Stony Brook University

The Physics of Complex Systems in Information and Biology

A Dissertation Presented

by

Dylan Walker

to

The Graduate School

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Physics

Stony Brook University

December 2008

Stony Brook University

The Graduate School

Dylan Walker

We, the dissertation committee for the above candidate for the Doctor of Philosophy degree, hereby recommend acceptance of this dissertation.

Sergei Maslov – Dissertation Advisor
Adjunct Professor, Department of Physics and Astronomy

Alexander Abanov – Chairperson of Defense
Associate Professor, Department of Physics and Astronomy

Gene Sprouse
Professor, Department of Physics and Astronomy

William Sherman
Assistant Scientist, Center for Functional Nanomaterials, Brookhaven National Laboratory

This dissertation is accepted by the Graduate School.

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

The Physics of Complex Systems in Information and Biology

by

Dylan Walker

Doctor of Philosophy

in

Physics

Stony Brook University

2008

Citation networks have re-emerged as a topic intense interest in the complex networks community with the recent availability of large-scale data sets. The ranking of citation networks is a necessary practice as a means to improve information navigability and search. Unlike many information networks, the aging characteristics of citation networks require the development of new ranking methods. To account for strong aging characteristics of citation networks, we modify the PageRank algorithm by initially distributing random surfers exponentially with age, in favor of more recent publications. The output of this algorithm, which we call CiteRank, is

interpreted as approximate traffic to individual publications in a simple model of how researchers find new information. We optimize parameters of our algorithm to achieve the best performance. The results are compared for two rather different citation networks: all American Physical Society publications between 1893-2003 and the set of high-energy physics theory (hep-th) preprints. Despite major differences between these two networks, we find that their optimal parameters for the CiteRank algorithm are remarkably similar. The advantages and performance of CiteRank over more conventional methods of ranking publications are discussed.

Collaborative voting systems have emerged as an abundant form of real-world, complex information systems that exist in a variety of online applications. These systems are comprised of large populations of users that collectively submit and vote on objects. While the specific properties of these systems vary widely, many of them share a core set of features and dynamical behaviors that govern their evolution. We study a subset of these systems that involve material of a time-critical nature as in the popular example of news items. We consider a general model system in which articles are introduced, voted on by a population of users, and subsequently expire after a proscribed period of time. To study the interaction between popularity and quality, we introduce simple stochastic models of user behavior that approximate differing user quality and susceptibility to the common notion of popular-

ity. We define a metric to quantify user reputation in a manner that is self-consistent, adaptable and content-blind and shows good correlation with the probability that a user behaves in an optimal fashion. We further construct a mechanism for ranking documents that take into account user reputation and provides substantial improvement in the time-critical performance of the system.

The structure of complex systems have been well studied in the context of both information and biological systems. More recently, dynamics in complex systems that occur over the background of the underlying network has received a great deal of attention. In particular, the study of fluctuations in complex systems has emerged as an issue central to understanding dynamical behavior. We approach the problem of collective effects of the underlying network on dynamical fluctuations by considering the protein-protein interaction networks for the system of the living cell. We consider two types of fluctuations in the mass-action equilibrium in protein binding networks. The first type is driven by relatively slow changes in total concentrations (copy numbers) of interacting proteins. The second type, to which we refer to as spontaneous, is caused by quickly decaying thermodynamic deviations away from the mass-action equilibrium of the system. As such they are amenable to methods of equilibrium statistical mechanics used in our study. We investigate the effects of network connectivity on these fluctuations by comparing them to different scenarios in which the interacting

pair is isolated form the rest of the network. Such comparison allows us to analytically derive upper and lower bounds on network fluctuations. The collective effects are shown to sometimes lead to relatively large amplification of spontaneous fluctuations as compared to the expectation for isolated dimers. As a consequence of this, the strength of both types of fluctuations is positively correlated with the overall network connectivity of proteins forming the complex. On the other hand, the relative amplitude of fluctuations is negatively correlated with the equilibrium concentration of the complex. Our general findings are illustrated using a curated network of protein-protein interactions and multi-protein complexes in bakers yeast with experimentally determined protein concentrations.

This thesis is dedicated to my wife, Kate, and my daughter, Madison.

Let the new sun rise and give hope where it once was forgotten.

Contents

List of Figures

List of Tables

Acknowledgements

This dissertation represents the culmination of my studies at Stony Brook University. I feel privileged to have been given the opportunity to work amongst some of the brightest and kindest scientific minds. With the pursuit of all great things comes great sacrifice and I am indebted to those people in my personal and professional life that have sacrificed their time and efforts to allow me to pursue my passion. To be sure, things have not always been easy for them and it is my sincerest hope that I will be able to return the favor and use my education for the betterment of the greater good.

Firstly, I would like to thank my advisor, Sergei Malsov. Sergei is one of the most brilliant intuitive physicists that I have met. He has taught me, in countless ways, to develop the intuitions that yield simplicity from complex problems. More than that, he has instilled in me an appreciation for the vast array of scientific disciplines that collectively comprise the field of complex systems. It is my hope to one day realize his profound ability to cut to the quick of a problem. His open-minded nature and relentless curiosity will serve to guide me in all my future endeavors.

I would like to thank my officemates, Koon-Kiu Yan, Huafeng Xie, Sebastian Reyes, and Dmitri Volja who have provided a comfortable and friendly

environment full of spirited discussion, both academic and otherwise. I have been fortunate to collaborate with my brilliant colleagues Koon-Kiu Yan and Huafeng Xie. They are excellent scientists, spirited teammates and great friends.

I am indebted to the faculty and my peers at Stony Brook University, who have helped me become the scientist I am today. It is rare to come across such a wide community of talented individuals that embrace the spirit of cooperation and teamwork so fully.

To Pat Peiliker, I extend my warmest thanks for helping me navigate the labyrinth of administration. She has been my savior in a number of circumstances and is truly the mortar that holds together the walls of our great university.

Finally, I wish to thank my lovely wife, Kate, without whom none of this would be possible. She has given me a life and a family that make all other things seem insignificant. I cannot imagine a life without her and I certainly wouldn't want to try.

Chapter 1

Introduction

Complex systems are comprised of a large number of heterogenous components that interact with one another. Complex systems are unique insomuch as they exhibit collective behavior that cannot be understood completely by understanding the local interactions between individual subcomponents. The study of complex systems has its historical origins in several converging fields. Formally, the mathematical study of complex networks underlying such systems can be traced to the development of graph theory in the early eighteenth century, with the pioneering work by Leonhard Euler on the famous *Konigsberg Bridge Problem*. Over the past several decades, advances in the study of complex systems have emerged in the fields of mathematics, statistical physics, computer science, game theory, biology, library science, and sociology.

With the growth of availability of large-scale data sets, complex systems have become an exciting interdisciplinary field of research that requires the development of new tools, methods and modeling techniques to better understand the complex dynamics of the world that surrounds us. In information

systems, the continuing development of online data repositories and emerging interactive systems provide unprecedented opportunity to study complex systems in ways that were never before possible. In biology, high throughput experimental techniques have yielded a wealth of novel empirical data that challenges our conventional understanding of life and sheds new light on the complexity of biological processes.

Throughout the past decade, the study of complex systems has been dominated by a focus on underlying network structure. Recent focus on processes that occur on the backbone of the underlying network have unveiled new areas of theoretical interest. This thesis examines a few studies in information and biological systems focused on the central theme of dynamics of complex systems. This chapter introduces some basic notions and conventions of complex networks that will serve as a foundational context for the chapters that follow. Finally, an outline of the remaining chapters is provided with a brief description of the studies presented and how they relate to the overarching focus.

1.1 Complex Networks

In general, networks are abstract representations of a relational structure that describe how items (nodes) relate to one another. In mathematics, a network or graph $G = (V, E)$ is a pair of finite sets of vertices (or nodes) V and edges E of connections between vertices. Defined in a less rigorous way, nodes and edges in a network can contain a variety of simple and more complex data structures. A node or edge may be defined to have several properties beyond

its unique identifier such as color and value, for example. In this text we will adopt the loose definition of networks and supplement a description of explicit properties wherever necessary. We will further make mention of several classes of networks that are conventionally recognized. In directed networks, edges originate from one node and terminate at another. By contrast, in undirected networks, edges (and thus implied relationships) between nodes are symmetric in that they do not originate from one node and terminate at another, but implicitly "go both ways". Weighted networks contain edges with numerical values that tend to represent the strength of a relationship, whereas edges in unweighted networks may be considered as equal strength and characterized with unit weight to indicate existence. It should be noted that the above statements are not rigorous principles to which networks must adhere, but rather standard conventions that are assumed throughout much of complex network literature.

Networks can be symbolically represented or stored in a number of ways, a few of which are delineated here. An *adjacency list* for a network G, contains, for each node u, a list of adjacent nodes $\{v_i\} = \{v_1, v_2, \ldots, v_k\}$ for which there is an edge (u, v_i). Similarly, an *edge list* is simply a list of all edges $\{(u, v)\}$. The aforementioned structures are useful for storing networks that are sparsely populated (i.e., those for which the ratio $|E|/|V|^2$ is relatively low). Alternatively, an *adjacency matrix* is a matrix:

$$A_{ij} = \begin{cases} 1 & \text{If an edge } i \to j \text{ exists;} \\ 0 & \text{otherwise.} \end{cases}$$

The adjacency matrix can be used to efficiently store dense networks and is a useful expression in analytical manipulations of networks, irrespective of the means of actual network storage. For undirected networks, the adjacency matrix is necessarily symmetric. A generalization of the adjacency matrix for weighted networks can be constructed by inserting the weight of the (i, j) edge in place of unity.

1.2 Examples of Networks

Complex networks are abundant in the natural world, with examples spanning several disciplines of study. Perhaps the most familiar example of a complex network is that of the world wide web, the directed network of hyperlinks that connect one webpage to another. Similarly, a well-known example of a physical network is the internet, the set of physically connected autonomous systems through which online data is routed. More abstract examples of information networks include citation networks, formed by a set of publications that cite on another, and co-actor networks, in which two actors are connected if they have co-starred in one or more performances. This latter type of network has become a well-known example in popular culture of a network that exhibits small-world behavior, as illustrated by the *six degrees of Kevin Bacon* game, that we will discuss briefly in the next section.

In the domain of social science, friendship networks describe connections between people that declare themselves as friends and are useful in the study of the structure of social relationships [1].

In ecology, food web networks describe the complex predator-prey relations

that govern the behavior of natural ecosystems and have been the subject of extensive scientific studies [2, 3]

In biology, gene regulatory networks detail the interactions between the genes of a species that regulate one another to produce the coordinated biological activity necessary to sustain life. The study of genetic regulatory networks has emerged as an important topic of both empirical and analytical studies, [4–6].

The above represent just a few examples of the multitude of complex networks that exist in our world. Indeed, there can be little doubt that the detailed study of complex networks is integral to our scientific understanding of a variety of systems. While the study of these systems has in the past been regulated to their respective disciplines, the necessity to develop a common set of techniques and descriptive language has led to a recent convergence and, ultimately, the formation of the interdisciplinary field of complex networks.

1.3 Network Structure and Properties

While the topology and features of small networks can be easily characterized graphically, the visualization of larger networks in a manner that reveals both local and global properties is significantly more challenging. An alternative description of network properties can be accomplished through the definition of several metrics that help to elucidate critical characteristics and distinguish the topological nature of large networks from one another. Several such metrics are defined below that have proven to be indispensable to our understanding of large networks.

1.3.1 Degree

The degree of a node i in a network is defined as the number of edges for which that node is a participant. For directed networks, this concept can be generalized to the in- and out-degree of a nodes, corresponding to the number of edges that terminate and originate from that node, respectively.

In terms of the adjacency matrix, the in-degree of the jth node can be defined by:

$$k_j^{in} = \sum_i A_{ij} \tag{1.1}$$

Similarly, the out-degree of the ith node is defined:

$$k_i^{out} = \sum_j A_{ij} \tag{1.2}$$

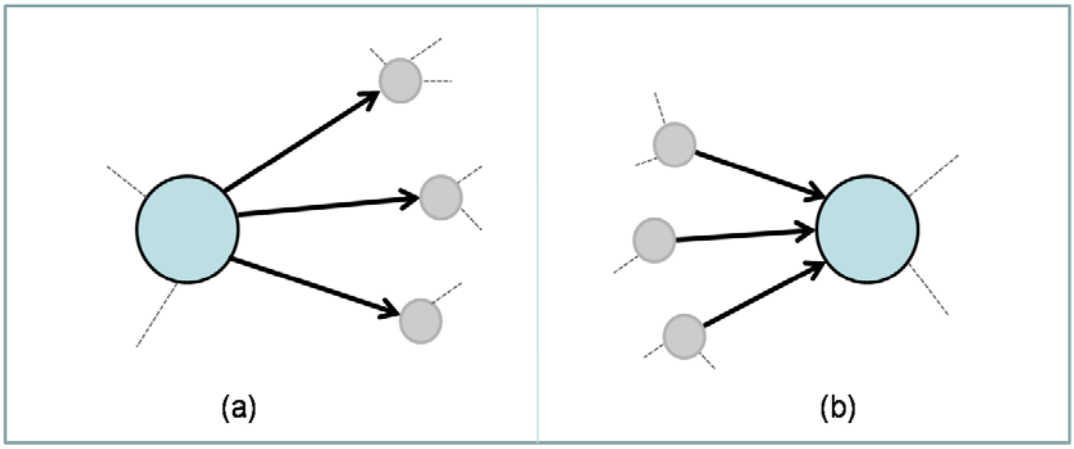

Figure 1.1: An illustration of in- and out-degree of the large colored node is shown in the two panels above. In panel (a), the out-degree of the node is $k^{out} = 3$. In panel (b), the in-degree of the node is $k^{in} = 3$.

For undirected networks, $k_i = k_i^{in} = k_i^{out}$ for all nodes in the network and

we simply speak of degree. For directed networks, the total degree is given by the sum of in- and out- degrees, $k_i = k_i^{in} + k_i^{out}$. An illustration of in- and out-degree is provided in fig. 1.1. For large networks, it is useful to consider the *degree distribution*, $P(k)$, that describes the statistical distribution of degrees for all nodes within a network.

Of particular interest is the finding that many real world networks exhibit long-tailed degree distributions that appear to be approximately power law in nature, $P(k) \sim k^{-\gamma}$ for some positive exponent γ. Networks exhibiting power law degree distribution are often referred to as *scale-free*, as the degree of a typical node in the network cannot be characterized by a single scale, a consequence of the diverging first moment of the degree.

Network	size (nodes)	γ^{in}	γ^{out}	Reference
www	2×10^8	2.1	2.71	Broder (2000)[7]
E.coli (metabolic)	778	2.2	2.2	Jeong (2000)[8]
Movie co-actor	2.12×10^5	2.3	2.3	Barabasi (1999)[9]
Word co-occurence	4.62×10^5	2.7	2.7	Cancho (2001)[10]

Table 1.1: A few examples of networks that exhibit power law degree distributions. For the directed networks shown above, γ^{in} (γ^{out}) corresponds to the power law exponent for in- (out-) degree distributions. For undirected networks in the table, the two are equal.

A few examples of networks with power law degree distribution are given in table 1.1. From the standpoint of statistical physics, power law distributions are particularly interesting as they are typically associated with critical behavior. Recently a great deal of study has been devoted to identify plausible mechanisms that generate power law degree distributions. A review of power

law behavior and generative mechanisms is provided by the author of [11].

One notable mechanism for generating power law degree distributions is a rich-get-richer phenomenon first generalized by Yule [12] and later studied in the context of complex networks by Barabasi and Albert under the name of *Preferential Attachment* [13]. In preferential attachment, complex networks are generated from smaller core networks by the successive addition of nodes. Nodes are created and attached to the network via a single edge to an existing node with probability proportional to the degree of the existing node.

$$Pr\{\text{new node } n \leftrightarrow i\} = \frac{k_i}{\sum_j k_j} \tag{1.3}$$

In many systems, this mechanism can be associated with an intuitively plausible scenario. For example, in the world wide web, it might seem reasonable to assume the probability to link to a webpage proportional to the chance to visit that page. We will revisit a form of the preferential attachment mechanism in chapter 3, in the context of popularity.

Beyond the degree distribution, there are a number of other metrics that help to reveal and characterize the topological properties of complex networks. Two networks with identical degree distributions may exhibit vastly different topologies. One distinguishing topological property of networks is the assortativity, the extent to which nodes in the network tend to be connected to other nodes with similar degree. An illustration of two networks with identical degree distribution, but different assortativity is shown in fig. 1.2.

A more microscopic view of topological properties can be attained through the examination of network motifs. A network motif is a small subgraph of a

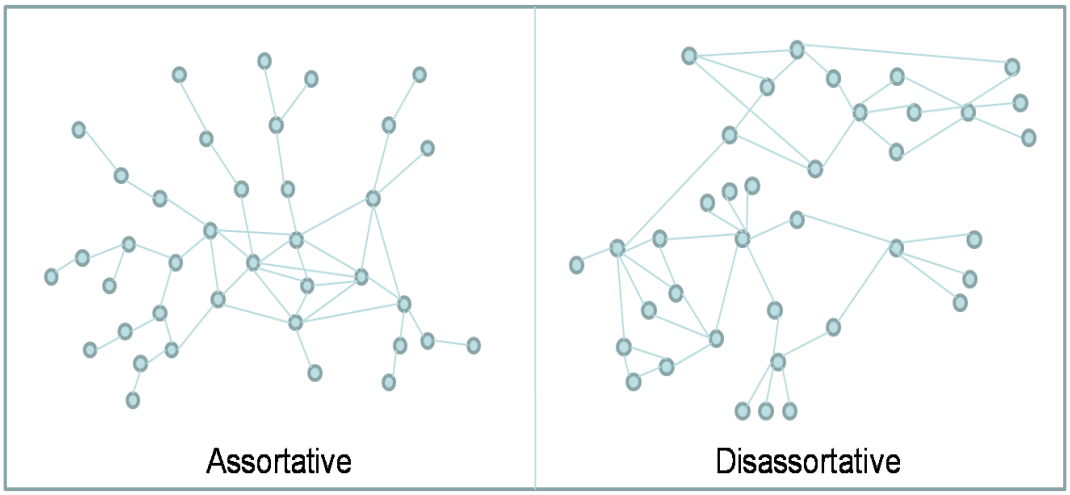

Figure 1.2: The topology of two networks with identical degree distributions. The assortative network features high degree nodes that tend to connect to one another and clump together. In contrast, in the disassortative network, nodes with high degree tend to be connected to nodes with low degree.

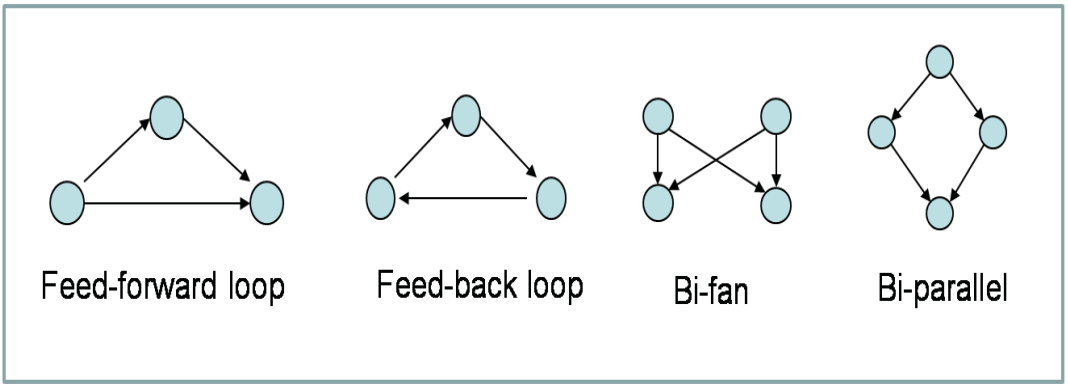

Figure 1.3: An illustration of several network motifs that commonly occur in directed complex networks.

few nodes that are connected together in a particular pattern. Motifs are particularly significant in the study of dynamical properties of complex systems, as they may be related to the function of a subsystem and, in isolation, lend themselves to tractable studies of dynamical behavior. An illustration of a few

network motifs in directed networks is provided in fig. 1.3. In undirected networks, one popular topological metric that is based on motifs is the *clustering*, the ratio of the number of triangles in a graph to the number of unordered triplets.

1.3.2 Connectedness

The connectedness of a network generally describes how nodes in the network are connected to one another through pathways. A simple pathway of length L in a network from some node u to another node v is defined as a sequence of adjacent nodes originating from u and ending in v, $\{u, u_1, \ldots, u_{L-1}, v\}$ in which no node is repeated. Pathways that are not simple, (i.e, ones that involve the repetition of a vertex), arise due to the existence of cycles. A cycle is a pathway that begins and terminates with the same node. In general, for two arbitrarily chosen nodes in a networks many pathways may exist between them.

A *component* of a network is a subset of nodes in the network that can all be reached from one another through pathways of any length. We say that all nodes in a component are connected to one another. An illustration of components of a network is show in panel (b) of fig. 1.4.

A *geodesic path* from u to v is the shortest path connecting u to v whose length will be referred to as the distance between u and v. An illustration of a geodesic path is shown in panel (a) of fig. 1.4. An interesting property of any network is the average distance between all nodes in components of the network.

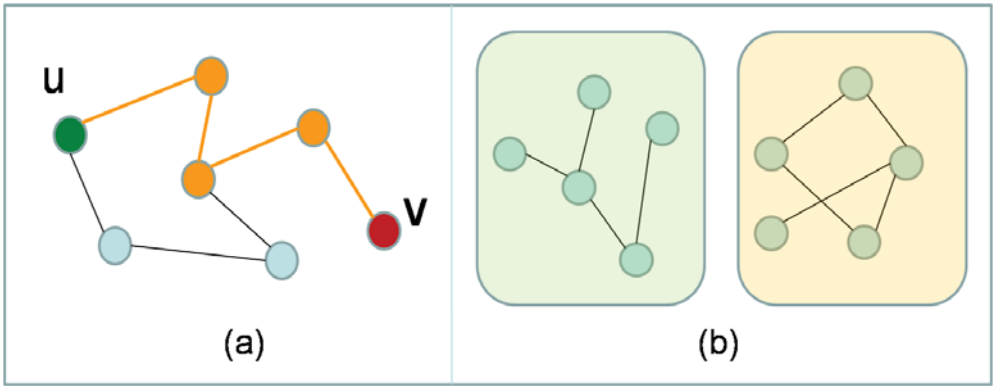

Figure 1.4: Panel (a) illustrates a geodesic path from node u (green) to v (red) via intermediary nodes (orange). Panel (b) illustrates two separate components of a single complex network. All nodes of each components are reachable from any other nodes in that component.

The study of the connectedness of real world networks has revealed a variety of networks that exhibit relatively low average distance $d \leq 6$. In the literature, such networks are commonly referred to as *small-world networks*. In a famous 1969 study of social connectedness, Stanley Milgram [14] conducted an experiment in which several randomly selected individuals from cities in the United States were given a letter detailing the nature of the study and a contact destination. Participants were asked to forward the letter to the contact, if known personally, or to a personal acquaintance that they believed likely to know the contact. The chain of intermediary acquaintances was recorded as the letter passed from one individual to the next. The remarkable result of the study showed that the average pathway between the original recipient and contact destination was approximately $L \sim 6$, resulting in the popular notion of *six degrees of separation*. In the example of movie co-star networks, this notion has received wide popularity in the form of various trivia games in

which the player is challenged to connect two actors to one another by citing pathways of co-stars. For example, one might challenge a player to connect Arnold Schwarzenegger to Bruce Willis. One solution is as follows: Arnold Schwarzenegger was in *The Terminator* with actor Michael Biehn. Michael Biehn was in *Planet Terror* with Bruce Willis. Thus, Arnold Schwarzenegger has a Bruce Willis number of 2. An online implementation of this game is available through the University of Virginia's Computer Science Department (oracleofbacon.org).

1.3.3 Randomness and Graphs

Attempts to understand the topological features of complex networks has led to the introduction of generative approaches that involve randomness. Indeed, early generative studies of graph structure focused on random graph formation, popularized by the well-known Erdős-Rényi (ER) Model [15], in which every pair of nodes is connected by an edge with a fixed probability p. While this model is a useful stepping stone in understanding both connectedness and stochasticity of graph generation as a function of the parameter p, it unfortunately fails at reproducing many of the features of real world networks. In particular, the limiting degree distribution of the ER model is Poissonian and it cannot produce graphs with small-world properties or scale-free degree distributions.

