Notes on the Exponential Random Graph Model: A Contribution to the Critique of Interdisciplinarity

Paolo Dini[1,2]

[1]Department of Media and Communications
London School of Economics and Political Science
e-mail: p.dini@lse.ac.uk

[2]School of Computer Science
University of Hertfordshire, Hatfield, United Kingdom

December 13, 2021

Abstract

A tutorial discussion is presented about the Exponential Random Graph Model (ERGM) of Social Network Analysis (SNA). The intended audience is post-graduate students and researchers in social science who are curious to understand better where quantitative models come from, for a more effective integration with qualitative methodologies. The discussion is traced back to Jaynes's distinction between objective and subjective interpretations of probability, where the former emphasizes likelihood of outcomes based on frequency distributions, while the latter emphasizes our incomplete knowledge or uncertainty about the outcome. Although within information theory both views lead to the same functional form for information entropy, when applying these concepts to graph theory the paper shows that the subjective view leads to the ERGM, while the objective view yields a different functional form for the 'graph entropy'. It is hoped that the critical perspective on interdisciplinarity developed throughout the paper lends credibility and insight to the conclusion that the subjective view of graph entropy is justified as an optimization principle for the most likely distribution of different graph metrics.

1 Introduction

The main motivation for this paper is to serve as a tutorial on the exponential random graph model (ERGM) for social networks analysis (SNA), in particular to complement the discussion of this model presented by Newman [6]. The intended audience is post-graduate students and researchers in social science who are curious to understand better where quantitative models come from, for a more effective integration with qualitative methodologies.

Although in [5] the point is made that there are many ERGMs, we refer here to the whole class of such models as a singular ERGM on the grounds that they all share the same mathematical structure and may vary only with respect to the number of parameters and network statistics

used to define them. In other words, in this paper ERGM refers to the general mathematical model and not to a specific instance used to model a specific network.

Whereas Newman's presentation is very clear and provides both motivation and mathematical details, it does not explain or derive the most fundamental equation upon which the ERGM rests, the equation for Gibbs's entropy (Eq. (15.22) in Newman [6]):

$$S = -\sum_{i=1}^{M} p_i \ln(p_i), \tag{1}$$

where ln stands for natural logarithm and the rest of the variables will be defined later.

Plausibly, Newman justifies this omission on the grounds that it would take the discussion 'some way away from our central topic of networks' ([6]: 568). However, the problem is that across information theory and physics the concept of entropy is amenable to different interpretations. Further, to these different interpretations there correspond different mathematical derivations of the same equation. Consequently, the use of Gibbs's entropy in the context of graph theory is confusing at several levels. First, it suggests that the physical concept of entropy is relevant, whereas this is not the case: it is the information theory concept of entropy as information uncertainty that applies. Second, the information theory derivation of the latter leads to the same functional form as Gibbs's entropy, but when the same derivation is applied to a graph theory context a different expression for entropy results. Third, the maximization of uncertainty with Lagrange multipliers is conceptually sound but harks back to the derivation of the canonical distribution of statistical mechanics, which is the original application of Gibbs's entropy in physics. This again suggests a link with physics, whereas the more important and relevant link of graph entropy is with information theory, in spite of any possible mismatch in functional form. Therefore, an especially careful discussion is necessary to keep these threads of interpretation distinct.

These problems with interpretations can be traced back to a long-standing debate between the 'subjective' and 'objective' interpretation of probability, for example as discussed by Jaynes [2, 3] and in an in-depth review by Uffink [11]. Put very simply, the objective view of probability is more relevant to physics and emphasizes likelihood of outcomes based on frequency distributions. The subjective view, also relevant to physics, emphasizes our incomplete knowledge or uncertainty about the outcome. When transported to information theory, the objective view yields a measure of information content or most efficient encoding in terms of word length in bits, whereas the subjective view maintains the same interpretation of information uncertainty. Finally, when applying these concepts to graph theory the ERGM derivation again relies on the subjective view in the same way, but the objective view yields a different functional form for the 'graph entropy'. Although this paper does not attempt to resolve this mismatch, the discussion and explanation of these different viewpoints in the three related areas of statistical physics, information theory, and graph theory will hopefully make the derivation and interpretation of the ERGM more understandable for students in network science.

These problems are exacerbated by the great and growing importance of SNA, over the past few decades, for studying online communities, because the majority of researchers and practitioners in this relatively new field have little mathematical and physics training. Consequently, they are likely to use the ERGM as a "black box" in the analysis of networks without understanding completely the assumptions upon which it rests. While for the more applied uses for which this model is employed this is not necessarily problematic, in my opinion it is difficult for social science researchers studying social networks to be able to critique the model and what it tells them without a better understanding of the fundamentals. Understanding the epistemological

foundations of the tools used in quantitative methodologies such as SNA is particularly important for social scientists since more often than not the quantitative methodologies are employed to answer research questions informed by different, qualitative epistemologies. A better understanding of the fundamental concepts upon which the quantitative models rest, therefore, may help integrate them with qualitative arguments. This latter point, which falls within the scope of critiques of interdisciplinarity, provides the second motivation for this paper.

2 Strategy

A discussion of the ERGM is particularly apt as a bridging topic between different disciplinary viewpoints. In fact, although entropy was first introduced as a fundamental concept in physics in the second half of the 19th Century, as eloquently discussed by Jaynes the information theory interpretation of entropy is 'more fundamental' than its physics counterpart (Jaynes [2], cited by Penfield [7]). Thus, at the basis of the ERGM we have an interplay between the concept of network, which already straddles the exact sciences, mathematics, and the social sciences, and the concept of information, which likewise is as 'fundamental' in the physical/natural sciences as in the human and social sciences.

This paper, therefore, follows a two-pronged strategy: on the one hand, it pays a great deal of attention to recounting the basic concepts of physical entropy and information entropy in an intuitively accessible but mathematically rigorous way; on the other hand, it relies on the power of analogy to compare and contrast physical and information entropy and, subsequently, to build up an intuitive understanding of 'graph entropy'. This last concept then forms the intuitive basis for the ERGM. Finally, where relevant the consequences of the analogies are followed to their logical conclusions, leading to an alternative formulation of Gibbs's entropy specific to networks. Throughout, no apologies are made for showing most of the mathematical details that are normally skipped – in particular taking care to keep them understandable and accessible to a non-mathematical audience.

In this linked progression of ideas from physics to graphs through information theory there is one step that is more difficult conceptually than it is mathematically. Whereas in information theory the mathematical expression for average information content is identical to the expression for maximum information uncertainty, hence making the connection with physical entropy easier to understand and believe, in the case of graphs the two expressions are different. The more important principle of maximum uncertainty still applies, thereby making the application of Gibbs's entropy's maximization principle to graph ensembles viable and effective in the derivation of the ERGM. However, the generalization of information content to graphs leads to a different equation and a different concept. Without necessarily invalidating the ERGM, this fact plants a seed of potential confusion that warrants further scrutiny and that should be exposed and clarified. Further, the mathematical and computer science consequences of this divergence should be explored to their conclusion. This paper begins to pose these questions.

Briefly, the derivation of the ERGM requires a conceptual leap, or a leap in abstraction level, between the basic physics and information theory concepts and Newman's Eq. (15.22): we need to develop an intuitive understanding of the analogue, for an ensemble of graphs, to the entropy of a system or to the average word length emitted by a communication source. Here are a few possible candidates:

- the information entropy of a graph ensemble (this is the most solid part of the ERGM derivation)

- the average graph "length" (clearly not a very good choice)
- the average graph size (this is the new interpretation that diverges from Gibbs's entropy for graphs)

At a more mundane level, whereas Shannon's entropy is based on a frequency distribution of different possible word lengths, the ERGM is based on a frequency distribution of different possible graphs of the same size, i.e. number of nodes. This is not necessarily a shortcoming, but we need to be extra-careful about how we interpret the results of a mathematical theory that has deviated from its intuitively accessible physics and information theory foundations.

