Network models of stochastic power-laws

by

George Jack Peterson

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

The text of this dissertation includes reprints of the material appearing in the following articles:

1. G.J. Peterson, S. Pressé, and K.A. Dill. Nonuniversal power law scaling in the probability distribution of scientific citations. *Proceedings of the National Academy of Sciences USA* 107 (37): 16023 – 16027 (2010).

2. G.J. Peterson, S. Pressé, K.S. Peterson, and K.A. Dill. The evolution of protein-protein interaction networks. Submitted to *PLoS Computational Biology*, October 2011.

K.A. Dill directed and supervised the research that forms the basis for the dissertation. S. Pressé helped develop the theoretical foundations for articles 1 and 2. K.S. Peterson performed statistical analyses for article 2.

Network models of stochastic power-laws

George Jack Peterson

Abstract

In this work, I investigate what power-law processes may have in common. I propose that their common feature is a specific type of positive feedback. Next, I discuss two specific power-law models for processes which appear to be quite different from one another: (1) citations of scientific papers, and (2) physical interactions between proteins in cells. Finally, I discuss a new theoretical framework for describing power-law phenomena, based on the principle of maximum entropy, combined with a symmetry relationship.

Contents

List of Figures

List of Tables

1 Introduction

Probability distributions with a heavy tail show up in a wide variety of real-world systems: the probability that a scientific paper will receive a certain number of citations [1], the distribution of the number of interaction partners of proteins [2], the sizes of fluctuations in the stock market [3], the intensity of solar flares [4], and many others. In these systems, a distribution $p(k)$ may have exponential behavior for small k and a power-law tail for large k. Power-law tails imply that very large events occur with an unusually high probability. This is of considerable practical interest: often, the most consequential events – the stock market crash, the breakthrough paper, the conserved hub protein – are the 'outliers' far out in the tail, rather than commonplace events closer to the mean. My research has focused on the question: *What mechanisms are shared by these disparate systems that generate similar large-scale statistical behavior?*

Over the past decade, there have been a number of efforts to explain the existence of power-law tails using models of growing networks [5, 6, 7, 8, 9, 10] (see [4] for review). Typically, these models start from an abstract mathematical framework (e.g., a rule for the formation of an edge between two nodes), then fit the tail of their model to the the tail of one or more real-world distributions. One shortcoming of this approach is that the physical explanation for the heavy tail is often unclear. I approached this problem from the opposite direction: I first built models for two specific power-law phenomena, then worked backwards to see what their underlying mechanisms have in common. My observations are as follows:

1. Power-law tails arise from positive feedback.

2. Exponential behavior observed near the origin can be explained by simple random models.

3. Competition between the random and positive feedback mechanisms causes the transition from exponential to power-law behavior.

1.1 Two ways to cite a paper

The first system I examined is the distribution of citations to scientific papers [11]. A natural representation for this system is a directed graph, where nodes represent papers, and an edge represents the citation of one paper (incoming) by another (outgoing). Citations are a convenient modeling testbed for two reasons: first, data is plentiful, and easy to download, and, second, the model can be mathematically very simple, because citations are only issued on the publication of a paper, and can not be reassigned later.

In this model, citations are issued via two competing mechanisms:

1. A 'direct' mechanism, where all papers have an equal chance of being located and cited.

2. An 'indirect' mechanism, where papers are found through the reference list of one of the k papers which have already cited them. If that the number of references per paper is some fixed value n, the probability of indirectly citing a paper with k citations proportional to k/n.

Suppose that the probability of searching through a paper's reference list and citing one of its references is an unknown parameter c. The indirect mechanism is *imitation*: an author is citing a paper because it has already been cited by another paper. The parameter c represents the probability of an author citing a paper via imitation, and $1 - c$ is the probability that the author will instead cite a paper independently.

The competition between these mechanisms can be encoded into a master equation for $p(k)$, which is sufficiently simple that it may be solved analytically. (See the following chapter for details of the derivation.) The large-k tail scales as

$$p(k) \sim k^{-(1+1/c)}. \tag{1}$$

This model predicts that the slope of the power-law exponent should be dependent on a dataset-specific parameter, c, rather than a universal constant, and the transition to power-law scaling occurs when the indirect overtakes the direct mechanism. Consistent with this prediction, we discovered that a variable exponent is a feature of several empirical citations datasets. We also found that the best-fit value of n (the number of references given out per paper) is consistent with an empirical measurement of this quantity made in [12].

1.2 A more severe test?

The basic idea behind this model seems sufficiently general that I wondered if it could be applied to fit the probability distributions for a variety of phenomena that have roughly the same shape as the citations distribution. For example, can stock price fluctuations be explained in the same way, as a competition between the 'rational actor' mind of a stock trader and the herd mentality that sets in once enough people have bought or sold a particular stock? Does the distribution of links to web sites arise because of the competition between the tendency to randomly surf the Internet, and the many paths available to a site that has been linked to many times?

As shown in Figure 1, at first glance, the answer to these questions appears to be 'yes'. The $p(k)$ derived for the citations distribution fits these probability distributions quite well. However, is this really a severe test of the hypothesis? It is possible that there are many models capable of fitting to the distribution, particularly for intensely studied systems such as the stock market.

A more convincing test would be to measure each system in several independent ways, and show that a two-mechanism model successfully reproduces these features (and that other models do not). The trouble is that for citations, stock prices, and web links, the only readily-accessible data is of *counts*. It is straightforward to find the number of citations a paper has, but very hard to assemble a full list of which papers

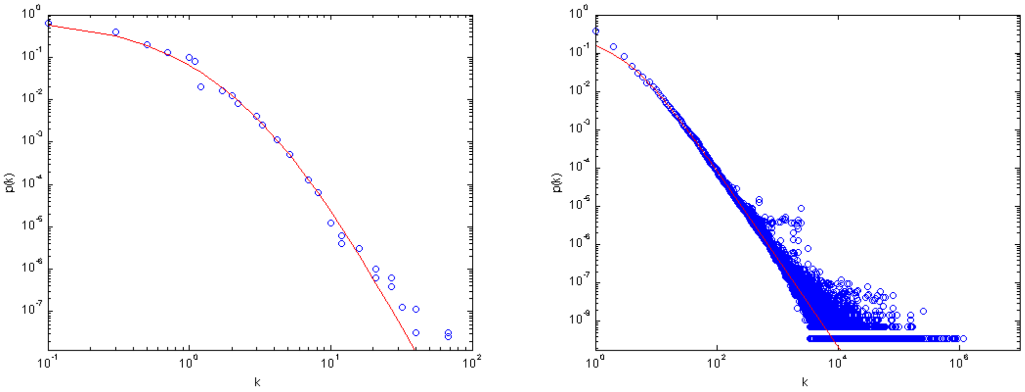


Figure 1: Curve fits of the citations model to (left) normalized 1-minute returns for the S&P 500 index ($c = 0.18$, $n = 0.25$) and (right) distribution of links to about 200 million websites in 1997 ($c = 0.76$, $n = 9.4$) [13]. Data is shown in blue, and best-fit $p(k)$ in red.

cited which other papers. Likewise, it is difficult merely to get access to minute-to-minute stock market data, and nearly impossible to get a complete list of buyers and sellers for each trade. This is a great advantage of protein-protein interaction (PPI) networks – it is simple to obtain a list of not only the number of interaction partners each protein has, but exactly which proteins interact with which other proteins.

1.3 Two mechanisms for PPI network evolution

PPI networks are an important souce of information about the complex machinery underlying cellular function. Clusters of interactions must be considered to analyze the functional role of proteins in cells, as proteins typically work in large-scale pathways to execute specific tasks. The probability distribution of the PPI network has the same general shape as the citations distribution; a recent statistical study showed that the tails of both distributions are best fit by power-laws [14].

In chapter 3, I discuss a model our group developed for the evolution of eukaryotic PPI networks. In this model, a cell's protein network evolves by two known biological mechanisms:

1. Gene duplication, which is followed by rapid diversification of duplicate interactions.

2. Neofunctionalization, in which a mutation leads to a new interaction with some other protein. Many interactions are nonspecific, arising from proteins that have exposed hydrophobic surface areas; these nonspecific interactions cause an increased likelihood of interacting with other proteins in the target protein's neighborhood.

The model is in good agreement on 10 different network properties compared to high-confidence experimental PPI networks in yeast, fruit flies, and humans. In this model, PPI networks evolve modular structures, with no need to invoke particular selection pressures. The model indicates that evolutionarily old proteins should have higher connectivities and be more centrally embedded in their networks. This suggests a way in which present-day proteomics data could provide insights into biological evolution.

1.4 A 'cost sharing' framework for social power-laws

Many of the real-world systems that exhibit power-laws are *social* phenomena, such as the fluctuations in the stock market, citations to scientific papers, the sizes of U.S. cities, etc. I was interested in whether power-laws might arise as a consequence of a general variational principle for stochastic processes. In chapter 4, I describe a model, based on the idea of communities of 'social particles', where the cost of adding a particle to the community is shared equally between the particle joining the community and the particles that are already members of the community. In this model, power-law probability distributions of community sizes arise as a natural consequence of the maximization of entropy, subject to this 'equal cost sharing' rule. I also explore a generalization in which there is unequal sharing of the costs of joining a community. Distributions change smoothly from exponential to power-law as a function of a sharing-inequality quantity. This chapter gives an interpretation of power-law distributions in terms of shared costs.

2 The probability distribution of scientific citations

Commonly observed in nature and in the social sciences are probability distribution functions that appear to involve dual underlying mechanisms, with a ‘tipping point’ between them. Examples of such probability distributions include the distributions of city sizes [15, 16]; fluctuations in stock market indices [3, 17]; U.S. firm sizes [18, 19]; degrees of Internet nodes [20, 14]; numbers of followers of religions [14]; gamma-ray intensities of solar flares [4]; sightings of bird species [14]; and citations of scientific papers [1, 21, 22, 23]. In these situations, a distribution $p(k)$ may have exponential behavior for small k and a power-law tail for large k. Here we develop a generative model for one such dual-mechanism process, scientific citations, for which databases are large and readily available. Here, k represents the number of citations a paper receives, ranging from 0 to hundreds or, sometimes, thousands. $p(k)$ is the distribution of the relative numbers of such citations, taken over a database of papers.

There have been several important studies of power-law tails of distributions, including those involving scientific citations. Price noted that highly cited scientific papers accumulate additional citations more quickly than papers that have fewer citations [24]. He called this ‘cumulative advantage’ (CA): the probability that a paper receives a citation is proportional to the number of citations it already has. Price showed that this rule asymptotically gives a power law for large k. Power-law tails have been widely explored in various contexts and under different names – ‘the rich get richer’, the Yule process [25, 26], the Matthew effect [27], or preferential attachment [5]. Barabási and Albert noted that networks, such as the World Wide Web, often have power-law distributions of vertex connectivities, called ‘scale-free’ behavior [5]. Their model, called preferential attachment, leads to a fixed power-law exponent of -3. Because many properties of physical systems near their critical points also display power-law behavior, and because such exponents are often *universal* (*i.e.,* independent of microscopic particulars of the system), it raises the question of which power-law distributions have universal exponents and which do not.

The tail of the scientific citations distribution has been fit by various distributions, including power law [1, 28], log-normal [29], and stretched exponential [30]. Recently, Clauset, Shalizi, and Newman proposed detailed statistical tests for determining whether various data sets have true power-law tails [14]. In agreement with Redner's earlier analysis [1], Clauset *et al.* confirm that the 1981 data set studied by Redner is indeed well-fit by a power-law.

Our interest here is not just in the large-k tails of such distribution functions. We are interested also in the small-k behavior and the tipping point between the two different regions. After all, the preponderance of scientific papers are not cited very commonly. Some previous models have explored both small-k and large-k regimes of citations. In 2001, Krapivsky and Redner developed a rate equation method to obtain solutions for several generalizations of the CA model, including results for nonlinear connection probabilities [31]. Krapivsky and Redner proposed a 'growing network with redirection' (GNR) for the citations network. They proposed that new papers could randomly cite existing papers, or could be *redirected* to one of the papers in its reference list. The GNR mechanism leads to a distribution with a *non-universal* scaling exponent, depending on the value of the redirection parameter. An analysis of this mechanism for arbitrary out-degree distribution was carried out by Rozenfeld and ben-Avraham [32]. Recently, Walker *et al.* proposed a redirection algorithm to rank traffic to *individual* papers, which, instead of an initial random attachment probability, used an exponentially decaying probability of citation, according to the age of the paper [33]. There have been many variations proposed of the basic CA model, including CA with error tolerance [6], with an attractiveness parameter [34], with a fitness parameter [7], with memory effects [8], with hierarchical organization

[35], with aging nodes [9], and a number of others. A useful overview of CA models, and power laws in general, is by Newman [4].

Here, we develop a model to address three points of particular interest to us. First, existing models focus on the power-law tail. We are interested here in the full distribution function and the nature of the transition, or the 'tipping point,' from one mechanism to the other. Second, we seek a mechanism that illuminates why the 'rich get richer' in scientific citations. Third, a strictly linear attachment rule predicts a single fixed exponent, $\gamma = 3$, where $p(k) \sim k^{-\gamma}$. Here, we ask whether the power-law exponent for scientific citations is a universal constant, as is often observed in the physics of critical phenomena, or whether the power-law exponent for citations is a non-universal parameter which varies from one dataset to another.

The two-mechanism model we propose here is similar to the GNR model studied in [31], generalized for an out-degree greater than one. A general treatment of the GNR model with arbitrary out-degree distribution was given in [32]. Here, we derive $p(k)$ explicitly for the case of a fixed out-degree, and analyze the 'tipping point' transition between the two mechanisms. We then fit our $p(k)$ to several citations datasets, and examine how the interactions between the two mechanisms produces different distributions (with different tipping points) for each dataset. By sorting our datasets according to h-index, we show that the scaling exponent, γ, decreases systematically with increasing values of h. We interpret the changes in the scaling exponent using a parameter of our model as an increasing bias towards indirect citation of well-known scientists.

2.1 A two-mechanism model

Consider a directed graph on which each node represents a scientific paper. Each edge represents a citation of one paper by another. An outgoing edge indicates *giving* a citation, and an incoming edge indicates *receiving* a citation. At a given time, the graph has N nodes, representing *old* papers that are already part of the graph. At each time step, a *new* paper is published (a node is added to the graph). Each new paper gives a fixed number of citations, n, distributed among the N old papers. Hence the total number of citations given is Nn, and the total number of citations received is also Nn. In general, we consider situations in which N is large. Let k be the number of incoming links (citations) that a paper has received. For example, a paper that has received no citations from other papers has $k = 0$. Some 'classic' papers have attracted more than $k = 1000$ citations. A given collection of papers will have a distribution, $p(k)$, of papers that have received $k = 0, 1, 2, \dots$ citations.

We first focus on a particular old paper, paper A. The probability that a new paper will randomly link to paper A is

$$r_{\text{direct}} = \frac{1}{N}. \tag{2}$$

We call Equation 2 the *direct mechanism* of citations.[1]

In addition, scientific papers are also cited by an *indirect mechanism*: the author of the new paper may first find a paper B and learn of paper A *via* B's reference list. On the citation graph, searching through B's reference list is a nearest-neighbor-link mechanism. Suppose there are already k incoming links to paper A. Because there are a total of nN incoming links to all papers, the probability that the author of the new paper randomly finds paper A, *via* the reference list of some other paper is

$$r_{\text{indirect}}(k) = \frac{k}{Nn}. \tag{3}$$

[1] Because each new paper will not cite an old paper more than once, the direct probability, Eq. 2, of the first citation is $1/N$, for the second citation is $1/(N-1)$, and so on, and for the n^{th} citation is $1/(N-n+1)$. For real-world graphs, however, N is of the order of $500,000$ and n is around 20. So, we assume $N \gg n$, and $1/(N-n+1) \sim 1/N$. Similarly, the indirect probability, as $Nn \gg n$, Eq. 3 is approximately $k/(Nn-n+1) \sim k/(Nn)$. Note also that, perhaps unrealistically, no special weight is given to the possibility of simultaneously citing both paper A and one of its references.

Given that the author of the new paper has found old paper A, the author will either cite a paper from A's reference list with probability c, or cite A itself with probability $1 - c$. If paper A currently has k citations, then the number of citations, $R(k)$, to paper A from a new paper, through either the direct or indirect mechanism, is

$$R(k) = n\Big[\,(1-c)\,r_{\text{direct}} + c\,r_{\text{indirect}}(k)\Big] = \frac{n(1-c)}{N} + \frac{kc}{N}. \tag{4}$$

Next, we compute the in-link distribution $p(k)$, the fraction of the N papers that have k incoming citations. The total number of papers having k citations is $Np(k)$.[2] We calculate $p(k)$ using a difference equation to express the flows into and out of the bin of papers having k citations for each time step (each time a new node is added). The population of the bin of papers with k citations increases every time a paper with $k - 1$ citations receives another citation and decreases every time a paper that already has k citations receives another citation,

$$\begin{aligned} p(k) &= N\Big[R(k-1)p(k-1) - R(k)p(k)\Big] \qquad (5)\\ &= \Big[n(1-c) + c(k-1)\Big]p(k-1) - \Big[n(1-c) + ck\Big]p(k). \end{aligned}$$

Equation 5 rearranges to:

$$p(k) = \frac{\alpha - 1 + k}{\alpha + 1/c + k} \cdot p(k-1). \tag{6}$$

where, to simplify the notation, we have defined

$$\alpha = \frac{n}{c} - n. \tag{7}$$

[2] The in-link distribution should be considered a function of both k and N, $p(k, N)$. However, we find that in the large N limit, the difference between $p(k, N)$ and $p(k, N - 1)$ decreases as $1/N$. It is therefore vanishingly small for very large N, and $\lim_{N\to\infty} p(k, N) = p(k)$.

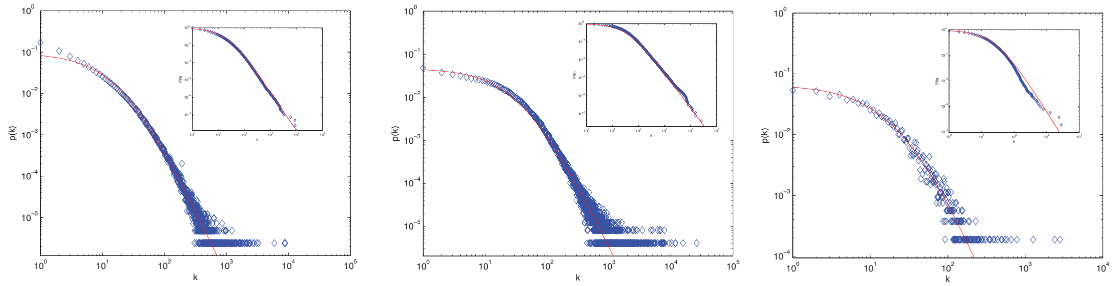


Figure 2: Probability of receiving exactly k citations (PDF) and at least k citations (CDF, inset) for datasets 1 (left), 2 (center), and 3 (right). Empirical data points are shown as blue diamonds, and best-fit curves as solid red lines.

The equation for $p(0)$ involves no inflow from a lesser bin. Instead, the inflow comes from the addition of a new paper per time step, which is 1 by definition. The outflow term is calculated as for other values of k. Therefore, $p(0) = 1 - n\,(1-c)\,p(0)$, which rearranges to:

$$p(0) = \frac{1}{n - nc + 1}. \tag{8}$$

Substituting in Equation 8 and applying Equation 6 recursively gives[3]

$$p(k) = \frac{1}{\alpha c + 1} \cdot \frac{(\alpha - 1 + k)!(\alpha + 1/c)!}{(\alpha - 1)!(\alpha + 1/c + k)!}. \tag{9}$$

When α is sufficiently large, we apply Stirling's approximation to Equation 9, which yields

$$\begin{aligned} p(k) \approx \frac{(\alpha + 1/c)^{\alpha + 1/c}}{(\alpha c + 1)\,(\alpha - 1)^{\alpha - 1}} \left(\frac{\alpha - 1 + k}{\alpha + 1/c + k} \right)^{\alpha + k} \\ \times (\alpha - 1 + k)^{-1} \left(\alpha + \frac{1}{c} + k \right)^{-1/c}. \end{aligned} \tag{10}$$

[3] The factorials in Equation 9 are understood to be gamma functions for non-integer $1/c$ values. To show that equation 9 is normalized, we use

$$\sum_{k=0}^{\infty} \frac{(\alpha - 1 + k)!}{(\alpha + 1/c + k)!} = (\alpha c + 1)\, \frac{(\alpha - 1)!}{(\alpha + 1/c)!}.$$

Substituting into 9, we find that $\sum_k p(k) = 1$, as required.

In the large-k tail ($k \gg \alpha$), we have

$$\left(\frac{\alpha - 1 + k}{\alpha + 1/c + k}\right)^{\alpha+k} \approx e^{-(1+1/c)},$$

and

$$(\alpha - 1 + k)^{-1}\left(\alpha + \frac{1}{c} + k\right)^{-1/c} \approx k^{-(1+1/c)}.$$

Therefore, Equation 10 becomes, in the large-k tail:

$$p(k) \approx \left[\frac{(\alpha + 1/c)^{\alpha+1/c} e^{-(1+1/c)}}{(\alpha c + 1)\,(\alpha - 1)^{\alpha-1}}\right] k^{-(1+1/c)}. \tag{11}$$

Equation 10 gives our model's prediction for the distribution of citations. It expresses both the direct and indirect citation mechanisms. Equation 11 indicates that once a paper's number of citations, k, is large enough, further citations of that paper undergo a sort of runaway growth because there are so many ways to find it through other papers that have already cited it; for scientific citations, 'the rich get richer.' The 'tipping point' where r_{indirect} overtakes r_{direct} happens at

$$k = \alpha. \tag{12}$$

For example, if $c = 1/2$ and the average paper in the database gives out $n = 15$ citations, then after any particular paper in that database has received 15 citations, it will begin to accumulate citations significantly faster than random – it will have 'tipped over' into the power-law scaling region. In this region, the power law exponent,

$$\gamma = 1 + \frac{1}{c}, \tag{13}$$

is determined by the parameter c. Hence, 'cumulative advantage' arises in our model because there are more routes (through the reference lists of other papers) for finding a classic paper than for finding a non-classic paper.

Table 1: Fitting parameters for datasets 1-3

Dataset	c	n	γ	α	N
1. All 1981 publications	0.454(4)	17.3(3)	3.20(2)	20.8(4)	415229
2. High h-index chemists	0.517(1)	42.0(1)	2.935(5)	39.2(1)	245461
3. *Phys. Rev. D* publications	0.48(3)	27(2)	3.1(1)	29(3)	5327

2.2 The datasets

Figure 2 shows fits to normalized empirical probability distribution functions (PDFs, the probability of receiving *exactly* k citations) and complementary cumulative distribution functions (CDFs, the probability of receiving *at least* k citations), $P(k) = \sum_{k'=k}^{\infty} p(k')$, for three datasets:

1. Citations of publications catalogued in the ISI Web of Science database in 1981 [1]
2. Citations of publications by authors on a 2007 list of the living highest h-index chemists [36]
3. Citations of publications in the *Physical Review D* journal from 1975-1994 [1]

Datasets 1 and 3 were downloaded from Sidney Redner's website[4]. We gathered dataset 2 from the ISI Web of Knowledge[5] using a Python script. Parameters for these fits are shown in Table 1, and plots of the datasets and best-fit $p(k)$ distributions are shown in Figure 2. We also sorted dataset 2 by h-index. Parameters for different h-index ranges are shown in Table 2, and fits are shown in Figure 3. The relation between our estimates of γ and h is shown in Figure 4. To obtain estimates and 95%

[4] http://physics.bu.edu/∼redner/projects/citation/index.html

[5] http://isiwebofknowledge.com

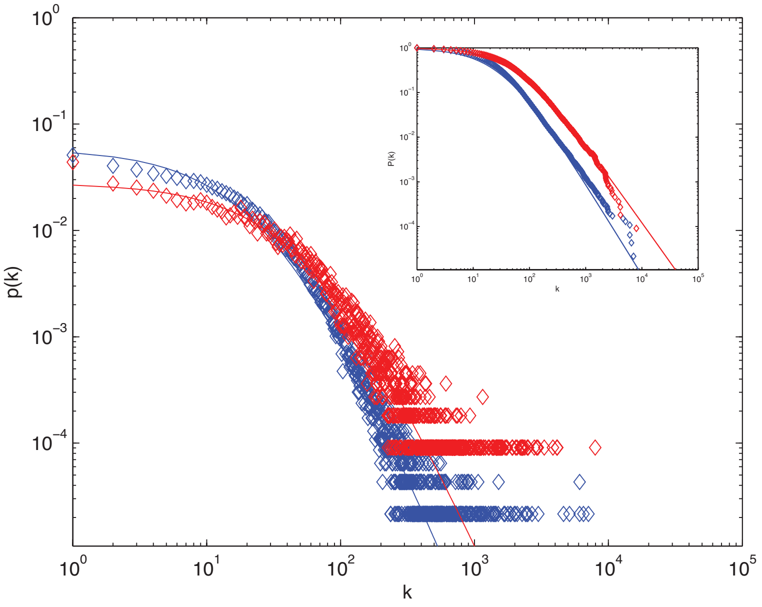


Figure 3: Comparison of the normalized PDFs and CDFs (inset) for chemists with $h = 100+$ (red) and chemists with $h = 50$-53 (blue).

confidence intervals of c and n, we used Matlab's implementation of the iteratively reweighted least squares algorithm, using bisquare weights [37]. All curve fitting was applied to the raw (not binned or log-transformed) data.

2.3 Results

Our model has two parameters: n, the average number of citations given out by all the papers in the database, and c, the chance of citing from a paper's reference list. The model power-law exponent is then fixed by the relationship $\gamma = 1 + 1/c$. Our best fit of dataset 1 gives a value of $n = 17.3 \pm 0.3$, in approximate agreement with the independent estimate of 15.01 found for papers published in 1980 [12]. Also, our predicted value of $\gamma = 3.20 \pm 0.02$ agrees with the best-fit power-law exponent previously found by Clauset, of $\gamma = 3.16$ [14]. Table 1 shows the best-fit parameter values for the three different datasets.

We explored the $p(k)$ distributions for small groups of scientists, as shown in Figure

3. We wanted to test an alternate hypothesis that some scientists might publish only low-k papers and others might publish only classic high-k papers. Our limited tests argue against this hypothesis. Figure 3 indicates that even highly cited scientists have more low-k papers than high-k papers. One reason is that every publication in the scientific literature is new for a while, and requires some time to become highly cited.

Interestingly, the slope of the power-law region differs between the two groups shown in Figure 2. To examine this difference in more detail, we parsed dataset 2 by h-index (Table 2). The h-index of a scientist is defined as the point where h of the scientist's papers have at least h citations each [38]. That is, h is defined by the requirement to satisfy the expression, $Np(h) = h$. There is no simple analytical relationship between a scientist's h-index and the parameters of our model.

From Table 2, we conclude that c increases with h-index, indicating that there is a bias towards selecting papers out of a reference list that were written by scientists who are already very highly cited (Figure 3). This bias may reflect the tendency of authors who, scanning a paper's references for further information, are more likely to select a paper written by an author they have previously heard of. The more highly cited the scientist, the lower his or her power-law exponent (*i.e.*, the fatter the tail); see Figure 4. The error bars are sufficiently small to indicate that these trends are real, and that there is not a single universal exponent, such as $\gamma = 3$; rather, the exponent depends on the subset of scientists examined. Note that, here, we consider a scientist to have authored a paper if his or her name appears anywhere in the list of authors. An interesting question for future work might be to examine whether this effect is changed by only considering the h-index of each paper's leading and/or corresponding author.

Our model bears some resemblance to Price's application of CA to scientific citations [24]. One key difference is that our two parameters both have physical meaning.

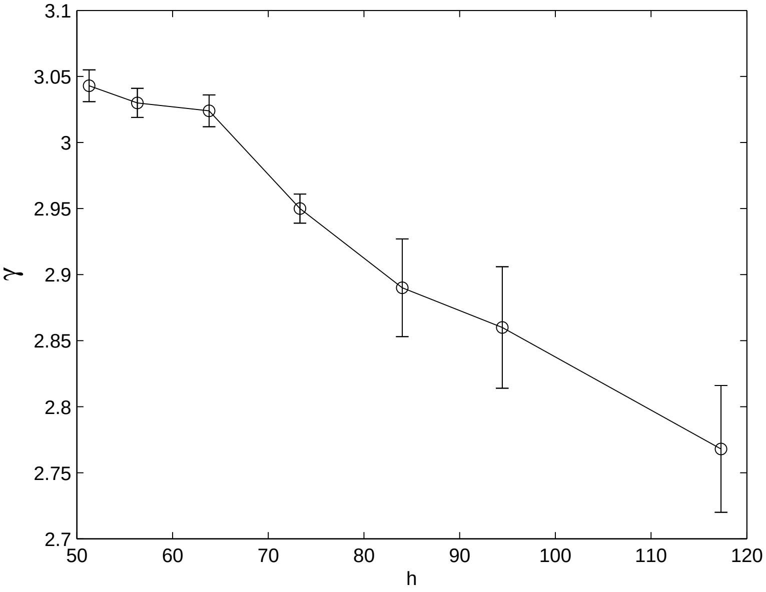


Figure 4: Power-law exponent γ plotted against h-index for subsets of dataset 2.

To avoid the issue of new papers having a citation probability of zero when $k = 0$, Price proposed that the citation probability should be proportional instead to $k + w$, where w is a constant that he refers to as a 'fudge factor.' He sets $w = 1$, although as later noted by Newman, there does not seem to be a good reason to choose this value [4]. The connection rule for our model is given by Equation 4, and suggests a simple interpretation: Price's constant arises from random connections, and the tipping point, Equation 12, is determined by the average size of the reference lists given out per paper, and the probability of searching through those reference lists.

This two-mechanism model also provides a justification for a CA mechanism. Barabási and Albert remarked that CA only produced a power law distribution when the connection probability was linearly proportional to k [5], but it was not clear what was special about linearity. The present model presents a possible explanation for the existence of this mechanism, and why the k dependence should be linear: k appears in r_{indirect} because a paper's k incoming citations are represented by k nearest-neighbor links on the graph.