In 1998, Watts and Strogatz (WS) introduced the first successful model of random graph generation that can lead to graphs with small world properties [16]. The model begins with an ordered ring graph in which each node is con-

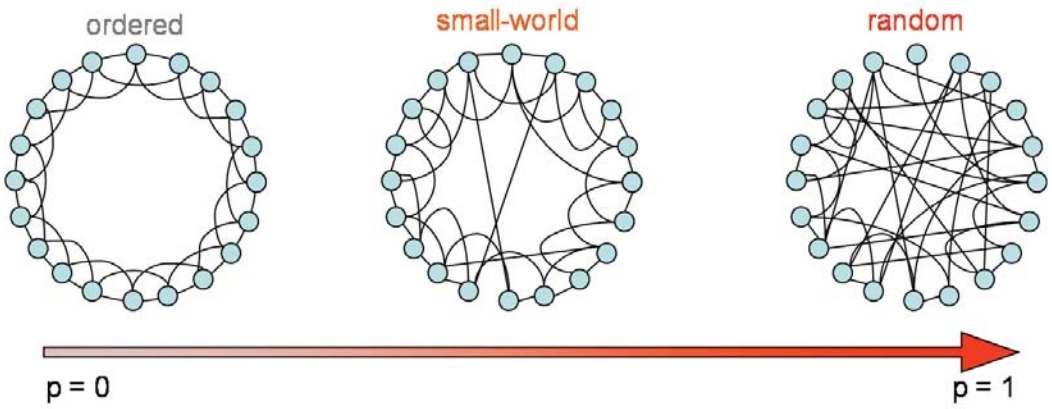

Figure 1.5: An illustration of the Watts-Strogatz Model for random graph generation. The model begins with an ordered ring network in which each node is connected to precisely k neighbors. With probability p, every edge is randomly rewired, resulting in the generation of a random graph. In the limit $p \to 0$, the graph is simply ordered, while, in the limit $p \to 1$, the random graph is equivalent to one produced by the ER model described in the text. For intermediary values of p, the model produces graphs with both small-world properties and high clustering.

nected to exactly k nearest neighbors on the ring. Each edge is subsequently randomly rewired with probability p. In the limit $p \to 0$, the graph is unchanged and possesses a homogeneous degree distribution, while in the limit $p \to 1$, the graph is completely random and equivalent to the ER case. In between these two extremes, however, the model generates graphs with high clustering and low average path lengths known as the small-world property. An illustration of the WS model is provided in fig. 1.5. Unfortunately, the WS model does note produce scale-free degree distributions typical of many real-world networks.

While more recent generative models, such as the aforementioned preferential attachment model, and various other models, such as duplication

and deletion models studied by the authors of [17–19], do successfully produce scale-free graphs, they are reliant upon the plausibility of their inherent generation mechanisms, which may not be realistic for all types of networks. Furthermore, these models do not explicitly address many of the topological features of graphs and associated metrics previously discussed.

An alternative approach to random graph generation was derived by Holland and Leinhardt [20] and throughout the past several decades has been generalized to a class of ensemble models commonly referred to as *Exponential Random Graph Models* (ERG). In ERG models, one considers an ensemble of all graphs with N nodes, and introduces the partition function:

$$Z = \sum_{G} e^{-H(G)} \tag{1.4}$$

where the sum is carried out over the ensemble of graphs. The probability of realizing some particular graph G is then given by the standard statistical relation:

$$P(G) = \frac{e^{-H(G)}}{Z} \tag{1.5}$$

The effective hamiltonian, $H(G)$, may then be defined as a linear combination of any number of desired topological features. For example, to encourage the realization of a graph with high clustering, we can include a term $H(G) = -\alpha_T T(G) + \ldots$ where $T(G)$ is the number of triangles in the graph G. The associated weight α_T characterizes the relative importance of this feature with respect to other terms in the hamiltonian. Thus, ERG models attempt to define a phase space for graphs, in analogy with proven methods of statistical physics. One advantage of such a method is that it allows for the generation

of several graphs with similar topological features. The study of ERG models is an active topic of research [21, 22]. Unfortunately, in many situations ERG models are not feasible. Furthermore, it may not be entirely clear which features of a graph are significant and ERG models cannot, in general, account for features that are not put into the hamiltonian "by hand".

One application of randomization that can help us discover which features are significant in a network is the random edge rewiring approach introduced by Maslov and Sneppen [23, 24]. In this approach, edges in a network are rewired in a manner that conserves both in- and out- degree distribution. This is accomplished by several edge-swapping iterations in which two edges are randomly selected, $A \to B$ and $C \to D$, and subsequently swapped such that $A \to D$ and $C \to B$, provided that the resultant edges do not already exist (if so, the swap is abandoned). The resulting edge-shuffled version of the network provides a null model whose topological features may be directly compared to those of the original network. For any topological feature that appears θ times in the original network and θ_r times in the randomized network, the quantity of interest is the ratio θ/θ_r, whose value is ~ 1 when the feature occurs no more in the real network than as expected by random chance in a network with the same in- and out- degree distributions. This rewiring technique can be extended to conserve local topological properties in the null model by prohibiting any edge exchanges that alter those properties.

As an illustration of this technique, the assortativity of a network can be quantified by considering degree-degree correlations, $d(k_0, k_1)$, the number of nodes with degree k_0 connected to neighbors with degree k_1. A plot of the *correlation profile*, $d(k_0, k_1)/d_r(k_0, k_1)$ for the internet (autonomous systems

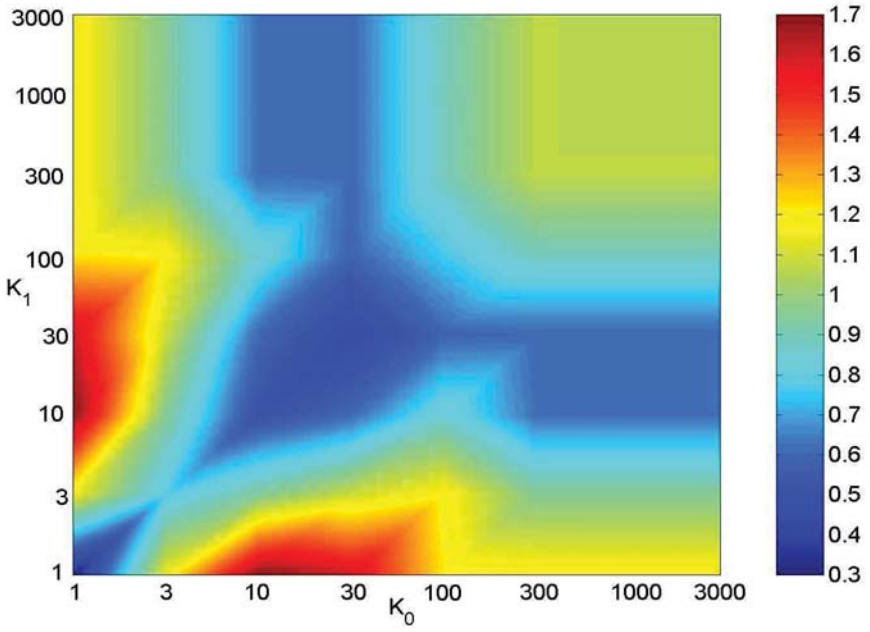

Figure 1.6: A correlation profile of the internet (autonomous systems level) derived from the random edge rewiring technique discussed in the text. The x-axis represents a nodes of degree k_0, while the y-axis represents the degree of its neighbor, k_1. For any point (k_0, k_1), the intensity is given by $d(k_0, k_1)/d_r(k_0, k_1)$, the ratio of occurrences in the real network to those in a randomized network with the same in- and out- degree distribution. The plot displays that nodes in the internet tend to be connected to neighbors with significantly different degree, indicating that the internet is a relatively disassortative network. This figure is reproduce from [24] with permission from the authors.

level) is reproduced with permission from the authors in fig. 1.6 Evidently, the internet is somewhat disassortative, as connections between nodes of equal degree are significantly under-represented in the range $k \leq 100$ than one would expect by random chance.

1.4 Network Algorithms

The study of networks has led to the development of several algorithms that address a variety of issues, including optimization and search. While a complete review of graph algorithms is beyond the scope of this text, we will briefly mention a few relevant graph algorithms.

Traversal in networks is the process by which an object moves from one node of the graph to another adjacent node by traveling through an adjacent edge. Mathematically, traversal in a directed or undirected graph can be accomplished through the multiplication of an initial occupation vector by the adjacency matrix. Explicitly, if $\boldsymbol{p}[t]$ (a vector of length N) describes a distribution of objects that "live" on the N nodes of a network at time t, and objects traverse an edge at each time step, then:

$$\boldsymbol{p}[t+1] = A \cdot \boldsymbol{p}[t] \tag{1.6}$$

describes the distribution of the objects on the N nodes at the next time step, $t+1$. Where the matrix A is the adjacency matrix described in eq. 1.1. The movement of objects along a graph can be generalized to a stochastic process with the introduction of random variables and accompanying probabilities that describe the behavior of the motion. We will see an example of this in chapter 2, in the context of random walks on citation networks.

One common problem related to traversal is that of find the shortest path between any two nodes in a network. As it turns out, the calculation of the shortest path from a source node to a destination node is computationally

no longer than the calculation from a source node to *any* other node in the network. Several algorithms exist to efficiently compute shortest paths with computation times that depend on the type of data structures employed.

A fast algorithm for computing shortest path distances in unweighted graphs is the *Burning Algorithm* [25]. This algorithm is so named because it models the spread of a fire from a source node s to other nodes in the graph. At each time step, the fire traverses from each burning node, along all edges, to unburnt nodes. At each step, we retain the list of unburnt nodes and the burning time step, t. When a node is burnt, it is removed from the list of unburnt nodes and the distance from the source is recorded in a distance matrix:

$$d[s, j] = t \tag{1.7}$$

The burning algorithm can be accomplished with a series of matrix multiplications and in $O(D)$ time where D is the largest shortest path from originating from node s in the network. However, the burning algorithm is not appropriate for weighted networks.

Several algorithms exist for shortest path determination in weighted networks. These algorithms typically take advantage of a property of shortest path problems referred to (in computer science) as *optimal substructure*. Problems with optimal substructure are those for which the solution can be constructed from the optimal solutions of subproblems. Such problems can effectively be solved using algorithms that take a greedy approach to the solution of subproblems. In the context of shortest paths, greedy algorithms find the shortest path from a source node s to an end node e by first finding the shortest

paths from s to neighboring nodes and then iterating.

One well known fast algorithm for shortest path determination in a weighted network from a node s to all other nodes is *Dijkstra's Algorithm* [26]. In the algorithm, the current estimate of shortest path distance $d[s, j]$ for all nodes is maintained and initially set as infinite (except for $d[s, s] = 0$). A list, Q, is maintained of all nodes whose shortest path distance estimate is not yet "finalized". At teach time step, the node $u = min(Q)$ with the minimum distance estimate is finalized and the node is removed from Q. Its outgoing neighbors are then explored $\{v_i\}$. For each neighbor, the distance estimate is updated according to:

$$d[s, v_i] = min(d[s, v_i], d[s, u] + w(u, v_i)) \tag{1.8}$$

where $w(u, v_i)$ is the weight of the outgoing edge from u to v_i. That is, if the pathway from s to v_i that includes the intermediary node u is shorter than the current estimate for the shortest path from s to v_i, it replaces it as the current estimate. This procedure is then repeated, until all nodes are removed from Q. The running time of Dijkstra's algorithm depends on the data structure employed to implement the list Q. For an implementation that uses a binary heap for Q, the running time is $O(NlogN + ElogN)$ for N nodes and E edges. Unfortunately, Dijkstra's algorithm cannot be used for networks that include negative weights. For such networks, other algorithms such as the *Bellman-Ford Algorithm* [27] are preferred. A more efficient pathfinding algorithm that involves heuristic estimations of distance from nodes to path endpoints is the well-known *A* Algorithm* [28] that is used extensively in artificial intelligence

applications.

Beyond traversal and path-finding algorithms, there are countless other graph algorithms that will not be mentioned here. One notable group of algorithms are those that attempt to find *community structure* within networks. A community can be loosely defined as a group of nodes within a network that are densely interconnected relative to connections to nodes outside of the community. Over the past two decades, the detection of community structure has been recognized as vital to the understanding of many complex systems. An overview of these algorithms is beyond the scope of this text, however a good review of community structure detection can be found here [29].

1.5 Outline

The remaining chapters of this work will focus on the study of three specific complex systems.

In chapter 2, we study the dynamics of citation networks. The ageing characteristics of citation networks are explored and the effect of aging on the conventional method of ranking is discussed. The diffusive ranking method of PageRank is considered and its positive attributes are discussed. A modified version of the PageRank algorithm, called CiteRank, is introduced to account for aging effects and to rank articles according to relevance to modern research. Optimization of the parameters of the CiteRank algorithm is discussed and the performance is compared to that of PageRank. The structure of recent citation growth is explored through the introduction of a mean-field model that explains differences in recent citation accrual in two age regimes that

correspond to direct and indirect means of dominant article discovery.

In chapter 3, we study time-critical collaborative document voting systems that occur in a variety of implementations in real online systems. A general model of the system is presented with the introduction of two simple models of user behavior that serve to characterize quality and popularity in mixed populations of users. Issues of ranking through voting mechanisms are discussed and an estimate for cooperation of the population is presented. The effects of popularity on rank performance are examined and nonlinear herding effects are observed and characterized by a mean-field model. A mechanism for user reputation is introduced that is self-consistent, content-independent and adaptive. Finally, a method of ranking articles is introduced that accounts for user reputation and provides a significant improvement in the time-critical rank performance.

In chapter 4, we study fluctuations in mass-action equilibrium of protein-protein-interaction (PPI) networks of living cells. We identify two types of fluctuations that arise from total protein concentration changes and changes in bound concentration as a result of thermal kinetics that we respectively refer to as *driven* and *spontaneous* fluctuations. Fluctuations are calculated for a real curated PPI network of Baker's Yeast. The collective effects of the underlying network are explored through the introduction of isolated dimer models that provide an upper- and lower-bound to fluctuation amplitudes. Amplification of fluctuations that arise as a consequence of collective effects are found to be highly significant. Finally, we compare noise amplitude of dimers to the simple predictors of abundance and network connectivity.

Chapter 2

Traffic Models and Ranking in Citation Networks

2.1 Background

A class of information networks that arise naturally in a variety of information systems are citation networks. Citation networks are typically formed in document publication systems where documents cite pertinent information from existing publications. Real-world citation networks can arise from systems that are quite diverse in character, such as patent systems, legal cases, and scientific journals. Despite this diversity, citation networks share some common characteristic features that distinguish them from other types of information networks.

Unlike objects in many modern information systems, documents in citation networks are rarely modified after the point of publication. Because of this, aging effects in citation networks can be significantly pronounced. Both the

information and citations in published documents can only depend on earlier works. Consequently, the topology of a citation network exhibits a natural time-arrow such that chains of citations naturally flow backwards in time to older publications. These issues have serious implications on the study of citation networks and will be addressed further in this chapter.

Serious quantitative study of scientific publication networks can be perhaps traced to several pioneering studies in the early 1960s. Eugene Garfield formed the *Institute for Scientific Information* (ISI) in 1960, and produced the *Science Citation Index* (SCI), the first systematic large-scale index of citations between scientific papers. While the advent of the SCI was intended as a useful tool for scientific researchers to trace citations between papers, the SCI allowed for the quantitative study and evaluation of research publications. That citation data might be used as a means of ranking articles, journals and even authors is a notion that has existed for some time within scientific publishing. However, with the advent of large-scale indexes such studies became statistical in nature. The authors of [30] studied and pioneered some of the precedents for bibliometrics in common use today, such as rank by number of incoming citations and journal impact factor. The authors of [31] and [32] recognized the network description of citation indices.

Modern citation network analysis has matured significantly with the growth of data access and availability. At present, the various citation indices (SCI and others) maintained by ISI has grown into the *Web of Science*, an online database containing citation data from more than 8,700 journals in a wide range of fields throughout the sciences, arts and the humanities [33]. Alternate online citation databases exist as well, such as *Scopus*.

In light of the vast growth in scientific publications and online accessibility, it is not surprising that the issue of information search, ranking and retrieval has come to the forefront of citation network analysis. Several freely available engines have been developed to search scientific literature, such as *CiteSeer*, and *CiteBase*, that rank publications based on their incoming citations. Other engines, such as *Google Scholar*, use ranking schemes that are not known to the public.

In this chapter, we examine two real-world citation networks, all Physical Review Journals up to 2003 (physrev), and a snapshot of the preprints in the Physics High Energy Theory arXiv (hep-th). The weaknesses of standard ranking by incoming citations are discussed. We consider application of the PageRank algorithm to citation networks and remark on its strength and weaknesses. The problem of aging in citation networks is considered and a new traffic-based algorithm, CiteRank, is presented to rank scientific publications by their current relevance. The advantages of CiteRank over traditional methods of ranking scientific publications are discussed.

2.2 Real Citation Networks

Throughout this chapter we examine two real-world citation networks

- **Hep-th**: An archive snapshot of the "high energy physics theory" archive (http://arxiv.org/archive/hep-th) from April 2003 (preprints ranging from 1992 to 2003). This dataset, containing around 28,000 papers and 350,000 citation links, was downloaded from [34]. We know the actual date of appearance of each of the entries in the preprint archive and thus

the age of each node is known with the resolution of 1 day.

- **Physrev**: Citation data between journals published by the American Physical Society. These journals include Phys. Rev Series I (1893-1912), Phys Rev. Series II (1913-1969), Phys Rev. Series III (1970-present). The latter is comprised of five topical sections: Phys. Rev. A,B,C, D and E (1990-2003). Additionally included are Phys. Rev. Lett., Rev. Mod. Phys., and Phys. Rev. Special Topics, Accelerators and Beams (1990-2003). This dataset contains around 380,000 papers and 3,100,000 citation links. We know only the year in which each paper was published and it ranges from 1893 to 2003.

These networks are, on the surface, quite different in nature. The Physical Review citation network (physrev) is comprised of a large number of peer-reviewed publications acquired over a period close to a hundred years. The high-energy physics archive citation network (hep-th) is comprised completely of a much smaller number of topically similar electronically submitted publication preprints, with no associated form of peer review.

General Features of the Citation Networks

In general, citation network are directed and we will adopt the convention for the directed adjacency matrix:

$$A_{ij} = \begin{cases} 1 & \text{if article j cites article i;} \\ 0 & \text{otherwise.} \end{cases}$$

The in- and out- degree is accordingly defined in the standard manner,

$k_i^{in} = \sum_j A_{ij}$ and $k_j^{out} = \sum_i A_{ij}$. The in-degree for the two citation networks is shown in Fig. 2.1.

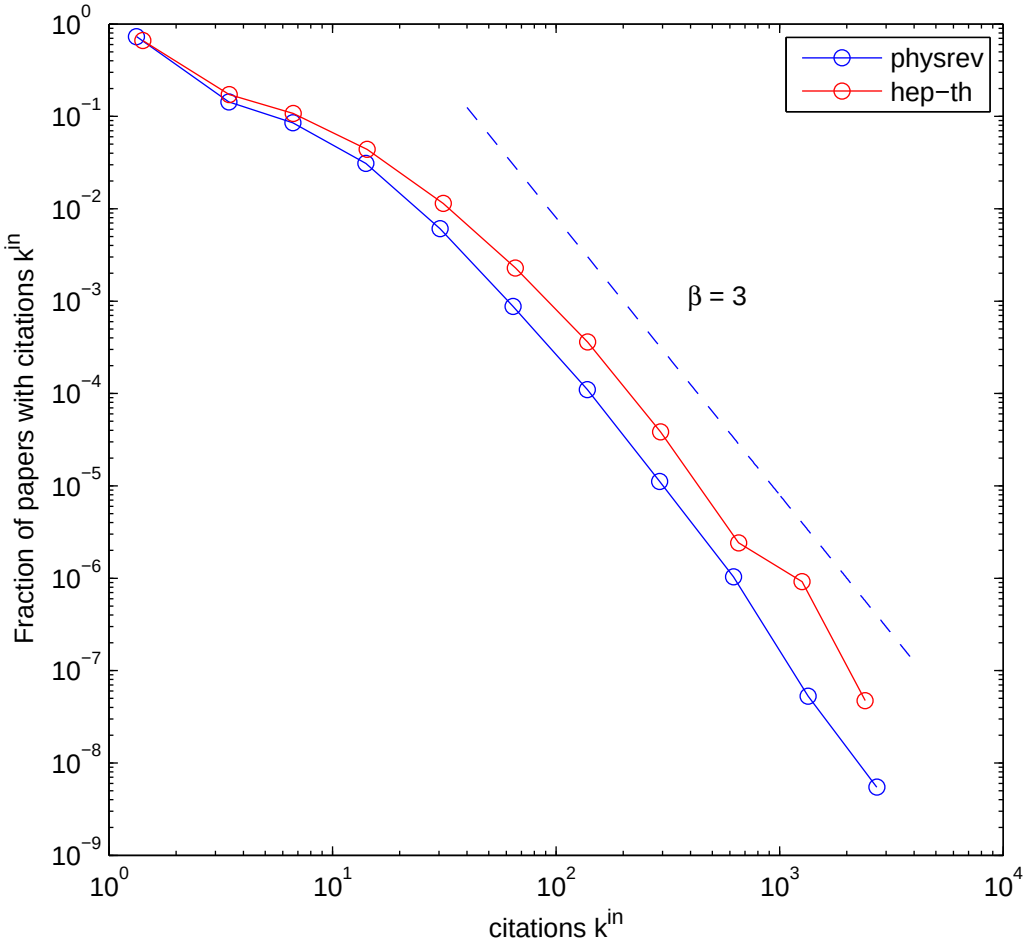

Figure 2.1:

For both networks, it has a broad distribution that is approximately fit by a power law of exponent $\beta = 3$. The power law form is in agreement with several in-degree studies of various citation networks. However, in a recent study of the physrev dataset [35], the author argues that a log-normal

distribution may better fit the precise nature of in-degree distribution. Studies of various generative models, that attempt to explain the long-tail behavior by modeling the growth of citations, are not conclusive. One possible reason for this failure is the inability for generative models to account for extreme cases, where publications experience citation growth that is far from the typical. Such cases are likely represented in datasets, like physrev, that span large periods of time over which the practice of science and scientific publishing has drastically evolved. If this is true, then small deviations from power law behavior in the tail of the data may be negligible. These issues aside, an observations of typical citation growth are noteworthy and relevant to the details of this study. Specifically, publications with high citations tend to be cited more frequently, a rich-get-richer phenomenon recognized early on by Merton and termed "cumulative advantage" [36]. We will return to this phenomenon later, both in the concept of publication visibility and aging.

Another property of interest for citation networks is the distribution of in-degree according to the publication age. While the in-degree of a particular paper is a function that monotonically increases with age, the average in-degree of papers of a particular age is not similarly constrained. The number of citations versus publication age is shown in Fig. 2.2 for the Physrev citation network.

