NETWORK-CENTRIC DATA MINING FOR MEDICAL APPLICATIONS

A Dissertation

Submitted to the Graduate School
of the University of Notre Dame
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

by

Darcy A. Davis,

________________________________________
Nitesh V. Chawla, Director

Graduate Program in Computer Science and Engineering
Notre Dame, Indiana
April 2012
NETWORK-CENTRIC DATA MINING FOR MEDICAL APPLICATIONS

Abstract

by

Darcy A. Davis

Faced with unsustainable costs and enormous amounts of under-utilized data, health care needs more efficient practices, research, and tools to harness the benefits of data. These methods create a feedback loop where computational tools guide and facilitate research, leading to improved biological knowledge and clinical standards, which will in turn generate better data. In order to facilitate the necessary changes, better tools are needed for assessing risk and optimizing treatments, which further require better understanding of disease interdependencies, genetic influence, and translation into a patient’s future. This dissertation explores network-centric data mining approaches for benefit in multiple stages of this feedback loop: from better understanding of disease mechanisms to development of novel clinical tools for personalized and prospective medicine. Applications include predicting personalized patient disease risk based on medical history, optimizing NICU nursing schedules to reduce negative effects, and predicting novel disease-gene interactions.
FIGURES

3.1 Example: Converting an ICD-9-CM code to a list of Disease Ontology codes. The ICD code *cerebral thrombosis with cerebral infarction* (434.01) and its hierarchical parents (434.0 and 434) are mapped to directly to DO terms using mappings provided by the DO disease descriptions. This direct mapping produces a starter list of DO codes \{12815, 12751, 10127\}. This list is then expanded to include ancestors by following all *is-a* links until root nodes are reached. The final list for ICD 434.01 includes 12 DO codes, as shown above.

3.2 Global network properties. (A) Degree distributions and (B) clustering spectrums of the phenotypic (PDN) and genetic (GDN) disease networks. The PDN has higher average degree and clustering coefficient due to very high edge density. Interestingly, the degree distribution of the GDN generally decreasing while the PDN is more uniform, indicating that many diseases are co-morbid with a large number of other diseases, often with few or no underlying shared genes.

3.3 The Phenotypic Disease Network. The phenotypic disease network (PDN) is constructed based on clinical history of 700,000 patients. Each node represents a unique disease, and two nodes are connected if the diseases co-morbid significantly more than randomly expected according to population prevalence. Black edges indicate hierarchically related diseases (*is-a* relationships). The accompanying table displays the most relevant Disease Ontology codes associated with each cluster. Purity corresponds to the percent of member nodes which are accurately described by the DO term, and completeness indicates the percentage of descendants of the DO term which belong to the cluster.

3.4 The Genetic Disease Network. The genetic disease network (GDN) is constructed on the same disease nodes present in the PDN, but edges instead indicate that the disease pair shares a significant number of gene associations. Black edges indicate hierarchically related diseases (*is-a* relationships). The accompanying table displays the most relevant Disease Ontology codes associated with each cluster. Purity corresponds to the percent of member nodes which are accurately described by the DO term, and completeness indicates the percentage of descendants of the DO term which belong to the cluster.
3.5 **The Multi-Relational Disease Network.** This network is created by overlaying the phenotypic (PDN) and genetic (GDN) networks, which contain the same disease nodes. Blue edges indicate phenotypic links, red edges are genetic, green edges are both genetic and phenotypic, and black edge are is-a relationships. The two-tone nodes indicate original cluster membership in the GDN (inner circle) and PDN (outer circle). Regions where multiple nodes share the same color pattern correspond to groups of diseases which cluster together in both the PDN and the GDN. These overlaps are common and in some cases quite large, such as the teal-and-green cluster containing the heart diseases. Still, none of the overlaps fully contain a PDN or GDN cluster. The overlapping regions are listed in the accompanying table, along with the most relevant Disease Ontology codes associated with the cluster.  

3.6 **Genetic vs. phenotypic mutual information.** Each data point represents a disease pair which is linked in both the PDN and the GDN. The plot illustrates the correlation between the mutual information edges weights in each respective network. There is some upward trend but the effect is far from linear. In aggregate, the values have a Pearson correlation of .473, a weak-to-moderate positive correlation.  

3.7 **Example: calculating the probabilistic weights for MRLP.** This toy example demonstrates how to calculate the probability of a given edge type closing a partial triad structure based on triad census counts. The left side shows the subset of the triad census which conforms to the given constraints. On the right is the partial structure for which a probability is being calculated. When calculating probability, MRLP assumes that the observed structure is completely correct except for the potential absence of the prediction target. Thus, the middle two patterns do not affect the probability. Since no phenotypic was observed between nodes A and C and phenotypic is not the target type, we assume that no phenotypic edge is present between A and C and the middle patterns are not relevant.  

3.8 **Link prediction performance.** (A) Receiver operating curves (ROC) and (B) precision-recall curves for the multi-relational link predictor (MRLP) and three traditional neighborhood-based link prediction methods: common neighbors, Jaccard coefficient, and the Adamic/Adar measure. MRLP is the best method with respect to area under the receiver operating curve (AUROC). The precision-recall curve, which is less biased, shows that MRLP is most accurate with the highest ranked predictions, but is not always optimal for lower prediction thresholds.
3.9 **Link predictor performance by individual disease.** Area under the receiver operating curve (AUROC) comparison of link predictor performance for each unique disease. The experiments were hold-one-out, where all genetic associations of the testing disease were removed. The x axis shows the performance of Adamic/Adar on the phenotypic data only, and the y axis is the performance using the MRLP on the multi-relational network. Each point which falls above the diagonal indicates that multi-relational evidence improved link prediction performance for the corresponding disease.

4.1 **Degree distributions of each homogenous projection of the disease-gene network.** (A) Disease-gene associations (B) Phenotypic comorbidity (C) Protein-protein interactions (D) DNA sequence similarity (E) Positive expression correlation (F) Negative expression correlation.

4.2 **An example of varying relative evidence between edge types.** Each series denotes the probability that a node pair \((d, g)\) has a disease-gene association \((G)\) with respect to the number of 2-hop paths from \(d\) to \(g\) conforming to a given pattern.

4.3 **Example: Construction of CW-MRLP feature vectors.** This toy example demonstrates how feature vectors are constructed for each disease-gene pair. For each pair, we find all unique 2-hop paths in the network between the gene and disease and record the pattern of interaction types. For example, disease 77 and gene MSH2 are associated, and MSH2 has a protein-protein interaction with BLM. Thus, there is a path with the pattern \((G, \text{gene, I})\) between 77 and BLM, passing through the intermediate gene MSH2. MSH2 and BLM also have correlated expression, forming a second path with pattern \((G, \text{gene, E+})\) from 77 to BLM through MSH2. These paths overlap, so they can either be counted as a pattern with compound edge types. In both cases, the number of unique paths of each pattern form the feature vector.

4.4 **Example: producing multiple AUROC values from a single ranking.** The network on the left is a toy heterogenous network with 5 nodes and 3 edge types. Solid lines represent training edges and dotted lines indicate holdout edges for evaluation. This example assumes that all node pairs were ranked in descending order with respect to prediction scores produced by some algorithm. The ranking is fabricated and for demonstration purposes only. The testing classes are determined based on the network. Dashes indicate training edges, which are already known and do not impact performance. Class 1 contains the holdout edges, while class 0 comprises edges which are neither known nor held out. The AUROC can be calculated for each class column, as provided below the table. Each edge type produces a different value for the same ranking.
4.5 **Lift in AUROC and AUPR attributable to each edge type.**
The lifts yielded by each edge type $x$ are in a separate bar graph. In
the graph for type $x$, each bar indicates the lift in CW-MRLP perfor-
ance yielded by adding the type $x$ dimension to a base network.
Each base network is a combination of homogeneous dimensions from
the disease-gene network. The x-axis indicates the set of dimensions
comprising each base network. Thus, each bar graph shows the lift
trends of an edge type $x$ across all combinations of homogeneous
dimensions of the disease gene network. Blue bars were calculated
based on AUROC performance and yellow bars were constructed from
AUPR. A lift of 1 is neutral, indicating no change in AUC. Bars in
the region above 1 indicate that adding dimension $x$ increased AUC
and bars below 1 indicate decreased AUC. This figure is continued
on the following page. 61

4.5 **Continued** 62

5.1 **Construction of input and evaluation vectors for the Medicare testing patients.** The original patient records are provided by
hospital visit. All splits between two time-consecutive visits form a
valid input/evaluation vector pair. We evaluate on all possible splits. 73

5.2 **A high-level overview of the CARE system.** 75

5.3 **Recall trends with respect to known testing patient data.**
(A) Performance with respect to the number of visits known by
CARE about the testing patient. (B) Performance with respect to
the total number of unique diseases known about the patient. (C)
Performance with respect to the length of time in days between the
patient’s last training visit and first testing visit. 87

5.4 **Calculating disease predictability: a toy example.** (A) A
feature vector is constructed for each disease $d$. Each patient is an
instance with two features: the rank of $d$ for the patient and a class
indicating whether the patient later develops $d$. (B) Precision and
recall are calculated for each rank threshold. (C) The AUPR is used
as a measure of the predictability of disease $d$. 90

5.5 **Comparison of diagnoses for patients who benefit from ho-
ogeneous vs heterogeneous training sets.** Comparison of past
and future diagnosis age distributions for patients in the age cate-
gory 20-35, compared to the baseline distribution of diagnoses for all
patients in the age category. (A) Averaged age distributions for pa-
tients in the 20-35 age category for whom performance is better with
age-homogeneous training sets. (B) Averaged age distributions for
patients in the 20-35 age category for whom performance is better
with age-heterogeneous training sets. 96
5.6 **Year-by-year performance of CARE.** (A) Recall of diagnoses occurring within each future year, categorized by the number of years of previous medical history. (B) Increased recall of CARE over the baseline, year-by-year. ......................................................... 98

6.1 **Example patient handoff network and calculation of mean repeat caregiver interval (MRCI).** (A) Each network represents the nursing care team of a single infant over the entire NICU stay. Each unique nurse (RN) participating in the care team is represented as a node in the network, and directed edges indicate the sequence in which responsibility for the patient is transferred between nurses. (B) Traversing the network in order starting from the first caregiver (RN A), 5 handoffs occur before a caregiver is repeated (RN C). Thus, the first repeat caregiver interval (RCI 1) is of length 5. The second RCI begins with the last nurse in RCI 2 (RN C), and 3 handoffs occur before another repeat. There are no additional handoffs, so this network has a total of 2 RCIs. The mean repeat caregiver interval (MRCI) is 4, the average of RCI 1 and RCI 2. .................. 106

6.2 **Nursing handoff networks and associated measures for 3 NICU patients.** Patient (A) was cared for by 20 nurses over a stay of 21 nursing shifts, indicating that only one nurse cared for the infant more than once. Patient (B) was hospitalized for 36 nursing shifts and cared for by 26 nurses, with some long repeat caregiver intervals. Finally, patient (C) was cared for by 19 nurses during 42 nursing shifts, with more frequent repeat caregivers. Based on the observed measurements, we hypothesize that (A) and (B) are less likely to result in effective, efficient communication than network (C). 108

6.3 **Nursing team size.** The number of nurses involved in a patient’s care team as a function of the length of stay. .......................... 109

6.4 **Percentage of newcomer nurses.** The proportion of shifts provided by newcomer nurses relative to the length of stay. .............. 110

