ABSTRACT

Title of dissertation: MAKING PREDICTIONS AND HANDLING ERRORS IN RECONSTRUCTED BIOLOGICAL NETWORKS

John Platig, Doctor of Philosophy, 2013

Dissertation directed by: Associate Professor Michelle Girvan
Department of Physics

In this thesis we present methods for applying techniques from complex network theory to analyze and interpret inferred biological interactions. With the advent of high throughput technologies such as gene microarrays and genome-wide sequencing, it is now possible to measure the activity of every gene in a cancer cell population under different conditions. How to extract important interactions from these experiments remains an outstanding question. Here we present a method to identify these key interactions by focusing on short paths in a transcription factor network.

We use a mutual information-based approach to infer the transcription factor network from gene expression microarrays, which measure perturbations in a Diffuse Large B Cell Lymphoma cell line. By focusing on the number of short paths between transcription factors and signature genes in the inferred network, we find a set of transcription factors whose biology is crucial to the continued survival of these lymphoma cells and also show that a subset of these factors have a distinct expression

pattern in patient tumors as well.

As many networks of interest are reconstructed from data containing errors, we introduce two simple models of false and missing links to characterize the effects of network misinformation on three commonly used centrality measures: degree centrality, betweenness centrality, and dynamical importance. We show that all three measures are especially robust to both false and missing links when the network has a power law in the tail of its degree distribution.

MAKING PREDICTIONS AND HANDLING ERRORS IN RECONSTRUCTED BIOLOGICAL NETWORKS

by

John Platig

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:
Associate Professor Michelle Girvan, Chair/Advisor
Distinguished University Professor Edward Ott, Co-Advisor
Associate Professor Wolfgang Losert
Professor Rajarshi Roy
Distinguished University Professor James Yorke

Dedication

For Elizabeth.

Acknowledgments

I would like to thank my advisor, Prof. Michelle Girvan, for her patience and persistence with my highly interdisciplinary project. She was always willing to invest the time to help, even as I plunged deeper into the world of cancer biology. I would also like to thank my co-advisor, Prof. Ed Ott, for his wisdom and concern for my success in graduate school. Thanks to Prof. Losert for setting up the UMD-NCI collaboration and my project in particular. I would also like to that my thesis committee for taking the time to review my manuscript.

I owe a great debt to the members of the Staudt lab. Thanks to Dr. Lou Staudt for investing the time to teach me how to think about molecular biology and functional genomics. Art Shaffer's constant insight and support, especially during the bleak times when the project outcome was uncertain, for the last five years made this thesis possible. Thanks to George Wright for always being willing to take my phone calls about statistics and for his significant contribution to the statistical methods in Chapter 2. I would also like to thank Michele Ceribelli, Yibin Yang, Ryan Young, and Daniel Hodson for their willingness to investigate the biological usefulness of my ideas.

The assistance from my fellow UMD graduate students proved invaluable. I thank Geet Duggal for developing an extremely fast computational approach to calculating mutual information and for teaching me the nuances of the R scripting language. My group members Shane Squires, Mark Herrera, Wai Lim Ku, and (formerly) Kimberly Glass provided timely assistance and valuable friendship, thank

you. I owe special thanks to Evan Berkowitz for his assistance with Mathematica. I am also very grateful for the support of Jane Hessing, Arlene Dyer, Lisa Errington, Sildiz Ali, and Ed Condon.

I am very grateful for the funding from the UMD-NCI Cancer Technology Partnership and the CRTA fellowship program at the NCI.

I am tremendously grateful for the unabated support of my family and fiancée. Harve, Peggy, Catherine, Elizabeth, and Anthony, thank you. My apologies to those I have forgotten to mention.

To God be the glory.

Table of Contents

List of Figures

List of Abbreviations

ABC DLBCL	Activated B Cell-like Diffuse Large B Cell Lymphoma
BATF	Basic leucine zipper transcription factor, ATF-like
cDNA	complementary Deoxyribonucleic Acid
DNA	Deoxyribonucleic Acid
ER	Erdos-Renyi
GCB DLBCL	Germinal Center B Cell-like Diffuse Large B Cell Lymphoma
ID3	Inhibitor of DNA binding 3
IRF4	Interferon Regulatory Factor 4
IRF7	Interferon Regulatory Factor 7
IRF9	Interferon Regulatory Factor 9
mRNA	messenger Ribonucleic Acid
MI	Mutual Information
NF-κB	Nuclear Factor kappa-light-chain-enchancer of activated B cells
NF-κB1	Nuclear Factor kappa-light-chain-enchancer of activated B cells 1
R-CHOP	Rituximab Cyclophosphamide Hydroxydaunomycin Oncovin Prednisone
RNA	Ribonucleic Acid
SF	Scale-Free
STAT3	Signal Transducer and Activator of Transcription 3
shRNA	short hairpin Ribonucleic Acid
TCF4	Transcription Factor 4
TF	Transcription Factor

Chapter 1

Introduction

The theory of complex networks—the intersection of graph theory, statistical physics, and computer science—has provided a principled approach to understanding patterns of connectivity in large biological and social networks [1]. Concurrently, the development of high throughput technologies in biology [2] and the proliferation of digital social interactions [3] have lead to an abundance of data on networked systems. Web platforms such as Facebook and Twitter have millions of users who interact with each other in real time, providing detailed, time-dependent social networks at an unprecedented level [4]. Similarly, genome sequencing is rapidly changing biology into a quantitative science in search of reliable analytic approaches for interpreting the vast amounts of data [5].

As these networks have thousand and even millions of nodes, we are faced with the challenge of finding effective tools to identify the key mechanisms and structures therein. We are entering an era in which data is cheap, but insight is expensive. However, this presents an exciting opportunity for physicists, who are accustomed to working with very large data sets and identifying the core behavior of systems with many variables.

This thesis aims to demonstrate that the ideas and techniques developed for complex networks can provide valuable insight into complex biological networks.

Using a data set measuring the gene regulatory network of a specific subtype of cancer cells, we identify both known and novel genetic interactions based on the underlying network structure, which we infer from the data using an information theoretic approach. In addition, we put forward a first effort to characterize the effect of link errors on network-based predictions that may arise in various applications involving complex networks.

1.1 Chapter 2: Identifying Therapeutic Targets in a Lymphoma Regulatory Network

The ability of human cells to carry out various functions is an emergent behavior of hundreds and sometimes thousands of different genes working in concert [6]. Many of these genes produce proteins that often only bind to a subset of other genes, who then produce proteins that bind to their own subset of "neighbors" [7]. These genes and their gene-protein interactions can be naturally represented as a network consisting of nodes (genes) and links (interactions) between genes.

In this thesis, we consider such a gene regulatory network inside a particularly aggressive subtype of Diffuse Large B Cell Lymphoma (DLBCL) called Activated B Cell like DLBCL. Patients with ABC DLBCL have only a $\sim$40% survival rate after 5 years, and the aggressive nature of this subtype is driven largely by genetic mutations and other changes in gene regulation that are unique to the ABC subtype [8]. We hypothesize that genetic mutations in the cancer cells have rewired the gene regulatory network in ABC DLBCL to enable new function abilities such as

preventing cell death. We believe these new functions are an emergent property of the rewired regulatory network and that the connectivity of this network may provide insight into new therapeutic targets in ABC DLBCL.

1.2 Chapter 3: Robustness of Network Measures to Link Errors

Much of the theory of complex networks assumes a complete and accurate knowledge of the links and nodes in a given network. However, we often reconstruct social and biological networks from incomplete or inaccurate data. The social networks on Facebook and Twitter, for example, may contain links between people who do not actually interact. In addition, there may be missing links between people whose face-to-face social interactions are significant. Biological networks are measured using experimental techniques, which have inherent systematic errors which can lead to false and missing links [9, 10]. To explore the effect of these types of link errors on three popular network centrality measures, we propose two simple stochastic models of link error and measure their effect on the ranking of nodes before and after introducing link errors. We develop a theoretical framework for analyzing these types of errors that can be naturally extended to networks beyond those discussed in this thesis.

1.3 Outline of Thesis

In Ch. 2, we discuss a network-based method to identify potential therapeutic targets from data measuring a Lymphoma cell line. In Ch. 3 we address the effect

of link errors on node centrality measures in networks. In the conclusion (Ch. 4), we summarize our findings and provide suggestions for future directions.

Chapter 2

Identifying Therapeutic Targets in a Lymphoma Regulatory Network

2.1 Introduction

As the human genome project approached completion in the late 1990's, it became increasingly apparent that the number of genes involved in the function of human cells was much larger than expected. Instead of a few thousand genes working together, there are tens of thousands interacting in a context-dependent manner. In addition to broadening our understanding of the biological world, the complete sequencing of the human genome provides the possibility of identifying mutations within genes that cause cancer. If it were possible to identify those genes that were mutated and inactivate them, one could kill the cancer cells. That hypothesis, unfortunately, has proven too simplistic, and we now have a growing body of knowledge about how mutated genes interact with one another, with unmutated genes, and with the cellular environment as a networked system to give rise to cancerous tissue [11]. Figure 2.1 illustrates the current understanding of genetic mutation in cancer.

The last ten years gave rise to technology that allows researchers to probe the properties of all ∼25,000 genes in a population of cancer cells. Current high-throughput experimental platforms can assess every gene's protein and mRNA levels across a cell population [12, 13]. These platforms make a systems approach to understanding cancer possible, and the cancer biology community has seen an explosion

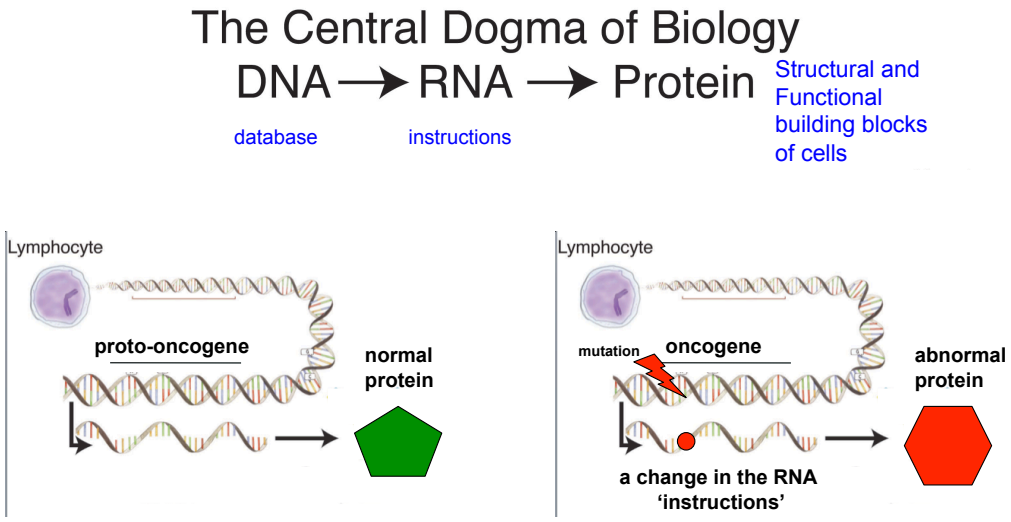

Figure 2.1: Each cell contains a complete set of DNA instructions that define the organism. Specific sections of the DNA–called genes–are copied into RNA molecules which are then used as instructions to make a specific protein. Changes to the DNA through mutations (changing DNA base pairs) can lead to changes in the RNA which then produce an abnormal protein that may acquire new functional abilities that promote the cancer phenotype. (Courtesy of L. Staudt and A. L. Shaffer III)

of such systems-level data as the cost of experiments has fallen. This rapid quantification of cancer biology now requires analytic methods that can highlight the relevant system behavior in a principled fashion. For example, the transcription factor regulatory system, the "command and control center" of the cell, can be naturally modeled as a network of genes that interact to perform cellular functions. In this chapter, we adapt techniques from the field of complex networks to analyze the transcriptional regulatory network in a lymphoma subtype and predict genes that may contribute to the aggressive nature of this disease.

2.2 Background

2.2.1 The need for a quantitative approach to Diffuse Large B-cell Lymphoma

Before the sequencing of the human genome, different cancer subtypes were classified based solely on their morphology; that is, how they looked under the microscope. In the case of non-Hodgkin's lymphomas, this kind of classification into distinct subtypes was "widely believed to be imprecise" [14]. A single morphological subtype, Diffuse Large B Cell Lymphoma (DLBCL), constitutes 40% of all non-Hodgkin's Lymphomas with ∼23,000 new diagnoses per year, a ∼50% cure rate, and ∼10,000 deaths per year [15]. With the development of microarray technology [16], Alizadeh and coauthors classified these DLBCLs based on the similarity of gene expression patterns between tumors [17]. By clustering gene expression of patient samples, they were able to show that DLBCLs comprise two distinct subgroups

(shown in Fig 2.2): an Activated B-cell-like (ABC) subtype, which has transcriptional programming similar to activated B cells, and a Germinal Center B-cell-like (GCB) subtype, which shares the same programming as healthy germinal center B cells.

This gene expression based classification is clinically relevant as well, as patients with the ABC subtype who are treated with chemotherapy (R-CHOP) have only a 40% progression-free survival rate after three years compared to 60% for those with the GCB subtype as shown by the survival curves in Fig. 2.3 [18]. In addition to comparing the mRNA levels of genes between patient samples, DNA microarrays have also proven to be an invaluable tool for measuring the changes in gene expression across the whole genome [16]. From a network viewpoint, a microarray measures the change in state for each gene in the regulatory network given a specific perturbation.

The experimental details are as follows: The DNA microarray is a glass slide on which small DNA fragments are fixed. Multiple copies of a given DNA fragment are fixed within a specific spot on the slide such that each DNA fragment corresponds to a specific mRNA. With the human genome completely sequenced, it is possible to identify which mRNA fragments are produced by each gene. Spots correspond to genes that are actively transcribed into mRNA.

Given this set-up, it is possible to compare the amount of mRNA produced by every gene in the genome under two different conditions (for example, cells with a drug treatment vs. cells without). First, the mRNA for cell population under condition one is extracted and each mRNA molecule is copied and an additional

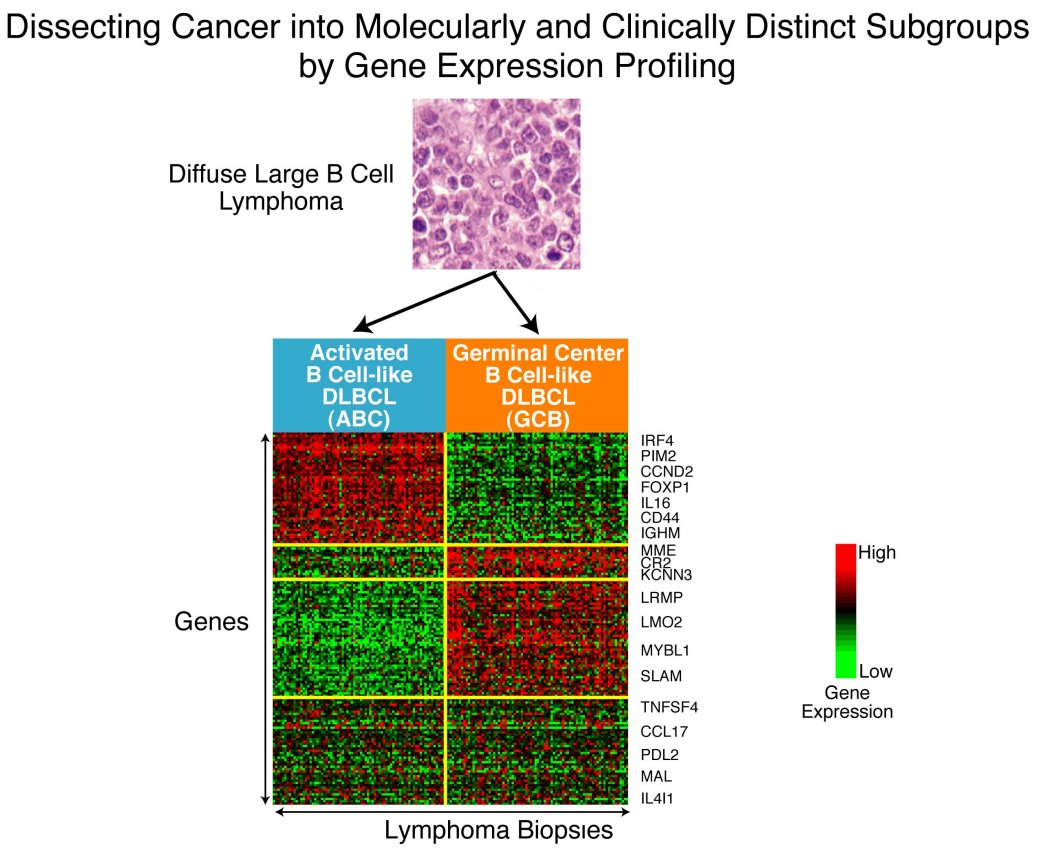

Figure 2.2: DLBCL patient samples that look identical under the microscope have gene expression profiles which belong to different subgroups that reflect their underlying biology. Samples are clustered based on the similarity of the gene expression between samples with yellow lines delineating the groups of coordinately expressed genes in the different subtypes [17].