Old publications in this distribution are heavily skewed both by the approximate exponential growth rate of articles and the emergence of new physical review journals. A particularly striking feature of this distribution that is not explained by growth in publications is the sharp drop-off in citations for papers that are only a few years old. It is intuitively clear that newer publications on

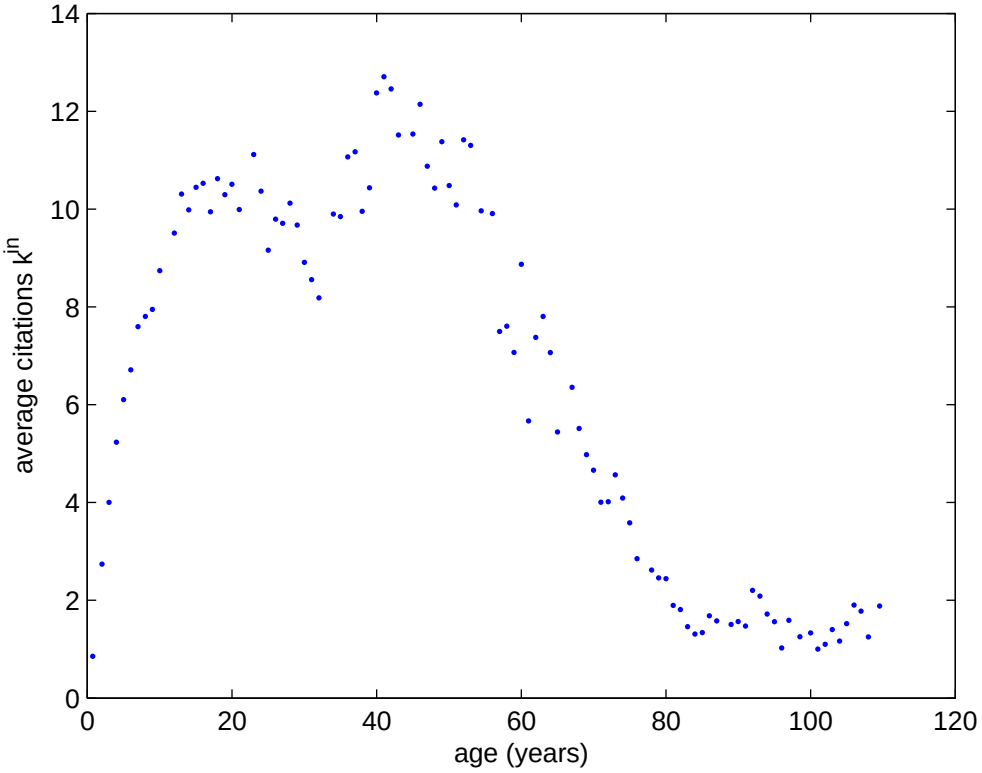

Figure 2.2: Average number of citations versus age of publications in the Physrev citation network. The x axis is the age of the publication. The y axis is the average number of citations for publications of a given age.

the average cannot compete with older ones in terms of raw citations alone, by the simple fact that they have not existed long enough to accrue as many citations. This observation is relevant to the task of ranking publications for the purposes of search, where it is precisely the newer publication that are likely to be pertinent to modern research. These consideration lead naturally to the problems inherent in the traditional method of ranking publications, by raw number of citations.

Ranking Publications

As previously mentioned, the interpretation of incoming citations as a measure of relative quality amongst papers in the network is a common practice. While naively, there seems to be some merit to this method, several issues preclude the use of raw number of citations as a "good" indicator of quality. As was shown in the prior section, new publications do not typically achieve large number of citations until they have existed for times in excess of several years or even decades. Indeed, the average age of the top ten papers ranked according to k^{in} is 31 years in the case of the physrev network, and 3.8 years in the case of the hep-th network. Furthermore, the treatment of all citations as equal is "too democratic". A more reasonable ranking metric should account for the quality of *citing* publications. In other words, citations originating from an eminent paper should contribute to the quality of a cited publication more than those originating from a lesser known paper.

The problem of ranking citation networks is not entirely divorced from that of ranking other information networks. In the analogous network of the world wide web, ranking of websites according to number of incoming links also suffers from the same issue of over-democratization. Sergey Brin and Lawrence Page addressed this issue with the well-known PageRank algorithm [37], the core of Google webpage ranking.

PageRank is a diffusive network algorithm that is mathematically equivalent to the simulation of a population of random walkers diffusing along the links of the network. In the algorithm, a population of random walkers are initially distributed homogeneously across the network. Each random walker,

situated at a node in the network will, with probability α, jump to a random page in the network and with probability $1 - \alpha$ follow an outgoing link to a neighboring node.

The PageRank number of a website is defined as the cumulative traffic through the node. Unlike ranking by in-degree, PageRank captures the *self-consistent popularity* of websites, because the number of random walkers occupying a given node is proportional to the in-degree of the node. Sites with high popularity (large PageRank number) have high occupancy and the traffic flowing out of the site through its outgoing links is proportional to the occupancy. Consequently, a link from a popular site contributes to the PageRank number of the linked site more than a link from a less popular site. Moreover, in PageRank, the effect of a link from a site that has a large number of outgoing links is diminished. In the context of a citation network this feature is in accordance with the notion that the citations from a publication collectively represent the preceding work and inspiration upon which that publication has been based. In other words, an incoming citation from a paper with 100 outgoing citations is "less meaningful" than one from a paper with only a few outgoing citations.

While PageRank is mathematically equivalent to a diffusion process, no such process is actually simulated to evaluate PageRank on a network. The calculation of PageRank for a network is performed by a computationally efficient and guaranteed convergent matrix equation:

$$\boldsymbol{P} = \alpha I + (1-\alpha)\alpha W + (1-\alpha)^2\alpha W^2 + (1-\alpha)^3\alpha W^3 + \dots \tag{2.1}$$

where I is identity and W is defined as the transfer matrix:

$$W_{ij} = \begin{cases} \frac{1}{k_j^{out}} & \text{if j links to i;} \\ 0 & \text{otherwise.} \end{cases}$$

The first term in the PageRank corresponds to the uniform distribution of random walkers across the network. Each of the terms that follow correspond to random walkers that arrive at a site by following chains of links of length one, two, three, and so forth. The parameter α represents the probability that a walker will cease to follow a chain of links and simply jump to a new site in the network at random. In application of PageRank to the web, Google selects this parameter to be $\alpha \simeq 0.15$ in agreement with the notion that a typically web surfer will follow approximately $1/0.15 = 6$ hyperlinks before abandoning their search.

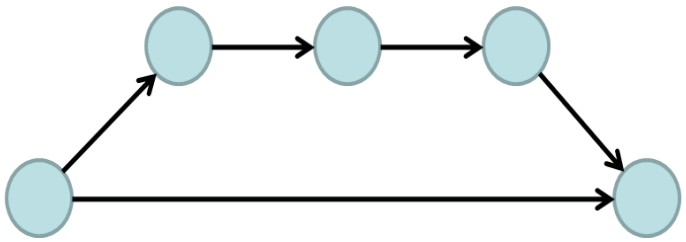

Figure 2.3: A feed-forward loop of length four. In citation networks, feed-forward loops are intimately related to indirect article discovery.

Ranking of publications in a citation network with the PageRank algorithm was first considered, for the physrev network, by the authors of [38]. In this study, the authors conjecture that the parameter α is closely related to the number of feed-forward loops in the network. A feed-forward loop in a citation network of length four is diagramed in Fig. 2.3.

Such a structure clearly suggests that the author of the leftmost paper in fig. 2.3 has read both the intermediary papers and the paper that completes the loop . An analysis of feed-forward loops in the physrev and hep-th networks is shown in Fig. 2.4. The feed-forward loops were analyzed using a modification of the burning algorithm discussed in chapter 1.

For both networks, we find a sharp drop off in feed-forward loops of length $L > 2$, in good agreement with the finding in [38], supporting the choice of parameter $\alpha \simeq 0.5$ of the PageRank algorithm.

We apply the PageRank algorithm to the physrev and hep-th citation networks. The results for the top ten publications for both networks are show in tables 2.1 and 2.2.

Comparison to the more traditional method of ranking by k^{in} is presented in Fig. 2.5. The two are well correlated, a result that can be understood on the basis that the larger the number of citations a paper has, the more likely it is to receive traffic from random walkers.

While the PageRank algorithm for citation networks addresses the issue of self-consistent popularity, it is nonetheless still susceptible to problems of aging. The mean age of the top ten publications according to PageRank is ~ 43 years in the physrev network and ~ 9 years in the hep-th network. Such a ranking is certainly useful as the papers in the top ten list represent undeniably significant advances in their respective scientific fields. Nearly all of the papers within the top ten for both networks show a broad range of citation activity, with citations from papers of all ages represented. Recent citations represent only a small fraction of total citations for these papers. With this in mind it is natural to interpret the PageRank of a publication as a "lifetime achievement

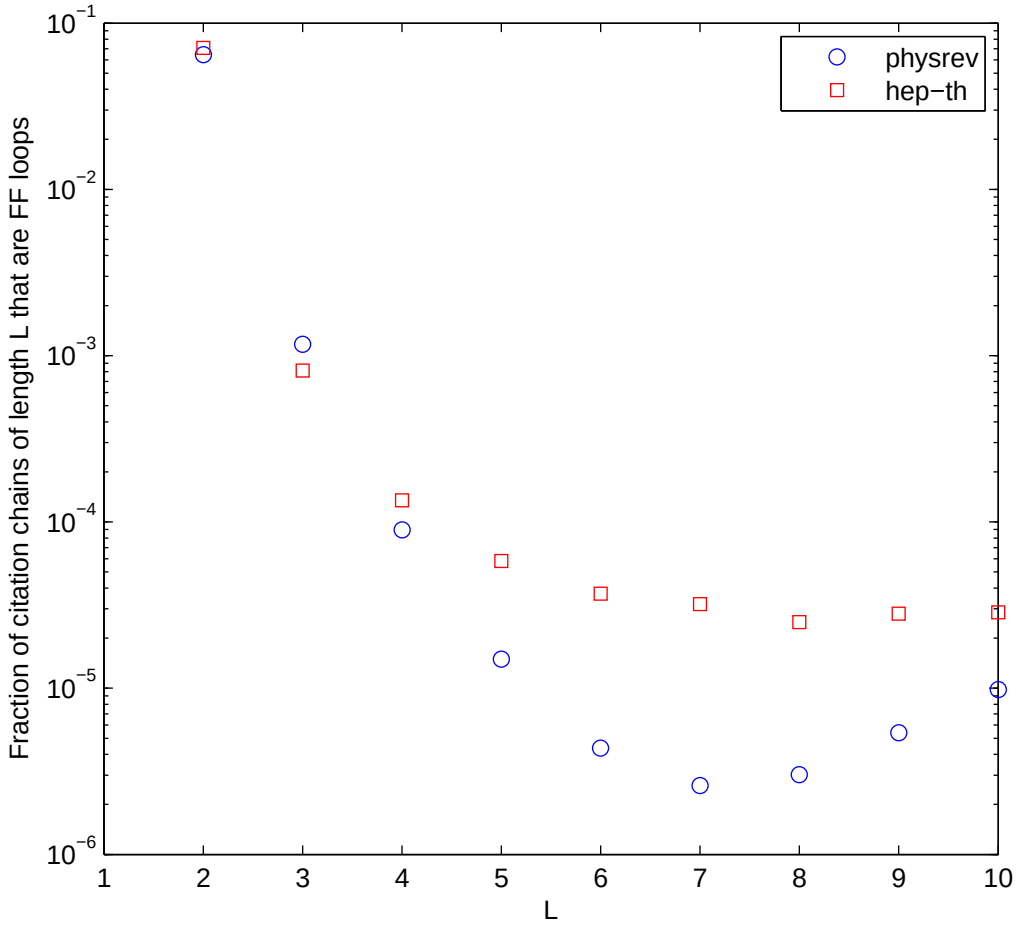

Figure 2.4: The x axis is length L of the citation chain. The y axis is the fraction of citations chains of length L that are feed-forward loops. Circles (squares) are the results for the physrev (hep-th) citation network. The appearance of feed-forward loops of length $L > 2$ drops off sharply for both networks.

award" [38]. In order to achieve a ranking that is more relevant to lines of current research, the issue of aging must be explicitly accounted for.

Aging in citation networks has serious implications on the network structure. The existence of a time-arrow in citation networks implies the absence

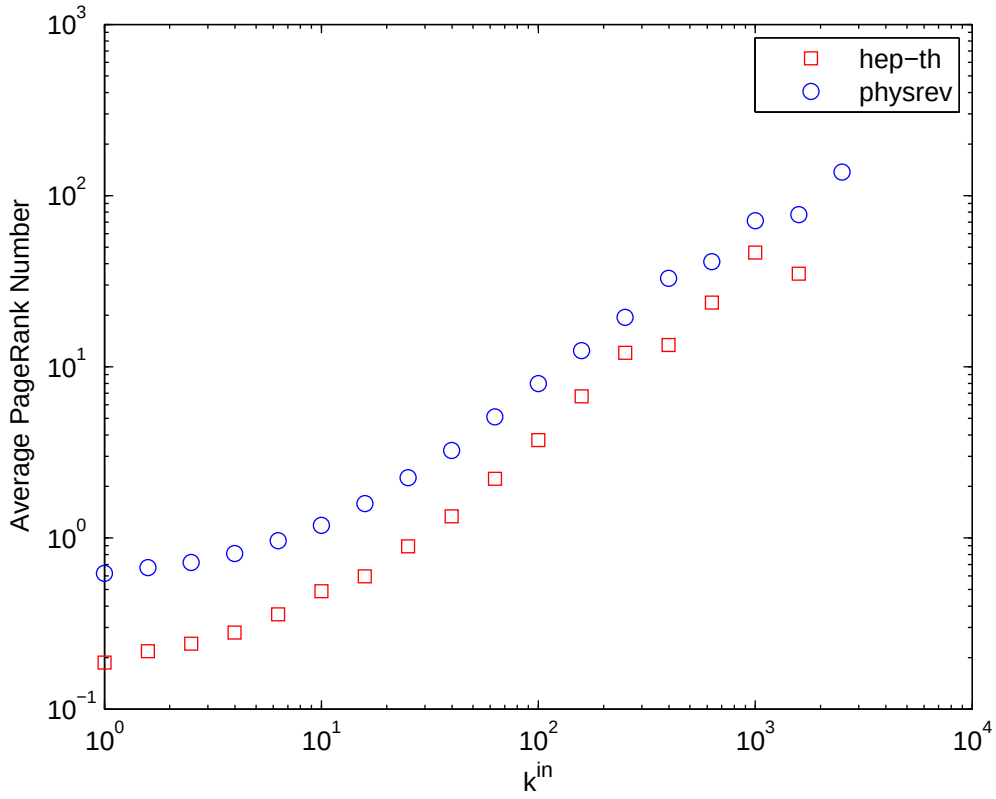

Figure 2.5: A scatter plot of the average PageRank as a function of the k^{in} for both the physrev and hep-th citation networks. In both cases, the average PageRank is well correlated with the in-degree. This arises because the larger the in-degree of a paper, the more likely it is easy to receive visits from a random walker in the network.

of directed loops that are typically present in other networks such as the world wide web. This means that random walkers diffusing on the network will tend to pile up on older papers, as there is no mechanism beyond random hopping to travel to earlier papers. Furthermore, aging significantly alters the spectral properties of the adjacency matrix which lie at the heart of the PageRank algorithm. The absence of directed loops means that the adjacency matrix can have only zero eigenvalues. To address these issues, we present the CiteRank

algorithm.

The CiteRank Algorithm

The success of the PageRank algorithm can be attributed, in part, to its ability to capture the behavior of people randomly browsing the network of web pages. The assumption that a typical web-surfer starts at a randomly selected webpage might be not completely unreasonable for the WWW, but it needs to be modified for citation networks. As all of us know, researchers typically start "surfing" scientific publications from a rather *recent* publication that caught their attention on a daily update of a preprint archive, a recent volume of a journal, or, perhaps, was featured in a news article in the popular media. Thus a more realistic model for the traffic along the citation network should take into account that researchers "surfing" the citation network preferentially start their quests from recent papers and progressively get to older and older papers with every step.

We introduce the CiteRank algorithm, an adaptation of the PageRank algorithm to citation networks. Our algorithm simulates the dynamics of a large number of researchers looking for new information. Every researcher, independent of one another, is assumed to start his/her search from a *recent* paper or review and to subsequently follow a chain of citations until satisfied or saturated with information. Explicitly, we define the following two-parameter CiteRank model of such a process, allowing one to estimate the traffic $T_i(\tau_{dir}, \alpha)$ to a given paper i. A recent paper is selected randomly from the whole population with a probability that is exponentially discounted according to the age

of the paper, with a characteristic decay time of τ_{dir}.

At every step of the path, with probability α the researcher is satisfied/saturated and halts his/her line of inquiry. With probability $(1 - \alpha)$ a random citation to an adjacent paper is followed. The predicted traffic, $T_i(\tau_{dir}, \alpha)$, to a paper is proportional to the rate at which it is visited (downloaded) if a large number of researchers independently follow such a simple-minded process.

While we interpret the output of the CiteRank algorithm as the traffic, its utility ultimately lies in the ability to successfully rank publications. High CiteRank traffic to a publication denotes its high relevance in the context of currently popular research directions, while the PageRank number is more of a "lifetime achievement award".

However, the more refined CiteRank algorithm surpasses both the conventional ranking, by number of citations, and the PageRank in its characterization of relevancy. Unlike the PageRank algorithm, the age of a citing paper is intrinsically accounted for: the effect of a recent citation to a paper is greater than that of an older citation to the same paper. Recent citations indicate the relevancy of a paper to current lines of research.

An algorithmic description of the aforementioned model can be understood as follows. Let ρ_i represent the probability of initially selecting the i^{th} paper in a citation network:

$$\rho_i = e^{-age_i/\tau_{dir}} \tag{2.2}$$

The probability that the researcher will encounter a paper by initial selection alone is given by $\boldsymbol{\rho}$. Similarly, the probability of encountering the paper after

following one link is $(1-\alpha)W \cdot \boldsymbol{\rho}$. The CiteRank traffic of the paper is then defined as the probability of encountering it via paths of any length. That is, given an initial distribution of new papers, $\boldsymbol{\rho}$, and transfer matrix, W, the CiteRank traffic is given by:

$$\boldsymbol{T} = I \cdot \boldsymbol{\rho} + (1-\alpha)W \cdot \boldsymbol{\rho} + (1-\alpha)^2 W^2 \cdot \boldsymbol{\rho} + \cdots \tag{2.3}$$

Practically, we calculate the CiteRank traffic on all papers in our dataset by taking successive terms in the above expansion to sufficient convergence ($< 10^{-10}$ of the average value).

Parameters of the CiteRank Model

In order to assess the viability of this ranking scheme and to select optimal parameters (τ_{dir}, α), we need a quantitative measure of its performance on real citation networks. Of course, evaluating the performance of any ranking scheme is a delicate, but often necessary, matter. One way to select the best performing α and τ_{dir} is to optimize the correlation between the predicted traffic, $T_i(\tau_{dir}, \alpha)$ and the actual traffic (e.g., downloads). Unfortunately, the actual traffic data for scientific publications are not readily available for these networks. However, it is reasonable to assume that traffic to a paper is positively correlated with the number of new citations it accrues over a recent time interval, Δk_{in}.

For lack of better intuition we first assume a linear relationship between actual traffic and number of recent citations accrued. This corresponds to a simple-minded scenario in which every researcher downloading a paper will,

with a certain small probability, add it to the citation list of the manuscript he/she is currently writing. It should be noted that we make no attempt to model network growth.

In order to compare CiteRank with actual citation accrual, we constructed an historical snapshot of both networks used in this study. In both cases, the most recent 10 percent of papers are pruned from the network. This corresponds to the last 4 years (2000-2003) in the physrev network and last 1 year in the hep-th network. The CiteRank traffic, T_i, of the remaining 90 percent of the papers is then evaluated and correlated with their actual accrual of new citations, Δk_{in}, originating at the most recent 10 percent of papers. The linear correlation of CiteRank traffic with recent citations for both networks is presented in Fig. 2.6.

Despite these significant differences in the nature of the networks considered, the general features of their correlation contours are outstandingly similar. In both cases, a single sharp peak in correlation is evident for particular values of the parameters. The value of the optimal parameters for both networks are:

hep-th: $\alpha = 0.48$, $\tau_{dir} = 1$ year

physrev: $\alpha = 0.50$, $\tau_{dir} = 2.6$ years

Remarkably, the value of α is nearly the same for the rather different networks considered and is in agreement with that proposed in [38] on purely empirical grounds. The difference in optimal parameter τ_{dir} for these networks is in agreement with the common-sense expectation of faster response time (and hence faster aging of citations) in preprint archives compared to peer-reviewed

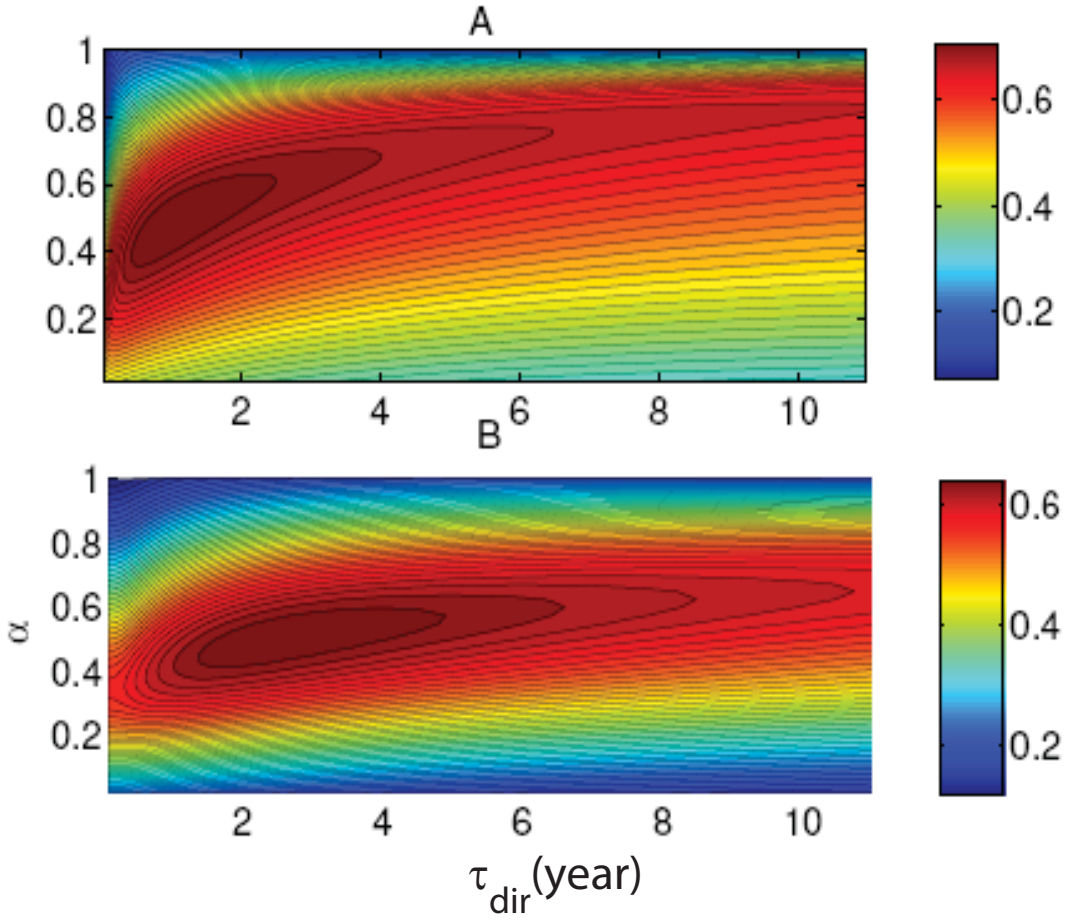

Figure 2.6: The Pearson (linear) correlation coefficient between the number of recent citations accrued (Δk_{in}) and CiteRank traffic (T_i) is calculated over the parameter space of the CiteRank model for the hep-th (A) and physrev (B) network. Both networks exhibit peaks in correlation coefficient in the α-τ_{dir} plane. The highest correlation is achieved for $\alpha = 0.48$, $\tau_{dir} = 1$ year in the hep-th network and $\alpha = 0.50$, $\tau_{dir} = 2.6$ years, in the physrev network.

publications. Another feature of Fig. 2.6 is that, in both networks, large values of the correlation coefficient are concentrated along a diagonally-positioned ridge. In other words, the best choice of α for a given τ_{dir} seems to rise linearly with τ_{dir}, a behavior that will be revisited later in this chapter.

While the correlation contour plots shown in Fig. 2.6 are a promising indication that the CiteRank model of traffic with optimized parameters provides a good zero-order approximation to the actual traffic along a citation network, they are to some extent predicated on the assumption of a linear relationship between actual traffic and Δk_{in}. One might readily ask how this model fares in the absence of such an assumption. While the assumption of a *linear* relationship may be unreasonable, a positive, monotonic relationship between these quantities is certainly expected. There is a statistical correlation method precisely adapted for such a situation, namely, the Spearman rank correlation. Under this relaxed correlation measure, only the rank of T_i are correlated with the rank of Δk_{in}. Numerical changes in T_i that do not lead to reordering have no effect on the value of the rank correlation coefficient. Another rationale for using rank correlations is that our ultimate goal is ranking publications, not modeling the traffic. Thus, we are currently not interested in individual T_i's, but only in their relative values. Spearman correlation contour plots are constructed for both networks and shown in Fig. 2.7.