The ERGM is derived by making use of a certain distribution of probability for graphs/networks of a given number of vertices/nodes. The fact that nothing is said about where the graphs are coming from or who/what is generating them should not stop us from imagining that given a certain number of nodes many networks can be constructed by linking the nodes in different ways. It happens that many of the graphs that can be constructed in this way also have the same number of edges or links. If, for a given fixed number of nodes, we look at the frequency distribution of graphs as a function of number of links and divide each such frequency value by the total number of graphs possible, we get a normalized probability distribution. This is the probability distribution that we are talking about here, that the ERGM can help us find, and that is central to using the ERGM as an analysis tool for networks obtained from empirical data. The usefulness of the ERGM is that it allows for the specification of different network characteristics through the quantification of their influence on this probability distribution, as we will discuss in detail below. For this reason we need to understand how to work with different kinds of probability distributions for different kinds of systems. Our starting point is the distribution of probability of the states of an isolated system.

3 Physics and Information Theory Background

3.1 Physical Entropy

3.1.1 Microcanonical Distribution

The fundamental postulate of statistical mechanics is that a system in equilibrium is equally likely to be in any one of its accessible states, at a given energy [8, 4]. If we take a system of N particles and use the function $g(N)$ to denote the number of its accessible states, the mathematical statement of the fundamental assumption is simply that the probability of any state i among the accessible ones is the same constant value given by

$$p_i = \frac{1}{g(N)}. \tag{2}$$

This trivially simple distribution is called the *microcanonical distribution.* The physical entropy of a system is ultimately a similarly trivial concept: it is the logarithm of the number of states that are accessible to that system at a given energy and with equal probability. We use the symbol S for entropy:

$$S = k_B \log(g), \tag{3}$$

where for the sake of completeness we have included Boltzmann's constant k_B whose numerical value and physical units do not matter in this discussion and can be taken as 1. This formula

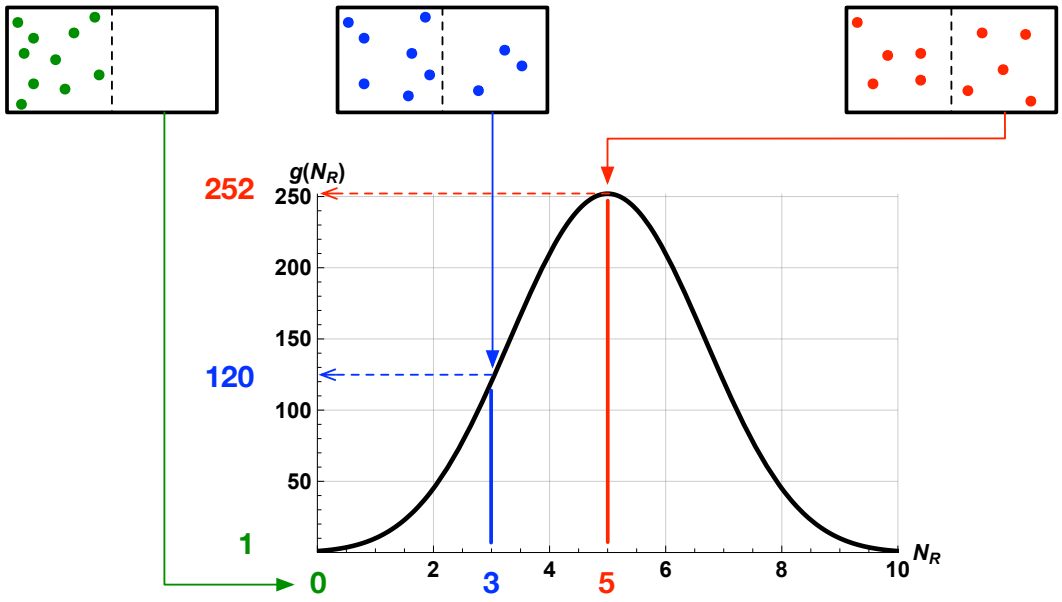

Figure 1: **Multiplicity of different configurations of 10 particles in a box**

is referred to as the Boltzmann entropy. g is called the multiplicity function and equals the number of accessible states (with equal probability) and by log we mean the natural logarithm or logarithm to the base e. We can use any other base, like 2 or 10, the concept does not change.

To get a sense for what this means, we take an insulated box with a gas of $N = 10$ identical free particles in the absence of gravity. We assume the particles have an initial random velocity, and therefore an initial energy. We assume there is no interaction between the particles, thus the potential energy of the system is zero. We assume further that the particles bounce off the walls with no energy loss. Therefore, the system only has kinetic energy, and since it is insulated its total kinetic energy is constant. We define a system configuration N_R as the number of particles that are in the *right* half of the box. Finally, we assume that we can tell the particles apart. We want to calculate the entropy of the system for different values of N_R, so we are looking for a function

$$S(N_R) = \log_2(g) = \log_2\left[g(N_R)\right] \tag{4}$$

where we have set $k_B = 1$ for convenience and we are using log base 2. Now choose the initial configuration to be $N_R = 0$, i.e. the 10 particles are all in the *left* half of the box. The particles will bounce around and fairly soon will visit both sides of the box. To get a sense of how likely it is for all the particles to find themselves again on the left-hand side at some future time, let's count the number of configurations for a few cases, as shown in Figure 1 (which shows a plot of the function $g(N_R)$):

- There is clearly only 1 way for the 10 particles to be on the left, so g(0) = 1. Notice that this follows from the fact that we are not subdividing our position resolution grid beyond the cells shown, left and right, thereby obtaining a so-called binary system:

  $$S(0) = \log_2[g(0)] = \log_2[1] = 0. \tag{5}$$

  Note that zero entropy means zero uncertainty: since all the particles are on the left, the system can only be in one state.
- Since we can tell the particles apart, if we allow only 1 particle on the right there will be 10 ways to do that, so g(1) = 10:

  $$S(1) = \log_2[g(1)] = \log_2[10] = 3.32. \tag{6}$$

- If $N_R = 2$, there are 45 ways for 2 of the 10 particles to be on the right:

$$S(2) = \log_2[g(2)] = \log_2[45] = 5.49. \tag{7}$$

More generally, the multiplicity function is given by the binomial coefficients:

$$\begin{aligned} g(N_R) &= \frac{N(N-1)(N-2)\cdots 1}{[(N-N_R)(N-N_R-1)(N-N_R-2)\cdots 1][N_R(N_R-1)(N_R-2)\cdots 1]} \\ &= \frac{N!}{(N-N_R)!N_R!}, \end{aligned} \tag{8}$$

where the exclamation mark indicates the factorial function. The rest of the cases can thus easily be calculated as follows:

$$\begin{aligned} S(3) &= \log_2[g(3)] = \log_2[120] = 6.91 \\ &\vdots \\ S(5) &= \log_2[g(5)] = \log_2[252] = 7.98 \\ &\vdots \\ S(7) &= \log_2[g(7)] = \log_2[120] = 6.91 \\ &\vdots \\ S(10) &= \log_2[g(10)] = \log_2[1] = 0 \end{aligned} \tag{9}$$

The point is that $N_R = 5$ will be by far the most frequent or probable configuration. In other words, equilibrium is just the most probable configuration, or the configuration with the highest entropy. By the same token, the configuration with the highest entropy also has the highest uncertainty, since we know the least about which of the 252 possible states the system might be in at any one time – for the simple reason that no other configuration has as many states. Finally, it is important to remember that whereas different configurations have different probabilities the different states corresponding to any one configuration are all equally probable.

The association of entropy with disorder is for the same reason: there are many fewer ways to keep a room in order (N_R close to zero) than to mix its contents into a chaotic jumble (N_R close to $N/2$). Thus, the most disordered 'configuration' is the most probable. To drive the point home, assume that we have 50 particles instead of 10. Now the case $N_R = 0$ still gives an entropy of $S(0) = 0$, but for $N_R = 25$ we get

$$S(25) = \log_2[g(25)] = \log_2[126,410,606,437,752] \approx 47 \tag{10}$$

So, clearly the number of states corresponding to the equilibrium configuration grows very quickly with system size. We also see why the log function was adopted on practical grounds in the definition of entropy, since it keeps its numerical value within relatively manageable bounds even for large systems.[1] The function $g(N_R)$ shown in Figure 1 quickly becomes much more spiked. As shown in Figure 2, this is most easily shown by plotting a normalized version of multiple curves, each for a different value of N, on the axes N_R/N and g/g_{max}.

