, meaning that the network is statistically stationary. However, in the first case the dynamics is locally balanced between pairs of users, while in the second case there is a directed current of messages flowing clockwise through the system.

In the present work we introduce a new type of quantity, called *entropy production*, to characterize the stationary properties of arbitrary social networks. In statistical physics this quantity is frequently used as a experimentally accessible measure for the irreversibility of a complex system dynamical system, see e.g. Tietz et al. (2006) and Andrieux et al. (2007). More specifically, in physics thermally equilibrated systems are known to be invariant under time reversal, i.e., watching a movie of the microscopic dynamics we would not be able to decide whether the movie is running forward or backward. Obviously such a system cannot exhibit any kind of directed currents because they would flow in opposite direction under time reversal. This means that in thermal equilibrium all microscopic processes are statistically undirected, obeying 'detailed balance' in the jargon of statistical physics. Systems out of equilibrium, however, do have directed currents, even if the dynamics is statistically stationary. It turns out that such currents can only be maintained if they are actively driven from outside, leading to a continuous increase of the entropy in the environment. Therefore, by quantifying the entropy production in the environment, one can measure how strongly a system is out of thermal equilibrium.

Here we show that the concept of entropy production can also be used in the completely different context of communication networks. More specifically, we associate with each pair of individuals i, j of a social network a quantity H_{ij} called entropy production, which depends on the number of messages sent from i to j and vice versa. Of course this quantity is not an entropy in the usual physical sense, it rather measures the *directionality* of the information exchange and vanishes for perfectly balanced communication. Therefore, the intuitive meaning of entropy production in the context of social networks is that of a measure for the directionality of communication flow. As we will see, this measure can be defined both globally and locally. For example, defining the entropy production of a node as the sum over the entropy of all its links, one can identify nodes contributing preferentially to balanced or unidirectional information transfer.

The concept of entropy production requires to make certain assumptions about the dynamics of the network. In particular, we ignore possible correlations between the messages by assuming that the individuals communicate randomly at constant rates. With this assumption each message sent from node i to node j increases the entropy by

$$\Delta H_{ij} := \ln \frac{w_{ij}}{w_{ji}}, \tag{1}$$

where w_{ij} and w_{ji} are the rates for messages from i to j and in opposite direction, respectively. In physics, this quantity can be interpreted as the minimal entropy produced by a machine that keeps a nonequilibrium process running (Schnakenberg 1976; Schreiber 2000; Andrieux and Gaspard 2004; Seifert 2005). In the present context it does not represent a physical entropy, it rather characterizes the directionality of communication, for example, by distinguishing the situations (a) and (b) in Fig. 1. This means that we use the term *entropy production* to highlight the analogy to physical systems but not in the sense of a direct correspondence.

Equation (1) is trivial to evaluate if the rates w_{ij} and w_{ji} are known. However, in realistic networks with data taking over a finite time span T, only the number of messages n_{ij} and n_{ji} exchanged between pairs of nodes are known. Although it is tempting to replace the rates w_{ij} by the relative frequencies n_{ij}/T and to approximate the entropy production by $\Delta H_{ij} = \ln n_{ij} - \ln n_{ji}$, it is easy to see that this approximation would diverge as soon as one of the count numbers vanishes. Therefore, this article deals to a large extent with the question how we can reasonably reconstruct the rates from the given number of messages. Bayesian inference enables the computation of rates from observations which leads to non-vanishing values ($w_{ij} > 0$) even in the case that no messages were sent ($n_{ij} = 0$), but only received ($n_{ji} > 0$) or vice versa ($w_{ji} > 0, n_{ji} = 0, n_{ij} > 0$). Accordingly, the entropy production is calculated and leads to finite values.

The remainder of this paper is structured as follows. Section 2 provides a brief overview on existing social network measures and revisits related work in order to show that entropy production fills a gap in characterizing social networks. Section 3 introduces variables describing the observed data from a measurement campaign of a social network. Further, assumptions on the network dynamics are summarized. Using these assumptions we define the entropy production as a function of the (unknown) rates at which the individuals exchange their messages. Applying a simulation model to construct arbitrary networks with given properties and well-known rates, we show the need for introducing entropy production as a new metric in Section 4, as it quantifies the social network characteristics differently than existing metrics. An estimator of the unknown rates based on the observed measurement data is then introduced by means of Bayesian inference. Section 5 shows the need to use

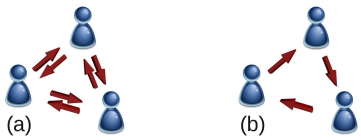


Fig. 1 Example of a stationary network with three users. (a) Each individual sends messages to randomly selected neighbors, leading to a statistically balanced stationary state with vanishing entropy production. (b) The individuals send messages to only one neighbor, generating a stationary but directed flow of information with positive entropy production.

Bayesian inference for computing the rates w_{ij} instead of using measured frequencies n_{ij}. Section 6 presents appropriate choice of the prior distribution for small-world networks, in which the number of messages follows a power-law distribution. Relevant parameters of the prior distribution are calculated which finally allows computing the entropy production. The general scheme is summarized for social networks manifesting small-world characteristics on the number of messages. The general scheme is applied exemplarily to the R mailing list. Finally, Section 7 concludes the work and gives an outlook on next steps in this research direction.

2 Related Work on Social Networks, their Characteristics, and Common Metrics

In literature, various measures of the characteristics of complex networks exist. We briefly revisit them in order to show qualitatively that entropy production fills a gap for measuring the directionality of the information exchange and quantifying the balance of communication or interaction. The quantities introduced in literature analyze mainly the topology of the graph itself by means of the adjacency matrix $\mathcal{A}$ with elements $\mathcal{A}_{ij} = 1 - \delta_{0,n_{ij}}$ using the Kronecker delta $\delta_{a,b}$. Extensions of several quantities exist for weighted networks. In that case, the directed link connecting the nodes i and j is weighted by the message rate w_{ij}. Instead of $\mathcal{A}$, the message rate matrix $\mathcal{W}$ is used. Thereby, beyond the topological effects, the metrics which allow to work on weighted networks give insights into the structure of the message diffusion, too. Those metrics are to be analyzed with the network entropy H for different network topologies and message exchange models $\mathcal{W}$. Section 4 concerns a quantitative analysis of those metrics with the entropy production for different network topologies and message exchange models $\mathcal{W}$, while Section 6.4 considers a real-world example. In addition, recent work on the dynamical properties of networks from an entropic viewpoint is revisited at the end of this section.

2.1 Principal Graph Characteristics

The basic quantities are the in- and out-degree of nodes corresponding to the number of incoming and outgoing links of nodes. We observe a strong correlation of the node degree with entropy production for the example of the R mailing list which will be discussed later in Section 6.4. Nodes with a large number of outgoing links tend to produce less entropy per message. Future work investigates for which kind of network topologies and message exchange models those quantities are correlated. Other principal characteristics of nodes are eccentricity and local clustering coefficient. Global network metrics are e.g. radius, diameter, average path length, or assortativity coefficient. Those metrics can be extended to weighted networks and need to be interrelated to entropy production per node and network entropy production, respectively.

Social networks show heavy-tailed in- and out-degrees (e.g. Bird et al. 2006; Broder et al. 2000; Mislove et al. 2007). Futhermore social networks form communities which are exhibited by a high clustering coefficient. Recent works aim at identifying those communities efficiently (Vasudevan and Deo 2012) or in dynamic social networks (Gilbert et al. 2011). As a social network evolves the effective diameter of the network shrinks because cross-community links are formed which lead to a huge largest connected component. We use the

Forest Fire model (Leskovec et al. 2005) to generate networks with those properties. This is later described in detail in Section 4.

2.2 Centrality Metrics

Centrality metrics quantify the 'importance' of nodes. Different variations exist like degree, (random walk) closeness, information, betweenness, or Eigenvector centrality like PageRank. Considering again the R mailing list, we observed a strong correlation between entropy production per node H_i and e.g. betweenness centrality. Nodes in the social network that have a high probability to occur on a randomly chosen shortest path between two randomly chosen nodes have a high betweenness. Hence, those nodes are also responsible for high entropy production in the network. However, node betweeness centrality is computationally expensive. Brandes (2001) proposes an approximation of weighted graphs in time $O(nm) + n^2 \log n$ for n vertices and m edges. Kourtellis et al. (2013) introduced a new graph centrality index to reach a speedup gain of up to 6 orders of magnitude for large networks while still allowing to identify the top central nodes properly. Beyond betweenness centrality, closeness measures how fast it will take to spread information from a single node to all other nodes sequentially. In contrast, closeness centrality and entropy production revealed no correlation which we will show in Section 6.4.