The optimal values for both networks are:

hep-th: $\alpha = 0.31$, $\tau_{dir} = 1.6$ year

physrev: $\alpha = 0.55$, $\tau_{dir} = 8$ years

These results roughly confirm the prediction of $\alpha \sim 0.5$ from fig. 2.6, however there is a more appreciable discrepancy in τ_{dir} between linear and rank correlation for both networks.

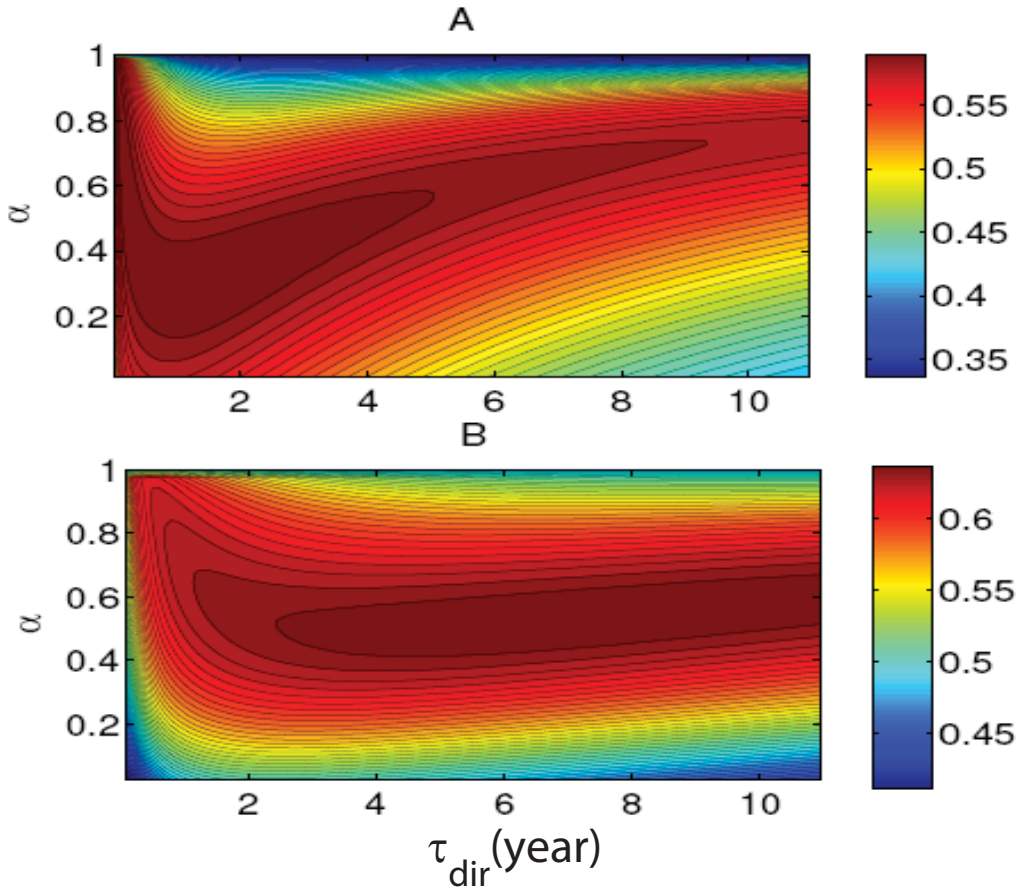

Figure 2.7: The Spearman rank correlation coefficient between recent citations accrued (Δk_{in}) and CiteRank traffic (T_i) for the hep-th (A) and physrev (B) network. Both networks exhibit similar behavior. There are more extended regions of good correlation relative to the linear correlation contours of fig. 2.6. This broadening is expected as a consequence of the more relaxed correlation measure. The highest rank correlation occurs for $\alpha = 0.31$, $\tau_{dir} = 1.6$ years, in the hep-th network and $\alpha = 0.55$, $\tau_{dir} = 8$ years, in the physrev network.

Qualitative Comparison of CiteRank to PageRank

We apply the CiteRank algorithm to the physrev and hep-th citation networks. The results for the top ten publications for both networks are show in tables 2.1 and 2.2.

A qualitative examination of CiteRank performance over the unmodified PageRank algorithm can be accomplished by direct comparison on the networks in question. As an example of this, the scatter plot of CiteRank vs. PageRank for all papers in the physrev network is shown in Fig. 2.8.

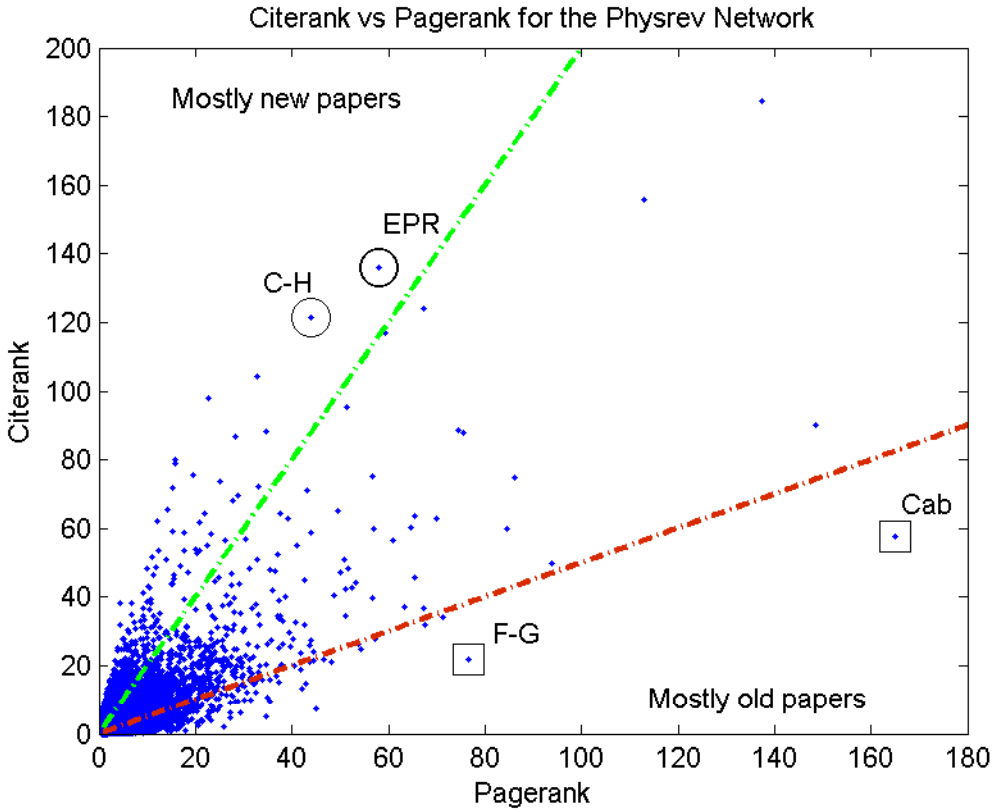

Figure 2.8: The scatter plot of CiteRank vs. PageRank for all papers in the physrev network. Two sectors of the data are distinguished according to the CiteRank to PageRank ratio: $\frac{CR}{PR} > 2$ (above the dashed green line) and $\frac{CR}{PR} < 1/2$ (below the dashed red line). The average publication year of papers in these sectors is 2000 and 1973, respectively. The sophistication of CiteRank goes beyond simple age classification, however. Particular examples that illustrate this sophistication are marked above and discussed in the main body of the text.

The positive correlation between the two algorithms is clearly evident in

the plot. Two sectors of the data are distinguished according to the ratio of CiteRank to PageRank, $\frac{CR}{PR}$. Papers with a relatively large (small) ratio have been marked comparatively higher (lower) by the aging effects inherent to the CiteRank algorithm. These sectors are distinguished in Fig. 2.8 above (below) the dashed green (red) line, respectively. In accordance with our claim that the CiteRank algorithm ranks papers of current relevance in research higher, the average publication year of papers in the high- (low-) ratio sector is found to be 2000 (1973). Of course, CiteRank is more sophisticated than a simple re-ranking according to publication age. For one thing, recent citations contribute greatly to a paper of *any* age. A particularly good example of this is the famous 1935 Einstein, Podolsky, Rosen paper [39](EPR) which receives both a large CiteRank and $\frac{CR}{PR}$ ratio despite its age. A quick glance at citing papers reveals approximately fifty citations to this paper throughout this year (2007) alone, indicating its clear connection to current lines of research. Another notable example of a publication in the high-ratio sector is a review paper of out-of-equilibrium pattern formation[40] by Cross and Hohenberg (C-H). It has a high ratio, despite being significantly older than papers in this sector. This paper is a good example of a class of *review* papers that serve to summarize the state of affairs regarding a particular topic that is of continuing interest to research. They are clearly of great use to the modern researcher and thus obtain their high $\frac{CR}{PR}$ by virtue of recent citations. Of further interest are papers that received high PageRank (lifetime achievement) but have a relatively low $\frac{CR}{PR}$ ratio. The wealth of these papers cover undeniably fundamental advancements in physics. Two explicit examples of this are the Feynman, Gell-Mann paper on Fermi interactions [41] (F-G) and the well known Cabibbo paper on leptonic

decay [42] (Cab). The low $\frac{CR}{PR}$ of these papers can be explained by a dearth of recent citations, which in turn is likely due to the incorporation of fundamental discoveries and advancements into textbooks and other published works that include more recent developments in addition to historical context.

A better physical understanding of the sophistication of the CiteRank algorithm may be gleaned from a simple quantitative analysis of the traffic dynamics in terms of its parameters. In both panels of Fig. 2.6, over a broad range of parameters, the optimal value of $\alpha(\tau_{dir})$ for a given value of τ_{dir} is positively correlated with τ_{dir}. This is an indication that these two parameters are entangled. In fact, this is to be expected as it is some admixture of the two parameters which leads to the exposure of a given paper to the researcher. An intuitive picture of this entanglement can be understood in terms of the penetration depth, which is a measure of how far back in time a random researcher following rules of the CiteRank algorithm is likely to get. The penetration depth is affected by both τ_{dir} - the average age of the initial paper at which he/she started following the chain of citations, and $1/\alpha$ - the mean number of steps on this chain of citations. For small τ_{dir} and large α, the penetration depth is small, implying that only very recent papers receive traffic. On the other hand, for large τ_{dir} and small α, the penetration depth is very large, indicating that most of the traffic is directed towards older papers.

To better understand how α and τ_{dir} influence the age distribution of CiteRank traffic, we performed the following quantitative analysis. Let $T_{tot}(t)$ denote the CiteRank model traffic to papers written exactly t years ago, where the meaning of the additional subscript shall be made clear in the lines that follow. As described by Eq. 2.3, two distinct processes contribute to $T_{tot}(t)$.

The first is the "direct" traffic $T_{dir}(t)$ due to the initial selection of papers in this age group, which is proportional to $\exp(-t/\tau_{dir})$. The second is the "indirect" traffic $T_{ind}(t)$ arriving via one of the incoming citation links, which is given by $T_{ind}(t) = (1-\alpha)\int_0^t T_{tot}(t')P_c(t',t)dt'$, where $P_c(t',t)$ is the fraction of citations originating from papers of age t' that cite papers of age t. It should be noted that $P_c(t',t)$ is an *empirical* distribution and, as such, is a *measured* property of the citation network under consideration. The integral takes into account the fact that incoming links to papers of age t can originate from all possible intermediate times. According to [35] and our own findings, $P_c(t',t)$ is reasonably well approximated by the exponential form $\frac{1}{\tau_c}\exp(-(t-t')/\tau_c)$. Taking the Fourier transform of the equation $T_{tot}(t) = T_{dir}(t) + T_{ind}(t)$, we have

$$T_{tot}(\omega) = T_{dir}(\omega) + (1-\alpha)T_{tot}(\omega)P_c(\omega). \tag{2.4}$$

Solving Eq. 2.4 and taking the inverse Fourier transform, yields

$$T_{tot}(t) \sim (\tau_c - \tau_{dir})\exp(-t/\tau_{dir}) + (1-\alpha)\tau_{dir}\exp(-\alpha t/\tau_c). \tag{2.5}$$

Thus, the traffic arriving at the subset of papers of age t is given by the superposition of two exponential functions.

Having an approximate analytical expression for $T_{tot}(t)$, we are now in a position to better understand what determines the optimal values of α and τ_{dir}.

Fig. 2.9 shows the age distribution of the number of recently acquired citations, Δk_{in}, for papers in the physrev dataset. The approximate CiteRank traffic, given by Eq. 2.5, is also displayed. It is calculated using the empirically

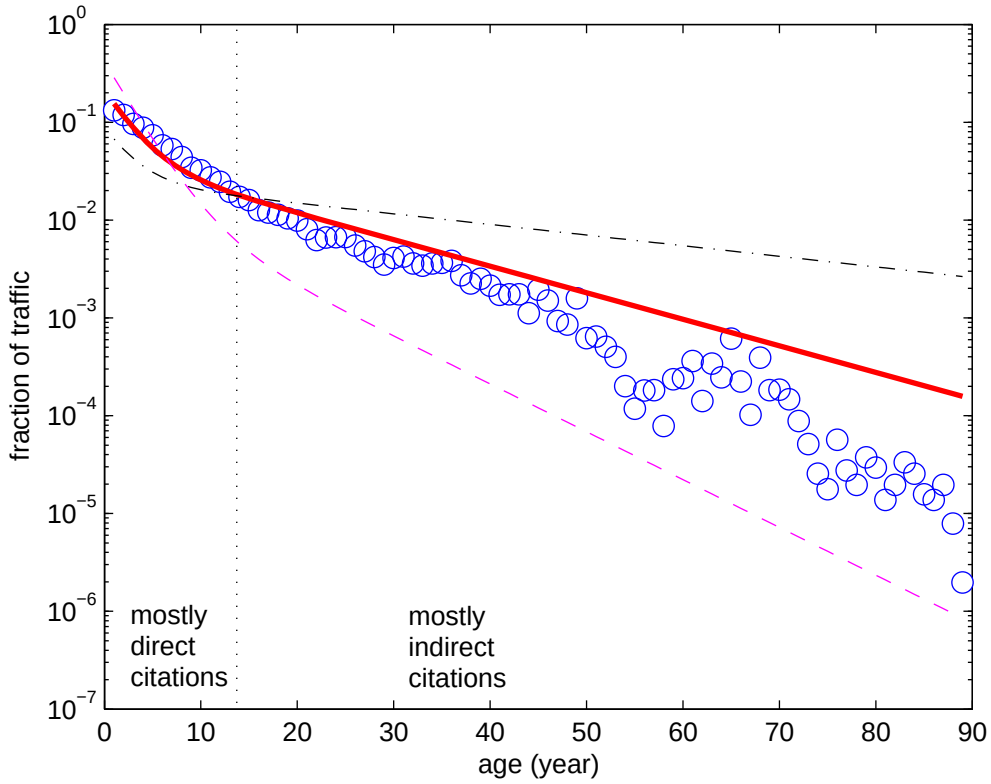

Figure 2.9: The age distribution of newly accrued citations Δk_{in} (blue) for the physrev network. Theoretical predictions [2.5] for the CiteRank traffic are calculated for the optimal $\tau_{dir} = 2.6$ and three values of $\alpha = 0.2$ (dot-dashed line), 0.5 (thick solid line), and 0.9 (dashed line). In agreement with Fig.2.6, the optimal value, $\alpha = 0.5$, provides the best agreement with Δk_{in}. All curves are normalized so that the sum of all data points is equal to 1.

determined value $\tau_c = 8$ years, optimal $\tau_{dir} = 2.6$ years and three values of $\alpha = 0.2, 0.5$ and 0.9. As one would expect, the profile of $\langle \Delta k_{in} \rangle$ vs t best agrees with the CiteRank plot for the optimal value $\alpha = 0.5$ [**?**]. Fig. 2.9 also provides some clues to the positive correlation between near-optimal choices of α and τ_{dir}, visible as diagonal "ridges" in Fig. 2.6 A and B. Indeed, if the value of α is chosen to be large, the contribution from the second term is

diminished; the use of a larger value of τ_{dir} can partially compensate for the loss of CiteRank traffic to older papers, and would thus be in reasonably good agreement with the Δk_{in} data.

Another encouraging observation is that, like Eq. 2.5, the age distribution of recently acquired citations shown in Fig. 2.9 has two regimes characterized by two different decay constants of about 5 and 16 years, with a crossover point around $t = 15$ years. Our interpretation of this fact is that papers are found and cited via two distinct mechanisms: researchers can either find a paper directly or by following citation links from earlier papers. For each of these mechanisms, the probability that a given paper is found decays with its age but the characteristic decay time for the direct discovery is shorter. While very recent papers, especially the ones altogether lacking citations, are for the most part discovered directly, older papers are mostly discovered by following citation links.

2.3 Conclusion and Outlook

Understanding the dynamics of citation networks is a challenging problem that touches on methods of statistical physics and the study of complex systems. Ranking publications is a difficult but necessary task that is vital to the search and navigation of the immense and ever growing body of scientific work. As we have shown, current methods of ranking impact of publications suffer from a number of undesirable features. The traditional method of ranking by incoming citations is shown to be both overly democratic, in treating all citations equally, and preferentially weighted towards older papers. Ranking according

to the PageRank algorithm addresses the the issue of over democratization via self-consistent popularity, but does not explicitly account for aging. Thus, while PageRank is useful to gauge the lifetime achievement of publications, it is not a good ranking algorithm to gauge the relevance of a publication to modern research. The CiteRank algorithm successfully addresses issues of aging and self-consistent popularity in a manner that is adaptable to natural variations in the properties of citation networks (such as aging timescale and degree distribution).

Motivated by the interpretation of CiteRank as a model of traffic, the age-dependent citation structure of the networks was analytically modeled using a mean-field method that was shown to be in good agreement with the empirical structure. The simple model reveals two distinct age regimes corresponding to likely citation via direct and indirect methods of discovery.

As with all real citation network studies, the results we obtain for real-world citation networks are susceptible to effects of data incompleteness, as the networks do not include citations to publications in external journals. In some extreme cases, this may lead to spurious ranking, particularly for the case of an esteemed publication with only a few outgoing citations to the known network. This effect could be partially alleviated with the inclusion of immediate downstream publications external to the network, whose CiteRank might be ignored.

Future extensions of this work might be developed to address the estimation of journal impact. One future direction of study that seems promising and feasible is the evolution of a CiteRank over long periods of time. For large citation networks this could be achieved by time-resolving the network

to points throughout its history.

Title	**Ref**	CR	PR	KR	k^{in}
Self-Consistent Equations Including Exchange and Correlation Effects	Phys. Rev. 140 A1133 (1965)	1	3	1	3104
Inhomogeneous Electron Gas	Phys. Rev. 136 B864 (1964)	2	4	2	2340
Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?	Phys. Rev. 47 777 (1935)	3	22	84	492
Self-interaction correction to density-functional approximations for many-electron systems	Phys. Rev. B 23 5048 (1981)	4	15	3	2079
Pattern formation outside of equilibrium	Rev. Mod. Phys. 65 851 (1993)	5	46	19	829
Self-organized criticality: An explanation of the 1/f noise	Phys. Rev. Lett. 59 381 (1987)	6	21	30	699
Bose-Einstein Condensation in a Gas of Sodium Atoms	Phys. Rev. Lett. 75 3969 (1995)	7	97	16	874
Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels	Phys. Rev. Lett. 70 1895 (1993)	8	250	82	495
Ground State of the Electron Gas by a Stochastic Method	Phys. Rev. Lett. 45 566 (1980)	9	32	4	1778
Theory of Superconductivity	Phys. Rev. 108 1175 (1957)	10	2	5	1364

Table 2.1: The top ten articles according to CiteRank in the physrev citation network. For each article, the CiteRank (CR), PageRank (PR), k^{in} Rank (KR) and number of citations, k^{in} are displayed.

Title	**Ref**	CR	PR	KR	k^{in}
The Large N Limit of Superconformal Field Theories and Supergravity	Adv. Theor. Math. Phys. 2 231 (1998)	1	2	1	2414
Anti De Sitter Space And Holography	Adv. Theor. Math. Phys. 2 253 (1998)	2	6	2	1775
An Alternative to Compactification	Phys. Rev. Lett. 83 4960 (1999)	3	11	9	1032
String Theory and Noncommutative Geometry	Jour. High En. Phys. 9909 32 (1999)	4	13	7	1144
Gauge Theory Correlators from Non-Critical String Theory	Phys. Lett. B428 105 (1998)	5	8	3	1641
Monopole Condensation, And Confinement In N=2 Supersymmetric Yang-Mills Theory	Nucl. Phys. B426 19 (1994)	6	1	4	1299
Dirichlet-Branes and Ramond-Ramond Charges	Phys. Rev. Lett. 75 4724 (1995)	7	3	6	1155
M Theory As A Matrix Model: A Conjecture	Phys. Rev. D 55 5112 (1997)	8	7	5	1199
String Theory Dynamics In Various Dimensions	Nucl. Phys. B 443 85 (1995)	9	4	8	1114
Large N Field Theories, String Theory and Gravity	Phys.Rept. 323 183 (2000)	10	29	11	807

Table 2.2: The top ten articles according to CiteRank in the hep-th citation network. For each article, the CiteRank (CR), PageRank (PR), k^{in} Rank (KR) and number of citations, k^{in} are displayed.

Chapter 3

Time-Critical Collaborative Document Voting Systems

3.1 Background

Many modern studies of collaborative filtering, ranking and recommendation systems owe a good deal to their academic ancestor, the study of voting systems. Early mathematical studies of voting systems were formalized in the eighteenth century in the context of methods of decision making of a body politic in an election. In general terms, an election describes the procedure by which constituents (or users) cast a ballot to exhibit their preference for a set of options. While the most common perception of an election is the simple scenario by which a single winner is declared by majority rule, a variety of election systems have been defined that break from this common paradigm. Many systems involve a ballot in the form of an ordered list of preferences. The problem of determining the most amenable choice/s from a list of many

such ballots is well connected to the modern problem of *rank aggregation* from a large set of partial rankings.

While an extensive examination of the advantages and drawbacks to various voting systems lies beyond the scope of this text, a good summary may be found in [43].

Recently, several online systems based on collaborative ranking have become popular as a means to identify time-critical information. These include news aggregation systems such as Digg, Reddit, and others. While the details of these systems may differ, they share a set of common dynamical features and as such are instances of a general class of *time-critical collaborative ranking systems*.

In these systems, users may introduce new items and cast their vote on existing ones. Frequently, the community of users is relatively persistent, while items are both rapidly introduced and effectively expire after a short period of time. The main output of many of these time-critical systems is a ranked list of current headlines that serves as a common starting point for the user community. As a consequence, popularity effects can be a significant factor in the ranking dynamics. That popularity effects play a role in user behavior is a well-accepted observation that many have noted in the scientific community [44, 45]. These negative popularity effects have been qualitatively discussed in the context of *information cascades* and *herding behavior*. Recent quantitative models of popularity effects due to social connections present in the Digg system have been studied by Lerman [46]. In contrast to these studies, we examine popularity effects in a more general model that is applicable to a wide variety of systems, even in the absence of explicit social connections.

We further address the time-critical rank performance of the system and its dependence on the population of users.

In this work we follow a reputation-based approach to time-critical collaborative ranking and investigate the dynamic interplay between quality and popularity through the introduction of simple user models. In particular, we address the question: What are the positive effects of popularity, if any? How well can a mixed user population successfully rank time-critical information?

Time-critical systems share much in common with standard collaborative ranking systems that exist in various forms throughout the web (e.g., collaborative movie ratings such as Movielens[47], Netflix [48]; collaborative product recommendation such as epinion.com, amazon.com; and many others) and have been a topic of intense study in several communities throughout the past decade. Despite this, relatively little attention has been paid to the time-critical performance.

Time-critical systems differ from standard user-based collaborative systems in several aspects. As is often the case with collaborative ranking, convergence of item ranks to their proper order is guaranteed only in the limit of large times. For news aggregation systems, this is clearly unacceptable as news items cease to be *news* in the large time limit. Furthermore, many machine learning approaches to collaborative systems are content-based and often require the existence of a training data set that is a *typical* representation of the data as a whole. Not only must such training sets be painstakingly built by hand, but in time-critical systems, the nature of typical data is subject to change and there is little guarantee that the content of the past will be a good representation of future content.