[1] There is a deeper, better reason: entropy is a so-called 'extensive' property of systems, such that the entropy of two systems that are joined together is the *sum* of their individual entropies (contrast this with temperature, which is an intensive property). But the number of accessible states of the same two systems when joined together is their *product*. Thus, the log function naturally accounts for the increase in entropy when two or more systems are joined.

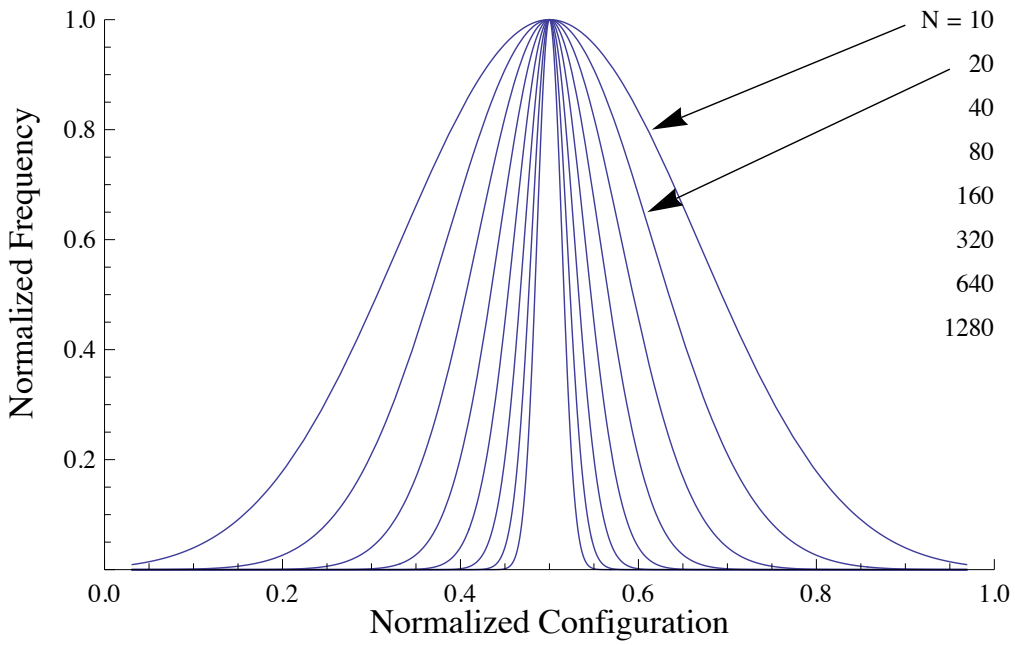

Figure 2: **Normalized frequency distributions for different size systems**

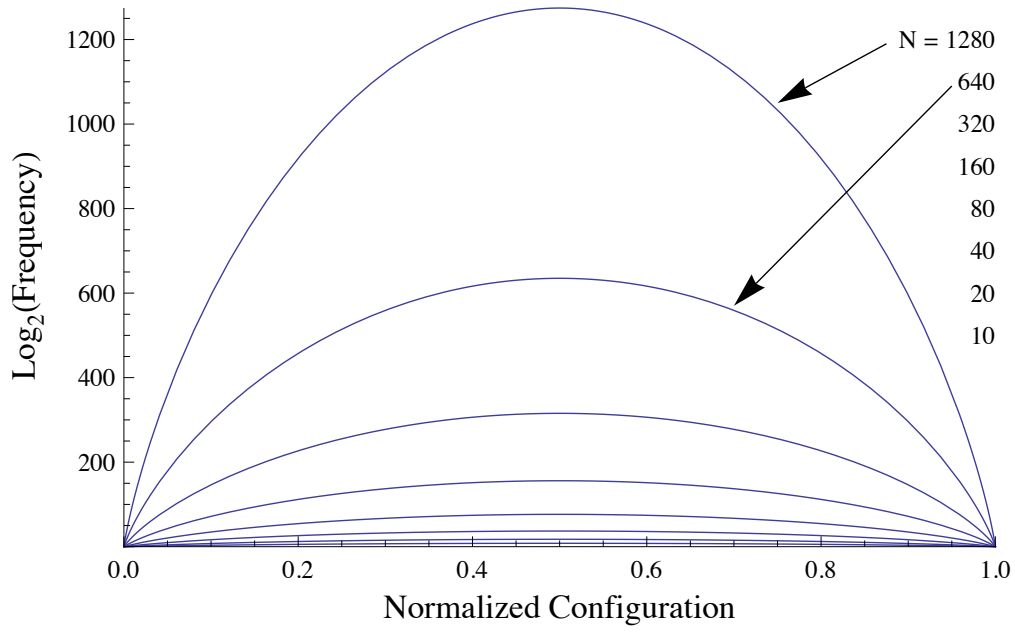

Figure 3: **Logarithm to base 2 of the frequency distributions for different size systems**

Figure 3 shows the logarithm of the same curves shown in Figure 2, i.e. the entropy. The number of states corresponding to the most probable configuration of the system with $N = 1280$ is rather large:

$$2^{1275} = 6.5 \times 10^{383}. \tag{11}$$

As a term of comparison, [7] provides the estimate that the number of atoms of the Earth is approximately 2^{170} ($= 10^{51}$) and in the visible universe 2^{265} ($= 10^{80}$). So for a system of a mere 1280 particles the number of configurations at maximum entropy is already unimaginable. If $N =$ Avogadro's number (6.02×10^{23}), i.e. the number of particles in a litre of air, the number of configurations possible for half of the particles to be in each half of the box is much, much larger than that. So it becomes easier to see why we never observe significant deviations from equilibrium in the distribution of air in a room, for example, under normal conditions.

The function we have been discussing is the Gaussian or normal distribution, and this is how it arises. This informal discussion should make it easier to see why 'the law of large numbers', otherwise known as the Central Limit Theorem [10], is so powerful. Note that, however, the system needs to be isolated for it to have a chance to reach equilibrim and for the argument to work. When these conditions do not apply the Gaussian still works, but progressively less the farther away the system is kept from equilibrium. This is why statistical methods tend not to work very well for small systems or for systems that are not near equilibrium or not isolated.

3.1.2 More General Formulation: Gibbs's Entropy

In the case of the isolated box of 10 free particles discussed above it is clear that the highest-entropy, equilibrium configuration is for $N_R = N/2$. However, by the fundamental postulate all states have equal probability, so the system will spend time also in the other configurations. The entropy of the system should therefore be an average weighted by the different probabilities.

Gibbs introduced the concept of an 'ensemble' of systems as an alternative to observing a single system for a very long time when calculating probabilities. The idea is to imagine that, if the ensemble is composed of a very large number ν of systems (much larger than the total number of states), ν_1 systems are in state 1, ν_2 are in state 2, ν_3 are in state 3, and so forth, with

$$\sum_i \nu_i = \nu. \tag{12}$$

This leads to a somewhat abstract derivation because the number of states can itself be very large. For example, even in our very small system of 10 particles the total number of possible states is $2^{10} = 1024$. So to make things more manageable and intuitively accessible let's instead apply the derivation to the configurations. We have 10 configurations whose probabilities are given by

$$p_i = \frac{g(N_R)}{2^N}, \tag{13}$$

where the subscript i refers to different values of N_R. Table 1 shows the values for the example system discussed.