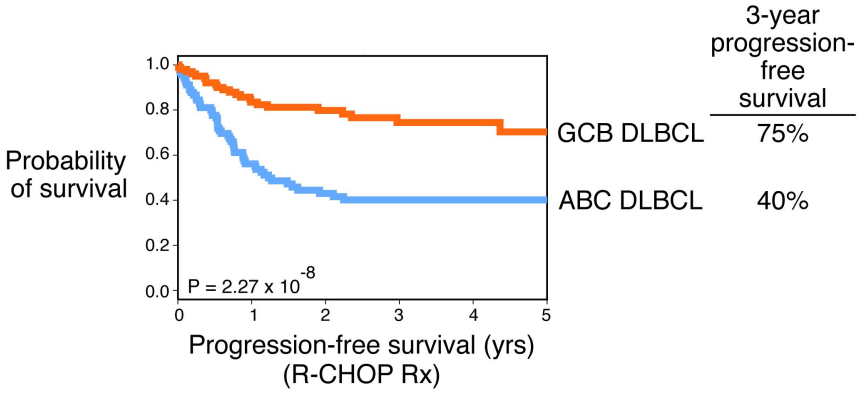

Figure 2.3: Kaplan Meyer survival curves for patients classified as having either ABC DLBCL (in blue) or GCB DLBCL (in orange). The biopsies of patients with either subtype look identical under the microscope, but have distinctly different gene expression profiles and survival outcomes after treatment with chemotherapy (R-CHOP) [17].

red fluorescent tail is added. Second, the cells in a second condition–exposure to a drug, for example–are extracted and copied with an added green fluorescent tail. The resulting complementary DNA (cDNA) is then placed on the slide and because of the highly specific nature of the binding affinity between complimentary DNA fragments, the tagged cDNA binds to the appropriate spot on the microarray. Finally, each spot is illuminated and the log_2 ratio of green/red fluorescence is reported as shown in Fig. 2.4.

2.2.2 Different molecular mechanisms drive different subtypes

At their core, cancers are genetic diseases: through the alteration of existing transcriptional machinery by genetic aberrations, healthy cells transform into cancer cells. From a network perspective, these aberrations lead to new network links or removal of networks links which can reinforce multiple cellular functions including the promotion of cell division and growth and the inhibition of cell death [19]. In ABC DLBCL, this network rewiring often leads to interactions between signaling pathways—or subnetworks in the network view—that do not exist in the transcriptional network of healthy cells [20].

For example, mutations upstream of the NF-κB pathway lead to constituative activitaton of the pathway. This pathway activation is considered a hallmark of the ABC DLBCL subtype and endows cells with the ability to minimize cell death via apoptosis (apoptosis is internally programmed cell death). This minimization has been hypothesized as one of the reasons why ABC DLBCL is so resistant to

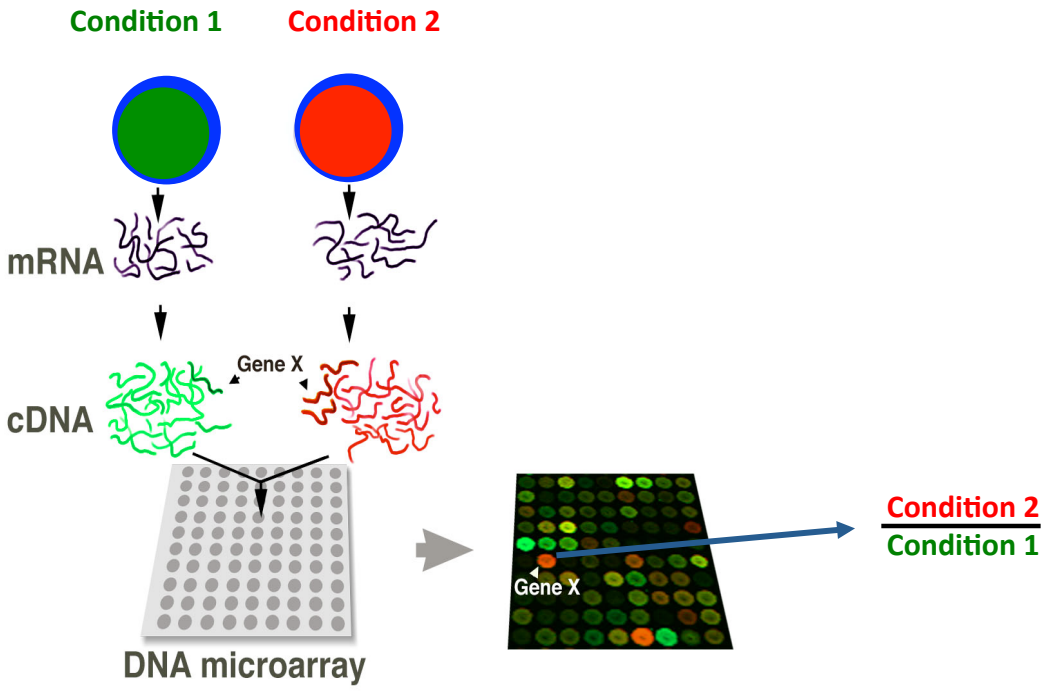

Figure 2.4: Method for profiling the log_2 ratio of each gene's mRNA levels—for all 23,000 genes—in cells under both conditions. The mRNA level for each gene under both conditions is then measured by the microarray. The log_2 ratio of those mRNA levels is then calculated to determine the change in expression—i.e., mRNA—level for each gene.

chemotherapy, which attempts to kill cancer cells via apoptosis. In addition, the NF-κB pathway interacts with the STAT3 pathway to further reinforce a survival response as shown in Fig. 2.5 [21].

Another critical component of the ABC DLBCL network is the activity of the transcription factor IRF4. While IRF4 gene expression is crucial to the survival of ABC DLBCL cells (though not of the GCB DLBCL subtype), it is not structually or genetically altered from copies of IRF4 in healthy cells [22]. It plays a complicated role in the ABC DLBCL subtype, upregulating genes in the NF-κB pathway and repressing genes associated with the interferon pathway, which tends to cause cell death when activated [23]. As such, IRF4 appears to be a crucial hub for the interaction between multiple pathways within ABC DLBCL.

These differences highlight the uniqueness of gene regulatory networks in specific cancer subtypes and the need to generate subtype specific networks if network topology is to be a useful tool for predicting therapeutic targets.

2.2.3 Systematic Approaches to Constructing Gene Regulatory Networks

As the price of microarray platforms has fallen, researchers have amassed large sets of microarrays measuring cancer cells under a host of different conditions. However, the best quantitative method for understanding these data has yet to be determined. Early work focused on hierarchical clustering of data sets in which genes that have highly correlated expression patterns are near each other in the result-

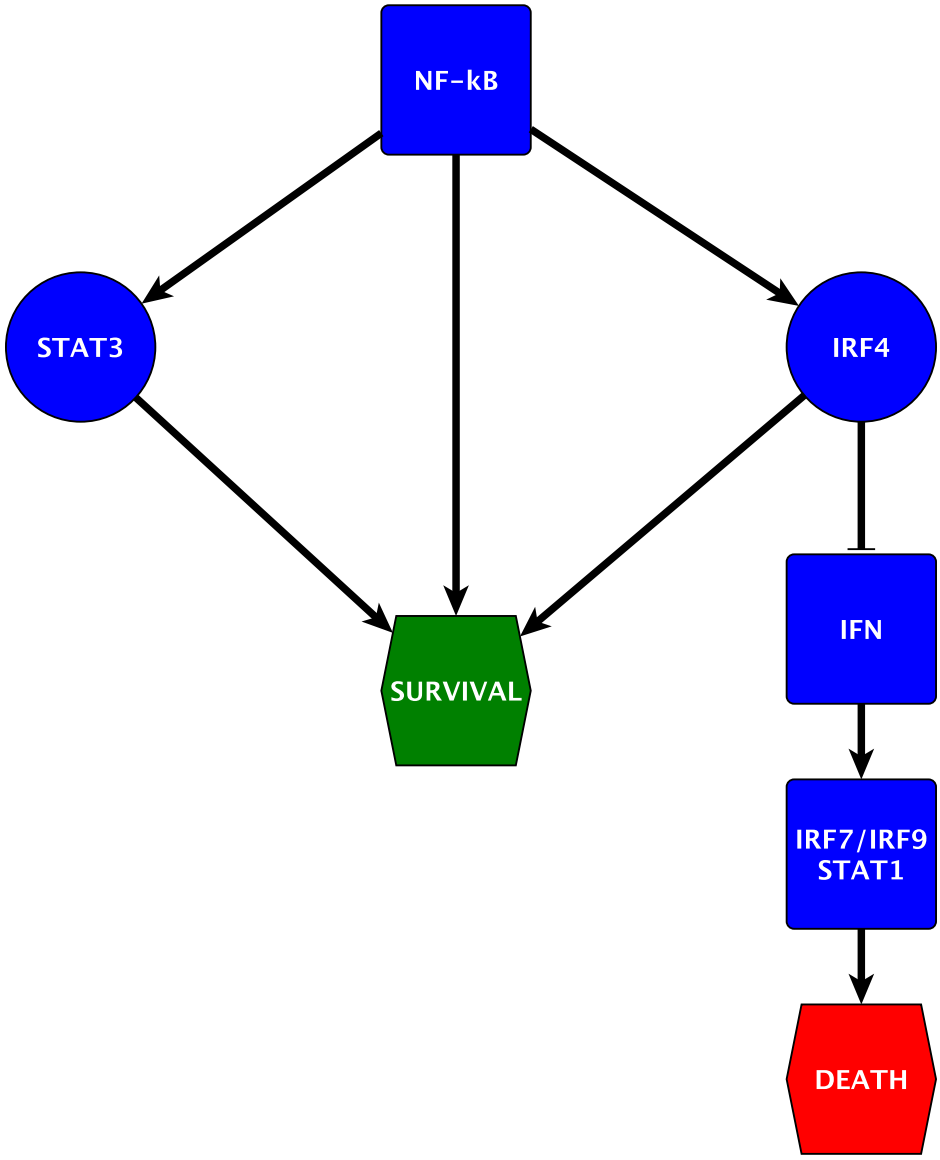

Figure 2.5: The NF-κB pathway is constitutively activated through upstream oncogenic mutations and interacts with the transcription factors STAT3 and IRF4 to promote pro-survival in ABC LDLBCL while IRF4 inhibits the interferon pathway to prevent signaling promoting cell death.

ing nested hierarchy [24]. This approach is computationally tractable for hundreds of microarrays (experimental perturbations), each of which measures as many as ∼40,000 unique mRNA levels. However the resulting clusters provide only a crude view of the transcriptional organization within human cells [6].

In light of this, work to develop methods that reflect the interactions between genes has emerged [25]. While genes can produce proteins that bind together into complexes before performing their cellular function, the amount of data required for estimating these higher order interactions from microarrays is beyond what is currently feasible for a single lab.

Instead, popular methods [26, 10, 27, 28] focus on the slightly idealized view that a TF produces a protein which then binds to its target gene, causing the target's mRNA levels to increase or decrease. Because cooperative interactions between multiple proteins may be required to influence a target gene's expression, the relationship between a TF's mRNA levels and the target's may not be linear. For example, one might imagine an XOR circuit in which two TFs independently activate the same target gene, but when both TFs are expressed their proteins bind together into a complex that ultimately represses the target gene. In order to capture such nonlinear relationships, the mutual information (MI) between pairs of

genes is calculated:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{2.1}$$

$$H(X) = -\sum_{x \in X} p(x)\ln(p(x)) \tag{2.2}$$

$$H(X,Y) = -\sum_{x \in X, y \in Y} p(x,y)\ln(p(x,y)) \tag{2.3}$$

where $p(x)$ is the probability that the random variable X—associated with a gene's expression level—is in state x, and $p(x,y)$ is the joint probability of $X = x$ and $Y = y$, for a gene pair. The best way to estimate the MI from a limited number of measurements ($N \lesssim 500$) remains an area of active research with methods for using kernel density estimation [29], spline fitting [10], or bin width maximization [30] all available.

Once a matrix of MI values is computed, a variety of methods for determining the statistical significance of a given MI value can be used. Margolin and coauthors [29] developed a shuffling scheme in which a gene's expression values are randomly reordered and the MI is recomputed. This is done many times for many (10^5) gene pairs to produce a null distribution which can be used to assign a p-value to each true MI value.

Faith and coauthors [10] put forward a simpler, less computationally intensive method that determines, for a given gene, the distribution of its MI values across all other genes, and calculates its mean and standard deviation. Using these, a combined z-score can be computed for a gene pair's particular MI value to determine its statistical significance. For a reconstructed network in *E. coli*, this has been shown to be more accurate in identifying true links and minimizing false positive

links than the method of Margolin et. al. [10].

2.3 Methods

2.3.1 Microarray Data

We consider 168 two-color Agilent microarrays each measuring the change in mRNA levels of every gene in the cell population in response to one of 23 different perturbations at a specific time point. These perturbations come in two basic types: (1) shRNA knockdowns which inactivate a specific node and all of its links and (2) small molecule inhibitors which disrupt the links in a specific subgraph. From a network perspective each microarray measures the response of each node (gene) to a particular perturbation.

For the shRNA knockdowns, a short hairpin RNA (shRNA) is designed to bind to the mRNA of a specific gene and bind or "knock down" the mRNA, effectively preventing the mRNA from being translated into protein. If a gene's protein is not produced, it can no longer perform its designated functions. Because these shRNAs inactivate the mRNA and not the protein of the targeted gene, it can take between 12 and 24 hours before existing protein is used up and the effect of the perturbation is realized.

Consequently, for the shRNA experiments, the microarray measurements of the change in gene expression are typically taken at 12, 24, 48 and 72 hours after the induction of the shRNA. Multiple shRNAs are designed to target the same gene and tested to make sure that any response by the cell is not the result of the

shRNA indiscriminately binding to mRNA that is not produced by the target gene. The cells are designed so that the shRNAs are not expressed until they are treated with doxocyclin, which ensure that cells in the population experience the shRNA perturbation at the same time.