Some alternative approaches to collaborative ranking do not depend on content stability, but rather on user persistence, through the introduction of user reputation [49]. These reputation approaches are more amenable to time-critical systems, because user communities tend to be more stable. Nevertheless, they are not immune to the problem of a changing environment. Indeed, a robust reputation approach should be both content-blind and adaptive.

In the study that follows, we investigate a simple model of time-critical collaborative ranking for a mixed user population. We present a natural measure of user reputation that exploits the innate feedback between popularity and quality.

3.2 The Model

We consider a system of N users voting on M objects. Any user may cast a vote for any object once and only once. At each time step, ΔM new articles are introduced (by random users) into the system and the oldest ΔM articles, with age given by a_{max}, expire. Each article is randomly assigned a fitness f in the range $(0, 1)$.

The situation is depicted in figure 3.1. At each time step, in random order, each user casts a single vote for an article of their choice, with a probability dictated by their behavior. We define two types of users that vote stochastically according to simple behaviors. A *quality* user of quality β will vote for an article j with probability:

$$P_q = \frac{f_j^\beta}{\sum_{j'} f_{j'}^\beta} \tag{3.1}$$

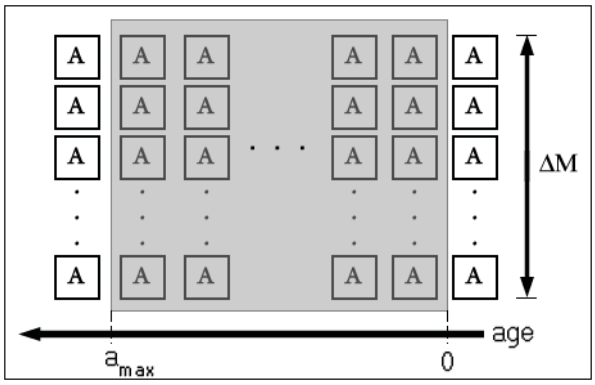

Figure 3.1: A diagram of article birth and expiration in the simulation. At each time step, ΔM new articles are born and the oldest ΔM articles with age given by a_{max} expire.

where the sum extends over all living articles for which the user has not yet voted. Similarly, a popularity user will vote for an article j with probability:

$$P_p = \frac{v_j}{\sum_{j'} v_{j'}} \tag{3.2}$$

where v_j is the total number of votes received by article j. A general population mixture of quality and popularity users may be described as:

$$N = n_p + \sum_{\beta} n_{q,\beta} \tag{3.3}$$

A rough measure of the time-critical performance of the system at any instant, for a given population mixture, is the rank correlation of article votes with article fitness for all living articles:

$$R_v(t) = rankcorr(\{f_j\}, \{v_j\}) \quad \forall_j \{a_j < a_{max}\} \tag{3.4}$$

For any population mixture, the rank correlation converges to a relatively stable value, $R_v(t_\infty)$, for simulation times well beyond the maximum article age. Unless otherwise specified, in the rest of this chapter, we will refer to the the steady state rank performance simply as the "rank performance" or "rank correlation" of the system.

It is important to note that the peak rank performance of an arbitrary population is limited by the size of population, N, and is a direct consequence of ties within the ranking. Indeed, even in the case of a fully cooperating population of users, the minimum number of votes required to provide a full ranking of M articles is given by the critical $n_v^c = M(M+1)/2$, where the ith best article receives exactly i votes. The constraint that no user may vote multiple times on the same article implies that population sizes $N < n_v^c$ will necessarily be suboptimal. Any comparison of performance across differing population sizes must take this limitation into account. For fully cooperating populations of sizes below or above this critical value, the best achievable rank correlation, R^{best} can be found from a simple algorithm where each vote is cast in such a way that the new correlation is maximized.

The best achievable rank correlation is displayed as a function of number of votes cast n_v, for several article set sizes , M, in fig. 3.2. As can be seen in the inset of the figure, the relevant quantity for any article set size is n_v/M, in accordance with the common sense notion that the number of votes necessary to achieve the best possible rank correlation scales linearly with the article set size.

Needless to say, in the case of real collaborative systems, explicit user cooperation is not typically present. However, the effective cooperation of any

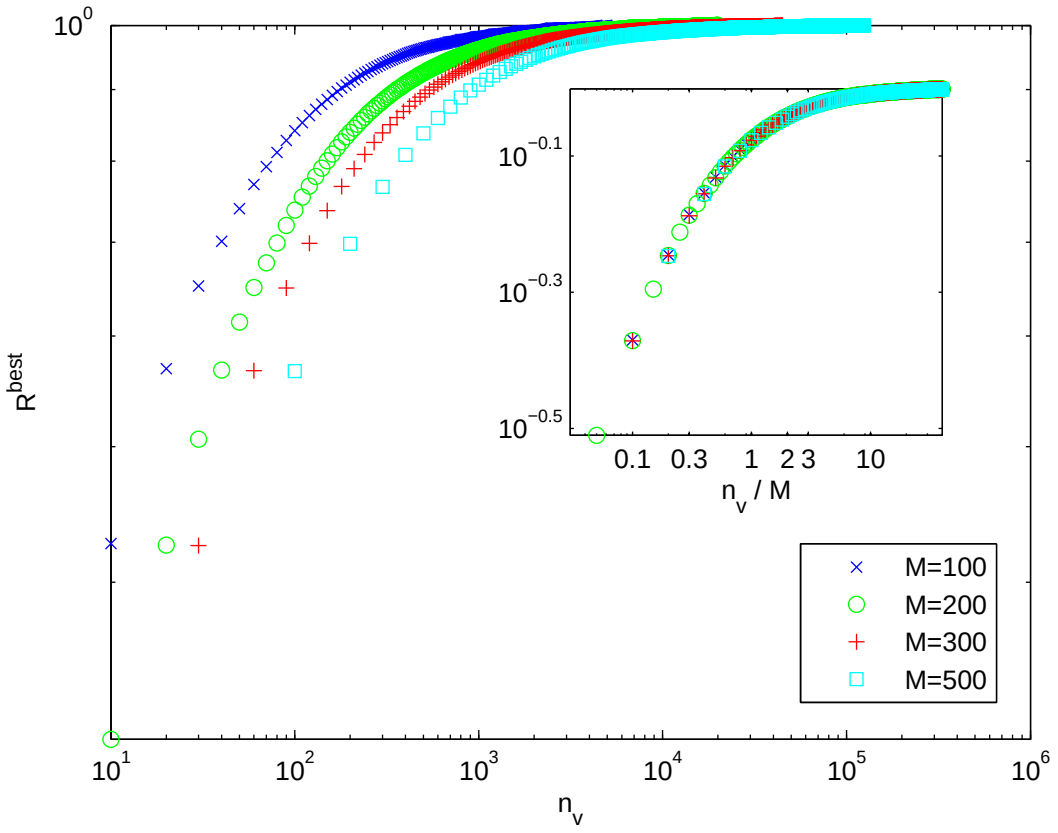

Figure 3.2: The best achievable rank performance given the number of votes. The x-axis is the number of votes cast, n_v and the y-axis is the maximal rank performance achievable by distributing n_v votes across M articles. The results are calculated, using the straightforward algorithm described in the text, for several cases where the number of articles is given by $M = 100, 200, 300, 500$. The inset shows the same figure with the x-axis scaled to n_v/M, revealing that only the ratio of number of votes to number of articles is relevant in determining the best achievable rank performance.

population may be defined as the ratio of its rank performance relative to the best possible, $C = R/R^{best}$.

For the simple case where all users in the population vote for the best article possible (all quality users with $\beta \to \infty$), the effective cooperation versus n_v/M

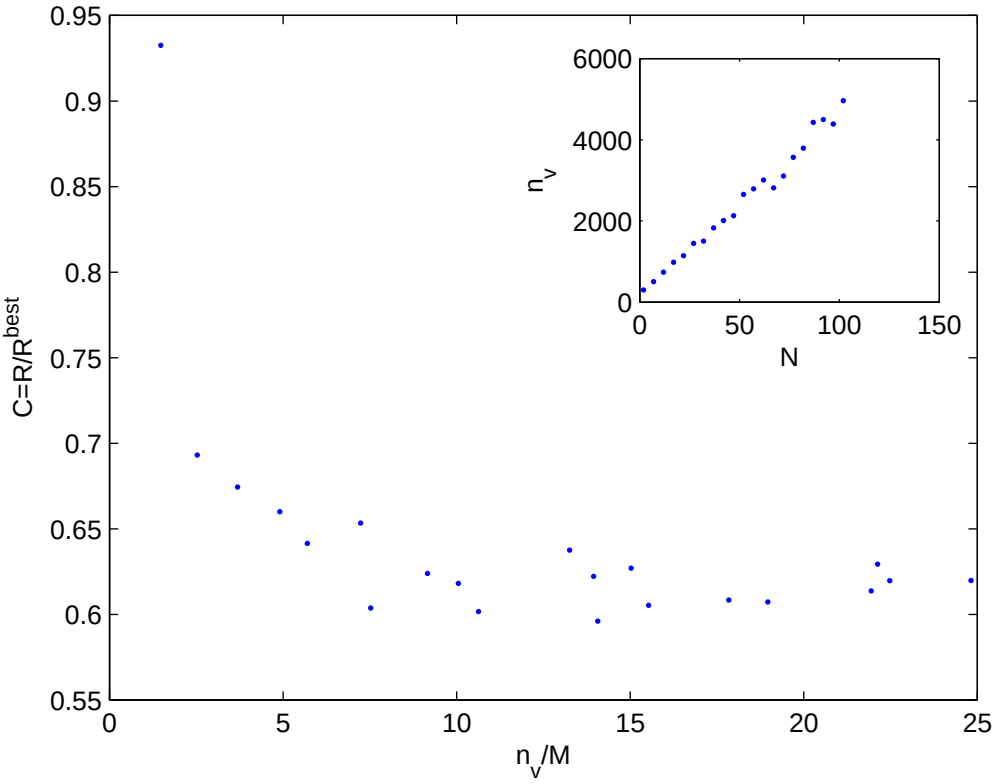

Figure 3.3: The effective cooperation versus the ratio of votes cast to article set size, n_v/M, for a population of "best" quality users ($\beta \to \infty$). The scatter plot is generated from several simulations of population sizes ranging from $N = 2$ to 200, with an article set size of $M = 200$. The number of votes cast scales linearly with population size, as can be seen from the inset. The effective cooperation of small populations is significantly higher, indicating that incidental cooperation is easier to achieve for small populations. For large populations, the cooperation tends to saturate about an average value of ~ 0.62 and is relatively insensitive to the size of the population.

is displayed in fig. 3.3. Apparently, small populations achieve an incidental cooperation more easily than large populations. For large populations, the cooperation is insensitive to the size of the population, with an average value of $C \sim 0.62$.

3.3 Negative Effects of Popularity

To understand the qualitative effects of popularity, we consider only simple population mixtures comprised of n_p popularity users and n_q quality users, all with quality $\beta = 1$. We define the quality fraction, $Q = n_q/N$, and determine how the steady-state rank correlation depends on quality fraction.

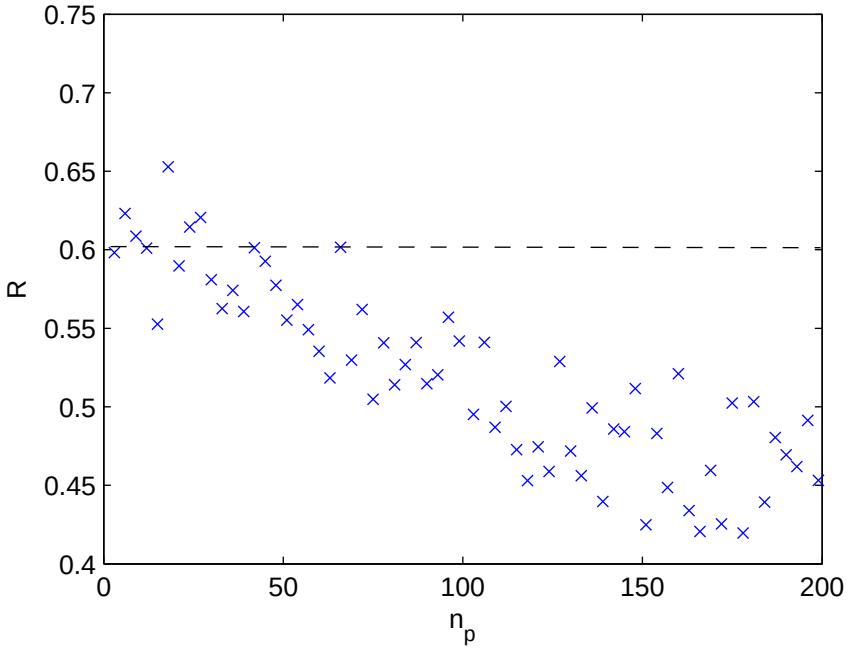

Figure 3.4: An example of negative popularity effects. The steady-state rank performance as a function of number of popularity users, n_p, for a simulation with constant number of quality users, $n_q = 100$. The dashed line indicates the average rank performance of quality users alone. In all cases, the addition of popularity users leads to a decrease in the rank performance of the system. The results shown are a typical example of the effect of increasing number of popularity users on the rank performance of the system.

The results, for a range of population fractions are displayed in figure 3.4. In all cases, it is clear that the overall rank performance of the system

decreases as the number of popularity users in the population grows. In other words, the existence of popularity always leads to a deleterious effect. This is particularly problematic because, in real world systems where top-ranked articles are displayed as an output, popularity effects are often unavoidable.

While quality users within the model behave independent of one another, popularity users by definition respond to the actions of the population as a whole. A popularity user responds both to the behavior of quality users and to the behavior of other popularity users. Thus, popularity users interact with the population and are susceptible to nonlinear effects that arise as a consequence of that interaction.

To quantitatively assess the nonlinear effects of popularity , we consider the case of a fixed number of popularity users and variable number of quality users of constant quality $\beta = 1$. Under these circumstances, we avoid the problem of comparing complex populations of differing size by examining the rank performance of the popularity users alone. To accomplish this, for each article i, we differentiate votes cast by popularity users from those cast by quality users:

$$v_i = v_i^p + v_i^q \tag{3.5}$$

The rank performance of popularity users is simply the steady state rank correlation of popularity votes with article fitness for all live articles, $R_v^p(t_\infty)$.

The results for $M = 1000$ articles and a fixed population of popularity users $n_p = 100$ are displayed in fig. 3.5. As the fraction of fitness users Q increases, the rank performance of votes cast by popularity users (displayed as (+)'s in the figure) improves. The effect is an example of *herding behavior*

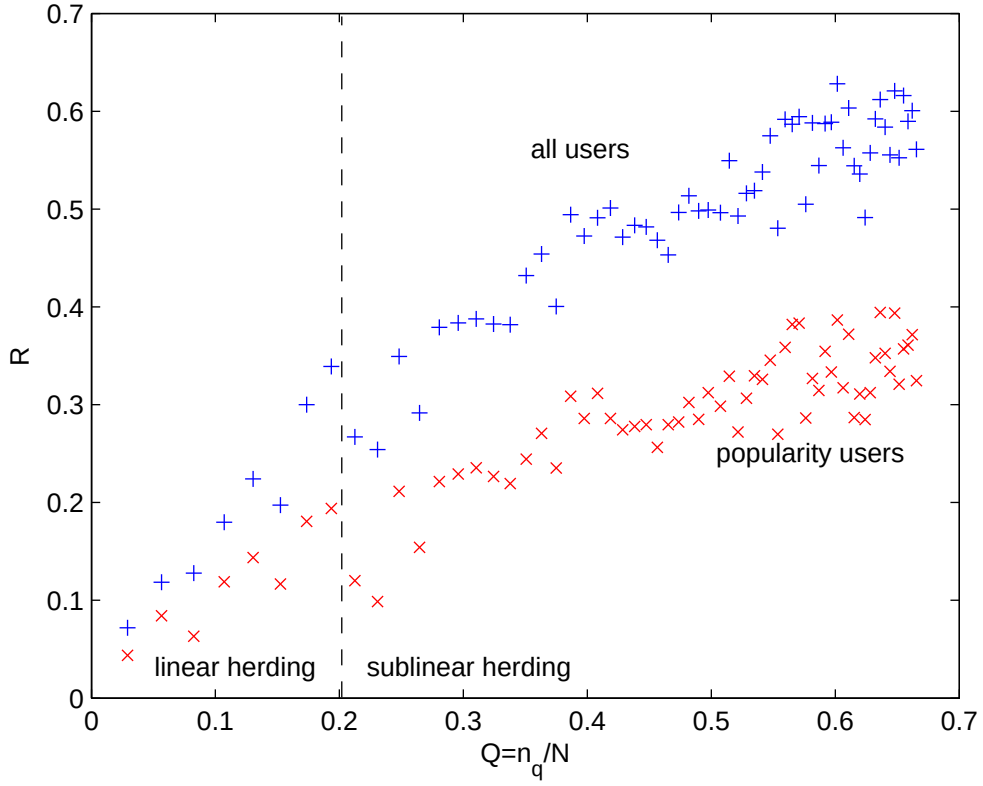

Figure 3.5: The rank performance versus the fraction of quality users, Q, in the population. The simulation was performed with a constant number of popularity users $n_p = 100$ and varying number of quality users n_q. The $(+)$'s display the rank performance for votes cast by the entire population of users. The (x)'s display the rank performance for votes cast by only the popularity users. As the number of fitness users increases, the rank performance of popularity users R_p increases, displaying a clear "herding" behavior as quality users guide popularity users to articles with high fitness. For small quality fraction, $Q \leq$, the herding effect is approximately linear with Q. For quality fractions $Q > 0.2$, the herding effect becomes sublinear, indicating that experts (or herders) are most effective when they constitute less than 20% of the population.

and may be intuitively described as follows. When the number of quality users is small, popularity users vote according to a random feedback loop. Initially,

the number of votes for all articles is approximately constant and popularity users are equally likely to vote for any article. Because the number of quality users is small, $v \sim v^q$ and there is relatively no correlation between the innate fitness of an article, f and the probability to receive a vote. As the system evolves, popularity users become arbitrarily biased in favor of articles with many votes, in a rich-get-richer phenomenon.

When the number of quality users is sufficiently large, however, the initial distribution of votes from quality users is biased towards articles with high fitness, while the distribution of popularity votes remains homogeneous. As the system evolves, popularity users respond to the bias of high fitness. In this way, quality users *herd* popularity users to higher fitness articles.

For quality fractions $Q \leq 0.2$, the herding is linear with Q, as indicated in the figure. For quality fractions $Q > 0.2$, herding becomes sublinear with Q, indicating that the addition of more quality users yields herding that is less effective. Slower herding behavior is a direct consequence of the saturating cooperation seen in fig. 3.3, as a large number of quality users have lower incidental cooperation. Evidently, a group of experts (quality users) yield the most "value" when they constitute less than 20% of the population.

While the addition of popularity users to a population negatively effects the rank performance of the system, the deleterious effects of popularity are mitigated by herding. To better understand the source of nonlinear herding effects, we consider a mean-field deterministic model of vote evolution.

3.4 Mean-Field Model for Mixed Populations

While the behavior of a numerical model of mixed populations is stochastic in nature, according to the definitions of user behavior, it is nonetheless possible to derive a mean-field deterministic model. In the derivations that follow, we adopt an article-centric approach and consider the number of quality- and popularity-user votes received by a particular article. The variable t denotes the age of the article we consider. When it is necessary to explicitly refer to the age of other articles, we will employ the notation $a_j[t]$ to refer to the age of the jth article when the ith article is age t. While this notation may seem confusing, it is employed solely as an explicit reminder that live articles in the model have varying age. In addition, we consider the time evolution of new articles after the system has been running for long periods of time (larger than several times the maximum article age), so that we may assume steady state dynamics.

Consider the number of quality votes received by the ith article as a function of time step:

$$v_i^q[t] = v_i^q[t-1] + \sum_{\text{q-users } m} (1 - A_{mi}) \frac{f_i}{\sum_j (1 - A_{mj}) f_j} \tag{3.6}$$

The first summation extends over all m quality users. The coefficient $(1 - A_{mi})$ is nonzero only when the mth user has not yet voted for the ith article. The sum in the denominator extends over all the articles j in the selection pool. In general this sum depends on the user in question, through the coefficent $(1 - A_{mj})$ as well as the fitness of the articles in the selection pool for that

user. In the mean-field case, we assume that the sum in the denominator is approximately constant:s

$$A = \sum_j (1 - A_{mj}) f_j \tag{3.7}$$

equivalent to the case where the fitness distribution of articles in the selection pool of all users is on the average the same. Given the quality fraction of the population $Q = n_q/N$, we replace the sum over quality users with $(QN - v_i^q[t-1])$, the amount of quality users that have not yet voted for the ith article, with the tacit assumption that $v_i^q[t] \leq QN$ so that this term is positive definite. The number of quality votes received by the ith article is thus

$$v_i^q[t] = v_i^q[t-1] + \frac{(QN - v_i^q[t-1]) f_i}{A} \tag{3.8}$$

or simply

$$v_i^q[t] = QN f_i/A + (1 - f_i/A) v_i^q[t-1] \tag{3.9}$$

The above equation is a finite geometric series with the solution:

$$v_i^q[t] = QN(1 - [1 - f_i/A]^{(t+1)}) \tag{3.10}$$

Similarly, we can calculate the number of popularity user votes received by the ith article:

$$v_i^p[t] = v_i^p[t-1] + \sum_{\text{p-users } m} \frac{(1 - A_{mi})(v_i^p[t-1] + v_i^q[t-1])}{\sum_j (1 - A_{mj})(v_j^p[a_j[t-1]] + v_j^q[a_j[t-1]])} \tag{3.11}$$

Where the first sum extends over all m popularity users. The variable $a_j[t]$ is the age of the jth article at time t. The sum in the denominator is simply the total number of votes cast up to time $t-1$ for all articles excluding those for which the mth user has already voted. Under the mean-field assumption, we treat this as a quantity independent of m and t:

$$B = \sum_j (1 - A_{mj})(v_j^p[t-1] + v_j^q[t-1]) \tag{3.12}$$

The sum over popularity users can be replaced by the quantity $((1-Q)N - v_i^p[t-1])$, so that the number of popularity user votes received by the ith article is

$$v_i^p[t] = v_i^p[t-1] + ((1-Q)N - v_i^p[t-1])\frac{v_i^p[t-1] + v_i^q[t-1]}{B} \tag{3.13}$$

In population growth models, the above equation is a form of an *inhibited growth* equation. Unlike in population growth models, however, the number of votes is guaranteed to be a monotonically increasing quantity. In general, such equations are not analytically tractable and can express a wealth of nonlinear dynamical phenomena. However, the model is certainly numerically viable. In comparison to the stochastic simulation, the mean-field model yields comparable predictions for vote evolution. Indeed, the herding that occurs with increasing population fraction can be examined using the mean field model.

Consider the popularity vote evolution of two articles with the same birth date, but vastly different fitness. Numerical computation of the mean field popularity vote of eq. 3.13 can be accomplished for any quality fraction, Q.