To keep things as simple as possible, let's assume now that the size of our ensemble of identical systems is exactly 1024. If we follow this line of reasoning, rather than using Eq. (8) for the entropy of a single configuration we can calculate the entropy based on the total number of possible ways in which the probability distribution shown in Table 1 can be realized by the 1024

	N_R	$g(N_R)$	$S(N_R)$	p_i
Configuration	0	1	0	0.000977
	1	10	3.32	0.00977
	2	45	5.49	0.0439
	3	120	6.91	0.117
	4	210	7.71	0.205
	5	252	7.98	0.246
	6	210	7.71	0.205
	7	120	6.91	0.117
	8	45	5.49	0.0439
	9	10	3.32	0.00977
	10	1	0	0.000977
Total	-	1024	-	1

Table 1: **Probability distribution of the configurations of the 10-particle system**

systems in the ensemble, each involving the same 10 configurations. This is just a generalization of the formula for the binomial coefficients:

$$\begin{aligned} g_\nu(N) &= \frac{\nu!}{\nu_1!\nu_2!\nu_3!\cdots\nu_{10}!} \\ &= \frac{2^N!}{g(N_1)!g(N_2)!g(N_3)!\cdots g(N_{10})!} \\ &= \frac{1024!}{1!\ 10!\ 45!\ 120!\ 210!\ 252!\ 210!\ 120!\ 45!\ 10!\ 1!} \end{aligned} \tag{14}$$

The subscript ν on the function g indicates that this is the multiplicity function for the whole ensemble. This number is too large to be written out in full (2.806×10^{823}), but in any case since we want the entropy of a single system we need to take the log and divide by the size of the ensemble:

$$S_N = \frac{\log_2[g_\nu(N)]}{2^N} = \frac{\log_2[2.806 \times 10^{823}]}{2^N} = \frac{2735.4}{1024} \approx 2.67. \tag{15}$$

This is the entropy of our system of 10 particles taking into account the Gaussian probability distribution of its 10 configurations. For larger systems the numbers in Eq. (14) become prohibitively large and an approximation is needed. Luckily a very good one is available, known as Stirling's approximation:

$$\log(N!) = N\log(N) - N, \tag{16}$$

which works for any base of the logarithm. Thus, using Eq. (14) and the elementary rules of logarithms we can approximate the entropy of the ensemble as

$$\begin{aligned} g_\nu(N) &= \frac{\nu!}{\nu_1!\nu_2!\nu_3!\cdots\nu_{10}!} \\ &= \log_2(\nu!) - \log_2(\nu_1!) - \log_2(\nu_2!) - \log_2(\nu_3!) - \cdots - \log_2(\nu_{10}!) \\ &= \nu\log_2(\nu) - \nu - \nu_1\log_2(\nu_1) + \nu_1 - \nu_2\log_2(\nu_2) + \nu_2 - \nu_3\log_2(\nu_3) + \nu_3 - \cdots - \\ &\qquad \nu_{10}\log_2(\nu_{10}) + \nu_{10} \\ &= \nu\log_2(\nu) - \nu_1\log_2(\nu_1) - \nu_2\log_2(\nu_2) - \nu_3\log_2(\nu_3) - \cdots - \nu_{10}\log_2(\nu_{10}) \\ &= \nu\log_2(\nu) - \nu p_1\log_2(\nu p_1) - \nu p_2\log_2(\nu p_2) - \nu p_3\log_2(\nu p_3) - \cdots - \nu p_{10}\log_2(\nu p_{10}) \\ &= \nu\Big[\log_2(\nu) - p_1\log_2(\nu) - p_2\log_2(\nu) - p_3\log_2(\nu) - \cdots - p_{10}\log_2(\nu) - \\ &\qquad p_1\log_2(p_1) - p_2\log_2(p_2) - p_3\log_2(p_3) - \cdots - p_{10}\log_2(p_{10})\Big] \end{aligned}$$

$$= \nu \Big[\log_2(\nu)(1 - p_1 - p_2 - p_3 - \cdots - p_{10}) - \\ p_1 \log_2(p_1) - p_2 \log_2(p_2) - p_3 \log_2(p_3) - \cdots - p_{10} \log_2(p_{10}) \Big]$$

$$= -\nu \sum_{i=1}^{10} p_i \log_2(p_i) = 2761.38, \tag{17}$$

where we have used the normalization condition for the probabilities in the second-to-the-last equation and the numbers from Table 1. Finally, the entropy of our system of N particles is

$$S_N = -\sum_{i=1}^{10} p_i \log_2(p_i) = \frac{2761.38}{1024} = 2.697, \tag{18}$$

which is pretty close to the exact value given by Eq. (15). For large N the approximation becomes increasingly close to the exact value. This is then where Gibbs's formula for the entropy comes from.

3.2 Information Entropy

3.2.1 Information Content

Now we switch gears to information theory, where a similar set of concepts applies, although we find a first "stretching" of the original physics concept of entropy.

Shannon [9] talked about 'symbols' transmitted over a communication channel. In digital communications such symbols can be encoded in a given 'alphabet' such as the binary $\{0, 1\}$, forming a finite string of 0s and 1s for each symbol. Such strings are called 'words' and the individual binary digits forming the words are called 'bits'. The 'information content' of a communication stream is the minimum length of the word that can encode all the symbols being transmitted. Confusingly, the same quantity also represents the maximum uncertainty for the set of symbols, and in that case it is called the 'information entropy'. The former is easier to understand than the latter.

To see how the notion of information content arises, let's assume we have an idealized situation where a communication source emits a finite stream of M different symbols at a constant rate. Let's call the total number of symbols emitted N, and let's assume a situation where $N = M$. If we are using a binary alphabet, then to encode the M symbols in this alphabet we need at least H bits, where H is found from

$$2^H = M, \tag{19}$$

or, solving for H,

$$H = \log_2(M). \tag{20}$$

Normally, however, the total number of symbols being transmitted (N) is much larger than the number of different symbols that make up the communication (M). Therefore, it is more convenient to talk about the probability of a given symbol being transmitted. To introduce this notion, let's start from the case where $N = M$. In this case, while the communication is taking place, the probability of each of the M words being emitted is uniform and it is given by

$$p_i = \frac{1}{M}, \qquad i = 1, \cdots, M. \tag{21}$$

We can now rewrite Eq. (20) in order to obtain the information content in terms of probability:

$$H = \log_2(M) = \log_2\left(\frac{1}{p_i}\right) = \log_2(p_i)^{-1} = -\log_2(p_i). \tag{22}$$

So, the reason S is a positive function and H has a negative sign is that S is expressed in terms of the number of accessible states, whereas H is expressed in terms of the probability of emitted symbols, and the logarithm of a number between 0 and 1 is negative. Multiplying this negative number by the minus sign then gives back a positive H, as expected. Note that the subscript i would normally mean that the probability varies for each symbol. In this case it is constant and equal for all the symbols, but we still use the subscript because we are going to need it below for the more general case. Thus, if the M symbols are encoded in a binary alphabet $\{0, 1\}$, then H is the *minimum* word size, in bits, needed to encode a set of M symbols with the binary alphabet $\{0, 1\}$. For example, if we had 512 different symbols, we would need a string of at least 9 bits to encode all of them:

$$2^9 = 512, \qquad \text{so} \qquad H = \log_2(512) = 9. \tag{23}$$

Of course, if the number of symbols is not a power of 2 then the value of H will not be an integer. This forces an increase in abstraction level in its interpretation relative to number of bits without changing the essence of the concept.

The fact that H as defined here is the minimum-length word that can encode all the symbols is significant. We could in fact have chosen to encode our M symbols with a longer string of length, say, $L > H$. By so doing we would have had left-over strings of length L after all the M symbols had been encoded. Therefore, we would have had the potential for explicitly introducing some irrelevant information or noise in our communication stream. This is in fact the case in applications where one uses a number of bits that encodes a number of symbols that is greater than or equal to M, rather than strictly equal to it. For example, if the number of symbols falls between two powers of 2 then one would use the larger power for the word length. Notwithstanding this practical limitation, the fact that H is the most efficient way to encode the given symbols justifies its name '*information content*'.