6.5 **Team centralization and size.** The relationships between gestational age category, NICU nursing team centralization index, and team size. The charted values are medians for each measure, with error bars representing interquartile ranges. ..................... 111

6.6 **Nursing dimension problem scores.** The relationships between MRCI values and rates of problem responses for each of the 9 questions contained in the nursing dimension of the Picker NICU Family Satisfaction Survey. Item 10, denoted as the RN problem score, is the combined count of problem scores from the 9 nursing dimension questions. Patients with MRCI values that exceeded the median where considered to have high values. .................. 113
<table>
<thead>
<tr>
<th>TABLES</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 DO CODES OMITTED FROM THE DISEASE-GENE NETWORK</td>
<td>22</td>
</tr>
<tr>
<td>4.1 BASIC STATISTICS FOR EACH HOMOGENEOUS PROJECTION OF THE DISEASE-GENE NETWORK</td>
<td>50</td>
</tr>
<tr>
<td>4.2 PROBABILITY OF OVERLAP BETWEEN EDGE TYPES</td>
<td>51</td>
</tr>
<tr>
<td>4.3 COMPARISON OF LINK PREDICTION METHODS (AUROC)</td>
<td>58</td>
</tr>
<tr>
<td>5.1 SAMPLE PATIENT RECORD IN THE MEDICARE DATASET</td>
<td>72</td>
</tr>
<tr>
<td>5.2 THE 20 MOST PREVALENT DISEASES</td>
<td>74</td>
</tr>
<tr>
<td>5.3 ICARE PERFORMANCE ON THE MEDICARE AND REGIONAL DATASETS</td>
<td>79</td>
</tr>
<tr>
<td>5.4 ICARE PERFORMANCE ON DEMOGRAPHIC PARTITIONS OF THE REGIONAL DATASET</td>
<td>80</td>
</tr>
<tr>
<td>5.5 EVALUATION OF PERFORMANCE ICARE AND TIME-SENSITIVE ICARE COMPARED WITH THE BASELINE RANKING</td>
<td>84</td>
</tr>
<tr>
<td>5.7 THE MOST PREDICTABLE AND LEAST PREDICTABLE ICD-9-CM DISEASES CODES</td>
<td>92</td>
</tr>
<tr>
<td>6.1 CHARACTERISTICS OF THE STUDY POPULATION</td>
<td>103</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

There are many people who have been instrumental to the realization of this thesis. First, I thank my advisor, Nitesh Chawla, for providing the opportunity to research problems that are interesting and important to me, and for offering guidance along the way. I also would like to thank my dissertation committee members: Prof. Scott Emrich, Prof. Predrag Radivojac, and Prof. Zoltán Toroczkai. Their questions and suggestions have added value to this document. My gratitude also goes out to the many collaborators who have been involved in these projects, in ways both large and small. I especially thank Predrag Radivojac, Dr. James Gray, and Ryan Lichtenwalter for invaluable contributions.

I am particularly indebted to the YSU Foundation, the Arthur J. Schmitt Foundation, and the Eck Center for Global Health for their support of my education. These groups have contributed more than just financial support, and I am grateful for many unique opportunities they have made possible. I also acknowledge CRA-W for organizing valuable peer networking opportunities for women in computer science.

Finally, I would like to thank the friends and family who have been continuously supportive. The members of the DIAL lab group are both friends and colleagues. Special thanks to the denizens of Nieuwland 384L; our group discussions provided many insights, ideas, and solutions that may have otherwise gone unexplored.
“All of life and human relations have become so incomprehensibly complex that, when you think about it, it becomes terrifying and your heart stands still.” Anton Chekhov. *In the Cart* (1897).

If Russian author and physician Anton Chekhov was overwhelmed by the complexity of life in 1897, imagine how he might feel about the medical practice in modern times. He practiced medicine in the early years of the progressive era of medicine, when diseases were still just phenotypes and preventative concepts like nutrition, exercise, hygiene, and public sanitation were radical new ideas [35]. Today, a disease is the product of a massive system of molecular interactions, influenced both by genetics and environmental factors. Furthermore, the biomedical structures underlying disease overlap and have complex effects on one another. This complex system or network view of disease is overwhelming, yes, but it also serves as an amazing resource for understanding and improving human life. Researchers in the emerging fields of bioinformatics and computational medicine are looking for answers in this complexity to keep hearts pumping, rather than making them stand still.

Despite the massive influx of data created by rapid advances in genomic technologies and increasing collection of clinical data, we still have a very incomplete picture of disease at the “systems” level [66]. These rich sources of data have enormous potential for increased understanding of disease mechanism and better healthcare,
but the size, complexity, and biases of the data also present many challenges. There is a strong need for scalable computational tools that can discover patterns without discounting the statistical complexity of heterogenous data or falling prey to the noise it introduces.

This dissertation explores multiple medical applications using network-centric data mining. We use this term to describe the intersection between the rich relational structures of network science and the rigorous inference tools of data mining. Disease data is particularly well suited to a network representation; the structure emerges naturally since biological phenotypes are products of molecular interaction [111]. On the other hand, adequately describing a disease within the constrained feature vector traditionally associated with data mining is very difficult. Listing clinically reported traits like organ system, symptoms, and patient statistics has some value, but provides little or no information about the underlying biology. This information may be useful for treating the symptoms, but not for eliminating the cause.

Network representations do have the drawback of being less studied than standard data representations like feature vectors. Many of the existing approaches for inference in networks are computationally expensive or intractable on large networks, or less statistically rigorous than methods from the data mining literature. For some of the applications in this work, we first build a network, and then turn it into a feature vector of structural properties. In our experiments, this strategy has proven very effective at harnessing the best of both domains.

The overarching goal of this work is to demonstrate that network-centric data mining is a suitable and effective approach to medical applications, both biological and clinical. On our way to this objective, we describe adjustments of existing tools and introduce new methods for dealing with the unique challenges in medical and biomedical domains.
1.1 Contributions

Within the broader domains of network-centric data mining and challenging medical applications, this dissertation makes the following specific contributions:

- A network exploration of the patterns of interplay between gene-disease associations and patient comorbidity.

- Multiple novel methods for multi-relational link prediction, both supervised and unsupervised. These have been applied to the task of gene-disease candidate detection, with promising results.

- An in-depth study of the added predictive value of diverse biological relationships for the disease-gene candidate detection task. The study is performed on a highly heterogeneous network containing disease comorbidity, known disease-gene associations, the human protein interaction network, protein sequence similarity, and genetic expression correlation.

- A novel time-sensitive extension to our previous work on CARE [27], a recommendation system which generates individual disease risk assessments. This extension makes the system more practical for long-term, diverse data such as public health records.

- Many additional CARE experiments including validation on a demographically diverse dataset, analyses of performance trends, and various studies that are of practical or clinical interest. The focused studies explore disease predictability, chronic diseases, age progression, and the effect of time span on CARE performance.

- A network study of the relationship between neonatal intensive care unit (NICU) nursing team organization and quality of care. This includes a novel topological measure which best captures the correlation between the nursing schedule and family satisfaction. In this study, network-centric data mining is harnessed for direct insight into a real clinical application.

1.2 Organization

The remainder of this dissertation is organized as follows. Chapter 2 is an introduction to basic data mining and network terminology, concepts, and relevant methods. Chapter 3 is centered around a combined genetic and phenotypic disease-disease network, which is used to explore the interplay between disease comorbidity and shared genetic associations. It also begins a discussion on link prediction in multi-relational networks by introducing a novel unsupervised method. Chapter 4
continues the link prediction theme on a highly heterogeneous disease-gene network constructed from five different data types. The multi-relational link predictor is adapted into a supervised method and used to evaluate the predictive value of each relationship type for disease-gene candidate detection. Chapter 5 presents new extensions and focused studies on CARE, a recommendation system for personalized disease risk assessments developed in earlier work. Chapter 6 exhibits an application of network-centric data mining to a clinical dataset from a neonatal intensive care unit (NICU). Specifically, it describes a study that examines the correlation between topological properties of the nursing schedule, modeled as a patient handoff network, and family satisfaction with care. Chapter 7 summarizes our contributions and results, then closes with interesting avenues for future work.
CHAPTER 2

NETWORK CONCEPTS

2.1 Network Basics

The term network is defined as a group of objects with some kind of connections or relationships among themselves. Networks are often represented using graphs. A graph is a mathematical structure \( G = (V, E) \) where \( V \) is a set of vertices and \( E \) a set of edges which connect some vertex pairs. The edges in a graph may be directed (used to represent to-from relationship) or undirected (a reciprocal relationship). Each vertex has a mathematical property called the degree, usually denoted as \( k \), which is defined as the total number of edges connecting to the vertex. In directed graphs, this number can be separated into the in-degree and the out-degree. The neighbors \( N_v \) of a vertex \( v \in V \) is the set of all vertices which share an edge with \( v \). It follows that \( |N_v| = k \). A graph can be weighted by assigning a numeric value to each edge [135]. In a labeled graph, each node is marked with unique identifier. An empty graph contains no edges, while a complete graph, or clique, includes an edge for all vertex pairs. A graph is called bipartite if the vertices can be partitioned into two subsets such that every edge in the graph is defined between one node from each subset. There is a large body of terminology describing other classes of graphs, most of which are outside the scope of this work. In network studies, the term node is often used in place of vertex, and the term link may be used interchangeably with edge.
The difference between graph theory and network science is subtle. Graph theory is the theoretical study of various abstract graph problems. Usually, network science refers to the study of phenomena [67], observed relationship patterns in the real world. Networks are usually complex, an umbrella term for structures which are neither regular, i.e. governed by a concise deterministic formula, nor generated by a random process [12, 36, 134, 8, 102].

Many interesting real-world systems form complex networks with multiple distinct types of inter-related objects and relationships. Networks with multiple node types are called multi-modal, while networks with multiple edge types are referred to as multi-dimensional [133]. Networks in this type can also be called multi-relational [127] or heterogeneous networks [50], umbrella terms which encapsulate structures which are multi-modal, multi-relational, or both.

2.2 Network Properties

Networks provide a very rich and expressive representation of the complex systems in the real world, but meaningful interpretation is a challenging problem. In this section, we describe measures commonly used to quantify the properties of a network. All of the definitions are for undirected networks.

2.2.1 Size and Density

The starting point for quantifying the network is always size, represented by the number of nodes $n$ and the number of edges $m$ present in the network. From these values, network density can be calculated as $\frac{2m}{n(n-1)}$. Conceptually, density is the percentage of all possible edges which were observed on the node set. Density can also be used as a local property for any subnetwork. Even these very simple metrics can be important for applications such as disease transmission [81], the spread of influence [20], and resource management. Also, network size and density are neces-
sary to interpret some more complicated properties due to correlation between the measures.

### 2.2.2 Property Distributions

Many different metrics have been proposed to measure the network topology. The *degree distribution* is the probability distribution of the node degrees. This metric has been heavily studied, especially in the context of *scale-free* networks [8, 51]. In scale-free networks, the degree distribution approximates a *power-law* of the form \( P(k) = k^{-\alpha} \) for each degree \( k \), where \( \alpha \) is a constant scaling exponent [5]. Power-law distributions have a distinctive long-tail shape, but show a linear trend when plotted on a log-log scale. The scale-freeness of various real networks has been a hotly debated topic in the literature. Guidelines and statistical tools for determining if a network has a power-law distribution are provided in [21].