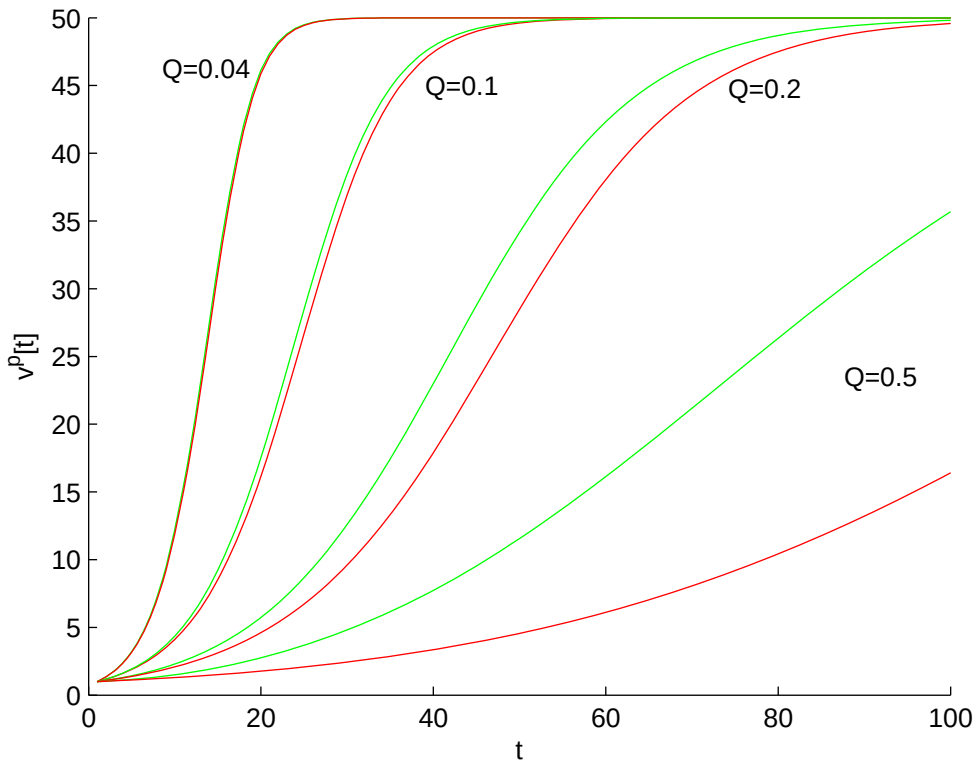

Figure 3.6: The expected number of votes as a function of article age as predicted by the mean-field model for articles with fitness $f = 0.1$ (red) and $f = 0.9$ (green) and different quality fractions. The vote evolution is a sigmoidal function of time, up to the saturation point (where the expected number of popularity votes is equal to the number of popularity users and thus cannot increase). The maximum difference between $f = 0.1$ and $f = 0.9$ sigmoids grows as the quality fraction, Q, increases.

In fig. 3.6, we examine the mean field prediction for popularity vote evolution of two articles with fitness $f = 0.1, 0.9$ for different quality fractions. When the quality fraction is low ($Q \leq 0.10$), there is little difference in the vote evolution curves for articles of differing fitness and hence, popularity users do not distinguish between articles of differing fitness. As quality fraction in-

creases, the difference in expected number of votes for articles of differing fitness also increases and herding occurs. For long times, the vote evolution of any article saturates to the maximum number of votes, which cannot exceed the number of popularity users. Prior to saturation point, the expected number of votes as a function of time is sigmoidal with a precise shape that is determined by the fitness and the mean field constants A and B. Using this model, one can gain a qualitative understanding of the dependence of herding behavior on the quality fraction, Q. Effective herding occurs when there is a significant difference between the vote evolution curves of articles with differing fitness. Because articles have a finite lifetime in the simulation, the evolution curves presented in fig. 3.6 will be abbreviated at $t = a_{max}$. Optimal herding occurs for Q values that are large enough to achieve significant difference between the vote evolution curves of all articles of differing fitness. As can be seen from the figure, for larger Q values, the vote evolution is slower and may be abbreviated by the article lifetime prior to achieving the maximal separation between article vote curves. This qualitatively explains the transition from linear to sublinear herding behavior seen in fig. 3.5.

With an approximate understanding of the dynamics of popularity, we now turn to the issue of whether it is possible to differentiate between users of quality based on their historical behavior.

3.5 Dynamic User Reputation

In real collaborative voting systems, the precise nature of the user population is not known. In general it is reasonable to expect the population to consist

of users with varying quality. Furthermore, we expect varying susceptibility of the population to the article popularity. In terms of the assumptions encompassed by our simple models of user behavior, we can therefore expect a real system to behave approximately similar to a population with quality fraction $Q = n_q/N$ of quality users with a distribution of user qualities $\{\beta\}_{n_q} = (\beta_1, \beta_2, \ldots, \beta_{n_q})$.

Given such a population with no a priori knowledge of its makeup and no knowledge of the innate fitness of articles, we would like to investigate whether it is possible, using only explicitly observable metrics of the system, to recover the makeup of the user population. Because we treat the user population as relatively persistent (i.e., all users participate actively in the system over long periods of time), we can exploit observations of user behavior of periods of time. Specifically, we wish to quantify the *reputation* of all users within the population, using a user based metric, S_i that we will refer to as the *user score*. A common sense notion of reputation suggests several properties to which the user score should adhere:

> Reputation should increase with user quality: The scores of any two users i and j should obey $S_i > S_j$ if user i has higher quality than user j, $\beta_i > \beta_j$. This implies that user score should increase monotonically with user quality.
>
> Reputation should be mutable and adaptive: In real-world systems the quality of a given user may change throughout time. In accordance with the first property, the user score must be capable of reflecting such a change. This implies that user score must not depend too heavily on the

"old" behavior of the user.

Reputation should be content-blind: Because the nature of article content is subject to change, the scoring of users must not depend in any way on the properties of article content.

We can define a user score that satisfies the above properties by considering the order in which an article accrues votes from users. Without much loss of generality, we assume a system with reasonable rank performance, so that high fitness articles are likely to receive a large number of votes. This is true even for many systems with somewhat low quality fractions, as a consequence of the herding behavior described in the prior section.

For illustrative purposes, consider the case of an article with high fitness f_i and isolate three quality users of quality $\{\beta_1, \beta_2, \beta_3\}$, in order of decreasing quality. At each time step, the probability that the three users have voted for article i is $(f_i^{\beta_1}/\sum_j f_j^{\beta_1}, f_i^{\beta_2}/\sum_j f_j^{\beta_2}, f_i^{\beta_3}/\sum_j f_j^{\beta_3})$ in order of decreasing probability. In the case where all three of the users vote for the article, the likeliest order of votes received by the ith article is simply $(m_1, \ldots, m_2, \ldots, m_3)$, where the dots represent possible votes by other users. In other words, the three users vote on the article in the same relative order as their qualities. The position of the mth user in this ordering can thus be recovered by counting the number of users that *follow* user m to article i.

We can further extend this notion to include all articles. In other words, we count the total number of users, F_m, that follow user m to any article. The number of followers has some features that make it an ideal candidate for user scoring. On the average, the F_m increases monotonically with β_m, the user

quality. Furthermore, F_m does not depend on the content of articles in the system. Unfortunately, F_m for any user grows with the lifetime of the system. Furthermore, in real systems, the activity of users may vary as a function of time. What really matters is not the number of followers, but the proportion of followers a user accrues. This suggests a normalization by the average number of followers, $F_m/\langle F_m \rangle$.

While the number of followers can be used as a measure of user reputation, it is in some sense, too democratic. A user that is followed by several other "good" users should be valued higher than a user that is followed by several "mediocre" users. This reasoning is in line with the notions of self-consistence introduced in the CiteRank system of chapter 2. Accordingly, we can weight each follower by it's user score to attain an update rule for the score of any user.

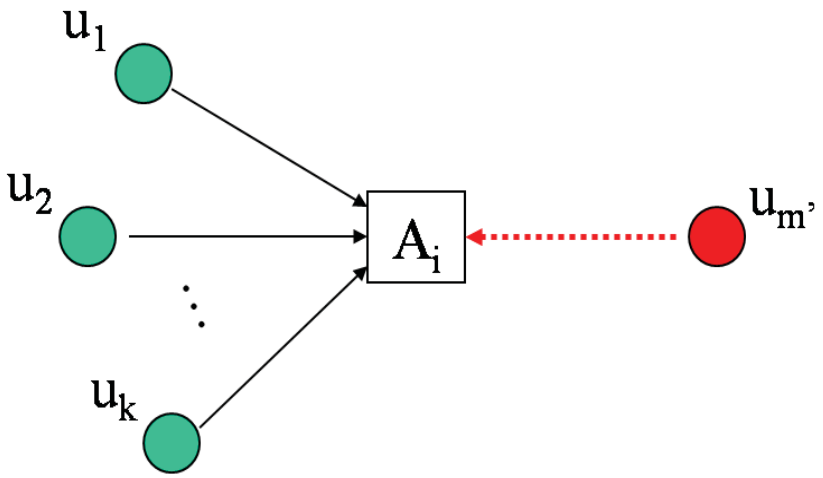

Figure 3.7: An illustration of the user score update rule. When a user $u_{m'}$ casts a vote to an article A_i, the score of users that previously voted for that article $(u_1, \ldots, u_k)$, are updated by the additive term $S^u_{m'}/\langle S^u \rangle$

As illustrated in fig. 3.7, for each vote that is cast $m' \to i$, by any user m' to any article i, we update the user scores of prior users that voted for article i:

$$\forall_{m \to i} : S^u_m = S^u_m + S^u_{m'}/\langle S^u \rangle \tag{3.14}$$

where the term $S^u_{m'}/\langle S^u \rangle$ is computed at the time the vote is cast. The above equation counts the number of followers of user m weighted by the relative score of the followers. At any instant, the relevant quantity of interest to determine the quality of a user m is $S^u_m/\langle S^u \rangle$.

Having devised a method to score users according to their quality, it is natural to ask how the user score compares to an empirical measure of user performance. One simple empirical measure of user performance is the probability that a user votes for the best article possible.

A comparison of user score to the probability a user votes for the best possible article is present in fig. 3.8 as a function of β/M. The probability to vote for the best article is estimated as the fraction of times a vote was cast to the best article over the total votes cast by that user. In real systems, of course, the fitness of articles are not known, and thus the probability to vote for the best article cannot be estimated. Nonetheless, as the figure affirms, the user score is a good proxy for user quality.

Given the user score as a proxy for user quality, it is possible to examine the effects of popularity in a different light. Specifically, we can ask how the addition of popularity users to a population effects the user score. To do so, we consider two groups of quality users: low quality users with quality β_L and high quality users with quality β_H. We examine how the average user score

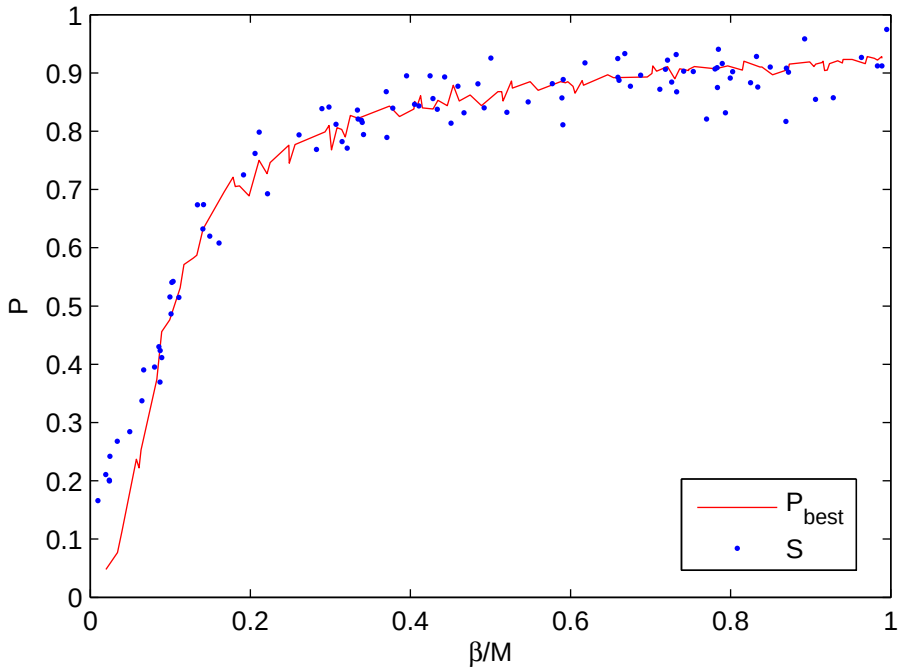

Figure 3.8: The solid line depicts the probability that a quality user of quality β/M will vote for the best article possible. The probability was estimated as the fraction of votes cast for best articles over total votes cast, in a numerical simulation containing $M = 150$ live articles, for 1000 time steps. The small circles depict a scatter plot of user score, S^u, (normalized for best fit) against user quality β/M for $n_q = 100$ quality users voting on $M = 150$ live articles, for 1000 time steps. The plot clearly indicates that user score, as defined in the text, is well correlated with the probability to vote for the best article, a reasonable empirical measure of user quality.

of low and high quality users changes as popularity users are added to the system.

The difference between average user score of high and low quality users as a function of the number of popularity users, n_p, is presented in fig. 3.9 for a system of $n_{\beta_L} = n_{\beta_H} = 25$. As can be seen from the figure, the addition of popularity users helps distinguish low quality users from high quality users.

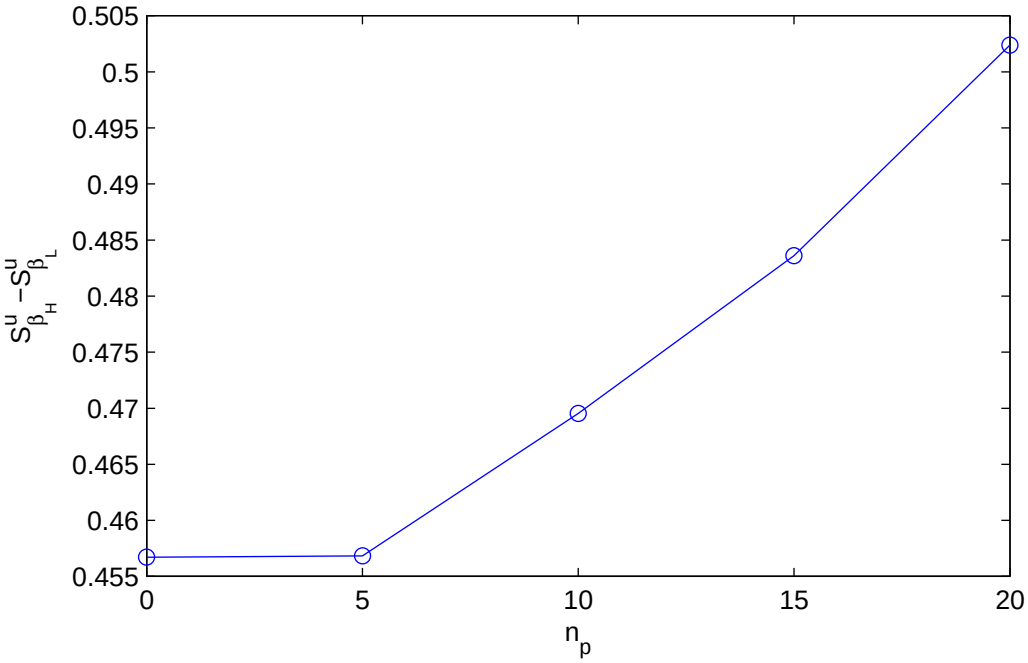

Figure 3.9: The difference between normalized user score of low ($\beta_L = 1$) and high ($\beta_H = 5$) quality users as a function of increasing number of popularity users, n_p. The simulation was performed for population size of $n_{\beta_L} = n_{\beta_H} = 25$ and averaged over several trials. The results shown are a typical example of the effect of popularity user addition and are qualitatively similar for a range of high and low user qualities and population sizes.

The results shown are a typical example of popularity effects on user score and are consistently reproducible for a range of user populations.

Intuitively, this effect can be understood by considering popularity users as a "background field of followers". On the average, high quality users receive more popularity followers than low quality users and consequently receive a high user score. In this way, popularity users provide useful feedback on the quality of other users in the population. In order to take advantage of this positive benefit of popularity, it is necessary to integrate user score into a

method for ranking articles. Given a method of assessing user quality, we would like to utilize this information to optimize the ranking of articles in the system.

3.6 An optimal method for ranking articles

Conventionally, the live articles in the system are ranked according to the raw number of votes received, irrespective of the type or quality of users from which the votes originated. A better method for ranking articles should take into account the quality or reputation of users.

To incorporate user reputation, we define the article score S_i^a. For each vote that is cast by a user m to an article i, the article score is updated according to:

$$S_i^a = S_i^a + \Theta(\frac{S_m^u}{\langle S_m^u \rangle} - s_c^u)\frac{S_m^u}{\langle S_m^u \rangle} \tag{3.15}$$

where the additive the additive term is evaluated at the time the vote is cast. The coefficient, $\Theta(S_m^u/\langle S_m^u \rangle > s_c^u)$ is a step function whose value is 1 when $S_m^u/\langle S_m^u \rangle > s_c^u$ and 0 otherwise and is included to ignore the noise effects from users with relative user score below some critical value. While the inclusion of this cutoff is not strictly necessary, it guards against the influence of a large number of low-scored users voting coherently, as occurs with populations of popularity users.

A comparison of the rank performance of the article score with that of raw votes alone is presented in fig. 3.10 as function of time, as an example of a typical scenario. The critical users score was heuristically chosen to be

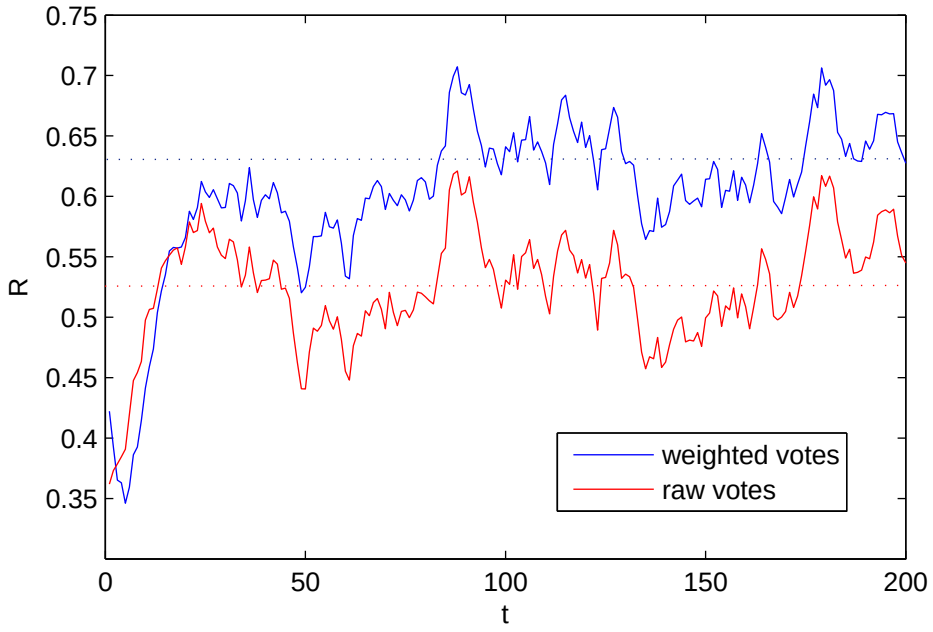

Figure 3.10: The rank performance over time for a typical system. The solid red line depicts rank performance using ranking by raw number of votes. The solid blue line depicts rank performance using weighted votes according to the article score scheme defined in the text, with a cutoff of $s_c^u = 0.8$. The article score ranking scheme, which assumes no prior knowledge about the user population, provides a significant performance advantage of (approximately $\sim 10\%$ in the above example). Dotted lines depict long-term averages for both ranking schemes. The use of user reputation in the article score scheme confers an advantage over ranking by only quality user votes.

$s_c^u = 0.8$. As can be seen from the figure, the performance of the article score is significantly better than the rank performance for raw votes alone, providing an average improvement in rank performance of $\sim 10\%$. Moreover, the article score rank performance is significantly better than the rank performance for votes cast by only quality users. For a population of quality user with varying qualities, $\{\beta_i\}$, this is to be expected, as the vote of a high quality user contributes more to an article's score than that of a low quality user, through the

use of user reputation (user score).

It should be noted that the effects of popularity are not excluded in the article score ranking scheme, as popularity users may contribute to the score of other users and may achieve high user score through herding. In fact, as we have seen, popularity users help to distinguish quality users from one another.

3.7 Conclusion

Understanding the dynamics of time-critical collaborative systems is an essential problem intimately linked to topics that span the fields of complex systems and computer science. The performance of such systems is intimately related to topics of recommendations, reputation, cooperation and rank aggregation. We have presented a general model of time-critical collaborative systems that is amenable to several real-world collaborative systems in operation today.

Popularity effects in these systems are often unavoidable and present a potentially worrisome obstacle to good performance. As we have seen, left untreated, popularity effects negatively impact the collaborative rank performance. At the same time, the existence of a small percentage of high quality users can mitigate popularity effects through herding behavior. We have derived a mean field model of vote evolution in mixed populations. A detailed examination of the dynamical properties of this model are a topic for future consideration.

Of further note is the finding that popularity effects can be harnessed, through the introduction of user reputation, to help distinguish the quality of users.

We have presented a method for estimating user reputation in a self-consistent manner that is both adaptable and independent of the nature or type of content within the system. Using this method, we have shown that the scoring of users is well correlated with their probabilistic behavior.

Finally, we have presented a method to dynamically rank documents that accounts for user reputation and provides significant improvement ($\sim 10\%$) in the rank performance of the system. The examination of systems with more realistic user behavior is a topic of future study.

Chapter 4

Dynamical Fluctuations and Noise in Protein Binding Network

4.1 Background

The study of dynamical fluctuations in complex systems has emerged as a topic of intense interest germane to the fields of biology [6], financial systems [50], traffic in information [51] and transportation [52] networks, and many others. Of particular interest is the nature of collective effects that arise as a consequence of the connectivity of the underlying network. By examining such fluctuations we can understand when the underlying network plays an important role and when, if possible, it may be ignored. A good candidate arena to study dynamical fluctuations is that of biomolecular processes taking place in cells.

A variety of biomolecular systems involve large numbers of interacting elements. The processes by which living systems sustain the functions necessary for life are quite complex. Within the cell this complexity arises from systems of gene expression, protein structure and protein-protein interaction. In the process of protein synthesis, expressed genes are transcribed from DNA to mRNA by RNA polymerase and are subsequently translated by the ribosome into protein expression. Expressed proteins undergo post-translational changes and interact with one another chemically to accomplish the biological functions of the cell. Throughout this process, several levels of complex interactions exist that may be viewed from a network perspective are briefly summarized here.

Genetic regulation is the process by which an expressed gene (or RNA or protein product) up- or down- regulates the expression of other genes. Genetic regulatory networks are therefore described by directed networks in which edges may carry positive or negative signs. These networks display rich topological features, such as positive and negative feedback loops, that can lead to a wealth of nonlinear dynamical behavior. The resulting gene expression is stochastic and may be viewed as the summary output of the complex processes that occur in genetic regulatory networks. From a functional standpoint, genetic regulatory networks help the cell dynamically respond to changing intra- and inter-cellular environments. Studies of fluctuations in genetic regulatory networks have recently been undertaken [53–55], though they are typically limited to understanding of simple topologies.

Metabolic networks describe the set of chemical reactions that break down molecules into metabolites (catabolic) or construct molecules from metabolites

(anabolic) with the assistance of enzymes in order to sustain the life of the cell. In the network representation, nodes are metabolites and edges are directed chemical reactions that transform one metabolite into another. Studies of fluctuations in metabolite concentration have recently been performed by the authors of [56]. These studies focus on linear metabolic pathways and small motif extensions therefrom.

On the other hand, reversible binding interactions occur between proteins and are governed by inter-protein affinity and kinetics. These binding interactions describe the formation of multi-protein complexes of n-mers from constituent monomer proteins. Proteins exist in several copy numbers within the cell and at any instant the number of all monomers and higher order complexes with the cell are stochastic quantities. The equilibrium values of protein copy numbers are dictated by the Law of Mass Action which states that the equilibrium concentration of a multi-protein complex is proportional to reactant concentrations. In this chapter, we study fluctuations in a network of reversible dimer interactions.

4.2 Network Description of Protein Protein Interactions

We adopt a network description of protein protein interactions (PPI) for a reversible dimer network of N distinct proteins represent by nodes. Edges are undirected and represent reversible binding interactions. The network is

described by an NxN symmetric adjacency matrix:

$$A_{ij} = \begin{cases} 1 & \text{if protein i binds to protein j;} \\ 0 & \text{otherwise.} \end{cases}$$

Each unique entry in the adjacency matrix correspond to the formation of a dimer (ij) from constituent proteins i and j. All dimers and monomers represented in the network may exist with multiple copy number within the cell. At any instant, the dynamical system may be described by the set of all protein instantaneous free ($\{F_i^*\}$) and bound dimer ($\{D_{ij}^*\}$) copy numbers. In this work, we treat the volume of the cell as constant and assume unit volume for simplicity. Thus, throughout this chapter, we will use the terms "copy number" and "concentration" interchangeably. The system at any instant is described by the state variables given by $\{C_i\}$, the set of total protein copy numbers, $\{D_{ij}^*\}$, the set of copy numbers for all dimers (ij) and $\{F_i^*\}$, the set of free protein copy numbers. At any instant, the system is constrained by mass conservation:

$$C_i = F_i^* + \sum_{i \neq j} D_{ij}^* + 2D_{ii}^* \tag{4.1}$$

so that F_i^* is not an independent variable. The latter term in the above equation pertains to homodimers, dimers that are formed from two copies of the same protein.