Let's now look at the second, more famous interpretation of H as uncertainty or information entropy. This interpretation arises from the fact that when the probability is uniform we know the least about which symbol will be emitted next by the source. Thus, in this case we expect the uncertainty to be maximum. On the other hand, if we know that some symbols are more likely than others, then there is less uncertainty in the composition of the symbol stream and we expect the uncertainty to decrease. As shown in Figure 4, in a case where we have 4 symbols at least 2 bits are needed to encode them (information content = 2). The fact that this number is also the maximum uncertainty becomes clear if we happen to know that Symbols 1 and 3 have zero probability of being emitted. As a result, there is less uncertainty since we know that out of the four symbols only 2 and 4 will be emitted. This matches the fact that now the least number of bits needed to encode two symbols is 1 (since it can take on two values, either 0 or 1). In the last case shown in the figure, where we know that only Symbol 3 will be emitted, there is no uncertainty and we don't need any bits to encode it. In all cases of course the sum of the probabilities must equal 1. In the third case we know more than in the second, since we know that Symbol 4 is more probable than Symbol 2, but not as much as in the fourth. This is reflected in the value of the uncertainty between 1 and 0. The fact that H also reflects our knowledge about the outcome justifies its name 'information uncertainty'. However, the fact that it has the same functional form as physical entropy, as we show below, justifies the name '*information entropy*'.

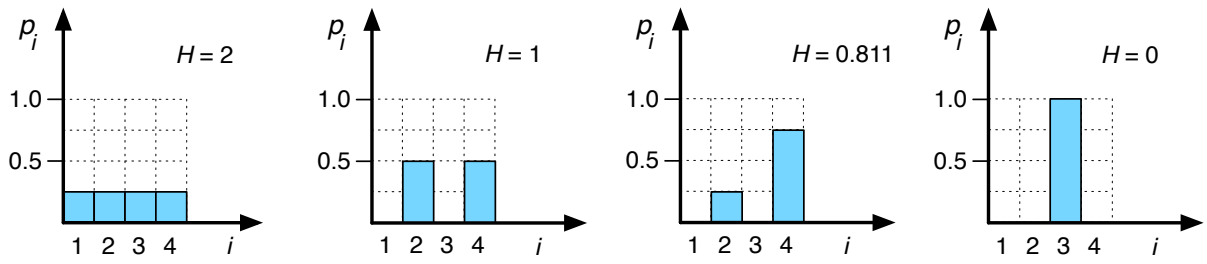

Figure 4: **(Left to right:) Decrease in uncertainty (information entropy) with increasing knowledge of most probable symbols**

3.2.2 Which Concept is More Fundamental: Information Entropy or Physical Entropy?

For a source that outputs an infinite sequence of bits to communicate a finite set of symbols M, Shannon generalized Eq. (22) to express an average word length. This brings information entropy a step closer to the original concept of physical entropy, since the possibility of multiple values of H is allowed, although the average is not quite the same thing as the most probable, especially to the extent of the huge spike of the Gaussian for large systems. On the other hand, if we look at the average word length "from below" then what we see is maximum information entropy in the sense just explained. Although in this more general case maximum uncertainty does not quite equate to most probable configuration, it turns out that the mathematical expression is uncannily similar – in fact, identical within a constant factor. This fact led Jaynes to say that information entropy is more fundamental than physical entropy:

> The mere fact that the same mathematical expression $-p_i \sum \log p_i$ occurs both in statistical mechanics and in information theory does not in itself establish any connection between these fields. This can be done only by finding new viewpoints from which thermodynamic entropy and information-theory entropy appear as the same *concept*. In this paper we suggest a reinterpreation of statistical mechanics which accomplishes this, so that information theory can be applied to the problem of justification of statistical mechanics. $\cdots$ Just as in applied statistics the crux of a problem is often the devising of some method of sampling that avoids bias, our problem is to find a probability assignment which avoids bias, while agreeing with whatever information is given. The great advance provided by information theory lies in the discovery that there is a unique, unambiguous criterion for the 'amount of uncertainty' represented by a discrete probability distribution, which agrees with our intuitive notions that a broad distribution represents more uncertainty than does a sharply peaked one, and satisfies all other conditions which make it reasonable. ([2] emphasis in original)

In other words, rather than trying to justify information entropy in terms of most probable outcomes, Jaynes suggests that we invert the logical dependence: we should justify most probable outcomes in terms of least bias and maximum uncertainty. It is in this sense that information theory is 'more fundamental' than statistical physics. Indeed, this formulation is consistent with all the examples discussed here from both disciplines. Be that as it may, the identical mathematical treatment between Gibbs's entropy, Shannon's information entropy, and the ERGM enables us to focus on the latter two while avoiding a detailed derivation of Gibbs's canonical distribution, which may be difficult to follow for non-physical scientists.

3.2.3 Average Information Content or Average Information Entropy

The derivation of the average word length or of maximum information entropy is easier to see for a large but finite total number of symbols N, where most/all symbols occur more than once. So this is the case where $N \gg M$. N_i is the number of occurrences of the symbol M_i, and therefore acts as a weight in the average. We first note that

$$\sum_{i=1}^{M} N_i = N. \tag{24}$$

Now we can say that each symbol will make a contribution to the required word length that is proportional to its frequency in the symbol stream, such that the average length H can be calculated from:

$$\begin{aligned} H &= \frac{\sum_{i=1}^{M} N_i[-\log_2(p_i)]}{\sum_{i=1}^{M} N_i} = \frac{\sum_{i=1}^{M} N_i[-\log_2(p_i)]}{N} \\ &= -\sum_{i=1}^{M} \frac{N_i}{N}[\log_2(p_i)] = -\sum_{i=1}^{M} p_i[\log_2(p_i)]. \end{aligned} \tag{25}$$

The simplicity of this derivation relative to the one for Gibbs's entropy is striking.

4 Taking Stock: An Alternative Formulation of the ERGM

4.1 A Disconnect

Eq. (25) is the same as Eq. (15.22) in Newman [6] up to a constant factor (since Newman uses the natural logarithm). We have seen that, incredibly, the same function was derived by completely different conceptual arguments and mathematical methods for physical entropy and for information entropy. Let's now see how entropy arises in the context of networks.

By a network we mean a set of edges or links that connect a set of nodes to one another in some way. Even if we limit ourselves to connected networks, i.e. without isolated nodes or islands, the number of ways in which a given set of nodes can be connected quickly becomes astronomical even for relatively small node sets. Allowing for isolated nodes and islands means that we are again dealing with the set $\{0, 1\}$ to indicate the absence or presence of a link between any two nodes, respectively. Therefore, if we assume that we are dealing with undirected links the number of possible networks for k nodes is

$$M = 2^{\frac{k(k-1)}{2}}. \tag{26}$$

For a network of 30 nodes, for example, this equation tells us that there are a possible 8.87×10^{130} different networks. In order to be able to talk about a probability distribution over this set of networks, the ensemble needs to be much bigger. This is a bit daunting but mathematically it is perfectly acceptable. As before, the sum of all the probabilities must add up to 1:

$$\sum_{i=1}^{M} p_i = 1. \tag{27}$$

As we saw before with physical and information systems, one possibility is for this distribution to be constant. This, however, is not very useful since we wish to develop a model that can

be representative of the empirical network we wish to analyse. Thus, we can impose additional constraints that are derived specifically from the empirical network to be modelled, such as various weighted average or 'expectation' values for network properties:

$$\langle x_j \rangle = \sum_{i=1}^{\nu} p_i x_{ij}, \tag{28}$$

where the angle brackets imply an averaging process or function, ν ($\gg M$) is the number of networks in the ensemble, j refers to a particular property – such as the number of triangles – and x_{ij} refers to that property for a particular network i belonging to the ensemble. If the left-hand side (LHS) is treated as a known network measure of the empirical network, this equation can be used as a constraint that can help us derive the corresponding probability distribution. However, even if we assume that we have a very rich data set with many such network measures known, we would need to specify M of them in order to solve for all the possible values of p_i. This is clearly unrealistic even for the smallest sets of nodes. For example, for a network of 4 nodes, $M = 64$; it is difficult to imagine that we can come up with 64 different network measures such as degree, number of diads, etc with an empirical network of only 4 nodes!