2.3 Symmetry Measures and Entropy Measures

Current developments introduce measures of symmetry and their relation with measures like graph entropy (Garrido 2011). The graph entropy is defined as the intrinsic Shannon entropy or information content of the node distribution quantifying the homogeneity of the node importance. More specifically, networks with equally important nodes have a high graph entropy, while networks with dominating hubs are characterized by a low graph entropy. This entropy differs from the entropy *production* studied in the present work in that the latter depends on rates of communication *between* the nodes, quantifying the asymmetry of data transmission across links. Hence, graph entropy measures the amount of information within the graph. Symmetry of complex networks means invariance of adjacency of nodes under the permutations on the node set itself and a symmetry index is defined in Mowshowitz and Dehmer (2010). This concept is close to entropy production, however, symmetry measures are defined on the network topology only. In a similar way, the symmetry based structure entropy of complex networks (Xiao et al. 2008) quantifies the heterogeneity of a network system concerning the automorphism partition of the node set into equivalent cells. Thereby, the probability that a node belongs to the cell is used while message rates are not part of this concept. Dehmer and Mowshowitz (2011) survey methods for measuring the entropy of graphs and their applicability and present later a taxonomy to the measurement of network complexity for a systematic interpretation of entropy based measures of graph complexity (Mowshowitz and Dehmer 2012). The entropy based measures depend on particular structural features, as the above mentioned graph entropy.

2.4 Dynamical Properties of Networks from an Entropic Viewpoint

As a key concept, entropy production quantifies the dynamical properties of a network and the flow of information. Kossinets and Watts (2006) shows on the example of email communication that average network properties appear to approach an equilibrium state, whereas individual properties are unstable. To this end, a network time series is constructed from the user interactions in terms of email exchange which allows to analyze network-level properties over time. In particular, the mean node degree, the fractional size of the largest component, the average path length, and the clustering coefficient are calculated. Kossinets and Watts found that averages of local network properties like clustering coefficient are more stable than global properties like average shortest path length. Related to entropy production, the convergence of those metrics supports our assumption that social networks reach a statistically stationary state which was also demonstrated for the network of Wikipedia authors by Hirth et al. (2012). Wang and De Wilde (2004) propose an email network model which formulates the network topology as a random process which allows to study the dynamics of the process. In Zhu et al. (2006), analytical results show a number of steady state properties on email exchange between nodes and the macroscopic networking behavior. The influence of the network topology on the dynamics of spreading processes is investigated on macroscopic but also on microscopic level in related work. Smilkov and Kocarev (2012) study an epidemic model on graphs in which the dynamics of individual nodes is described by a discrete-time Markov chain. Thereby numerical results are given for the Enron email network with respect to the reactive process and the contact process. As a main result, it is shown that under certain circumstances local information of the topology, i.e. without knowing the whole network, is sufficient to derive characteristics of the spreading process. An interesting model based on interarrival times is provided by Mihaljev et al. (2011) which allows to indirectly learn about the process of sending messages on networks with unknown topologies. Entropy production utilizes also only local information of a node about its neighbors and the message exchange with them in order to quantify globally the balance of information flow in the network.

The dynamical properties of entropy in complex networks has recently drawn attention in the research community. Gómez-Gardeñes and Latora (2008) analyze the entropy rate for a dynamical process on a graph and in particular diffusion processes with node degrees as local information by random walkers related to information flow in a social network. Entropy rate of the diffusion process is defined to capture the dynamics, i.e. the bias in the random walker, and the graph topology. Therefore, entropy rate also allows to characterize complex networks. The design of random walks that maximize entropy rate on a given graph is proposed by Sinatra et al. (2011). The goal is to use local information on the graph only to construct the maximal-entropy random walk which has its application for example in spreading information in a network. For quantifying the complexity of a system changing in time, West et al. (2012) consider approximate entropy whose notion was introduced by Pincus and Huang (1992). Thereby, the concept of approximate entropy is modified in order to cope with networks generated by a dynamical process and West et al. find that approximate entropy relates to the amount of information generated by the underlying dynamics. The quantity entropy production is complementary to the approaches above and its relation to entropy rate or approximate entropy is an interesting path for future work.

3 Entropy Production: Definition and Model Assumptions

3.1 Definition

As outlined above, the entropy production is defined in such a way that each message sent from node i to j produces an entropy of

$$\Delta H_{ij} \;:=\; \ln \frac{w_{ij}}{w_{ji}} \,. \tag{2}$$

Since node i sends n_{ij} messages to node j during the observation period, the total entropy produced by messages $i \to j$ is given by $n_{ij}\Delta H_{ij}$, while messages in opposite direction produce the entropy $n_{ji}\Delta H_{ji}$. Adding the two contributions we obtain the link entropy

$$H_{ij} \;=\; n_{ij}\Delta H_{ij} + n_{ji}\Delta H_{ji} \;=\; (n_{ij} - n_{ji}) \ln \frac{w_{ij}}{w_{ji}} \,. \tag{3}$$

We chose Eq. (2) as a logarithmic ratio of rates because it complies with the usual definition of entropy production in statistical physics. The logarithmic ratio can be motivated as follows. Firstly, a reasonable measure for the directionality of the traffic along a link should increase linearly with the intensity of the mutual communication. As $(n_{ij} - n_{ji})$ in Eq. (3) is linear in the communication intensity, this means that ΔH_{ij} should be dimensionless and asymptotically constant. Since rates carry the dimension 1/time, this suggests that ΔH_{ij} should depend on the *ratio* of the forward and backward rate. Moreover, a measure of directionality should respond to the imbalance of message flow but not on the direction itself, i.e. it should be invariant under links reversal $i \leftrightarrow j$. Obviously, the logarithm in Eq. (3) is the most natural function which obeys these requirements.

In statistical physics the logarithm has in addition a direct information-theoretic meaning as it quantifies the amount of information needed to specify physical states in the environment. In case of social networks this could be translated into the difference of information needed to specify one of the messages in forward and in backward direction, respectively (not to be confused with the unknown information content of the messages themselves). In the present case, however, this information-theoretic interpretation is not needed and therefore it is not required to use a logarithm. In fact, various other functionals may be also suitable and should be studied in the future. However, the logarithm is simple and special in so far as it responds only weakly to extreme rate asymmetries. Moreover, it behaves nicely under Bayesian regularization, as will be discussed below.

As the entropy is symmetric ($H_{ij} = H_{ji}$) it can equally be attributed to the corresponding nodes, allowing us to define an entropy production per node i by considering all N links of node i

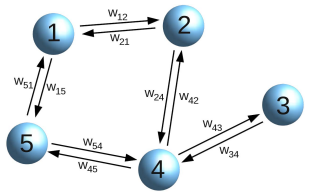
$$H_i \;=\; \frac{1}{2} \sum_{j=1}^{N} H_{ij} \tag{4}$$

as well as the entropy production of the total network

$$H \;=\; \sum_i H_i \;=\; \frac{1}{2} \sum_{i,j=1}^{N} H_{ij} \,. \tag{5}$$

3.2 Observed Data

Let us consider a social network of individuals communicating by directed messages (e.g. emails, Twitter or Facebook messages). Suppose that we monitor M messages over a finite time span, recording sender and receiver identifiers in a file. Such a data set can be represented as a graph of nodes (individuals) connected by directed links (messages) as depicted in Fig. 2.



$$\{w_{ij}\} = \begin{pmatrix} 0 & 0.3 & 0 & 0 & 2.0 \\ 0.5 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 3.0 & 0 \\ 0 & 0.7 & 0.2 & 0 & 1.5 \\ 0.2 & 0 & 0 & 1.5 & 0 \end{pmatrix}, \quad \{n_{ij}\} = \begin{pmatrix} 0 & 0 & 0 & 0 & 5 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 4 & 0 \\ 0 & 1 & 1 & 0 & 3 \\ 0 & 0 & 0 & 2 & 0 \end{pmatrix}$$

Fig. 2 Example of a directed social network with $N = 5$ participants. It is assumed that node i sends messages to node j randomly with the rate w_{ij}. Observing the network for a finite time span the number of recorded messages from i to j is n_{ij}. In order to compute the entropy production, one has to estimate the unknown rates w_{ij} from the numbers n_{ij}.

Enumerating the individuals in the list by $i = 1 \ldots N$, let n_{ij} be the number of messages sent from i to j. These numbers constitute a $N \times N$ connectivity matrix which is the starting point for the subsequent analysis. If at least one message is sent from i to j, the two nodes are said to be connected by a directed link. Obviously, the number of recorded messages M and the total number of directed links L are the following using the Kronecker delta $\delta_{a,b}$.

$$M = \sum_{i,j=1}^{N} n_{ij}\,, \qquad L = \sum_{i,j=1}^{N} (1 - \delta_{0,n_{ij}}) \tag{6}$$

Note that $M \geq L$ since two individuals can communicate several times during the observation period. The statistics of multiple communication is described by the probability distribution

$$P(n) := \frac{\sum_{i,j=1}^{N} \delta_{n,n_{ij}}}{N(N-1)} \tag{7}$$

of the matrix elements n_{ij}. In the present work we are particularly interested in social media networks with a small-world topology, where this distribution follows a power law,

$$P(n) \sim n^{-1-\alpha}\,. \tag{8}$$

3.3 Assumptions on Stationary Network Dynamics

In a realistic social network the messages are causally connected and mutually correlated. As this information is usually not available and requires semantic analysis of the messages, let us consider the messages as uncorrelated instantaneous events which occur randomly like the clicks of a Geiger counter. More specifically, we start with the following assumptions:

- *Stationarity:* We assume that the size of the social network is approximately constant during data taking. This means that the total number N_{tot} of participants in the system is constant. This assumption is e.g. valid for networks following a logistic growth model (Hoßfeld et al. 2011a). Note that N_{tot} may be larger than the actual number of participants N communicating during data taking.
- *Effective rates:* Messages are sent from node i to node j at a constant *rate* (probability per unit time), denoted as $w_{ij} \geq 0$.
- *Reversibility:* If node i can communicate with note j, node j can also communicate with node i. This assumption is typically true in social networks. Hence if w_{ij} is nonzero, then the rate in opposite direction w_{ji} is also nonzero.