The dynamical equation that governs the rate of formation of a dimer from its constituent proteins is determined by chemical kinetics and is given by the

simple balance relation

$$\frac{d}{dt}D^*_{ij} = r^{(\text{on})}_{ij} F^*_i F^*_j - r^{(\text{off})}_{ij} D^*_{ij} \tag{4.2}$$

where the $r^{(\text{on})}_{ij}$ and $r^{(\text{off})}_{ij}$ are the kinetic rate constants for association and dissociation.

From the above equation it is clear that dynamic equilibrium is achieved when the left hand side vanishes so that

$$D_{ij} = \frac{F_i F_j}{K_{ij}} \tag{4.3}$$

where we have adopted the convention that unstarred variables denote equilibrium concentrations. The term $K_{ij} = r^{(\text{off})}_{ij}/r^{(\text{on})}_{ij}$ is referred to as the dissociation constant. At a given temperature, it is related to the binding energy, ϵ_{ij}, of a dimer by $\epsilon_{ij} = -k_B T ln(K_{ij}/K^{(0)})$.

An alternate and very useful examination of the system that does not explicitly involve time dependence can be accomplished with the introduction of the partition function

$$Z(\{C_i\}) = \sum_{\{D^*_{ij}\}} N_S(\{D^*_{ij}\}) \exp(-\sum_{i<j} \frac{\epsilon_{ij} D^*_{ij}}{k_B T}) \tag{4.4}$$

where the sum extends over all possible occupation states D^*_{ij} for a particular set of total protein abundances C_i. The combinatorial factor $N_S(\{D^*_{ij}\})$ counts the number of microstates of individual labeled proteins resulting in a given occupation state $\{D_{ij}\}$. For example, for a single heterodimer AB, $N_S(D^*_{AB})$

is the combinatorial factor:

$$N_S(D^*_{AB}) = \binom{C_A}{D^*_{AB}}\binom{C_B}{D^*_{AB}} D^*_{AB}! = \frac{C_A! C_B!}{D^*_{AB}! F^*_A! F^*_B!} \tag{4.5}$$

Using the Stirling's approximation for factorials in $N_S(\{D^*_{ij}\})$ one gets a concise expression for the free energy of a given occupation state $\{D^*_{km}\}$:

$$\begin{aligned} G &= \sum_{(km)\epsilon E} \{\epsilon_{km} D^*_{km} + k_B T D^*_{km}[\log(D^*_{km}) - 1]\} && (4.6) \\ &+ k_B T \sum_{i=1}^{N} \{F^*_i[\log(F^*_i) - 1] - C_i[\log(C_i) - 1]\} && (4.7) \end{aligned}$$

where the first sum runs over all E edges (dimers) and the second sum runs over all nodes (proteins) in the network. For brevity we have suppressed volume-dependent entropy terms that are not relevant to our discussion here.

The requirement of zero first derivative of the free energy with respect to dimer copy number relates equilibrium free (F_i) and bound (D_{ij}) concentrations in the system via $D_{ij} = F_i F_j / K_{ij}$, the same equilibrium relation found by considering the dynamical equation 4.2. Insertion of equilibrium dimer concentration into the conservation of mass (eq. 4.1), gives the Law of Mass Action (LMA):

$$F_i = \frac{C_i}{1 + \sum_j A_{ij} F_j / K_{ij}} \tag{4.8}$$

The above set of nonlinear equations, while not analytically tractable, readily yield a numerical solution (via iteration) for equilibrium free concentrations and, in accordance with eq. 4.3, equilibrium dimer concentrations as well. It

should be further noted that the above approach is further extendible to systems involving higher-order multi-protein complexes. For a simple illustration of this, consider a system of homogenous dissociation in which a protein A participates in dimer AB and trimer ABC. The free concentration of protein A is thus determined by $F_A = C_A/(1 + F_B/K_d + F_B F_C/K_d^2)$.

Throughout this chapter, we will make use of both the temporal and partition function formalisms in order to assess the effect of fluctuations.

We are generally interested in studying fluctuations of any of the independent state variables $\{C_i\}$ or $\{D^*_{ij}\}$. From the standpoint of their role in the cell, fluctuations in the former are quite different from fluctuations in the latter. Total protein concentrations in the cell are regulated by the various mechanisms that control protein expression as a function of the changing intra- and extra-cellular environment. Such regulation is a consequence of the complicated systems of genetic regulatory networks and is further influenced by stochasticity in systems of protein production and degradation. We will not study these systems here, but rather we adopt empirical estimates of total protein abundance and fluctuation as an input and examine the effect of fluctuating abundance on the protein binding network. We will refer to fluctuations in protein binding that arise as a result of total abundance fluctuations as *driven* fluctuations. Conversely, fluctuations in dimer concentration occur even when total protein concentration is static, as a result of the thermal kinetics of protein-protein collisions and binding energy. We will refer to this latter type as *spontaneous* fluctuations. It should be further remarked that these two types of fluctuations differ in typical magnitude and timescales. Driven fluctuations are usually somewhat larger than the spontaneous noise.

They also change relatively slowly on timescales (tens of minutes) that are large compared to the relaxation time of the mass action equilibrium which are rarely slower than seconds.

In general, for either type of fluctuation we are interested in calculating the noise in dimer concentration as quantified by the deviation from equilibrium, δD_{ij}. The magnitude of the noise is given by the second moment of, $\langle \delta D_{ij}^2 \rangle = \langle (D_{ij}^* - D_{ij})^2 \rangle$.

4.3 The Empirical PPI Network

To illustrate general principles with a concrete example, in this study, we used a curated genome-wide network of PPI in baker's yeast (*S. cerevisiae*), which, according to the BIOGRID database [57], were independently confirmed in at least two published experiments. A genome-wide set of protein abundances for bakers yeast was experimentally studied in [58, 59] during log-phase growth of the medium. Protein abundances were found to be highly variable, ranging over orders of magnitude from copy numbers of 50 up to 10^6. We retain only interactions in the PPI network between proteins of known total concentration, yielding a network of 4085 heterodimers formed from 1740 constituent proteins. The same network was previously used by others in [60, 61].

Another assumption (justified in these earlier studies) is that in the absence of large-scale experimental data on the strength of protein-protein interactions we use a set of evolutionary-motivated [61] dissociation constants:

$$K_{ij} = \frac{max(C_i, C_j)}{20} \tag{4.9}$$

for all interactions in our network. The evolutionary-motivated *association* is defined to be the weakest association strength necessary to keep a sizable fraction of the rate-limiting protein in a given interacting pair bound in the dimer. The denominator 20 is chosen to reproduce the average association strength, $\langle 1/K_{ij} \rangle = 1/5$ nM, in a set of experimentally measured dissociation constants from the PINT database [62], which are assumed to be representative for all biologically functional interactions among yeast proteins. This choice is further justified by a relative lack of sensitivity of equilibrium concentrations to details of assignment of dissociation constants to individual interactions (see Fig. 5 in Ref. [61]).

To incorporate the effect of higher-order multi-protein complexes in our study, we use a list of manually curated yeast protein complexes obtained from the MIPS CYGD [63] database (May 2006) formed from 3 or more constituent proteins with known total concentration, yielding a set of 81 multi-protein complexes formed from 2004 constituent proteins. While the constituents of multi-protein complexes are known, the detailed structure of their binding is not. In light of this, we assume a constant dissociation of $5nM$ for binding within multi-protein complexes and a minimum number of binding interactions necessary to cohere the constituents. Under these assumptions, it is possible to include the effect of multi-protein complexes on free and dimer equilibria using the techniques described in the previous section.

4.4 Driven Fluctuations

The effect of large scale (twofold) static changes in total protein abundance on network equilibria concentration has recently been studied by the authors of [61]. In the study, the authors performed a series of numerical experiments in which they systematically perturbed the individual protein abundance of all proteins within the network by a factor of two, $C_i \rightarrow 2C_i$, and observed the resulting cascading change in all equilibrium free concentrations. In particular, they found that such waves of perturbation exponentially decay with network distance from the source of the perturbation. Interpreting such large total concentration changes as cellular signals, the biological implications of this finding suggest that problems of cross-talk between functional systems are naturally mitigated by this decay. Interestingly, the authors noted that despite this general decay with network distance, the discovery of several proteins at larger network distances (~ 4) that exhibited significant free concentration changes ($\sim 20\%$) in response to the source perturbation. The authors dubbed the pair of source perturbed protein and responding protein as "concentration-coupled". While it is not clear that these concentrated coupled pairs are employed for real cellular signaling, their existence suggests that the network response to changes in protein abundance can be significant.

In contrast to large scale static perturbation, small perturbations of total protein concentration within the cell occur as a consequence of stochasticity in production and degradation and possibly as a result of noise originating in genetic regulatory systems. This latter variety of fluctuation may be considered as a type of intra-cellular *noise* in protein abundance. The relative response

of free concentration of proteins to changes in total abundance was originally presented by the authors of [60] for the response F_m to small static changes in total concentrations C_k. It is quantified by the matrix

$$\Lambda_{km} = \frac{\partial C_k}{\partial \log F_m} = D_{km}(1 - \delta_{km}) + C_k \delta_{km} \tag{4.10}$$

which follows directly from the mass conservation equation eq. 4.1. In the above formula and in the derivation that follows, we have omitted the homodimer case for the sake of simplicity, though it require only a trivial modification. It follows that an arbitrary number of small perturbations δC_m add up to

$$\frac{\delta F_i}{F_i} = \sum (\Lambda^{-1})_{im} \delta C_m \qquad . \tag{4.11}$$

Due to bilinear dependence of D_{ij} on F_i and F_j, one also has

$$\frac{\delta D_{ij}}{D_{ij}} = \frac{\delta F_i}{F_i} + \frac{\delta F_j}{F_j} \qquad . \tag{4.12}$$

Thus, in general, the amplitude of driven fluctuations is given by:

$$\frac{\langle \delta D_{ij}^2 \rangle}{D_{ij}} = D_{ij} \langle (\sum_k (\Lambda^{-1})_{ik} \delta C_k + \sum_m (\Lambda^{-1})_{jm} \delta C_m)^2 \rangle \tag{4.13}$$

The evaluation of the above expression requires the full matrix of cross-correlations $\langle \delta C_k \delta C_m \rangle$ which is currently experimentally unknown. Indeed, measurement of the noise profile of cross-correlations presents a significant challenge to the empirical community. Despite this current lack of knowledge, it is nonetheless still possible to examine two interesting cases of noise profiles.

For the simplest case of uncorrelated driving fluctuations $\langle \delta C_k \delta C_m \rangle \propto C_k^2 \delta_{mk}$ (the so-called intrinsic noise [6]), the driven response becomes:

$$\left(\frac{\langle \delta D_{ij}^2 \rangle}{D_{ij}} \right)_{\text{int}} \propto D_{ij} \sum_k [(\Lambda^{-1})_{ik} + (\Lambda^{-1})_{jk}]^2 C_k^2 \tag{4.14}$$

Alternatively, for completely coherent driving fluctuations (so called extrinsic noise), $\langle \delta C_k \delta C_m \rangle \propto C_k C_m$, the driven response becomes:

$$\left(\frac{\langle \delta D_{ij}^2 \rangle}{D_{ij}} \right)_{\text{ext}} \propto D_{ij} \sum_k [((\Lambda^{-1})_{ik} C_k)^2 + ((\Lambda^{-1})_{jk} C_k)^2 + 2\Lambda_{ik}^{-1} C_k \sum_m \Lambda_{jm}^{-1} C_m] \tag{4.15}$$

Empirical examinations of noise in protein abundance were considered by the authors of [59] who examined cell-to-cell variability and found variability in relative protein abundances of $\delta C_i / C_i \sim 20\%$. With this assumption, the constant of proportionality in the above equations for driven intrinsic and extrinsic noise in dimer concentration is $(0.2)^2$. Of particular interest, is the magnitude of the response of driven dimer noise $\delta D/D$ to the incident driving noise $\delta C/C$. The relative response is characterized by the ratio $(\delta D/D)/(\delta C/C)$. The results for intrinsic and extrinsic driven noise amplitudes are presented in fig. 4.1. As can be seen in the figure, the intrinsic noise response is significantly larger than the extrinsic case. Evidently, in most cases, coherent driving fluctuations lead to fluctuations in dimer concentration that tend to cancel one another, yielding a smaller response. In the intrinsic case, the opposite seems to be the case, as the amplitude of the dimer response is magnified relative to that of the driving fluctuations.

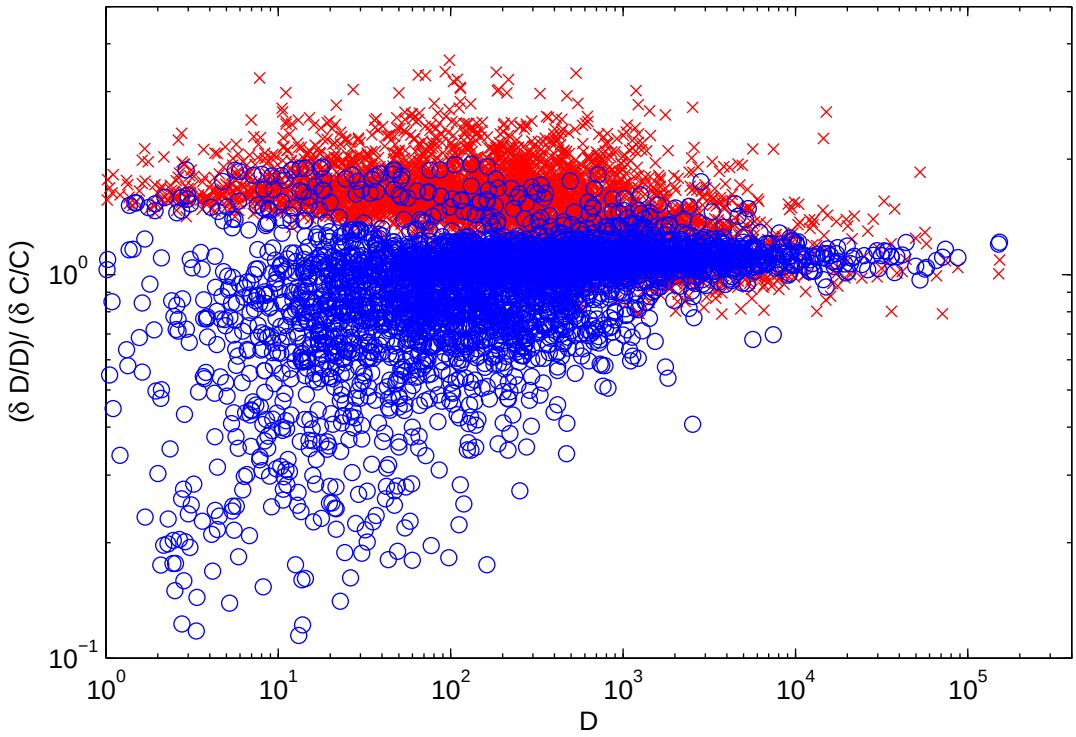

Figure 4.1: The driven driven noise response for intrinsic (red X's) and extrinsic (blue circles) incident fluctuations in total abundance in the PPI network of *S. Cerevisiae*. The results are shown for dimers with equilibrium concentration $D \geq 1$ copy per cell. In general, the noise response to intrinsic driving fluctuations is significantly larger than in the extrinsic case. In most cases, extrinsic driving fluctuations lead to fluctuations in dimer concentration that tend to cancel one another, as evidenced by a response that is typically less than unity. In contrast, intrinsic driving fluctuations lead to dimer response that is amplified relative to the amplitude of the driving noise.

4.5 Spontaneous Fluctuations

As previously mentioned, spontaneous fluctuations in free and bound protein concentrations occur even while the total abundance of all proteins within the network remains fixed. These fluctuations arise as a consequence of chemical

kinetics and thermal stochasticity of molecular collisions and, as such, are well suited to a statistical physics treatment. Employing the partition function approach of eq. 4.4 and the corresponding free energy of eq. 4.7, we can calculate the *generalized susceptibility* from the second derivative of the free energy with respect to dimer concentration.

$$\begin{aligned} \Gamma_{(ij)(km)} &= \frac{D_{ij}}{k_B T} \frac{\partial^2 G}{\partial D_{ij} \partial D_{mk}} \\ &= \delta_{ik} D_{ij}/F_i + \delta_{jm} D_{ij}/F_j + \delta_{ik}\delta_{jm}. \end{aligned} \tag{4.16}$$

where Γ is an $E \times E$ matrix that characterizes the response of the system to perturbations and the pair indices (ij) denote dimers (edges) within the network. The last term in the above connects the perturbative response of neighboring dimers that share a common constituent protein. In accordance with the Fluctuation Dissipation Theorem (FDT) [64] and as derived by the author of [65] for an arbitrary network of reversible chemical reactions, the spontaneous noise for a dimer (ij) is given by the corresponding diagonal element of the inverse susceptibility:

$$\eta \equiv \frac{\langle \delta D_{ij}^2 \rangle}{D_{ij}} = \left(\Gamma^{-1}\right)_{(ij)(ij)} \tag{4.17}$$

Of particular note is that spontaneous noise is independent of temperature, a well-known outcome of FDT. Furthermore, a direct consequence of eq. 4.17 is that spontaneous fluctuations for a dimer linked to the rest of the network involve contributions from other dimers, through the inverse of Γ, the so-called collective effects of the network. To address the impact of collective effects on

the noise, it seems natural to compare the noise of a dimer in the network to the noise for an isolated dimer (*isol-F*) with the same equilibrium concentrations F_i, F_j, and D_{ij}. Such an isolated dimer corresponds to a matrix Γ that is diagonal and has a trivial inverse such that:

$$\eta^{\mathrm{isol-F}} = [\Gamma_{(ij)(ij)}]^{-1} = [\frac{D_{ij}}{F_i} + \frac{D_{ij}}{F_j} + 1]^{-1} \tag{4.18}$$

It can further be shown that the real noise of a dimer in the network always exceeds the isolated dimer noise prediction, $\eta > \eta^{\mathrm{isol-F}}$, by the following convexity argument.

For brevity, we assume the edge notation $\mu = (ij)$, $\nu = (mk)$. The matrix Γ is symmetrized by the diagonal matrix:

$$Q_{\mu\nu} = \sqrt{D_{\mu\nu}}\delta_{\mu\nu} \tag{4.19}$$

and is diagonalized by a unitary transformation U so that:

$$\Gamma = QU\Gamma_D U^{-1} Q^{-1} \tag{4.20}$$

From the convexity of the functional form $f(x) = x^{-1}$ it follows that

$$\Gamma^{-1}_{\mu\mu} \geq (\Gamma_{\mu\mu})^{-1} \tag{4.21}$$

where

$$\Gamma^{-1}_{\mu\mu} = \sum_{\alpha} U^2_{\mu\alpha} (\Gamma_D)^{-1}_{\alpha\alpha} \tag{4.22}$$

and

$$(\Gamma_{\mu\mu})^{-1} = (\sum_{\alpha} U^2_{\mu\alpha}(\Gamma_D)_{\alpha\alpha})^{-1} \tag{4.23}$$

Clearly then, collective effects act to amplify thermal fluctuations. This is related to propagation of static perturbations, studied in [60], as fluctuations from neighboring dimers contribute to a dimer's own noise. We define the amplification factor for a dimer (ij):

$$R = \eta/\eta^{\text{isol}-\text{F}} \tag{4.24}$$

A cumulative histogram of amplification factors for the PPI network of baker's yeast is examined in Fig. 4.2.

Collective amplification of thermal noise presents a worrisome theoretical possibility. Can amplification occur without limit? To address this question, it is fruitful to develop an alternative formalism in which the magnitude of fluctuations are calculated directly from the partition function. Using Eq. 4.4 it is straightforward to show, by a change of variables, that higher moments of D_{ij} can be related to the lower moments evaluated at a reduced system size. Indeed, in calculation of $\langle D_{ij}\rangle$ the combinatorial factor containing $C_i!C_j!/D_{ij}!$ becomes $D_{ij}C_i!C_j!/D_{ij}! = C_iC_j(C_i-1)!(C_j-1)!/(D_{ij}-1)!$. As a result one has the following *exact* equality:

$$\langle D_{ij}\rangle|_{C_i,C_j} = C_iC_j\frac{Z(C_i-1,C_j-1)}{Z(C_i,C_j)} \tag{4.25}$$

Here for the sake of brevity we omitted the concentrations other than C_i and C_j as parameters of the statistical sum $Z(\{C_k\})$. A similar expression for a

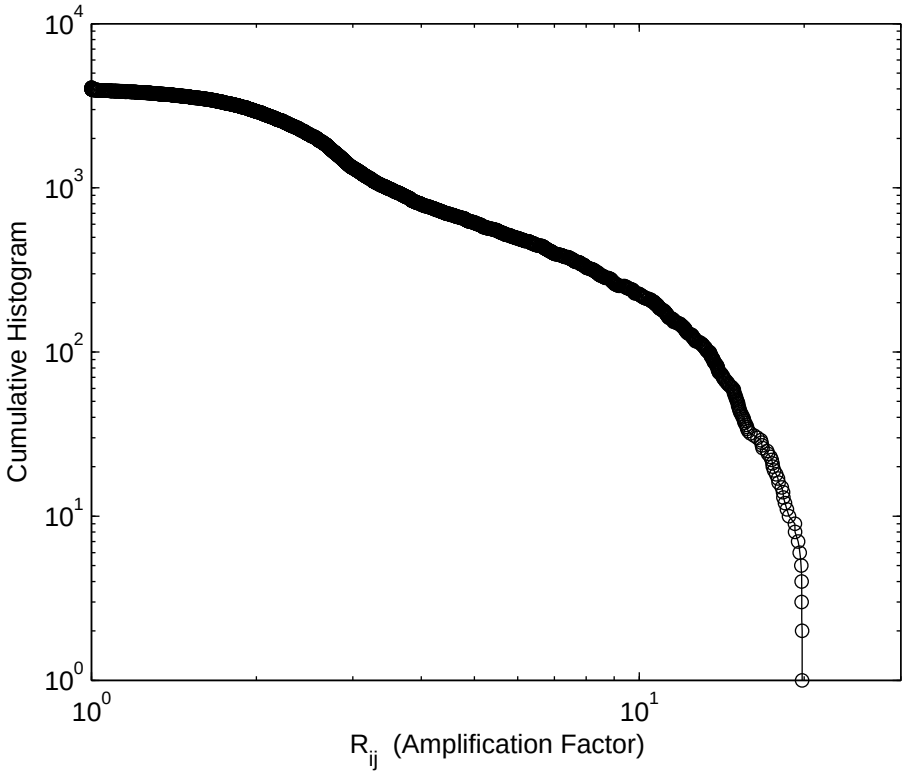

Figure 4.2: Cumulative histogram of amplification factors for spontaneous (thermal) noise of equilibrium dimer concentrations D_{ij} in the PPI network of *S. Cerevisiae*. Collective effects lead to significant noise amplification relative to the isolated case. Amplification factors range as high as 20, while, for the bulk of the dimers in the network, amplification is of order 1.

higher moment

$$\langle D_{ij}(D_{ij}-1)\rangle = C_i(C_i-1)C_j(C_j-1)\frac{Z(C_i-2,C_j-2)}{Z(C_i,C_j)}$$

may be rewritten as

$$\langle D_{ij}(D_{ij}-1)\rangle = \langle D_{ij}\rangle|_{C_i,C_j}\langle D_{ij}\rangle|_{C_i-1,C_j-1} \tag{4.26}$$

where the latter moment is evaluated in a system for which the copy number of proteins i and j (C_i and C_j) are reduced by exactly one. It follows that apart from Eq. 4.17, the noise may be alternatively expressed as:

$$\eta = 1 + \langle D_{ij} \rangle|_{C_i-1, C_j-1} - \langle D_{ij} \rangle|_{C_i, C_j} \tag{4.27}$$

The above expression for thermal noise hints at an intimate connection between the dynamic and static perturbations of the mass-action equilibrium. This connection can be made even more explicit by expanding the 2nd term to first order in total concentration:

$$\eta \simeq 1 - D_{ij}[(\Lambda^{-1})_{ii} + (\Lambda^{-1})_{jj} + 2(\Lambda^{-1})_{ij}] \tag{4.28}$$

where we have expressed the derivatives of the expansion using the definition of the matrix Λ given ineq. 4.10.