The *clustering coefficient* \( C_i \) of a vertex \( i \) is calculated by \( C_i = \frac{2|\{e_{j,k}\}|}{k_i(k_i-1)} \) such that \( \{e_{j,k}\} \) is the set of all edges satisfying \( v_j, v_k \in N_i, e_{j,k} \in E \). This calculation is essentially the density measure applied only to the set \( N_i \). The *average clustering coefficient* \( C = \frac{1}{n} \sum_{i=1}^{n} C_i \) is the sum over all node values. A more informative global property is the *clustering spectrum*, which is the probability distribution over all values of \( k \) of \( C(k) = \sum_{i:k_i=k} C_i \).

### 2.2.3 Path Properties

Another group of network characteristics are based on path lengths. A *path* is a sequence of vertices such that any two consecutive vertices in the sequence are connected. The *shortest path* between vertices \( i \) and \( j \) corresponds to the shortest path that begins with \( i \) and ends with \( j \). Shortest paths are often not unique. The *All-pairs shortest paths* algorithm finds the shortest path length for all vertex pairs.

The network *diameter* is the longest shortest path length in the network. The
characteristic path length, also called the mean average path length is the mean of the all-pairs shortest paths.

2.2.4 Centralization Measures

There are various centrality measures which attempt to characterize the relative importance of each vertex in a network. Each metric makes a different assumption about the characteristics of important vertices. We describe the three most widely used centrality measures.

The degree centrality of a vertex $i$, defined as $C_D(i) = k_i$, is equivalent to the degree of $i$. Thus, degree centrality assumes that high degree vertices are most important. Closeness centrality attempts to quantify the closeness of a vertex $i$ to all other vertices in the network [107]. Formally, $C_C(i) = \frac{1}{\sum_{j \in V \setminus i} dist(i, j)}$ where $dist(i, j)$ is some measure of the distance from $i$ to $j$. Traditionally, the distance is the shortest path length, but other variants exist [91]. Finally, betweenness centrality [37] considers vertices important if they often belong to shortest paths. Formally, $C_B(i) = \sum_{s \neq t \neq i \in V} \frac{\theta_{st}(i)}{\theta_{st}}$ where $\theta_{st}$ is the number of shortest paths from $s$ to $t$ and $\theta_{st}(i)$ is the number of these shortest paths which include $i$.

2.3 Link Prediction

In network science, the link prediction task can be broadly generalized as follows: Given disjoint source node $s$ and target node $t$, predict if the node pair has a relationship, or in the case of dynamic interactions, will form one in the near future [75]. For many real world scenarios, link prediction can be applied to anticipate future behavior or to identify probable relationships that are difficult or expensive to observe directly. In social networks, link prediction can be used to predict relationships that will form, uncover relationships that probably exist but have not been observed, or even to assist individuals in forming new connections [64]. In
biomedicine, where exhaustive, reliable experimentation is usually not viable, link prediction techniques such as disease-gene candidate detection are valuable for navigating incomplete data, as well as guiding lab resources toward the most probable interactions.

This section details the link prediction methods used throughout this dissertation and briefly mentions other approaches from the broader literature. Also, we discuss some early work related to multi-relational link prediction, a topic which will be of central importance in Chapters 3-4.

2.3.1 Unsupervised Link Prediction

There are many existing approaches to link prediction for standard networks with homogeneous edges, formulated for various link formation hypotheses. A survey of these methods is provided in [75]. While some methods also incorporate node or edge features [4, 92], this work is primarily concerned with topological link predictors. Topological methods rely only on the structure of the network to draw conclusions. The common topological link prediction methods can be divided into two types: neighborhood methods and path methods. The neighborhood methods are typically limited only to connections among the immediate neighbors of the source and target nodes, while path methods allow more global influences.

Neighborhood Methods

Many traditional link prediction scores are derived from the immediate node neighborhoods. The *preferential attachment* [9, 88] link prediction score for a node pair is the product of their degrees. *Common neighbors* [89] is another simple method which counts the common neighbors of $s$ and $t$, which is the equivalent to the number of paths of length 2 between the nodes. It can also be expressed as the size of the intersection between the neighbor sets of $s$ and $t$, $|N_s \cap N_t|$. *Jaccard’s*
Coefficient [110] is the size of the intersection of their neighbors divided by the size of the union of their neighbors,

\[
\frac{|N_s \cap N_t|}{|N_s \cup N_t|} \tag{2.1}
\]

Another variation of common neighbors is the Adamic/Adar measure [1], which weights the impact of neighbor nodes inversely with respect to their total number of connections. Specifically,

\[
score(s, t) = \sum_{n \in N_s \cap N_t} \frac{1}{\log(|N_n|)} \tag{2.2}
\]

This inverse frequency approach is based on the assumption that rare relationships are more specific and have more impact on similarity.

Path Methods

A second class of link prediction methods are calculated based on paths between nodes. The PageRank algorithm of Google fame, first introduced in the academic sphere in [16], represents the significance of a node in a network based on the significance of other nodes that link to it. If we assume that linking to nodes that are important is desirable, an assumption implicit in preferential attachment prediction, then the PageRank of the target node represents a useful metric. For our experiments, we perform the original, unoptimized PageRank calculation iteratively, checking for convergence by calculating the Pearson correlation coefficient \( r \) between the vectors of PageRank scores from successive iterations.

Rooted PageRank [75] is another link predictor derived from the original PageRank in which prediction outputs correspond to the probability of visiting the target node in the prediction during a random walk from the source. A parameter \( \alpha \), the probability of restarting the walk at the source, allows the walker to avoid getting trapped in directed networks or dense areas. We use \( \alpha = 0.15 \). Again, prediction
scores are determined after the walks converge. Especially with low to moderate values of $\alpha$, this may take many walk steps. In addition to the parameter, the rate of convergence depends on the size and local density of the network.

The PropFlow predictor introduced in [76] is a path-based predictor that models the link prediction score as being propagated radially outward from the source. Starting from the source node with a score of 1, all neighboring nodes are given an equal share of the score (in the unweighted case), or $1/|N_s|$. The scores continue outward, summing together for nodes which are reached by multiple paths.

The Katz score [62] is the sum of all paths from $s$ to $t$, with the weight of each path exponentially dampened with respect to length to place higher weight on short paths.

2.3.2 Bipartite Link Prediction

While preferential attachment and the path methods apply naturally to bipartite networks, the common neighbor methods are triangle-based and require modification. One possibility to extend the common neighbor methods to the bipartite case, proposed by Huang et al. [58], is to replace $N_t$ with $\bigcap_{n \in N_t} N_n$, the set of neighbors of $t$’s neighbors. This formulation is not necessarily symmetric, i.e. in some cases $\text{score}(s, t) \neq \text{score}(t, s)$. In our work, we instead use symmetric formulations based on paths of length 3 between $s$ and $t$.

First, we note that for nodes $s$ and $t$ in a single-mode network, each common neighbor $n$ belongs to a unique path $(s, n, t)$ of length two from $s$ to $t$. In a bipartite network, nodes are always connected by paths of odd length. Thus, to extend common neighbors link prediction to bipartite networks, we simply count the number of unique paths $(s, n1, n2, t)$ of length three from $s$ to $t$. Similarly, Jaccard’s coefficient is calculated as the number of unique paths of length three from $s$ to $t$.
divided by the total number of unique paths of length three starting at either $s$ or $t$. For Adamic/Adar, the log term is replaced with

$$\log(|N_{n1} \cup N_{n2}|)$$

which, in bipartite networks, is equivalent to

$$\log(|N_{n1}| + |N_{n2}|)$$

### 2.3.3 Supervised Link Prediction

A strong argument promoting a supervised approach to link prediction has been presented by previous work [76]. All of the unsupervised link prediction methods can only perform well in networks for which the link formation mechanism conforms to the *a priori* scoring function. A well-designed classification framework is not domain-specific, can support multiple decision boundaries, and can more flexibly combine the information provided by individual topological features. The link prediction problem is also extremely imbalanced, beyond the bounds of most problems studied by the imbalance community.

When performing studies on supervised link prediction, some researchers have coped with imbalance by heavily undersampling the minority class in both the training and the testing set [4, 132]. However, evaluation results on a modified testing distribution cannot be used to extrapolate the performance on the real-world distribution [76]. Luckily, there are other well established approaches for combatting imbalance in a supervised setting without modifying the testing distribution. In [76], Lichtenwalter *et al.* combine the unsupervised link prediction metrics as features in a classification scheme, a combination they refer to as high performance link prediction (HPLP). Their work also contributes a general supervised link prediction framework which controls for variance and imbalance. The link prediction literature also included many matrix factorization approaches [109, 108, 65, 3, 47, 70].
which are beyond the scope of this dissertation. Generally, these methods tend to be prohibitively resource intensive and better suited to very small networks.

2.3.4 Multi-Relational Link Prediction

Extending the link prediction problem to multi-relational network structures is straightforward: Given disjoint source node $s$, target node $t$, and edge type $x$, predict whether a link of type $x$ exists or will form between $s$ and $t$. However, generating good predictions is not nearly so simple. The link prediction methods described in Sections 2.3.1-2.3.2 are not directly applicable to multi-relational data, although there are two naive solutions. Link prediction can be performed on homogeneous projections of the network; that is, edges of type $x$ are predicted using only links of the same type. The second option is to treat all nodes and edges types indistinguishably. Neither of these approaches is actually multi-relational, since node and edge types have no influence on the results. In fact, the second approach is forced to make the same predictions for all edge types, an obvious disadvantage. Nonetheless, these approaches have been used extensively to reduce naturally multi-relational data to existing network methods.

A small but growing body of recent research is introducing new methods designed to capture some of the diverse evidence in multi-relational structures. Some approaches have been very application specific [139]. Others extend latent feature models to the multi-relational case. In [82], a single set of latent features is generated for each node, but the features are weighted differently for each relationship type. The authors present good experimental results on three small multi-relational networks (14, 90, and 234 nodes, respectively). However, this approach is computationally expensive and highly sensitive to initialization values, which may limit scalability. In [70], the authors briefly discuss a similar extension, but do not provide
any experimental results on multi-relational data.

2.4 Collaborative Filtering

Collaborative filtering is a data mining technique designed to predict a user’s opinion about an item or service based on the known preferences of a large group of users [125]. Most applications assume that the input is partial preference information for each user. That is, the user’s opinion or rating is known about a few items, but usually unknown for a strong majority of the item set. The basic principle behind collaborative filtering is that users who have similar taste on some items are likely to have similar taste on other items. Most collaborative filtering methods are not inherently limited to recommending products and services. In the case of binary preferences (e.g. like or did not like), collaborative filtering is equivalent to bipartite link prediction [58]. For applications with a range of preference values, collaborative filtering can be formulated as a bipartite link prediction problem on weighted networks, with the more challenging goal of predicting missing edge values rather than presence. It follows that collaborative filtering can be applied to any problem which can be modeled with this structure.

The first automated collaborative filtering systems were GroupLens [69, 104] and Ringo [112], which recommended internet news articles and music, respectively. These systems are part of the larger class of memory-based algorithms, which make predictions using the entire user database. This is typically accomplished by calculating a weight of similarity between the active user and all others, and the active user’s opinion is determined by the weighted average of the others’ opinions. In many cases, only a limited number of ‘nearest neighbors’ are included in the calculation. The most common similarity metrics are the Pearson correlation coefficient [104] and vector similarity [110, 13]. Memory-based algorithms are simple, easily
updated, and generally produce good results. These advantages come at the cost of high resource consumption, since the entire database must be retained and used. While correlation is usually cited as the superior method, other results have shown vector similarity to perform equally well or better [46].