Here is where Newman introduces the maximization of Gibbs's entropy as the criterion that makes up for all the missing constraints and enables us to derive a valid and useful probability distribution. Newman's Eq. (15.22) is expressed in terms of the natural logarithm (which in this paper we simply write as 'log') rather than the base-2 logarithm:

$$S = \sum_{i=1}^{M} p_i \ln(p_i) \qquad \text{Eq. (15.22) in [6].} \tag{29}$$

This is where we see the power of information theory, since clearly such a probability distribution has nothing at all to do with the likelihood of a given network relative to the others but, rather, with the least bias or greatest uncertainty of the distribution. Put otherwise, although there is nothing that would lead us to expect a sharply peaked probability distribution such as the Gaussian for the physical example we discussed earlier, it is still entirely plausible and legitimate to seek the probability distribution that introduces the least bias beyond the explicit constraints enforced from the empirical data. This is precisely the basis of the information theory interpretation of entropy, which we have already seen at work for Shannon's communication channel. The only proviso is that the maximum of the uncertainty distribution may not be very sharp at all: therefore, whereas on the one hand it may be more difficult to find it, on the other hand any probability distribution in the vicinity of the maximum of Eq. (29) will do (this point is reflected in how convergence is treated in current implementations of ERGM methods).

Although it would therefore seem that the maximization of Gibbs's entropy is all we need to worry about, we now realize that since we are maximizing uncertainty rather than likelihood we should really be using an equation that is derived from information theory rather than physics. Since we saw that the information theory derivation led to the same equation (Eq. (25)) as the equation from the physics derivation (Eq. (18)), we might expect the same to happen here. In fact, it doesn't. Thus, the rest of this paper is concerned with deriving an expression for the entropy of an ensemble of networks and with posing some questions about the possible consequences of such a different formulation.

4.2 Graph Entropy

Newman's own derivation implicitly refers to the information theory derivation even if it continues to invoke concepts from physics, as we now show. Eq. (29) is the starting point of the

calculus part of the derivation of the ERGM. Before we delve into the details of the mathematics it is worth examining the concepts and assumptions the derivation is based on. We do this by comparing the network concepts to their analogues in information theory and physics.

First, Newman introduces the concept of an ensemble of networks, all of which have the same number of nodes, n. The average of an ensemble is a very useful concept in statistical physics because in a frequency-based framework it is equivalent to long-time averages of a single system. Although the ERGM was not developed specifically to analyse time-dependent network topologies, the concept is still useful. Even though we did not mention ensembles when talking about Shannon's entropy, we invoked the concept implicitly when deriving the average word length. In information theory the analogue to an ensemble of networks is provided by the fact that the communication source emits many copies of every symbol. We saw that the frequency distribution of the various symbols is not in general uniform, which justified the derivation of Eq. (25).

Second, Newman writes Eq. (29), explicitly saying that each network has a certain probability of occurrence in the ensemble. Thus, S in Eq. (29) is necessarily the 'average entropy of the network ensemble'. This is a rather odd notion, to be sure, but one that we can cope with by analogy. We have already demystified Eq. (25) by making recourse to a particular finite ensemble of N symbols in which each is different from the others (so that $N = M$). We can do the same here as a first step: assume we have an ensemble of M networks, each different from the others. Then, the "word length" needed to "encode" this family of networks is precisely the same as the H given by Eq. (22). However, now H becomes a graph metric that is analogous to word length. Clearly, "graph length" is meaningless, so we need to come up with a better generalization of this metric.

Here we notice that the fact that we are using the logarithm function in information theory is incidental, the more important and intuitively accessible concept is that of word length. So we could imagine as something analogous a quantity that represents the most efficient way to encode the M networks. I would argue that this is not going to be the logarithm of M but, rather, something else. A plausible candidate is the inverse of the function that yields the M networks from the smallest possible number of nodes, k. Let's find this function by retracing our steps.

In information theory, given M symbols we want the smallest-size binary string that can encode them. So we are solving Eq. (19) for H, obtaining Eq. (20). In graph theory, therefore, we want something analogous to length. But networks are not one-dimensional. In fact they are not "dimensional" at all, necessarily, because they do not require a metric space for their definition; they are *topological* objects. Therefore, a suitably analogous notion to length could be number of nodes. Thus, for simple bidirectional graphs we want to solve Eq. (26) for k, obtaining

$$k = \frac{1}{2} \pm \frac{1}{2}\sqrt{1 + 8\log_2(M)} = \frac{1}{2} \pm \frac{1}{2}\sqrt{1 + 8\log_2\left(\frac{1}{p_i}\right)} = \frac{1}{2}\left[1 \pm \sqrt{1 - 8\log_2(p_i)}\right] \tag{30}$$

This seems like a better analogue to Eq. (22) for networks than Eq. (3). This view is strengthened by the interpretation we have already discussed of Shannon's entropy or information content as the shortest or most efficient encoding of M symbols. So Eq. (30) provides a possible candidate for 'network entropy' as the smallest and most efficient encoding, with networks of k nodes, of M networks with n nodes. This is consistent also with Newman's discussion of his Eq. (15.22) which, however, harks back to physics in a way that seems unjustified for the reasons stated:

> The Gibbs entropy is precisely a measure of the amount of "assumption" that goes into a particular choice of distribution $P(G)$, or more precisely it is the amount of "anti-assumption"

> or ignorance, and by maximizing it we minimize unjustified assumptions as much as possible. ([6]: 568)

I agree entirely with the sentiment expressed here, but the optimization Newman is talking about is from information theory, not from physics, even if Shannon used physical entropy as initial inspiration. Therefore, we are justified in carrying the development further to obtain the 'average entropy of a network ensemble':

$$K = \sum_{i=1}^{M} \left(\frac{N_i}{\sum_{i=1}^{M} N_j} \right) \frac{1}{2} \left[1 + \sqrt{1 - 8 \log_2(p_i)} \right] = \frac{1}{2} \sum_{i=1}^{M} \left(\frac{N_i}{N} \right) \frac{1}{2} \left[1 + \sqrt{1 - 8 \log_2(p_i)} \right]$$
$$= \frac{1}{2} \sum_{i=1}^{M} p_i \frac{1}{2} \left[1 + \sqrt{1 - 8 \log_2(p_i)} \right] \tag{31}$$

This quantity K seems more appropriate than the S of Eq. (29) to represent the average network entropy, where we have chosen to use the positive branch of the square-root. For simple directed graphs the same expressions apply but with a 4 wherever there is an 8. The advantage of this formulation is that now K has acquired a concrete meaning: it is the average number of nodes that can most efficiently encode the M networks. However, the disadvantage is that the M networks had a different, larger set of nodes. Further, the topology of the encoding probably has nothing at all to do with the topology of the original networks. So there is some more work to do, and it is not yet clear whether anything useful will come out of this exercise.

The analogous situation for Shannon's entropy would obtain if the symbols to be encoded were themselves binary words. For example, assume we have 256 words, all of them of the same length of 20 bits. What is the information content of this sample? Well, it is $\log_2(256) = 8$. So it takes an 8-bit word to encode all 256 original words, in spite of the fact that they are 20 bits long. Another way to think of that is that with an 8-bit word one can count in binary from 0 (= 0000 0000) to 255 (= 1111 1111). Clearly, the mapping between the 8-bit encoding and the original 256 20-bit words is arbitrary since it depends on the order in which the latter are sampled. An arguably desirable sampling would be to sort both sets of binary numbers in increasing order and then map the two sets of words 1-1. Finally, notice that the 256 20-bit words were taken from a possible set of $2^{20} = 1,048,576$ words or 'symbols'.

In summary, the need for this whole argument is a consequence of the fact that by invoking Shannon's entropy (even though Newman calls it the Gibbs entropy) the ERGM implicitly assumes that each network in the ensemble is analogous to a separate "symbol". If we then keep the analogy rigorous both conceptually and mathematically we can't avoid the encoding of an ensemble of N networks by a set of much smaller networks whose average size is given by Eq. (31).