2.4 Conclusion

We have developed a model of scientific citations, involving both direct and indirect routes to finding and citing papers. This two-mechanism model predicts exponential behavior in the small-k region and power law tails in the large-k region. One parameter of the model, n, is the average number of citations given out per paper. Our best-fit value of n is consistent with an independent, empirical measure of it made by Biglu [12]. Our other parameter, c, defines the power-law exponent, $\gamma = 1 + 1/c$, which is in agreement with data previously evaluated in [14]. Two key findings here are: (1) the tipping point for a paper to reach 'classic-paper' status, *i.e.* its power-law citation region, is about 21 citations for the ISI Web of Science database, and (2) the power-law exponent is not a universal feature of all scientific citations. The exponent diminishes systematically with increasing h-index of a scientist. Our model describes systems that are governed by random choices in the small-k region, cumulative advantage in the high-k region, and a tipping point between them.

Table 2: Fitting parameters for h-index ranges within dataset 2

h range	c	n	γ	α	N
100+	0.57(1)	80(3)	2.77(5)	60(2)	11029
90-99	0.54(1)	77(3)	2.86(5)	66(3)	11476
80-89	0.53(1)	60(2)	2.89(4)	53(2)	15408
70-79	0.513(3)	40.6(4)	2.95(1)	38.5(4)	54236
60-69	0.494(2)	48.7(4)	3.02(1)	49.9(5)	56052
54-59	0.493(3)	34.9(3)	3.03(1)	35.9(4)	44715
50-53	0.489(3)	31.3(3)	3.04(1)	32.7(4)	46421

3 The evolution of protein-protein interaction networks

We are interested in the evolution of protein-protein interaction (PPI) networks. PPI network evolution accompanies cellular evolution, and may be important for processes such as the emergence of antibiotic resistance in bacteria [39, 40], the growth of cancer cells [41], and biological speciation [42, 43, 44]. In recent years, increasingly large volumes of experimental PPI data have become available [45, 46, 47, 48], and a variety of computational techniques have been created to process and analyze these data [49, 50, 51, 52, 53, 54, 55, 56]. Although these techniques are diverse, and the experimental data are noisy [57], a general picture emerging from these studies is that the evolutionary pressures shaping protein networks are deeply interlinked with the networks' topology [58]. Our aim here is to construct a minimal model of PPI network evolution which accurately captures a broad panel of topological properties.

In this work, we describe an evolutionary model for eukaryotic PPI networks. In our model, protein networks evolve by two known biological mechanisms: (i) a gene can duplicate, putting one copy under new selective pressures that allow it to establish new relationships to other proteins in the cell, and (ii) a protein undergoes a mutation that causes it to develop new binding or new functional relationships with existing proteins. In addition, we allow for the possibility that once a mutated protein develops a new relationship with another protein (called the target), the mutant protein can also more readily establish relationships with other proteins in the target's neighborhood. One goal is to see if random changes based on these mechanisms could generate networks with the properties of present-day PPI networks. Another goal is then to draw inferences about the evolutionary histories of PPI networks.

3.1 Model

We represent a PPI network as a graph. Each node on the graph represents one protein. A link (edge) between two nodes represents a physical interaction between the two corresponding proteins. The links are undirected and unweighted. To model the evolution of the PPI graph, we simulate a series of steps in time. At time t, one protein in the network is subjected to either a gene duplication or a neofunctionalizing mutation, leading to an altered network by time $t + \Delta t$. We refer to this model as the DUNE (DUplication & NEofunctionalization) model.

3.1.1 Gene duplication

One mechanism by which PPI networks change is gene duplication (DU) [59, 60, 61]. In DU, an existing gene is copied, creating a new, identical gene. In our model, duplications occur at a rate d, which is assumed to be constant for each organism. All genes are accessible to duplication, with equal likelihood. For simplicity, we assume that one gene codes for one protein. One of the copies continues to perform the same biological function and remains under the same selective pressures as before. The other copy is superfluous, since it is no longer essential for the functioning of the cell [62].

The superfluous copy of a protein/gene is under less selective pressure; it is free to lose its previous function and to develop some other function within the cell. Due to this reduced selective pressure, further mutations to the superfluous protein are more readily accepted, including those that would otherwise have been harmful to the organism [63, 64]. Hence, a superfluous protein diverges rapidly after its DU event [65]. This well-known process is referred to as the *post-duplication divergence*. Following [66], we assume that the link of each such superfluous protein/gene to its former neighbors is deleted with probability ϕ. The post-duplication divergence tends to be fast; for simplicity, we assume the divergence occurs within the same time step as the DU. The divergence is *asymmetric* [67, 68]: one of the proteins diversifies rapidly, while the other protein retains its prior activity. We delete links from the original or the duplicate with equal probability because the proteins are identical.[6] In our model, ϕ is an adjustable parameter.

In many cases, the post-duplication divergence results in a protein which has lost all its links. These 'orphan' proteins correspond to silenced or deleted genes in our model. As discussed below, our model predicts that the gene loss rate should be slightly higher than the duplication rate in yeast, and slightly lower in flies and humans.

We simulate a DU event at time t as follows:

1a) Duplicate a randomly-chosen gene with probability $d\Delta t$.

2a) Choose either the original (50%) or duplicate (50%), and delete each of its links with probability ϕ.

3a) Move on to the next time interval, time $t + \Delta t$.

3.1.2 Neofunctionalization

Our model also takes into account that DNA can be changed by random mutations. Most such mutations do not lead to changes in the PPI network structure. However, some protein mutations lead to new interactions with some other protein (which we call the *target protein*). The formation of a novel interaction is called a *neofunctionalization* (NE) event. NE refers to the creation of new interactions, not to the disappearance of old ones. Functional deletions tend to be deleterious to organisms [69]. We do not account for loss-of-function mutations (link deletions) except during post-duplication divergence because damaged alleles will, in general, be eliminated by purifying selection. In our model, NE mutations occur at a rate μ, which is assumed to be constant. All proteins are equally likely to be mutated.

[6] As discussed in the supporting information (SI), this is closely related to the idea of *subfunctionalization*, where divergence freely occurs until redundancy is eliminated.

How does the mutated protein choose a target protein to which it links? We define a probability q that any protein in the network is selected for receiving the new link from the mutant protein. To account for the possibility of multimerization, the mutated protein may also link to itself. Random choice dictates that $q = 1/N$ (see SI).

Many PPI's are driven by a simple geometric compatibility between the surfaces of the proteins [70]. The simplest example is the case of PPI's between flat, hydrophobic surfaces [71], a type of interaction which is very common [72]. These PPI's have a simple planar interface, and the binding sites on the individual proteins are geometrically quite similar to one another. One consequence of these similar-surface interactions is that if protein A can bind to proteins B and C, then there is a greater-than-random chance that B and C will interact with each other. We refer to this property as *transitivity*: if A binds B, and A binds C, then B binds C. The number of triangles in the PPI network should correlate roughly with transitivity. As discussed below, the number of triangles (as quantified by the global clustering coefficient) is about 45 times higher in real PPI networks than in an equally-dense random graph. This suggests that transitivity is quite common in PPI networks.[7]

A concrete example of transitivity is provided by the evolution of the retinoic acid receptor (RAR), an example of neofunctionalization which has been characterized in detail [74]. Three paralogs of RAR exist in vertebrates (RARα, β, and γ), as a result of an ancient duplication. The interaction profiles of these proteins are quite different. Previous work indicates that RARβ retained the role of the ancestral RAR [74], while RARα and γ evolved new functionality. RARα has several interactions not found in RARβ. RARα has novel interactions with a histone deacetylase (HDAC3) as well

[7] Another source of transitivity is gene duplication. If A binds B, then A is copied to create a duplicate protein A', then A' will (initially) also bind B. If A interacts with A', then a triangle is formed. However, duplication is unlikely to be the primary source of transitivity; recent evidence shows that, due to the post-duplication divergence, duplicates tend to participate in fewer triangles than other proteins [73].

as seven of HDAC3's nearest-neighbors (HDAC4, MBD1, Q15959, NRIP1, Q59FP9, NR2E3, GATA2). None of these interactions are found in RARβ. The probability that all of these novel interactions were created independently is very low. RARα has 65 known PPI's and HDAC3 has 83, and the present-day size of the human PPI network is a little over 3000 proteins. Therefore, the chance of RARα randomly evolving novel interactions with 7 of HDAC3's neighbors is less than 1 in a billion. This strongly suggests that when a protein evolves an interaction to a target, it has a greater-than-random chance of also linking to other, neighboring proteins.

How do similar-surface interactions affect the evolution of PPI networks? First, consider how an interaction triangle would form. Suppose proteins A and B bind due to physically similar binding sites. Protein X mutates and evolves the capacity to bind A. There is a reasonable chance that X has a surface which is similar to both A and B. If so, protein X is likely to also bind to B, forming a triangle. Denote the probability that two proteins interact due to a simple binding site similarity by a. The probability that A binds B (and X binds A) in this manner is a. Assuming these probabilities are identical and independent, the probability that X binds B is a^2.

So far, we have discussed transitivity as it affects the PPI's in which protein A is directly involved (A's first-neighbors). We now introduce a third protein to the above example, resulting in a chain of interactions: protein A binds B, B binds C, but C does not bind A. Protein X mutates and gains an interaction with A (with probability a^2). What is the probability that X will also bind C? The probability that B binds C due to surface similarity is a. Thus, X will bind C (A's second-neighbor) with probability a^3. In general, the probability that X will bind one of A's j^{th} neighbors is a^{j+1}. We refer to this process as *assimilation*, and the 'assimilation parameter' a is a constant which varies between species.[8] Assimilation is assumed to act on a much

[8] As discussed in SI, it is primarily mutliple-partner proteins which bind to their partners at different times and/or locations which are affected by this process; consequently, at most one link is created by assimilation at the first-neighbor level, second-neighbor level, etc.

shorter time scale than DU and NE; in our model, it is instantaneous.

As discussed above, evidence suggests that a protein which evolves a PPI to a target protein should also have a greater-than-random chance of forming interactions with its neighbors. Although this is not direct proof that the assimilation mechanism exists exactly as proposed here, our proposed mechanism makes several predictions that could be tested experimentally: (1) the probability of a protein assimilating into a new pathway should be a^2 (at the first-neighbor level), a^3 (at the second-neighbor level), and so on, where a is a constant which varies between species; (2) weak, nonspecific binding and planar interfaces should be overrepresented in interaction triangles (and longer cycles) between non-duplicate proteins; (3) competitive inhibitors should be overrepresented in interaction triangles; and (4) domain shuffling should be associated with assimilation. (See SI for discussion of (3) and (4).)

We simulate an NE event at time t as follows:

1b) Mutate a randomly-chosen gene with probability $\mu\Delta t$.

2b) Link to a randomly-chosen target protein.

3b) Add a second link to one of the target's first-neighbor proteins, chosen randomly, with probability a^2.

4b) Add a link to one of the target's second-neighbor proteins, with probability a^3, etc.

5b) Move on to the next time interval, time $t + \Delta t$.

3.1.3 Model simulation and parameters

A flowchart of how PPI networks evolve in our model is shown in Figure 5. To simulate the network's evolution, one of the two mechanisms above is used at each time step, using Gillespie's algorithm [75]. We call each possible time series a *trajectory.* We begin each trajectory starting from two proteins sharing a link (the simplest configuration that is still technically a network). Each simulated trajectory ends when

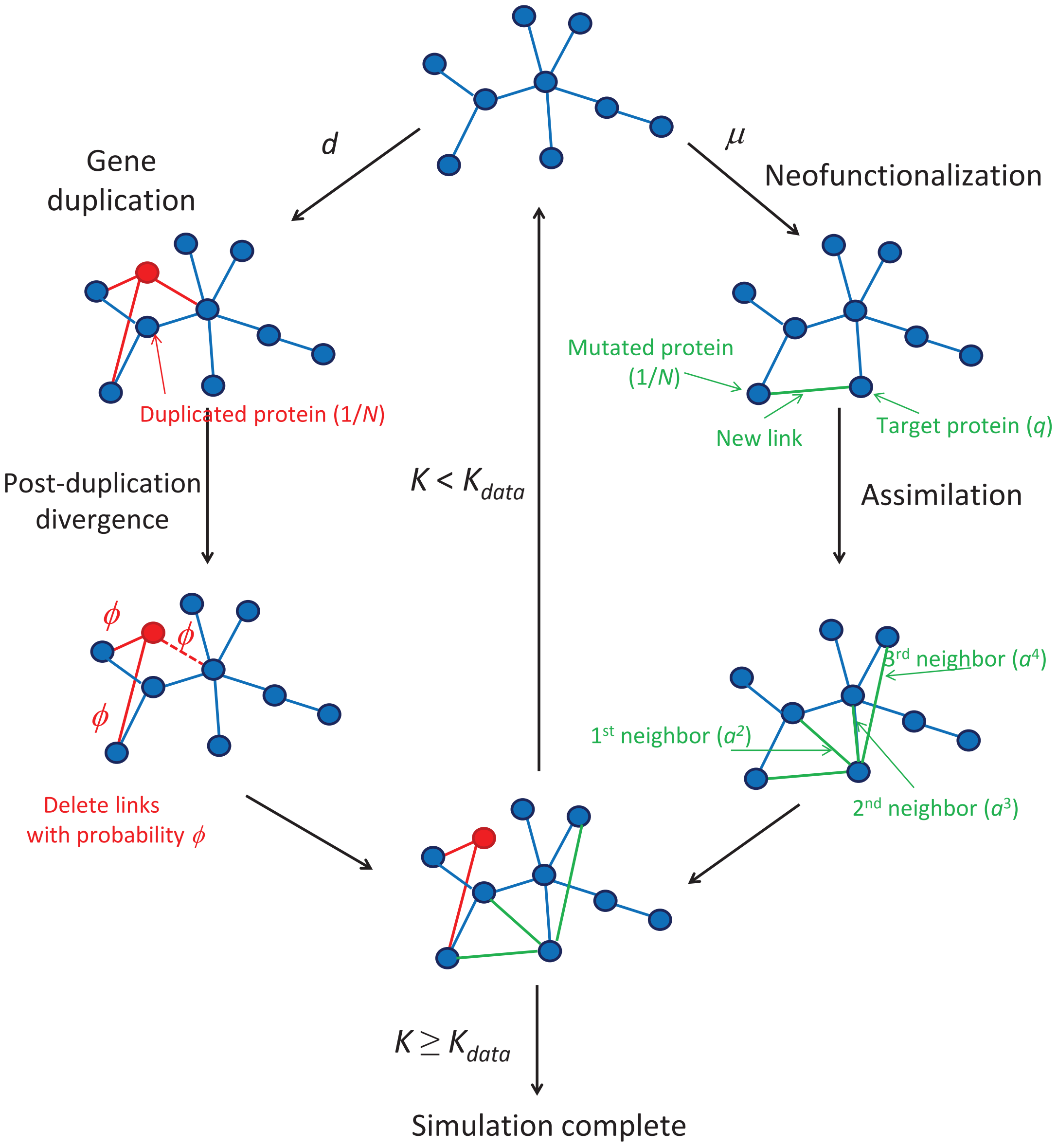


Figure 5: DUNE model flowchart. At each time step, the simulated network undergoes a DU or NE event. Red nodes/links indicate nodes/links that have been created by DU during the current time step. Green links indicate links that have been created by NE during the current time step. A dashed line indicates a duplicated link that has been deleted during the post-duplication divergence (with probability ϕ). Only up to 3 neighbors are shown for the assimilation mechanism; however, the actual simulations included up to 20th neighbors. The simulated network evolves until its number of links (K) meets or exceeds the number of links in the data (K_{data}).

	N_{data}	K_{data}	d	μ	ϕ	a
Yeast	2170	3819	0.01	7.86×10^{-4}	0.555	0.690
Fly	878	1140	0.0014	5.89×10^{-4}	0.866	0.546
Human	3165	5547	0.0037	7.62×10^{-4}	0.652	0.727

Table 3: Network sizes and model parameters. N and K are the numbers of proteins and links, respectively. (K_{data} is used to stop the simulation. N_{data} is not used as a constraint.) d and μ have units of per gene per million years (Myr). ϕ and a are probabilities (unitless). K and d are constraints from the data, while μ, ϕ, and a are adjustable parameters. We used Monte Carlo simulations to optimize the parameter values, by minimizing the total symmetric mean absolute percentage error values of the simulated versus the experimental data (see SI for details). Our values of μ are substantially lower than d because μ is the rate of mutations leading to the creation of a new PPI (rather than being a simple mutation rate, which would be much higher).

the model network has grown to have the same total number of links, K, as found in the experimental data, K_{data}. Here, we perform sets of simulations for three different organisms: yeast (*Saccharomyces cerevisiae*), fruit flies (*Drosophila melanogaster*), and humans (*Homo sapiens*). Because evolution is stochastic, there are different possible trajectories, even for identical starting conditions and parameters. We simulated 50 trajectories for each organism. Our figures below show the median values of each feature as a heavy line, and individual trajectories as light lines.

For a given data set, the number of links (K_{data}) is known. We estimate the DU rate d from literature values. There have been several empirical estimates of DU rates, mostly falling within an order of magnitude of each other [65, 76, 77, 78, 79, 78, 80, 81]. We averaged together the literature values to estimate d for each species (Table 7).

The quantity μ is not as well known. Its value relative to d has been the topic of considerable debate [62, 82, 83, 84]. Although, in principle, μ is a measurable quantity, it has proven difficult to obtain an accurate value, in part because the fixation rate of NE alleles varies with population size [85, 86]. In the absence of a consensus order-of-magnitude estimate, in our model, we treat μ as a fitting parameter. Consistent with the findings of [87] and [82], our best-fit values of μ are within an order of magnitude of each other for yeast, fruit fly, and human networks. Best-fit parameter values are given in Table 7.

3.2 Results

3.2.1 Present-day network topology

One test of an evolutionary model is its predictions for present-day PPI network topologies. Current large-scale PPI data sets have a high level of noise, resulting in significant problems with false positives and negatives [57, 88]. To mitigate this, we compare only to 'high-confidence' experimental PPI network data gathered in small-scale experiments (see Methods). We computed 10 topological features, quantifying various static and dynamic aspects of the networks' global and local structures: degree, closeness, eigenvalues, betweenness, modularity, diameter, error tolerance, largest component size, clustering coefficients, and assortativity. 7 of these properties are described below (see SI for the others). These network properties are largely uncorrelated (see SI).

The *degree* k of a node is the number of links connected to it. For protein networks, a protein's degree is the number of proteins with which it has a direct relationship. Some proteins interact with few other proteins, while other proteins (called 'hubs') interact with many other proteins. Previous work indicates that hubs have structural and functional characteristics that distinguish them from non-hubs, such as increased proportion of disordered surface residues and repetitive domain structures [91]. The high degree of a protein hub could indicate that protein has unusual biological significance [92]. The network's overall link density is described by its mean degree, $\langle k \rangle$ (Table 4). The *degree distribution* $p(k)$ is the probability that a protein will have k links. PPI networks have a few hub proteins and many relatively isolated proteins. The heavy tail of the degree distribution shows that PPI networks have significantly more hubs than random networks have. Simulated and experimental degree distribu-

	Q	D	f_1	$\langle C \rangle$	$\langle k \rangle$
Yeast data	**0.75**	**15**	**0.89**	**0.09**	**3.65**
DUNE	0.74(7)	17(6)	0.8(1)	0.041(9)	4.0(8)
Vázquez	0.80(4)	21(5)	0.2(1)	0.045(5)	2.6(4)
Berg	0.518(4)	12.0(7)	0.990(3)	0.0027(9)	4.10(3)
RG	0.910(3)	36(3)	0.987(6)	0.475(8)	5.31(8)
MpK	0.58(6)	24(5)	1.000(2)	0.08(3)	4.4(6)
ER	0.588(8)	13.0(9)	0.995(2)	0.002(1)	3.5(6)
Fly data	**0.86**	**23**	**0.73**	**0.10**	**2.93**
DUNE	0.82(2)	20(2)	0.81(3)	0.09(1)	2.36(9)
Human data	**0.75**	**15**	**0.88**	**0.08**	**3.69**
DUNE	0.74(6)	17(2)	0.88(4)	0.09(1)	3.7(4)

Table 4: Modularity Q, diameter D, fraction of nodes in the largest component f_1, global clustering coefficient $\langle C \rangle$, and $\langle k \rangle$ is the average degree of proteins the largest component. 'Data' is the empirical data, 'DUNE' is the model described here, 'Vázquez' is the duplication-only model of [66], 'Berg' is the link dynamics model [2], 'RG' is random geometric [89], 'MpK' is the physical desolvation model presented in [88], and 'ER' is an Erdős-Rényi random graph [90]. Simulated values are the median (± standard deviation) over 50 simulations. (See SI for details of each model's setup and optimization.)

tions are compared in Fig. 6. (For quantitative comparisons, see SI.)

Component refers to a set of reachable proteins. If any protein is reachable from any other protein (by hopping from neighbor to neighbor), then the network only has one component. If there is no path leading from protein A to B, then A and B are in different components. The fraction of nodes in the largest component (f_1) is a measure of network fragmentation (Table 4 and Fig. 19). Note that, although silent genes (proteins with no links) exist in real systems, these genes do not appear in data sets consisting only of PPI's. Therefore, calculations of f_1 for all models exclude orphan proteins (proteins with $k = 0$).

Gene loss, the silencing or deletion of genes, is known to play an important role in evolution. The loss of a functioning gene will damage an organism, making the gene loss unlikely to be passed on. The exception is if the gene is redundant. Consistent with this reasoning, evidence suggests that many gene loss events are losses of one

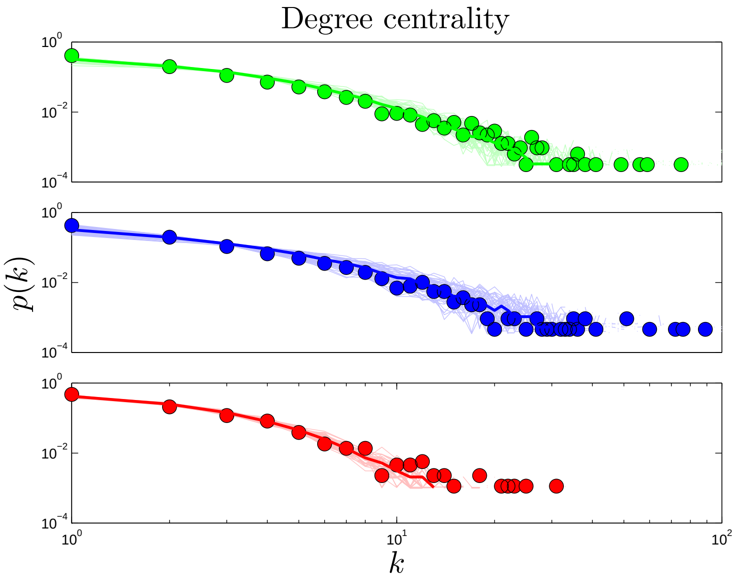


Figure 6: Degree (k) distributions in human (green), yeast (blue), and fly (red). Heavy lines are the median values from 50 simulations, and light lines are results of individual simulations. Points represent high-confidence empirical data for each organism (see Methods). Unless otherwise noted, color coding in the same in all plots. Quantitative comparisons between simulation and experiment (for DUNE and several other models) are detailed in SI.

copy of a duplicated gene [93, 67]. Although empirical estimates of the gene loss rate varied considerably, a consistent finding across several studies is that the rates of gene duplication and loss are of the same order-of-magnitude [65, 80, 77]. This broad picture is in good agreement with our model. In our model, a gene is considered lost when it has degree zero. Our model predicts that the ratio of orphan to non-orphan proteins is 1.6 ± 0.4 in yeast, 0.58 ± 0.06 in flies, and 0.67 ± 0.09 in humans. The gene loss rate has been previously estimated to be about half the duplication rate in both flies and humans [65, 80], consistent with our model's prediction.

The *distance* between nodes i and j is defined as the number of node-to-node steps that it takes along the shortest path to get from node i to j. The *closeness centrality* of a node i, ℓ_i, is the inverse of the average distance from node i to all other nodes in the same component. The *diameter*, D, of a network is the longest distance in the network. Simulated closeness distributions are compared to experiments in

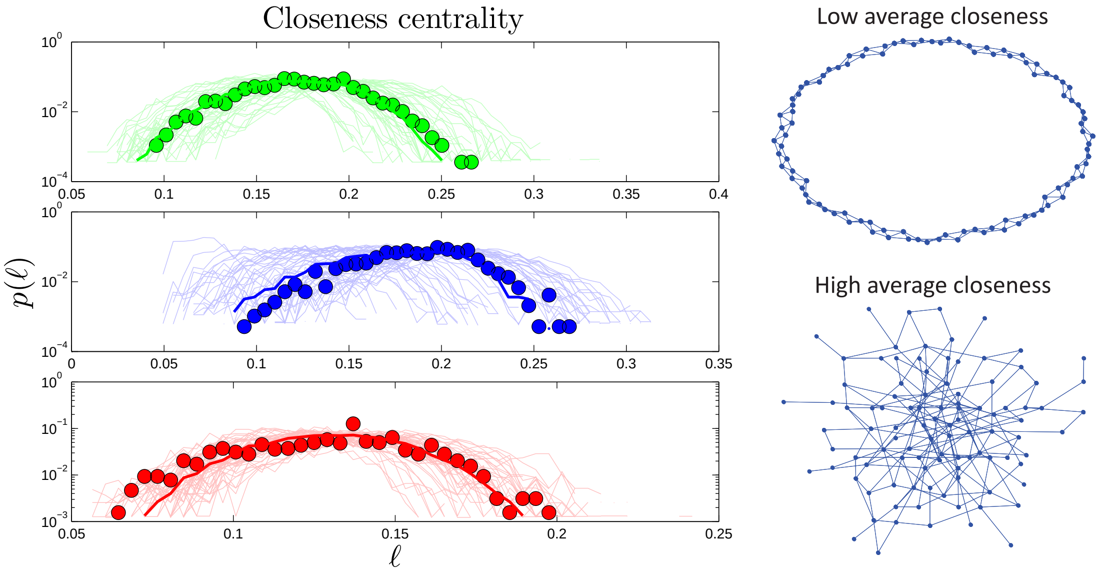


Figure 7: (Left) Closeness (ℓ) distributions in human (green), yeast (blue), and fly (red). Heavy lines are the median values from 50 simulations, and light lines are results of individual simulations. (Right) Examples of networks with low average closeness $\langle \ell \rangle = 0.06$ (top; each node is generally far away from most other nodes because there are no 'short cuts') and high average closeness $\langle \ell \rangle = 0.28$ (bottom; the random connections allow each node to be only a short distance from the other nodes). Note that both networks pictured here have the same number of nodes ($N = 100$) and roughly the same average degree (top: $\langle k \rangle = 4$, bottom: $\langle k \rangle = 3.7$).

Fig. 7. Interestingly, proteins have about 'six degrees of separation', similar to social networks [94, 95]. The closeness distribution $p(\ell)$ has a peak around $1/\ell \approx 5 - 7$.

Another property of a network is its *modularity* [96]. Networks are modular if they have high densities of links (defining regions called modules), connected by lower densities of links (between modules). One way to quantify the extent of modular organization in a network is to compute the modularity index, Q [97, 98]:

$$Q \equiv \frac{1}{K} \sum_{i,j}^{N} \left(A_{ij} - \frac{k_i k_j}{K} \right) \delta(u_i, u_j), \tag{14}$$

where k_i and k_j are the degrees of nodes i and j, u_i and u_j denote the modules to which nodes i and j belong, $\delta(u_i, u_j) = 1$ if $u_i = u_j$ and $\delta(u_i, u_j) = 0$ otherwise, and $A_{ij} = 1$ if nodes i and j share a link, and $A_{ij} = 0$ otherwise. Q quantifies the difference between the actual within-module link density to the expected link density in a randomly connected network. Q ranges between -1 and 1; positive values of Q indicate that the number of links within modules is greater than random.[9] As shown in Table 4, PPI networks are highly modular, and our simulated Q values are in good agreement with those of experimental data.

The *clustering coefficient*, C_i, for a protein i, is a measure of mutual connectivity of the neighbors of protein i. C_i is defined as the ratio of the actual number of links between neighbors of protein i to the maximum possible number of links between them,

$$C_i = \frac{\#\text{ edges between neighbors of node } i}{k_i(k_i - 1)}. \tag{15}$$

In a PPI network, clustering is thought to reflect the high likelihood that proteins of similar function are mutually connected [100]. The average (or global) clustering coefficient, $\langle C \rangle$, quantifies the extent of clustering in the network as a whole. As shown in Table 4, PPI networks have large global clustering coefficient values; the yeast PPI network, for example, has a value of $\langle C \rangle$ which is 45 times higher than that of a random graph of equivalent link density. In flies and humans, our simulated networks have $\langle C \rangle$ values in excellent agreement with the data; in yeast, our predicted value is slightly low.

A network is said to be 'hierarchically clustered' if the clustering coefficient and degree obey a power-law relation, $C \sim k^{-\xi}$ [101] (Fig. 13), indicating that nodes are organized into small-scale modules, and the small-scale modules are in turn organized into larger-scale modules following the same pattern [102]. By plotting the median clustering coefficient (per node) against degree, we observed a trend consistent with hierarchical clustering, although data in the tail is very limited.

[9] The numerical value of Q required for a network to be considered 'modular' depends on the number of nodes and links and method of computation. To calibrate baseline Q values given our particular network data, we used the null model described in [99]. Our non-modular baseline values are $Q = 0.603$ for the human PPI net, $Q = 0.590$ for yeast, and $Q = 0.722$ for flies (see SI).

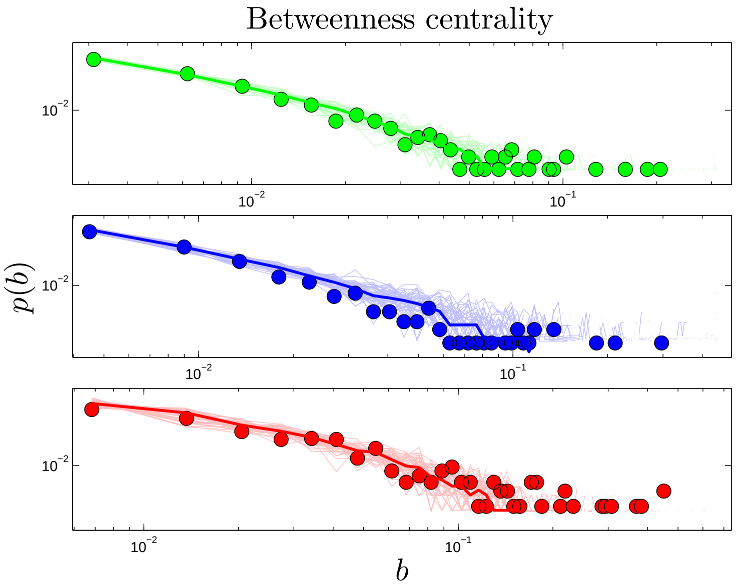


Figure 8: Betweenness (b) distributions in human (green), yeast (blue), and fly (red). Heavy lines are the median values from 50 simulations, and light lines are results of individual simulations.