The second widely-used class of collaborative filtering algorithms are model-based, where predictions are generated by a model of user preferences which was preconstructed on the user database. The model-based algorithms are faster and more scalable, in general. However, model building tends to be expensive, leading to inflexibility for introducing new data. The quality of predictions for model-based methods widely vary [118]. Well known model-based methods include Bayesian clustering or models [13], Personality Diagnosis [96], Singular Value Decomposition [42, 95], and the Aspect model [55, 54]. There are also many content-based recommender systems, which take advantage of user and item features. The networks used in our work do not have node content, so these methods are not relevant. Many of the mentioned algorithms and other techniques can be found in surveys by Adomavicius [2] and Su [125].
CHAPTER 3

THE DISEASE-DISEASE NETWORK

The saying “treat the disease, not the symptoms” is widespread, a cliche for eliminating or repairing the root of a problem rather than mitigating the negative effects. It is taken for granted that prevention is the best course of action. It is ironic, then, that many of today’s best “disease treatments” are actually symptom suppressors, and the disease roots, i.e. biological mechanisms, are not fully understood. The clever phrase doesn’t capture the challenge of recognizing and defining the root causes of a problem, or of finding an effective treatment that doesn’t produce even worse symptoms. In medicine, both of these tasks involve combing through an immensely complicated web of molecular processes. Rapidly evolving technologies are continuously improving the resolution and reliability of this network. However, the network is very large, and each lab experiment costs time and money. Furthermore, researcher are frequently discovering new dimensions of importance. This chapter describes an exploration of a subset of this biological network to better understand the relationship between gene-disease association and phenotypic comorbidity of diseases.

3.1 Motivation

Many diseases do not occur in isolation. Diseases with similar genetic, environmental, and lifestyle risk factors may be co-morbid in patients, or the disease
products themselves may be risk factors for additional conditions. Also, many serious chronic diseases, such as cancer and diabetes, are complex diseases influenced by a combination of environment and epistasis between many genes [10, 34, 111]. In this way, diseases may share many distinct types of relationships with varying levels of impact for important problems such as patient risk or drug efficacy. Thus, a singular view of dependencies among diseases is not sufficient. Rather, disease mechanisms form a complex system. The underlying goal is to combine all available information and develop the most complete models of interaction between these many factors, simultaneously using information widely applicable and patient specific.

Schadt [111] suggests that diseases can be seen as emergent from a complex network of underlying molecular activity influenced by genes and environment. Indeed, complex networks are a natural way of representing any data with complicated dependency relationships. Unfortunately, most network studies and standard tools are insufficient for the task, limited to treating all relationship types equally or separate analysis of each type. Both of these approaches represent a loss of information. In this study, we use patient medical histories (phenotype data) and previously discovered disease-gene associations to construct, analyze, and compare disease-disease networks. We then take a novel approach to studying interplay between patients, diseases, and genes by merging the heterogeneous data into a multi-relational network and analyzing the structure of interaction between shared genes and clinical comorbidity. Finally, we demonstrate how the multi-relational structure can be applied to enhance the link prediction task of determining good targets for further gene association research.
3.2 Related Work

Both gene-based [41] and patient-based [52] disease-disease networks, constructed similarly to ours, have been previously studied. These separate studies explore different questions, while our approach is to compare and combine the networks and take the composite view. In [94], Park et al. begin exploring relationships between the network links, showing that genetic association is correlated with comorbidity and thus justifying integrated study. However, they do not take advantage of the network structure, and there are still many questions to be addressed for useful inference between the networks. Also, as our networks will show, diseases show far more co-morbidities than genetic links to other diseases, so direct inference based on shared genetic association only applies to a limited subset of co-morbidities. Park et al. acknowledge that many disease pairs share genes but are not co-morbid, and we will further show that there are far more disease co-morbidities without significant gene overlap. Our explicit integration of the networks facilitates inference based on a neighborhood of interactions, providing a richer pool of data than pairwise correlation.

Many other studies have explored integrating diverse evidence to answer biological questions, using various types of data [31, 84, 103, 129, 128]. We have already mentioned some of the limitations of simple correlation studies, particularly with respect to inference tasks. Another approach which has been used is classification using diverse evidence, such as the work on predicting gene-disease associations performed by Radivojac et al. in [103]. Classification has proven to be a good tool for many tasks, but we claim that network-based inference has certain differences that may be advantageous for biological data. Most currently available data, particularly on the molecular level, is incomplete, biased, and noisy, which corresponds to a great deal of missing or unreliable data. Classification methods must impute
values for missing data to create positive and negative profiles for each decision class, which can hurt performance [106]. Most complex networks methods, including the link prediction method we will introduce, apply naturally to whatever partial information is known.

3.3 Data

We use a compilation of known disease-gene associations extracted from OMIM, Swiss-Prot [7], and HPRD [97]. The diseases are classified by Disease Ontology (DO) codes and the gene names are based on the HUGO Gene Nomenclature [100]. The Disease Ontology [93] project is intended to develop a controlled medical vocabulary to unify diverse medical languages and ontologies such as UMLS, ICD, and SNOMED. It is implemented as a directed acyclic graph indicating the hierarchical structure of the disease terms. For more information about the data, see [103].

A second dataset consists of real patient medical histories collected from a group of 77 physicians within a regional health system. This includes data for the last 12 years, from 1997 to 2009, with a total of 5.5 million visits for approximately 700,000 patients. Each data record is a single visit represented by an anonymized patient ID and a primary diagnosis code, as defined by the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM\(^1\)).

ICD-9-CM codes are 3-5 digits and have a hierarchical structure, wherein each 5-digit code is a subset of a 4-digit code, which is further a subset of a 3-digit code. In general, these groups are not based on comorbidity; they are often comprised of specific forms or complications of the same disease or injury. The 3-digit codes can be further separated into large categories based on a major disease type (e.g.

\(^1\)The International Statistical Classification of Diseases and Related Health Problems (ICD) provides codes to classify diseases and a wide variety of signs, symptoms, abnormal findings, social circumstances, and external causes of injury or disease. It is published by the World Health Organization.
neoplasms} or organ system {e.g. diseases of the circulatory system}. For example, the ICD-9 code of 434 corresponds to occlusion of cerebral arteries. The specific version 434.0 corresponds to cerebral thrombosis, which may occur without cerebral infarction (434.00) or with cerebral infarction (434.01). In some cases, such as the previous example, each 5-digit code is distinct from the others. Other 5-digit codes may be substitutes for one another, either by design or due to lack of clarity in the code definition or medical situation. For example, diabetes with unspecified complication (250.09) is a substitute for other diabetic complications (250.1-250.8). For all codes, the 3 or 4-digit parent code is also a correct substitution, with some loss of detail.

For consistency with the first dataset, the ICD-9-CM codes in the clinical data were converted to Disease Ontology codes. The DO term descriptions includes mappings to equivalent ICD-9-CM codes for many terms. The matches may not be exact. The conversion process results in a list of DO codes for each single ICD-9-CM code, and a DO code may apply to multiple ICD-9-CM codes as well. For an ICD code icd, the list begins with all DO codes which are directly mapped to icd or icd’s parent codes. Each DO code in this starter list is then expanded to include all of its direct ancestor codes. This can be done by following all is-a links (provided by the ontology) until root nodes are reached. An example of this process for ICD-9-CM code 434.01 is shown in Figure 3.1. Some terms, both ICD and DO, have no mappings provided by the DO, and were excluded from the study. The DO is a work in progress, so the mappings may contain errors or omissions. Using parent codes in the mapping process helps to reduce omissions. Any potential errors discovered during the mapping process were reported to the DO project leaders, but we did not manually modify the mappings. However, some overly generic DO codes were omitted from the study. These are listed in Table 3.1.
Figure 3.1. **Example: Converting an ICD-9-CM code to a list of Disease Ontology codes.** The ICD code *cerebral thrombosis with cerebral infarction* (434.01) and its hierarchical parents (434.0 and 434) are mapped to directly to DO terms using mappings provided by the DO disease descriptions. This direct mapping produces a starter list of DO codes \{12815, 12751, 10127\}. This list is then expanded to include ancestors by following all *is-a* links until root nodes are reached. The final list for ICD 434.01 includes 12 DO codes, as shown above.
TABLE 3.1

DO CODES OMITTED FROM THE DISEASE-GENE NETWORK

<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>acellular anatomical structure disease</td>
</tr>
<tr>
<td>2</td>
<td>disease of anatomical structure</td>
</tr>
<tr>
<td>3</td>
<td>disease of environmental origin</td>
</tr>
<tr>
<td>4</td>
<td>disease</td>
</tr>
<tr>
<td>5</td>
<td>disease of material anatomical entity</td>
</tr>
<tr>
<td>6</td>
<td>biological macromolecule</td>
</tr>
<tr>
<td>7</td>
<td>disease of anatomical entity</td>
</tr>
<tr>
<td>8</td>
<td>disease of physical anatomical entity</td>
</tr>
<tr>
<td>9</td>
<td>disease of body substance</td>
</tr>
<tr>
<td>10</td>
<td>disease of body</td>
</tr>
<tr>
<td>11</td>
<td>disease of male body</td>
</tr>
<tr>
<td>12</td>
<td>disease of female body</td>
</tr>
<tr>
<td>63</td>
<td>temp holding</td>
</tr>
</tbody>
</table>

3.4 Network Construction

We constructed a phenotypic disease network (PDN) [52] in which nodes are diseases and edges indicate comorbidity of the diseases. Comorbidity can be broadly defined as co-occurrence in the same patients significantly more than chance. The comorbidity was measured using the real patient data described in 3.3. Edges are included between disease pairs for which the co-occurrence (joint probability) is significantly greater than the random expectation based on population prevalence of the diseases (product of marginal probabilities). Statistical significance is determined by a one-tailed two proportion z-test with 95% confidence. Additionally, diseases are required to have a minimum co-occurrence in 2 patients to avoid noise from lone rare events. Diseases with no significant relationships are omitted. Our phenotypic disease network consists of 437 unique diseases nodes and 40,579 comor-
bidity relationships, creating a very dense network.

We also constructed a genetic disease network (GDN) [41] from gene-disease associations reported in previous studies. Nodes are unique diseases, which are connected when the diseases share a significant number of genetic associations. Similar to the PDN, disease pairs have an edge if they share significantly more gene associations than randomly expected based on the generality of the diseases. We approximated the generality with the marginal probability of the disease being associated with a random gene from the dataset. Again, significance was decided by a two proportion z-test with 95% confidence. The genetic disease network has 399 nodes connected with 7817 significant genetic links.

For the methods in this paper, we primarily utilized unweighted networks. However, we found a weighting scheme to be useful for some observations. We weighted the edges using a mutual information metric which quantifies how much greater the edge relationship is with respect to chance. The mutual information weight \( w(d_1, d_2) \) between two diseases \( d_1 \) and \( d_2 \) is defined as

\[
w(d_1, d_2) = \log \left( \frac{p(d_1, d_2)}{p(d_1)p(d_2)} \right)
\]

(3.1)

where the numerator is the observed co-occurrence (joint probability) and the denominator is the random expectation of co-occurrence (product of marginal probabilities). Co-occurrence refers to the number of shared patients in the PDN, or the number of shared genes in the GDN. This weighting scheme is used to avoid bias based on disease prevalence.

Diseases considered for inclusion in the networks were limited to those which appeared in both datasets; that is, diseases which are associated with at least one gene and occur in at least one patient. However, we do not necessarily require significant relationships in both. The overlap of the network is 399 nodes; all diseases in the GDN also had at least one significant comorbidity. However, the PDN contains
38 additional diseases that have significant co-morbidities, have some known gene associations, but are not sufficiently genetically similar to any other diseases.