With these assumptions, the average number of communications from i to j is given by $\langle n_{ij} \rangle = w_{ij}T$, where T denotes the observation time. It has to be noted that the assumptions above can be relaxed at the cost of complexer derivations of the message rates from observed measurements, see Section 6. For the sake of simplicity, throughout this paper we shall assume that these assumptions are valid or at least approximately fulfilled.

4 Entropy Production vs. Existing Metrics: Example of ForestFire Graphs

The differences between entropy production and existing metrics for characterizing social networks were qualitatively discussed in Section 2. To quantify the differences but also to understand similarities, arbitrary networks with given network characteristics and well-known message rates are considered. The intention is to support the reader with the interpretation of the entropy production metric and its relation to properties of a simple synthetic network which can be easily understood. As a result, we found that for typical small world networks, complex metrics like betweenness centrality can be approximated by entropy production which is a strong outcome due to the computational effort for computing betweenness centrality (Brandes 2001; Kourtellis et al. 2013). In the following, a simulation model for the network topology and the message rates is introduced, before the numerical results are presented.

4.1 Simulation Model

In order to evaluate the relevance of entropy production for real networks and to compare entropy production with existing online social network metrics, we need a graph model that generates realistic online social network graphs. Online social network graphs have several properties such as a degree power law, i.e. heavy tailed in- and out-degrees, communities, a densification power law and a shrinking diameters. The densification power law describes the evolution of the number of edges $e(t)$ and nodes $n(t)$ of the graph at time t

$$e(t) \propto n(t)^a, \tag{9}$$

where a is an exponent that is generally between 1 and 2. As the network grows and gets more dense, the effective diameter is shrinking. This enables small world phenomena to

be present even in large networks. A model that produces graphs with these properties is the Forest Fire Model which was first proposed in Leskovec et al. (2005). In literature, there are also other topology models provided as for example by Sallaberry et al. (2013) which allows to generate artificial social networks having community structures with small-world and scale-free properties. Nevertheless, the aim of Section 4 is to illustrate on one hand the differences between entropy production and other metrics and on the other hand to illustrate the meaning of the entropy production metric. Therefore, we have used the Forest Fire model as an example, as its properties are well researched and due to its simplicity to get an impression about entropy production. Along the same line of argumentation, we have used a simple message rate model of constant outgoing email traffic rate which is known as equal contact model proposed by Zhu et al. (2006). In contrast, Section 6.4 provides the analysis of the R mailing list as a real world example including correlations between message rate and network topology.

4.1.1 Forest Fire Model

The Forest Fire model is based on a process that finds neighbors of new nodes by virally burning through existing edges. The most basic version of the model is defined as follows. A new node v forms an out-link to a random node w in the existing graph. Then it burns links outward from w linking to the nodes with a certain probability p and recursively discovering the linked nodes. Similar to a citation network where an author not only is interested in the literature a paper A refers to but also in the papers that refer to paper A, in-links are also burned backward with a certain probability p_b in the same fashion as the out-links. For a detailed description of the algorithm refer to the paper by Leskovec et al. (2005).

Such we obtain graphs with the desired properties. Highly-linked nodes can easily be reached by new nodes during the discovering process, such heavy-tailed degrees are derived. Communities emerge, cause every new node copies several neighbors of its ambassador. New nodes form a lot of links near the community of the ambassador. During the burning process they also form a few links beyond the community and significantly fewer links in remote parts of the graphs. Such the densification power law is derived. Finally graphs generated by the Forest Fire Model also exhibit a shrinking diameter as they grow with proper choice of parameters.

For easy access and comparison with the synthetic social network graphs we generate random graphs by creating an edge between each pair of nodes with probability p.

4.1.2 Message Rate Model

Neither the Forest Fire Model nor random graphs define rates on links. Furthermore the generated graphs can include node pairs which have links only in one direction, which would not allow the calculation of entropy production. Hence, we distinguish the structure obtained by the graph generation models and the rates. Such we neglect the directions of the edges in the generated graphs to obtain an undirected graph structure. To define rates for an initial evaluation of the entropy production metric we assume limited capacity per user. Such we set the total outgoing rate of each node to a constant c. The total outgoing rate is equally divided among the (undirected) links of the node, such that

$$\sum_{j \in V, i \neq j} w_{ij} = c \tag{10}$$

holds for each node i. Hence, we set the outgoing rate w_{ij} on each link of node i to

$$w_{ij} = \frac{c}{\sum_{j \in V, i \neq j}}. \tag{11}$$

4.2 Numerical Analysis: Common Metrics for Characterizing the Social Networks

To find if the entropy production metric provides additional value for (social) network analysis, we simulate different networks and compare entropy production to existing network metrics. Several metrics measure quantities of nodes in a graph. An overview is given in Section 2. We compare entropy production against two centrality metrics, namely PageRank and betweenness centrality. Another widely used metric that reflects the connectedness of the neighborhood of a node is the clustering coefficient. PageRank was developed to rank websites based on the link structure of the web. According to Moler (2011) PageRank is the limiting probability that an infinitely dedicated random walker visits any particular page. The random walker follows the links randomly according to the probabilities given by the weights of the links. The stationary probability distribution of the websites being visited is basically derived by the eigenvector v of the weighted adjacency matrix $\mathcal{W}$ with eigenvalue $\lambda = 1$:

$$\mathcal{W} v = \lambda v. \tag{12}$$

This solution is extended such that the random walker jumps with a certain probability to a random website to cope with dead ends. For further information refer to the literature, e.g. Page et al. (1999).

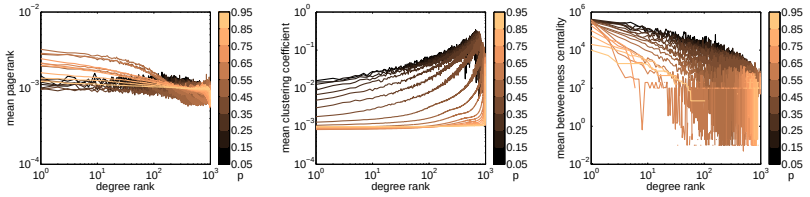
The clustering coefficient of a node quantifies in how far the neighbors of a node are fully meshed. If d_i is the degree of node i, $d_i(d_i - 1)$ is the number of edges that could exist among the nodes in the neighborhood of node i. The local clustering coefficient of node i is then defined as the fraction of actual number of edges in the neighborhood of i and $d_i(d_i - 1)$. For details and a definition of the weighted clustering coefficient please refer to Fagiolo (2007).

Another centrality measure is the betweenness centrality. The betweenness centrality of a node v determines the number of shortest paths in the graph passing v. Let σ_{st} denote the number of shortest paths from $s \in V$ to $t \in V$. $\sigma_{st}(v)$ is the number of shortest paths from s to t that traverse some v. The length of a path is the sum of the weights of its edges. The betweenness centrality for a node v is then defined as:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}. \tag{13}$$

For more details and a definition of weighted betweenness centrality refer to Brandes (2001).

Figure 3 shows the mean scores of graph metrics for each node sorted by decreasing node degree. 20 runs were conducted with burning probability p from 0.05 to 0.95 resulting in



(a) PageRank (b) Clustering Coefficient (c) Betweenness Centrality

Fig. 3 Mean scores of graphs metrics in the Forest Fire model for each node sorted by decreasing node degree. 20 runs were conducted with burning probability p from 0.05 to 0.95 depicted by the color.

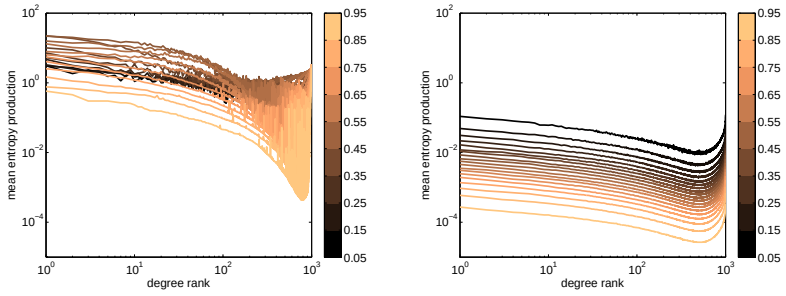
very sparse graphs and dense clique-like graphs, respectively. Each of the lines represents one burning probability depicted by its color. Figure 3(a) shows the mean scores of the PageRank metric. Very sparse graphs show PageRank values independent of the node degree. As the graphs get denser with increasing p, higher degree nodes show higher PageRank scores. This is expected, since such nodes are visited more frequently by a random walker. As p approaches 1 graphs get very dense, such the probability of a random walker hitting high degree nodes decreases, resulting in less difference in the PageRank scores.

Figure 3(b) shows the mean scores of the clustering coefficient. For dense networks the network consists of one big cluster resulting in a clustering coefficient independent of the node degree. If we do not consider edge weights the clustering in a fully meshed network would be 1 for each node. However, we calculate values dependent on the edge weights, which tend to be smaller for dense networks. For the same reason the cluster coefficient increases for sparse networks. Since the edge weights are indirect proportional to the node degree, the weighted clustering coefficient increases with the degree rank. Very low degree nodes in sparse graphs which might not be part of a clique show again lower clustering coefficients.