The *betweenness* of a node measures the extent to which it 'bridges' between different modules. *Betweenness centrality*, b, is defined as:

$$b_i \equiv \frac{\#\text{ shortest paths passing through node } i}{\#\text{ total shortest paths}}. \tag{16}$$

Betweenness has been proposed as a uniquely functionally-relevant metric for PPI networks because it relates local and global topology. It has been argued that knocking out a protein that has high betweenness may be more lethal to an organism than knocking out a protein of high degree [103]. Betweenness distributions are shown in Fig. 8.

If a network's well-connected nodes are mostly attached to poorly-connected nodes, the network is called *disassortative*. A simple way to quantify disassortativity is by determining the median degree of a protein's neighbors (n) as a function of its degree (k). Previous work has found that yeast networks are disassortative [99]. It has been argued that disassortativity is an essential feature of PPI network

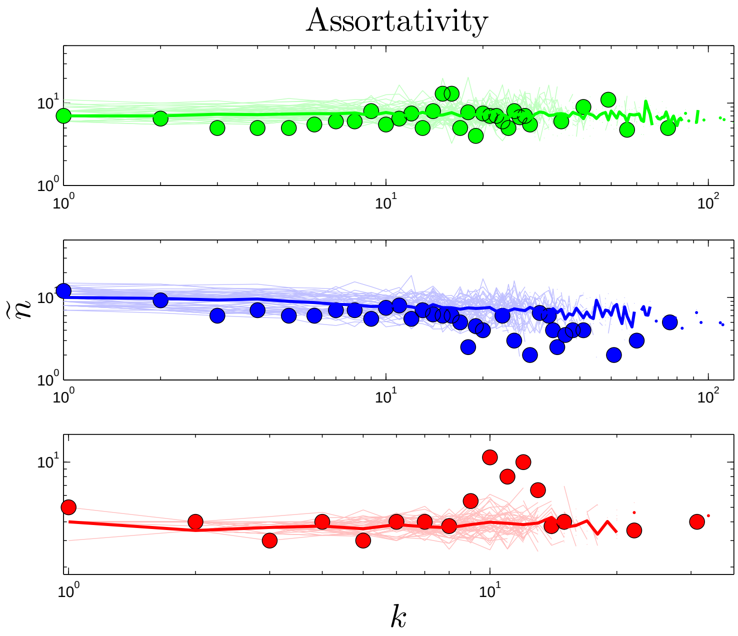


Figure 9: Median nearest-neighbor degree vs. degree in human (green), yeast (blue), and fly (red). Heavy lines are the median values from 50 simulations, and light lines are results of individual simulations.

evolution, and recent modeling efforts have heavily emphasized this feature [104, 105]. However, it was argued in [106] that disassortativity may simply be an artifact of the yeast two-hybrid technique, and [107] pointed out that this trend is quite different among different yeast datasets, and in some cases is completely reversed, resulting in *assortative* mixing, where high degree proteins prefer to link to other high-degree proteins. As shown in Fig. 9 and Table 5, the empirical data shows no evidence of disassortativity in flies or humans, and even the trend in yeast is quite weak. This conclusion is based solely on analysis of the empirical data, and casts further doubt on the role of disassortative mixing in PPI network evolution.

3.2.2 Evolutionary trajectories

We now consider the full time trajectories and the question of how PPI networks evolve in time. The static snapshots show a rich-get-richer structure: protein nets tend to have both more well-connected nodes and more poorly connected nodes than random networks have. In our model, the rich-get-richer property has two bases: duplication and assimilation. The equal duplication chance per protein means the probability for a protein with k links to acquire a new link via duplication of one of its interaction partners is proportional to k. Likewise, the probability of a protein to receive a link from the first-neighbor assimilation probability a is proportional to its degree k. 'Rich' proteins get richer because the probability of acquiring new links rises with the number of existing links.

First, we discuss two dynamical quantities for which experimental evidence exists: the rate of gene loss, and the relation between a protein's age and its centrality. Gene losses in our model correspond to 'orphan' proteins which have no interactions with other proteins. As shown in Fig. 19, the fraction of orphan proteins grows quickly at first, then levels off. This is consistent with the findings of [80]: in humans, while the overall duplication rate is higher than the loss rate, when only data from the past 200 Myr are considered, the loss rate is slightly higher than the duplication rate. In our model, after the initial rapid expansion, the rate of gene loss stabilizes relative to the duplication rate.

Our model shows that a protein's age correlates with certain network properties. Consistent with earlier work [31, 108, 109], we find that older proteins tend to be more highly connected. Interestingly, our model suggests that other centrality scores (betweenness and closeness) are also correlated with age. Fig. 18 shows DUNE's prediction that a protein's age correlates with degree, betweenness, and closeness centrality. We confirmed this prediction by following the evolutionary trajectories of individual proteins (Fig. 20). This suggests a way that present-day data could be used describe the evolution of PPI networks. These results are consistent with the eigenvalue-based aging method described in [109] (Fig. 21). Phylogenetic protein age estimates indicate that older proteins tend to have a higher degree [108, 109], which the DUNE model correctly predicts. Interestingly, the eigenvalue-based scores

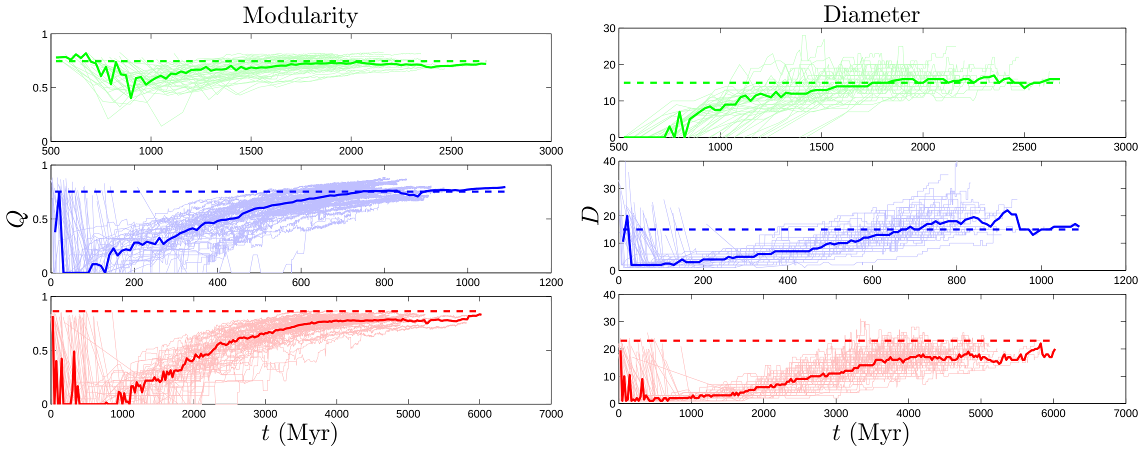


Figure 10: Network modularity Q and diameter D are both predicted to grow with time in human (green), yeast (blue), and fly (red). Light lines indicate the evolutionary trajectories of 50 individual simulations, and the heavy line is the median value. The modularity and diameter of the empirical data are shown as dashed horizontal lines. Time traces occasionally do not start at $t = 0$ because these simulations spend the first few time steps in a completely disconnected state, so the dynamical quantities are undefined. (See Fig. 19 for other dynamical plots.)

are only modestly correlated with other centrality scores (0.36 degree, 0.47 betweenness, and 0.10 closeness correlations). Using the eigenvalue method in tandem with our centrality-based method could provide stronger age-discriminating power for PPI networks than either method alone.

In contrast to the two dynamical quantities discussed so far, most structural properties of PPI networks have only been measured for the present-day network. Although our model accurately reproduces the present-day values of these quantities, there is no direct evidence that the simulated trajectories are correct; rather, these are predictions of our model. Fig. 10 shows that both modularity Q and diameter D increase with time. These are not predictions that can be tested yet for biological systems, since there is no time-resolved data yet available for PPI evolution. Time-resolved data is only currently available for various social networks (links to websites, co-authorship networks, etc.). Interestingly, the diameters of social networks are found to shrink over time [110]. Our model predicts that PPI networks differ from these social networks in that their diameters grow over time. In addition to Q and D, we tracked the evolutionary trajectories of several other quantities: the evolution of the global clustering coefficient, the rate of signal propagation, the size of the largest connected component (Fig. 19), as well as betweenness and degree values for individual nodes (Fig. 20). These are discussed in more detail in the SI.

3.3 Discussion

The relevance of selection to PPI network evolution has been a topic of considerable debate [111], particularly in the context of higher-order network features, such as modularity. A number of authors have argued that specific selection programs are required to generate modular networks, such as oscillation between different evolutionary goals [112, 113, 114, 115, 116, 117]. However, previous work has shown that gene duplication by itself, in the absence of both natural selection and neofunctionalization, can generate modular networks [118, 119].[10] Unfortunately, duplication-only models err in their predictions of other network properties (Tables 4 and 6). A well-known problem with duplication models is that they generate excessively fragmented networks, with only about 20% of the proteins in the largest component. This is in sharp contrast to real PPI networks, which have 73% to 89% of their proteins in the largest component. Neofunctionalization-only models have most of their proteins in the largest component, but are significantly less modular than real networks. As shown in Table 4, by modeling duplication and neofunctionalization simultaneously, the DUNE model generates networks which have the modularity found in duplication-only models, while retaining most proteins in the largest component. This lends support to the idea that gene duplication contributes to the modularity found in real biological networks, and that protein modules can arise under neutral evolution, without requiring complicated assumptions about selective pressures. This is consistent

[10] Consistent with the findings of [118, 119], modularity in our model is primarily generated by gene duplications (Fig. 16; see SI for sensitivity analysis).

with recent experiments, which characterized a real-world fitness landscape, showing that it is primarily shaped by neutral evolution [120].

Previous estimates of NE rates in eukaryotes have varied widely, generally falling in the range of 100 to 1000 changes/genome/Myr [62, 2, 82], or on the order of 0.1 change/gene/Myr. However, more recent empirical work has identified several problems with the methods used to obtain these estimates, suggesting that *de novo* link creation is much less common than previously thought [84]. This is consistent with our model. The best-fit values of our NE rate μ are in the range of 10^{-5} to 10^{-4}/gene/Myr (Table 7), which in all three organisms are considerably slower than the duplication rates d.

Biologically, many of the interactions created by our neofunctionalization mechanism are expected to initially be weak, non-functional interactions. The results of [121] suggest that strong functional interactions are correlated with hydrophobicity, which in turn is correlated with promiscuity. We posit that initially weak, non-functional interactions are an essential feature of PPI evolution, as they provide the 'raw material' for the subsequent evolution of functional interactions. If this reasoning is correct, one consequence should be that hub proteins are, on average, more important to the cell than non-hub proteins. This has been found to be true: both degree [92] and betweenness centrality [103] have positive correlations with essentiality, indicating that hub proteins are often critical to the cell's survival.

In addition to the factors discussed here, there is an essentially limitless list of biological factors which could potentially be relevant for PPI network evolution: protein copy numbers, alternative splicing, post-translational modifications, protein stability, noise in gene expression levels, chromosomal organization, and many others. Our model does not attempt to address all these issues. Rather, we have proposed a *minimal model*. Minimal models, which have been widely used in physics and biophysics, are based on an idea best expressed by Mark Kac, who said that: "models are, for the most part, caricatures of reality, but if they are good, they portray some of the features of the real world... (T)he main role of models is not so much to explain and to predict ... as to polarize thinking and to pose sharp questions" [122]. In particular, in the present work, we hypothesize a mechanism of assimilation. We expect this mechanism to exist, from both geometric and biological considerations, but, at the present time, there is no direct evidence of it. Its value here is that it generates more of the known data on PPI nets than other models with which we have compared. Because cellular evolution is 'data poor', we believe there can be considerable value in theoretical modeling that postulates mechanisms in advance of experimental data, provided it leads to testable hypotheses.

We have described here a model for how eukaryotic protein networks evolved. The model, called DUNE, assumes two well-known biological mechanisms: (1) gene duplications, leading to a superfluous copy of a protein that can change rapidly under new selective pressures, giving new relationships with other proteins and (2) a protein can undergo random mutations, leading to neofunctionalization, the *de novo* creation of new relationships with other proteins. We assume these changes are otherwise random. The model shows good agreement with 10 topological properties in yeast, fruit flies, and humans. One finding is that PPI networks can evolve modular structures, just from these random forces, in the absence of specific selection pressures. We also find that the most central proteins also tend to be the oldest. This suggests that looking at the structures of present-day protein networks can give insight into their evolutionary history.

3.4 Methods

Genome-wide PPI screens have a high level of noise [57], and specific interactions correlate poorly between data sets [88]. We found that several large-scale features differed substantially between types of large-scale experiments (see SI). Due to concerns about the accuracy and precision of data obtained through large-scale screens, we chose to work with 'high-confidence' data sets consisting only of pairwise interactions confirmed in small-scale experiments, which we downloaded from the public HitPredict database [123]. We found sufficient high-confidence data in yeast (*S. cerevisiae*), fruit flies (*D. melanogaster*), and humans (*H. sapiens*).

All simulations and network feature calculations were carried out in Matlab. Our scripts are freely available for download at `http://ppi.tinybike.net`. We computed betweenness centralities, clustering coefficients, shortest paths, and component sizes using the MatlabBGL package. Modularity values were calculated with the algorithm of [124]. All comparisons (except the degree distribution) are between the largest connected components of the simulated and experimental data.

Due to the human network's somewhat larger size, most dynamical features were calculated once per 50 time steps for the human network, but were updated at every time step in the yeast and fly networks. For dynamical plots, the y coordinates of the trend line are medians-of-medians. The amount of time elapsed per time step (the x coordinate) varies between simulations. We binned the time coordinates to the nearest 10 million years for yeast and fly, and 25 million years for human. When multiple values from the same simulation fell within the same bin, we used the median value. We then calculated the median value between simulations. Scatter plot trend lines are calculated in a similar way. The trend line represents the median response variable (C, b, or ℓ) value over all nodes within a single simulation with degree k. The y coordinate of the trend line is therefore the median (across 50 simulations) of these median response variables. This median-of-medians includes all simulations that have nodes of a given degree.

3.5 Supporting Information

3.5.1 Subfunctionalization

One model for the fate of duplicate genes is *subfunctionalization* (SF). In SF, the original and the duplicate genes are both free to lose their redundant functions, so they can evolve freely until they exactly reproduce the ancestral function [143]. The post-duplication divergence in our model is similar in spirit to SF, but it differs in two significant ways: (1) in our model, the link loss is completely asymmetric, and (2) a fraction $(1-\phi)$ of the redundant links are retained, so, unlike SF, not all of the redundancy is eliminated. For the first point, empirical evidence suggests that the divergence is asymmetric [68], although the assumption of *complete* asymmetry would likely need to be revisited to build a finer-grained model. Second, genetic regulatory networks have been shown to be robust to random link deletions, indicating that these networks retain some degree of redundancy [144, 145, 146]. *In silico* evidence suggests that a more accurate picture may be of a transient period of functional divergence, followed by prolonged neofunctionalization, resulting in only a partial loss of redundancy [147]. This is consistent with our model.

3.5.2 Large-scale duplications

Our model does not explicitly consider simultaneous duplication of multiple genes (chromosomal duplications, whole genome duplications, etc.). However, as shown in Table 7, duplication rates in our model are considerably higher than neofunctionalization rates, so that, on average, there are multiple duplications per neofunctionalization event. Sequential duplications of this type may be thought of as an (imperfect) representation of multi-gene duplications. The advantage of this approach is that we do not require separate rates for each duplication scale (one could imagine an extremely detailed model which included separate rates for gene duplication, gene-pair duplication, gene-triplet duplication, etc.). The downside is that our implementation of larger-scale duplications will generally include some genes which have been duplicated multiple times, and others which have not been duplicated at all. A potential mitigating factor is that the rate of gene loss (and evolution in general) following genome duplication is very high [93, 67], so even a completely faithful large-scale duplication would likely be altered within short order.

3.5.3 Randomness conjecture

We describe here a test of our randomness conjecture, $q = 1/N$. The implication is that the average number of NE links per protein should be independent of N. Another possibility is that q is constant, implying that the number of NE links per protein is proportional to N. To test this, we compared NE rates in the human and yeast PPI networks. The total number of proteins in humans is estimated to be 22740 [148] and in yeast, 5616 [149]. The number of mutations to coding DNA is approximately 0.004/genome/replication in humans and 0.0027/genome/replication in yeast [150]. If the number of new links is proportional to N, then, based on the number of proteins and the mutation rate, there should be roughly 600% more links created by NE in humans than in yeast. However, by counting the number of nonredundant interactions in duplicate gene pairs, it has been shown empirically that the average number of links created by NE per protein is only about 8% higher in humans than in yeast [87]. These results support the conjecture that the probability for a protein to receive a new link via point mutation is approximately independent of N, as previously noted in [82]. Due to the finite copy number of proteins, as well as the compartmentalization of eukaryotic cells, we regard it as unlikely that proteins will simultaneously acquire multiple links to targets in different locations in the cell, or which are involved in divergent biological processes.

3.5.4 Assimilation is driven by single-interface proteins

Closely related to the randomness conjecture is the number of ‘extra’ links created by each assimilation event, which should also independent of N. Proteins with multiple interaction partners may interact with their partners simultaneously (so-called ‘party hubs’) or at different times/locations (‘date hubs’) [151]. Since they bind to several partners simultaneously, party hubs typically have multiple binding sites, each specific to one binding partner [152]. This specificity suggests that a protein which evolved the capability to bind to a party hub would be unlikely to undergo assimilation. By contrast, the binding sites of date hubs are often disordered regions which are able to form transient interactions with multiple partners [153, 154]. If a protein evolves the capability to bind to a date hub, it is likely to share the physical characteristics of the hub’s neighbors, leading to assimilation. However, to avoid competition for the same binding site, the interaction partners of date hubs tend not to be coexpressed [152]. One consequence of this is that assimilating proteins will likely only bind one of the target protein’s neighbors – whichever neighbor happens to be present at that time and place. Although the capability to bind to the hub protein’s other neighbors may initially be present, these will presumably remain unused in the cell. Our expectation is that the assimilating protein will therefore be unlikely to retain this capability, as it evolves. Similarly, only a single extra link should be generated at the second-neighbor level, third-neighbor level, etc. Consistent with the evidence discussed above, the number of links created by assimilation is approximately independent of the total network size. Party hubs typically are centrally-located within modules, while date hubs often function to stitch together large-scale modules in the cell. It may be that duplication-only models are unrealistically fragmented (Table 4) because their modules are not properly attached with date hubs; instead, the modules are disconnected components.

3.5.5 Domain shuffling and assimilation

One example of a known biological mechanism which should lead to assimilation is domain shuffling, the copy-and-pasting of part of one protein into another [155, 156]. The neofunctionalization mechanism described here is quite general, and includes domain shuffling, among other methods of PPI creation. A PPI formed via domain shuffling will often be the result of a binding site duplication. Assuming the binding is due to simple surface similarity, the initial link will be to the protein which had its domain copied. The likelihood of binding to neighbors of the original protein should depend only on the probability that each interaction is due to surface similarity because the copied binding site will be identical to the original.

The role of domain shuffling in assimilation raises the question of whether domains should be modeled explicitly, rather than representing proteins as integral units. Previous work indicates that overall PPI network topology is robust to the details of domain shuffling [157]. Moreover, while proteins which have experienced domain shuffling have a higher average degree than other proteins, high- and low-degree proteins are equally likely to acquire new interactions this way [158]. Because the creation of new links by domain shuffling should be topologically very similar to the creation of new links by other neofunctionalization events, we believe our model is a reasonable implementation of this mechanism, as it applies to the evolution of network topology.

3.5.6 Network rewiring

Some higher-order features of the network are simply a result of its degree sequence, and other features might be important in their own right. As discussed in [99], it is possible to isolate the effects of the degree sequence by 'rewiring' (detaching then reattaching links) the network at random, subject to the restriction that the degree sequence must be preserved. If a property contains extra information about the network's structure, then it should be different in the rewired network. On the other hand, if the network is rewired many times, and the property is always the same, then it is likely to just be a result of the degree sequence. We used a script downloaded from `http://www.cmth.bnl.gov/∼maslov/matlab.htm` to randomly rewire the empirical network $4K_{\text{data}}$ times. As expected, modularity is decreased by random rewiring. Upon rewiring, we find $Q = 0.603 \pm 0.002$ in humans, $Q = 0.590 \pm 0.003$ in yeast, and $Q = 0.722 \pm 0.007$ in flies (median ± standard deviation from 50 repeats of the rewiring algorithm). Rewiring also shrinks the diameters of PPI networks to $D = 13 \pm 0.9$ in humans, $D = 12 \pm 1.0$ in yeast, and $D = 15 \pm 1.1$ in flies. These results suggest that these features contain important structural information about the network, and are not merely consequences of the degree sequence.

One reason we are interested in calculating Q and D is simply to check that the values are comparable between the simulated and experimental networks. However, on a more qualitative level, we would also like to have some idea of what the threshold is for a network to be considered 'modular' or 'small-diameter'. The rewired Q and D values are useful because these features are dependent on the size of the network (number of nodes N and links K); given an identical network construction method, Q and D will generally be different in sparse versus dense networks. We use these Q and D values as baseline values with which the experimental and simulated networks can be compared; we considered Q and D values differing from the rewired values by more than a standard deviation to be significantly different.

3.5.7 Eigenvalues

The connectivity of a network can be expressed by its *adjacency matrix*, an $N \times N$ matrix $\mathbf{A}$, in which the entries A_{ij} equal 1 if a link exists between proteins i and j, and 0 otherwise. If $\mathbf{A}$ is normalized by column, then the entries describe the rates of a transition from i to j in one time step. The distribution of eigenvalues $p(\lambda)$ is called

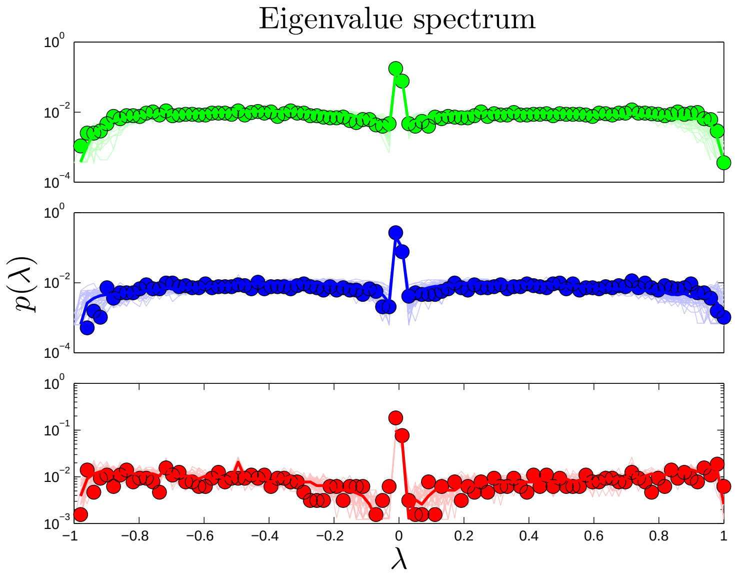


Figure 11: Eigenvalue (λ) distributions in human (green), yeast (blue), and fly (red). Heavy lines are the median values from 50 simulations, and light lines are results of individual simulations.

the network's *spectrum* (Fig. 11). This matrix and its eigenvalues can be interpreted in terms of a process in which a random walker starts on one node i, and, over a series of time steps, reaches another node j. Intuitively, this can be thought of as a signal propagation rate: if one protein is affected by an external signal, how long does it take that signal to diffuse through the network? The eigenvalues λ of this 'walk matrix' describe the rate at which a random walk on the network reaches steady-state. The second-largest eigenvalue (λ_2) determines the rate of convergence of the random walk (Fig. 19). A larger value of λ_2 indicates a slower signal propagation rate.

3.5.8 Error tolerance

We measured the 'error tolerance' as described in [66]: we examined the decrease of f_1 when nodes (and their accompanying edges) were deleted from the network either (1) at random, or (2) according to their degree, starting with the most well-connected node. Results for the simulated and experimental networks were very similar (Fig. 12).

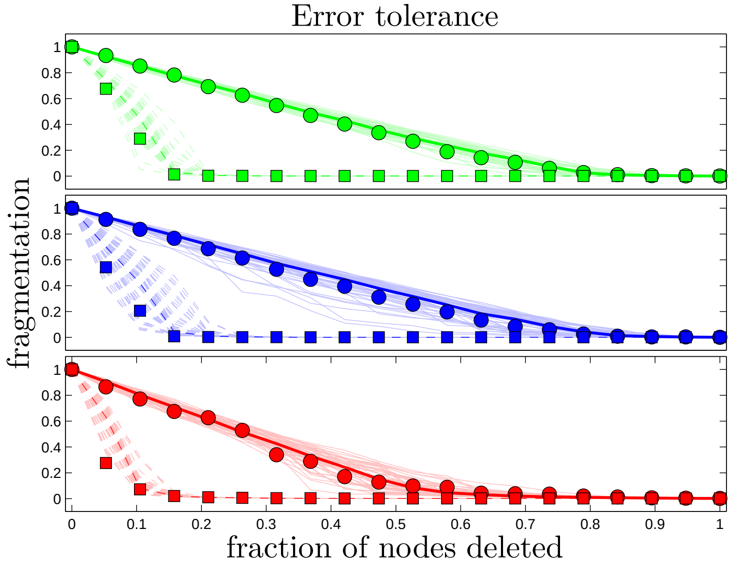


Figure 12: Error tolerance in human (green), yeast (blue), and fly (red). Circles indicate proteins deleted randomly, and squares indicate proteins deleted starting with the most well-connected protein and removing proteins in descending order.

3.5.9 Simulation length: early versus late evolution

The total time elapsed during our simulations varies considerably, with yeast and human simulations running about 1 to 2 billion years, and the fly simulations about 5 billion years. This is compared to the rough estimate of 3.5 billion years since the origin of life on Earth [159]. The exceptionally long duration of the fly simulations are due to the very low gene duplication rate ($d = 0.001$/gene/Myr). The aim of our model is to describe the evolution of PPI networks with all their present-day machinery. Gene duplication, in the form in which it exists today, certainly would not have existed at the origin of life! The initial state in our model consists of two interacting proteins. Biologically, these are two polypeptides (or, more likely, RNA molecules) in a pre-biotic soup, that happen to interact in a way that is mutually beneficial. Each of these molecules has the ability to replicate. This autonomous replication of individual proteins corresponds to 'gene duplication' in the very early stages of evolution. However, this is a very different conceptual underpinning for the duplication mechanism, and it seems unlikely to share the present-day values of the duplication rate. Because, in the early stages of evolution, each time step represents a very long duration in real time, it is likely that this accounts for the discrepancy in total time elapsed.

3.5.10 Empirical data

We downloaded large-scale data sets from BioGRID [160], and used the Wilcoxon rank-sum test to compare aggregate statistical features across various experimental types in yeast (*S. cerevisiae*) and humans (*H. sapiens*) [151]. As expected, we found that data obtained by affinity capture was significantly different than pairwise experimental data (primarily yeast two-hybrid and *in vitro* complexation), as the affinity capture interactions represent entire complexes, which is somewhat different information than the pairwise interactions we are attempting to capture using our model. However, more surprisingly, the only feature to show significant agreement between pair-wise techniques was the eigenvalue distribution of the walk matrix ($P > 0.05$). Further sub-dividing the individual techniques into smaller data sets containing only results obtained in single experiments, we discovered that, again, only the spectra agreed between different screens.

Note that, due to the small size of the fly network, there may be too many missing links to obtain an accurate description the network's large-scale topology. Although, by appropriate parameter tuning, our model is able to accurately reproduce the fly network, it is possible that different parameters will be required to match the fly network once it becomes more fully characterized experimentally. The data sets considered here do not include interactions which are enabled through post-translational modifications. Although these data sets are far from complete, and may be susceptible to false-positive detections, these appear to be the most accurate data available at the present time.

	γ	β	ξ	α	δ
Yeast	2.8(2)	2.4(1)	1.8(2)	1.2(2)	0.32(6)
Fly	3.1(2)	2.2(4)	0.8(5)	0.8(7)	0.0(3)
Human	2.8(1)	2.3(1)	1.5(3)	1.3(2)	0.0(1)

Table 5: Scaling exponents. Distributional exponents ($p(k) \sim k^{-\gamma}$, $p(b) \sim b^{-\beta}$) were estimated using the maximum likelihood method of [14]. Other exponents ($\widetilde{C} \sim k^{-\xi}$, $\widetilde{b} \sim k^{\alpha}$, $\widetilde{n} \sim k^{-\delta}$) were estimated using nonlinear regression. Note that, due to the relatively small sizes of the data sets, there is considerable uncertainty in these estimates.

3.5.11 Fitting functions

The degree distribution obeys a power law in its tail, $p(k) \sim k^{-\gamma}$ [14], with $\gamma \approx 3$ (Table 5), implying that hub proteins are more common than would be expected for a randomly connected network, which would have an exponentially decaying $p(k)$. The closeness distribution $p(\ell)$ is approximately Gaussian, with mean 0.17 and standard deviation 0.03 in humans, mean 0.19 and standard deviation 0.03 in yeast, and mean 0.13 and standard deviation 0.03 in flies. Closeness is a measure of distance, indicating that the distances within the network are essentially a random walk in 'node space'. The betweenness distribution also follows a power law in its tail. This is an indication of modular structure, due to the overrepresentation of 'bridge' proteins, relative to a randomly connected network.

All species examined show a power law decay in median clustering coefficient as a function of degree, $\widetilde{C} \sim k^{-\xi}$. Poorly-connected proteins therefore tend to have *higher* clustering coefficients, meaning that a greater fraction of their neighbors are mutually connected.

Disassortative mixing was quantified for the yeast PPI network in [99] as a power law *decrease* in median neighbor degree, $\widetilde{n} \sim k^{-\delta}$. This is consistent with our data, although the very small estimated value of $\delta = 0.32$ indicates only a slight negative relation (Table 5). Interestingly, $\delta = 0$ in both human and fly networks, indicating that disassortativity may be a trait unique to the yeast network.