In both networks, the diseases are classified by Disease Ontology (DO) codes, which have a hierarchical structure. The structure is arranged such that a code may be a subset of other codes at many levels of generality, creating long chains of ‘is-a’ relationships. For example, toxic pneumonitis is a pneumonia is a non-neoplastic lung disorder. Obviously, the is-a relationship is fundamentally different from other edges in the networks and should be treated as such. These links are essential to the structure of the network, so they were included but not weighted.

3.5 Network Analysis and Comparison

For each network, we calculated the degree distribution and spectrum of clustering coefficients, which are shown in Figure 3.2. The extremely high density of the PDN suggests that diseases generally have more co-morbidities than genetic associations. Thus, it is unsurprising that the phenotypic network has higher average degree and clustering coefficient. More interesting, however, is the remarkable difference in degree distribution. While the degree distribution in the genetic network is generally a decreasing function, the phenotypic degree distribution is more uniform. Neither of these networks have a power-law degree distribution [21]. Since these networks mostly contain the same nodes, this difference indicates that many conditions are highly co-morbid despite few or no shared genes. For example, migraines do not have significant genetic link to any other disease, but are co-morbid with more than 200 conditions. Note that a lack of genetic edges does not mean the conditions are not genetic, but rather that their set of associated genes is not highly similar to the sets associated with other diseases.

Both networks were computationally clustered using Walktrap, a hierarchical
Figure 3.2. **Global network properties.** (A) Degree distributions and (B) clustering spectrums of the phenotypic (PDN) and genetic (GDN) disease networks. The PDN has higher average degree and clustering coefficient due to very high edge density. Interestingly, the degree distribution of the GDN generally decreasing while the PDN is more uniform, indicating that many diseases are co-morbid with a large number of other diseases, often with few or no underlying shared genes.

clustering tool for networks based on the intuition that random walks are often trapped within dense network regions corresponding to clusters. Algorithm details are provided in [99]; we use the implementation provided by the authors with the default parameters. The reported clusters correspond to the partition with the highest modularity [90]. The clustered networks are provided in Figures 3.3-3.4, along with limited descriptions of the clusters. Due to the high density, the visual representations are limited to strongest 10% of edges in each network according to the mutual information weights. All of the nodes remain present. The reduced networks are for visual clarity only; the clusters and associated descriptions correspond to the full network. Dynamic, fully labeled representations of the networks are available at [http://www.nd.edu/~dial/plosone/diseasenetworks/](http://www.nd.edu/~dial/plosone/diseasenetworks/) in Cytoscape format, an open source tool for visualizing and analyzing networks [119].
The central theme of each cluster is described by finding the DO term(s) that are most pure or complete within the cluster. Each node has a DO code which is further associated with a hierarchy of more general terms. Purity and completeness of each network cluster are calculated with respect to each DO term associated with the cluster members. We define \textit{purity} as the percent of all cluster members which belong to a given term and \textit{completeness} as the percent of all nodes belonging to the term that also belong to the cluster. Intuitively, the purity indicates the homogeneity of the cluster, while completeness measures the uniqueness relative to other clusters.

We now provide a detailed example of this calculation for a cluster of disease terms \{coccidiosis, malaria, arthropod diseases, helminthiasis, parasitic intestinal diseases\}, which corresponds to the five blue nodes on the upper right side of Figure 3.3. These are all children of the DO term \textit{infectious diseases}, so the purity with respect to that term is \(\frac{5}{5} = 1\). However, these are only 5 of the 21 \textit{infectious diseases} in the network, so the completeness is \(\frac{5}{21} = 0.238\). Similarly, 4 of the 5 terms are \textit{parasitic diseases} (purity: \(\frac{4}{5} = 0.8\)) and only one other parasitic disease is included elsewhere in the network (completeness: \(\frac{4}{5} = 0.8\)).

Purity and completeness can be determined for all cluster-DO term pairs. The terms provided in Figures 3.3-3.4 are those that best describe each cluster. While the determination of the best terms was subjective, most clusters had clear winners. Otherwise, the clusters have been marked as “mixed”.
Figure 3.3. The Phenotypic Disease Network. The phenotypic disease network (PDN) is constructed based on clinical history of 700,000 patients. Each node represents a unique disease, and two nodes are connected if the diseases co-morbid significantly more than randomly expected according to population prevalence. Black edges indicate hierarchically related diseases (is-a relationships). The accompanying table displays the most relevant Disease Ontology codes associated with each cluster. Purity corresponds to the percent of member nodes which are accurately described by the DO term, and completeness indicates the percentage of descendants of the DO term which belong to the cluster.
The Genetic Disease Network. The genetic disease network (GDN) is constructed on the same disease nodes present in the PDN, but edges instead indicate that the disease pair shares a significant number of gene associations. Black edges indicate hierarchically related diseases (is-a relationships). The accompanying table displays the most relevant Disease Ontology codes associated with each cluster. Purity corresponds to the percent of member nodes which are accurately described by the DO term, and completeness indicates the percentage of descendants of the DO term which belong to the cluster.
The PDN was partitioned into 10 clusters, four of which are of acceptable quality due to size, purity, and completeness. By acceptable, we simply mean that the cluster is non-trivial and has a reasonably specific universal theme. These clusters can be roughly classified as neuromuscular and neuro-degenerative disease, sensation disorders, malignant neoplasms, and female reproductive system disorders. Four of the remaining clusters are tiny groups of related conditions, consisting almost entirely of is-a edges. The final two clusters are enormous “catch-all” clusters of mixed conditions, accounting for 56.98% of all disease nodes in combination. One of these large clusters contains the famously complex relationships between heart disease, diabetes, strokes, obesity, and many other chronic diseases believed to be lifestyle influenced. The other large cluster contains most of the congenital deformities. Both of these clusters additionally contain many other disease categories which were not easily separable, forming a chaotic picture of intra-connections across disease families and organ systems.

The GDN was separated into 11 clusters, with seven high quality clusters. Again, neoplasms and nervous system disease form fairly pure clusters. Genetic clusters also form for heart disease, endocrine diseases, and diseases of the hematopoietic system. Similar to the PDN, there are 3 tiny clusters, but only one mixed cluster of moderate size, accounting for about 19.55% of the nodes. Overall, clusters in the GDN are more specific and separated than the PDN, although in both cases there are many conditions which do not form distinct modules.

3.6 Network Integration

The PDN and GDN provide insight into the way genes associate with diseases and the way diseases occur in patients, but little information about the interplay between the two mechanisms. The structures suggest that the two networks ex-
press different information for understanding disease mechanisms, which is not unexpected. There are multiple possible reasons, both biological and artificial, which we can speculate are behind these differences. Two diseases being associated with the same gene might not have a practical effect, especially if the diseases are associated with different loci, alleles, or expression levels, or if the gene multi-functional. This corresponds to a connection in the GDN and none in the PDN. In the other direction, some diseases may be co-morbid only because they are influenced by the same environmental conditions. Finally, both networks are likely to have collection biases, although we expect that the PDN is more complete. However, these differences do not preclude the possibility of important patterns and dependencies between the structures. We combined the individual network information into a multi-relational disease network (MRDN) and probabilistically analyzed local structures to draw more specific conclusions about how genetic influences relate to disease comorbidity.

We previously mentioned that the patient-based and gene-based networks were constructed from the same pool of diseases. While some diseases have significant relationships only in the PDN, the majority of the node sets overlap. Thus, the networks contain many of same disease nodes but with a different pattern of connections and weights. This allows them to be easily overlaid and represented as a single multi-relational network with multiple edge types. The edge type can be thought of as a nominal edge attribute. If both a phenotypic and a genetic edge is present between two diseases, it was treated as a single edge of type ‘both’. The is-a relationship is explicitly treated as a unique edge type. This fundamental disease relationship supersedes any other correlations. Thus, the MRDN has four possible edge types: (G)enetic, (P)henotypic, (B)oth genetic and phenotypic, and (I)s-a. The clusters from the original networks can be retained as separate node attributes.
Figure 3.5. The Multi-Relational Disease Network. This network is created by overlaying the phenotypic (PDN) and genetic (GDN) networks, which contain the same disease nodes. Blue edges indicate phenotypic links, red edges are genetic, green edges are both genetic and phenotypic, and black edge are is-a relationships. The two-tone nodes indicate original cluster membership in the GDN (inner circle) and PDN (outer circle). Regions where multiple nodes share the same color pattern correspond to groups of diseases which cluster together in both the PDN and the GDN. These overlaps are common and in some cases quite large, such as the teal-and-green cluster containing the heart diseases. Still, none of the overlaps fully contain a PDN or GDN cluster. The overlapping regions are listed in the accompanying table, along with the most relevant Disease Ontology codes associated with the cluster.
The integrated MRDN is shown in Figure 3.5. The edge colors represent the relationship type. The two-tone nodes indicate original cluster membership in the GDN (inner circle) and PDN (outer circle). Areas matching the white background color indicate that the node was omitted from one of the networks. Groups of nodes with a matching two-tone pattern are overlaps between clusters found in the separate networks. It is visually apparent that substantial overlap between the cluster results is common. However, the clusters never fully overlap, nor are they contained within each other. In general, phenotypic influence tends to extend far beyond the bounds of genetic similarity. Consistent with Park et al.’s result that sharing genes is correlated with comorbidity [94], we observe that 72% of genetic edges underlie a phenotypic edge. In Figure 3.6, we plot the genetic mutual information versus the phenotypic mutual information for the 4465 disease pairs which have both relationships.

In order to determine if genetic mutual information and phenotypic mutual information are significantly correlated for the disease pairs in our networks, we used a Monte Carlo permutation test. Each permutation was determined by randomly pairing each genetic mutual information value with a phenotypic value using a Fisher-Yates shuffle of each value set, respectively. We generated 1 million permutations, each with a corresponding Pearson correlation value. The correlation values were within the range \([-0.0702, 0.0699]\) with a mean value of \(1 \times 10^{-5}\). The observed Pearson correlation in our networks was 0.473, which falls well outside the range generated from the permutations, which corresponds to probability \(p = 0\) of random observation. We concluded with high confidence that genetic and phenotypic mutual information of the disease pairs is significantly positively correlated. Despite the strong significance, it is a weak-to-moderate correlation in the general sense, i.e. the strongest genetic relationships do not necessarily translate to high comorbidity,
Figure 3.6. **Genetic vs. phenotypic mutual information.** Each data point represents a disease pair which is linked in both the PDN and the GDN. The plot illustrates the correlation between the mutual information edges weights in each respective network. There is some upward trend but the effect is far from linear. In aggregate, the values have a Pearson correlation of .473, a weak-to-moderate positive correlation.

Despite visually different structures, there are definite dependencies between genetic association and comorbidity, but pair-wise correlations are weak indicators on their own. Nonetheless, even weak evidence can be very valuable for inference tasks, particularly in combination with complementary evidence, which is our approach. Furthermore, we suspect that the PDN can be very valuable for inference in regions of the GDN which have been sparsely studied, perhaps due to rarity or low morbidity.

3.7 Multi-relational Local Structure

Construction and manual observation of the multi-relational network has already confirmed significant interplay between the genetic and phenotypic networks. However, there are still many questions about the basic rules and probabilities that
govern these influences, particularly in terms of strength. In addition to furthering biological understanding of disease mechanisms, understanding the probabilistic properties of the network structure will be instrumental to locating additional genetic associations or recognizing the role of genetics in poorly understood co-morbidities.