Figure 3(c) shows the mean scores of the betweenness centrality. As expected the betweenness centrality decreases with the node degree, since high degree nodes are located centrally in the graph. If the graph gets denser central high degree nodes get less important, because the number of shortest paths connecting nodes without traversing central high degree nodes increases. In a very dense graph a node v only is part of a few shortest paths in the graph. The length of a path is the sum of the weights of its edges. If node v has low degree and therefore high edge weights, there might be even shorter paths without v that connect the direct neighbors of v. Hence, in this case v does not lie on any shortest path, resulting in betweenness centrality of 0. Thus the log-log-plot in Figure 3(c) shows no value.

4.3 Numerical Analysis: Entropy Production

In the previous section we saw how existing network metrics rank nodes in Forest Fire networks with varying density. Here we investigate entropy production as a new metric and compare it to the existing metrics. Figure 4 shows the mean scores of the entropy production metric dependent on the degree rank for Forest Fire and random graphs of different density. The mean scores of the entropy production metric in the Forest Fire graph are depicted



(a) Forest Fire Graph (b) Random Graph

Fig. 4 Mean scores of the entropy production metric for Forest Fire and random graphs dependent on the degree rank. 20 runs were conducted with parameter p varying from 0.05 to 0.95 depicted by the color.

in Figure 4(a). In Forest Fire graphs with equal node rates low degree nodes tend to have lower entropy production. High degree nodes have many links with low outgoing rates. Each neighbor having less links and therefore higher rates contributes to the entropy production of the high degree nodes. Thus high degree nodes achieve high entropy production scores. On the other hand very low degree nodes also show high entropy production scores. Low degree nodes have high outgoing rates. If they are connected to nodes with high degree, the entropy production on the links is also high. Here the high value stems from the high entropy production of single links.

The same relation of entropy production and node degree can be observed for random networks, cf. Figure 4(b). Furthermore, in the random graph the entropy production is correlated with the sparsity of a the graph. If the graph is fully meshed, the entropy production is zero, since in our simulation model the rates are then equal for all links and all neighbors, resulting in no entropy being produced. If the graph gets sparser hub nodes with higher degree than the nodes in their neighborhood exist with higher probability that contribute to entropy production, as explained above. The entropy production scores in the random graph are at least one magnitude lower than in the Forest Fire graph. This can be explained by the fact, that the node degrees in the random network follow a binomial distribution with mean $p \cdot N$ with N nodes in the network. Hence, the difference between degrees and rates of neighbored nodes tend to be lower than in the network with heavy-tailed node degree distribution. This leads to less entropy production of the links and nodes in the random network compared to the Forest Fire network.

To show the difference between entropy production and the existing metrics we compare the metrics in Forest Fire graphs with fixed burning probability $p = 0.45$. Figure 5 shows the mean score of the normalized metrics dependent on the entropy production rank. Betweenness centrality seems to have a high correlation with entropy production for nodes that have a high rank. For low rank nodes the value of betweeness centrality fluctuate a lot which let's assume that the scores differ for these nodes.

Clustering coefficient and entropy production seem to be negative correlated. In highly clustered regions node degrees are similar, leading to similar rates and low entropy production scores. Bridge builders that connect different possibly distinct communities show high entropy production but low clustering coefficient.

PageRank and entropy production show a positive correlation. The PageRank scores only have a slight gradient compared to entropy production. Such we can identify a point of interception after which PageRank scores are higher than entropy production. This suggests to take a closer look at the correlation between entropy production and the existing metrics.

4.4 Correlation between Entropy Production and Existing Social Network Metrics

To further study the similarity of entropy production to the existing metrics we investigate the correlation between the entropy production metric and existing metrics. Figure 6 shows the correlation of entropy production with the existing metrics for a fraction of x percent of nodes with highest entropy production scores. For the sake of clarity the results are only plotted for p in range [0.3,0.75].

Figure 6(a) shows the correlation of entropy production with the PageRank score, dependent on the top nodes according to entropy production metrics. PageRank shows low correlation for sparse graphs with low burning probability p. The correlation increases above 0.5 as the graphs get denser, but drops below 0.5 for very dense graphs again. Hence, dependent on the density of the graph PageRank and entropy production are correlated and rank scores similarly or are uncorrelated. This shows that entropy production offers additional value to identify and rank nodes in complex networks.

Figure 6(b) shows the correlation of entropy production with the clustering coefficient, dependent of the percentage of nodes with highest entropy production. As seen in the previous Section 4.3 the clustering coefficient is negative correlated with entropy production for nodes with high entropy score in graphs with $p < 0.6$. The correlation is less than 0.5 if we look at more than the top 10% of nodes with highest entropy production and can argue that the metrics are uncorrelated for these subgraphs.

Figure 6(c) shows the correlation of entropy production with the betweenness centrality metric for the top x-% of nodes with highest entropy production. In the previous section 4.3 we found that betweenness centrality has a high correlation with entropy production in a Forest Fire graph with burning probability $p = 0.45$. As depicted in Figure 6(c), entropy production and betweenness centrality are correlated for sparse graphs with burning probabilities

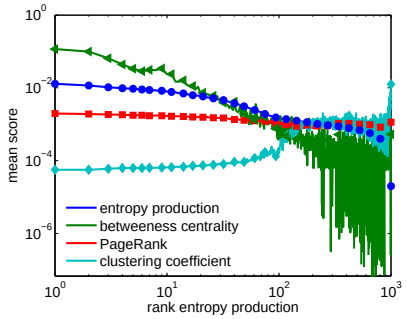
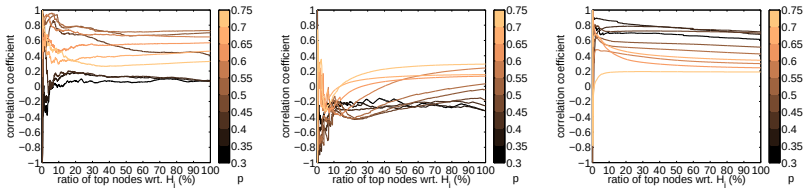
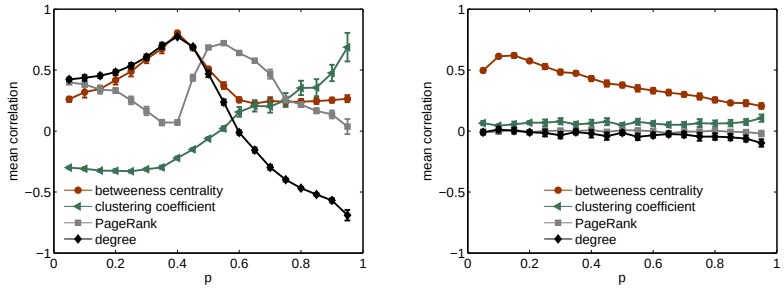


Fig. 5 Normalized score of common metrics dependent on the entropy production rank. Normalized mean value of 20 Forest Fire networks with fixed burning probability $p = 0.45$.



(a) PageRank (b) Clustering Coefficient (c) Betweenness Centrality

Fig. 6 Correlation of entropy production with common metrics for nodes with highest entropy production scores. Correlation has been calculated for Forest Fire graphs with burning probability $p = \{k \cdot 0.05 | k = 6, 7, \ldots, 15\}$.



(a) Forest Fire Graph (b) Random Graph

Fig. 7 Correlation of entropy production with existing metrics dependent on the graph density. Mean correlation together with 95% confidence intervals have been calculated for 20 Forest Fire and random networks.

$p < 0.5$. As the graphs get denser the correlation remains only for the entropy production nodes with high rank. For dense graphs with $p > 0.7$ betweenness centrality and entropy production are uncorrelated.

Hence, the correlation of entropy production with existing metrics in Forest Fire graphs depends heavily on the density of the graph. Figure 7 shows the correlation of entropy production with common metrics dependent on the graph density, given by parameter p. For each p the mean correlation together with 95% confidence intervals have been calculated for 20 Forest Fire and random networks. Figure 7(a) shows the correlation of entropy production for Forest Fire graphs with increasing density. Dependent on the density entropy production is correlated with existing metrics. For instance betweenness centrality and the node degree show high correlation with entropy production if p is in the range $p = [0.3,0.45]$. Interesting is, that if p is in that range most realistic networks are generated. Furthermore, for such realistic networks PageRank is uncorrelated, whereas it is correlated with entropy production for slightly denser graphs. The clustering coefficient is correlated with entropy production only for very dense graphs. For very dense graphs the node degree is also correlated with entropy production but negatively. Figure 7(b) shows the correlation of entropy production for random graphs with increasing density. Except of betweenness centrality entropy production is uncorrelated with the existing metrics independent of the density of the random graph, because there is no structure in the graph. Betweenness centrality is correlated with entropy production in random networks for sparse graphs. The correlation decreases with

the graph density. This depends on the fact that for increasing p the weighted betweenness centrality approaches 0 for more an more nodes, as described in Section 4.2.

All in all we find that entropy production offers additional value to rank nodes of complex networks compared to the existing social network metrics, we see the need for entropy production as new metric.

The correlation of entropy production and betweenness centrality in some especially realistic networks still has to be understood and might help to develop improved algorithms to identify central nodes.