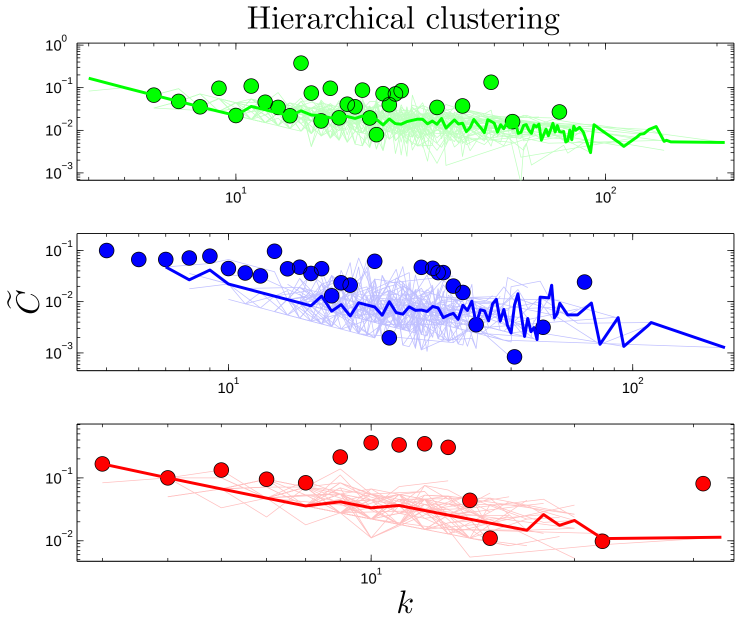


Figure 13: Median clustering coefficient vs. degree in human (green), yeast (blue), and fly (red). Heavy lines are the median values from 50 simulations, and light lines are results of individual simulations.

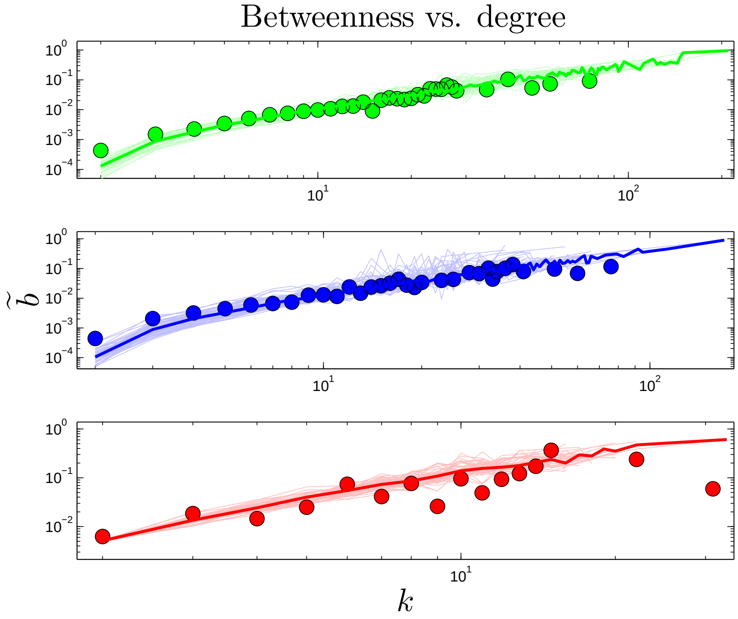


Figure 14: Median betweenness vs. degree in human (green), yeast (blue), and fly (red).

3.5.12 Principal component analysis

We examined six features which calculate a value for each node in the network: degree centrality, clustering coefficients, closeness centrality, eigenvalue spectrum, betweenness centrality, and mean nearest-neighbor degree. To quantify the independence of these features, we used principal component analysis (PCA) [161]. Each feature assigns a value to each node in the network, giving a $6 \times N$ data matrix, where each row represents a feature (signal), and each column is a node (sample). We subtract the mean and divide by the standard deviation of each row. This results in a standardized data matrix, denoted by $\mathbf{Y}$. The 6×6 correlation matrix for each species is defined as $\mathbf{C} \equiv \frac{1}{N-1}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$:

$$\mathbf{C}_{\mathrm{h}} = \begin{bmatrix} 1 & 0.10 & 0.87 & 0.56 & 0.02 & -0.12 \\ 0.10 & 1 & -0.04 & 0.09 & 0.01 & 0.03 \\ 0.87 & -0.04 & 1 & 0.44 & -0.01 & -0.08 \\ 0.56 & 0.09 & 0.44 & 1 & 0.00 & 0.34 \\ 0.02 & 0.01 & -0.01 & 0.00 & 1 & -0.02 \\ -0.12 & 0.03 & -0.08 & 0.34 & -0.02 & 1 \end{bmatrix}, \tag{17}$$

$$\mathbf{C}_{\mathrm{y}} = \begin{bmatrix} 1 & 0.03 & 0.91 & 0.43 & -0.02 & -0.21 \\ 0.03 & 1 & -0.06 & 0.04 & 0.01 & 0.03 \\ 0.91 & -0.06 & 1 & 0.39 & -0.01 & -0.14 \\ 0.43 & 0.04 & 0.39 & 1 & -0.03 & 0.34 \\ -0.02 & 0.01 & -0.01 & -0.03 & 1 & -0.04 \\ -0.21 & 0.03 & -0.14 & 0.34 & -0.04 & 1 \end{bmatrix}, \tag{18}$$

$$\mathbf{C}_{\mathrm{f}} = \begin{bmatrix} 1 & 0.15 & 0.62 & 0.36 & -0.11 & -0.15 \\ 0.15 & 1 & -0.08 & 0.06 & -0.10 & -0.02 \\ 0.62 & -0.08 & 1 & 0.40 & 0.02 & -0.16 \\ 0.36 & 0.06 & 0.40 & 1 & 0.00 & 0.30 \\ -0.11 & -0.10 & 0.02 & 0.00 & 1 & -0.05 \\ -0.15 & -0.02 & -0.16 & 0.30 & -0.05 & 1 \end{bmatrix}. \tag{19}$$

The entries of each $\mathbf{C}$ are (from left-to-right, and top-to-bottom): degree centrality, clustering coefficients, betweenness centrality, closeness centrality, eigenvalue spectrum, and mean nearest-neighbor degree. Many of the off-diagonal elements of the $\mathbf{C}$ matrices are close to zero, suggesting that the features are to a large extent independent of one another.

To perform PCA, we diagonalized each correlation matrix,

$$\mathbf{C} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^{\mathrm{T}}, \tag{20}$$

where $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues and $\mathbf{S}$ has the eigenvectors of $\mathbf{C}$ as its columns. As shown in Fig. 15, the degree and betweenness show similar loadings on the first two principal components, reflecting the nearly linear relation between these centrality scores (Fig. 14).

The eigenvalue matrices $\mathbf{\Lambda}$ are given by:

$$\mathbf{\Lambda}_{\mathrm{h}} = \begin{bmatrix} 2.27 & & & & & \\ & 1.23 & & & & \\ & & 1.02 & & & \\ & & & 0.98 & & \\ & & & & 0.40 & \\ & & & & & 0.11 \end{bmatrix}, \tag{21}$$

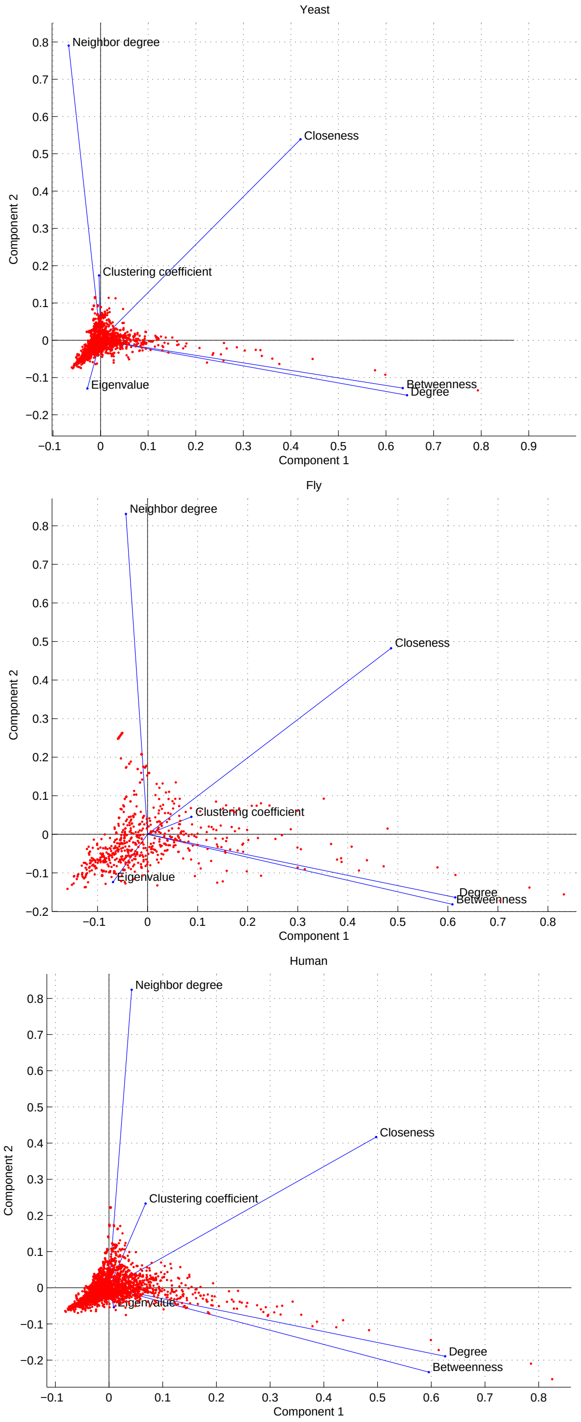


Figure 15: Principal component analysis. Shown are the factor loadings and scores on the first two principal components. Data scores are shown in red, and blue lines represent feature loadings.

$$\mathbf{\Lambda}_{\mathrm{y}} = \begin{bmatrix} 2.20 & & & & & \\ & 1.30 & & & & \\ & & 1.01 & & & \\ & & & 0.98 & & \\ & & & & 0.43 & \\ & & & & & 0.08 \end{bmatrix}, \tag{22}$$

$$\mathbf{\Lambda}_{\mathrm{f}} = \begin{bmatrix} 1.95 & & & & & \\ & 1.24 & & & & \\ & & 1.13 & & & \\ & & & 0.91 & & \\ & & & & 0.44 & \\ & & & & & 0.32 \end{bmatrix}. \tag{23}$$

(Zeros have been suppressed for clarity.) The fraction of variance explained by the ith principal component is given by $\Lambda_{ii}/\sum_j \Lambda_{jj}$. The closer the number of components required to explain most of the variance is to the total number of input signals, the more independent the signals are. In yeast and humans, 4 components are required to explain 90% of the variance; in fruit flies, it requires 5 components. Linear transformations are able to only modestly reduce the dimensionality of the problem, suggesting that each feature contributes unique information about the network's structure. This does not, of course, rule out the possibility of the existence of other independent, informative features, a far more complicated question which is outside the scope of this current work.

3.5.13 Sensitivity analysis

The DUNE model has four parameters. One parameter, the DU rate d, is estimated from empirical data. The other three are adjustable parameters: the NE rate μ,

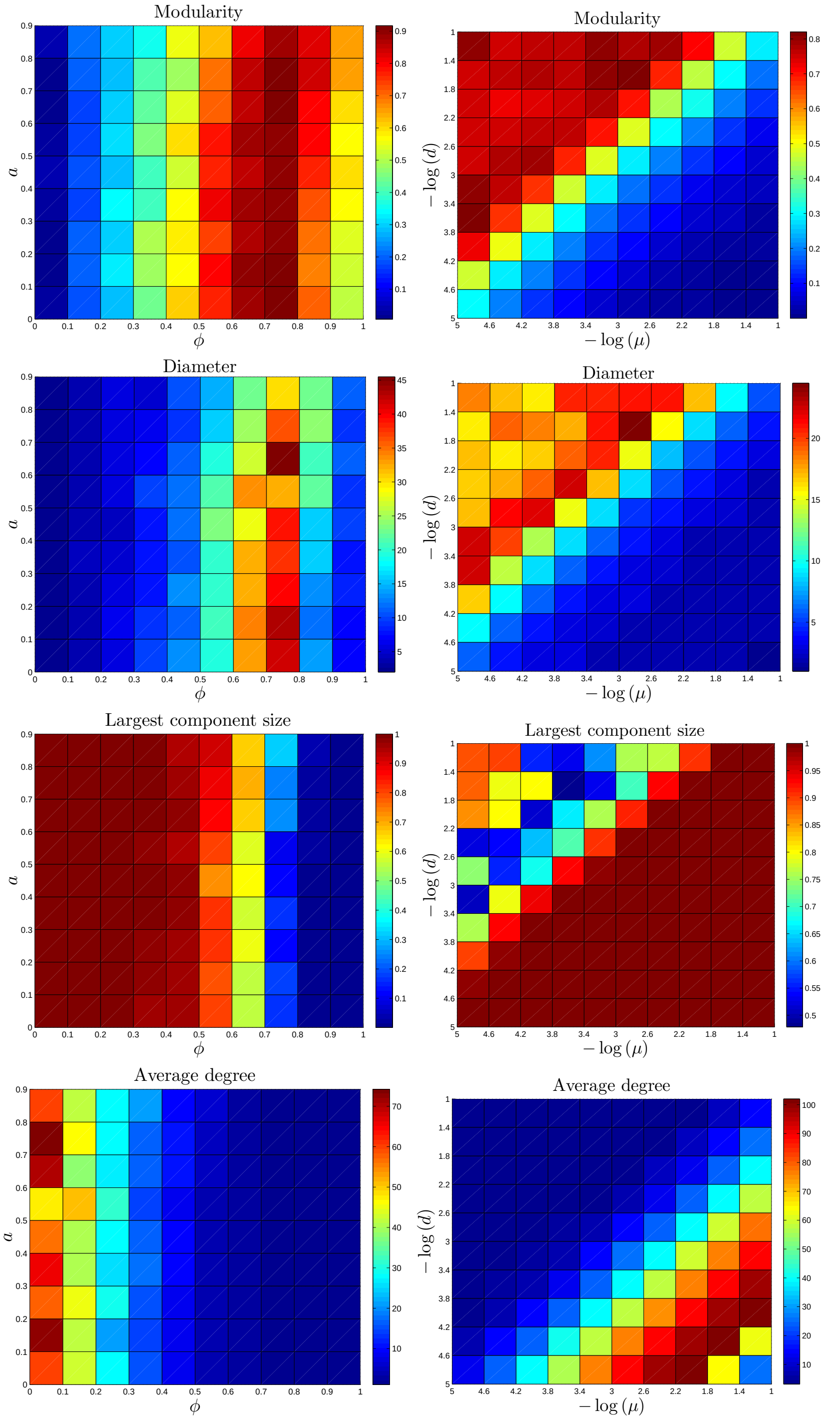


Figure 16: Sensitivity analysis. Heat maps represent median values for 10 simulations per parameter combination of the yeast network. Left: ϕ and a are varied, d and μ values are kept fixed. Right: d and μ varied, ϕ and a kept fixed.

the divergence probability ϕ, and the assimilation probability a. To gain a better understanding of how these parameters affect the final network structure, starting with each organism's set of parameters, we systematically adjusted all 4 parameters. Results are shown in Fig. 16.

As expected, the divergence parameter ϕ was positively correlated with both the modularity Q and the diameter. When $\phi \approx 0$, the network quickly reaches a fully-connected state, where all proteins are linked to all other proteins. Consequently, the network is not organized into pathways, and all distances in the network are equal to 1. The 'thinning out' of duplicate links is therefore essential to generate non-trivial network features. The opposite occurs when the NE rate μ is too low: $f_1 \approx 0$, and the network evolves towards a completely disconnected state.

Why does gene duplication lead to modularity? Consider an initially uniform network. When a node is duplicated at random, this causes the original node, the copy, and their immediate neighbors to share more links internally than they do with the rest of the network. Subsequent duplications amplify this effect: if a node that has 10 links within a module but only 2 external links is duplicated, there are now 20 internal and 4 external links (prior to post-duplication divergence).

Interestingly, Q has a weak negative correlation with the assimilation parameter a. When a is large, the probability to link to distant neighbors of the target protein is relatively high, and mutated proteins have a non-negligible chance to generate links to proteins outside of their target's pathway. This causes pathways to blur at the edges; their member proteins will share a higher number of links to other pathways than for a low a network. Although the modularity is reduced for a high a network, it does not disappear entirely. Similarly, there is a sharp decrease in Q as the NE rate surpasses the DU rate, indicating the important role of the DU mechanism in modular organization.

Diameter is also negatively correlated with a. When a single NE event has a significant chance to generate links to the target protein's neighbors, this tends to reduce the overall separation of proteins in the network.

3.5.14 Comparison to other models

Our model is rooted in previous modeling efforts. The basic framework for our model combines the gene duplication mechanism described in [66] with a link creation mechanism inspired by [2]. The principal difference between our model and previous models is that our model considers duplication and mutation simultaneously. All previous models attempted to construct the PPI network from a single mechanism. Another significant difference is the existence of the assimilation mechanism. To the best of our knowledge, previous work has not explicitly modeled proteins integrating into biological pathways.

We compare the DUNE model to four models previously proposed for PPI networks. Two were evolutionary models: (1) the Vázquez model of DU followed by rapid loss-of-function mutations [66];[11] and (2) the Berg 'link dynamics' model of point mutations coupled with a PA-like 'rich-get-richer' rule for assigning new interactions [2]. Two others were static models (models of present-day networks that do not simulate the network's evolutionary path) that consider the primary organizing principle to be nonspecific interactions between proteins: (1) random geometric (RG), a mathematical model where proteins are randomly scattered in a 2 to 4 dimensional box, and any proteins close enough to one another form an interaction [89]; and (2) the 'MpK' desolvation model, which assigns interactions based on proteins' exposed hydrophobic surface areas [88]. For reference, we also calculated results for an Erdős-Rényi (ER) random graph with N and $\langle k \rangle$ set by the data [90].

These models were originally validated against different features of the empirical

[11] A slightly different DU model is presented by Pastor-Satorras [162]. However, because the Vázquez model has been shown to be a better fit to experimental data [163], we have limited our DU-only comparison here to the Vázquez model.

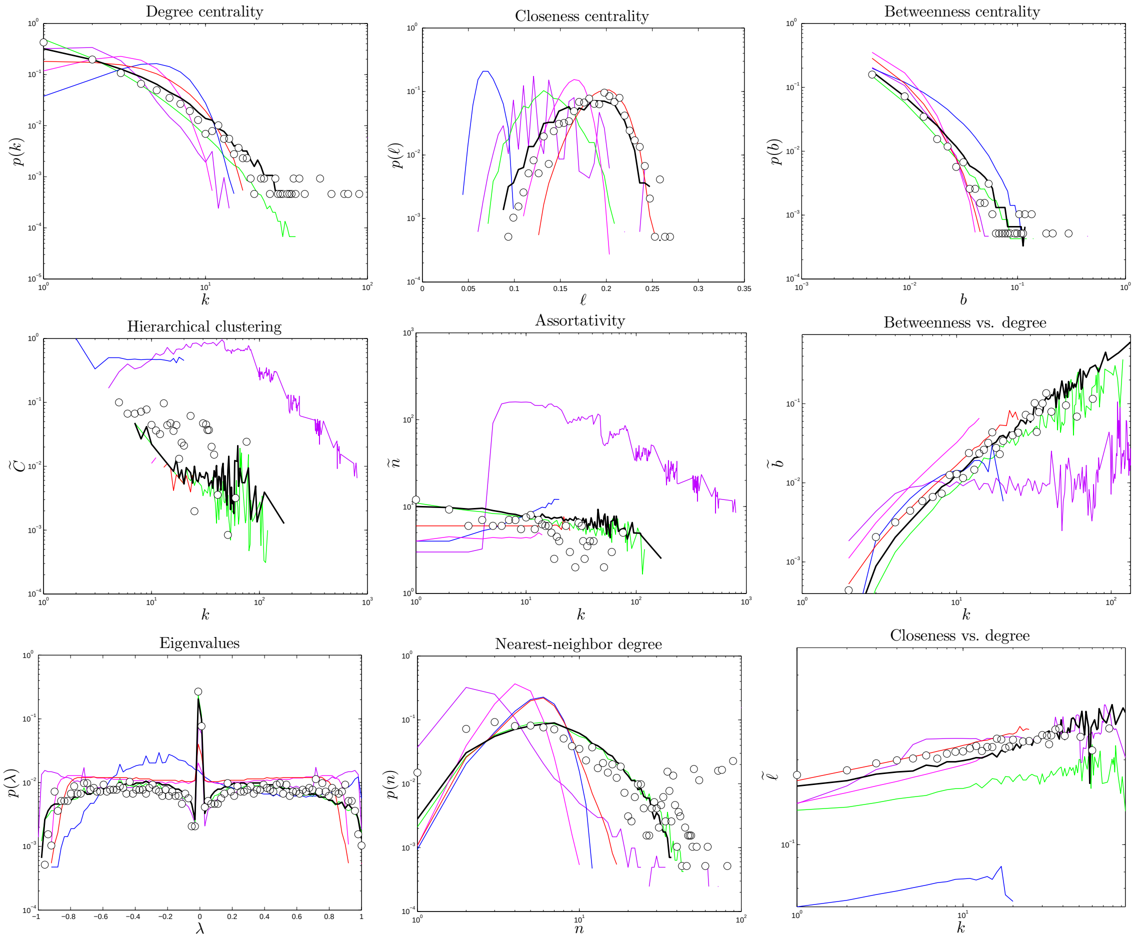


Figure 17: Comparison of five other models to the yeast PPI network: Vázquez [66] (green), Berg [2] (red), random geometric [89] (dark blue), MpK desolvation [88] (purple), and ER random graph [90] (brown). For reference, DUNE model results are shown as a black line. Dots represent high-confidence experimental yeast data, and solid lines are median values over 50 simulations.

network, making it difficult to directly compare them. To characterize these models in greater detail, we coded each of these models, and ran 50 simulations of each model with identical parameters and starting conditions. Using Matlab, we coded the Vázquez [66], Berg [2], RG [89], and MpK [88] models as described in the original papers. Since each model was originally parametrized for older yeast PPI data sets, we re-optimized the parameters for our yeast data as follows. We used a Monte Carlo simulation to adjust each model's parameters to minimize the total symmetric mean absolute percentage error values (SMAPE; see below) for the yeast HitPredict data set.

For the Vázquez model, we used a value of 0.582 for the post-duplication divergence probability, and a value of 0.083 for the dimerization probability. As noted by previous authors, duplication-only simulations produce networks which are extremely fragmented [118]. We observed that the Vazquez simulations typically had around 20% of their nodes in the largest connected component (Table 4). Since most of the network features we examined are limited to the largest component, in order to make a reasonable comparison of the Vazquez simulation results to the data, we allowed the simulated network to grow until its number of links met or exceeded 5 times the number of links in the data, $K \geq 5K_{\text{data}}$. Since the largest component is not always exactly 20% of the total nodes, this stopping condition is somewhat arbitrary; however, results for this model seem robust to small changes in the stopping condition. For the Berg model, we used our empirically estimated duplication rate of 0.01/gene/Myr, and found best-fit values of 24.5/gene/Myr for the mutation rate, and $N_{\text{data}} - 98$ proteins for the initial network size. For the RG model, we used a $45.5 \times 45.5 \times 45.5$ 'box' with a maximum interaction radius of 3.92. For the MpK model, the number of exposed surface residues was 19, the fraction of exposed hydrophobic residues was $M = 0.230 \pm 0.110$ (mean ± standard deviation), and the best-fit linear equation relating M to the binding threshold was $1.09M + 1.04$.

The Vázquez simulations were initialized with 2 connected nodes, and the simulation was allowed to run until $K \geq K_{\text{data}}$. The Berg simulations were initialized with $N_{\text{data}} - 98$ randomly connected nodes, then run until $N = N_{\text{data}}$. The RG and MpK models (which are not evolutionary models and therefore create the network all at once) were set up as described in the original papers.

To characterize the networks, we computed several network properties:

Single-value modularity Q, diameter D, fraction of nodes in the largest component f_1, global clustering coefficient $\langle C \rangle$, and mean degree in the largest component $\langle k \rangle$ (Table 4)

Distributional degree $p(k)$, betweenness $p(b)$, closeness $p(\ell)$, eigenvalue $p(\lambda)$, and nearest-neighbor degree $p(n)$ distributions

Scatter plot median closeness vs. degree $\widetilde{\ell}(k)$, median clustering coefficient vs. degree $\widetilde{C}(k)$, median betweenness vs. degree $\widetilde{b}(k)$, and median nearest-neighbor degree vs. degree $\widetilde{n}(k)$

and compared these features to those of empirical data from yeast. As shown in Fig. 17, we found that none of the previous models capture the full set of network properties.

To quantify agreement with the data for non-single-value features, we calculated the symmetric mean absolute percentage error (SMAPE) between simulation and experiment [164, 165]:

$$\text{SMAPE} = \frac{1}{Y} \sum_{i}^{Y} \frac{\left| y_i - y_i^{\text{data}} \right|}{y_i + y_i^{\text{data}}}, \tag{24}$$

where Y is the number of data points, and y_i and y_i^{data} denote the ith point of the response variable (Table 6) in the simulated and experimental data, respectively. For the distributional features, Y is the number of bins (arbitrarily chosen to be 100) minus the number of bins in which $y_i + y_i^{\text{data}} = 0$. For non-distributional (scatter plot) features, Y is the number of k values with values for both simulation and experiment. There are many possible measures of accuracy (such as the widely-used root mean squared error); we used SMAPE for two reasons. First, because it relies on absolute value, SMAPE does not over-emphasize the impact of outliers. Second, dividing by $y_i + y_i^{\text{data}}$ ensures that the magnitude of the response variable does not overwhelm the sum. This is significant for the non-distrbutional features. For example, in a plot of median betweenness vs. degree (Fig. 14), we are just as interested in the overlap of the low-betweenness, low-degree region of the curve as we are with the high-betweenness, high-degree region. SMAPE values are collected in Table 6. The total SMAPE values shown in Table 6 indicate that, although previous models can accurately reproduce certain features of the PPI network, only the DUNE model provides a reasonable across-the-board fit.

	$p(k)$	$p(\ell)$	$p(b)$	$\widetilde{C}(k)$	$\widetilde{n}(k)$	$\widetilde{b}(k)$	$p(\lambda)$	E.T.	E.T. (k)	$p(n)$	$\widetilde{\ell}(k)$	Total
DUNE	**0.4686**	**0.3728**	**0.3605**	0.6034	0.2889	0.2432	0.1371	0.1028	0.1955	0.5630	0.1434	**3.4791**
Vázquez	0.5252	0.8139	0.3893	0.6254	0.2714	0.3456	**0.1233**	0.1006	**0.1504**	**0.5032**	0.2604	4.1088
Berg	0.6984	0.4500	0.7182	0.6668	**0.1461**	**0.1870**	0.2949	0.1230	0.3011	0.8557	**0.0658**	4.5070
RG	0.8121	0.9790	0.6087	0.8330	0.2172	0.2262	0.3291	0.3128	0.4477	0.8907	0.4927	6.1492
MpK	0.8170	0.7917	0.6809	0.8908	0.8818	0.5798	0.1800	0.1855	0.1846	0.8225	0.1650	6.1796
ER	0.8292	0.7027	0.7785	**0.5273**	0.2211	0.4095	0.2771	**0.0716**	0.3201	0.9298	0.0681	5.1350

Table 6: Symmetric mean absolute percentage error (SMAPE) of simulation versus experiment in yeast (Eq. 24). ‘E.T.’ is the error tolerance curve with random protein removal, and ‘E.T. (k)’ is the error tolerance curve with highest-degree proteins removed first. ‘DUNE’ is the model described here, ‘Vázquez’ is the DU-only model of [66], ‘Berg’ is the link dynamics model [2], ‘RG’ is random geometric [89], ‘MpK’ is the physical desolvation model presented in [88], and ‘ER’ is an Erdős-Rényi random graph [90]. For each comparison, the lowest value is shown in bold.

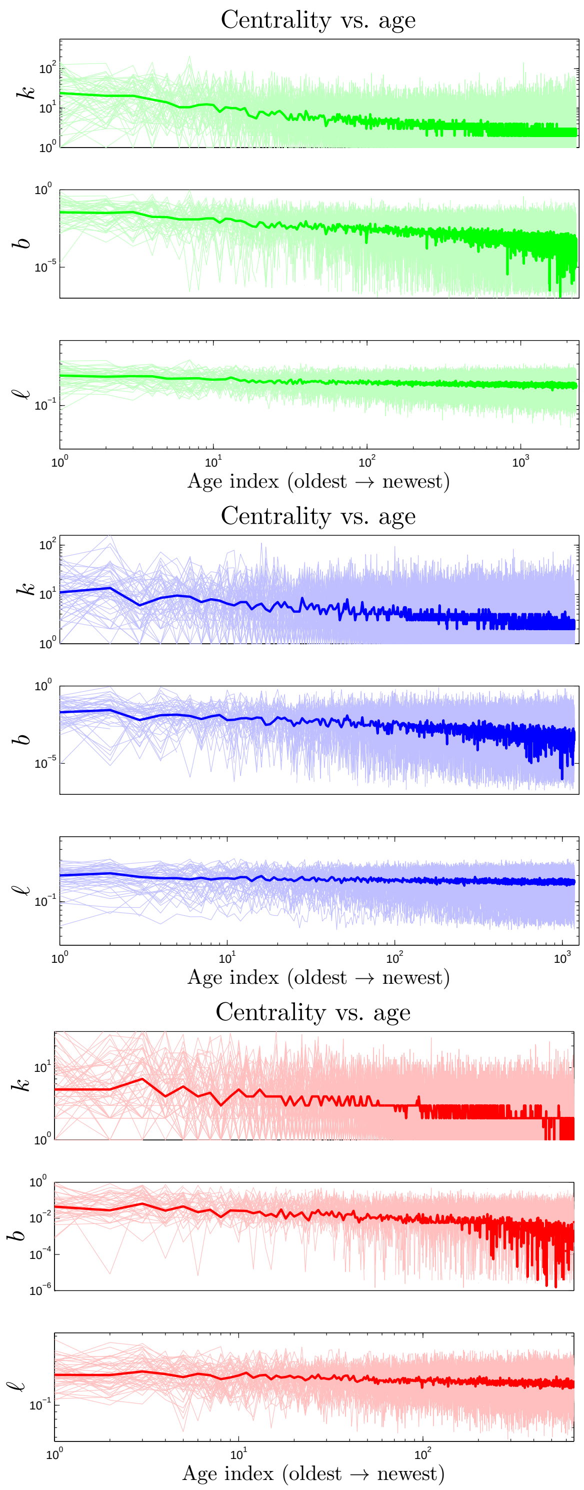


Figure 18: Older proteins are more central in PPI networks. Simulations of a protein's age index (time since introduction into the network) vs. degree (k), betweenness (b), and closeness (ℓ) centrality, for human (green), yeast (blue), and fly (red). The oldest proteins are on the *left* in this figure, and the proteins get younger moving to the right. There is an approximately monotonic increase in centrality with age.