We approach these global questions through the local substructures, which can provide manageable and interpretable insights into the global structure [131]. For this study, we counted the occurrence of each unique 3-node structure, traditionally called triad census [56, 30] and more recently defined as counting 3-node graphlets [102]. Triad census has been widely used in social network analysis, often for evaluating local structure hypotheses such as transitivity [133]. The triad census trivially extends to multi-relational networks; the only difference is the number of unique structures. While a traditional directed network yields 16 possible structures, our network has 30 unique connected triad patterns of the four edge types and single unlabeled node type. Of course, the hypothesis space becomes increasingly complex with each additional relation. For this work, our use of triad counts is more similar to the context of recent work on graphlet distribution, where substructure counts are used to thoroughly characterize the local structure [101].

Instead of hypothesizing about which structures are important, we wish to probabilistically weight all relationship patterns. The triad census provides the probability of each structure, which further translates to the probability that a partial triad is closed by each edge type. Specifically, for three nodes \((s, n, t)\) and an edge type \(x\), we first count all triads with the same pattern as \((s, n, t)\), then count all triads with the same pattern plus \(x\) added between \(s\) and \(t\). We can determine \(P(x \subset \text{edge}\_type(s, t)|\text{pattern}(s, n, t))\) by dividing the first count by the second count. This probability assumes that the observed pattern is correct except for the
Figure 3.7. **Example: calculating the probabilistic weights for MRLP.** This toy example demonstrates how to calculate the probability of a given edge type closing a partial triad structure based on triad census counts. The left side shows the subset of the triad census which conforms to the given constraints. On the right is the partial structure for which a probability is being calculated. When calculating probability, MRLP assumes that the observed structure is completely correct except for the potential absence of the prediction target. Thus, the middle two patterns do not affect the probability. Since no phenotypic was observed between nodes A and C and phenotypic is not the target type, we assume that no phenotypic edge is present between A and C and the middle patterns are not relevant.

potential absence of type \( x \), which simplifies the calculation. A pictorial example is shown in Figure 3.7.

3.8 Multi-Relational Link Prediction

One of the great challenges for studying biological networks and systems biology in general is the incompleteness and noise of the data. In any large scale molecular context, especially considering phenomena such as epistasis, experimentally exhausting all combinations is not a viable option. Even experimental studies, especially high-throughput methods, may be inconsistent or can result in high false
positive rates, thus requiring many trials or diverse evidence to be reliable. Computational approaches such as disease-gene prioritization are essential for targeting future experiments, directing time and money towards the most likely successes.

In complex networks, finding missing associations is the link prediction problem, which can be broadly generalized as follows: Given two nodes $s$ and $t$ that are not connected by an edge, predict whether the edge actually does exist, or in the case of dynamic interactions, will form in the near future. Usually, this prediction is in the form of a score for each disease pair. The scores are then ranked to determine the nodes pairs that are relatively most likely to have an edge. Many link prediction methods exist for networks; a survey of these methods can be found in [75, 38]. For this work, we focus on unsupervised topological models.

However, most traditional link prediction methods have no direct applicability to multi-relational networks other than treating all edges equally, which can be detrimental to their performance for many reasons. Different link types contain different information by nature, and various combinations introduce different amounts of evidence to the link prediction task. This is particularly troublesome when the link types have very different frequency or distribution, which is clearly the case in our multi-relational disease network. In a sufficiently complicated system, some edge types may be irrelevant or redundant with respect to certain prediction tasks. Even if these barriers could be overcome, treating all edges equally provides no information about the type of link being predicted.

We propose a novel multi-relational link prediction (MRLP) method which addresses all of these issues to predict the location and type of new edges. The most important component of our MRLP method is an appropriate weighting scheme for different edge type combinations. In Section 3.7, we explained how a triad census can be used to place a probability on local substructures, which conveniently trans-
lates to a non-arbitrary, data-justified weighting scheme. To account for frequency disparity, the probabilistic weights are normalized by the marginal probabilities of the edge types involved. Ours is a general algorithm for multi-relational networks, which can also be trivially extended to multiple node types.

Our multi-relational link prediction (MRLP) approach is a weighted extension of the neighborhood methods. Nodes $s$ and $t$ form a partial triad with each common neighbor $n \in N_s \cap N_t$, and each partial triad provides a probabilistic weight based on the triad census. We can simply sum the weights over all neighbors, which is equivalent to weighted common neighbors. Prediction scores are calculated individually for each link type of interest. Formally, the prediction score for edge type $x$ between nodes $s$ and $t$ is

$$\text{score}_x(s, t) = \sum_{n \in N_s \cap N_t} w_n$$

where

$$w_n = \frac{\sigma|P(x) - P(x \subset \text{edge_type}(s, t)|\text{pattern}(s, n, t))|}{P(\text{edge_type}(s, n))P(\text{edge_type}(t, n))}$$

in which pattern$(s, n, t)$ describes the node and edge type pattern of the network path $(s, n, t)$. Also,

$$\sigma = \begin{cases} 1 & P(x \subset \text{edge_type}(s, t)|\text{pattern}(s, n, t)) > P(x) \\ 0 & P(x \subset \text{edge_type}(s, t)|\text{pattern}(s, n, t)) = P(x) \\ -1 & P(x \subset \text{edge_type}(s, t)|\text{pattern}(s, n, t)) < P(x) \end{cases}$$

where the sign is determined by statistical comparison rather than numerical. Statistical significance is determined by a two-tailed two proportion z-test with 99% confidence. As mentioned earlier, the denominator of the weight term is a normalization factor to account for the frequency disparity between edge types. The weighting scheme can suffer from the “zero frequency” or low frequency problem,
which is particularly problematic in networks with a disproportionately large number of objects and relationship types or many overlapping type combinations. Of course, the best solution is a larger sample, but this is often not available in practice. The problem, which is common to other probabilistic models such as Naive Bayes, can be combated with many existing approaches such as smoothing operations. In our experiments, we set $\sigma = 0$ if the target pattern occurs less than 5 times or if type $x$ occurs less than 10 times, which corresponds to frequencies too low for a valid z-test. We assume that due to very low frequency, removing the influence of these patterns does not substantially effect the performance.

Equation 3.2 can be extended to include the inverse frequency principle of the Adamic/Adar measure, since it has been shown to increase predictive performance in many cases. The integration is direct except that the degree of a neighboring node $n$ only counts edges with the same types as the edges connecting $n$ to $s$ or $t$. The prediction score becomes

$$
\text{score}_x(s, t) = \sum_{n \in \mathcal{N}_s \cap \mathcal{N}_t} \frac{1}{w_n} \log \left\{ \begin{array}{ll}
|N_n(t1)| & t1 = t2 \\
|N_n(t1)| + |N_n(t2)| & t1 \neq t2
\end{array} \right. 
$$

(3.5)

where $t1 = \text{edge\_type}(s, n)$, $t2 = \text{edge\_type}(t, n)$, and $|N_n(y)|$ is the number of edges of $n$ with edge type $y$. Unless otherwise noted, MRLP refers to the formulation in Equation 3.5 in our experiments.

3.8.1 Results

We applied our probabilistically weighted MRLP to the multi-relational disease network. For this application, we only generated prediction scores for genetic links, since 'is-a' relationships are known and we assume that the patient data completely represents significant co-morbidities. We compared our performance to traditional
neighborhood-based link prediction methods as applied to the genetic disease network (GDN). The algorithms used are Common Neighbors, Jaccard coefficient, and the Adamic/Adar measure (details in Section 2.3.1). These methods provide a baseline for how well the genetic links in our network can be predicted without the benefit of multi-relational analysis. For all experiments, we use 10-fold cross validation, holding out 10% of the genetic edges for each run. Structure probabilities were calculated on the remaining training network, consisting of all non-held-out edges, and predictions scores were generated for all disease pairs which did not have a genetic edge in the training network. The held out edges form the positive class for testing, while all other pairs are assumed to be negatives. The comparative performance is shown by the receiver operating characteristic (ROC) curves and precision-recall curve [29] in Figure 3.8. The MRLP outperforms the traditional methods with respect to AUROC. The precision-recall curve, which is potentially less biased by the extreme imbalance between actual links versus all possible pairs, shows that MRLP performs particularly well on the top 50 rankings, but is not optimal for all decision boundaries, as one would expect. The drop in precision as recall increases is expected, as it is accompanied by an increase in false positives. When predicting possible genetic links for further investigation, it is also important to have fewer false positives, and thus the operating range might be constrained to top-50 or top-100 rankings (precision at 50 or at 100, for instance).

The MRLP also reaches 100% recall with higher precision than the other methods, which we hypothesize is due to an improved ability to distinguish between edges with very low genetic evidence. These results indicate that phenotypic information can help improve the prediction of genetic links between diseases, even though less than 12% of the phenotypic relationships coincide with an underlying genetic association.
Figure 3.8. **Link prediction performance.** (A) Receiver operating curves (ROC) and (B) precision-recall curves for the multi-relational link predictor (MRLP) and three traditional neighborhood-based link prediction methods: common neighbors, Jaccard coefficient, and the Adamic/Adar measure. MRLP is the best method with respect to area under the receiver operating curve (AUROC). The precision-recall curve, which is less biased, shows that MRLP is most accurate with the highest ranked predictions, but is not always optimal for lower prediction thresholds.
We then applied the MRLP to a more difficult problem: a disease with no known genetic associations, only a phenotypic profile. Such a disease is disconnected in the GDN, and thus cannot be predicted by the baseline algorithms applied to the GDN as in previous experiment. The link predictions can be made based on the PDN, but phenotypic evidence alone is weak. The multi-relational approach provides a connection while allowing the genetic associations of the other diseases in the network to still play a role. Experimentally, we simulated this scenario by holding out all genetic associations for each disease individually, and then using the MRLP to predict the correct locations of the removed associations. Figure 3.9 shows the AUROC achieved for each disease using the MRLP versus Adamic/Adar applied to phenotypes only. Similar trends hold for MRLP versus the other benchmark algorithms, slightly shifted leftward due to lower average performance. The strong majority of diseases fall above the diagonal, indicating that the multi-relational approach improved the predictions for that disease. Using MRLP in this way, the genetic associations were most easily predicted for alopecia, hypothyroidism, and complications of diabetes mellitus. The most poorly predicted were schizophrenia, polymyositis, and frontotemporal dementia.

3.9 Introducing Disease Ontology and Clustering Priors into MRLP

We now describe two additional weighting schemes for MRLP which incorporate information about the Disease Ontology hierarchy or a clustering structure. Neither of the methods in this section showed better performance relative to the probabilistic weighting scheme described in 3.8. Nonetheless, the ideas are novel and interesting, and could be valuable for future work.

For the experiments in this section, we assume that each pair of nodes $s$ and $t$ has a prior prediction score $\text{prior}(s, t)$ independent of the MRLP score. We then
Figure 3.9. **Link predictor performance by individual disease.** Area under the receiver operating curve (AUROC) comparison of link predictor performance for each unique disease. The experiments were hold-one-out, where all genetic associations of the testing disease were removed. The x axis shows the performance of Adamic/Adar on the phenotypic data only, and the y axis is the performance using the MRLP on the multi-relational network. Each point which falls above the diagonal indicates that multi-relational evidence improved link prediction performance for the corresponding disease.
incorporate the prior into the final link prediction score and determine the effect on performance. We try two different methods for incorporating the priors. The first is a multiplicative prior, which is simply

\[ \text{score}(s, t) = \text{prior}(s, t) \times \text{MRLP}_\text{score} \]

The second method is a more fine-grained additive prior, calculated by

\[ \text{score}(s, t) = \sigma \text{prior}(s, t) + \text{MRLP}_\text{score} \]

where \( \sigma \) is a scaling factor which determines the prior’s strength of influence. The additive approach is sensitive to the scale of scores, so we normalize both the priors and the MRLP prediction scores to the range \([0,1]\). To determine the value of \( \sigma \), we use a 10% holdout validation set in addition to the 10% testing set. After choosing \( \sigma \), we reintroduce the validation set to the network before calculating MRLP scores for the test pairs. This avoids performance loss caused by removal of the validation set.