In addition to the 13 shRNA knockdown experiments, there are 10 perturbations using small molecule inhibitors, which target proteins within a specific pathway. Because these inhibitors are taken up by the cells quickly and begin inactivating the protein of interest, their effect is observed more immediately than the shRNA perturbations. Microarray measurements for the small molecule inhibitors are typically taken at 1, 3, 6 and 24 hours after treating the cells. For the purposes of this thesis, we treat each microarray as an independent "snapshot" of the cell and do not attempt to model the time dependent nature of the perturbations.

2.3.2 Calculating Mutual Information and Z-scores

Given a set of relative changes in gene expression for all genes under different conditions, we would like to identify gene pairs whose expression levels are significantly correlated. Because of an abundance of nonlinear interactions between gene expression levels [30], we choose not to use the traditional correlation coefficient, which identifies only linear relationships between random variables. Instead, we choose the mutual information (MI) as a measure of the linear and non-linear dependencies between gene pairs. Given random variables X and Y that represent the expression level of a pair of genes, their mutual information is given by

Eqs. (2.1-2.3). In Eq. (2.2), $p(x)$ is the probability that X has expression level x. This probability is estimated from a histogram of the expression values with four equally spaced bins ranging from the minimum to maximum value of X. The joint probability $p(x, y)$ (Eq. (2.3)) is estimated from a 16-bin histogram. Given 159 measurements for each gene, we chose the the number of bins such that the average number of data points per bin in the joint histogram is approximately 10.

After calculating the matrix M_{ij} containing the MI between genes i and j for all ∼484 billion gene pairs, we generate a network in which links exist between two genes if they have statistically significant mutual information. To determine this, we calculate the z-score for the mutual information between the pair by combining the z-scores for the individual genes. The z-score for an individual gene, i, is

$$z_{ij}^{(row)} = \frac{M_{ij} - \mu_i}{\sigma_i} \tag{2.4}$$

$$z_{ij}^{(col)} = \frac{M_{ij} - \mu_j}{\sigma_j} \tag{2.5}$$

where μ_i and σ_i are the mean and standard deviation of row i and μ_j and σ_j are the mean and standard deviation of column j. A z-score for the pair ij is then:

$$Z_{ij} = \sqrt{\left(z_{ij}^{(row)}\right)^2 + \left(z_{ij}^{(col)}\right)^2}$$

We link the genes in the networks if and only if $Z_{ij} \geq 3$.

2.3.3 Identifying Regulator Genes to Target

Given a network consisting of links reconstructed from gene expression microarray data, we want to identify those transcription factors (TFs) that are close

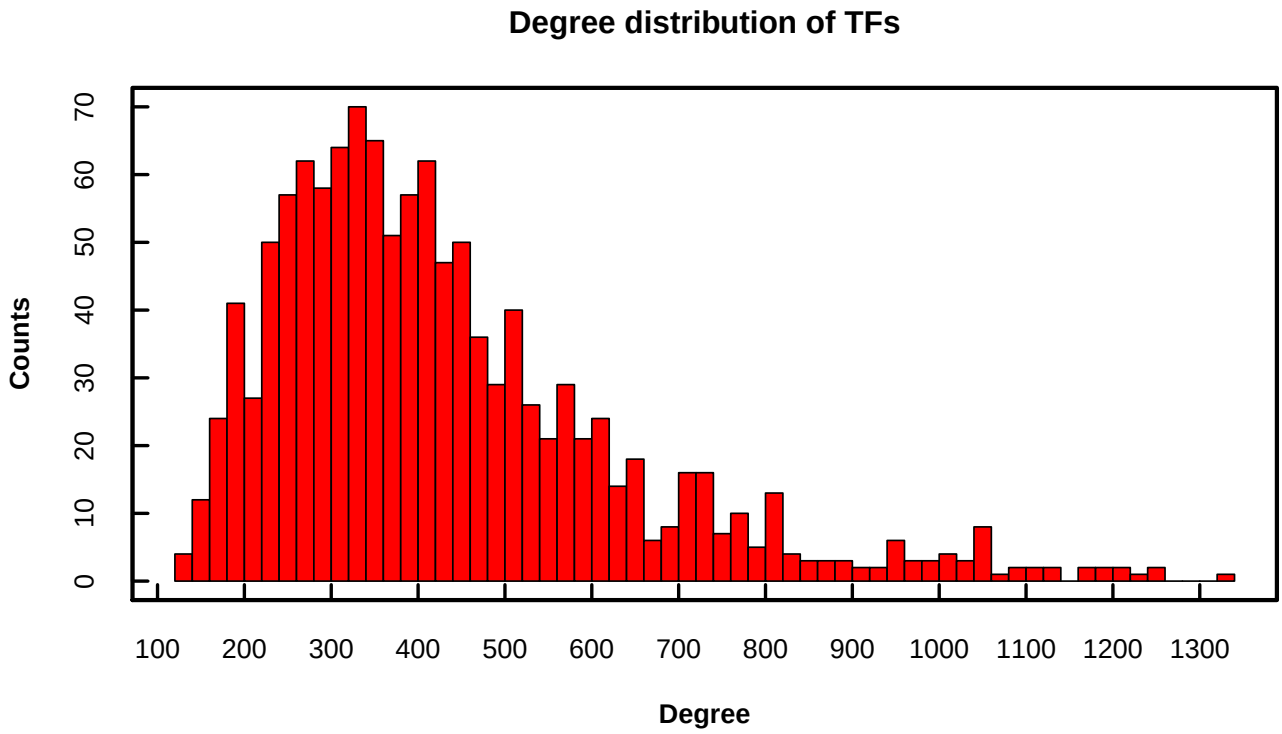

Figure 2.6: Degree distribution of the 1201 transcription factors from the mutual information derived network where TFs are connected to genes if they have a mutual information value that is 3 standard deviations or more away from the mean.

in network space to sets of genes that are implicated in the aberrant behavior ABC DLBCL cells. There are 1201 TFs that are expressed in the cells and included in the network. Genes were designated as a TF by manually curating a list of known B-cell TFs in addition to a general list of TFs included in Carro et. al [31]. We exclude links that do not connect to a TF since those links would reflect post-translational interactions, which are not reliably measured by the microarray experiments.

To determine the potential network drivers of different cell functions in ABC

lymphoma, we focus on short paths from a TF to sets of genes whose coordinate mRNA expression levels reflect important cell physiology (see Fig. 2.7). These sets of genes, *gene signatures*, are compiled from the literature [32] and are derived from two general types of gene expression microarray experiments: (i) profiling experiments where patient samples are hierarchically clustered to determine differentially expressed genes (as in [17]), (ii) perturbation experiments where the normal cell state is altered via inactivation of a specific gene or pathway(s) and the resulting changes in gene expression are captured with microarray measurements after induction of the perturbation.

Those genes with similar expression levels are clustered together into a gene signature, as shown in Fig. 2.8. The entire signature database includes 317 signatures comprised of 14,000 unique genes. Signatures range in size from 10 to 4600 genes. Note that individual genes also show up in multiple signatures, reflecting the biological reality that a single gene can be involved in multiple cellular functions.

2.3.4 Calculating p-values for First and Second Neighbors

We wish to use the gene interaction network topology to identify transcriptional regulators who are significantly connected to genes in a specific signature. By looking at the first and second neighbors—genes that are one or two links away, respectively—of a regulator, we can ask how many of those neighbors are in a particular signature. Defining the set of first or second neighbors of regulator i as R_i and the set of genes in a given signature j as S_j, the size of the intersection of those two

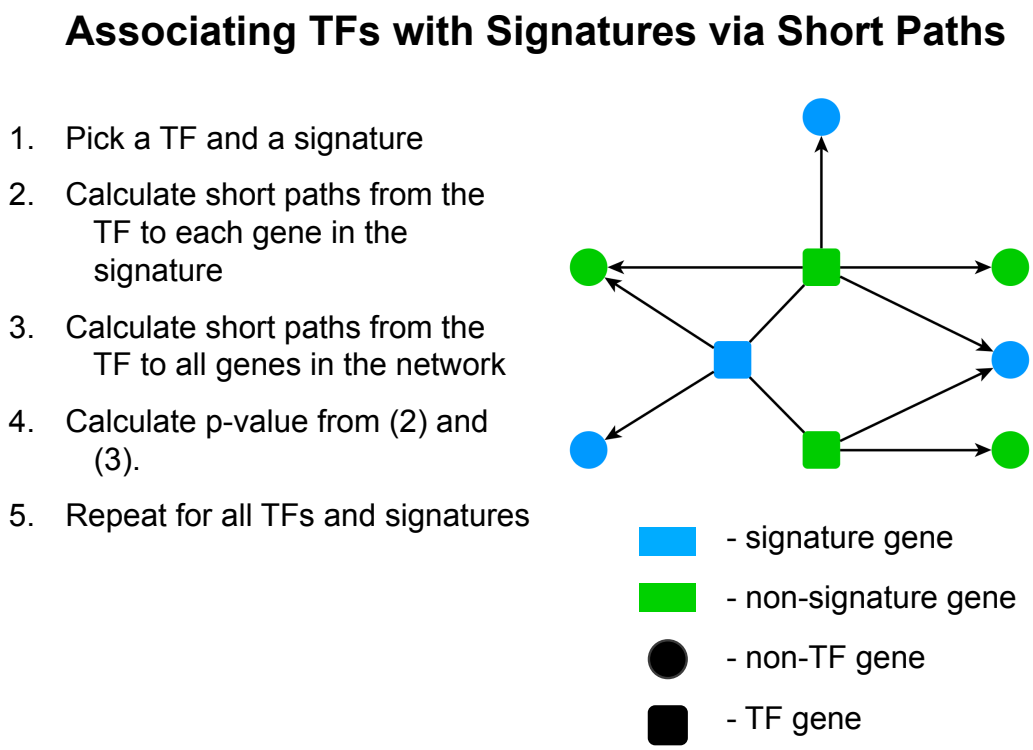

Figure 2.7: Schematic of the algorithm for choosing transcription factors to target based on an inferred regulatory network. A total of 1201 transcription factors and 317 signatures are considered reflecting multiple aspects of lymphoma and B cell biology.

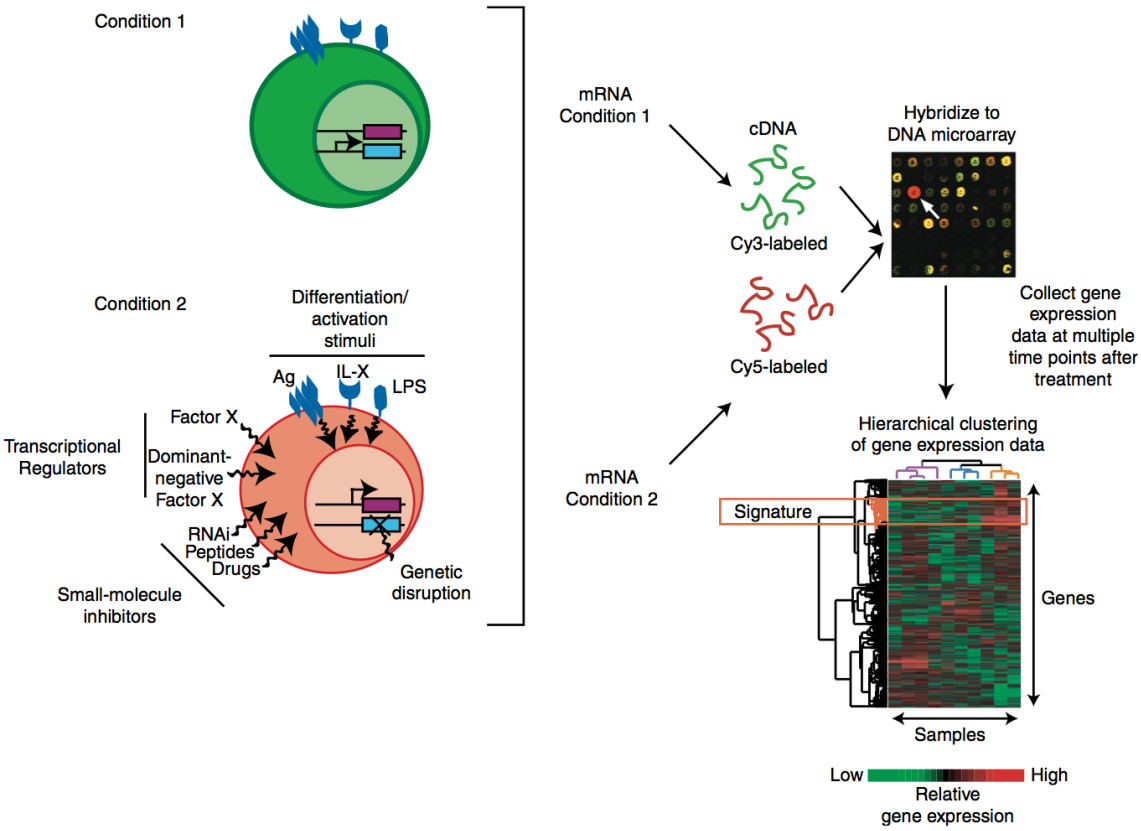

Figure 2.8: Genes whose mRNA levels change coordinately in response to a specific molecular perturbation are hierarchically clustered. Those genes in tightly correlated clusters are identified as part of a signature. (Figure courtesy of the Staudt Lab)

sets is then denoted X_{ij}. Our goal is to estimate the significance of this intersection by constructing an appropriate null distribution for signature j. The expectation value of this distribution

$$\langle X_{ij} \rangle = np_{ij} \tag{2.6}$$

is written in terms of the n genes in signature j and the probability p_{ij} that a gene in signature j is also in neighborhood i. We fit p_{ij} with a logistic regression:

$$p_{ij} = \frac{1}{1 + e^{-v_{ij}}} \tag{2.7}$$

$$v_{ij} = a_j + b_i \tag{2.8}$$

where we assume v_{ij} is a linear combination of the weights a_j and b_i. a_j reflects the tendency of genes in signature j to appear in any given neighborhood, and b_i reflects how likely genes in neighborhood i will also appear in a signature, i.e., how highly connected a regulator is in the network. If there were no correlation between signature j and neighborhood i then $p_{ij} \to p_i$, indicating that the probability of a gene being in neighborhood i and signature j does not depend on the specific signature of interest and thus the probability distribution characterizing X_{ij} would follow a binomial distribution:

$$n = S_j \tag{2.9}$$

$$k = X_{ij} \tag{2.10}$$

$$\text{Binomial}(n, k) = \binom{n}{k} p_i^k (1 - p_i)^{n-k}. \tag{2.11}$$

However, the signatures are chosen for their biological relevance and include genes whose mRNA levels are often coordinate, causing the same subset of genes

from signature j to show up more frequently in a neighborhood. When this is the case, the binomial distribution no longer applies as the signature genes do not appear in a neighborhood independent of each other. Rather, the probability of one gene being in X_{ij} may be highly predictive of other signature genes also being in X_{ij}. Because this behavior is a product of how the signatures are constructed, we would like to account for this effect in our null model of the probability of X_{ij} occuring. In [33] it is shown that a betabinomial distribution models the case where binomial "trials", i.e., signature genes are in neighborhood i, are correlated with each other. The null model for a given signature j that accounts for all of these effects can be written in terms of the following betabinomial distribution, f:

$$f(n, \alpha, \beta) = \binom{n}{k} \frac{B(k+\alpha, n-k+\beta)}{B(\alpha, \beta)} \tag{2.12}$$

$$B(\alpha, \beta) = \int_0^1 p_{ij}^{\alpha-1}(1-p_{ij})^{\beta-1}\,\mathrm{d}p_{ij} \tag{2.13}$$

where the shape parameters α and β allow the betabinomial distribution to take on a variety of different concave and convex forms depending on how highly connected a regulator is with the rest of the network and whether or not there is an all or nothing effect with a given signature, i.e., a link to one signature gene is highly indicative of links to many signature genes. Given the number of genes n in signature j, α, and β, we can calculate a p-value for the significance of X_{ij}:

$$\text{p-value} = \text{Prob}[f(n, \alpha, \beta) \geq X_{ij}]. \tag{2.14}$$

The only remaining task is to estimate α and β for each signature, which can be accomplished using a logistic regression whose parameters we can relate to α and

β. We choose to reparamaterize in terms of a_j and β, which makes for more stable computation as suggested in [34], with an estimate for b_i in terms of p_i and the total number of genes in the network, N,

$$p_i = \frac{R_i}{N} \tag{2.15}$$

$$b_i = \ln\left(\frac{p_i}{1-p_i}\right). \tag{2.16}$$

We can then equate the mean of the regression probability and the mean of the betabinomial distribution to solve for α in terms of β and v:

$$v = a_j + b_i \tag{2.17}$$

$$n\frac{1}{1+e^{-v}} = n\frac{\alpha}{\alpha+\beta} \tag{2.18}$$

$$\alpha = e^v\beta. \tag{2.19}$$

The probability of observing X_{ij} is then:

$$\mathrm{P}(X_{ij}=k) = \binom{n}{k}\frac{B(k+\beta e^v, n-k+\beta)}{B(\beta e^v,\beta)}. \tag{2.20}$$

To finally fit the model, we find a_j and β for each signature that maximizes the sum of log likelihoods:

$$\sum_{R_i} \ln[\mathrm{P}(X_{ij}=k)]. \tag{2.21}$$

We then have a p-value for each signature and regulator neighborhood—for first neighbors and second neighbors separately. The p-values for first and second neighbors can be combined using Fisher's method [35] to obtain a shortest path p-value, p_{tot}, as follows:

$$p_{tot} = p_1p_2(1-\ln(p_1p_2)) \tag{2.22}$$

where p_1 and p_2 are the p-values from the first and second neighbors. A supplementary table containing the p-value results for the ABCDLBCL-4 signature vs. all 1201 TFs is included in Appendix 1.