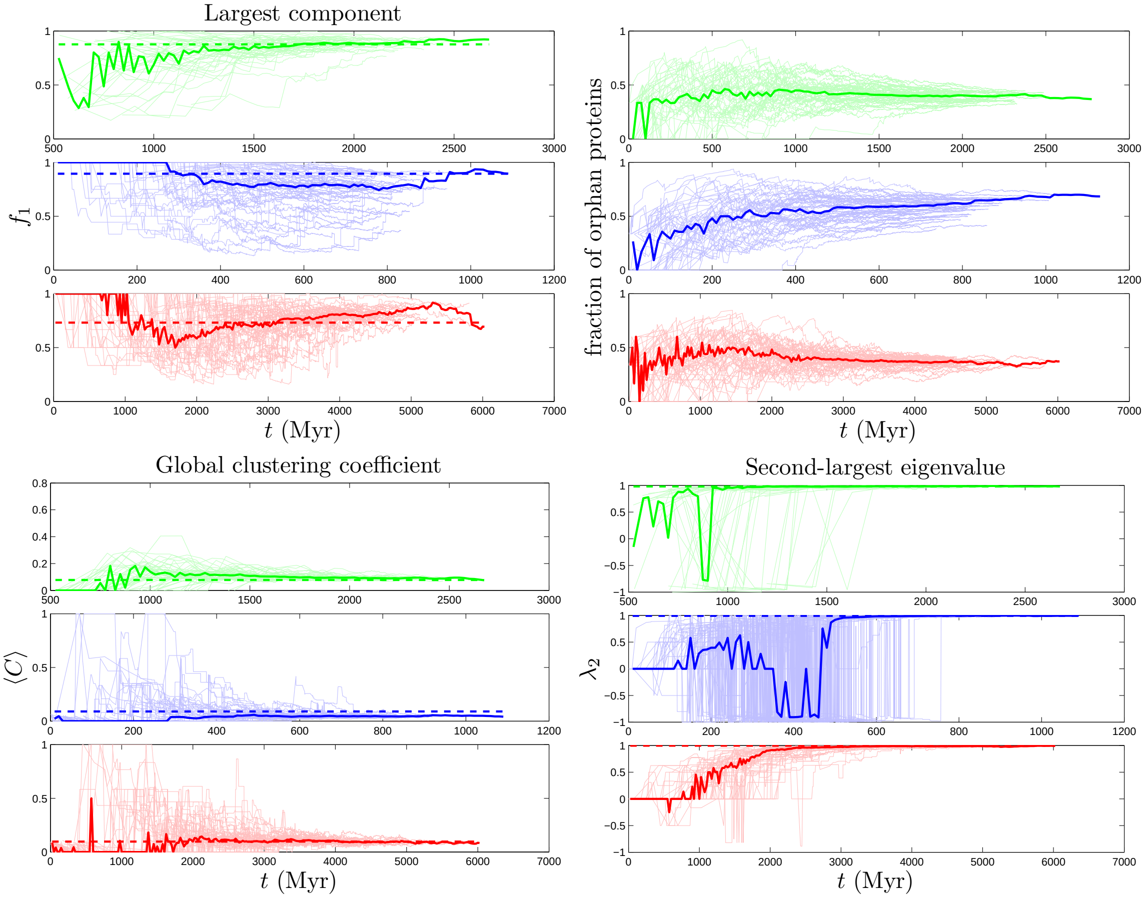


Figure 19: Evolution of several dynamical features in human (green), yeast (blue), and fly (red). The increasing value of $\langle C \rangle$ over time indicates that the network becomes more clustered as it grows. λ_2 approaches 1 over time, indicating that signal propagation is slower and slower as the network grows larger and more modular. Light lines indicate the evolutionary trajectories of 50 individual simulations, and the heavy line is the median value. Empirical data values are shown as a dashed line, where available.

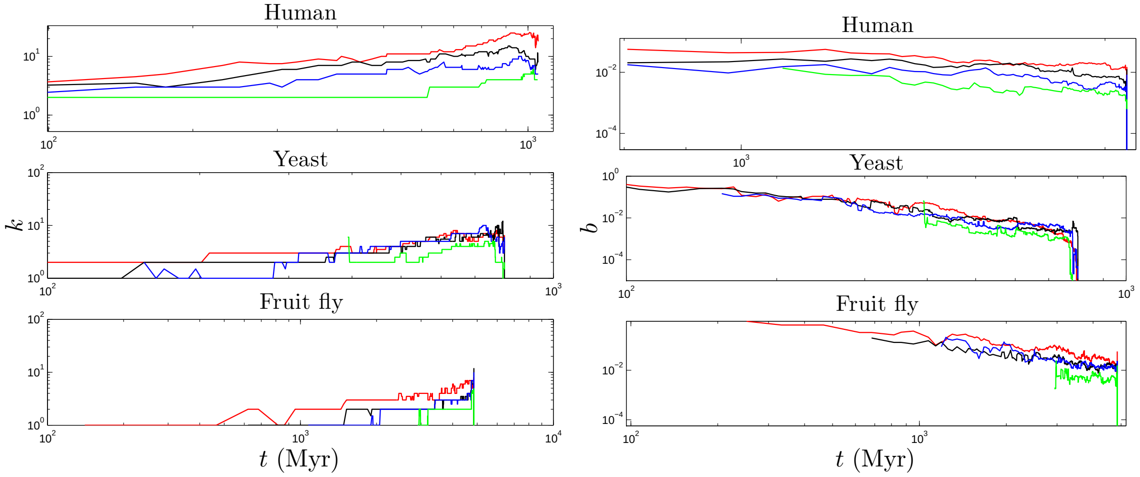


Figure 20: Evolution of degree and betweenness for proteins introduced to the network at different times in humans (top), yeast (middle), and flies (bottom). The 1st protein (one of the two initial proteins) is shown in red, the 6th protein in black, the 11th protein in blue, and the 101st protein in green. Curves are median values from 50 simulations.

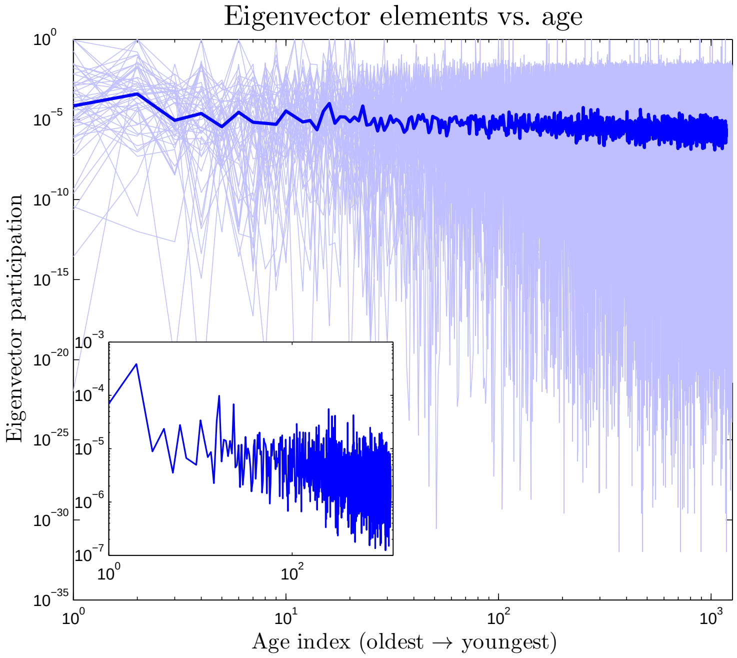


Figure 21: Elements of the eigenvector of the Laplacian matrix (defined as $\mathbf{K} - \mathbf{A}$, where $\mathbf{K}$ is a diagonal matrix with the degree of node i as element K_{ii}) associated with the largest eigenvalue vs. protein age index (time of introduction) in the yeast simulation. Details of this method are discussed in [109]. Heavy lines are the median values from 50 simulations, and light lines are results of individual simulations. The inset plot shows the trend line with a rescaled y-axis.

4 Power-law distributions derived from maximum entropy and a symmetry relationship

Power-law probability distributions are ubiquitous in nature, especially in social systems. For example, the fraction p_k of U.S. cities with a population of k people scales as $p_k \sim k^{-2.37}$ [15, 16, 14]. Other examples of power-law distributions include incomes [18], Internet links [13, 20], fluctuations in stock market prices [125, 3, 17], company sizes [19], numbers of citations received by scientific papers [24, 1, 11], and many others [14].

Here, our interest is in how power-law distributions might arise from stochastic processes, particularly in social physics. Our approach is based on the principle of maximum entropy (MaxEnt) [126, 127, 128]. MaxEnt is widely used, not only in thermal physics, but also in image analysis [129, 130, 131], drawing inferences [132, 133], and in nonequilibrium statistical mechanics [134, 135, 136]. We show here that the same principle, with a 'cost-sharing' type of constraint, leads to power-law distributions.

4.1 Theory

4.1.1 Clustering can be framed in terms of 'joining costs'

We focus on a problem of particle clustering, which provides a convenient language for comparing people joining cities to the statistical physics of growing colloids and polymers. We first describe a standard growth mechanism, which we express in terms of the 'joining costs' of a particle to a growing cluster. One particle may stick to another, forming a cluster of size 2. Another particle may join the cluster, forming a size 3 cluster, and so on. If there are N total particles, then the equilibrium of clusters of various sizes can be written as

$$\begin{aligned} 1 + 1 &\overset{K_1}{\rightleftharpoons} 2 \\ 1 + 2 &\overset{K_2}{\rightleftharpoons} 3 \\ &\vdots \\ 1 + (N-1) &\overset{K_{N-1}}{\rightleftharpoons} N \end{aligned} \tag{25}$$

where K_k is the k^{th} equilibrium binding constant. Let n_k be the number of clusters of size k at equilibrium, so that

$$K_k = \frac{n_{k+1}}{n_k n_1}. \tag{26}$$

Rearranging and solving for n_k yields

$$n_k = n_{k-1} n_1 K_{k-1} = {n_1}^{k-1} K_{k-1} K_{k-2} \cdots K_1. \tag{27}$$

The probability distribution, p_k, of cluster sizes is the ratio of the number of size k clusters to the total number of clusters of all sizes,

$$p_k = Q^{-1} n_k, \tag{28}$$

where Q is the grand canonical partition function,

$$Q = \sum_{k=1}^{N} {n_1}^{k-1} \prod_{j=1}^{k-1} K_j. \tag{29}$$

Statistical physics provides an alternative language for expressing the logarithms of populations in terms of energies, free energies or chemical potentials, which are cost-like additive quantities. Here, we define a dimensionless chemical-potential-like quantity, μ_k, that we call the *joining cost* for a particle to join a size-k cluster,

$$\mu_k \equiv \ln\left(K_k n_1\right). \tag{30}$$

Re-expressing Q in terms of these costs gives

$$Q = \sum_k \mathrm{e}^{-\sum_{j=1}^{k-1} \mu_j} = \sum_k \mathrm{e}^{-w_k}, \tag{31}$$

where

$$w_k \equiv \sum_{j=1}^{k-1} \mu_j \tag{32}$$

is the dimensionless cost of assembling the whole cluster. In this language, if w_k is positive, the distribution will be dominated by small clusters; if w_k is negative, the system will preferentially populate large clusters. We use this cost language for social physics below. To proceed further, we need to know how w_k depends on k.

4.1.2 Independent particle clusters are exponentially distributed

First, for illustration, we treat a standard problem of colloid assembly or polymerization. Assume the cost μ_k for a particle to join a cluster is independent of k. Then

$$\mu_k = \mu^\circ, \tag{33}$$

where the constant μ° is the cost of adding one particle to a cluster of any size. A cluster of size k requires $k-1$ particle additions, so the total cost of assembling the cluster is

$$w_k = \mu^\circ\left(k-1\right), \tag{34}$$

and the average cost of adding a particle, taken over all cluster sizes, is

$$\langle w \rangle \equiv \sum_k w_k p_k = \mu^\circ \left(\langle k \rangle - 1 \right). \tag{35}$$

To predict the probability distribution of cluster sizes, we maximize the entropy,

$$S = -\sum_k p_k \ln p_k, \tag{36}$$

subject to two constraints: (1) a fixed known value of $\langle w \rangle$ (Eq. 35) and (2) normalization (ensuring the probabilities sum to 1),

$$\sum_k p_k = 1. \tag{37}$$

A constraint on $\langle w \rangle$ gives the extremum function

$$\langle w \rangle - S = \mu^\circ \sum_k (k-1) p_k + \sum_k p_k \ln p_k, \tag{38}$$

so that the optimization condition is

$$\sum_k dp_k^* \big[\ln p_k^* + 1 + \alpha + \mu^\circ \left(k - 1 \right) \big] = 0, \tag{39}$$

where α and μ° are the Lagrange multipliers that enforce normalization and constraint 35, respectively. Solving Eq. 39 gives the equilibrium probability distribution, p_k^*, which maximizes the entropy and satisfies the average cost and normalization constraints. The solution is the standard exponential distribution of cluster sizes,

$$p_k^* = \mathrm{e}^{-1-\alpha} \mathrm{e}^{-\mu^\circ (k-1)} = Q^{-1} \, \mathrm{e}^{-\mu^\circ k}, \tag{40}$$

where Q is the grand canonical partition function,

$$Q = \sum_{k=1}^{N} \mathrm{e}^{-\mu^\circ k} = \frac{1 - \mathrm{e}^{-\mu^\circ N}}{1 - \mathrm{e}^{-\mu^\circ}}. \tag{41}$$

4.1.3 Equal cost sharing leads to power-law distributions

Now, we consider a more general notion of a particle's joining cost when it enters a k-mer cluster. A person is more likely to join a larger city than a smaller city because of greater opportunities of jobs, infrastructure, services, entertainment, and other economic and social factors. Existing citizens have already paid some of the cost of entry for the new joiner 'particle'. So, relative to the cost of joining an independent-particle cluster (μ°), the cost μ_k of joining a 'social-particle' cluster of size k is reduced to

$$\mu_k = \mu^\circ - kr_k. \tag{42}$$

Eq. 42 expresses the idea of *cost sharing*, namely that the cost of joining a cluster is reduced because the existing k member particles provide a discount of r_k each to the joiner particle.

There are two non-arbitrary limiting cases for how we might choose the value of r_k: **(1) No sharing**, $r_k = 0$, and the particles are independent, as described above, or **(2) Equal sharing**, where each member pays the same amount as the joiner when it enters the cluster,

$$r_k = \mu_k. \tag{43}$$

Substituting Eq. 43 into 42 and solving for μ_k yields

$$\mu_k = \frac{\mu^\circ}{1 + k}, \tag{44}$$

which expresses how the costs diminish with cluster size in social-particle systems.

The total cost to assemble a social cluster of k particles is

$$w_k = \mu^\circ \sum_{j=1}^{k-1} \frac{1}{1+j}. \tag{45}$$

The sum in Eq. 44 can be written

$$\sum_{j=1}^{k-1} \frac{1}{1+j} = \psi\left(k+1\right) + \gamma - 1. \tag{46}$$

where $\psi(k) \equiv d \ln \Gamma(k)/dk$ is the digamma function, and $\gamma = 0.5772...$ is Euler's constant. $\psi(k+1)$ has asymptotic series [137]

$$\psi\left(k+1\right) \sim \ln\left(k+\frac{1}{2}\right) + \mathcal{O}\left(k^{-2}\right). \tag{47}$$

Dropping the order k^{-2} corrections, Eq. 44 becomes

$$w_k \approx \mu^\circ \left[\ln\left(k+\frac{1}{2}\right) + \gamma - 1\right], \tag{48}$$

and the average joining cost per particle is

$$\langle w \rangle \approx \mu^\circ \left[\left\langle \ln\left(k+\frac{1}{2}\right)\right\rangle + \gamma - 1\right]. \tag{49}$$

To predict the social-particle probability distribution, we maximize the entropy subject to constraint 49 on $\langle w \rangle$,

$$\sum_k dp_k^* \left\{\ln p_k^* + 1 + \alpha + \mu^\circ \left[\ln\left(k+\frac{1}{2}\right) + \gamma - 1\right]\right\} \approx 0. \tag{50}$$

Solving for p_k^* yields

$$p_k^* \sim \left(k+\frac{1}{2}\right)^{-\mu^\circ}, \tag{51}$$

giving a power-law distribution. The scaling exponent μ° is the 'un-discounted cost' of adding one particle to a cluster. Scaling form 51 is accurate for most values of k. The exact distribution is given by

$$p_k^* = Q^{-1} \mathrm{e}^{-\mu^\circ \psi(k+1)}, \tag{52}$$

with partition function

$$Q = \sum_k \mathrm{e}^{-\mu^\circ \psi(k+1)}. \tag{53}$$

The present work shows how power-law distributions can emerge naturally from random clustering of particles that equally share the joining cost. The 'rich-get-richer' aspect of power-law size distributions is that social particles are more attracted to bigger clusters than to smaller clusters. We have expressed this in a language of 'costs': the power-law arises because member particles equally share the joining costs with joiner particles. The power-law exponent is the 'un-reduced' cost μ°.[12]

4.1.4 A generalized model for partial cost sharing

Described above are two limiting cases: no cost sharing (independent particles) or full cost sharing (the members pay the same as the joiner). What if the member particles only pay a fraction s of the cost that the joiner particle pays? Now the cost reduction is

$$r_k = s\mu_k. \tag{54}$$

Combining Eq. 54 with Eq. 42 and solving for μ_k, we find

$$\mu_k = \frac{\mu^\circ}{1 + sk}, \tag{55}$$

[12] Our term 'costs' here can alternatively be thought of in terms of 'economies of scale'. Making more widgets decreases the cost-per-widget. That is, if the cost of a widget factory is μ°, then the cost per widget for the first widget is μ° and the cost per widget for making two widgets is $\mu^\circ/2$. In this sense, the cost of the second is 'shared' by the first.

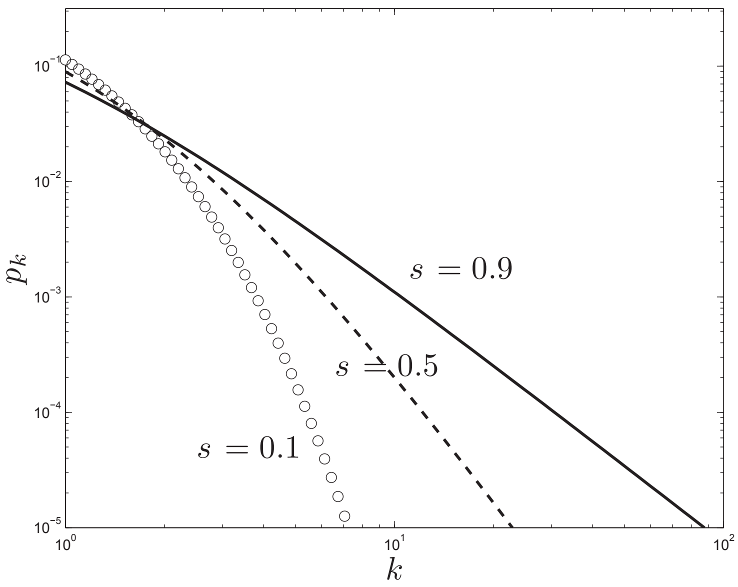


Figure 22: Probability distributions for various sharing values, with $\mu^\circ = 2$. At $s = 0.1$, the particles are barely helpful at all, and the entering particle must pay most of its joining cost, so the distribution is nearly exponential. At $s = 0.9$, the particles are very helpful, and the distribution is a power-law.

so the total cost of assembling a partially-social cluster of size k is

$$w_k = \mu^\circ \sum_{j=1}^{k-1} \frac{1}{1+sj}. \tag{56}$$

The sum in Eq. 56 may be written

$$\sum_{j=1}^{k-1} \frac{1}{1+sj} = \frac{1}{s}\left[\psi\left(k+\frac{1}{s}\right) - \psi\left(1+\frac{1}{s}\right)\right], \tag{57}$$

As with the social-particle case, we asymptotically expand $\psi(k+1/s)$ and drop the order k^{-2} corrections, yielding

$$w_k \approx \frac{\mu^\circ}{s}\left[\ln\left(k+\frac{1}{s}-\frac{1}{2}\right) - \psi\left(1+\frac{1}{s}\right)\right]. \tag{58}$$

Maximizing the entropy subject to a constraint on $\langle w \rangle$ gives

$$p_k^* \sim \left(k + \frac{1}{s} - \frac{1}{2} \right)^{-\mu^\circ/s}. \tag{59}$$

As with the social-particle p_k, this scaling form is accurate for most values of k. The exact distribution is given by

$$p_k^* = Q^{-1} \mathrm{e}^{-\frac{\mu^\circ}{s} \psi\left(k + \frac{1}{s}\right)}, \tag{60}$$

where

$$Q = \sum_k \mathrm{e}^{-\frac{\mu^\circ}{s} \psi\left(k + \frac{1}{s}\right)}. \tag{61}$$

We call s the sharing parameter: $s = 0$ involves no sharing (*i.e.*, independent particles) and $s = 1$ involves full equal sharing between the joiner particle and all members. The s parameter controls the shape of the distribution. For $s = 0$, the entry cost is the same for all clusters, and the distribution is exponential; there are very few large clusters. For $s = 1$, the particles are social, and the entry cost is reduced for larger clusters, resulting in a power-law distribution; there are more large clusters in this case. Fig. 24 shows that varying s changes the distribution smoothly from exponential to power-law. We refer to the limit $s \to \infty$ as 'super-social particles': the existing members pay the full cost, and the joiner particle pays nothing. For super-social particles, we obtain a uniform distribution; all cluster sizes are equally probable.

4.2 Results

We fit Eq. 60 to several empirical data sets, representing a wide variety of systems:

- Number of citations by 1997 to all scientific papers published in 1981 and stored in the Institute for Scientific Information (ISI) database [1], number of citations to all papers in the ISI database for a 2007 list of the living highest h-index chemists [36], and number of citations of papers published in the *Physical Review D* journal from 1975 - 1994 [1]

- Number of pairwise, physical protein-protein interactions (PPI) of proteins detected in small-scale PPI network data, in yeast (*Saccharomyces cerevisiae*), fruit flies (*Drosophila melanogaster*), and humans (*Homo sapiens*) [123]

- Number of deaths resulting from terrorist attacks from February 1968 to June 2006 [138]

- Number of links to web pages in 1997 [13]

- Populations of cities in the United States, as of the 2000 census [14]

- Number of people affected per electrical blackout in the U.S. between 1984 and 2002 [4]

- Count of unique word use in the novel *Moby Dick* [4]

- Counts of surnames in the U.S., as of the 1990 census

- Daily, weekly, and monthly fluctuations in the closing price of the S&P 500, from 1950 to 2010, in 2010 U.S. dollars

For each data set, we obtained best-fit parameter values (for μ° and s) and 95% confidence intervals by fitting Eq. 60 to the data using nonlinear regression in Matlab. Best-fit parameter values are shown in Table 7, and results are plotted against the data in Fig. 23. To reduce noise in the U.S. city size data set, we binned the data, using size 200 bins. The blackout data was binned using 100 equal sized bins. Stock price fluctuations are reported as magnitudes (absolute values), and are rounded to the nearest nickel, to reduce noise. All other distributions are raw (whole number) counts.

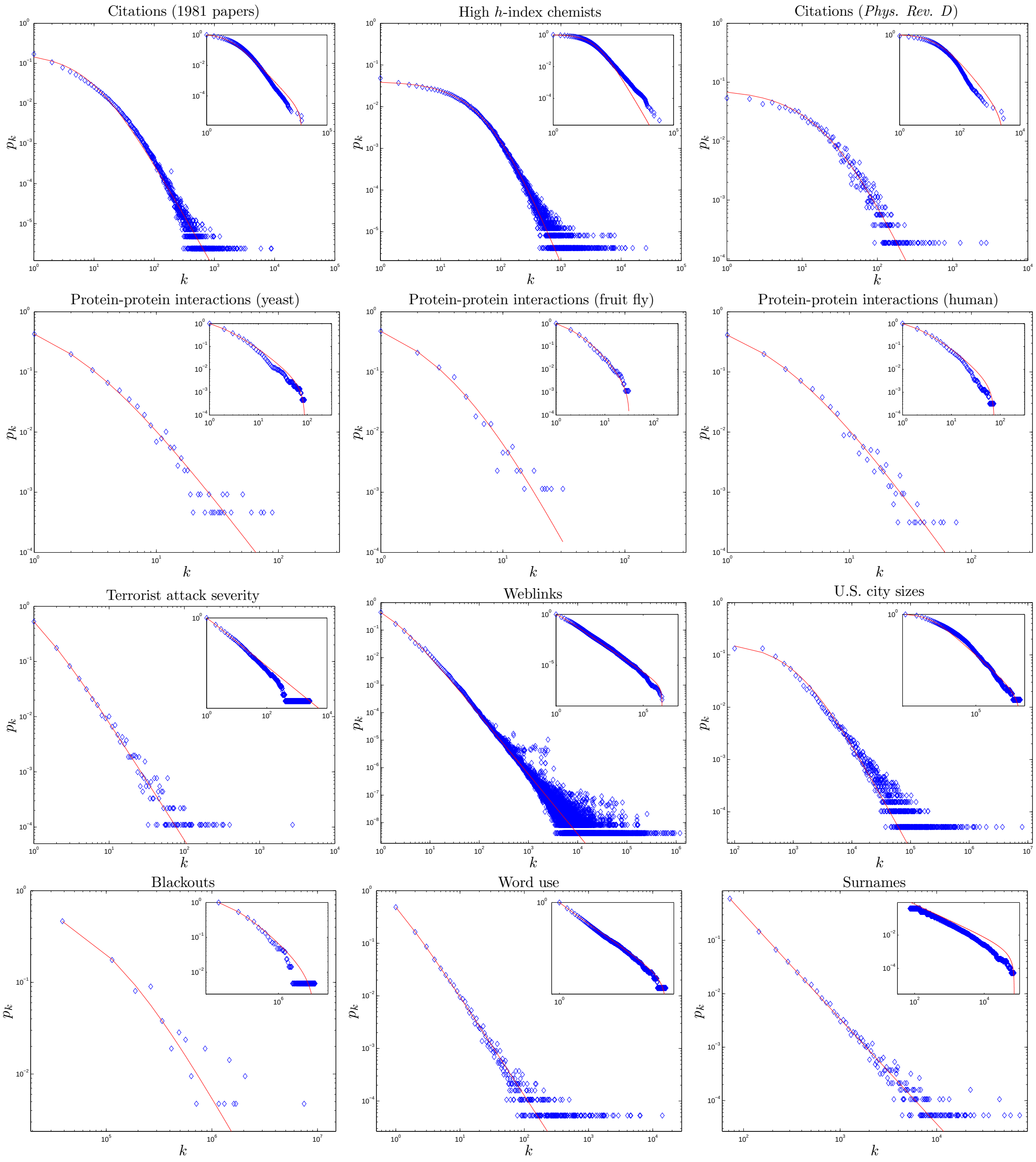


Figure 23: Partially-social p_k (Eq. 60) fitted to several empirical distributions, using the parameters listed in Table 7. Points are empirical data, and lines represent best-fit distributions. Main plots show the probability distribution function (the probability of exactly k events), and inset plots show the complementary cumulative distribution function (the probability of at least k events), $\sum_{j=k}^{\infty} p_j$. To help visualize the tail, all plots have log-log axes.

Data set	μ°	s	N
1. Citations (1981)	0.2616(4)	0.0975(3)	415,229
2. Citations (chem)	0.05476(7)	0.01523(5)	245,461
3. Citations (*PRD*)	0.120(3)	0.044(2)	5,327
4. PPI (yeast)	1.08(3)	0.40(2)	2170
5. PPI (fruit fly)	1.0(1)	0.24(7)	878
6. PPI (human)	0.95(3)	0.33(2)	3165
7. Terrorist attacks	2.260(8)	1.045(5)	9,101
8. Weblinks	1.49652(5)	0.68823(4)	241,428,853
9. U.S. city sizes	0.001688(9)	0.000818(8)	19,447
10. Blackouts	0.000023(3)	0.000011(3)	211
11. Word use	2.501(4)	1.324(3)	18,855
12. Surnames*	1.962(4)	1	88,799
13. Stocks (daily)	0.358(8)	0.063(7)	15,601

Table 7: Fitting parameters μ° and s and data set sizes N. The surnames data set appeared to be pure power-law, so we fixed $s = 1$ and only fitted μ°.

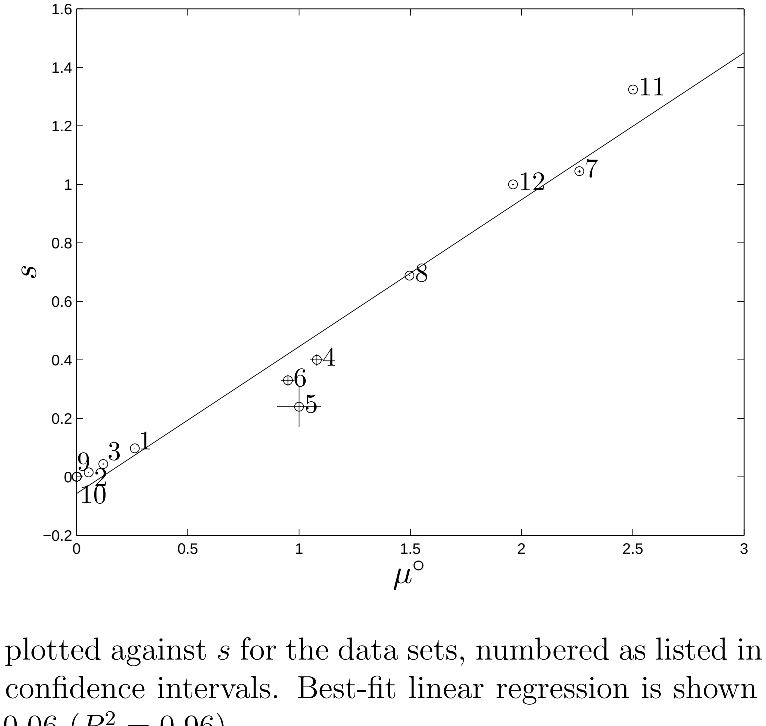


Figure 24: μ° plotted against s for the data sets, numbered as listed in Table 7. Error bars are 95% confidence intervals. Best-fit linear regression is shown as a solid line, $s = 0.50\,\mu^\circ - 0.06$ ($R^2 = 0.96$).