3.9.1 Disease Ontology Priors

The diseases in our phenotypic (PDN) and genetic (GDN) disease networks are all contained within the Disease Ontology (DO), which is implemented as a fully connected directed acyclic graph. Ignoring the direction of edges, there is at least one path in the DO between every pair of diseases. We will refer to the length of the minimum path between two nodes as the ontological distance. In general, the ontological distance provides some information about the relatedness of the disease. For example, all nodes with a distance of 1 have a direct parent-child relationship; one disease is a sub-classification of the other. We hypothesize that node pairs with shorter distance in the DO are more likely to be biologically similar, and thus more likely to share associated genes. It logically follows that the ontological distance
could be valuable for predicting genetic association. We used the inverted minimum
distance between disease pairs as ontological link prediction weights. Specifically,
the weight for a pair of diseases is their ontological distance divided by the DO
diameter.

We found that the DO weights negatively impacted the performance of MRLP
when introduced as prior, for both the multiplicative and additive cases. We weren’t
surprised by this result. The Disease Ontology is valuable for high level observations,
but is not intended for mathematical precision. Also, the minimum distance within
the DO hierarchy is already represented within the networks, and is already directly
used by the MRLP if the distance is 2. Greater distances could be much more
elegantly implemented by simply allowing MRLP to consider larger graphlets, which
is future work for us.

3.9.2 Hierarchical Clustering Priors

In many hierarchical clustering methods, including the Walktrap method that
we used for the disease networks, all nodes begin in separate clusters and clusters
are gradually merged until the network is whole again. For most studies, the goal
is to find the best clustering based on some performance criterion. However, for the
purpose of this work, we are interested in the order in which clusters are merged.
In the hierarchical process we just described, all node pairs are eventually in the
same cluster at some level of the tree. We hypothesize that diseases which reach
the same cluster earlier are more likely to have a link. We define each merging of
two clusters to be a hierarchical level, where the first two nodes are merged at level
1. We further define the hierarchical level where two nodes first belong to the same
cluster as the introduction level. To calculate a clustering link prediction weight for
a disease pair, we use their inverted introduction level. Specifically, the weight for
two diseases is their introduction level divided by the maximum level.

Each disease pair has two clustering priors from the PDN and the GDN, respectively. We introduce them both individually and simultaneously. However, in all experimental cases the priors did not result in performance gains. In this case, we suspect that using clustering information was somewhat redundant, since the principles governing clustering (dense connections between nodes) are similar to those used by MRLP and all of the neighbor based clustering methods.

3.10 Conclusion

Despite many differences in structure between the genetic relationships between diseases and the observed patterns of comorbidity in patients, we observed that there are interesting dependency patterns between the two relationships. Due to these dependencies, we hypothesized that phenotypes could aid inference about the genetic structure between diseases. We developed and applied a novel multi-relational link prediction method to the disease-disease networks, and the results support our hypothesis. Incorporating phenotypic information improved performance over genetic relationships alone for predicting missing genetic links in the network.
4.1 Motivation

We began our discussion of computational disease-gene prediction in Chapter 3, but our work with the disease-disease network has a number of limitations in this context. The most obvious is the network itself, which does not explicitly represent genes. The genetic associations shared by disease pairs were combined and abstracted into a single weighted edge. Thus, the individual value of each gene was lost. Further, the edge threshold throws away weak genetic similarities. Link prediction in this structure can suggest disease pairs that may have a more shared genes, but cannot suggest which genes they are likely to be, beyond naive solutions such as listing all observed associations of both diseases. These predictions are of minimal value for guiding lab experiments.

The scope of the disease-disease network was also very narrow. A wide range of biological features have been implicated for disease-gene identification, including (but not limited to) protein function and interactions, protein sequence, chromosomal position, and gene expression [123]. A good multi-relational model for predicting disease-gene candidates should incorporate as many of these dimensions as possible. A more complicated model also acts as a much stronger testing ground for quantifying the performance and flexibility of multi-relational methods such as MRLP.
Finally, the MRLP method has a number of weaknesses. First, it is bounded by the same *a priori* assumptions about link formation that govern Adamic/Adar. Of course, MRLP is only one of many ways to formulate or extend a topological link predictor for multi-relational data, but however weighted, extended, and otherwise modified, unsupervised link predictors are still domain-specific and inflexible. In heterogeneous information networks, the issue is exacerbated by the fact that multiple relationships may have vastly different formation mechanisms within the same system. For many real problems, such as the disease-gene network described in this paper, it is very unlikely that heterogeneous information types will share the same properties or distribution. In fact, even a single link type may form based on more than one mechanism. Also, the normalization method which controls for imbalance between edge types is quite rudimentary and untested.

In this chapter, we address these limitations by reformulating the disease-disease network as a disease-gene network and adding link types describing protein-protein interactions, protein sequence similarity, and gene expression correlation. We also adapt MRLP for use with a flexible supervised framework which can support multiple decision boundaries and controls for variance and imbalance in a principled manner. With these improved tools, we revisit the problem of disease-gene candidate prediction. In this context, we perform an exhaustive study of the predictive value of each dimension of the disease-gene network, both alone and in combination.

4.2 Data

The data described in 3.3 is also used in this study. Additionally, we use a protein interaction map assembled from physical interactions recorded in HPRD [97], the Online Predicted Human Interaction Database (OPHID) [17], and studies by Rual [105] and Stelzl [122]. Proteins can be easily mapped to the name of their
coding gene in the HUGO Gene Nomenclature. The protein sequences used in this study are consistent with those in [103].

Gene expression profiles were also used, derived from microarray data originally presented by Su et al.[124]. It is now available from the Gene Expression Omnibus, accession GSE1133, and has been used in multiple previous studies [80, 74, 87]. A complete description of the data collection found in [124]. We used the normalization process detailed in [87].

4.3 Network Construction

A multi-relational disease-gene network was constructed, encapsulating all of the data types described in 4.2. Each node in the network is either a disease or a gene, uniquely labeled with the corresponding DO code or HUGO gene name, respectively. Each edge in the network has a label that indicates the type of relationship it represents. For the purpose of this study, we define a network dimension as the set of all edges with a given set of labels. Thus, a homogenous dimension consists of all edges of single edge type, ignoring all other edge types.

Since our application is disease-gene prediction, the node set consists of the diseases and genes present in the gene association data. The phenotypic disease network (PDN) described in 3.4 is preserved in the disease-gene network. The is-a edges imposed by the DO hierarchy (see Section 3.4) are also preserved. The genetic disease network (GDN) is not preserved; disease gene associations are explicitly represented as edges. Similarly, protein-protein interactions form edges connecting their corresponding coding genes. The disease-gene associations and protein interactions are binary, so no inclusion threshold is required.

On the protein sequences, we first measured the sequence similarity for all gene pairs. We used BLAST, a Basic Local Alignment Search Tool, a heuristic algorithm
that aligns sequences based on short seed matches [6] and calculates the statistical significance of the matches between sequences. BLAST reports an *E*-value describing the random expectation of the match score between sequences. A lower E-value corresponds to higher sequence similarity. For values < 0.01, an E-value closely approximates the p-value of observing the match score in random sequences. Additional details about the statistics can be found in [86]. Edges with an E-value less than 0.01 were included in the network.

Edges are also present between gene pairs for Pearson correlation between expression values was > 0.2 or < −0.2. This threshold is unusual is several ways, so more explanation is merited. Many approaches have been suggested for choosing an appropriate threshold in co-expression networks [98], but the right choice really depends on the application goals. The target of our study is identifying disease-genes, so we began by profiling the predictive power of the gene co-expression network for all positive thresholds Θ = 0.05x for x = 0, .., 19. For each threshold, the corresponding co-expression network was combined with the disease-gene association network and the Adamic/Adar score was calculated for each disease-gene pair. Pairs with or without a genetic association were assigned to classes 1 or 0, respectively. The predictive power was then quantified by the AUROC of the scores, sorted from highest to lowest. The winning threshold was 0.2 with AUROC = 0.772. We also found that imposing an upper bound of −0.2 and including only the strongest negative correlations had similar power with AUROC = 0.766. Combined indistinguishably, the positive and negative correlations showed lower performance (0.760), so the co-expression edge labels also indicate the direction of correlation.
TABLE 4.1

BASIC STATISTICS FOR EACH HOMOGENEOUS PROJECTION OF THE DISEASE-GENE NETWORK

<table>
<thead>
<tr>
<th>Edge Type</th>
<th>Abbr</th>
<th>Diseases</th>
<th>Genes</th>
<th>Edges</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene-disease association</td>
<td>G</td>
<td>2184</td>
<td>2000</td>
<td>40746</td>
<td>0.009</td>
</tr>
<tr>
<td>phenotypic comorbidity</td>
<td>P</td>
<td>702</td>
<td>-</td>
<td>74523</td>
<td>0.151</td>
</tr>
<tr>
<td>is-a</td>
<td>F</td>
<td>2184</td>
<td>-</td>
<td>23025</td>
<td>0.005</td>
</tr>
<tr>
<td>protein-protein interaction</td>
<td>I</td>
<td>-</td>
<td>1132</td>
<td>2450</td>
<td>0.002</td>
</tr>
<tr>
<td>sequence similarity</td>
<td>S</td>
<td>-</td>
<td>1346</td>
<td>4548</td>
<td>0.003</td>
</tr>
<tr>
<td>positive expression correlation</td>
<td>E+</td>
<td>-</td>
<td>1665</td>
<td>566963</td>
<td>0.205</td>
</tr>
<tr>
<td>negative expression correlation</td>
<td>E-</td>
<td>-</td>
<td>1665</td>
<td>258970</td>
<td>0.093</td>
</tr>
</tbody>
</table>

4.4 Network Description

Table 4.1 displays basic descriptive statistics for each homogeneous projection of the network. The counts ignore nodes with degree $k = 0$ in the projection. The abbreviations column provides a legend of edge type labels for use in results and figures. The short abbreviations will make compound edge types and triad patterns easy to represent.

Degree distributions for each projection are provided in Figure 4.1. The phenotypic dimension of this network is unchanged from Chapter 3, and shows a chaotic degree distribution. The disease-gene associations, protein interactions, and sequence similarities also have many low degree nodes and few high degree nodes. The degree distributions of the genetic correlation projections are somewhat surprising. Both follow an upward arrow shape with a peak near $k = 200$. The peak at such a high number is due to our very permissive threshold, which causes the extreme density of these networks. The positive correlations show a second peak at the far right of the distribution near $k = 1000$. This indicates that there are very
TABLE 4.2

PROBABILITY OF OVERLAP BETWEEN EDGE TYPES

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>S</th>
<th>E+</th>
<th>E-</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1</td>
<td>0.080</td>
<td>0.395</td>
<td>0.117</td>
</tr>
<tr>
<td>S</td>
<td>0.043</td>
<td>1</td>
<td>0.399</td>
<td>0.088</td>
</tr>
<tr>
<td>E+</td>
<td>0.002</td>
<td>0.003</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E-</td>
<td>0.001</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Numbers represent $P(X|Y)$ where $X$ and $Y$ correspond to the column and row edge types, respectively.