2.4 Biological Results

The resistance of ABC DLBCL to chemotherapy emphasizes the need for new therapies built on a better understanding of the molecular mechanisms that drive the disease. In the context of this project, we aim to identify new transcription factors (TFs), like IRF4, that are significantly connected to the set of uniquely expressed genes in the ABC DLBCL. Even though we included over 300 gene signatures reflecting a wide range of cellular biology, we focus on the defining ABCDLBCL-4 signature which contains the 288 genes that are more highly expressed in the ABC subtype than in the GCB subtype based on hierarchical clustering of patient gene expression, as we hypothesize that part of the molecular "programming" that uniquely drives ABC DLBCL also enables its resistance to chemotherapy.

Given the reconstructed network, we rank TFs in by the p-value associated with the number of first neighbors who are also in the ABCDLBCL-4 signature or the p-value associated with combined number of first and second neighbors who are also in the signature. Given these criteria, there are 75 TFs (out of a total of 1201 TFs in the network) who have an associated p-value < 0.025 with the ABCDLCBL-4 signature. We chose to consider both 1st neighbors and 1st and 2nd neighbors combined because we are interested in TFs that are locally or globally connected to

the signature genes.

Within the network, one might consider the situation where a TF is not significantly connected to signature genes directly, but through a secondary TF which is highly connected to genes in the signature. We hypothesize that this may be the case for the TF ID3, which does not have a significant number of first neighbors ($P < 0.14$) but does have a significant number of second neighbors in the signature ($P < 0.005$). In a previous study of Burkitt's Lymphoma, ID3 was shown to have mutations that prevented its normal inhibition of the factor TCF3 which lead to activation of a pro-survival pathway [36]. It is possible that ID3 plays an important role in ABC DLBCL as well.

In addition to ID3, there are multiple TFs out of the 75 predicted whose lymphoma biology is known, and show the robustness of this single cell line network in predicting true tumor behavior. For example, IRF4 is a master regulator in the ABC DLBCL subtype and its activity is critical to the survival of ABC DLBCL cell lines [23]; STAT3 works in conjunction with the NF-κB pathway to promote cell growth. In addition, a subset of ABC DLBCL patients show very high levels of STAT3 [21], although the significance of this remains an unanswered question. IRF7 and IRF9 are critical to the control of the interferon pathway in ABC DLCBL [23] and, furthermore, IRF7 and IRF9 are the second and third ranked TFs associated with the interferon signature, which contains the list of genes involved in the interferon pathway. Finally, NFKB1 is an important regulator in the NF-κB pathways, which is a hallmark of the ABC subtype [8].

Beyond known regulators of the ABC DLBCL subtype, the top ranked TF

associated with the ABCDLBCL-4 signature may prove an important part in the pathogenesis of this disease. The role of BATF in ABC DLBCL is currently unknown, but recent work [37] has shown that BATF works directly with IRF4 in normal T cells. Experiments are ongoing to determine the role of BATF in ABC DLBCL.

Given that the network was reconstructed from perturbations within one ABC DLBCL cell line, it is possible that the network might reflect the fact that a cancer cell line, while originally derived from a patient biopsy, has certain evolutionary traits that allow it to live successfully in the lab but do not reflect the actual tumor environment. To investigate this possibility, we compared the gene expression in actual tumor biopsies for the 75 TFs predicted from the network. Of the 75 TFs, 24 had 1.3-fold higher ($P < 0.01$) gene expression in patients with the ABC subtype compared to those with the GCB subtype and 8 TFs had 1.3-fold lower ($P < 0.01$) gene expression in the ABC subtype as shown in Fig. 2.9. This would suggest that the network has predicted TFs that are either activated or repressed in the ABC phenotype. An shRNA screen is in progress to determine the dependence of this tumor type on these factors by determining whether the knockdown of TFs associated with ABCDLBCL-4 via the network leads to the death of ABC DLBCL cells.

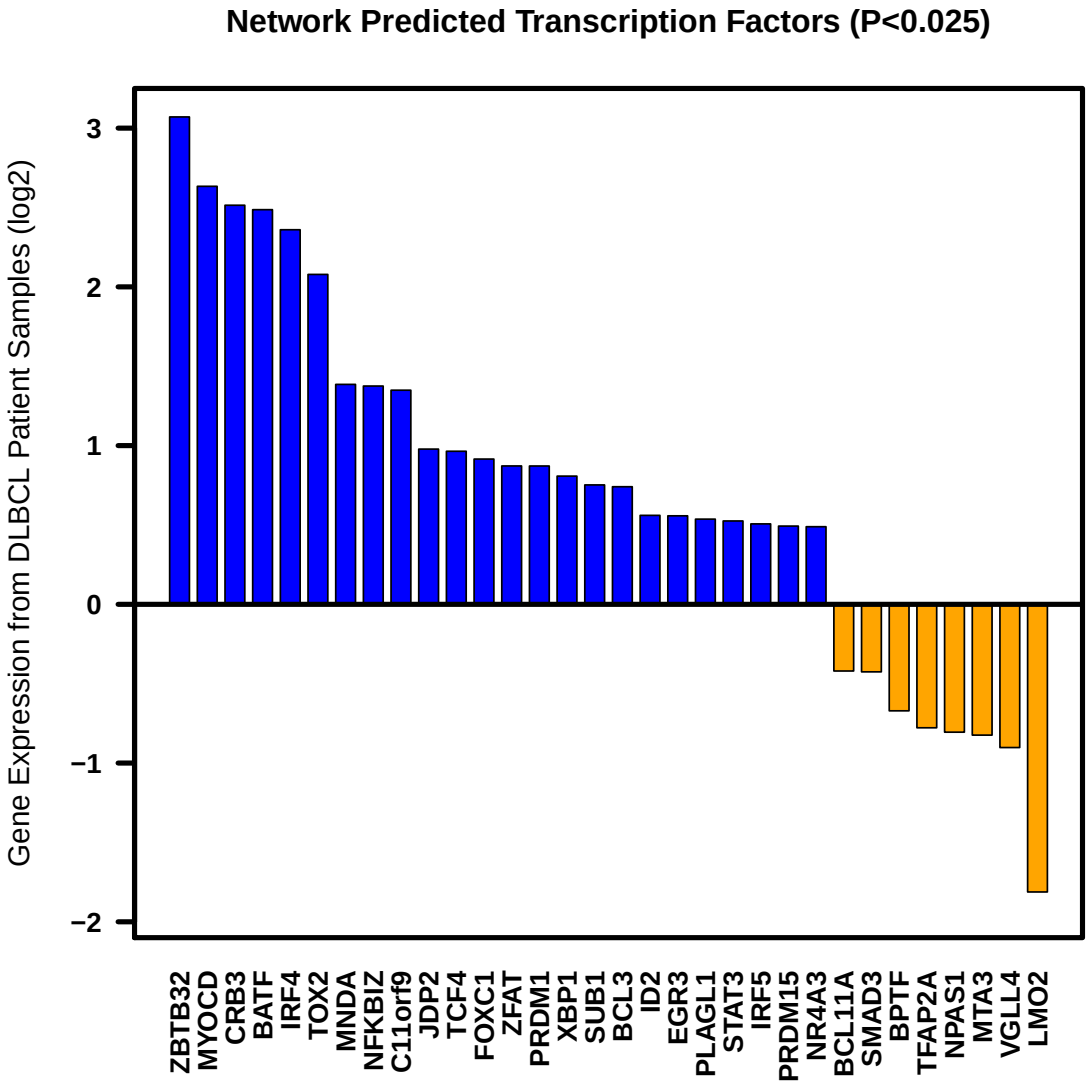

Figure 2.9: log2 ratio of TF gene expression in ABC vs. GCB patient samples for network predicted TFs ($P < 0.025$) where each TF had 1.3-fold higher gene expression in either the ABC subtype or the GCB subtype ($P < 0.01$).

2.4.1 Predicting Known TF Links

Ideally, we would have enough experimental perturbations such that the network structure would remain very robust to the removal of one set of perturbation experiments (typically measured over 4 time points). However, we instead have a very limited set of unique perturbations (∼30). We imagine that the removal of a particular perturbation may lead to the removal of a particular subset of links in the network. Conversely, by adding new perturbations, we can improve the resolution in the nework.

Unlike network reconstruction efforts in model organisms such as yeast and *E. coli*, there is no gold standard to compare the accuracy of our network in ABC DLBCL as we add/remove perturbations. Instead, as a first effort to test the network's predictive power, we remove the perturbations of TCF4, whose neighbors were originally enriched for the ABCDLBCL-4 signature and highly expressed in patient tumors as shown in Fig. 2.9. There were 12 microarrays measuring the response to TCF4 perturbations including the shRNA knockdown of TCF4 (8 arrays) and the over-expression of TCF4 protein (4 arrays). We then reconstructed the network with the remaining 168 microarrays and reran the network signature analysis focusing on the set of genes that have been experimentally validated as TCF4 target genes, meaning that their gene expression levels change when TCF4's levels change and confirmation that TCF4 protein binds and regulates that gene as well. These TCF4-dependent target genes were organized into a signature and were significantly associated with TCF4 first neighbors. However, unlike the network

reconstructed from the arrays that included direct TCF4 perturbations, there was no longer a significant association between TCF4 and the ABCDLCBL-4 signature. This suggests that TCF4 perturbations directly affect a subset of ABC DLBCL-specific genes independent of other perturbations, and that our network resolution is limited by the size of our perturbation set.

2.4.2 Conclusions and Future Work

We have shown that the topology of a gene regulatory network reconstructed using an information theory approach from a single cancer cell line and a limited number of network perturbations can capture important known tumor biology and provide a principled way to predict the roles of previously unknown transcription factors. The success of using simple network properties such as first and second neighbors suggests more advanced techniques may provide further insight. We hope that others will build on this work and continue to employ tools from graph theory and complex networks, especially as the quality of gene regulatory networks improves.

Chapter 3

Robustness of Network Measures to Link Errors

3.1 Abstract

In various applications involving complex networks, network measures are employed to assess the relative importance of network nodes. However, the robustness of such measures in the presence of link inaccuracies has not been well characterized. Here we present two simple stochastic models of false and missing links and study the effect of link errors on three commonly used node centrality measures: degree centrality, betweenness centrality, and dynamical importance. We perform numerical simulations to assess robustness of these three centrality measures. We also develop an analytical theory, which we compare with our simulations, obtaining very good agreement.

3.2 Introduction

As applications of network science continue to grow and the cost of large data sets decreases, complex network models are increasingly moving from a useful means for building insights [38] to a powerful tool for control and prediction [39, 40]. However, the false and missing links that often plague these data sets may pose a challenge to the application of complex network models. For example, networks

based on mobile phone records [41] may miss important highly connected hubs due to a lack of institutional phone numbers, while social media-based networks may show friendships between people where no face-to-face friendship exists. Thus, characterizing the reliability of network properties inferred from measured data with link errors can be an important issue. This is a challenging problem as there is often no "true" network to compare against and only an estimate of the link errors can be made.

Biological networks, in particular, are often constructed from noisy data. For example, recent high-throughput technologies such as yeast two-hybrid screening now make it possible to test potential interactions between proteins in a organism; however, depending the stringency of the screening, the number of reported protein-protein interactions can vary dramatically [42]. When the number of reported interactions is on the high end, many false links are likely to be included, and when the number of reported interactions is on the low end, many true links are likely missed. Further, interactions can also be missed when they are conditioned on other events in the cell. Link errors are also common in the reconstruction of gene regulatory networks from gene expression microarray data; in particular, false links are frequently inferred from non-causal correlations [10, 43].

While much attention has been devoted to improving network reconstruction algorithms to limit the number of false and missing links [44, 10, 45, 46], in this paper we aim to provide a step toward understanding the effect of these link errors on the conclusions we draw from network analysis.

3.3 Approach

We study the effects of false and missing links on three different network measures of node importance: degree centrality, betweenness centrality, and dynamical importance, which are described below. In general, our goal is to understand the extent to which a node importance measure calculated using a noisy network correlates with its value in the true network. In particular, we wish to determine how measures of node importance differ in their robustness to false and missing links. In this section we describe the different measures of node importance considered, the different types of "truth" networks studied, and the different models for false and missing links employed. We limit our considerations to unweighted, undirected networks with no self-links.

3.3.1 Centrality Measures

The number of links connected to a node is its degree, the most basic centrality measure. The use of degree has been especially popular in identifying the function of genes in genetic regulatory networks. Genes with many links can play important roles in multiple biological functions [47]. In social networks a node's number of acquaintances or friends reflects the local influence of that node.

More global measures of node centrality account for a node's neighbors, neighbors of neighbors, and so on. *Betweenness centrality* is such a measure. The betweenness centrality of a node i is defined as [48]

$$g(i) = \sum_{j \neq l} \frac{\sigma_{jl}(i)}{\sigma_{jl}}, \tag{3.1}$$

where $\sigma_{jl}(i)$ is the number of shortest paths between nodes j and l going through i, and σ_{jl} is the total number of shortest paths between j and l. By summing over all pairs j,l we have the fraction of shortest paths that run through i.

One might consider a centrality measure that effectively takes into account all paths instead of only shortest paths. There are multiple eigenvalue metrics [49] that account for such paths. We focus on the *dynamical importance*. The dynamical importance of node i is defined in terms of the decrease, $-\Delta\lambda_i$, of the largest eigenvalue, λ, of the network's adjacency matrix upon the removal of node i [50]:

$$I_i \equiv \frac{\Delta\lambda_i}{\lambda}. \tag{3.2}$$

The dynamical importance measure is motivated by the observation that the largest eigenvalue of the adjacency matrix plays an important role in various processes on networks, including synchronization of oscillators [51] and phase transitions in boolean models of gene regulatory networks [52].