As shown in Fig. 23, Eq. 60 is a good fit to these 12 data sets. Eq. 60 may be useful simply as a fitting function, even when the data sets are not clearly related to the idea of cost-sharing (*e.g.*, the number of people affected by blackouts). Eq. 60 is quite versatile, in that it can fit both pure exponential and pure power-law distributions, as well as more complicated distributions which appear to be exponential close to the origin, but power-law in the tail. The ratio μ°/s suggests whether the data is better represented by a power-law (if the ratio is small) or an exponential (if it is large).

However, in many cases, the underlying system may be interpretable in terms of shared costs. There is an extensive literature of models that generate power laws [139, 4]; here, we briefly discuss several well-known power-law distributions, and interpret them in terms of cost-sharing.

In many situations involving people, the idea of cost-sharing arises naturally. For example, a well-studied power-law distribution is of the number of citations to scientific papers. Suppose the author of paper A is preparing to cite paper C. Every paper in the set {B} that has already cited paper C can be thought of as a member of a community that paper A is about to join. The larger is the community {B}, the more likely it is for paper A to join that community, and cite paper C. The community {B} has already paid a higher cost by finding paper C in sea of options that was larger at the time. In this way, any paper B has lowered the cost for the author of paper A to find and cite paper C. Note that the sharing values found for the three citations data sets examined here are relatively small ($s = 0.098$, 0.015, and 0.044), suggesting that each citation a paper receives only causes a minor reduction in the cost of citing the paper. This may be because many papers are found essentially at random, rather than through the reference lists of other papers.

For the distribution of links to webpages, the cost of making a link primarily represents the difficulty of finding the site to begin with. Most visits to websites occur because a person browing the web clicks on an existing link to the site, which is a sharp contrast with the situation for scientific citations, and is reflected in the relatively high sharing value for this data set ($s = 0.69$). In this case, the cost of making a link is substantially reduced by each previously made link.

The distribution of U.S. city sizes also has a power-law tail. It is more likely a person will choose to move to Los Angeles, CA (population: 9 million) than to Fields, Oregon (population: 86). Similar to the cost amortization obtained through an economy of scale, the people of LA have already paid the development costs of creating the companies, jobs, infrastructure, and services that attract new individuals. Thus, the marginal cost of adding one more person is reduced, relative to the cost of adding a person to a smaller city. The sharing value $s = 0.00082$ is quite small for this data set, possibly because each person makes a relatively small contribution to the overall cost amortization. Put another way, when considering a town's economic climate, infrastructure, etc., a town with population 86 is probably not terribly different, on average, from a town of 172.

It has been argued that terrorism is intended for the survivors, rather than the victims, since the goal of terrorist attacks is usually to attract attention to a cause, whether that is religious fundamentalism, animal rights, national separatism, etc. Because of this, the immediate result of terrorism (x number of people dead) might be thought of as a proxy for the fearful reaction of viewers to news of terrorism (k number of people terrified). In general, the more people killed, the more fearful the reaction. However, once terrorism becomes commonplace, terrorists need to kill more and more people to generate the same reaction. The news that 10 people were killed in a bombing by terrorist group A on Monday might grab a viewer's attention. Terrorist group B sees that group A was successful in getting attention, so they so a similar bombing on Thursday that also kills 10 people. Terrorist group C then does another copycat bombing the followng Monday that kills another 10 people. Intuitvely, group B might be expected to receive less attention than group A, and group C might receive less still. In recognition of this effect, it seems likely that groups B and C will try and kill progressively larger numbers of people, in order to receive the same amount of attention to their cause. The very high value of $s = 1.045$ for terrorist attack severity suggests that this effect may be fairly strong.

It has been observed that the degree distributions (the probability for a protein to have k physical interactions with other proteins) of PPI networks in cells have power-law tails [2]. The cost of forming an interaction between two proteins represents the energy barriers required for two proteins to stick together. Previous work suggests that one underlying force shaping the degree distribution of present-day PPI networks are the fractions of exposed hydrophobic surface area per protein [88]. μ° is the energy barrier that a protein's first interaction partner is required to overcome, which represents the cost to the proteins to have an energetically-unfavorable patch of exposed hydrophobic surface area. A second interaction partner can now take advantage of the pre-existing binding site, so it is easier to form the second link, compared to the first. Alternatively, the second partner could require a completely different binding site than the first; the medium-sized values of s for the three PPI networks examined (0.40, 0.24, and 0.33) suggest that both mechanisms are probably common.

4.3 Discussion

We have shown how power-law distributions can arise naturally from the maximization of entropy subject to a symmetry relationship in which all particles share the cost incurred when a new particle joins a cluster. It has been noted before that MaxEnt with logarithmic constraints leads to power-law distributions [140, 141].[13] Exponential distributions result from constraints on linear averages such as $\langle k \rangle$, while power-law distributions result from constraints on logarithmic averages such as $\langle \ln k \rangle$. Here, we have described how such constraints can be interpreted as a type symmetry

[13] Logarithmic constraints have also been justified in terms of the information content of language [142], to explain the observation that word-use frequency obeys a power-law [15, 4].

of sharing that is natural in the social realm.

Many previous studies on generative mechanisms for power-laws have investigated a family of 'proportional attachment' (PA) rules [25, 26, 24, 5]. In the context of particle clustering, a PA rule says that the probability of a cluster acquiring a new particle is proportional to the number of particles it already contains. A premise of the PA rule is that the joining particle is cognizant of the populations of the clusters in the system. By contrast, our cost-sharing framework does not assume the particles know anything about the system, so the present approach may be useful for modeling systems composed of 'uninformed' particles.

References

[1] S. Redner. How popular is your paper? An empirical study of the citation distribution. *Eur. Phys. Jour. B.*, 4:131–134, 1998.

[2] J. Berg, M. Lassig, and A. Wagner. Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evolutionary Biology*, 4:51–63, 2004.

[3] P. Gopikrishnan, V. Plerou, L.A.N. Amaral, M. Meyer, and H.E. Stanley. Scaling of the distributions of fluctuations of financial market indices. *Phys. Rev. E*, 60:5305–5316, 1999.

[4] M.E.J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46:323, 2005.

[5] A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[6] R. Albert, H. Jeong, and A.L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–382, 2000.

[7] G. Bianconi and A.L. Barabási. Competition and multiscaling in evolving networks. *Europhys. Lett.*, 54:436–442, 2001.

[8] K. Klemm and V.M. Eguíluz. Highly clustered scale-free networks. *Phys. Rev. E*, 65, 2002.

[9] K.B. Hajra and P. Sen. Modelling aging characteristics in citation networks. *Physica A*, 368:575–582, 2006.

[10] J. Leskovec and C. Faloutsos. Scalable modeling of real graphs using kronecker multiplication. *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[11] G.J. Peterson, S. Pressé, and K.A. Dill. Nonuniversal power law scaling in the probability distribution of scientific citations. *Proc. Natl. Acad. Sci. USA*, 107:16023–16027, 2010.

[12] M.H. Biglu. The influence of references per paper in the SCI to Impact Factors and the Matthew Effect. *Scientometrics*, 74:453–470, 2007.

[13] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer Networks*, 33:309, 2000.

[14] A. Clauset, C.R. Shalizi, and M.E.J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51:661–703, 2009.

[15] George K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley: Cambridge, 1949.

[16] X. Gabaix. Zipf's law for cities: an explanation. *Q.J. Econ.*, 114:739–767, 2001.

[17] V. Plerou, P. Gopikrishnan, L.A.N. Amaral, M. Meyer, and H. E. Stanley. Scaling of the distribution of price fluctuations of individual companies. *Phys. Rev. E*, 60:6519–6529, 1999.

[18] K. Okuyama, M. Takayasu, and H. Takayasu. Zipf's law in income distribution of companies. *Physica A*, 269:125–131, 1999.

[19] R. Axtell. Zipf distribution of U.S. firm sizes. *Science*, 293:1818–1820, 2001.

[20] P. Holme, J. Karlin, and S. Forrest. Radial structure of the Internet. *Proc. R. Soc. A*, 463:1231–1246, 2007.

[21] M.E.J. Newman. Scientific collaborations networks. *Phys. Rev. E*, 64, 2001.

[22] A.L. Barabási, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A*, 311:590–614, 2002.

[23] S. Redner. Citations statistics from 110 years of physical review. *Physics Today*, 58:49, 2004.

[24] D.J. de Solla Price. A general theory of bibliometric and other cumulative advantage processes. *J. Am. Soc. Inform. Sci.*, 27:292–306, 1976.

[25] G.U. Yule. A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis. *Philos. Trans. R. Soc. London B*, 213:21–87, 1925.

[26] H.A. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955.

[27] R.K. Merton. The Matthew effect in science. *Science*, 159:56–63, 1968.

[28] S. Lehmann, B. Lautrup, and A.D. Jackson. Citations networks in high energy physics. *Physical Review E*, 68:026113, 2003.

[29] S. Redner. Citations statistics from 110 years of *Physical Review*. *Physics Today*, 58:49–54, 2005.

[30] J. Laherrère and D. Sornette. Stretched exponential distributions in nature and economy: “fat tails.

[31] P.L. Krapivsky and S. Redner. Organization of growing random networks. *Physical Review E*, 63:066123, 2001.

[32] H.D. Rozenfeld and D. ben-Avraham. Designer nets from local strategies. *Physical Review E*, 70:056107, 2004.

[33] D. Walker, H. Xie, K. Yan, and S. Maslov. Ranking scientific publications using a model of network traffic. *J. Stat. Mech.*, 2007:P06010, 2007.

[34] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Structure of growing networks: Exact solution of the Barabási-Albert's model. *Phys. Rev. Lett.*, 85:4633–4636, 2000.

[35] E. Ravasz and A.L. Barabási. Hierarchical organization in complex networks. *Phys. Rev. E*, 67, 2003.

[36] A. Peterson and H. Schaefer. H-index ranking of living chemists. *Chemistry World*, 2007.

[37] F. Mosteller and J.W. Tukey. *Data Analysis and Regression.* Addison-Wesley, Reading, 1977.

[38] J.E. Hirsch. An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. USA*, 102:16569–16572, 2005.

[39] D. Hughes. Microbial genetics: Exploiting genomics, genetics and chemistry to combat antibiotic resistance. *Nature Reviews Genetics*, 4:432–441, 2003.

[40] R.T. Cirz, J.K. Chin, D.R. Andes, V. de Crécy-Lagard, W.A. Craig, and F.E. Romesberg. Inhibition of mutation and combating the evolution of antibiotic resistance. *PLoS Biology*, 3:e176, 2005.

[41] I.W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J.L. Wrana. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology*, 27:199–204, 2009.

[42] M. Lynch, M. O'Hely, B. Walsh, and A. Force. The probability of preservation of a newly arisen gene duplicate. *Genetics*, 159(4):1789–1804, 2001.

[43] C.-T. Ting, S.-C. Tsaur, S. Sun, W.E. Browne, Y.-C. Chen, N.H. Patel, and C.-I. Wu. Gene duplication and speciation in *Drosophila*: evidence from the *Odysseus* locus. *Proc. Natl. Acad. Sci. USA*, 101(33):12232–12235, 2004.

[44] J. Dutkowski and J. Tiuryn. Phylogeny-guided interaction mapping in seven eukaryotes. *BMC Bioinformatics*, 10:393, 2009.

[45] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci. USA*, 97:1143–1147, 2000.

[46] P. Uetz, L. Giot, G. Cagney, T. Mansfield, R. Judson, J. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. Rothberg. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae. Nature*, 403:623–627, 2000.

[47] N.J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A.P. Tikuisis, T. Punna, J. Peregran-Alvarez, M. Shales, X. Zhang, M. Davey, M.D. Robinson, A. Paccanaro, J.E. Bray, A. Sheung, B. Beattie, D.P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M.M. Canete, J. Vlasblom, S. Wu, C. Orsi, S.R. Collins, S. Chandran, R. Haw, J.J. Rilstone, K. Gandi, N.J. Thompson, G. Musso, P. St. Onge, S. Ghanny, M.H.Y. Lam, G. Butland, Atlaf Ul A.M., S. Kanaya, A. Shilatifard, E. O'Shea, J.S. Weissman, J.C. Ingles, T.R. Hughes, J. Parkinson, M. Gerstein, S.J. Wodak, A. Emili, and J.F. Greenblatt. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae. Nature*, 440:637–643, 2006.

[48] H. Yu, P. Braun, M.A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.F. Rual, A. Dricot, A. Vazquez, R.R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A.S. de Smet, A. Motyl, M.E. Hudson, J. Park, X. Xin, M.E. Cusick, T. Moore, C. Boone, M. Snyder, F.P. Roth, A.L. Barabási, J. Tavernier, D.E. Hill, and M. Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science*, 5898:104–110, 2008.

[49] E. Marcotte, M. Pellegrini, H. Ng, D. Rice, T. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Nature*, 402:86–90, 1999.

[50] M. Pellegrini, E. Marcotte, M. Thompson, D. Eisenberg, and T. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA*, 96:4285–4288, 1999.

[51] A. Valencia and F. Pazos. Computational methods for the prediction of protein interactions. *Current Opinion in Structural Biology*, 12:368–373, 2002.

[52] S. Gomez, W. Noble, and A. Rzhetsky. Learning to predict protein-protein interactions from protein sequences. *Bioinformatics*, 19:1875–1881, 2003.

[53] R. Jothi, M. Kann, and T. Przytycka. Predicting protein-protein interaction by searching evolutionary tree automorphism space. *Bioinformatics*, 21:241–250, 2005.

[54] Y. Liu, N. Liu, and H. Zhao. Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics*, 21:3279–3285, 2005.

[55] B. Shoemaker and A. Panchenko. Deciphering protein-protein interactions. part II. computational methods to predict protein and domain interaction partners. *PLoS Computational Biology*, 3:e43, 2007.

[56] L. Burger and E. van Nimwegen. Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Mol. Syst. Biol.*, 4:165, 2008.

[57] C.M. Deane, Ł. Salwiński, I. Xenarios, and D. Eisenberg. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Molecular & Cellular Proteomics*, 1(5):349–356, 2002.

[58] T. Yamada and P. Bork. Evolution of biomolecular networks – lessons from metabolic and protein interactions. *Nature Reviews Molecular Cell Biology*, 10:791–803, 2009.

[59] Susumu Ohno. *Evolution by Gene Duplication.* Springer-Verlag, New York, 1970.

[60] J. Zhang. Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, 18(6):292–298, 2003.

[61] H. Xiao, N. Jiang, E. Schaffner, E.J. Stockinger, and E. van der Knaap. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science*, 319:1527–1530, 2008.

[62] A. Wagner. How the global structure of protein interaction networks evolves. *Proc. R. Soc. Lond. B*, 270:457–466, 2003.

[63] A.L. Koch. Enzyme evolution. i. the importance of untranslatable intermediates. *Genetics*, 72(2):297–316, 1972.

[64] J.S. Taylor and J. Raes. Duplication and divergence: the evolution of new genes and old ideas. *Annual Review of Genetics*, 38:615–643, 2004.

[65] M. Lynch and J.S. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290:1151–1155, 2000.

[66] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani. Modeling of protein interaction networks. *ComPlexUs*, 1:38–44, 2003.

[67] M. Kellis, B.W. Birren, and E.S. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, 428:617–624, 2004.

[68] X. Gu, Z. Zhang, and W. Huang. Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc. Natl. Acad. Sci. USA*, 102:707–712, 2005.

[69] M. Lynch and B. Walsh. *Genetics and Analysis of Quantitative Traits.* Sinauer, Sunderland, MA, 1998.

[70] S. Jones and J.M. Thornton. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA*, 93:13–20, 1996.

[71] A. Tovchigrechko and I.A. Vakser. How common is the funnel-like energy landscape in protein-protein interactions? *Protein Science*, 10(8):1572–1583, 2001.

[72] F. Wu, F. Towfic, D. Dobbs, and V. Honavar. Analysis of protein-protein dimeric interfaces. In *IEEE International Conference on Bioinformatics and Biomedicine*. Fremont, CA, 2007.

[73] A.E. Vinogradov and O.V. Anatskaya. Loss of protein interactions and regulatory divergence in yeast whole-genome duplicates. *Genomics*, 93(6):534–542, 2009.

[74] H. Escriva, S. Bertrand, P. Germain, M. Robinson-Rechavi, M. Umbhauer, J. Cartry, M. Duffraisse, L. Holland, H. Gronemeyer, and V. Laudet. Neofunctionalization in vertebrates: The example of retinoic acid receptors. *PLoS Genetics*, 2(7):e102, 2006.

[75] D.T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81:2340–2361, 1977.

[76] Z. Gu, A. Cavalcanti, F.C. Chen, P. Bouman, and W.H. Li. Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. *Molecular Biology and Evolution*, 19:256–262, 2002.

[77] L. Gao and H. Innan. Very low gene duplication rate in the yeast genome. *Science*, 306(5700):1367–1370, 2004.

[78] M. Lynch and J.S. Conery. The evolutionary demography of duplicate genes. *Journal of Structural and Functional Genomics*, 3:35–44, 2003.

[79] N. Osada and H. Innan. Duplication and gene conversion in the *Drosophila melanogaster* genome. *PLoS Genetics*, 4(12):e1000305, 2008.

[80] J.A. Cotton and R.D.M. Page. Rates and patterns of gene duplication and loss in the human genome. *Proceedings of the Royal Society B*, 272, 2005.

[81] D. Pan and L. Zhang. Quantifying the major mechanisms of recent gene duplications in the human and mouse genomes: a novel strategy to estimate gene duplication rates. *Genome Biol.*, 8(8):R158, 2007.

[82] P. Beltrao and L. Serrano. Specificity and evolvability in eukaryotic protein interaction networks. *PLoS Computational Biology*, 3, 2007.

[83] M. Lynch, W. Sung, K. Morris, N. Coffey, C.R. Landry, E.B. Dopman, W.J. Dickinson, K. Okamoto, S. Kulkarni, D.L. Hartl, and W.K. Thomas.

A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad. Sci. USA*, 105(27):9272–9277, 2008.

[84] T.A. Gibson and D.S. Goldberg. Questioning the ubiquity of neofunctionalization. *PLoS Computational Biology*, 5:e1000252, 2009.

[85] M. Kimura. Some problems of stochastic processes in genetics. *Ann. Math. Stat.*, 28:882–901, 1957.

[86] J.B. Walsh. How often do duplicated genes evolve new functions? *Genetics*, 139:421–428, 1995.

[87] X. He and J. Zhang. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, 169:1157–1164, 2005.

[88] E.J. Deeds, O. Ashenberg, and E.I. Shakhnovich. A simple physical model for scaling in protein-protein interaction networks. *Proc. Natl. Acad. Sci. USA*, 103:311–316, 2006.

[89] N. Pržulj, D.G. Corneil, and I. Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20:3508–3515, 2004.

[90] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5:17–61, 1960.

[91] A. Patil, K. Kinoshita, and H. Nakamura. Domain distribution and intrinsic disorder in hubs in the human protein-protein interaction network. *Protein Science*, 19:1461–1468, 2010.

[92] H. Jeong, S.P. Mason, A.L. Barabási, and Z.N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001.

[93] H.M. Ku, T. Vision, J. Liu, and S.D. Tanksley. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl. Acad. Sci. USA*, 97(16):9121–9126, 2000.

[94] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.

[95] J. Leskovec and E. Horvitz. In *Proceedings of the 17th international conference on World Wide Web*. ACM, New York, 2008.

[96] S.H. Yook, Z.N. Oltvai, and A.L. Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4:928–942, 2004.

[97] M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, 2004.

[98] M.E.J. Newman. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 103:8577–8582, 2006.

[99] Sergei Maslov and Kim Sneppen. Specificity and Stability in Topology of Protein Networks. *Science*, 296(5569):910–913, 2002.

[100] C. von Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 31:399–403, 2002.

[101] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555, 2002.

[102] A.L. Barabási and Z.N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5:101–113, 2004.

[103] M.P. Joy, A. Brock, D.E. Ingber, and S. Huang. High-betweenness proteins in the yeast protein interaction network. *Journal of Biomedicine and Biotechnology*, 2:96–103, 2005.

[104] D. Zhao, Z. Liu, and J. Wang. Duplication: a mechanism producing disassortative mixing networks in biology. *Chin. Phys. Lett.*, 24(10):2766, 2007.

[105] X. Wan, S. Cai, J. Zhou, and Z. Liu. Emergence of modularity and disassortativity in protein-protein interaction networks. *Chaos*, 20:045113, 2010.

[106] P. Aloy and R.B. Russell. Potential artefacts in protein-interaction networks. *FEBS Lett.*, 530:253–254, 2002.

[107] L. Hakes, J.W. Pinney, D.L. Robertson, and S.C. Lovell. Protein-protein interaction networks and biology - what's the connection? *Nature Biotechnol.*, 26:69–72, 2008.

[108] C.R. Woese. Bacterial evolution. *Microbiol. Rev.*, 51(2):221–271, 1987.

[109] G. Zhu, H. Yang, R. Yang, J. Ren, B. Li, and Y.C. Lai. Uncovering evolutionary ages of nodes in complex networks. 2011.

[110] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. *Knowledge Discovery in Databases: PKDD 2005*, 2005.

[111] M. Lynch. The evolution of genetic networks by non-adaptive processes. *Nat. Rev. Genet.*, 8(10):803–813, 2007.

[112] H. Lipson, J.B. Pollack, and N.P. Suh. On the origin of modular variation. *Evolution*, 56(8):1549–1556, 2002.

[113] N. Kashtan and U. Alon. Spontaneous evolution of modularity and network motifs. *Proc. Natl. Acad. Sci. USA*, 102(9):13773–13778, 2005.

[114] K. Komurov and M. White. Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. *Molecular Systems Biology*, 3:110, 2007.

[115] G.P. Wagner, M. Pavlicev, and J.M. Cheverud. The road to modularity. *Nature Reviews Genetics*, 8:921–931, 2007.

[116] C. Espinosa-Soto and A. Wagner. Specialization can drive the evolution of modularity. *PLoS Computational Biology*, 6(3):e1000719, 2010.

[117] O.S. Soyer. Fate of a duplicate in a network context. In K. Dittmar and D. Liberles, editors, *Evolution After Gene Duplication*, pages 215–228. Wiley-Blackwell, 2010.

[118] J. Hallinan. Gene duplication and hierarchical modularity in intracellular interaction networks. *BioSystems*, 74:51–62, 2004.

[119] R.V. Solé and S. Valverde. Spontaneous emergence of modularity in cellular networks. *J. R. Soc. Interface*, 5(18):129–133, 2008.

[120] R.T. Hietpas, J.D. Jensen, and D.N.A. Bolon. Experimental evolution of a fitness landscape. *Proc. Natl. Acad. Sci. USA*, 108(19):7896–7901, 2011.

[121] M. Heo, S. Maslov, and E. Shakhnovich. Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. *Proc. Natl. Acad. Sci. USA*, 108(10):4258–4263, 2011.

[122] M. Kac. Some mathematical models in science. *Science*, 166(3906):695–699, 1969.

[123] A. Patil, K. Nakai, and H. Nakamura. Hitpredict: a database of quality assessed protein-protein interactions in nine species. *Nucleic Acids Research*, 39 (suppl 1):D744–D749, 2011.

[124] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008:P10008+, 2008.

[125] R.N. Mantegna and H.E. Stanley. Scaling behaviour in the dynamics of an economic index. *Nature*, 376:46–49, 1995.

[126] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[127] E.T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.

[128] J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, 26(1):26–37, 1980.

[129] S.J. Wernecke and L.R. D'Addario. Maximum entropy image recontruction. *IEEE Transactions on Computers*, C-26(4):351–364, 1977.

[130] S.F. Gull and G.J. Daniell. Image reconstruction from incomplete and noisy data. *Nature*, 272:686–690, 1978.

[131] J. Skilling and R.K. Bryan. Maximum entropy image reconstruction – general algorithm. *R.A.S. Monthly Notices*, 211(1):111, 1984.

[132] E.T. Jaynes. Where do we stand on maximum entropy? In R.D. Levine and M. Tribus, editors, *The Maximum Entropy Formalism.* MIT Press, Cambridge, MA, 1979.

[133] J. Skilling and S.F. Gull. Bayesian maximum entropy image reconstruction. *Lecture Notes – Monograph Series*, 20:341–367, 1991.

[134] G. Stock, K. Ghosh, and K.A. Dill. Maximum caliber: A variational approach applied to two-state dynamics. *J. Chem. Phys.*, 128:194102, 2008.

[135] D. Wu, K. Ghosh, M. Inamdar, H.J. Lee, K.A. Dill, and R. Phillips. Trajectory approach to two-state kinetics of single particles on sculpted energy landscapes. *Physical Review Letters*, 103(5):050603(1–4), 2009.

[136] S. Pressé, K. Ghosh, and K.A. Dill. Modeling stochastic dynamics in biochemical systems with feedback using maximum caliber. *Journal of Physical Chemistry B*, 115:6202–6212, 2011.

[137] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables.* Dover, New York, 1972.

[138] A. Clauset, M. Young, and K.S. Gleditsch. On the frequency of severe terrorist events. *J. Conflict Resolution*, 51:58, 2007.

[139] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[140] M. Milakovic. A statistical equilibrium model of wealth distribution. *Computing in Economics and Finance*, 214, 2001.

[141] John Harte. *Maximum entropy and ecology: a theory of abundance, distribution, and energetics.* Oxford University Press, 2011.

[142] B. Mandelbrot. An informational theory of the statistical structure of language. In W. Jackson, editor, *Communication Theory*, pages 486–502. Butterworth, Woburn, MA, 1953.

[143] M. Lynch and A. Force. The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154:459–473, 2000.

[144] M.L. Siegal and A. Bergman. Waddington's canalization revisited: developmental stability and evolution. *Proc. Natl. Acad. Sci. USA*, 99(16):10528–10532, 2002.

[145] A. Bergman and M.L. Siegal. Evolutionary capacitance as a general feature of complex gene networks. *Nature*, 424(6948):549–552, 2003.

[146] A. Wagner and J. Wright. Alternative routes and mutational robustness in complex regulatory networks. *BioSystems*, 88:163–172, 2007.

[147] T. MacCarthy and A. Bergman. The limits of subfunctionalization. *BMC Evolutionary Biology*, 7:213, 2007.

[148] S. Scherer. *A Short Guide to the Human Genome.* Cold Spring Harbor, NY, 2008.

[149] K.P. Byrne and K.H. Wolfe. The yeast gene order browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.*, 15:1456–1461, 2005.

[150] J.W. Drake, B. Charlesworth, D. Charlesworth, and J.F. Crow. Rates of spontaneous mutation. *Genetics*, 148:1667–1686, 1998.

[151] J.D. Han, N. Bertin, T. Hao, D.S. Goldberg, G.F. Berriz, L.V. Zhang, D. Dupuy, A.J. Walhout, M.E. Cusick, F.P. Roth, and M. Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88–93, 2004.

[152] P.M. Kim, L.J. Lu, Y. Xia, and M.B. Gerstein. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, 314(5807):1938–1941, 2006.

[153] D. Ekman, S. Light, A.K. Bjorklund, and A. Elofsson. What properties characterize the hub proteins of the protein-protein interaction network of *Saccharomyces cerevisiae*? *Genome Biol.*, 7(6):R45, 2006.

[154] G.P. Singh, M. Ganapathi, and D. Dash. Role of intrinsic disorder in transient interactions of hub proteins. *Proteins*, 66(4):761–765, 2007.

[155] J.V. Moran, R.J. DeBardinis, and H.H. Kazazian. Exon shuffling by L1 retrotransposition. *Science*, 283:1530–1534, 1999.

[156] L. Patthy. Genome evolution and the evolution of exon-shuffling – a review. *Gene*, 238:103–114, 1999.

[157] K. Evlampiev and H. Isambert. Modeling protein network evolution under genome duplication and domain shuffling. *BMC Syst. Biol.*, 1:49, 2007.

[158] D.B. Cancherini, G.S. Franca, and S.J. de Souza. The role of exon shuffling in shaping protein-protein interaction networks. *BMC Genomics*, 11(Suppl 5):S11, 2010.

[159] Neil A. Campbell, Jane B. Reece, and Lawrence G. Mitchell. *Biology*. Benjamin/Cummings, Menlo Park, CA, 5th edition, 1999.

[160] C. Stark, B.J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, 34:D535–9, 2006.

[161] J. Shlens. A tutorial on principal component analysis. 2009.

[162] R. Pastor-Satorras, E. Smith, and R.V. Solé. Evolving protein interaction networks through gene duplication. *Journal of Theoretical Biology*, 222:199–210, 2003.

[163] M. Middendorf, E. Ziv, and C.H. Wiggins. Inferring network mechanisms: the *Drosophila melanogaster* protein interaction network. *Proc. Natl. Acad. Sci. USA*, 102(9):3192–3197, 2005.

[164] M. O'Connor and M. Lawrence. Judgmental forecasting and the use of available information. In G. Wright and P. Goodwin, editors, *Forecasting with Judgment*. Wiley, Chichester, 1998.

[165] M. Hibon and S. Makridakis. The M-3 competition: results, conclusions, and implications. *International Journal of Forecasting*, 16:461–476, 2000.

[166] A. Baldassarri. *Statistics of persistent extreme events*. PhD thesis, Université de Paris-Sud U.F.R. Scientifique d'Orsay, Paris, France, 1999.

[167] L. Onsager. Crystal statistics. I. A two-dimensional model with an order-disorder transition. *Physical Review*, 65:114–149, 1944.

[168] B. Kaufman. Crystal statistics. II. Partition function evaluated by spinor analysis. *Physical Review*, 76:1232–1243, 1949.

[169] G. Sobotta. Transfer tensor method in statistical mechanics. *Physica A*, 129:343–359, 1985.