5 Naive Estimate vs. Bayesian Inference

In real networks, the rates for messages between nodes are unknown and have to be estimated from measurements over a certain time period. We show the need for Bayesian inference instead of naïve estimates of those rates, but highlight that for large measurement campaigns over a sufficient time period (or number of messages) the results based on naïve estimate are close to the results from Bayesian inference.

5.1 Naïve Estimate

The entropy production depends on the message numbers n_{ij} and the rates w_{ij}. While the numbers n_{ij} can be determined directly from the given data, the rates w_{ij} are usually not known in advance. Of course, in the limit of infinite observation time the relative frequencies of messages converge to the corresponding rates, i.e.

$$w_{ij} \;=\; \lim_{T\to\infty} \frac{n_{ij}}{T}\,. \tag{14}$$

For finite observation time the count numbers n_{ij} are scattered around their mean value $\langle n_{ij}\rangle = T w_{ij}$. Therefore it is tempting to approximate the entropy production by replacing the ratio of the rates with the ratio of the relative frequencies, i.e.

$$H_{ij}^{\text{naive}} \;\approx\; (n_{ij}-n_{ji}) \ln \frac{n_{ij}}{n_{ji}}\,. \tag{15}$$

However, this naïve estimator is useless for two reasons. Firstly, the nonlinear logarithm does not commute with the linear average and is thus expected to generate systematic deviations. Secondly, in realistic data sets there may be one-way communications with $n_{ij} > 0$ and $n_{ji} = 0$, producing diverging contributions in the naïve estimator (15). However, observing no messages in opposite direction does not mean that the actual rate is zero, it only means that the rate is small. In the following we suggest a possible solution to this problem by using standard methods of Bayesian inference, following similar ideas recently addressed in a different context (Zeraati et al. 2012).

5.2 Bayesian Inference

As the messages are assumed to occur randomly and independently like the clicks of a Geiger counter, we expect the number of messages n for a given rate w to be distributed according to the Poisson distribution

$$P(n|w) \;=\; \frac{(Tw)^n e^{-Tw}}{n!}\,, \tag{16}$$

where T is the observation time. But instead of n for given w, we need an estimate of the rate w for given n. According to Bayes formula (Box and Tiao 1973) the corresponding conditional probability distribution is given by the posterior

$$P(w|n) \;=\; \frac{P(n|w)P(w)}{P(n)}\,, \tag{17}$$

where $P(w)$ is the prior distribution and

$$P(n) = \int_0^\infty dw\, P(n|w)P(w) \tag{18}$$

is the normalizing marginal likelihood. The prior distribution expresses our belief how the rates are statistically distributed and introduces an element of ambiguity as will be discussed below. Having chosen an appropriate prior the expectation value $\langle \ln w\rangle$ for given n reads

$$\langle \ln w\rangle_n \;=\; \int_0^\infty dw\, \ln w P(w|n)\,. \tag{19}$$

This allows us to estimate the entropy production of the directed link $i \to j$ by

$$H_{ij} \;\approx\; (n_{ij} - n_{ji})\Big[\langle \ln w\rangle_{n_{ij}} - \langle \ln w\rangle_{n_{ji}}\Big]\,. \tag{20}$$

As we will see, this estimator does not diverge if $n_{ji} = 0$.

5.3 Difference: Naive Estimate vs. Bayesian Inference

In the previous section we described a naïve estimate H^{naive} of the entropy production using the relative frequencies of messages. The naïve estimator diverges, if we count a finite number of messages n_{ij} from i to j in the observation period in one direction but count $n_{ji} = 0$ messages in the opposite direction. To cope with this problem we set $n_{ji} := \epsilon$ assuming that a small number of messages $\epsilon > 0$ was observed in the opposite direction. Such $\frac{n_{ij}}{n_{ji}}$ is a finite value, making sure that H_{ij}^{naive} does not diverge.

To compare the naïve estimate with Bayesian inference, we conduct another experiment using Forest Fire and random graphs. We use the graphs with fixed burning probability $p = 0.45$ and use fixed node capacity and equal outgoing rates $w_{ij} = w_i$ as in Section 4.

We simulate an observation of the communication in these networks by generating messages according to the edge weights over a given period of time M. We assume that a node i sends a message with probability w_i in each time step $t = 1,2,...,M$. Such the number of messages sent by a node with outgoing link rates w_i over the period of time M follows a binomial

distribution with parameters M and w_i and mean $M \cdot w_i$. Such for each run of the simulation we draw a random number $X \sim B(M,w_i)$ for each link ij that follows a binomial distribution according to its rate w_{ij}. If X is greater than zero we set $n_{ij} = X$, otherwise we set $n_{ij} = \epsilon$.

We calculate the naïve estimates H_{ij}^{naive} according to equation Eq. (15). Figure 8 shows the the results of a Forest Fire graph and a random graph with $p = 0.45$. To investigate the impact of the length of the observation period, we simulate the periods $M = 100$ and $M = 1000$. The length of the simulation period is depicted by the color. Furthermore we try to identify the impact of ϵ on the estimate. Figure 8(a) depicts the naïve estimate of the entropy production H^{naive} for each node in the Forest Fire graph dependent on the rank of the entropy production derived by Bayesian inference. Regardless of ϵ and the observation period, the naïve estimate tends to decrease with the entropy rank. That means, that H^{naive} is a useful estimate for the entropy production in Forest Fire like networks. Figure 8(b) shows the same results for random graphs. Obviously in random networks the estimator does not work, since mean values of H^{naive} seem to be independent from the entropy rank. If we look at the Forest Fire graph and compare a short observation period $M = 100$ with a long observation period $M = 1000$ we can see that the naïve entropy production scores lie closer together and that the estimates fluctuate less for high entropy nodes. That means that the accuracy of the estimator can be increased if the observation period is extended.

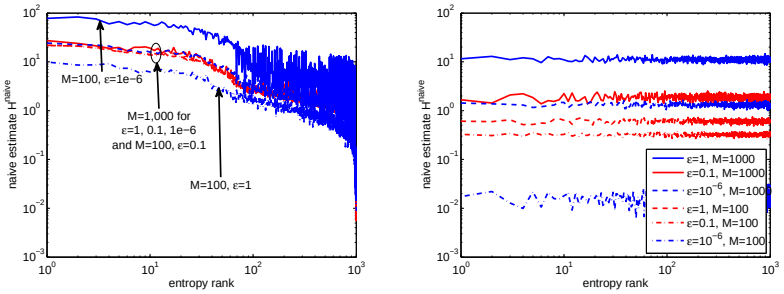
To further investigate the impact of ϵ on the naïve estimate and to compare it to the Bayesian inference, we calculate the absolute difference of the naïve estimate and the Bayesian inference for a short observation period $M = 10$ and varying $\epsilon = \{1, 10^{-1}, 10^{-6}\}$. Figure 9 shows the results for (a) a Forest Fire graph and (b) a random graph with fixed $p = 0.45$. One would expect that the difference of the estimate increases if we use a large ϵ, since we use a relative large value for a rate that is expected to be close to zero. But that is not the case. The results for the random graph show exactly the contrary behavior, the absolute difference increases, while ϵ decreases. This can be explained by the fact that it frequently happens that in the short observation period no message is recorded, although it has a decent rate. Such, $\epsilon = 1$ would be the best estimate. Furthermore the quotient $\frac{n_{ij}}{\epsilon}$ can get very large for small ϵ, leading to a large difference if no message is recorded. Why the absolute difference first decreases with epsilon for values $\epsilon < 1$ but then gets higher for very small $\epsilon = 10^{-6}$ still has to be cleared. The probability that no message is recorded in the observation period depends on the message rate w_i and the length of the observation period M is $P(X = 0) = (1 - w_i)^M$.

6 Retrieving Weights from Measurements on the Example of Small World Networks

6.1 Choice of the Prior Distribution in Bayesian Inference

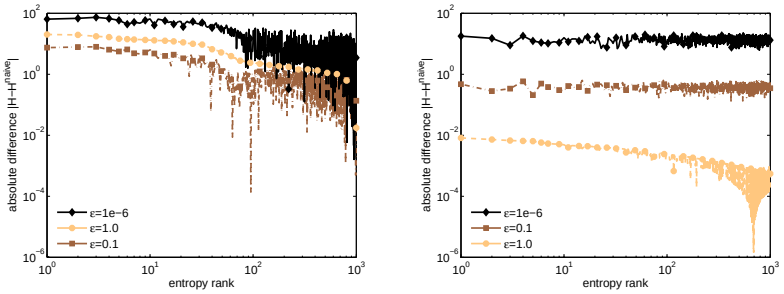
The prior should be as much as possible in accordance with the available data. In the example to be discussed below, where we investigate a small-world network with message numbers distributed according to Eq. (8) with an exponent $\alpha > 1$, it would be natural to postulate a power-law distribution of the rates $P(w) \sim w^{-1-\alpha}$. Since such a distribution can only be normalized with a suitable lower cutoff, a natural choice for the prior would be the inverse gamma distribution

$$P(w) = \frac{\beta^\alpha w^{-\alpha-1} e^{-\beta/w}}{\Gamma(\alpha)}, \tag{21}$$



(a) Forest Fire Graph (b) Random Graph

Fig. 8 Naïve estimate of the entropy production dependent on rank of actual entropy production. Comparison of a short observation period $M = 100$ with a long observation period $M = 1000$ with different ϵ in a Forest Fire and a random graph with fixed $p = 0.45$.