3.3.2 Model Networks

In our investigations of the effects of network noise in the form of link errors on the aforementioned centrality measures, we two types of widely-studied networks as our "truth" networks. The first type is an Erdos-Renyi (ER) [53] random network. To construct an ER network with M links we randomly choose M pairs of nodes and draw an edge between each pair. This kind of network exhibits a Poisson degree distribution if the number of nodes is large. The other type of truth network we explore is the scale-free (SF) network, which exhibits a power-law degree distribution. To

construct our SF networks, we start with a directed variant of the Barabasi-Albert preferential attachment model [54]. Our network begins with a small random seed network to which a single new node is added at every time step. When each new node is added, two directed links originating from it are made to existing nodes in the network. These connections are formed such that the probability of linking to an existing node is proportional to its current in-degree. We then convert this directed network to an undirected network.The resultant network exhibits a degree distribution that is power law in its tail with exponent $\gamma = -2.5$ (in contrast to the $\gamma = -3$ exponent for the original Barabási-Albert construction [54]).

3.3.3 Link Error Models

In order to explore how link errors affect centrality measures, we consider two models for creating missing and false links.

For both of our link error models, denoted Model 1 and Model 2, we create missing links by randomly selecting $M\delta$ $(0 \leq \delta \leq 1)$ of the M true links and deleting them. Models 1 and 2, however, differ in how false links are created. In Model 1 we create false links by connecting $M\alpha$ node pairs randomly selected from among the $N^2 - N - M$ node pairs not connected by a true link. From the nodes' point of view, the expected number of its links that get deleted is proportional to its degree in the truth network, while the expected number of links added is independent of its degree. In our second model of noisy networks (Model 2) , *both* the deletion and addition of links occur in proportion to node degree. Thus in Model 2 false links

are added between node pairs where each node in the pair is randomly selected with probability proportional to the node's true degree. That is, we randomly choose two nodes with probability proportional to their degree; if the two choices do not already have a connecting link, we add a link between them. We repeat this process until $M\alpha$ links have been added.

In what follows we will vary the link deletion fraction δ and the link addition fraction α to numerically (Sec. 3.4) and analytically (Sec. 3.5) explore the effects of missing and false links. While $0 \leq \delta \leq 1$, note that α can be larger than 1. Here we restrict ourselves to $0 \leq \alpha \leq 1$ and hence do not consider noisy networks for which the number of false links exceeds the number of true links. Because some centrality measures are not well-defined when there are multiple disconnected components in a network, only nodes in the giant connected component (GCC) of both the true and noisy network are considered.

3.4 Simulation Results

In this section, we report results of numerical simulations investigating the robustness of network centrality measures in the face of link errors for the two different types of truth networks and the two link error models considered. Using the methods described in Section 3.3, we generated Erdos-Renyi networks with with $N = 2500$ nodes and average degree $\langle k \rangle = 6$ and scale free networks with 2500 nodes and average degree $\langle k \rangle = 4$. Starting with each of these truth networks, we then produced noisy variants for different values of δ (the fraction of true links deleted)

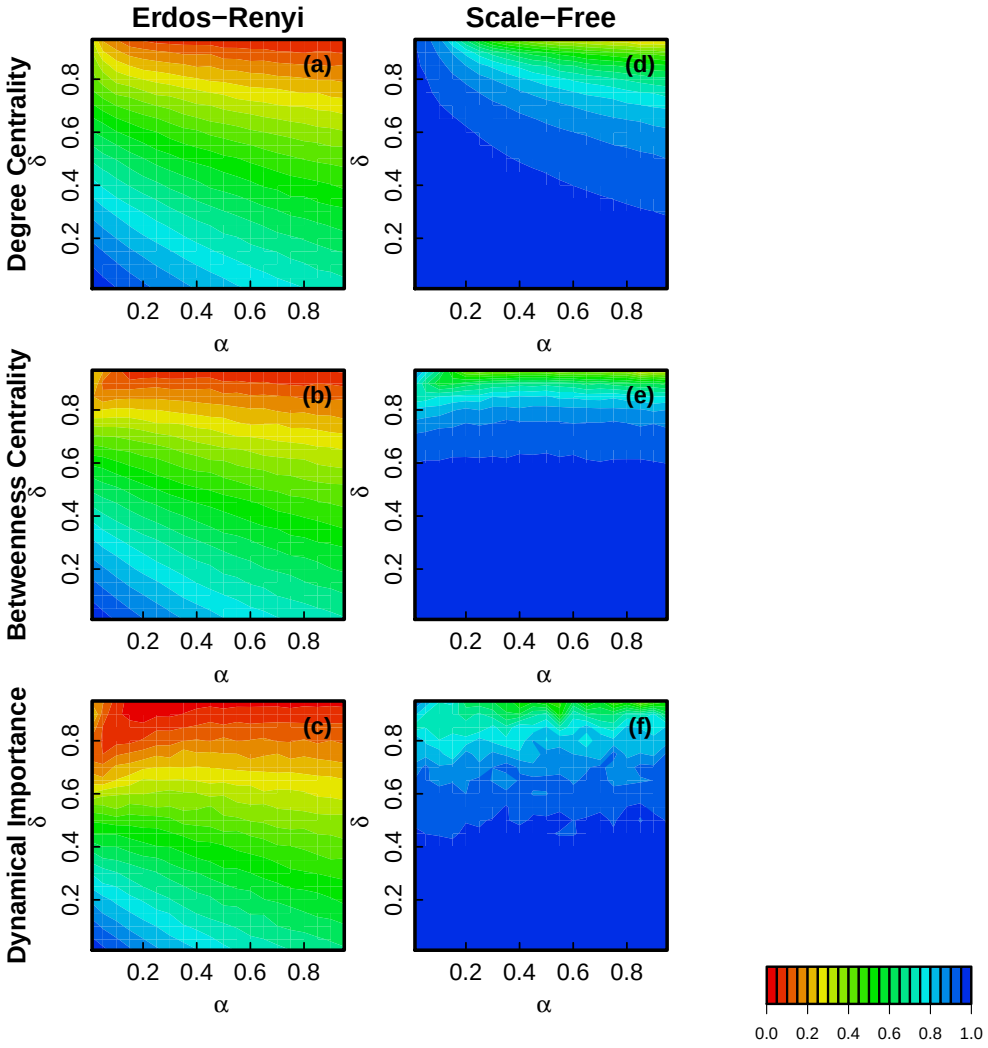

Figure 3.1: Model 1: Contour map of correlation before and after introduction of link errors where δ is the fraction of missing links and α is the fraction of false links for Erdos-Renyi and Scale-Free Networks. False links are added randomly (Model 1), missing links proportional to the original degree. True and noisy network measures are perfectly correlated when ρ is 1 (blue) and not correlated when ρ is 0 (red).

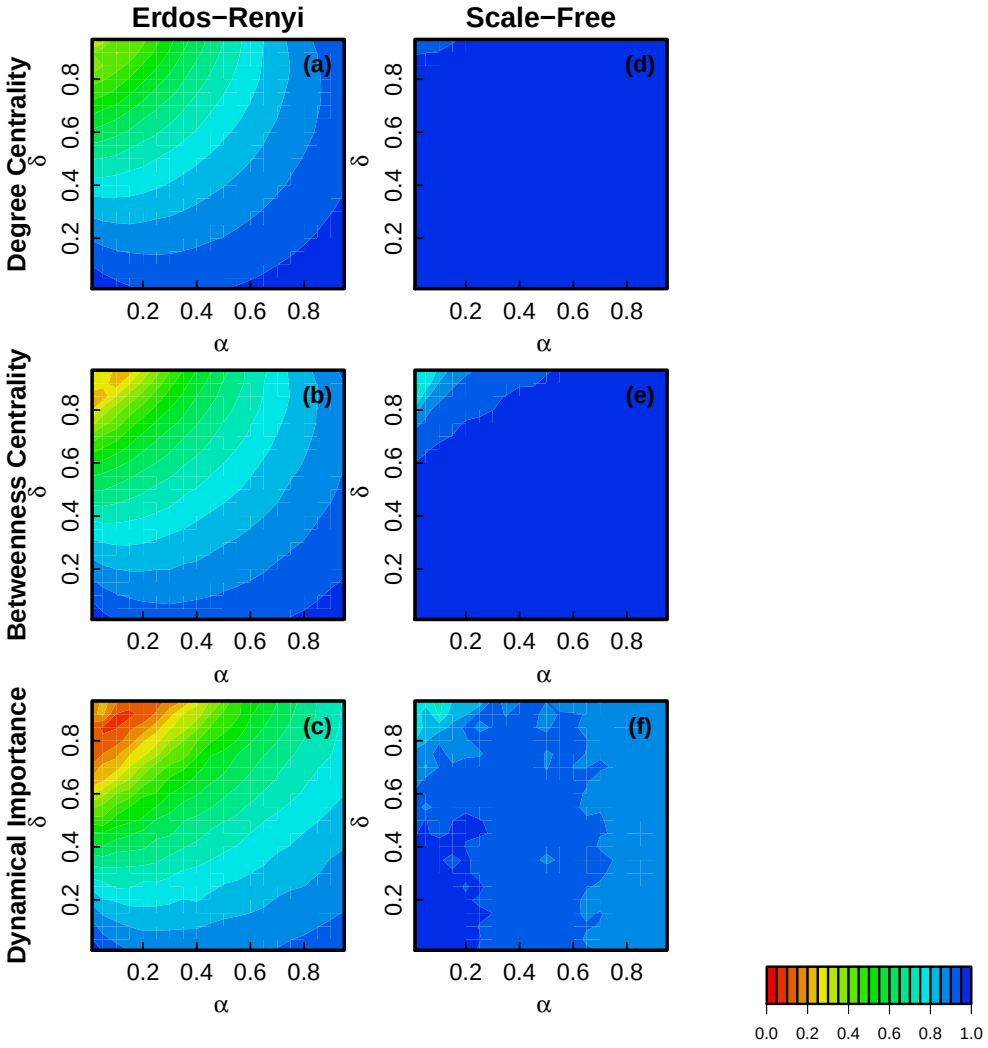

Figure 3.2: Model 2: Contour map of correlation before and after introduction of link errors where α is the fraction of missing links and α is the fraction of false links for Erdos-Renyi and Scale-Free Networks. False and missing links are proportional to the original degree. True and noisy network measures are perfectly correlated when ρ is 1 (blue) and not correlated when ρ is 0 (red).

and α (the fraction of false links added). For all values of α and δ, results from the noisy networks were averaged over 25 realizations.

3.4.1 Centrality Correlations

To assess the effect of link noise on a node's centrality measure C, we calculated the Pearson correlation ρ between the true measure C_T and noisy measure C_N [1]:

$$\rho(C_T, C_N) = \frac{\langle C_T C_N \rangle - \langle C_T \rangle \langle C_N \rangle}{\sqrt{(\langle C_T^2 \rangle - \langle C_T \rangle^2)(\langle C_N^2 \rangle - \langle C_N \rangle^2)}}, \tag{3.3}$$

where C denotes either the node's degree centrality, betweenness centrality, or dynamical importance, and $\langle ... \rangle$ indicates an average over nodes in the giant connected component of the network. We used the standard definitions of degree centrality and betweenness centrality from Section 2 when calculating the correlation ρ. In the case of dynamical importance, we employed a perturbation-based large-N approximation [50] of Eq. (3.2) using the left and right eigenvectors, u and v, associated with the largest eigenvalue of the network adjacency matrix:

$$\hat{I}_i = \frac{v_i u_i}{v^T u}. \tag{3.4}$$

The computational feasibility of Eq. (3.4) makes it amenable to application in very large networks ($N \approx 50\,000$), and it extends naturally to directed networks (only

[1] We only consider the correlation for each individual centrality measure before and after link errors are added. We do not study correlations between the different centrality measures, since we regard the latter issue as being more context-dependent. E.g., it may be more appropriate to choose a centrality measure because its character makes it more indicative of the particular processes that the network is experiencing, than to choose it because it is (by some criterion) more robust.

undirected networks are considered in this work).

Simulation results are shown in **??**. Figures 3.1 and 3.2 show heat maps of the correlation ρ in α, δ-space for the three node centrality measures and the two types of network considered, with Fig. 3.1 showing the system behavior for Model 1 link errors and Fig. 3.2 showing the system behavior for Model 2 link errors. In order to more quantitatively compare results from the various cases, Fig. 3.3 shows plots of ρ versus δ with α held fixed at $\alpha = 0.5$ (Figs. 3.3(a) and 3.3(c)) and of ρ versus α with δ held fixed at $\delta = 0.5$ (Figs. 3.3(b) and 3.3(d)).

In Figs. 3.1(a,b,c), which show Model 1 results for our ER network, we see that for low link deletions $\delta \lesssim 0.5$ all of the centrality measure correlations decrease as α and δ are increased, but that this decrease of correlation is somewhat faster when δ is increases as compared to when α is increased. At higher false deletion error, $\delta \gtrsim 0.5$, the betweenness, and especially the dynamical importance, become even less sensitive to false link additions (α).

Results using Model 1 link errors on SF truth networks (Figs. 3.1(d,e,f)), show that all three centrality measures are significantly more robust to link errors as compared to our results for the ER network, with very small error for values of $\delta \lesssim 0.7$. In addition the insensitivity of ρ to α for betweenness centrality and dynamical importance found for the ER network still applies.

Looking at Fig. 3.2, which shows results for Model 2 link errors, we again see that the centrality measures for the SF network (Figs. 3.2(d,e,f)) are very much more robust to link errors than is the case for the ER network (Figs. 3.2(a,b,c)) with perceptible SF error only appearing near $(\alpha, \delta) \approx (0, 1)$. Furthermore, particularly

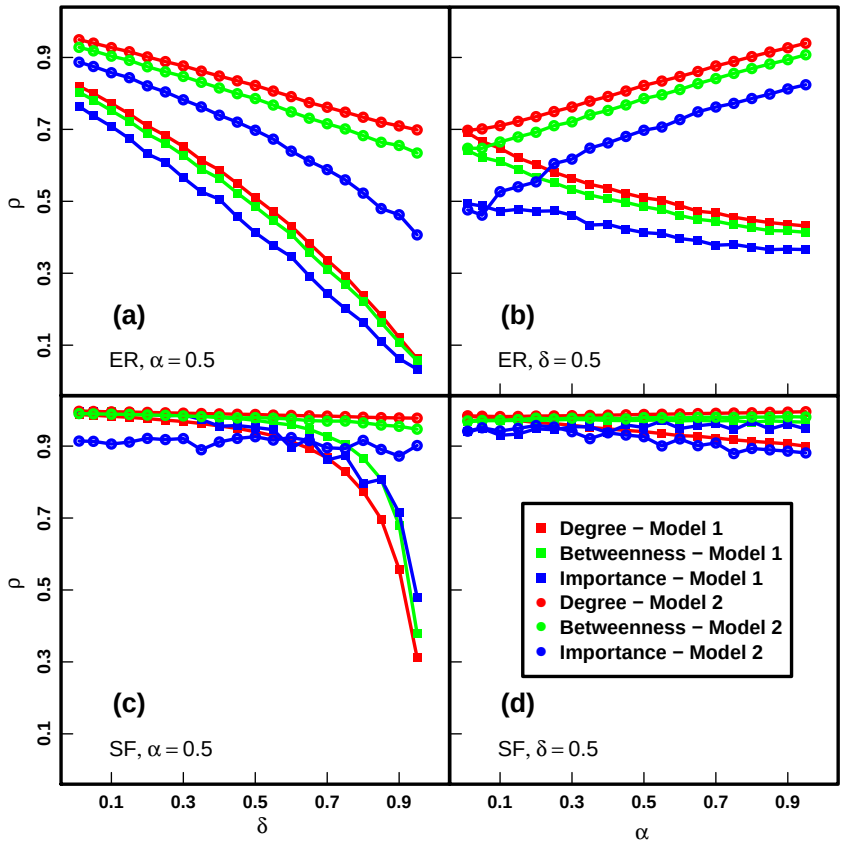

Figure 3.3: The centrality correlation for the three measures studied: degree centrality (red), betweenness centrality (green), and dynamical importance (blue). Squares correspond to results in which noise is added according to Model 1, circles correspond to results in which noise is added according to Model 2. Panels (a) and (b) are for Erdos-Reyni truth networks, and panels (c) and (d) are for scale-free truth networks. In (a) and (c), the fraction of false edges is fixed at $\alpha = 0.5$, and the fraction of true edges deleted, δ is varied. In (b) and (d), the fraction of true edges deleted is fixed at $\delta = 0.5$ and the fraction of false edges added, α, is varied.

for the ER network, we still see that the correlation between the centrality measures of the true and noisy networks decreases when links are deleted (δ is increased). In contrast, with false link additions (α increasing), we find that the correlation actually *increases*. E.g., for the case of degree (Fig. 3.2(a)), this occurs because Model 2 noise is added in proportion to the signal we are measuring (the true degree), and this effect can also be seen for the betweenness centrality and dynamical importance measures (Figs. 3.2(b,c)).