5 Derivation of the ERGM

Newman starts with a set G of all simple undirected graphs of k nodes. So the size of G is given by Eq. (26) and is therefore pretty large, as shown in Table 2.

For comparison with information theory we have added also a column that shows the number of symbols that can be encoded with a binary alphabet in a string of k bits. A real network could have a few tens or a few hundred nodes, whereas a subset of the Internet could easily have millions. So G can indeed grow to astronomical proportions (and beyond!). But this is not all.

In order to make the ERGM derivation work, for a given choice of k we need an ensemble of N networks taken from G, and the ensemble needs to be, for example, thousands of times larger than G in order to make the probabilities based on frequentist statistics meaningful. Thus, each network G_i in G will appear with some frequency in our ensemble, given by N_i/N just as in the case of information theory. The first equation we will need is therefore the normalization condition Eq. (27) which we re-write in a slightly different form:

k	M = Size of G (number of graphs of k nodes)	2^k
1	1	2
2	2	4
3	8	8
4	64	16
5	1024	32
6	32,768	64
7	2,097,152	128
8	268,435,465	256
9	68,719,476,736	512
10	35,184,372,088,832	1024
...	...	...

Table 2: **Comparison of the size of the set of encodable symbols for different information content in graph theory and information theory**

$$\sum_{n=1}^{M} p(G_i) = 1. \tag{32}$$

Any network measure x_j, such as the number of triangles in a given network, for instance, will therefore have an average or mean value given by (re-writing Eq. (28))

$$\langle x_j \rangle = \sum_{i=1}^{M} p(G_i) x_j(G_i). \tag{33}$$

As explained by Newman, if we want to arrive at a probability distribution from which we can construct a network that is likely to have a set of desired characteristics, we can set the mean value of x_j to something we want and treat the above equation as a constraint on $p(G_i)$. In order to determine this function completely we would need to specify a value of p for each network G_i. As we can see from Table 2, for any realistic number of nodes the number of constraints that we would need to impose quickly becomes unreasonable. Here is where Newman invokes a maximization principle, namely the maximization of Gibbs's entropy given as Eq. (29). As we have seen this equation is not Gibbs's entropy, it is Shannon's entropy inspired by but not identical to Gibbs's entropy. However, as we discussed the principle of optimization of uncertainty is still valid.

Using the visualization of the method of Lagrange multipliers shown in Figure 5, if we want to maximize a function such as an inverted paraboloid straightforward application of calculus will give us the very top, i.e. the absolute maximum of this function. If instead we want to maximize it subject to the constraint provided by a given curve, then the maximum will necessarily need to be on this curve, and may not be the absolute maximum. In Figure 5 the constraint curve is a Sin function (imagine its vertical projection or "shadow" down to the x-y plane).

The method of Lagrange multipliers says that this constrained maximum will occur where the gradient (vector of steepest ascent) of the surface is perpendicular to the constraint curve. It

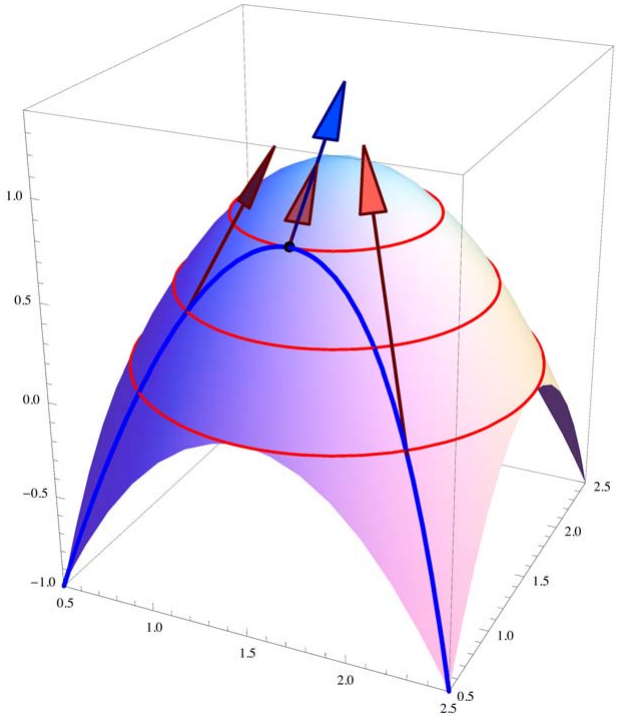

Figure 5: **Visualization of the maximization of a paraboloid subject to a constraint (blue Sin function)**

does not actually say this, but this formulation is easier to see in the figure. Note that the red gradient vectors are perpendicular to the red circles, which are visualizations of contour curves (or level sets, i.e. curves of constant elevation) of the paraboloid. So what the method says is that if we view the constraint curve as the contour curve of a *different* surface, not shown here, then the constrained maximum will occur where the two gradient vectors are parallel.[2] The condition for two vectors to be parallel constrains their direction to be the same, but not their magnitudes, which can differ by a constant factor. In other words:

$$Blue\ Vector = \alpha \times Red\ Vector, \tag{34}$$

where α is the Lagrange multiplier. In other words, the Lagrange multiplier arises naturally when enforcing a condition that can only occur at the maximum of a given function subject to the given constraint.

In our case we are dealing with a higher-dimensional problem with multiple constraints but the concept remains the same. In the standard derivation of the ERGM we want to maximize Eq. (29) subject to the constraints provided by Eqs. (32) and (33). In other words, we need to maximize the following quantity:

$$\sum_{i=1}^{M} p(G_i) \log p(G_i) - \alpha \left[1 - \sum_{i=1}^{M} p(G_i) \right] - \sum_{j=1}^{J} \beta_j \left[\langle x_j \rangle - \sum_{i=1}^{M} p(G_i) x_j(G_i) \right], \tag{35}$$

where J is the total number of network features that we want to fix. Differentiating with respect

[2]We should clarify that we have taken some artistic licence in creating Figure 5: none of the arrows shown is actually supposed to be tangent to the surface. Gradient vectors are defined as 2-D vectors lying in the x-y plane, but they look nicer when drawn on the surface. In any case, the fact that the two gradient vectors are parallel is true regardless of how we choose to draw them.

to the probability of a particular graph G_i and setting equal to zero gives

$$-\log p(G_i) - 1 + \alpha + \sum_{j=1}^{J} \beta_j x_j(G_i) = 0. \tag{36}$$

Therefore,

$$\begin{aligned} p(G_i) &= e^{\left[\alpha - 1 + \sum_{j=1}^{J} \beta_j x_j(G_i)\right]} = e^{\alpha - 1} e^{\left[\sum_{j=1}^{J} \beta_j x_j(G_i)\right]} \\ &= \frac{e^{\left[\sum_{j=1}^{J} \beta_j x_j(G_i)\right]}}{e^{1-\alpha}} = \frac{e^{H(G_i)}}{Z}, \end{aligned} \tag{37}$$

whence we see where the term 'exponential' in the name of the model comes from. The last step in the above involves definitions that conveniently parallel analogous quantities in the canonical distribution of statistical mechanics. In particular, the H shown here has nothing to do with Eq. (25) and in the context of the equation above is called the 'graph Hamiltonian'. We can make a few observations:

- This probability distribution deserves an in-depth analysis and discussion to understand both mathematically and physically what it is telling us. We would need to review the canonical distribution of statistical mechanics to achieve that. For the purposes of this paper, suffice it to say that the numerator can be visualized as a probability distribution that decreases as H increases. The Hamiltonian is the total mechanical energy of frictionless systems and generalizes to the (quantized) system energy in quantum mechanics, so it is probably invoked here by analogy since associating the linear combination of fixed network characteristics with Lagrange multipliers as coefficients with energy seems odd, to say the least. Perhaps as the number of constraints grows the network is assumed to be less probable, which would be analogous to being higher-energy.
- Z is called the partition function and it is just a normalizing factor, found explicitly with the help of Eq. (32):

  $$Z = \sum_{j=1}^{J} e^{H(G_j)}. \tag{38}$$

  Z is the function that makes the ERGM unusable for anything but the simplest graphs, given the huge value of J.
- We can see *a posteriori* the appeal of invoking the Gibbs entropy, because the resulting derivation is indeed beautiful, and mirrors the canonical distribution of statistical mechanics exactly.