(a) Forest Fire Graph (b) Random Graph

Fig. 9 Absolute difference of the naïve estimate for entropy production for a short observation period with $M = 10$. The absolute difference is calculated for different $\epsilon = \{1, 10^{-1}, 10^{-6}\}$ in a Forest Fire and a random graph with fixed $p = 0.45$.

where the parameter β plays the role of a lower cutoff for the rate w. With this prior distribution the integration can be carried out, giving the posterior

$$P(w|n) \;=\; \frac{(\beta/T)^{\frac{\alpha-n}{2}}\, w^{n-\alpha-1} e^{-Tw-\frac{\beta}{w}}}{2K_{n-\alpha}(z)}, \tag{22}$$

where $K_\nu(z)$ is the modified Bessel function of the second kind and $z = 2\sqrt{\beta T}$. Inserting this result into Eq. (19) we obtain an estimate of $\ln w$ for given n, namely

$$\langle \ln w \rangle_n \;=\; \frac{1}{2}\ln\frac{\beta}{T} + \frac{K^{(1,0)}_{n-\alpha}(z)}{K_{n-\alpha}(z)}, \tag{23}$$

where $K^{(1,0)}_\nu(z) = \frac{\partial}{\partial\nu}K_\nu(z)$. The estimator for the entropy production is given by

$$H_{ij} \;\approx\; (n_{ij} - n_{ji})\Big[\frac{K^{(1,0)}_{n_{ij}-\alpha}(z)}{K_{n_{ij}-\alpha}(z)} - \frac{K^{(1,0)}_{n_{ji}-\alpha}(z)}{K_{n_{ji}-\alpha}(z)}\Big]. \tag{24}$$

6.2 Estimating the Cutoff Parameter $z = 2\sqrt{\beta T}$

Eq. (23) depends on the exponent α, which can be obtained from the distribution of messages per link, and the lower cutoff β, which can be determined as follows. On the one hand, the probability to have no link between two nodes for a given rate w is $1 - P(0|w)$. Therefore, the total number of links L can be estimated by

$$L \approx L_{\rm tot} \int_0^\infty dw \left(1 - P(0|w)\right) P(w) = L_{\rm tot}\Big(1 - \frac{2(T\beta)^{\alpha/2}}{\Gamma(\alpha)} K_\alpha(2\sqrt{T\beta})\Big)\,. \tag{25}$$

Here $L_{tot} = N_{\rm tot}(N_{\rm tot} - 1)$ is the unknown total number of potential links in the stationary network which may exceed the actual number of links L established during the finite observation time T. On the other hand, it is obvious that the total number of messages M can be estimated by

$$M \approx L_{\rm tot} \sum_{n=0}^{\infty} n \int_0^\infty dw\, P(n|w) P(w) = L_{\rm tot} \frac{T\beta}{\alpha - 1}\,. \tag{26}$$

This relation can be used to eliminate $L_{\rm tot}$, turning Eq. (25) into

$$\frac{L\, z^2}{4M(\alpha - 1)} \;\approx\; 1 - \frac{2(z/2)^\alpha}{\Gamma(\alpha)} K_\alpha(z)\,. \tag{27}$$

For given M,L,α this approximation interpreted as an equation allows us to numerically determine $z = 2\sqrt{\beta T}$.

6.3 Summary of the Procedure

The procedure to calculate the entropy production can be summarized as follows.

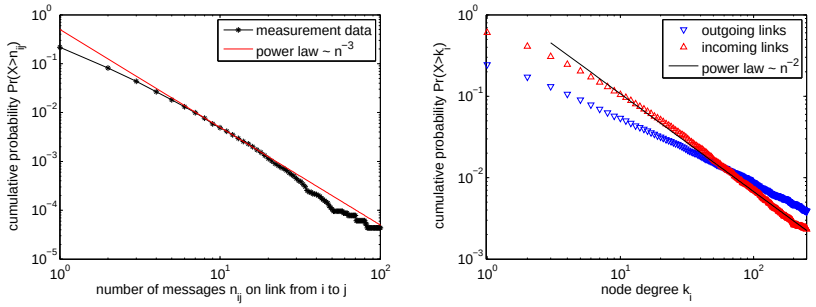
1. In the given data set of M messages, identify all participants (nodes) and label them from $1,\ldots,N$.
2. Determine the numbers n_{ij} how often a message is sent from i to j and count the number L of nonzero entries (links) in the matrix $\{n_{ij}\}$.
3. Plot a histogram of the numbers n_{ij}. If it exhibits a power law $P(n) \sim n^{-1-\alpha}$ estimate the exponent α.
4. Solve Eq.(27) numerically for z.
5. Compute the numbers $\chi_n = K^{(1,0)}_{n-\alpha}(z)/K_{n-\alpha}(z)$.
6. Associate with each directed link $i \to j$ the entropy production $H_{ij} = (n_{ij} - n_{ji})(\chi_{n_{ij}} - \chi_{n_{ji}})$.
7. Compute H_i and H according to Eqs. (4) and (5).

6.4 Real World Example: R Mailing List Archive

To demonstrate the concepts introduced above, we analyzed the mailing lists archive for the programming language **R** (R Mailing Lists 2013), recording senders and receivers of all messages over the past 15 years.

6.4.1 Network Characteristics of R Mailing List Results

In this mailing list $N = 23462$ individuals (nodes) have exchanged $M = 168778$ directed comments (undirected activities like opening a new thread are ignored). The connectivity matrix n_{ij} has $L = 114713$ nonzero entries (links). Their statistical distribution shown in Fig. 10 confirms a small-world topology with $\alpha \approx 2$. Interestingly, the node degree distribution of outgoing and incoming links in Fig. 10(b) seems to exhibit slightly different exponents. A similar phenomenon was observed some time ago in email communication networks (Ebel et al. 2002) with an in-degree and out-degree power-law exponent of 1.49 and 2.03, respectively. In a later work, Bird et al. (2006) provides similar result on the node degree for the email social network on the Apache HTTP server project. The in-degree and out-degree distribution show typical long-tailed, small world characteristics. In general, Mislove et al. (2007) analyzes different social networks including Flickr, YouTube, and LiveJournal and they show behavior consistent with a power-law network. However, social networks show much stronger correlation between in-degree and out-degree than Web graphs. Broder et al. (2000) consider the connection of web sites and also observe a power-law of in-degree and out-degree with an exponent of 2.1 and 2.72, respectively. Thus, in-degree and out-degree powerlaw exponents differ significantly, while the study from Mislove et al. (2007) for social networks shows that they are quite similar. Thus, the distribution of outgoing links and incoming link is similar, while in the web the incoming links are significantly more concentrated on a few high-degree nodes than the outgoing links.



(a) Messages per Link (b) Links per Node

Fig. 10 (a) Probability $Pr(X > n_{ij})$ that a directed link $i \to j$ carries more than n_{ij} messages. The data is consistent with a power law $P(n_{ij}) \sim n_{ij}^{-3}$. (b) Probability $Pr(X > k_i)$ that a node i is connected with more than k_i outgoing or incoming links. For incoming links, it is $P(k_i) \sim k_i^{-2}$.

From that perspective, the network characteristics of the R mailing list under investigation here as real-world example appears as a typical social network. The correlation between the in-degree and the out-degree per node is very strong with 0.949, see Table 1 which was also observed by Bird et al. (2006) (0.971) for the Apache HTTP server project mailing list.

The R mailing list leads to a very sparse graph which is quantified in terms of density measuring how close the network is to complete. A density value of 1 describes a complete graph with all possible edges, however, the R mailing list leads to $2.084 \cdot 10^{-4}$. The diameter of the network is 9 and the average path length is 3.256. The average clustering coefficient being computed on the directed R mailing list graph with message rates as link weights is 0.272 which is inline with the results from Chen et al. (2006) who derived a value of 0.267 for the W3C mailing list archive. Those results clearly emphasize the small world characteristics of the mailing lists. From the power-law exponent of the node degree, the corresponding parameters for a comparable Forest Fire graph can be estimated according to the results given by Leskovec et al. (2005). For the R mailing list, we obtain roughly $p = 0.4$ and $a = 2$ which indicates a densifying graph with slowly decreasing diameter.

The results of the synthetic graphs in Section 4 showed for sparse graphs and values of $p < 0.5$ a high correlation between betweeness centrality and entropy production per node. Therefore, we also expect this result for the R mailing list which is discussed in Section 6.4.2.