Figure 3.3 shows graphs of the correlation ρ along two slices through α-δ space: (i) $\alpha = 0.5$ with δ varying from 0 to 1 (Figs 3.3(a) and 3.3(c)), and (ii) $\delta = 0.5$ with α varying from 0 to 1 (Figs. 3.3(b) and 3.3(c)). Referring to Figs 3.3(a), we see that for ER networks at $\alpha = 0.5$, as true links are deleted, the correlation decreases more slowly for Model 2 than for Model 1. As already seen in Fig. 3.2(a), Fig. 3.3(b) shows a pronounced increase of the correlation for ER networks with increase of Model 2 link error additions (α) at fixed $\delta = 0.5$. Figure 3.3(c) shows that for scale-free networks at $\alpha = 0.5$ the correlations are relatively insensitive to link deletion for Model 2, while Model 1 shows significant decrease only for relatively large $\delta \gtrsim 0.6$. Finally, we see from Fig. 3.3(d) that at fixed $\delta = 0.5$ the scale-free network is largely unaffected by the addition of false links for both Model 1 and Model 2.

3.4.2 Overlap of Highly Ranked Nodes

In addition to correlation robustness, we have also characterized the effect of link errors on the overlap between the top 10% of nodes in the truth and noisy networks when ranked based on a given centrality measure. This consideration of overlap is motivated by the fact that node-ranking is often used to select nodes for further study or experimental validation (e.g., see the gene network study of human glioma in Ref. [55]).

With this motivation, we have studied the effects of missing and false links on the ranking of the nodes based on the three centrality measures from Section 3.3. To do this, we consider the overlap of the top 10% of the nodes in the giant components of the true and noisy networks and average over 25 network realizations. Results for Model 1 and Model 2 link errors are shown in Figure 3.4 and 3.5.

While the correlation results, Figs. 1-3, show a striking contrast between the ER and the SF networks, with the SF networks being very much more robust to link errors, this result is no longer true when we focus on overlap (Figs. 3.4 and 3.5) with the SF and ER networks now *both* showing substantial dependence of the overlap on α and δ. Similar to the correlation for ER networks with Model 2 link errors, Figs 3.2(a,b,c), we now see from Fig. 3.5 that the overlap with Model 2 link errors shows substantial decrease with increasing δ, and increase with increasing α, applying for both ER and SF networks.

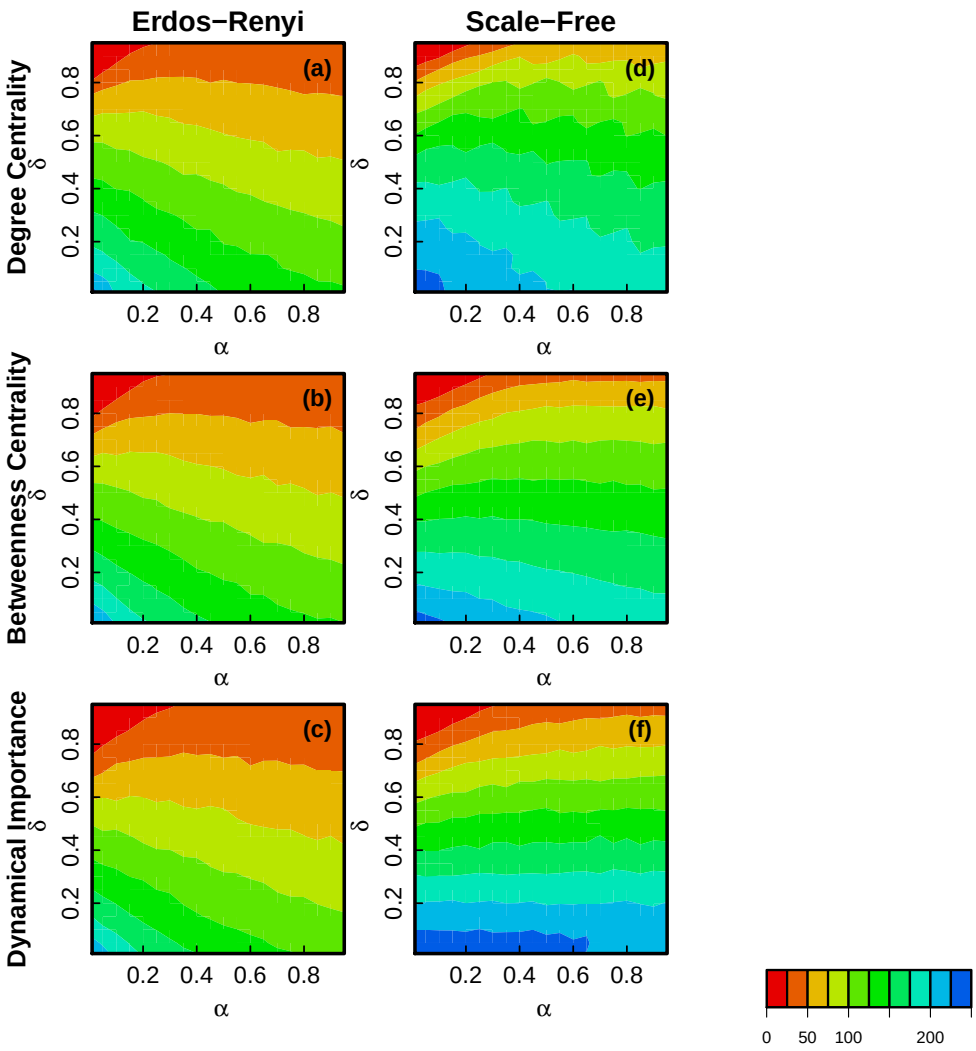

Figure 3.4: Model 1: Contour map of the overlap between the top 10% (250) nodes in the true network and the top 10% in the noisy network as ranked by each node's centrality measure before and after introduction of link errors where α is the fraction of missing links and α is the fraction of false links for Erdos-Renyi and Scale-Free Networks. False links are added randomly, missing links proportional to the original degree.

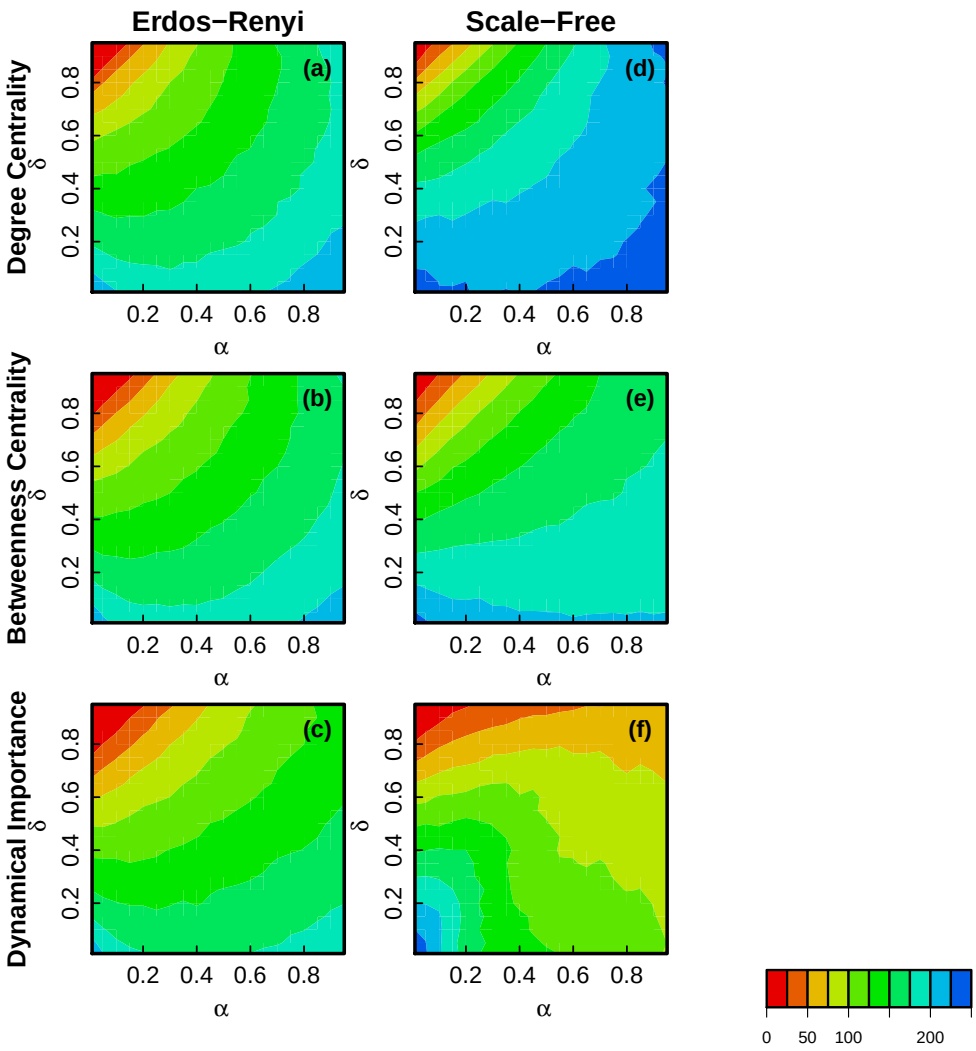

Figure 3.5: Model 2: Contour map of the overlap between the top 10% (250) nodes in the true network and the top 10% in the noisy network as ranked by each node's centrality measure before and after introduction of link errors where α is the fraction of missing links and α is the fraction of false links for Erdos-Renyi and Scale-Free Networks. The number of false and missing links for each node is proportional to the original degree of that node.

3.4.3 Centrality Changes for Individual Nodes

In sections 3.4.1 and 3.4.2 we explored the robustness of the three different centrality measures by assessing the effect of link errors on centrality correlation (Eq. (3.3)) and on overlap of highly ranked nodes, both of which are population-wide characterizations. In some cases, however, we may be interested in how the centrality of a specific node in the noisy network is related to its centrality in the true network. In this section, we address such situations.

To assess the effects of link errors on the centrality of a specific node with true degree k, we consider the set of nodes in the true network with degree k, and for each node in that set, take the ratio of its centrality measure (e.g., betweenness) in the noisy network to the same measure in true network. After repeating the process for 500 realizations of the noisy network (constructed from a single underlying truth network, randomly generated as described in section 3.3.2), we obtain a distribution of the noisy/true centrality ratios for a given value of k. We do this for both ER and SF true networks. For both models, we focus on an example in which the expected number of false links added is equal to the expected number of true links deleted ($\alpha = \delta = 0.3$). The first, second and third quartiles of this distribution are plotted for the three centrality measures with noise generated according to Model 1 in Fig. 3.6 and noise generated according to Model 2 in Fig. 3.7.

For Model 1, in which false links are added independent of node degree and true links are removed proportional to node degree, we see that, in both ER and SF networks, the median noisy/true ratio exhibits a general downward trend as k

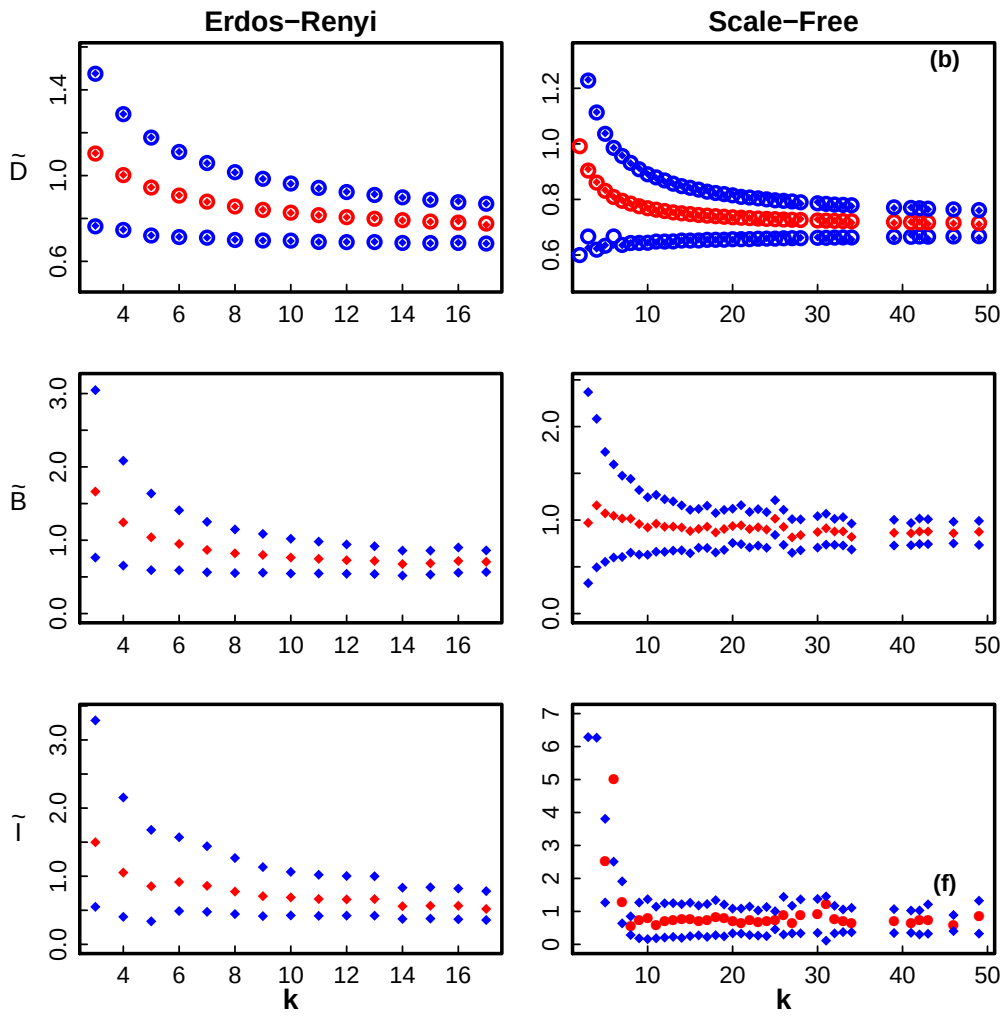

Figure 3.6: Model 1: The first (lower blue), second (red) and third (upper blue) quartiles for the ratio of noisy/true degree ($\tilde{D}$), betweenness ($\tilde{B}$), and dynamical importance ($\tilde{I}$) versus degree (k) in ER (left column) and SF (right column) networks for $\alpha = \delta = 0.3$. The open circles are derived from the theory described in 3.5. Results are averaged over 500 realizations of the noise model with the same underlying true network.