It should be remarked that, despite the approximation used in Eq. 4.28, this approach is in good agreement with the FDT formalism first introduced. One notes that this expression for noise explicitly depends only on the total and dimer concentrations used to define the matrix Λ. This suggests the definition of a new isolated model (*isol-C*), consisting of an isolated (ij) dimer formed by proteins with the same C_i, C_j and D_{ij}. This is only possible through changes in the dissociation constant and free concentrations of constituent proteins i and j. It is important to mention that this model is distinct from the *isol-F* benchmark defined earlier, in which each isolated dimer has the same equilibrium free and dimer concentrations (yet different C_i and C_j) as

the corresponding dimer in the network. For an *isol-C* dimer, the matrix Λ is 2x2 and trivially invertible. The noise is given by:

$$\eta^{\text{isol}-\text{C}} = \left(\frac{D_{ij}}{C_i - D_{ij}} + \frac{D_{ij}}{C_j - D_{ij}} + 1 \right)^{-1} \tag{4.29}$$

A comparison with the *isol-F* model reveals that a dimer in the *isol-C* model has an equilibrium free concentration $\tilde{F}_i = F_i + \sum_k D_{ik}$ and similarly for protein j. In other words, the contribution of neighboring dimers to the noise of dimer (ij) has been included by absorbing them into an effective free concentration. Thus, while the *isol-F* model completely ignores the effect of neighboring dimers, the *isol-C* model brings neighboring sources of noise one step closer to dimer (ij). Consequently, the noise of a dimer in the *isol-C* model always exceeds the noise of a corresponding dimer in the real network. The real noise for a dimer in a network falls somewhere between the bounds of these two isolated dimer scenarios.

A summary of the real noise amplitude and that of the lower- and upper-bound models is given in Fig. 4.3. The actual spontaneous fluctuations achieved are a result of real network topology and the distribution of total protein concentration. It is natural to ask how these fluctuations compare to their minimally and maximally achievable values. This suggests the coordinate transformation:

$$\eta \equiv (1 - \zeta)\eta^{\text{isol}-\text{F}} + \zeta\eta^{\text{isol}-\text{C}} \tag{4.30}$$

A histogram of ζ for the PPI network of yeast is shown in Fig. 4.4. Of particular note is the large pileup against the upper limit of amplification. In real PPI networks, it would seem that collective effects lead to amplification

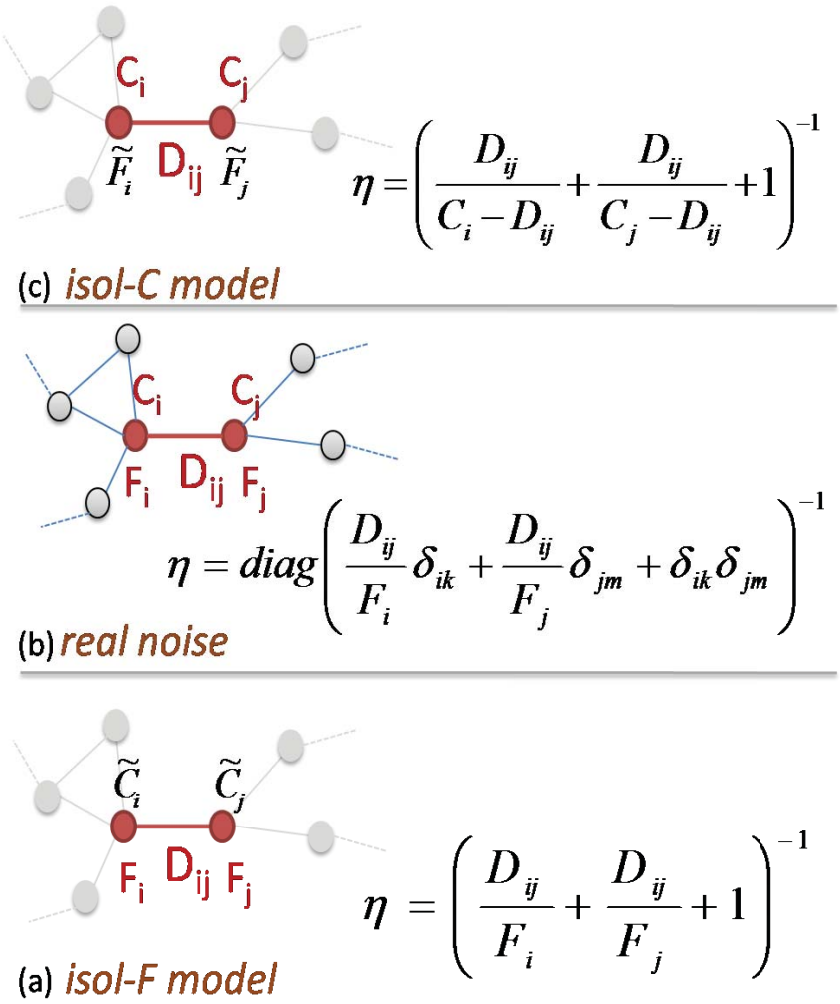

Figure 4.3: A comparison of the noise in a network dimer to two isolated dimer models defined in the text. (a) The *isol-F* model: Each dimer (ij) is isolated and has the same protein free- (F_i, F_j) and dimer- concentrations (D_{ij}) as the corresponding dimer in the network. This model ignores the contribution of other dimers to the noise of dimer (ij) (b) The noise of a dimer (ij) in the network is given by the $(ij),(ij)$ diagonal element of the inverse of the matrix Γ as described in the text. (c) The *isol-C* model: Each dimer (ij) is isolated and has the same protein total- (C_i, C_j) and dimer- concentrations (D_{ij}) as the corresponding dimer in the network. The real noise is bound below and above by the isolated models $\eta^{\rm isol-F} < \eta < \eta^{\rm isol-C}$

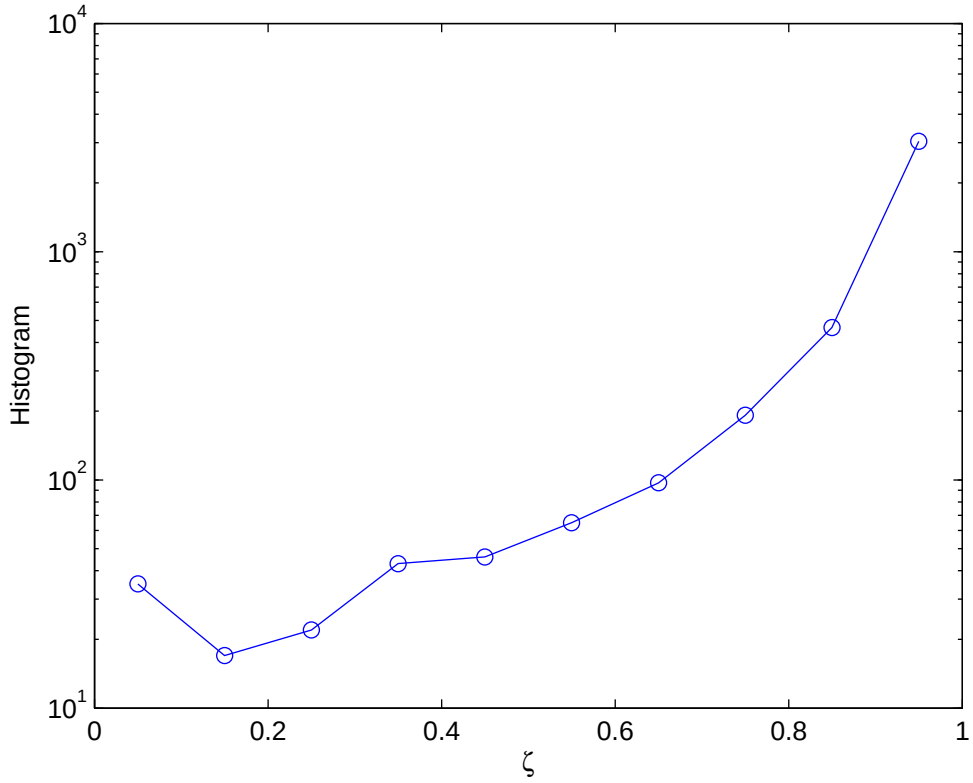

Figure 4.4: Histogram of the spontaneous noise coordinate ζ in the PPI network of bakers yeast. The coordinate describes the position of noise amplitude relative to its lower ($\zeta = 0$) and upper ($\zeta = 1$) limits described in the text.

quite close the maximally achievable limit.

Given numerical calculations for the spontaneous and driven noise studied above, it is natural to ask how the noise amplitude relates to simple predictors such as abundance and network connectivity (number of connections a dimer has to the network). With high statistical significance, we find that the relative amplitude ($\sqrt{\langle \delta D_{ij}^2 \rangle}/D_{ij}$) of both spontaneous and driven (intrinsic) noise is negatively correlated with dimer abundance D_{ij} (Spearman coefficient of $r = -0.98$, $r = -0.45$, respectively). Furthermore, we found that rela-

tive amplitude of both spontaneous and driven (intrinsic) noise are positively correlated with dimer connectivity ($r = 0.42$, $r = 0.33$). These results are consistent with the overall scenario that we investigated above in which any type of noise propagates throughout the network and where the existence of network connections(both direct and, to some extent, indirect) to noisy partners positively contribute to fluctuations of individual dimers.

4.6 Conclusion and Outlook

We have presented a formalism to study dynamical fluctuations in protein binding networks and have characterized the collective network effects on driven and spontaneous noise. For the case of driven noise, we have quantified the collective effects of the network in terms of correlations in total abundance fluctuations. The empirical measurement of total abundance fluctuations in real cells are currently unknown and are a topic for future investigation. We have calculated the dimer noise response for the two important cases of independent (intrinsic) and coherent (extrinsic) driving fluctuations in the real PPI network of *S. cerevisiae*. The response to extrinsic fluctuations is, in most cases, less than unity, while the response to intrinsic fluctuations tend to be amplified. The ramifications of this result hint at both a robustness to coherent fluctuations and susceptibility to independent fluctuations that remains to be understood in a biologically meaningful context.

For the case of spontaneous noise, the introduction of an isolated dimer model (isol-F) allows us to quantify collective network effects in terms of noise amplification factors that are found to be quite significant, ranging as high as

20 for the most extreme cases. The definition of an alternate isolated dimer model (isol-C) that absorbs the noise contribution from neighboring dimers into an effective free concentration allow for the specification of an upper bound on spontaneous dimer noise.

For both spontaneous and intrinsic driven noise, there is a positive correlation with noise amplitude and network connectivity, in agreement with the common-sense notion that fluctuations in neighboring dimers contribute to the noise of a dimer. Possible extensions of this work include more detailed examination of correlation between noise and a dimer's local topology, as well as the study of protein fluctuations that are functionally related.

Bibliography

[1] John P. Scott. *Social Network Analysis: A Handbook*. SAGE Publications, January 2000. ISBN 0761963391. URL http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20\&path=ASIN/0761963391.

[2] Stuart L. Pimm. *Food Webs*. University Of Chicago Press, June 2002. ISBN 0226668320. URL http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20\&path=ASIN/0226668320.

[3] Jennifer A. Dunne, Richard J. Williams, and Neo D. Martinez. Food-web structure and network theory: The role of connectance and size. *Proceedings of the National Academy of Sciences*, 99(20):12917–12922, October 2002. URL http://www.pnas.org/content/99/20/12917.abstract.

[4] Harleyh Mcadams and Adam Arkin. Stochastic mechanisms in geneexpression. *PNAS*, 94(3):814–819, February 1997. URL http://www.pnas.org/cgi/content/abstract/94/3/814.

[5] M. Kaern, W. J. Blake, and J. J. Collins. The engineering of gene regulatory networks. *Annu Rev Biomed Eng*, 5:179–206, 2003. ISSN 1523-9829. doi: 10.1146/annurev.bioeng.5.040202.121553. URL http://dx.doi.org/10.1146/annurev.bioeng.5.040202.121553.

[6] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, August 2002. ISSN 1095-9203. doi: 10.1126/science.1070919. URL http://dx.doi.org/10.1126/science.1070919.

[7] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer Networks*, 33(1):309–320, June 2000. ISSN 1389-1286. doi: 10.1016/S1389-1286(00)00083-9. URL http://portal.acm.org/citation.cfm?id=979923.

[8] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, October 2000. ISSN 0028-0836. doi: 10.1038/35036627. URL http://dx.doi.org/10.1038/35036627.

[9] Albert-Laszlo Barabási, Reka Albert, and Hawoong Jeong. Mean-field theory for scale-free random networks, July 1999. URL http://arxiv.org/abs/cond-mat/9907068.

[10] R. Ferrer I Cancho and R. V. Solé. The small world of human language. *Proc R Soc Lond B Biol Sci*, 268(1482):2261–2265, November 2001. ISSN 0962-8452. doi: 10.1098/rspb.2001.1800. URL http://dx.doi.org/10.1098/rspb.2001.1800.

[11] M. E. J. Newman. Power laws, pareto distributions and zipf's law, December 2004. URL http://arxiv.org/abs/cond-mat/0412004.

[12] G. U. Yule. A mathematical theory of evolution based on the conclusions of dr. j. c. willis. *Philosophical Transactions of the Royal Society B*, 213: 21–87, 1925.

[13] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks, October 1999. URL http://arxiv.org/abs/cond-mat/9910332.

[14] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969. URL http://www.jstor.org/stable/2786545.

[15] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[16] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998. ISSN 0028-0836. doi: 10.1038/30918. URL http://dx.doi.org/10.1038/30918.

[17] Ricard V. Sole, Romualdo Pastor-Satorras, Eric Smith, and Thomas B. Kepler. A model of large-scale proteome evolution. *Advances in Complex Systems*, 5, 2002. URL http://arxiv.org/PS_cache/cond-mat/pdf/0207/0207311v1.pdf.

[18] Andrey Rzhetsky and Shawn M. Gomez. Birth of scale-free molecular networks and the number of distinct dna and protein domains per genome. *Bioinformatics*, 17(10):988–996, October 2001. doi:

10.1093/bioinformatics/17.10.988. URL http://dx.doi.org/10.1093/bioinformatics/17.10.988.

[19] Jacob B. Axelsen, Koon-Kiu Yan, and Sergei Maslov. Parameters of proteome evolution from histograms of amino-acid sequence identities of paralogous proteins. *Biology Direct*, 2:32+, November 2007. ISSN 1745-6150. doi: 10.1186/1745-6150-2-32. URL http://dx.doi.org/10.1186/1745-6150-2-32.

[20] Paul W. Holland and Samuel Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981. doi: 10.2307/2287037. URL http://dx.doi.org/10.2307/2287037.

[21] Juyong Park and M. E. J. Newman. Solution of the 2-star model of a network. *Physical Review E*, 70:066146+, 2004.

[22] Garry Robins, Tom Snijders, Peng Wang, Mark Handcock, and Philippa Pattison. Recent developments in exponential random graph (p*) models for social networks. *Social Networks*, 29(2):192–215, May 2007. doi: 10.1016/j.socnet.2006.08.003. URL http://dx.doi.org/10.1016/j.socnet.2006.08.003.

[23] Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, May 2002. doi: 10.1126/science.1065103. URL http://dx.doi.org/10.1126/science.1065103.

[24] Sergei Maslov, Kim Sneppen, and Alexei Zaliznyak. Pattern detection in complex networks: Correlation profile of the internet. Nov 2002. URL http://arxiv.org/abs/cond-mat/0205379.

[25] H. J. Herrmann, D. C. Hong, and H. E. Stanley. Backbone and elastic backbone of percolation clusters obtained by the new method of 'burning'. *Journal of Physics A: Mathematical and General*, 17(5):L261–L266, 1984. doi: 10.1088/0305-4470/17/5/008. URL http://dx.doi.org/10.1088/0305-4470/17/5/008.

[26] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, December 1959. doi: 10.1007/BF01386390. URL http://dx.doi.org/10.1007/BF01386390.

[27] Richard Bellman. On a routing problem. *Quarterly of Applied Mathematics*, 16:87–90, 1958.

[28] P. E. Hart, N. J. Nilsson, and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *Systems Science and Cybernetics, IEEE Transactions on*, 4(2):100–107, 1968. doi: 10.1109/TSSC.1968.300136. URL http://dx.doi.org/10.1109/TSSC.1968.300136.

[29] M. Newman. Detecting community structure in networks. *The European Physical Journal B - Condensed Matter*, 38(2):321–330, March 2004. ISSN 1434-6028. doi: 10.1140/epjb/e2004-00124-y. URL http://dx.doi.org/10.1140/epjb/e2004-00124-y.

[30] E. Garfield and I. H. Sher. New factors in the evaluation of scientific literature through citation indexing. *Am. Doc.,*, 14:191+, 1963.

[31] De S. Price. Networks of scientific papers. *Science*, 149:510+, 1965. URL http://www.sciencemag.org/cgi/content/citation/149/3683/510.

[32] Ralph Garner. *Three Drexel Information Science Research Studies*. Drexel Press, 1967.

[33] Web of science. URL http://scientific.thomson.com/products/wos/.

[34] Kdd cup 2003 arxiv hep-th dataset.

[35] S. Redner. Citation statistics from 110 years of physical review. *Physics Today*, 58:49+, 2005.

[36] Robert K. Merton. Matthew effect in science. *Science*, 159:56+, 1968.

[37] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30 (1-7):107–117, April 1998. doi: 10.1016/S0169-7552(98)00110-X. URL http://portal.acm.org/citation.cfm?id=297810.297827.

[38] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with google, Apr 2006. URL http://arxiv.org/abs/physics/0604130.

[39] A. Einstein, B. Podolsky, and N. Rosen. Can quantum-mechanical description of physical reality be considered complete? *Physical Review*, 47(10):777–780, 1935. URL http://scitation.aip.org/getabs/servlet/GetabsServlet?prog=normal\&id=PHRVAO000047000010000777000001\&idtype=cvips\&gifs=yes.

[40] M. C. Cross and P. C. Hohenberg. Pattern formation outside of equilibrium. *Reviews of Modern Physics*, 65(3):851–1112, 1993. doi: 10.1103/RevModPhys.65.851. URL http://scitation.aip.org/getabs/servlet/GetabsServlet?prog=normal\&id=RMPHAT000065000003000851000001\&idtype=cvips\&gifs=yes.

[41] R. P. Feynman and M. Gell-Mann. Theory of the fermi interaction. *Phys. Rev.*, 109:193+, 1958.

[42] Nicola Cabibbo. Unitary symmetry and leptonic decays. *hys. Rev. Lett.*, 10:531+.

[43] Hannu Nurmi. Voting procedures: A summary analysis. *British Journal of Political Science*, 13(2):181–208, 1983. doi: 10.2307/193949. URL http://dx.doi.org/10.2307/193949.

[44] Sitabhra Sinha and Raj K. Pan. How a "hit" is born: The emergence of popularity from the dynamics of collective choice, Apr 2007. URL http://arxiv.org/abs/0704.2955.

[45] Santo Fortunato, Alessandro Flammini, Filippo Menczer, and Alessandro Vespignani. The egalitarian effect of search engines, Nov 2005. URL http://arxiv.org/abs/cs/0511005.

[46] Kristina Lerman. Dynamics of collaborative document rating systems. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 46–55, New York, NY, USA, 2007. ACM. ISBN 9781595938480. doi: 10.1145/1348549.1348555. URL http://portal.acm.org/citation.cfm?id=1348549.1348555.

[47] Movielens. URL http://movielens.umn.edu/.

[48] Netflix. Netflix prize: Home. URL http://netflixprize.com.

[49] Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. Reputation systems. *Commun. ACM*, 43(12):45–48, December 2000. ISSN 0001-0782. doi: 10.1145/355112.355122. URL http://portal.acm.org/citation.cfm?id=355122.

[50] Laurent Laloux, Pierre Cizeau, Jean-Philippe Bouchaud, and Marc Potters. Noise dressing of financial correlation matrices, Oct 1998. URL http://arxiv.org/abs/cond-mat/9810255.

[51] M. Takayasu, H. Takayasu, and T. Sato. Critical behaviors and 1/f noise in information traffic. *Physica A Statistical Mechanics and its Applications*, 233:824–834, February 1996. URL http://adsabs.harvard.edu/cgi-bin/nph-bib_query?bibcode=1996PhyA..233..824T.

[52] Kai Nagel and Maya Paczuski. Emergent traffic jams. *Physical Review E*, 51(4):2909+, April 1995. doi: 10.1103/PhysRevE.51.2909. URL http://dx.doi.org/10.1103/PhysRevE.51.2909.

[53] Juan M. Pedraza and Alexander van Oudenaarden. Noise propagation in gene networks. *Science*, 307(5717):1965–1969, March 2005. doi: 10.1126/science.1109090. URL http://dx.doi.org/10.1126/science.1109090.

[54] D. Orrell and H. Bolouri. Control of internal and external noise in genetic regulatory networks. *J Theor Biol*, 230(3):301–312, October 2004. ISSN 0022-5193. doi: 10.1016/j.jtbi.2004.05.013. URL http://dx.doi.org/10.1016/j.jtbi.2004.05.013.

[55] Keun Y. Kim, David Lepzelter, and Jin Wang. Single molecule dynamics and statistical fluctuations of gene regulatory networks: A repressilator. *The Journal of Chemical Physics*, 126(3), 2007. doi: 10.1063/1.2424933. URL http://scitation.aip.org/getabs/servlet/GetabsServlet?prog=normal\&id=JCPSA6000126000003034702000001\&idtype=cvips\&gifs=yes.

[56] Erel Levine and Terence Hwa. Stochastic fluctuations in metabolic pathways. *PNAS*, 104(22):9224–9229, May 2007. doi: 10.1073/pnas.0610987104. URL http://dx.doi.org/10.1073/pnas.0610987104.

[57] 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(Database issue), January 2006. ISSN 1362-4962. URL http://view.ncbi.nlm.nih.gov/pubmed/16381927.

[58] S. Ghaemmaghami, W. K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman. Global analysis of protein expression in yeast. *Nature*, 425(6959):737–741, October 2003. ISSN 1476-4687. doi: 10.1038/nature02046. URL http://dx.doi.org/10.1038/nature02046.

[59] John R. S. Newman, Sina Ghaemmaghami, Jan Ihmels, David K. Breslow, Matthew Noble, Joseph L. Derisi, and Jonathan S. Weissman. Single-cell proteomic analysis of s. cerevisiae reveals the architecture of biological

noise. *Nature*, 441(7095):840–846, May 2006. ISSN 0028-0836. doi: 10.1038/nature04785. URL http://dx.doi.org/10.1038/nature04785.

[60] Sergei Maslov, Kim Sneppen, and I. Ispolatov. Spreading out of perturbations in reversible reaction networks. *New J. Phys.*, 9(8):273+, August 2007. ISSN 1367-2630. doi: 10.1088/1367-2630/9/8/273. URL http://dx.doi.org/10.1088/1367-2630/9/8/273.

[61] Sergei Maslov and I. Ispolatov. Propagation of large concentration changes in reversible protein-binding networks. *PNAS*, pages 0702905104+, August 2007. doi: 10.1073/pnas.0702905104. URL http://dx.doi.org/10.1073/pnas.0702905104.

[62] M. D. Kumar and M. M. Gromiha. Pint: Protein-protein interactions thermodynamic database. *Nucleic Acids Res*, 34(Database issue), January 2006. ISSN 1362-4962. URL http://view.ncbi.nlm.nih.gov/pubmed/16381844.

[63] H. W. Mewes, D. Frishman, U. Güldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Münsterkötter, S. Rudd, and B. Weil. Mips: a database for genomes and protein sequences. *Nucleic Acids Res*, 30(1):31–34, January 2002. ISSN 1362-4962. doi: 10.1093/nar/30.1.31. URL http://dx.doi.org/10.1093/nar/30.1.31.

[64] Herbert B. Callen and Theodore A. Welton. Irreversibility and generalized noise. *Physical Review*, 83(1):34+, July 1951. doi: 10.1103/PhysRev.83.34. URL http://dx.doi.org/10.1103/PhysRev.83.34.

[65] I. Prigogine. Sur les fluctuations de l'equilibre chimique. *Physica*, 16:134–136, February 1950. doi: 10.1016/0031-8914(50)90071-1. URL http://dx.doi.org/10.1016/0031-8914(50)90071-1.