Beside the network topology, the message rates between users describe the activity in the network and determine entropy production. We have analyzed the correlation between the node degree d_i and the message rates for all nodes i which is 0.996 and 0.972 for incoming and outgoing messages, respectively. In contrast to the simplistic model in Section 4 used for illustration, we observe now strong correlations between the message rates and the network topology in the real-world example. Similiar to Bird et al. (2006), we als see a strong dependency between the number of messages sent and the number of different users responding to them. As a result, we observe a correlation coefficient of 0.933. As an important characteristic of the message dissemination we take a closer look whether the messages of a user are equally distributed over all links to neighbors. Figure 11 shows the complementary cumulative distribution function (CCDF) of the message rate for any node's links. To be more precise, for each node i, we consider the message rates X on all links to i's neigh-

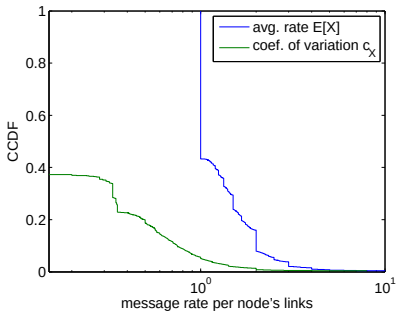


Fig. 11 For each node i, we consider the message rates on the links to neighbors and compute the mean message rate $E[X]$ per link and the coefficient of variation c_X. We can observe strong deviations of the message rates.

bors and calculate the average message rate $E[X]$ per link and the coefficient of variation c_X over those links. It can be seen in Fig. 11 that there are many nodes with the minimum message rate $E[X] = 1$. The minimum value occurs as there is at least one message required to establish a link between two users. On the other hand, there are a few users accounting for the bulk of messages sent or received which is also a typical phenomenon in such small world networks (Bird et al. 2006). The heterogeneity of a node's message dissemination to its neighbors is indicated by the coefficient of variation c_X in Fig. 11. For 37.87 % of all nodes, there is only a single neighbor, i.e. a node degree of 1, and therefore no variation in the message rate, i.e. $c_X = 0$. The large values of c_X show that there are strong differences in the communication of a user with its neighbors and that the total message rate per user is rather heterogeneously split. Therefore, we expect large differences of the entropy production which quantifies those dynamical properties and user interactions.

6.4.2 Entropy Production in R Mailing List

Since the entropy production of a link H_{ij} depends on two integers n_{ij} and n_{ji}, the entropy production H_{ij} produces a discrete set of values. The upper panel of Fig. 12 shows how these values are distributed and how often they occur. As can be seen, the entropy production varies over five orders of magnitude and is distributed irregularly with count numbers ranging from 1 to 10^5.

Let us now turn to the question how the node entropy $H_i = \frac{1}{2}\sum_{j=1}^{N} H_{ij}$ is correlated with the number of outgoing and incoming messages $n_i^{\text{out}} = \sum_j n_{ij}$, $n_i^{\text{in}} = \sum_j n_{ji}$. Since the entropy production is expected to grow with the number of messages, it is reasonable to define the node entropy production per message

$$h_i := \frac{H_i}{n_i^{\text{out}} + n_i^{\text{in}}}. \tag{28}$$

Figure 13 shows how the entropy production per node is distributed depending on the number of sent and received messages. As expected, the entropy is minimal if these numbers coincide. Plotting the entropy production of a node versus the difference of outgoing and incoming messages $\Delta n_i = n_i^{\text{out}} - n_i^{\text{in}}$ one finds again an asymmetric distribution (see right panel). This indicates that nodes with a large number of outgoing links tend to produce less entropy per message than individuals who preferentially receive messages.

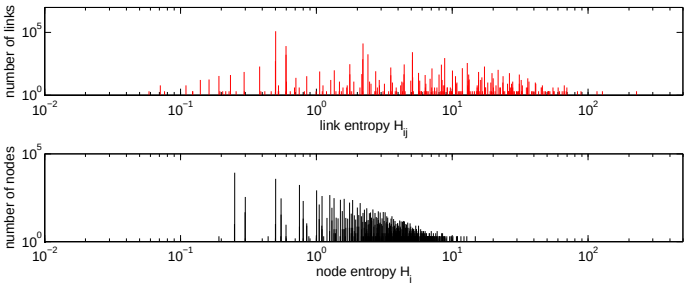


Fig. 12 Upper panel: Histogram of link entropy H_{ij}. Lower panel: Histogram of node entropy H_i.

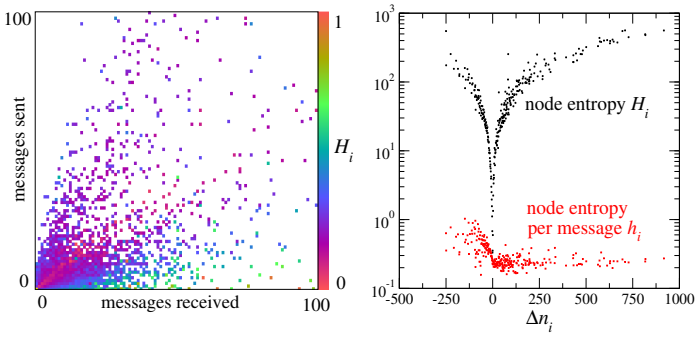


Fig. 13 Left: Average entropy production per node represented by a continuous color scale in a two-dimensional plane spanned by the number of incoming and outgoing messages. Right: Average entropy production of a node (black) and the same data divided by the total number of incoming and outgoing messages (red) as a function of the difference between outgoing and incoming messages.

Due to the definition of entropy production, there seems to be a strong correlation between Δn_i and H_i which is however not valid in general. For a fixed value of Δn_i, the entropy production is determined by the factor $\ln \frac{w_{ij}}{w_{ji}}$, i.e., the absolute number of incoming and outgoing messages impacts H_{ij}. Nevertheless, for the provided R mailing data, we observe a clear relationship between Δn_i and the entropy production per node H_i. In particular, the correlation coefficient is very high with 0.940.

We have investigated additional correlations between H_i and other node metrics which are summarized in Table 1 (column 'corr. with H_i). As a result, very high correlation coefficients between different centrality measures and entropy production are observed. For example, betweeness centrality and H_i lead to a Pearson's linear correlation coefficient of 0.940. Due to the properties of the R topology, this was an expected result according to Section 4. However, it has to be noted that the real message rates between nodes are now taken into account which are strongly varying, see Figure 11, in contrast to the equal contact model for message rates in Section 4. We conclude that for typical small world networks like the email social network we observe a strong correlation. The fast computation of entropy production in $O(n)$ allows therefore to approximate centrality measures like PageRank or betweenness centrality. Approximation by Brandes (2001) to compute betweeness centrality of weighted graphs requires time $O(nm) + n^2 \log n$ for n vertices and m edges. Entropy production also

Table 1 Pearson's linear correlation coefficient of entropy production H_i, out degree d_i^{out}, and in degree d_i^{in} with other metrics.

node metric	corr. with H_i	corr. with d_i^{out}	corr. with d_i^{in}
in degree	0.9173	0.9492	1.0000
out degree	0.9591	1.0000	0.9492
degree	0.9575	0.9963	0.9726
eccentricity	0.0525	0.0778	0.1214
closeness centrality	0.0390	0.0608	0.0969
betweenness centrality	0.9804	0.9524	0.9001
pagerank	0.9404	0.9499	0.9615
eigenvector centrality	0.6548	0.7065	0.8573
clustering coefficient	-0.0227	-0.0393	-0.0014

outperforms novel approaches as suggested by Kourtellis et al. (2013) in order to identify high betweeness centrality nodes in large scale networks. Therefore, entropy production may improve existing applications e.g. on real-time monitoring of social networks or on the identification of leaders or spammers in social networks (Fazeen et al. 2011).

7 Conclusions

In the current Internet, social networks like Facebook gain more and more popularity and attract millions of users. In a social network the users are connected to each other and as a key feature social media platforms allow interactions between users like exchange of messages. In complex network research, the majority of existing quantities analyze the structural properties of the emerging network topology and the growth dynamics, respectively. For social networks, however, beyond the network topology the interaction among individuals needs to be characterized. Further, the dynamics of a stationary state of a social network is not uniquely given, rather there is a large variety of possible realizations. Hence, there is a gap in describing dynamical properties of social networks in stationary conditions, i.e., when the network topology as well as the probability for receiving and sending messages do not change in the long-term limit. Inspired from statistical physics, we introduce a quantity called entropy production to characterize the stationary properties of arbitrary social networks. The entropy production measures the directionality of the information exchange and vanishes for perfectly balanced communication. Defining the entropy production of a node as the sum over the entropy of all its links, one can identify nodes contributing preferentially to balanced or unidirectional information transfer. Hence, entropy production is a valuable measure for link and node analysis and rating and can be used to detect hidden structures and interactions in networks. As first major contribution, we find that entropy production offers additional value to rank nodes of complex networks compared to the existing social network metrics. In particular, we found a high correlation of entropy production and betweenness centrality for especially realistic networks that may allow a fast computation and identification of central nodes. The second major contribution describes a procedure how to compute entropy production for measurements obtained from real networks by taking into account Bayesian inference and to which degree naïve estimates can be used for close approximation.

Since the application of entropy production is not limited to social media networks, but can be used for communication networks or interaction graphs in general, it can be applied for a variety of different purposes like anomaly detection (Bilgin and Yener 2010), leader and spammer detection in communication and social networks (Fazeen et al. 2011), but also characterization of traffic flows in the Internet, e.g. for BitTorrent swarms (Hoßfeld et al. 2011b). Future work addresses the application of entropy production to such use cases but also to analytically derive the interdependency between the quantity with centrality or symmetry measures for various network topologies and message exchange models.

References

D. Andrieux and P. Gaspard. Fluctuation theorem and onsager reciprocity relations. *J Chem Phys*, 121(13), 2004.

D. Andrieux, P. Gaspard, S. Ciliberto, N. Garnier, S. Joubaud, and A. Petrosyan. Entropy production and time asymmetry in nonequilibrium fluctuations. *Phys. Rev. Lett.*, 98:150601, 2007.