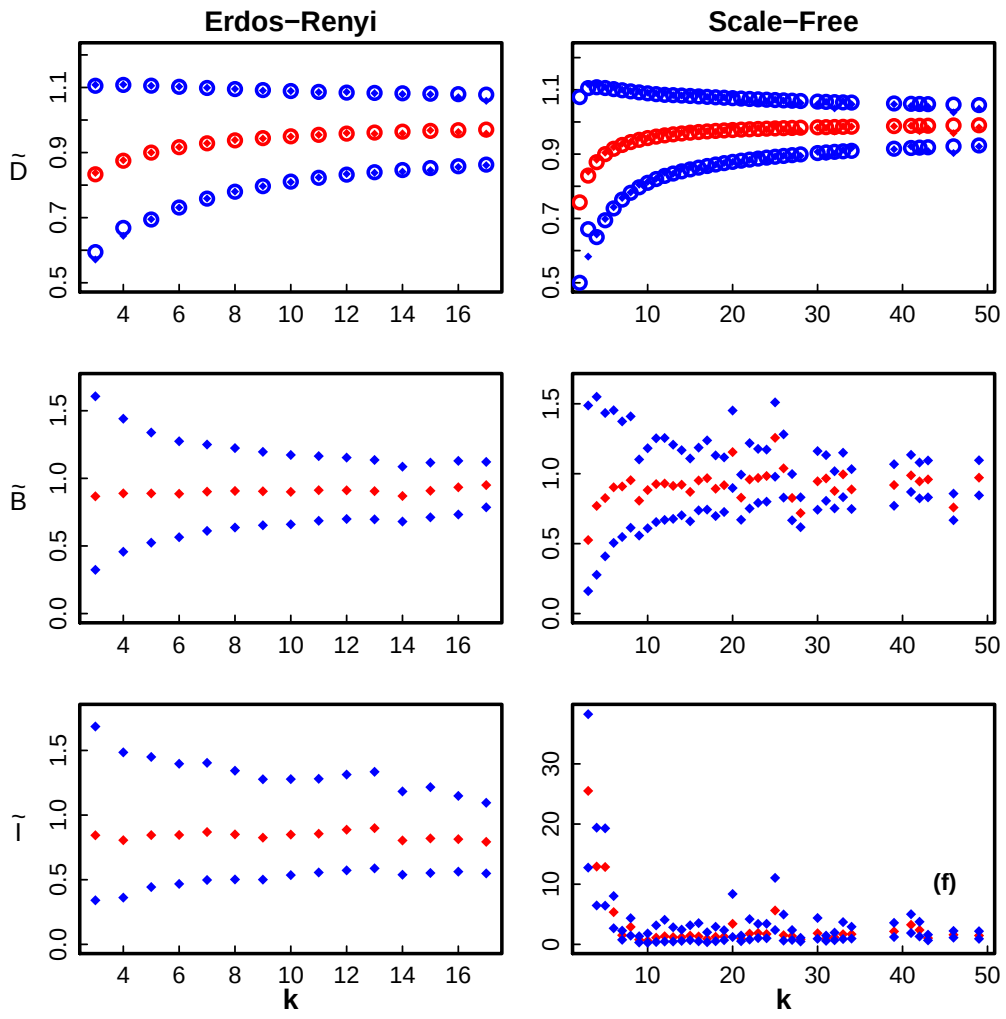

Figure 3.7: Model 2: The first (lower blue), second (red) and third (upper blue) quartiles for the ratio of noisy/true degree ($\tilde{D}$), betweenness ($\tilde{B}$), and dynamical importance ($\tilde{I}$) versus degree (k) in ER (left column) and SF (right column) networks for $\alpha = 0.3$, $\delta = 0.3$. The solid curves for the degree are derived from the theory described in 3.5. Results are averaged over 500 independent realizations of the noise model with the same underlying true network.

increases. This trend occurs because low degree nodes are less likely to have links removed than high degree nodes while being equally likely to have links added. For the betweenness and dynamical importance measures (Figs. 3.6(c, d, e, and f)), we see that for large degrees the median ratios are very close to 1. Since the first and third quartile boundaries are also reasonably close to 1 at large k, this implies that for higher degree nodes, the betweenness and dynamical importance of a specific node in the noisy network are good predictors of its corresponding measure in the true network, in the case ($\alpha = \delta$) that the total number of links in the noisy network is approximately equal to the number of links in the true network.

For Model 2, in which both link additions and deletions are proportional to the original degree, the median noisy/true ratio approaches one at large degree for all centrality measures, as shown in Fig. 3.7. For degree and betweenness centrality, we see that the first and third quartiles are roughly symmetric about the median, whereas for dynamical importance, the third quartile is significantly further from the median than the first, indicating a right-skewed distribution of noisy/true ratios for a given value of degree k. The scatter observed for betweenness and dynamical importance ratios in SF networks (Figs. 3.7(d and f)) occurs because the noisy networks are built from a single random realization of the true network.

3.5 Analysis of Degree Centrality

In this section, we derive analytic approximations for the Pearson correlation ρ of degree centrality before and after the addition of noise and for the probability

distribution function of the noisy node degree. We derive our analytic approximations for both models of link errors studied and show that the predictions of our analytic approximations are consistent with the numerical results presented in the preceding section.

As described in Sec. 3.3.3 and employed in our numerical simulations (Sec. 3.4), for both Model 1 and Model 2 we use a micro-canonical procedure in which, for given values of δ and α, the number of falsely deleted links is precisely $M\delta$ (or rather the integer nearest to $M\delta$) and the number of falsely added links $M\alpha$. However, because this procedure is hard to analyze, to facilitate the theory we employ a closely related canonical procedures that should yield results that are good approximations to the actual Model 1 and Model 2 results. Specifically, for link deletion, each one of the M true links is deleted with probability δ. Thus the *average* number of missing links is $M\delta$ with fluctuations whose ratio to the average decreases as $(M\delta)^{-1/2}$, and we expect a good approximation for link deletions when $M\delta \gg 1$. Similarly for Model 1 link addition, each of the $N^2 - N - M$ pairs of truly unconnected node pairs is connected with probability $M\alpha/(N^2 - N - M)$, creating on average $M\alpha$ false links; while Model 2 addition of truly unconnected node pairs is done with a probability that is proportional to the product of the true degrees of the node pairs.

With the canonical framework assumed [2], we first consider the creation of missing links. If we delete each true link with a probability δ, then the probability

[2]While the canonical framework is more easily treated by theory, the micro-canonical framework allows faster numerics, and that is why it is used in 3.4

that s links are deleted from a node with true degree k is

$$p_D(s|k) = \binom{k}{s}\delta^s(1-\delta)^{k-s}. \tag{3.5}$$

Next we consider Model 1 link addition. For large N, the probability of adding r false links to a node is approximately given by a Poisson distribution,

$$p_A(r) = \frac{u^r}{r!}e^{-u}, \tag{3.6}$$

where u is the average number of false links per node, $u = 2M\alpha/N$, while for Model 2 the probability that a randomly chosen node has r false links is

$$p_A(r|k) = \binom{k}{r}\alpha^r(1-\alpha)^{k-r}. \tag{3.7}$$

From our knowledge of p_A and p_D, Eqs. (3.5,3.6,3.7), we obtain the joint probability that a randomly chosen node has true degree k and noisy degree n. Since $n = k - s + r$,

$$p(n|k) = \sum_r p_A(r)p_D(s = k + r - n|k), \tag{3.8}$$

$$p(n,k) = p_0(k)p(n|k) \tag{3.9}$$

where $p_0(k)$ is the probability that a randomly chosen node has degree k. In particular for Model 1,

$$p(n|k) = \sum_r \frac{u^r e^{-u}}{r!}\binom{k}{k+r-n}\delta^{k+r-n}(1-\delta)^{n-r}, \tag{3.10}$$

while for Model 2,

$$p(n|k) = \sum_r \binom{k}{r}\alpha^r(1-\alpha)^{k-r}\binom{k}{k+r-n}\delta^{k+r-n}(1-\delta)^{n-r} \tag{3.11}$$

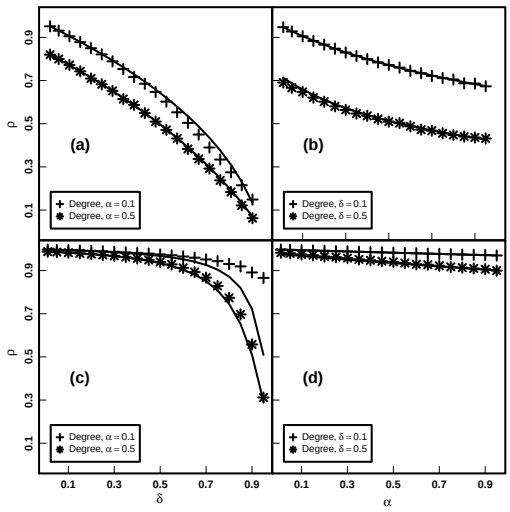

Figure 3.8: Model 1: Pearson correlation between the true degree centrality k and the noisy degree n as a function of missing link fraction, δ, and false link fraction, α. Markers reflect simulation results and theoretical results are plotted as solid lines.

The Pearson correlation of ensemble averaged degree, between the true and noisy networks is then:

$$\rho(k,n) = \frac{\langle kn \rangle - \langle k \rangle \langle n \rangle}{\sqrt{(\langle k^2 \rangle - \langle k \rangle^2)(\langle n^2 \rangle - \langle n \rangle^2)}} \tag{3.12}$$

and we use our theory for $p(n,k)$ along with

$$\langle k^x n^y \rangle = \sum_{k,n} k^x n^y p(n,k), \tag{3.13}$$

to obtain an analytical prediction of $\rho(k,n)$.

In order to compare our analysis to our numerical simulations, we take as our $p_0(k)$ the specific numerically generated degree distributions obtained from building our ER and SF networks. We then use the appropriate forms for $p(n,k)$ from Eq. (3.10) (for Model 1) and Eq. (3.11) (for Model 2) to calculate the expected correlations for degree centrality (Eq. (3.12))

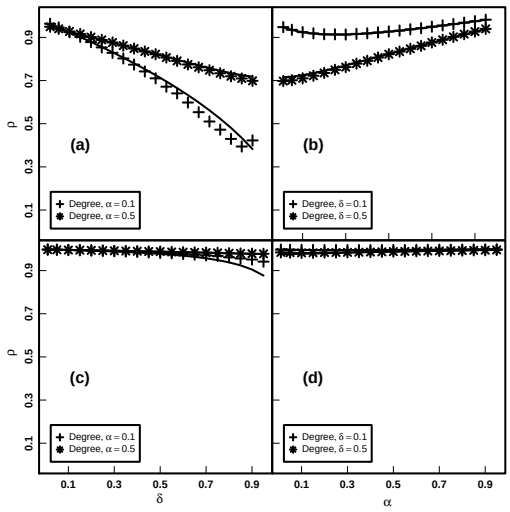

Figure 3.9: Model 2: Pearson correlation between the true degree centrality k and the degree noisy n as a function of missing link fraction, δ, and false link fraction, α. Markers reflect simulation results and theoretical results are plotted as solid lines.

For Model 1, Fig. 3.8 shows comparisons between our numerical simulations (plotted as symbols) and our theory (plotted as lines). We see that the analytical results are in good agreement with the numerical calculations. In SF as compared to ER networks, the degree remains strongly correlated in the presence of many false and missing links, with the correlation being driven by the resilient 'hub' nodes in the SF networks.

For Model 2, Fig. 3.9 shows the theory and simulation of correlation for the degree centrality match well, with a slight discrepancy when $\delta \to 1$.

The derived forms of $p(n|k)$ for the two models (Eqs. 3.10 and 3.11) provide theoretical predictions for how the degree centrality n of a specific node in the noisy network relates to its degree centrality k in the true network. In order to compare

the theory to the simulation results discussed in Sec 3.4.3, we used Eqs. (3.10) and (3.11) to find the first, second, and third quartiles of the distribution as a function of k, again taking $p_0(k)$ as the specific numerically generated degree distributions obtained from building the ER and SF networks. Figures 3.6(a,b) and 3.7(a,b) show that the theoretical predictions (plotted as open circles) are in very good agreement with the numerical results (plotted as solid diamonds).

3.6 Discussion and Conclusions

In this chapter we have investigated the effect of two types of link errors on three node centrality measures. We propose two simple models of link error (labeled Model 1 and Model 2) and study their effect for two types of network topology (Erdos-Renyi and scale-free). In Model 1, the probability that a link is deleted depends on the original number of links to which that link is connected, while false links do not depend on the structure of the original network. Model 2 follows the same formulation as Model 1 for deleting links, but in Model 2 the addition of false links is performed with a probability that is proportional to the product of the true degrees of the node pairs.

We have developed methods for assessing the robustness of node centrality to link errors by comparing the centrality measure for each node before and after link error. We compare in two ways: (i) by calculating the correlation between the nodes' centrality measures in the true and noisy network (Sec. 3.4.1 and 3.5), and (ii) by calculating the overlap between the top 10% of the nodes as determined by

their centrality measures in the true and noisy networks (Sec 3.4.2). In the case of correlation we have obtained analytical results (Sec 3.5) which are in good agreement with our numerical simulation results. The analytical and numerical results for the correlation suggest that degree centrality, betweenness centrality, and dynamical importance are relatively robust to the presence of false edges when the network is scale-free. Our result for the relative insensitivity of SF networks to δ is consistent with previous work showing that the size of the giant component in SF networks has a high tolerance to the random removal of nodes [56]. We note, however, that when considering the overlap between top ranked nodes in the true and noisy networks, the much larger SF robustness as compared to the ER case no longer applies.

One common link error not addressed here is that of false edges which complete triangles. This is a common problem in network reconstruction using microarray data for gene regulatory networks [29, 43]. In addition to affecting node centrality, these false links may substantially skew the enrichment for network motifs, which are often of interest in biological networks [7, 57].

Chapter 4

Conclusion

The number of large measurable networks now available has brought fresh opportunities and challenges for the field of complex networks. We are now able to test many theoretical ideas on real world networks. However, errors in the measurements of these networks makes clear the need for quantitative methods to test the robustness of the theoretical techniques.

In Chapter 2, we have seen that the first and second neighbors of transcription factors in a reconstructed gene regulatory network can provide valuable insight into the biology of cancer cells. Furthermore, the network provided insight into patient tumor biology despite the fact that the data was taken from laboratory isolated cell lines. We hope this network based approach to predicting the role of transcription factors in ABC DLBCL will advance the systematic study of gene regulation in cancer.

In Chapter 3, we showed there is a reasonable expectation for success in ranking genes based on their first and second neighbors in a network inferred from gene expression, which has been shown to produce false links. For two simple stochastic models of false and missing links, we demonstrated that ranking nodes based on their centrality—degree, betweenness, or dynamical importance—remains relatively robust for a large number of link errors as measured by correlating or ranking nodes

between networks with and without errors. In addition we found centrality measures in networks exhibiting a power law in the tail of their degree distribution to be very robust.

We hope the applications of complex networks in this thesis will inspire other scientists to tackle both the theoretical and applied challenges of elucidating the key mechanisms in biological networks. While the network reconstruction method in Ch. 2 is specifically designed for gene expression microarrays, the network method and statistical model for identifying transcription factors can be applied to networks measured from many of the newly emerging, more accurate technologies measuring biological networks. The theory developed in Ch. 3 can be readily applied to directed networks, and other, more application-specific stochastic models for link error may be relevant where the experimental uncertainty is well characterized.