6 Discussion

The inconsistency we have highlighted motivates us to dig further, and to look at what would happen if we started the derivation of the ERGM from Eq. (31) instead of Eq. (29). However, the algebra involved in applying the method of Lagrange multipliers to Eq. (31) quickly becomes unmanageable, making it impossible to separate – as nicely as Eq. (37) does – the different multipliers associated with the normalization condition and with the expectation values of the other network measures.

Further, it does not seem likely that different functional forms for graph entropy would make a big difference in the final probability distribution. This is for three reasons:

1. First, the huge dimensionality of the problem suggests that the probability distribution for Eq. (31) is likely to be smooth and with a maximum that is as shallow as that of Eq. (29), such that the broad ranges of networks near each maximum are likely to overlap. Admittedly, this is only a conjecture.

2. Second, even if this were not true the main purpose of the model is at least as much to distinguish between networks of different characteristics as it is to model a single network. It is not as important for the absolute quantification of the model to be accurate if it is effective at relative quantification.

3. Third, the effect of the Lagrange multipliers (constraints) on the probability distribution is so strong that the choice of different functions to represent the distribution may indeed be second-order in comparison.

We can demonstrate the third point with an example from statistical physics. The canonical distribution[3] can be derived following an identical formulation to what is presented here (except for a constant factor). In other words, Eq. (29) is maximized subject to a normalization constraint and a constraint on the average energy of the system. The Lagrange multiplier associated with the normalization constraint gives rise to the partition function, as we have just derived, whereas the Lagrange multiplier associated with the energy constraint is none other than the (inverse) temperature of the system! Thus we see that the rather abstract concept of Lagrange multiplier can actually have a very physical and almost mundane interpretation. It follows that the possible physical, topological (or otherwise) interpretations of the β factors in the graph Hamiltonian could potentially be very useful quantities and are worth further study – albeit not in this paper.

In order to get a feel for the impact of the β factors we now show an example from physics, i.e. we show the effect of different temperatures on the probability distribution of the states that a model system in contact with a reservoir can assume. For the sake of simplicity in this discussion the Boltzmann constant is assumed to equal 1 and it is not shown explicitly. The canonical probability distribution exactly analogous to Eq. (37) is given by

$$p_i = \frac{e^{-\beta E_i}}{Z} = \frac{e^{-\beta E_i}}{\sum_i e^{-\beta E_i}} = \frac{e^{-E_i/T}}{\sum_i e^{-E_i/T}} \tag{39}$$

Since Z is just a normalizing constant, this expression is simply saying that the probability of our model system to assume a state given by energy E_i decreases with increasing values of E_i. In other words, low-energy states are more probable than high-energy states. In order to produce a plot of this function, it is easier to treat the summation in the denominator as an integral over all possible energies:

$$Z \approx \int_0^\infty e^{-E/T} dE = \left[-Te^{-E/T}\right]_0^\infty = 0 - (-T) = T. \tag{40}$$

Thus, the function we are plotting in Figure 6 for different constant values of the temperature is just

$$p(E) = \frac{e^{-\beta E}}{Z} = \frac{e^{-\beta E}}{\sum_i e^{-\beta E_i}} \approx \frac{1}{T} e^{-E/T}. \tag{41}$$

[3]Gibbs ([1]: Preface) called this distribution canonical 'on account of its unique importance in the theory of statistical equilibrium'.

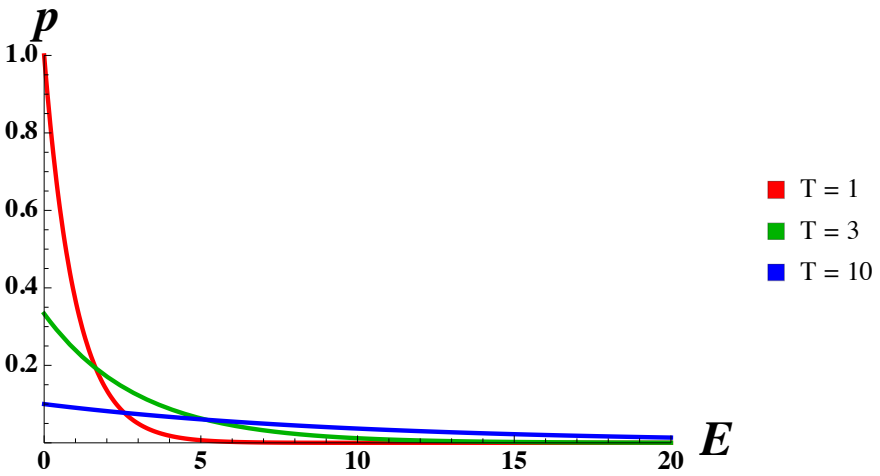

Figure 6: **Canonical probability distribution at different temperatures**

Due to the normalization constraint, the area under each of the three curves shown in Figure 6 is equal to 1. This plot shows that the probability of the model system assuming a high energy is vanishingly small at low temperatures and becomes noticeable at higher temperatures. The main point, however, is that this plot shows that even a modest increase in temperature causes a huge change in the probability distribution (although this effect decreases proportionately at higher temperatures), supporting the conjecture that using a slightly different function from Eq. (41) is likely to have a second-order effect in comparison.

7 Conclusion

The importance of the constraints in the maximization process of the average entropy of a network ensemble provides an intuitive justification for the claim of universality attributed to Gibbs's entropy. In other words, in spite of the derivation of a function for the average entropy of a network ensemble (Eq. (31)) that is different from the equation for Gibbs's entropy (Eq. (29)), it seems safe to use the latter as the basis of the ERGM.

At the same time, the search for an expression for the average entropy of a network ensemble that is consistent with information theory has uncovered the possibility of encoding graphs with a set of 'topological words' that can be regarded as a generalization of the familiar one-dimensional digital words expressed in bits. These topological words are nothing more than the adjacency matrices of a set of graphs that can encode a set of symbols (which can themselves be graphs) most efficiently. Whether or not these topological words can be regarded as 'primitives' of a formalism optimized for the mathematical analysis and formal manipulation of graphs is to be determined and is left as an open research question.

In conclusion, in this paper we have sought to demonstrate that a thorough analysis of the intuitive and mathematical foundations of concepts that straddle multiple disciplines, such as entropy, uncertainty, networks, and information, can lead to a stronger understanding of the fundamental ideas in each discipline and to new and interesting research questions purely through the power of analogy and the enforcement of self-consistency.

References

[1] J W Gibbs. *Elementary Principles in Statistical Mechanics Developed with Especial Reference to the Rational Foundation of Thermodynamics*. Edward Arnold, London, 1902.

[2] E T Jaynes. Information Theory and Statistical Mechanics. *The Physical Review*, 106(4):620–630, 1957.

[3] E T Jaynes. Gibbs vs Boltzmann Entropies. *American Journal of Physics*, 33:391–398, 1965.

[4] C Kittel and H Kroemer. *Thermal Physics, 2nd Ed.* W H Freeman, New York, 1980.

[5] D Lusher, J Koskinen, and G Robins, editors. *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge University Press, Cambridge, 2013.

[6] M Newman. *Networks, an Introduction*. Oxford University Press, Oxford, 2010.

[7] P Penfield. Information and entropy, 2008. MIT Open Courseware: (Accessed 16 March 2018) http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-050j-information-and-entropy-spring-2008/.

[8] F Reif. *Fundamental of Statistical and Thermal Physics*. McGraw-Hill, New York, 1965.

[9] C E Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:379–423, 1948.

[10] D Sornette. *Critical Phenomena in Natural Sciences: Chaos, Fractals, Self-organization and Disorder – Concepts and Tools*. Springer, Heidelberg, 2000.

[11] J Uffink. Can the maximum entropy principle be explained as a consistency requirement? *Studies in the History and Philosophy of Science Part B: Studies in the History and Philosophy of Modern Physics*, 26:223–262, 1995.