N. G. Barnes and J. Andonian. The 2011 fortune 500 and social media adoption: Have america's largest companies reached a social media plateau?, 2011. `http://www.umassd.edu/cmr/socialmedia/2011fortune500/`.

C. Bilgin and B. Yener. Dynamic network evolution: Models, clustering, anomaly detection. Technical report, Rensselaer University, NY, 2010.

C. Bird, A. Gourley, P. Devanbu, M. Gertz, and A. Swaminathan. Mining email social networks. In *Proceedings of the 2006 international workshop on Mining software repositories*, pages 137–143. ACM, 2006.

G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. John Wiley & Sons, New York, 1973.

U. Brandes. A faster algorithm for betweenness centrality*. *Journal of Mathematical Sociology*, 25(2): 163–177, 2001.

A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer networks*, 33(1):309–320, 2000.

H. Chen, H. Shen, J. Xiong, S. Tan, and X. Cheng. Social network structure behind the mailing lists: Ict-iiis at trec 2006 expert finding track. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006*, volume Special Publication 500-272. National Institute of Standards and Technology (NIST), 2006.

M. Dehmer and A. Mowshowitz. A history of graph entropy measures. *Information Sciences*, 181(1):57–78, 2011.

H. Ebel, L.-I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Phys. Rev. E*, 66, Sep 2002. doi: 10.1103/PhysRevE.66.035103.

G. Fagiolo. Clustering in complex directed networks. *Physical Review E*, 76(2):026107, 2007.

M. Fazeen, R. Dantu, and P. Guturu. Identification of leaders, lurkers, associates and spammers in a social network: context-dependent and context-independent approaches. *Social Network Analysis and Mining*, 1(3):241–254, 2011.

A. Garrido. Symmetry in complex networks. *Symmetry*, 3(1), 2011.

F. Gilbert, P. Simonetto, F. Zaidi, F. Jourdan, and R. Bourqui. Communities and hierarchical structures in dynamic social networks: analysis and visualization. *Social Network Analysis and Mining*, 1(2):83–95, 2011.

J. Gómez-Gardeñes and V. Latora. Entropy rate of diffusion processes on complex networks. *Physical Review E*, 78(6):065102, 2008.

M. Hirth, F. Lehrieder, S. Oberste-Vorth, T. Hoßfeld, and P. Tran-Gia. Wikipedia and its Network of Authors from a Social Network Perspective. In *International Conference on Communications and Electronics (ICCE)*, Hue, Vietnam, 2012.

T. Hoßfeld, M. Hirth, and P. Tran-Gia. Modeling of Crowdsourcing Platforms and Granularity of Work Organization in Future Internet. In *International Teletraffic Congress (ITC)*, San Francisco, USA, 2011a.

T. Hoßfeld, F. Lehrieder, D. Hock, S. Oechsner, Z. Despotovic, W. Kellerer, and M. Michel. Characterization of BitTorrent Swarms and their Distribution in the Internet. *Computer Networks*, 55(5), 2011b.

E. M. Jin, M. Girvan, and M. E. J. Newman. Structure of growing social networks. *Phys. Rev. E*, 64, 2001.

G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.

N. Kourtellis, T. Alahakoon, R. Simha, A. Iamnitchi, and R. Tripathi. Identifying high betweenness centrality nodes in large social networks. *Social Network Analysis and Mining*, 3(4):899–914, 2013.

J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187, 2005.

T. Mihaljev, L. de Arcangelis, and H. J. Herrmann. Interarrival times of message propagation on directed networks. *Phys. Rev. E*, 84:026112, 2011.

A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, 2007.

C. Moler. Experiments with matlab. *The MathWorks, Co*, 2011.

A. Mowshowitz and M. Dehmer. A symmetry index for graphs. *Symmetry: Culture and Science*, 21(4), 2010.

A. Mowshowitz and M. Dehmer. Entropy and the complexity of graphs revisited. *Entropy*, 14(3):559–570, 2012.

L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999.

S. M. Pincus and W.-M. Huang. Approximate entropy: statistical properties and applications. *Communications in Statistics-Theory and Methods*, 21(11):3061–3077, 1992.

R Mailing Lists. `http://tolstoy.newcastle.edu.au/R/`, Jan. 2013.

A. Sallaberry, F. Zaidi, and G. Melançon. Model for generating artificial social networks having community structures with small-world and scale-free properties. *Social Network Analysis and Mining*, 3(3):597–609, 2013.

J. Schnakenberg. Network theory of microscopic and macroscopic behavior of master equation systems. *Rev. Mod. Phys.*, 48, 1976.

T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85, 2000.

U. Seifert. Entropy production along a stochastic trajectory and an integral fluctuation theorem. *Phys. Rev. Lett.*, 95, 2005.

R. Sinatra, J. Gómez-Gardeñes, R. Lambiotte, V. Nicosia, and V. Latora. Maximal-entropy random walks in complex networks with limited information. *Phys. Rev. E*, 83:030103, 2011.

D. Smilkov and L. Kocarev. Influence of the network topology on epidemic spreading. *Phys. Rev. E*, 85: 016114, 2012.

C. Tietz, S. Schuler, T. Speck, U. Seifert, and J. Wrachtrup. Measurement of stochastic entropy production. *Phys. Rev. Lett.*, 97:050602, 2006.

M. Vasudevan and N. Deo. Efficient community identification in complex networks. *Social Network Analysis and Mining*, 2(4):345–359, 2012.

J. Wang and P. De Wilde. Properties of evolving e-mail networks. *Phys. Rev. E*, 70:066121, 2004.

J. West, L. Lacasa, S. Severini, and A. Teschendorff. Approximate entropy of network parameters. *Phys. Rev. E*, 85:046111, 2012.

Y.-H. Xiao, W.-T. Wu, H. Wang, M. Xiong, and W. Wang. Symmetry-based structure entropy of complex networks. *Physica A: Statistical Mechanics and its Applications*, 387(11), 2008.

S. Zeraati, F. H. Jafarpour, and H. Hinrichsen. Entropy production of nonequilibrium steady states with irreversible transitions. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(12):L12001, 2012.

C. Zhu, A. Kuh, J. Wang, and P. De Wilde. Analysis of an evolving email network. *Phys. Rev. E*, 74:046109, 2006.

Appendix: Notion of Variables

	variables describing the measurement data
T	measurement period over which messages between individuals are recorded
n_{ij}	number of messages sent from i to j during time T
N	total number of individuals, i.e. nodes in the graph, communicating during T
M	total number of recorded messages, i.e. directed link, $M = \sum_{i,j=1}^{N} n_{ij}$
$\delta_{a,b}$	Kronecker delta defined by $\delta_{a,b} = \begin{cases} 1 & a = b \\ 0 & a \neq b \end{cases}$
L	total number of directed links, $L = \sum_{i,j=1}^{N} 1 - \delta_{0,n_{ij}}$
n_i^{out}	number of outgoing messages from node i, $n_i^{\text{out}} = \sum_{j=1}^{N} n_{ij}$
n_i^{in}	number of incoming messages to node i, $n_i^{\text{in}} = \sum_{j=1}^{N} n_{ji}$
d_i^{out}	outgoing degree of node i
d_i^{in}	incoming degree of node i
d_i	degree of node i
Δn_i	difference of outgoing and incoming messages of node i
$P(n)$	probability that n messages are sent on a link, $P(n) = \sum_{i,j=1}^{N} \delta_{n,n_{ij}}/N(N-1)$
$\mathcal{A}$	adjacency matrix with matrix elements $\mathcal{A}_{ij} = 1 - \delta_{0,n_{ij}}$
	variables describing entropy production
w_{ij}	message rate from i to j estimated by measured n_{ij} over T
$\mathcal{W}$	rate matrix with matrix elements $\mathcal{W}_{ij} = w_{ij}$
ΔH_{ij}	amount of entropy increased for each message sent from i to j, $\Delta H_{ij} = \ln \frac{w_{ij}}{w_{ji}}$
H_{ij}	entropy production per link, $H_{ij} = \left(n_{ij} - n_{ji}\right) \ln \frac{w_{ij}}{w_{ji}}$
H_i	entropy production per node i, $H_i = \frac{1}{2} \sum_{j=1}^{N} H_{ij}$
H	entropy production of total network, $H = \sum_{i=1}^{N} H_i$
h_i	node entropy production per message, $h_i = \frac{H_i}{n_i^{\text{out}} + n_i^{\text{in}}}$
	variables for estimating message rates
$P(w\|n)$	posterior distribution of message rates w conditional on observed messages n
$P(w)$	prior distribution of message rates; assumed to follow a power law in social networks with $P(w) \sim w^{-1-\alpha}$; normalization with suitable lower cutoff leads to inverse gamma distribution $P(w) = \frac{\beta^\alpha}{\Gamma(\alpha)} w^{-\alpha-1} e^{-\beta/w}$
α	shape parameter of the inverse gamma distribution
β	lower cutoff parameter for the rate w concerning inverse gamma distribution
$P(n)$	normalizing marginal likelihood
$\langle \ln w \rangle_n$	expectation value for given n, $\langle \ln w \rangle_n = \int_0^\infty dw \, \ln w P(w\|n)$
$K_\nu(z)$	modified Bessel function of the second kind and $z = 2\sqrt{\beta T}$