[170] M.E. Kilmer, C.D. Martin, and L. Perrone. A third-order generalization of the matrix SVD as a product of third-order tensors. Technical Report TR-2008-4, Tufts University Department of Computer Science, Medford, MA, October 2008.

[171] J. Adler. Critical temperatures of the d=3 s=1/2 Ising model: the effect of confluent corrections to scaling. *Journal of Physics A*, 16:3585–3599, 1983.

[172] F. Livet. The cluster updating Monte-Carlo algorithm applied to the 3d Ising problem. *Europhysics Letters*, 16:139–142, 1991.

[173] H.W.J. Blöte and G. Kamieniarz. Simulations of the 3d Ising model at criticality. *Acta Physica Polonica A*, 85:395–398, 1994.

[174] H.W.J. Blöte, E. Luijten, and J.R. Heringa. Ising universality in three dimensions: a Monte Carlo study. *Journal of Physics A*, 28:6289–6313, 1995.

A Cost-sharing in growing network models

The independent and social particles frameworks are similar to two well-known models of growing networks, *random* and *preferential attachment* networks. Here, we consider the case of *directed* edges (distinguishing between outgoing and incoming edges), although our results are easily adapted to undirected networks.

A.1 Random networks

In a growing random network, we begin with a single node, and at every time step, a new node is added to the network, and forms a single edge to a previously existing node. There is no limit to the number of edges that a node may acquire, and each node is equally likely to receive an edge from the newly added node. The probability per time step (q) of any node to acquire one additional edge is therefore

$$q = \frac{1}{N}, \tag{62}$$

where N, the total number of nodes, increases as we add nodes to the network. q is referred to as the *connection* or *attachment kernel*.

The probability $p_{k,N}$ that a node will have degree k when there are N total nodes in the network is

$$p_{k,N} \equiv \frac{\text{\# nodes with } k \text{ edges}}{N}. \tag{63}$$

To determine $p_{k,N}$, we will write a difference equation describing the flows into and out of a bin of nodes containing k incoming edges [4]. The total number of nodes with k edges is $Np_{k,N}$. If each new node makes 1 outgoing edge, and 1 node is added per time step, then the number of nodes moving from the k bin to the $k+1$ bin is $Nqp_{k,N} = p_{k,N}$. The number of nodes moving from the $k-1$ bin to the k bin is

$Nqp_{k-1,N} = p_{k-1,N}$. Summing these contributions, we find

$$(N+1)p_{k,N+1} = Np_{k,N} + p_{k-1,N} - p_{k,N}, \tag{64}$$

which rearranges to:

$$Np_{k,N+1} - Np_{k,N} = -p_{k,N+1} + p_{k-1,N} - p_{k,N}. \tag{65}$$

Dividing both sides by N, we see that the difference between $p_{k,N+1}$ and $p_{k,N}$ decays as $1/N$:

$$p_{k,N+1} - p_{k,N} = \frac{1}{N}\big(- p_{k,N+1} + p_{k-1,N} - p_{k,N}\big). \tag{66}$$

Therefore, for very large networks ($N \to \infty$),

$$\lim_{N\to\infty} \big(p_{k,N+1} - p_{k,N}\big) = 0, \tag{67}$$

and $p_{k,N}$ reaches a steady-state, which we will denote by p_k. This is referred to as the *degree distribution*.

At steady-state, Eq. 64 simplifies to

$$p_k = p_{k-1} - p_k, \tag{68}$$

which can be rearranged to give a recursion relation between p_k and p_{k-1}:

$$p_k = \frac{1}{2}\, p_{k-1}. \tag{69}$$

Recursively substituting in smaller and smaller k values into Eq. 69 yields

$$p_k = \left(\frac{1}{2}\right)^k p_0. \tag{70}$$

A single node enters the p_0 bin at each time step by definition:

$$p_0 = 1 - p_0 = \frac{1}{2}. \tag{71}$$

Putting this all together, we see that the probability distribution decays exponentially with k,

$$p_k = 2^{-(k+1)}. \tag{72}$$

Eq. (72) is written as a product of independent factors. We are free to rewrite each factor of $1/2$ as an *additive* contribution,

$$\frac{1}{2} = \mathrm{e}^{-\ln 2}, \tag{73}$$

so that the potential is:

$$\mu \equiv \ln 2, \tag{74}$$

which is the work required to add a single edge to any node.[14] Because we are adding edges at random (*i.e.*, independent of the degree of the node receiving the edge), μ is independent of k, so that the work required to add k edges onto a node is

$$w_k = \mu k. \tag{75}$$

The optimization condition is therefore given by Eq. 39, so that p_k^* is exponential,

$$p_k^* = Q^{-1}\mathrm{e}^{-\mu k}, \tag{76}$$

and the partition function is:

$$Q = \frac{1}{1 - \mathrm{e}^{-\mu}}. \tag{77}$$

[14] This generalizes in a straightforward way to the case where the entering node creates a fixed number of edges n per time step. In that case, the potential is $\mu = \ln(1 + 1/n)$.

Substituting Eq. 74 into Eq. (77) gives

$$Q = 2, \tag{78}$$

and we recover Eq. 72 for the probability distribution.

A.2 Preferential attachment networks

Another widely-studied class of networks are those in which the probability of a node acquiring a new edge is proportional to the number of edges that is already has [24, 5]. This mechanism is often referred to as *preferential attachment*. Preferential attachment is a specific example of a model where the 'rich get richer', and has a power law degree distribution with a fixed exponent of 3. Although it is not exactly analogous to our social particles model, it leads to similar form of constraint.

In this mechanism, nodes that are 'rich' – *i.e.*, have many edges connecting them to other nodes – become richer, acquiring new edges faster than poorly-connected nodes do. The preferential attachment mechanism gives a power-law distribution for the probabilities, p_k, of the number k of edges attached to a node, with an exponent of 3, $p_k \sim k^{-3}$.

Formally, preferential attachment networks are defined by two things: growth (one node is added per time step, which makes a connection to one of the N pre-existing nodes), and the connection kernel,

$$q_k = \frac{k+1}{2N}, \tag{79}$$

where k is the number of incoming edges.

Our strategy for determining the steady-state ($N \to \infty$) p_k for the preferential attachment mechanism is the same as for the random network. The number of nodes going from the k bin to the $k+1$ bin is

$$Nq_k p_k = \frac{k+1}{2} \cdot p_k. \tag{80}$$

This is the outflow from the k bin. The number of nodes going from the $k-1$ bin to the k bin is

$$Nq_{k-1} p_{k-1} = \frac{k}{2} \cdot p_{k-1}. \tag{81}$$

This is the inflow to the k bin. Therefore we write a recursion equation for p_k,

$$p_k = \frac{k}{k+3} \cdot p_{k-1}, \tag{82}$$

which we may step through to obtain

$$p_k = \frac{k}{k+3} \cdot \frac{k-1}{k+2} \cdots \frac{1}{4} \cdot p_0. \tag{83}$$

Since 1 new node (with 0 incoming edges) is added per time step, the inflow into the $k = 0$ bin is simply 1. The outflow is

$$Nq_0 p_0 = \frac{1}{2} \cdot p_0 \tag{84}$$

which yields:

$$p_0 = \frac{2}{3}. \tag{85}$$

Substituting (85) into (83), we find

$$p_k = 4 \cdot \frac{k!}{(k+3)!} = \frac{4}{(k+1)(k+2)(k+3)}. \tag{86}$$

These are the multiplicative contributions to p_k. For large k, this scales as

$$p_k \sim k^{-3}. \tag{87}$$

PA gives an exponent of 3 because the attachment probability is equal to the node's degree (k), divided by the total number of edges in the network (normalization). So, in the recursion equation (Eq. 83), the denominator is always equal to the numerator + 3. Therefore, solving the recursion explicitly, there are exactly 3 terms in the denominator which depend on k that don't cancel out: $(k+1)(k+2)(k+3)$. For large k, the + 1,2,3 is negligible, and so the scaling is $p_k \sim k^{-3}$. Thus, the exponent of 3 always appears when the attachment rule is a linear proportionality.

The multiplicative terms in Eq. (83) may be re-written as:

$$p_k = p_0 \exp\left[-\sum_{j=1}^{k} \ln\left(1+\frac{3}{j}\right)\right]. \tag{88}$$

Therefore, the work required to add the j^{th} edge to a node that already has $j-1$ edges is

$$\mu_j = \ln\left(1+\frac{3}{j}\right). \tag{89}$$

The total work w_k to add k edges to a node is calculated by summing over Eq. (89),

$$w_k = \ln\big[(k+1)(k+2)(k+3)\big] - \ln 6, \tag{90}$$

which, similar to the w_k derived for social particles, is logarithmic. For large k, Eq. 91 is approximately

$$w_k \approx 3\left[\ln(k+2) - \frac{\ln 6}{3}\right], \tag{91}$$

leading to the fixed scaling exponent of 3.

B Sampling with double replacement

Here we discuss a simple statistical problem which leads to power-law distributions. Although the problem is described as picking balls from a barrel, the following derivation applies equally well to the formation of edges between nodes on a graph.

If we start with a barrel with a mixture of red (r) and blue (b) balls in a barrel, our initial probability of drawing a blue ball will be $b/(b+r)$. The standard 'sampling with replacement' rule gives, for the probability of drawing k blue balls in a row,

$$p(k) = \left(1 + \frac{r}{b}\right)^{-k}, \tag{92}$$

which decays exponentially because each selection is independent. However, if every time a ball is picked an *extra* ball of that color is placed into the barrel, the probabilities are no longer independent. Instead, this 'sampling with double replacement' rule gives

$$p(k) = \frac{(b+r-1)!}{(b-1)!} \cdot \frac{(k+b-1)!}{(k+b+r-1)!}. \tag{93}$$

How does Eq. 93 scale for relatively large values of k? First, apply Stirling's approximation to the factorials containing k,

$$\frac{(k+b-1)!}{(k+b+r-1)!} \approx \mathrm{e}^{r} \left(\frac{k+b+r-1}{k+b-1}\right)^{1-k-b} (k+b+r-1)^{-r}, \tag{94}$$

then use the limit

$$\lim_{k\to\infty} \left(\frac{k+b+r-1}{k+b-1}\right)^{1-k-b} = \mathrm{e}^{-r} \tag{95}$$

to obtain:

$$p(k) \approx \frac{(b+r-1)!}{(b-1)!} \cdot (k+b+r-1)^{-r}. \tag{96}$$

If we assume that the barrel initially contains the same number of red and blue balls,

$r = b = n$, this expression simplifies to

$$p(k) \approx \frac{(2n-1)!}{(n-1)!} \cdot (k + 2n - 1)^{-r} \,. \tag{97}$$

where the scaling form is valid for the probability of picking many balls of the same color in a row, $k \gg n$.

This is just the probability of picking k balls of the same color in a row. What is the *joint* distribution $p(k_b, k_r)$? First, note that each selection is a Bernoulli trial, where the possible outcomes are blue or red. The *order* of selection does not affect the total probability. For example, if four balls have been picked, the probability of three blue balls ($k_b = 3$) and one red ball ($k_r = 1$) does not depend on the order in which they are picked. There are four total permutations of the four trials that result in the state $k_b = 3$, $k_r = 1$:

$$p(b, b, b, r) = p(b, b, r, b) = p(b, r, b, b) = p(r, b, b, b).$$

The sequences by which we can arrive at this state are mutually exclusive, so the total probability of this state is the sum of the four permutations,

$$p(k_b = 3, k_r = 1) = p(b, b, b, r) + p(b, b, r, b) + p(b, r, b, b) + p(r, b, b, b) = 4 \cdot p(b, b, b, r).$$

This rule holds for arbitrary values of k_b and k_r. In general, for $k_b + k_r$ balls picked, there are

$$\begin{pmatrix} k_b + k_r \\ k_b \end{pmatrix} = \frac{(k_b + k_r)!}{k_b! k_r!}$$

equivalent permutations. Since the permutations are equal, we may pick any one we like to calculate $p(k_b, k_r)$:

$$\begin{aligned} p(k_b, k_r) &= \begin{pmatrix} k_b + k_r \\ k_b \end{pmatrix} \frac{b(b+1)\cdots(b+k_b-1)\, r(r+1)\cdots(r+k_r-1)}{(b+r)(b+r+1)\cdots(b+r+k_b+k_r-1)} \\ &= \frac{(k_b+k_r)!}{k_b!k_r!} \cdot \frac{(b+k_b-1)!}{(b-1)!} \cdot \frac{(r+k_r-1)!}{(r-1)!} \cdot \frac{(b+r-1)!}{(b+r+k_b+k_r-1)!}. \end{aligned}$$

This may be rearranged to the more intuitive form:

$$p(k_b, k_r) = \frac{\begin{pmatrix} b+k_b-1 \\ k_b \end{pmatrix} \begin{pmatrix} r+k_r-1 \\ k_r \end{pmatrix}}{\begin{pmatrix} b+r+k_b+k_r-1 \\ k_b+k_r \end{pmatrix}}. \tag{98}$$

This is the Bose-Einstein distribution [166].

In general, if there are N different color balls, with initial numbers $n_1, n_2, \ldots, n_N$, we may calculate the multivariate $p(k_1, k_2, \ldots, k_N)$ using the same strategy as in the two-color case:

$$p(k_1, k_2, \ldots, k_N) = \frac{\prod_i \begin{pmatrix} n_i + k_i - 1 \\ k_i \end{pmatrix}}{\begin{pmatrix} \sum_i (n_i + k_i - 1) \\ \sum_i k_i \end{pmatrix}}. \tag{99}$$

C Critical points of Ising models

A well-known source of power-laws in statistical physics are those power-laws associated with critical phenomena. These power-laws manifest as divergences in various thermodynamic quantities, and are somewhat different than the power-laws discussed thus far. One particularly well-known model of criticality is the Ising model. Ising models were originally designed as a simplified models of ferromagnetics, and they exhibit spontaneous magnetization in higher dimensions. They have been exhaustively studied by physicists over the past century, and have been a useful model for studying critical phenomena and universality. Although there are already very accurate ways to calculate the critical points of Ising models, here we present a novel method of locating critical points, using the decomposition of higher-order tensors.

We consider lattices where particles take one of two spin values ($s = +1$ or $s = -1$). We generalize a recently developed singular value decomposition for third-order tensors to fourth- and sixth-order tensors, and apply this method to the problem of finding critical points in Ising models. We calculate the exact critical temperature of a 2-D square lattice by decomposing its fourth-order transfer tensor, and compute upper and lower bounds on the critical temperature for the 3-D cubic lattice by decomposing its sixth-order transfer tensor. This method also provides a novel approximation of the 3-D critical temperature, which is accurate to 3 decimal points.

C.1 Linear chain

The Hamiltonian ($\mathcal{H}$) for the 1-D Ising model, in the absence of an applied magnetic field, is

$$\mathcal{H} = -J \sum_{i=1}^{N} s_i s_{i+1}, \tag{100}$$

where N is the length of the chain and J is the coupling constant. The partition function is given by

$$Q = \sum_{\{s\}=\pm 1} \mathrm{e}^{-\mathcal{H}/(k_B T)} = \sum_{\{s\}=\pm 1} \prod_{i=1}^{N} \mathrm{e}^{K s_i s_{i+1}}, \tag{101}$$

where we have defined $K \equiv J/(k_B T)$. The energy of a single lattice site (site i) is given by

$$E(s_i, s_{i+1}) = K s_i s_{i+1}. \tag{102}$$

It is straightforward to calculate Q using a 2×2 *transfer matrix* (A), which contains all combinations of $s_i = \pm 1$ and $s_{i+1} = \pm 1$ in its rows and columns:

$$A = \begin{bmatrix} \mathrm{e}^{E(1,1)} & \mathrm{e}^{E(1,-1)} \\ \mathrm{e}^{E(-1,1)} & \mathrm{e}^{E(-1,-1)} \end{bmatrix} = \begin{bmatrix} \mathrm{e}^{K} & \mathrm{e}^{-K} \\ \mathrm{e}^{-K} & \mathrm{e}^{K} \end{bmatrix}. \tag{103}$$

The partition function can then be evaluated by matrix multiplication,

$$Q = \begin{bmatrix} 1 & 1 \end{bmatrix} A^N \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \tag{104}$$

where $\begin{bmatrix} 1 & 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ are boundary vectors. The eigenvalues of A are:

$$\lambda_1 = 2\cosh K, \tag{105}$$

$$\lambda_2 = 2\sinh K. \tag{106}$$

Note that $\lambda_1 > \lambda_2$ for all temperatures (Fig. S1).

Assuming periodic boundary conditions, the partition function can be calculated

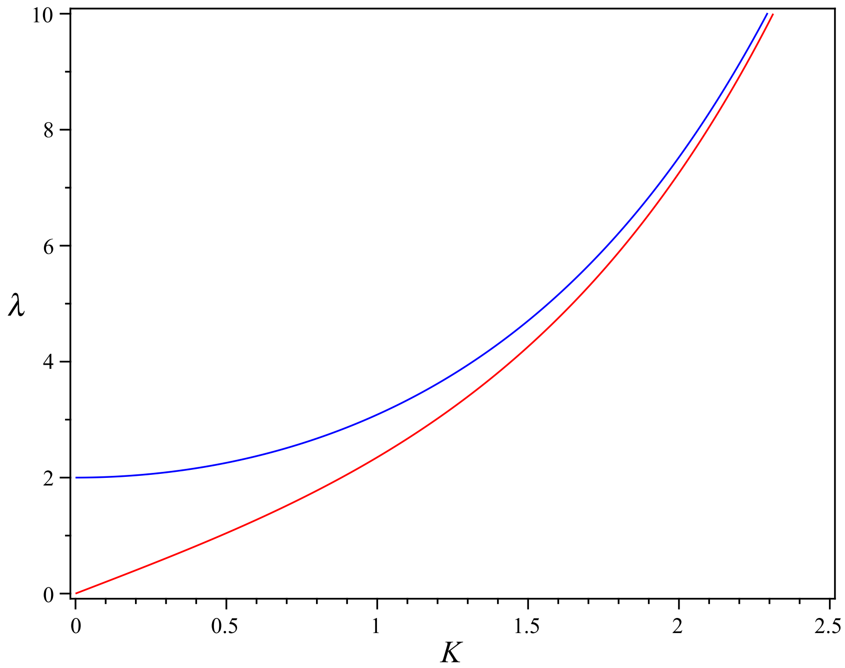


Figure S1: Transfer matrix eigenvalues for a linear (1-D) Ising model. λ_1 (blue) is greater than λ_2 (red) for all values of K.

directly from the eigenvalues,

$$Q = \lambda_1^N + \lambda_2^N. \tag{107}$$

Because $\lambda_1 > \lambda_2$, as N becomes very large, Q is determined by λ_1:

$$\lim_{N\to\infty} Q = \lambda_1^N. \tag{108}$$

The second derivatives of $\ln Q$ are continuous. Therefore, the 1-D model lacks a critical point.

C.2 Square lattice

A 2-D square ($N \times N$) lattice has Hamiltonian

$$\mathcal{H} = -J \sum_{i,j} s_{i,j} \left(s_{i+1,j} + s_{i,j+1} \right). \tag{109}$$

Similar to 1-D, the 2-D partition function can be evaluated by diagonalizing its transfer matrix [167, 168]. A complication is that, in 2-D, the size of the transfer matrix grows exponentially ($2^N \times 2^N$) with the size of the lattice.

One way to avoid this exponential growth in matrix size is to instead use a higher-order *transfer tensor* [169], so that the 2-D model is represented by a fourth-order ($2 \times 2 \times 2 \times 2$) tensor, 3-D by a sixth-order ($2 \times 2 \times 2 \times 2 \times 2 \times 2$) tensor, and so on. This technique has been limited by the lack of analytical diagonalization techniques for higher-order tensors. However, an equivalent of the matrix singular value decomposition (SVD) was recently invented for third-order tensors [170]. We show here that this technique generalizes in a straightforward way to tensors of fourth and higher order.

We construct a $2 \times 2 \times 2 \times 2$ transfer tensor $\mathcal{A}$ as described in [169]. The sites in the transformed lattice are distinguished from the original sites by half-integer indices. The four tensor indices represent all possible spin states (±1) for the four transformed lattice sites ($s_{i+\frac{1}{2},j}$, $s_{i-\frac{1}{2},j}$, $s_{i,j+\frac{1}{2}}$, and $s_{i,j-\frac{1}{2}}$). Using the transformed lattice, the energy of site $s_{i,j}$ is given by

$$E(s_{i+\frac{1}{2},j}, s_{i-\frac{1}{2},j}, s_{i,j+\frac{1}{2}}, s_{i,j-\frac{1}{2}}) = K\big(s_{i,j-\frac{1}{2}} + s_{i,j+\frac{1}{2}}\big)\big(s_{i-\frac{1}{2},j} + s_{i+\frac{1}{2},j}\big). \tag{110}$$

The four 2 × 2 matrices which form the 'faces' of $\mathcal{A}$ are

$$A_1 = \begin{bmatrix} e^{E(1,1,1,1)} & e^{E(1,-1,1,1)} \\ e^{E(-1,1,1,1)} & e^{E(-1,-1,1,1)} \end{bmatrix} = \begin{bmatrix} e^{4K} & 1 \\ 1 & e^{-4K} \end{bmatrix}, \tag{111}$$

$$A_2 = \begin{bmatrix} e^{E(1,1,-1,1)} & e^{E(1,-1,-1,1)} \\ e^{E(-1,1,-1,1)} & e^{E(-1,-1,-1,1)} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \tag{112}$$

$$A_3 = \begin{bmatrix} e^{E(1,1,1,-1)} & e^{E(1,-1,1,-1)} \\ e^{E(-1,1,1,-1)} & e^{E(-1,-1,1,-1)} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \tag{113}$$

$$A_4 = \begin{bmatrix} e^{E(1,1,-1,-1)} & e^{E(1,-1,-1,-1)} \\ e^{E(-1,1,-1,-1)} & e^{E(-1,-1,-1,-1)} \end{bmatrix} = \begin{bmatrix} e^{-4K} & 1 \\ 1 & e^{4K} \end{bmatrix}. \tag{114}$$

The 2-D partition function can be calculated by tensor multiplication, with appropriate boundary conditions. Unfortunately, this product appears to be analytically intractible as $N \to \infty$, preventing direct evaluation of the partition function. However, partial diagonalization of $\mathcal{A}$ is sufficient to locate the critical point, as shown below.

Kilmer's SVD for third-order tensors can be recursively generalized to fourth-order tensors. We will use a tilde to denote tensors which have been 'unfolded' as described in [170]. First, $\mathcal{A}$ is unfolded along its fourth index, giving us a 4 × 4 × 2 matrix-of-third-order-tensors, with front face given by

$$\widetilde{\mathcal{A}}_1 = \begin{bmatrix} A_1 & A_3 \\ A_3 & A_1 \end{bmatrix}, \tag{115}$$

and back face:

$$\widetilde{\mathcal{A}}_2 = \begin{bmatrix} A_2 & A_4 \\ A_4 & A_2 \end{bmatrix}. \tag{116}$$

Next, the third-order tensors $\widetilde{\mathcal{A}}_1$ and $\widetilde{\mathcal{A}}_2$ are individually unfolded, resulting in an

8×8 matrix,

$$\begin{aligned}\tilde{\tilde{\mathcal{A}}} &= \begin{bmatrix} A_1 & A_2 & A_3 & A_4 \\ A_2 & A_1 & A_4 & A_3 \\ A_4 & A_4 & A_1 & A_2 \\ A_4 & A_3 & A_2 & A_1 \end{bmatrix} \\ &= \begin{bmatrix} \mathrm{e}^{4K} & 1 & 1 & 1 & 1 & 1 & \mathrm{e}^{-4K} & 1 \\ 1 & \mathrm{e}^{-4K} & 1 & 1 & 1 & 1 & 1 & \mathrm{e}^{4K} \\ 1 & 1 & \mathrm{e}^{4K} & 1 & \mathrm{e}^{-4K} & 1 & 1 & 1 \\ 1 & 1 & 1 & \mathrm{e}^{-4K} & 1 & \mathrm{e}^{4K} & 1 & 1 \\ 1 & 1 & \mathrm{e}^{-4K} & 1 & \mathrm{e}^{4K} & 1 & 1 & 1 \\ 1 & 1 & 1 & \mathrm{e}^{4K} & 1 & \mathrm{e}^{-4K} & 1 & 1 \\ \mathrm{e}^{-4K} & 1 & 1 & 1 & 1 & 1 & \mathrm{e}^{4K} & 1 \\ 1 & \mathrm{e}^{4K} & 1 & 1 & 1 & 1 & 1 & \mathrm{e}^{-4K} \end{bmatrix} . \end{aligned} \tag{117}$$

Diagonalizing $\tilde{\tilde{\mathcal{A}}}$ yields 8 eigenvalues:

$$\lambda_1 = 2\left(\cosh 4K + 3\right), \tag{118}$$

$$\lambda_2 = \lambda_3 = \lambda_4 = 2\left(\cosh 4K - 1\right), \tag{119}$$

$$\lambda_5 = \lambda_7 = \sinh 4K, \tag{120}$$

$$\lambda_6 = \lambda_8 = -\sinh 4K. \tag{121}$$

If re-folded, the eigenvalue matrix becomes a fourth-order tensor that is diagonal on each face, but is not fully diagonal (*i.e.*, all zeros except $\mathcal{A}_{1111}$ and $\mathcal{A}_{2222}$).

The eigenvalues weight the contributions of the eigenvectors to the partition function, so the dominant contribution is from the largest eigenvalue. However, in this

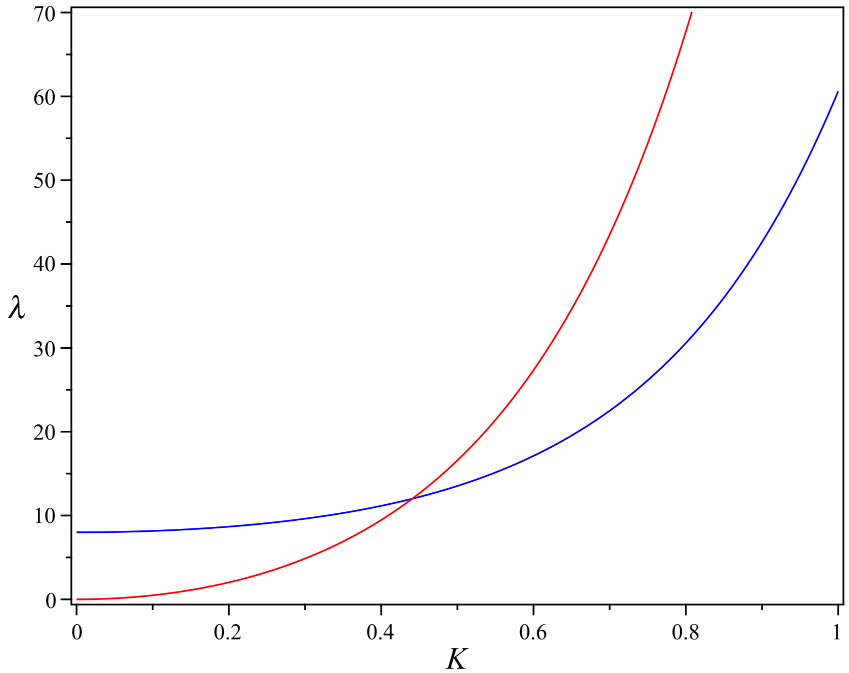


Figure S2: Largest eigenvalues for a square (2-D) Ising model. The intersection of λ_1 (blue) and $3\lambda_2$ (red) gives the exact value of the critical temperature, $K_c = \frac{1}{2}\ln(1+\sqrt{2}) \approx 0.4407$.

case, three smaller eigenvalues are equal ($\lambda_2 = \lambda_3 = \lambda_4$). At low temperatures, their net contribution ($3\lambda_2$) is larger than that of λ_1 (Fig. S2). The critical point is at the intersection of λ_1 and $3\lambda_2$. Setting $\lambda_1 = 3\lambda_2$ and solving for K, we recover the exact critical temperature,

$$K_c = \frac{1}{2}\ln\left(1+\sqrt{2}\right) \approx 0.4407. \tag{122}$$

C.3 Cubic lattice

The Hamiltonian of the 3-D cubic ($N \times N \times N$) lattice is

$$\mathcal{H} = -J\sum_{i,j,k} s_{i,j,k}\left(s_{i+1,j,k} + s_{i,j+1,k} + s_{i,j,k+1}\right). \tag{123}$$

We construct a $2 \times 2 \times 2 \times 2 \times 2 \times 2$ transfer tensor $\mathcal{A}$ using the same strategy as before. The six indices represent all possible spin states for the six transformed lattice sites. The energy of site $s_{i,j,k}$ is given by

$$E(s_{i,j-\frac{1}{2},k}, s_{i-\frac{1}{2},j,k}, s_{i,j,k-\frac{1}{2}}, s_{i,j+\frac{1}{2},k}, s_{i+\frac{1}{2},j,k}, s_{i,j,k+\frac{1}{2}}) = \tag{124}$$

$$K\Big\{s_{i,j-\frac{1}{2},k}\big(s_{i-\frac{1}{2},j,k} + s_{i,j,k-\frac{1}{2}} + s_{i+\frac{1}{2},j,k} + s_{i,j,k+\frac{1}{2}}\big) + s_{i-\frac{1}{2},j,k}\big(s_{i,j,k-\frac{1}{2}} + s_{i,j+\frac{1}{2},k} + s_{i,j,k+\frac{1}{2}}\big) \tag{125}$$

$$+ s_{i,j,k-\frac{1}{2}}\big(s_{i,j+\frac{1}{2},k} + s_{i+\frac{1}{2},j,k}\big) + s_{i,j+\frac{1}{2},k}\big(s_{i+\frac{1}{2}} + s_{i,j,k+\frac{1}{2}}\big) + s_{i+\frac{1}{2},j,k}s_{i,j,k+\frac{1}{2}}\Big\}. \tag{126}$$