Appendix A1

Transcription Factors associated with the ABCDLBCL-4 Signature

Appendix 1

TF	1st_neighbor_pval	2nd_neighbor_pval	combined_pval
ESR2	0.003663264	0.006107323	0.000261933
LMO2	0.001067544	0.030189567	0.000365559
ZNF219	0.00129633	0.028168394	0.000409623
TCF7	0.001210298	0.035297307	0.000472522
BATF	0.000401573	0.128796858	0.000562193
CBFA2T3	0.000769215	0.115089227	0.000914692
RELB	0.016392025	0.00652693	0.001085172
POU2AF1	0.033312982	0.003453433	0.001158517
CEBPD	0.00032453	0.359743443	0.001173954
XBP1	0.083700015	0.002055746	0.001663471
MNDA	0.008905215	0.020043249	0.001719029
VGLL4	0.014673846	0.013937994	0.00194192
KLF6	0.002309483	0.093029853	0.002029388
ZNF215	0.001692226	0.150133343	0.00235715
ELK3	0.004881936	0.052703873	0.002383928
ZFAT	0.013667248	0.021153321	0.002644963
STAT3	0.002487039	0.119099324	0.002702708
RBPJ	0.067671152	0.004799907	0.002933814
TRIM22	0.009821305	0.034144771	0.003018232
ZEB2	0.002078057	0.181126117	0.003344185
C11orf9	0.014146995	0.028840797	0.003592214
TFAP2B	0.1784078	0.002368992	0.003706177
MTA3	0.009322245	0.049835656	0.004029944
ID2	0.020427185	0.023555367	0.004156966
TOX2	0.011181081	0.046750666	0.004472658
USP7	0.002035649	0.270561932	0.004683848
ZNF589	0.00303983	0.182643467	0.004717119
SMAD3	0.107305603	0.005746902	0.005174622
BCL3	0.006716707	0.095218698	0.005343322
PWWP2B	0.179604188	0.003571794	0.005357684
E2F2	0.008532861	0.076740055	0.005455347
JDP2	0.003556838	0.187136505	0.00553445
ID3	0.138463988	0.005309192	0.006039447
CREG1	0.008695977	0.088726504	0.006301431
TFAP2A	0.443126573	0.001759769	0.00636042
SUB1	0.015365853	0.052852392	0.006591069
NFKB1	0.054915572	0.015076573	0.006703464
PLAGL1	0.008126645	0.108445617	0.007080457
IRF9	0.160499088	0.005900539	0.007540431
IRF7	0.061651503	0.015466676	0.007585751

TF	1st_neighbor_pval	2nd_neighbor_pval	combined_pval
FOXC1	0.172899123	0.005526668	0.007599745
ZBTB32	0.007936692	0.151073551	0.009263958
PRDM1	0.025205709	0.049760674	0.009634187
NPAS1	0.01258378	0.109009178	0.010413854
CIITA	0.097014846	0.0146164	0.010718015
ELL	0.002661738	0.618855882	0.012203781
IRF5	0.004690171	0.355536772	0.012333728
JUN	0.028557139	0.058542171	0.012361026
SMARCA2	0.316424545	0.005484175	0.012766041
NFKBIZ	0.016056682	0.121585858	0.014131991
BPTF	0.021741833	0.093346258	0.014612447
E2F5	0.00538526	0.3925047	0.015132891
TCF4	0.023104891	0.091875335	0.015188488
NR4A3	0.057773055	0.039007756	0.015989785
NOTCH1	0.167948827	0.013746093	0.016324617
EGR3	0.14566512	0.015881054	0.016352998
NFE2L3	0.098437129	0.02515496	0.017335843
ZHX3	0.06683224	0.039979971	0.018503117
THRA	0.085814024	0.031719757	0.018799189
BCL11A	0.038066713	0.071828793	0.018871728
CRB3	0.019568895	0.145105358	0.019490992
SNAI3	0.103511258	0.027872191	0.019757645
MYOCD	0.006970942	0.445480074	0.021037975
EGR2	0.06161607	0.051418548	0.021399948
IFRD1	0.017527772	0.188371336	0.022165531
IRF4	0.008747722	0.379378418	0.022262427
TCF3	0.149540319	0.022291572	0.022346828
SOX4	0.499945704	0.006751379	0.022585179
NOTCH4	0.037061533	0.095163231	0.023444463
PRDM15	0.207801045	0.018326102	0.02502206
IRF8	0.124117346	0.030889582	0.025165415
VDR	0.058274271	0.068530253	0.026050228
ETV6	0.07092219	0.058706779	0.026985927
SETD8	0.075607094	0.060682082	0.029291197
POLE4	0.061651503	0.076319678	0.029920898
FOXL1	0.196933994	0.024387479	0.03044241
SFN	0.016026348	0.299859503	0.030458047
MYB	0.044865259	0.114293278	0.032167079
EGR1	0.113607641	0.047022871	0.033292997
ZNF710	0.026738274	0.205689223	0.034115367
CREM	0.019168721	0.288031946	0.034226823

TF	1st_neighbor_pval	2nd_neighbor_pval	combined_pval
EBF1	0.031858393	0.175060528	0.034517397
HOXA7	0.029620788	0.190819623	0.034906494
IKZF2	0.024313926	0.236601957	0.035425737
ATF5	0.008564434	0.678493676	0.035725601
SNAI1	0.024929863	0.237749527	0.036322385
ID1	0.032189896	0.189160947	0.037151005
STAT5A	0.011920315	0.526294669	0.038089558
TBX21	0.157109439	0.040012005	0.038153776
TSC22D3	0.10718155	0.06186796	0.039892615
ZHX2	0.137054243	0.050559196	0.041382016
HDAC4	0.048938722	0.148937465	0.043160065
FOSL2	0.104210726	0.070121148	0.043251415
SALL2	0.377333226	0.019456512	0.043419623
ELL3	0.025519466	0.292714697	0.044049167
MAFF	0.074793945	0.10640949	0.046427407
CHD1	0.204802846	0.039625637	0.047183077
FLI1	0.215203724	0.038409286	0.047905611
SIRT3	0.839181419	0.0098619	0.047954051
SHOX2	0.146515583	0.058187788	0.049146503
PHF21B	0.184424311	0.046421033	0.049316763
RCAN1	0.156479032	0.055013921	0.049542082
DNMT3A	0.105106059	0.082735166	0.049957426

Bibliography

[1] M. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[2] Q.C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C.A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, et al. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, 490(7421):556–560, 2012.

[3] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

[4] Kevin Lewis, Jason Kaufman, Marco Gonzalez, Andreas Wimmer, and Nicholas Christakis. Tastes, ties, and time: A new social network dataset using facebook. com. *Social Networks*, 30(4):330–342, 2008.

[5] Eric S Lander and Robert A Weinberg. Journey to the center of biology. *Science*, 287(5459):1777–1782, 2000.

[6] Juan M Vaquerizas, Sarah K Kummerfeld, Sarah A Teichmann, and Nicholas M Luscombe. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10(4):252–263, 2009.

[7] M.B. Gerstein, A. Kundaje, M. Hariharan, S.G. Landt, K.K. Yan, C. Cheng, X.J. Mu, E. Khurana, J. Rozowsky, R. Alexander, et al. Architecture of the human regulatory network derived from encode data. *Nature*, 489(7414):91–100, 2012.

[8] Arthur L. Shaffer, Ryan M. Young, and Louis M. Staudt. Pathogenesis of human b cell lymphomas. *Annual Review of Immunology*, 30(1):565–610, 2012. PMID: 22224767.

[9] Adam A Margolin, Kai Wang, Wei Keat Lim, Manjunath Kustagi, Ilya Nemenman, and Andrea Califano. Reverse engineering cellular networks. *Nature Protocols*, 1(2):662–671, 2006.

[10] Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. Large-scale mapping and validation of transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, 5(1):e8, 01 2007.

[11] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[12] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

[13] Ioannis Xenarios, Danny W. Rice, Lukasz Salwinski, Marisa K. Baron, Edward M. Marcotte, and David Eisenberg. Dip: the database of interacting proteins. *Nucleic Acids Research*, 28(1):289–291, 2000.

[14] JO Armitage. A clinical evaluation of the international lymphoma study group classification of non-hodgkins lymphoma. *Blood*, 89(11):3909–3918, 1997.

[15] Staudt lab data repository.

[16] Patrick O Brown and David Botstein. Exploring the new world of the genome with dna microarrays. *Nature genetics*, 21:33–37, 1999.

[17] Ash A Alizadeh, Michael B Eisen, R Eric Davis, Chi Ma, Izidore S Lossos, Andreas Rosenwald, Jennifer C Boldrick, Hajeer Sabet, Truc Tran, Xin Yu, et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.

[18] Andreas Rosenwald, George Wright, Wing C Chan, Joseph M Connors, Elias Campo, Richard I Fisher, Randy D Gascoyne, H Konrad Muller-Hermelink, Erlend B Smeland, Jena M Giltnane, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine*, 346(25):1937–1947, 2002.

[19] R Eric Davis, Vu N Ngo, Georg Lenz, Pavel Tolar, Ryan M Young, Paul B Romesser, Holger Kohlhammer, Laurence Lamy, Hong Zhao, Yandan Yang, et al. Chronic active b-cell-receptor signalling in diffuse large b-cell lymphoma. *Nature*, 463(7277):88–92, 2010.

[20] Vu N Ngo, Ryan M Young, Roland Schmitz, Sameer Jhavar, Wenming Xiao, Kian-Huat Lim, Holger Kohlhammer, Weihong Xu, Yandan Yang, Hong Zhao, et al. Oncogenically active myd88 mutations in human lymphoma. *Nature*, 470(7332):115–119, 2010.

[21] Lloyd T Lam, George Wright, R Eric Davis, Georg Lenz, Pedro Farinha, Lenny Dang, John W Chan, Andreas Rosenwald, Randy D Gascoyne, and Louis M Staudt. Cooperative signaling through the signal transducer and activator of transcription 3 and nuclear factor-κb pathways in subtypes of diffuse large b-cell lymphoma. *Blood*, 111(7):3701–3713, 2008.

[22] Arthur L Shaffer, NC Tolga Emre, Paul B Romesser, and Louis M Staudt. Irf4: Immunity. malignancy! therapy? *Clinical Cancer Research*, 15(9):2954–2961, 2009.

[23] Yibin Yang, Arthur L Shaffer, NC Emre, Michele Ceribelli, Meili Zhang, George Wright, Wenming Xiao, John Powell, John Platig, Holger Kohlhammer, et al. Exploiting synthetic lethality for the therapy of abc diffuse large b cell lymphoma. *Cancer cell*, 21(6):723–737, 2012.

[24] Michael B Eisen, Paul T Spellman, Patrick O Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.

[25] Hidde De Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of computational biology*, 9(1):67–103, 2002.

[26] Katia Basso, Adam A Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. Reverse engineering of regulatory networks in human b cells. *Nature genetics*, 37(4):382–390, 2005.

[27] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

[28] Diego di Bernardo, Michael J Thompson, Timothy S Gardner, Sarah E Chobot, Erin L Eastwood, Andrew P Wojtovich, Sean J Elliott, Scott E Schaus, and James J Collins. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nature biotechnology*, 23(3):377–383, 2005.

[29] A.A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R.D. Favera, and A. Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(Suppl 1):S7, 2006.

[30] David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.

[31] Maria Stella Carro, Wei Keat Lim, Mariano Javier Alvarez, Robert J Bollo, Xudong Zhao, Evan Y Snyder, Erik P Sulman, Sandrine L Anne, Fiona Doetsch, Howard Colman, et al. The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, 463(7279):318–325, 2009.

[32] Arthur L Shaffer, George Wright, Liming Yang, John Powell, Vu Ngo, Laurence Lamy, Lloyd T Lam, R Eric Davis, and Louis M Staudt. A library of gene expression signatures to illuminate normal and pathological lymphoid biology. *Immunological reviews*, 210(1):67–85, 2006.

[33] DA Williams. 394: The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, pages 949–952, 1975.

[34] DA Griffiths. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics*, pages 637–648, 1973.

[35] RA Fisher. Questions and answers# 14. *The American Statistician*, 2(5):30–31, 1948.

[36] Roland Schmitz, Ryan M Young, Michele Ceribelli, Sameer Jhavar, Wenming Xiao, Meili Zhang, George Wright, Arthur L Shaffer, Daniel J Hodson, Eric Buras, et al. Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature*, 490(7418):116–120, 2012.

[37] Elke Glasmacher, Smita Agrawal, Abraham B Chang, Theresa L Murphy, Wenwen Zeng, Bryan Vander Lugt, Aly A Khan, Maria Ciofani, Chauncey J Spooner, Sascha Rutz, et al. A genomic regulatory element that directs assembly and function of immune-specific ap-1–irf complexes. *Science*, 338(6109):975–980, 2012.

[38] S.H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.

[39] Y.Y. Liu, J.J. Slotine, and A.L. Barabási. Controllability of complex networks. *Nature*, 473(7346):167–173, 2011.

[40] E.E. Schadt, S.H. Friend, and D.A. Shaywitz. A network view of disease and compound screening. *Nature Reviews Drug Discovery*, 8(4):286–295, 2009.

[41] J.-P. Onnela, J. Saramki, J. Hyvnen, G. Szab, D. Lazer, K. Kaski, J. Kertsz, and A.-L. Barabsi. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

[42] Haiyuan Yu, Pascal Braun, Muhammed A Yıldırım, Irma Lemmens, Kavitha Venkatesan, Julie Sahalie, Tomoko Hirozane-Kishikawa, Fana Gebreab, Na Li, Nicolas Simonis, et al. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, 2008.

[43] Wai Lim Ku, Geet Duggal, Yuan Li, Michelle Girvan, and Edward Ott. Interpreting patterns of gene expression: Signatures of coregulation, the data processing inequality, and triplet motifs. *PLoS ONE*, 7(2):e31969, 02 2012.

[44] Roger Guimer and Marta Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073–22078, 2009.

[45] K. Basso, A.A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano. Reverse engineering of regulatory networks in human b cells. *Nature genetics*, 37(4):382–390, 2005.

[46] T.R. Lezon, J.R. Banavar, M. Cieplak, A. Maritan, and N.V. Fedoroff. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proceedings of the National Academy of Sciences*, 103(50):19033–19038, 2006.

[47] M Madan Babu, Nicholas M Luscombe, L Aravind, Mark Gerstein, and Sarah A Teichmann. Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology*, 14(3):283 – 291, 2004.

[48] Stanley Wasserman and Katherine Faust. *Social network analysis: methods and applications*. Cambridge University Press, Cambridge; New York, 1994.

[49] C. MacCluer. The many proofs and applications of perron's theorem. *SIAM Review*, 42(3):487–498, 2000.

[50] Juan G. Restrepo, Edward Ott, and Brian R. Hunt. Characterizing the dynamical importance of network nodes and links. *Phys. Rev. Lett.*, 97:094102, Sep 2006.

[51] J.G. Restrepo, E. Ott, and B.R. Hunt. Emergence of synchronization in complex networks of interacting dynamical systems. *Physica D: Nonlinear Phenomena*, 224(1):114–122, 2006.

[52] Andrew Pomerance, Edward Ott, Michelle Girvan, and Wolfgang Losert. The effect of network topology on the stability of discrete state models of genetic control. *Proceedings of the National Academy of Sciences*, 106(20):8209–8214, 2009.

[53] P. Erdős and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl*, 5:17–61, 1960.

[54] A.L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[55] M.S. Carro, W.K. Lim, M.J. Alvarez, R.J. Bollo, X. Zhao, E.Y. Snyder, E.P. Sulman, S.L. Anne, F. Doetsch, H. Colman, et al. The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, 463(7279):318–325, 2009.

[56] Rka Albert, Hawoong Jeong, and Albert-Lszl Barabsi. The internet's achilles' heel: Error and attack tolerance of complex networks. *Nature*, 406:200–0, 2000.

[57] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science Signalling*, 298(5594):824, 2002.