There are sixteen 2 × 2 'faces' of $\mathcal{A}$:

$$A_1 = \begin{bmatrix} \mathrm{e}^{12K} & \mathrm{e}^{4K} \\ \mathrm{e}^{4K} & 1 \end{bmatrix} A_2 = \begin{bmatrix} \mathrm{e}^{4K} & 1 \\ 1 & 1 \end{bmatrix} A_3 = \begin{bmatrix} \mathrm{e}^{4K} & 1 \\ \mathrm{e}^{-4K} & \mathrm{e}^{-4K} \end{bmatrix} A_4 = \begin{bmatrix} 1 & 1 \\ \mathrm{e}^{-4K} & 1 \end{bmatrix} \tag{127}$$

$$A_5 = \begin{bmatrix} \mathrm{e}^{4K} & \mathrm{e}^{-4K} \\ 1 & \mathrm{e}^{-4K} \end{bmatrix} A_6 = \begin{bmatrix} 1 & \mathrm{e}^{-4K} \\ 1 & 1 \end{bmatrix} A_7 = \begin{bmatrix} 1 & \mathrm{e}^{-4K} \\ \mathrm{e}^{-4K} & \mathrm{e}^{-4K} \end{bmatrix} A_8 = \begin{bmatrix} 1 & 1 \\ 1 & \mathrm{e}^{4K} \end{bmatrix}$$

$$A_9 = \begin{bmatrix} \mathrm{e}^{4K} & 1 \\ 1 & 1 \end{bmatrix} A_{10} = \begin{bmatrix} \mathrm{e}^{-4K} & \mathrm{e}^{-4K} \\ \mathrm{e}^{-4K} & 1 \end{bmatrix} A_{11} = \begin{bmatrix} 1 & 1 \\ \mathrm{e}^{-4K} & 1 \end{bmatrix} A_{12} = \begin{bmatrix} \mathrm{e}^{-4K} & 1 \\ \mathrm{e}^{-4K} & \mathrm{e}^{4K} \end{bmatrix}$$

$$A_{13} = \begin{bmatrix} 1 & \mathrm{e}^{-4K} \\ 1 & 1 \end{bmatrix} A_{14} = \begin{bmatrix} \mathrm{e}^{-4K} & \mathrm{e}^{-4K} \\ 1 & \mathrm{e}^{4K} \end{bmatrix} A_{15} = \begin{bmatrix} 1 & 1 \\ 1 & \mathrm{e}^{4K} \end{bmatrix} A_{16} = \begin{bmatrix} 1 & \mathrm{e}^{4K} \\ \mathrm{e}^{4K} & \mathrm{e}^{12K} \end{bmatrix}$$

Kilmer's SVD generalizes in a straightforward way to sixth-order tensors. After four recursive unfoldings, $\mathcal{A}$ is transformed into a 32×32 matrix,

$$\begin{bmatrix}
A_1 & A_2 & A_3 & A_4 & A_5 & A_6 & A_7 & A_8 & A_9 & A_{10} & A_{11} & A_{12} & A_{13} & A_{14} & A_{15} & A_{16} \\
A_2 & A_1 & A_4 & A_3 & A_6 & A_5 & A_8 & A_7 & A_{10} & A_9 & A_{12} & A_{11} & A_{14} & A_{13} & A_{16} & A_{15} \\
A_3 & A_4 & A_1 & A_2 & A_7 & A_8 & A_5 & A_6 & A_{11} & A_{12} & A_9 & A_{10} & A_{15} & A_{16} & A_{13} & A_{14} \\
A_4 & A_3 & A_2 & A_1 & A_8 & A_7 & A_6 & A_5 & A_{12} & A_{11} & A_{10} & A_9 & A_{16} & A_{15} & A_{14} & A_{13} \\
A_5 & A_6 & A_7 & A_8 & A_1 & A_2 & A_3 & A_4 & A_{13} & A_{14} & A_{15} & A_{16} & A_9 & A_{10} & A_{11} & A_{12} \\
A_6 & A_5 & A_8 & A_7 & A_2 & A_1 & A_4 & A_3 & A_{14} & A_{13} & A_{16} & A_{15} & A_{10} & A_9 & A_{12} & A_{11} \\
A_7 & A_8 & A_5 & A_6 & A_3 & A_4 & A_1 & A_2 & A_{15} & A_{16} & A_{13} & A_{14} & A_{11} & A_{12} & A_9 & A_{10} \\
A_8 & A_7 & A_6 & A_5 & A_4 & A_3 & A_2 & A_1 & A_{16} & A_{15} & A_{14} & A_{13} & A_{12} & A_{11} & A_{10} & A_9 \\
A_9 & A_{10} & A_{11} & A_{12} & A_{13} & A_{14} & A_{15} & A_{16} & A_1 & A_2 & A_3 & A_4 & A_5 & A_6 & A_7 & A_8 \\
A_{10} & A_9 & A_{12} & A_{11} & A_{14} & A_{13} & A_{16} & A_{15} & A_2 & A_1 & A_4 & A_3 & A_6 & A_5 & A_8 & A_7 \\
A_{11} & A_{12} & A_9 & A_{10} & A_{15} & A_{16} & A_{13} & A_{14} & A_3 & A_4 & A_1 & A_2 & A_7 & A_8 & A_5 & A_6 \\
A_{12} & A_{11} & A_{10} & A_9 & A_{16} & A_{15} & A_{14} & A_{13} & A_4 & A_3 & A_2 & A_1 & A_8 & A_7 & A_6 & A_5 \\
A_{13} & A_{14} & A_{15} & A_{16} & A_9 & A_{10} & A_{11} & A_{12} & A_5 & A_6 & A_7 & A_8 & A_1 & A_2 & A_3 & A_4 \\
A_{14} & A_{13} & A_{16} & A_{15} & A_{10} & A_9 & A_{12} & A_{11} & A_6 & A_5 & A_8 & A_7 & A_2 & A_1 & A_4 & A_3 \\
A_{15} & A_{16} & A_{13} & A_{14} & A_{11} & A_{12} & A_9 & A_{10} & A_7 & A_8 & A_5 & A_6 & A_3 & A_4 & A_1 & A_2 \\
A_{16} & A_{15} & A_{14} & A_{13} & A_{12} & A_{11} & A_{10} & A_9 & A_8 & A_7 & A_6 & A_5 & A_4 & A_3 & A_2 & A_1
\end{bmatrix} \tag{128}$$

which we diagonalize using the discrete Fourier transform method of [170]. The three largest eigenvalues of 128 are

$$\lambda_1 = \mathrm{e}^{12K} + 6\,\mathrm{e}^{4K} + 9\,\mathrm{e}^{-4K} + 16, \tag{129}$$

$$\lambda_2 = \mathrm{e}^{12K} + 2\,\mathrm{e}^{4K} - 3\,\mathrm{e}^{-4K}, \tag{130}$$

$$\lambda_3 = \mathrm{e}^{12K} - 2\,\mathrm{e}^{4K} + \mathrm{e}^{-4K}, \tag{131}$$

where λ_1 is unique, λ_2 is 5-fold degenerate, and λ_3 is 7-fold degenerate.

As shown in Fig. S3, there are three distinct temperature ranges for the cubic lattice. Below the value of K where $\lambda_1 = 5\lambda_2$ (that is, at high temperatures), λ_1 is dominant:

$$K_- = -\frac{1}{36}\ln\left[1 + \frac{1}{2}\left(440 + 18\sqrt{606}\right)^{1/3} - 7\left(440 + 18\sqrt{606}\right)^{-1/3}\right] \approx 0.1436. \tag{132}$$

Above the value of K where $5\lambda_2 = 7\lambda_3$ (at low temperatures), $7\lambda_3$ is the largest eigenvalue:

$$K_+ = \frac{1}{8}\ln(11) \approx 0.2997. \tag{133}$$

For intermediate temperatures ($K_- < K < K_+$), $5\lambda_2$ is dominant. This specifies a *range* where the critical temperature K_c will be located – it is somewhere between K_- and K_+. A simple approximation of K_c is the center of the region bounded by the three eigenvalues. The intersection $\lambda_1 = 7\lambda_3$ is located at

$$K_0 = -\frac{1}{4}\ln\left(-3 + 2\sqrt{3}\right) \approx 0.1919, \tag{134}$$

and the center of mass is therefore:

$$K_c = \frac{\int_{K_-}^{K_0} K\,(5\lambda_2 - \lambda_1)\,dK + \int_{K_0}^{K_+} K\,(5\lambda_2 - 7\lambda_3)\,dK}{\int_{K_-}^{K_0} (5\lambda_2 - \lambda_1)\,dK + \int_{K_0}^{K_+} (5\lambda_2 - 7\lambda_3)\,dK} \approx 0.2212. \tag{135}$$

This rough estimate is quite close to previous calculations of the critical temperature

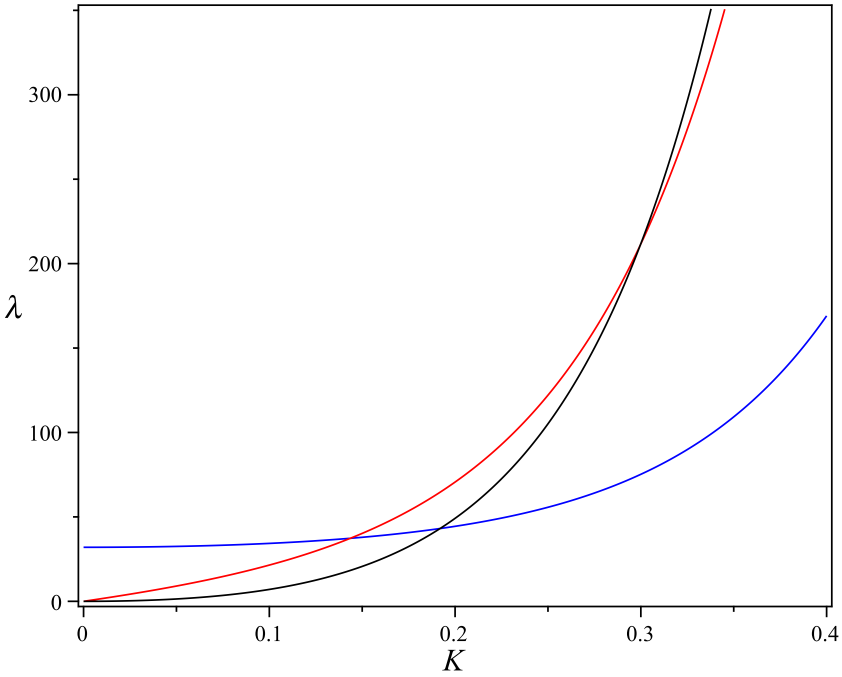


Figure S3: Three largest eigenvalues for a cubic (3-D) Ising model. The intersection of λ_1 (blue) with $5\lambda_2$ (red) gives the lower bound of the range of K where the critical point is located, while the intersection of $5\lambda_2$ (red) with $7\lambda_3$ gives the upper bound. The centroid of the region enclosed by λ_1, $5\lambda_2$, and $7\lambda_3$ provides a reasonably accurate estimate of the critical temperature, $K_c \approx 0.2212$.

($K_c \approx 0.2216$) from series expansions and Monte Carlo simulations [171, 172, 173, 174].

D A two-particle transition matrix

Finally, I will briefly discuss a purely mathematical topic. Although this work is not directly connected to power-laws, I hope that a future application of this framework could be to help model correlated, heavy-tailed dynamical phenomena.

A single particle hopping between two states (u_1 and u_2) starts in some initial state: $u_1(0) = 1$ and $u_2(0) = 0$ if the particle is initially in state 1, or $u_1(0) = 0$ and $u_2(0) = 1$ if the particle is initially in state 2. At each time step, the probability that a particle in state 1 jumps to state 2 is σ_{21}, and the probability that it stays in state 1 is σ_{11}. The probability that a particle in state 2 jumps to state 1 is σ_{12}, and the probability that it stays in state 2 is σ_{22}. After a single time step, we can calculate the particle's probability of being in either state:

$$u_1(1) = \sigma_{11}u_1(0) + \sigma_{12}u_2(0) \tag{136}$$

$$u_2(1) = \sigma_{21}u_1(0) + \sigma_{22}u_2(0) \tag{137}$$

And after two time steps:

$$u_1(2) = \sigma_{11}u_1(1) + \sigma_{12}u_2(1) = \sigma_{11}\left(\sigma_{11}u_1(0) + \sigma_{12}u_2(0)\right) + \sigma_{12}\left(\sigma_{21}u_1(0) + \sigma_{22}u_2(0)\right)$$

$$u_2(2) = \sigma_{21}u_1(1) + \sigma_{22}u_2(1) = \sigma_{21}\left(\sigma_{11}u_1(0) + \sigma_{12}u_2(0)\right) + \sigma_{22}\left(\sigma_{21}u_1(0) + \sigma_{22}u_2(0)\right)$$

Clearly, after a few time steps, these equations are going to turn into a confusing mess. A way of cleaning this up is to write the coefficients as a matrix:

$$\begin{bmatrix} u_1(1) \\ u_2(1) \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \begin{bmatrix} u_1(0) \\ u_2(0) \end{bmatrix}$$

$$\begin{bmatrix} u_1(2) \\ u_2(2) \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \begin{bmatrix} u_1(1) \\ u_2(1) \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}^2 \begin{bmatrix} u_1(0) \\ u_2(0) \end{bmatrix}$$

And for τ time steps,

$$\mathbf{u}_\tau = \mathbf{G}\mathbf{u}_{\tau-1} = \mathbf{G}^\tau \mathbf{u}$$

where unsubscripted $\mathbf{u}$ is the initial state vector, $\mathbf{u}_\tau$ is the state vector at time τ, and $\mathbf{G}$ is the *transition matrix* for this process. Since the total probability of the particle either staying *or* jumping is 1 (those are the only two options available), the columns of $\mathbf{G}$ sum to 1:

$$\sigma_{11} + \sigma_{21} = 1$$

$$\sigma_{12} + \sigma_{22} = 1$$

The partition function, Q, for a single particle after τ time steps is the sum of the elements of the final state vector:

$$Q = u_1(\tau) + u_2(\tau) = \begin{bmatrix} 1 & 1 \end{bmatrix} \mathbf{G}^\tau \mathbf{u}$$

The $\begin{bmatrix} 1 & 1 \end{bmatrix}$ is just a convenient way to 'flatten' $\mathbf{u}_\tau$ into a sum. The statistics of N independent particles may be calculated by multiplying their partition functions together:

$$Q^N = \left(\begin{bmatrix} 1 & 1 \end{bmatrix} \mathbf{G}^\tau \mathbf{u} \right)^N$$

If pairs of particles are coupled, then the problem becomes more complicated. To consider two particles (u and v), consider a table of conditional jump/stay probabilities:

We can use these conditional probabilities to write the equivalent of equations (1)

u	v	$P(u \text{ stays})$	$P(u \text{ jumps})$	$P(v \text{ stays})$	$P(v \text{ jumps})$
1	1	$\sigma_{11\|v=1}$	$\sigma_{21\|v=1}$	$\sigma_{11\|u=1}$	$\sigma_{12\|u=1}$
2	1	$\sigma_{22\|v=1}$	$\sigma_{12\|v=1}$	$\sigma_{11\|u=2}$	$\sigma_{12\|u=2}$
1	2	$\sigma_{11\|v=2}$	$\sigma_{21\|v=2}$	$\sigma_{22\|u=1}$	$\sigma_{21\|u=1}$
2	2	$\sigma_{22\|v=2}$	$\sigma_{12\|v=2}$	$\sigma_{22\|u=2}$	$\sigma_{21\|u=2}$

and (2) for coupled particles:

$$\begin{aligned}
u_1(1) &= \big(\sigma_{11|v=1}u_1(0) + \sigma_{12|v=1}u_2(0)\big)\, v_1(0) + \big(\sigma_{11|v=2}u_1(0) + \sigma_{12|v=2}u_2(0)\big)\, v_2(0) \\
u_2(1) &= \big(\sigma_{21|v=1}u_1(0) + \sigma_{22|v=1}u_2(0)\big)\, v_1(0) + \big(\sigma_{21|v=2}u_1(0) + \sigma_{22|v=2}u_2(0)\big)\, v_2(0)
\end{aligned}$$

$$\begin{aligned}
v_1(1) &= \big(\sigma_{11|u=1}v_1(0) + \sigma_{12|u=1}v_2(0)\big)\, u_1(0) + \big(\sigma_{11|u=2}v_1(0) + \sigma_{12|u=2}v_2(0)\big)\, u_2(0) \\
v_2(1) &= \big(\sigma_{21|u=1}v_1(0) + \sigma_{22|u=1}v_2(0)\big)\, u_1(0) + \big(\sigma_{21|u=2}v_1(0) + \sigma_{22|u=2}v_2(0)\big)\, u_2(0)
\end{aligned}$$

Iterating through more time steps this way would be an exercise in notational slapstick, so we won't. Instead, taking a cue from the single particle transition matrix, we will try to write these equations in matrix form:

$$\begin{bmatrix} u_1(1) \\ u_2(1) \end{bmatrix} = \begin{bmatrix} \sigma_{11|v=1}u_1(0) + \sigma_{12|v=1}u_2(0) & \sigma_{11|v=2}u_1(0) + \sigma_{12|v=2}u_2(0) \\ \sigma_{21|v=1}u_1(0) + \sigma_{22|v=1}u_2(0) & \sigma_{21|v=2}u_1(0) + \sigma_{22|v=2}u_2(0) \end{bmatrix} \begin{bmatrix} v_1(0) \\ v_2(0) \end{bmatrix}$$

$$\begin{bmatrix} v_1(1) \\ v_2(1) \end{bmatrix} = \begin{bmatrix} \sigma_{11|u=1}v_1(0) + \sigma_{12|u=1}v_2(0) & \sigma_{11|u=2}v_1(0) + \sigma_{12|u=2}v_2(0) \\ \sigma_{21|u=1}v_1(0) + \sigma_{22|u=1}v_2(0) & \sigma_{21|u=2}v_1(0) + \sigma_{22|u=2}v_2(0) \end{bmatrix} \begin{bmatrix} u_1(0) \\ u_2(0) \end{bmatrix}$$

These equations can be further structured by writing the *columns* of the matrices in matrix form:

$$\begin{bmatrix} u_1(1) \\ u_2(1) \end{bmatrix} = \left[\begin{bmatrix} \sigma_{11|v=1} & \sigma_{12|v=1} \\ \sigma_{21|v=1} & \sigma_{22|v=1} \end{bmatrix} \begin{bmatrix} u_1(0) \\ u_2(0) \end{bmatrix} \quad \begin{bmatrix} \sigma_{11|v=2} & \sigma_{12|v=2} \\ \sigma_{21|v=2} & \sigma_{22|v=2} \end{bmatrix} \begin{bmatrix} u_1(0) \\ u_2(0) \end{bmatrix} \right] \begin{bmatrix} v_1(0) \\ v_2(0) \end{bmatrix}$$

$$\begin{bmatrix} v_1(1) \\ v_2(1) \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} \sigma_{11|u=1} & \sigma_{12|u=1} \\ \sigma_{21|u=1} & \sigma_{22|u=1} \end{bmatrix} \begin{bmatrix} v_1(0) \\ v_2(0) \end{bmatrix} & \begin{bmatrix} \sigma_{11|u=2} & \sigma_{12|u=2} \\ \sigma_{21|u=2} & \sigma_{22|u=2} \end{bmatrix} \begin{bmatrix} v_1(0) \\ v_2(0) \end{bmatrix} \end{bmatrix} \begin{bmatrix} u_1(0) \\ u_2(0) \end{bmatrix}$$

Therefore, two coupled particles require four 'transition matrices' to describe their time evolution:

$$\mathbf{G}_1 = \begin{bmatrix} \sigma_{11|v=1} & \sigma_{12|v=1} \\ \sigma_{21|v=1} & \sigma_{22|v=1} \end{bmatrix} \qquad \mathbf{G}_2 = \begin{bmatrix} \sigma_{11|v=2} & \sigma_{12|v=2} \\ \sigma_{21|v=2} & \sigma_{22|v=2} \end{bmatrix}$$

$$\mathbf{G}_3 = \begin{bmatrix} \sigma_{11|u=1} & \sigma_{12|u=1} \\ \sigma_{21|u=1} & \sigma_{22|u=1} \end{bmatrix} \qquad \mathbf{G}_4 = \begin{bmatrix} \sigma_{11|u=2} & \sigma_{12|u=2} \\ \sigma_{21|u=2} & \sigma_{22|u=2} \end{bmatrix}$$

This allows us to write equations (3) through (6) in a more compact form:

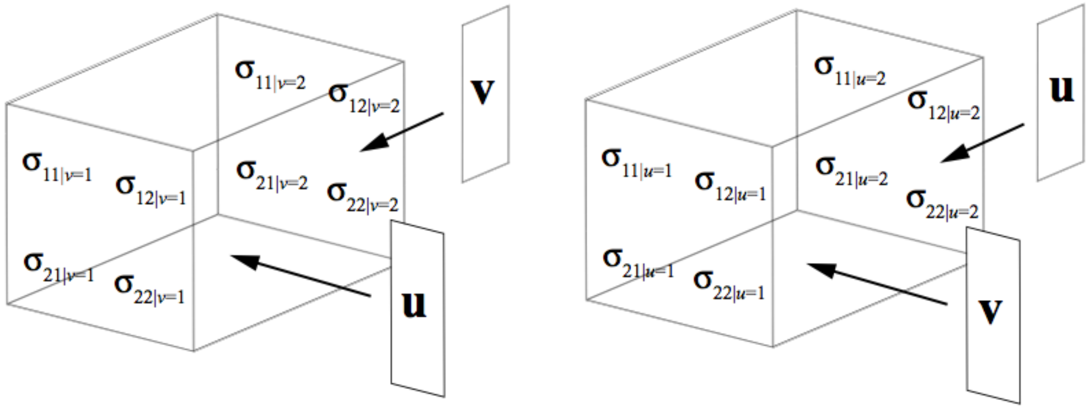
$$\mathbf{u}_1 = \begin{bmatrix} \mathbf{G}_1\mathbf{u} & \mathbf{G}_2\mathbf{u} \end{bmatrix} \mathbf{v} \tag{138}$$

$$\mathbf{v}_1 = \begin{bmatrix} \mathbf{G}_3\mathbf{v} & \mathbf{G}_4\mathbf{v} \end{bmatrix} \mathbf{u} \tag{139}$$

Equations (7) and (8) are multilinear: the vectors **u** and **v** both multiply the transition matrices. To evolve the system for many time steps, it will be useful to break apart the multilinearity, to avoid dealing with equations containing matrices-of-matrices-of-matrices-of... How can we do this? The matrix-of-matrices notation suggests that each matrix is a component of a larger structure. Making use of 1 additional index, we obtain two third-order ($2 \times 2 \times 2$) *transition tensors*:

$\mathbf{G}_1$ and $\mathbf{G}_2$ are the front and back 'faces' of tensor $\mathcal{A}$ (the left 'box' pictured above), and $\mathbf{G}_3$ and $\mathbf{G}_4$ are the front and back 'faces' of tensor $\mathcal{B}$ (the right 'box').

Viewed in tensor form, it is clear how to break apart the multilinearity: the vectors **u** and **v** are simply operated on by different *modes* of the tensors. We see that **u** multiplies mode 1 (rows) of $\mathcal{A}$, and **v** multiplies mode 3 (tubes). **u** multiplies mode 3 of $\mathcal{B}$, and **v** multiplies mode 1. Therefore, we 'slice' $\mathcal{B}$ along mode 2 (columns),



producing a modified pair of transition matrices:

$$\mathbf{G}_3' \equiv \begin{bmatrix} \sigma_{11|u=1} & \sigma_{11|u=2} \\ \sigma_{21|u=1} & \sigma_{21|u=2} \end{bmatrix} \qquad \mathbf{G}_4' \equiv \begin{bmatrix} \sigma_{12|u=1} & \sigma_{12|u=2} \\ \sigma_{22|u=1} & \sigma_{22|u=2} \end{bmatrix}$$

We now have a full set of transition matrices that operate identically on the state vectors $\mathbf{u}$ and $\mathbf{v}$.

It is clear from the image of the tensors that to multiply different modes of the tensors, the vectors $\mathbf{u}$ and $\mathbf{v}$ must be oriented *orthogonally* to each other. Therefore, we can separate out the multilinearity by taking the Kronecker product[15] ($\otimes$) of $\mathbf{v}$ with $\mathbf{u}$,

$$\mathbf{u}_1 = \begin{bmatrix} \mathbf{G}_1\mathbf{u} & \mathbf{G}_2\mathbf{u} \end{bmatrix} \mathbf{v} = \begin{bmatrix} \mathbf{G}_1 & \mathbf{G}_2 \end{bmatrix} (\mathbf{v} \otimes \mathbf{u})$$

$$\mathbf{v}_1 = \begin{bmatrix} \mathbf{G}_3'\mathbf{u} & \mathbf{G}_4'\mathbf{u} \end{bmatrix} \mathbf{v} = \begin{bmatrix} \mathbf{G}_3' & \mathbf{G}_4' \end{bmatrix} (\mathbf{v} \otimes \mathbf{u})$$

For notational simplicity, we will denote the 'unfolded' tensor $\mathcal{A}$ as $\mathbf{A} \equiv \begin{bmatrix} \mathbf{G}_1 & \mathbf{G}_2 \end{bmatrix}$.

[15] Let $\mathbf{A} = (a_{ij})$ be an $n \times n$ matrix and let $\mathbf{B}$ be an $m \times m$ matrix. Then the *Kronecker product* of $\mathbf{A}$ and $\mathbf{B}$ is the $mn \times mn$ block matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{n1}\mathbf{B} & \cdots & a_{nn}\mathbf{B} \end{bmatrix}.$$

The Kronecker product of vectors $\mathbf{v}$ and $\mathbf{u}$ is $\mathbf{v} \otimes \mathbf{u} = \begin{bmatrix} u_1v_1 & u_2v_1 & u_1v_2 & u_2v_2 \end{bmatrix}^T = \text{vec}\left(\mathbf{u}\mathbf{v}^T\right)$.

The unfolded tensor $\mathcal{B}$ is $\mathbf{B} \equiv \begin{bmatrix} \mathbf{G}_3' & \mathbf{G}_4' \end{bmatrix}$. After the next time step:

$$\mathbf{u}_2 = \mathbf{A}\left(\mathbf{v}_1 \otimes \mathbf{u}_1\right) = \mathbf{A}\left(\mathbf{B}\left(\mathbf{v} \otimes \mathbf{u}\right) \otimes \mathbf{A}\left(\mathbf{v} \otimes \mathbf{u}\right)\right) = \mathbf{A}\left(\mathbf{B} \otimes \mathbf{A}\right)\left(\mathbf{v} \otimes \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{u}\right)$$

$$\mathbf{v}_2 = \mathbf{B}\left(\mathbf{v}_1 \otimes \mathbf{u}_1\right) = \mathbf{B}\left(\mathbf{B}\left(\mathbf{v} \otimes \mathbf{u}\right) \otimes \mathbf{A}\left(\mathbf{v} \otimes \mathbf{u}\right)\right) = \mathbf{B}\left(\mathbf{B} \otimes \mathbf{A}\right)\left(\mathbf{v} \otimes \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{u}\right)$$

using the identity $\left(\mathbf{M}_1 \otimes \mathbf{M}_2\right)\left(\mathbf{M}_3 \otimes \mathbf{M}_4\right) = \mathbf{M}_1\mathbf{M}_3 \otimes \mathbf{M}_2\mathbf{M}_4$. After another time step,

$$\mathbf{u}_3 = \mathbf{A}\left(\mathbf{v}_2 \otimes \mathbf{u}_2\right) = \mathbf{A}\left(\mathbf{B} \otimes \mathbf{A}\right)\left(\mathbf{B} \otimes \mathbf{A}\right)^{\otimes 2}\left(\mathbf{v} \otimes \mathbf{u}\right)^{\otimes 4}$$

$$\mathbf{v}_3 = \mathbf{B}\left(\mathbf{v}_2 \otimes \mathbf{u}_2\right) = \mathbf{A}\left(\mathbf{B} \otimes \mathbf{A}\right)\left(\mathbf{B} \otimes \mathbf{A}\right)^{\otimes 2}\left(\mathbf{v} \otimes \mathbf{u}\right)^{\otimes 4}$$

where the exponent $\otimes n$ is used to indicate n Kronecker products. By now the pattern is becoming clear:

$$\mathbf{u}_4 = \mathbf{A}\left(\mathbf{v}_3 \otimes \mathbf{u}_3\right) = \mathbf{A}\left(\mathbf{B} \otimes \mathbf{A}\right)\left(\mathbf{B} \otimes \mathbf{A}\right)^{\otimes 2}\left(\mathbf{B} \otimes \mathbf{A}\right)^{\otimes 4}\left(\mathbf{v} \otimes \mathbf{u}\right)^{\otimes 8}$$

$$\mathbf{v}_4 = \mathbf{B}\left(\mathbf{v}_3 \otimes \mathbf{u}_3\right) = \mathbf{B}\left(\mathbf{B} \otimes \mathbf{A}\right)\left(\mathbf{B} \otimes \mathbf{A}\right)^{\otimes 2}\left(\mathbf{B} \otimes \mathbf{A}\right)^{\otimes 4}\left(\mathbf{v} \otimes \mathbf{u}\right)^{\otimes 8}$$

For τ time-steps:

$$\mathbf{u}_\tau = \mathbf{A}\left(\mathbf{v}_{\tau-1} \otimes \mathbf{u}_{\tau-1}\right) = \mathbf{A}\left(\mathbf{B} \otimes \mathbf{A}\right)\left(\mathbf{B} \otimes \mathbf{A}\right)^{\otimes 2}\left(\mathbf{B} \otimes \mathbf{A}\right)^{\otimes 4} \cdots \left(\mathbf{B} \otimes \mathbf{A}\right)^{\otimes \frac{1}{2}2^{\tau-1}}\left(\mathbf{v} \otimes \mathbf{u}\right)^{\otimes \frac{1}{2}2^{\tau}}$$

$$\mathbf{v}_\tau = \mathbf{B}\left(\mathbf{v}_{\tau-1} \otimes \mathbf{u}_{\tau-1}\right) = \mathbf{B}\left(\mathbf{B} \otimes \mathbf{A}\right)\left(\mathbf{B} \otimes \mathbf{A}\right)^{\otimes 2}\left(\mathbf{B} \otimes \mathbf{A}\right)^{\otimes 4} \cdots \left(\mathbf{B} \otimes \mathbf{A}\right)^{\otimes \frac{1}{2}2^{\tau-1}}\left(\mathbf{v} \otimes \mathbf{u}\right)^{\otimes \frac{1}{2}2^{\tau}}$$

Whether or not this framework will be useful depends on whether an analytical method can be found to calculate repeated Kronecker products. Although this appears to be an intractible problem at the present time, multilinear algebra is still a young and very much active area of mathematics research. If a technique is developed that allows 'Kronecker powers' to be calculated as easily as ordinary matrix powers are today, then this method may prove to be quite useful.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Jack Peterson
Author Signature

2/23/12
Date