A high number of nodes with many positive expression correlations. The cause of this peak is not currently known, and it does not appear for the negative correlations.

The protein-protein interactions, sequence similarities, and co-expression edges are all defined on gene-gene pairs and can overlap in arbitrary ways. Obviously, a gene pair cannot simultaneously have positive and negative expression correlation. The probability of overlap between these types are presented in Table 4.2. Each value in the table represents $P(X|Y)$ where $X$ and $Y$ correspond to the column and row edge types, respectively. All of observed values are significantly greater ($p < 0.01$) than the random expectation $P(X)$ (equivalent to the density of type $X$, see Table 4.1). However, no network dimension comes close to fully covering another, so all edge types provide some novel connections.

4.5 Classification Weighted MRLP

4.5.1 Support for a Supervised Approach

One of the basic assumptions of the neighborhood based link predictors, including our MRLP method, is that evidence of a given type is linearly additive. That
Figure 4.1. **Degree distributions of each homogenous projection of the disease-gene network.** (A) Disease-gene associations (B) Phenotypic comorbidity (C) Protein-protein interactions (D) DNA sequence similarity (E) Positive expression correlation (F) Negative expression correlation
Figure 4.2. **An example of varying relative evidence between edge types.** Each series denotes the probability that a node pair \((d, g)\) has a disease-gene association \((G)\) with respect to the number of 2-hop paths from \(d\) to \(g\) conforming to a given pattern.

is, each additional common neighbor (or similar unit of evidence) adds a consistent amount to the score, regardless of the total number seen. Even Adamic/Adar, where nodes have different value, is linearly additive with respect to a given degree. In practice for homogeneous networks, it only matters that the real function of likelihood with respect to evidence be increasing, since most evaluation metrics are concerned with ranking rather than the distance between scores. Considering the success of neighborhood based methods, this has been a reasonable assumption for many networks of interest.

In the case of multiple edge types, this assumption falls apart, especially if relatively weak forms of evidence are introduced. An example is shown in Figure 4.2. Each series denotes the probability that a node pair \((s, t)\) has a disease-gene association with respect to the number of common neighbors connected to \(s\) and \(t\) by a given pair of edge types. For simplicity, overlapping edge types are counted separately rather than a compound type. In this example, the distributions are not
linear, are not always increasing, and vary widely between type patterns. It is clear that relative evidence of one type versus another is not a static value, which points to a clear weakness of the MRLP probabilistic weighting scheme.

To address these issues, we learn a classifier on the triad patterns employed by MRLP (see Sections 3.7-3.8). We refer to this approach as classification weighted MRLP (CW-MRLP). This supervised approach allows the classifier to determine the relative value of the features dynamically at each step while building the model, as well as allowing for multiple decision boundaries. CW-MRLP is still limited to the same type of local features used by the neighborhood methods, but has the freedom to learn their value rather than making an a priori assumption.

4.5.2 Constructing the Feature Vectors

Feature vectors are constructed from a given training network. In all of our experiments, the training network is just the disease-gene network with some disease-gene associations held out for evaluation. The holdout scheme varies and will be detailed in Section 4.6. Each instance is a disease gene pair \((d, g)\), and each feature is a count of all 2-hop paths from \(d\) to \(g\) conforming to a given pattern \((\text{edge}_\text{type}, \text{node}_\text{type}, \text{edge}_\text{type})\). Overlapping edges can either be counted as compound types or as separate paths. The better method is domain-dependent and probably best established through experiments. The feature vector is assigned to class 1 if \(d\) and \(g\) are associated, and class 0 otherwise. Any other features defined on the pair \((d, g)\) could be added to the vectors but are beyond the scope of this study. Feature vector construction on a 10-node sample network is demonstrated in Figure 4.3.

For the studies in this chapter, all patterns containing an \(\text{is} - \text{a}\) edge are not included as features because they cause leakage. Leakage is a data mining mistake
Figure 4.3. **Example: Construction of CW-MRLP feature vectors.** This toy example demonstrates how features vectors are constructed for each disease-gene pair. For each pair, we find all unique 2-hop paths in the network between the gene and disease and record the pattern of interaction types. For example, disease 77 and gene MSH2 are associated, and MSH2 has a protein-protein interaction with BLM. Thus, there is a path with the pattern (G, gene, I) between 77 and BLM, passing through the intermediate gene MSH2. MSH2 and BLM also have correlated expression, forming a second path with pattern (G, gene, E+) from 77 to BLM through MSH2. These paths overlap, so they can either be counted (A) separately or (B) as a pattern with compound edge types. In both cases, the number of unique paths of each pattern form the feature vector.
defined as directly introducing information about the class labels into the features [63]. Each disease code is linked to all of the disease-gene associations of its child codes. Thus, a held out disease-gene association can be trivially predicted with 100% confidence if the disease has any child node connecting to the gene. In the other direction, children may not have all of their parents’ connections, but the candidate set is still limited only to the disease-gene associations of their parents. This is a clear source of leakage since some of the parent associations may only be present because of the child node’s association. This effect may be removable, but doing so in a fair way is both difficult and uninteresting.

4.5.3 Classification Framework

The training and testing sets described in the previous section are general, so arbitrary classification model that accepts continuous features and binary class labels can be applied. For our experiments, we use a modified version of the supervised framework described in [76], which was specifically designed to meet the challenges of link prediction. Bagging [14] is employed to reduce variance, but the chief challenge of the problem is class imbalance. To combat this without unfairly modifying the testing distribution, we undersample only the training set constructed within each fold so that the positives, links that actually exist, represent 25% of the data visible to the classifier. Each member of the ensemble for each fold sees all of the positives class instances from that fold and a different selection of negative class instances to achieve the indicated positive class representation. We replace random forests [15] with random subspaces [53] within each bag. For each of the 10 bags, we construct trees in each of 10 subspaces. The switch to random subspaces is motivated by faster run times and lower memory usage, traits that were essential due to the sheer volume of models generated for this study. Also, based on preliminary
experiments, performance is comparable or better than random forests.

4.5.4 Results on Three Dimensions of the Network

Early on in our work with the disease-gene network, CW-MRLP was applied to a three dimension subset of the disease-gene network to evaluate the performance compared to MRLP. Specifically, disease gene associations, comorbidity, and protein-protein interactions were included. Seven common topological link predictors (see Section 2.3) were also evaluated using a 10-fold cross validation holdout theme. We generated performance results for the homogeneous link prediction methods for both of the naive approaches: separately considering homogenous single-dimensional projections or treating edge types equally in a combined network.

In the first case, which we will call the *homogeneous baseline*, the standard link prediction methods were directly applicable, and prediction scores were generated individually for each edge type. The disease-gene association projection is bipartite, so modified versions of the Jaccard coefficient, common neighbors, and Adamic/Adar methods were applied for those three values (see Section 2.3.2).

When treating all edge types as indistinguishable in a single network (the *combined baseline*), only a single set of prediction scores are produced. However, this set produces different performance values when evaluated with respect to each edge type, since the class values (link present or not present) may change. Since held out edges may fall in a position where another edge type is already present in the training network, we produce prediction scores for all pairs in the combined network, regardless of the edge status. For meaningful comparison, we maintain information about the edge types and evaluate each link type separately only on node pairs that did not have a link of that type in the training set. Figure 4.4 demonstrates how separate performance values for each edge type can be produced by a single ranking.
Figure 4.4. **Example: producing multiple AUROC values from a single ranking.** The network on the left is a toy heterogeneous network with 5 nodes and 3 edge types. Solid lines represent training edges and dotted lines indicate holdout edges for evaluation. This example assumes that all node pairs were ranked in descending order with respect to prediction scores produced by some algorithm. The ranking is fabricated and for demonstration purposes only. The testing classes are determined based on the network. Dashes indicate training edges, which are already known and do not impact performance. Class 1 contains the holdout edges, while class 0 comprises edges which are neither known nor held out. The AUROC can be calculated for each class column, as provided below the table. Each edge type produces a different value for the same ranking.

<table>
<thead>
<tr>
<th>Pair</th>
<th>Rank</th>
<th>Testing Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>(B,D)</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>(D,E)</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>(A,B)</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>(C,E)</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>(C,D)</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>(A,D)</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>(B,C)</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>(A,C)</td>
<td>8</td>
<td>-</td>
</tr>
<tr>
<td>(A,E)</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>(B,E)</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

AUROC(■)=0.70)  
AUROC(▲)=0.58  
AUROC(▲▲)=0.86

TABLE 4.3

**COMPARISON OF LINK PREDICTION METHODS (AUROC)**

<table>
<thead>
<tr>
<th></th>
<th>PA</th>
<th>PR</th>
<th>RPR</th>
<th>PF</th>
<th>JC</th>
<th>CN</th>
<th>AA</th>
<th>MRLP</th>
<th>CW-MRLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homogeneous</td>
<td>0.903</td>
<td>0.786</td>
<td>0.933</td>
<td>0.951</td>
<td>0.957</td>
<td>0.951</td>
<td>0.956</td>
<td>0.962</td>
<td>0.971</td>
</tr>
<tr>
<td>Combined</td>
<td>0.829</td>
<td>0.632</td>
<td>0.826</td>
<td>0.886</td>
<td>0.554</td>
<td>0.821</td>
<td>0.857</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The methods applied are preferential attachment (PA), PageRank (PR), rooted PageRank (RPR), PropFlow (PF), Jaccard coefficient (JC), common neighbors (CN), Adamic/Adar (AA), the multi-relational link predictor (MRLP), and classification weighted MRLP (CW-MRLP).
The performance results in terms of AUROC are presented in Table 4.3. The methods applied are preferential attachment (PA), PageRank (PR), rooted PageRank (RPR), PropFlow (PF), Jaccard coefficient (JC), common neighbors (CN), and Adamic/Adar (AA), the multi-relational link predictor (MRLP), and classification weighted MRLP (CW-MRLP). The bold number indicate the best performance. Both MRLP and CW-MRLP effectively used diverse evidence to outperform the homogeneous link predictors in all cases. However, the poor performance results for the combined baseline are a cautionary tale for naively combining multi-relational data. CW-MRLP achieved substantial gains over MRLP, which was consistent with our expectations.

4.6 Experiments

In this section, we describe a large experiment designed to exhaustively quantify the value of each dimension and sub-dimension of the disease-gene network for the disease-gene candidate prediction. A caveat is that the predictive value and relative importance of each type is dependent on the classification method. Nonetheless, the results provide a novel view into the power, redundancy, and synergy of biological data types, both alone and in combination.

4.6.1 The Value of Diverse Evidence

For this experiment, we used 10-fold cross validation, holding out 10% of the disease-gene edges from the training network for each run. The training set contained feature vectors for all disease-gene pairs. The testing set was comprised of all feature vectors belonging to class 0 in the training set, but the feature vectors corresponding to held out associations were converted to class 1. Vectors of class 1 in the training set were ignored since they correspond to already known disease-gene associations. This setup is an interesting case of training on the testing set.
ever, there is no unfair advantage since all of the target instances are labeled with the wrong class. Also, this scenario accurately models the disease gene candidate detection task in the real world. Undiscovered associations are indistinguishable from independent pairs.

We applied our supervised multi-relational link prediction procedure to all combinations of homogeneous dimensions from the disease-gene network. The disease-gene association dimension was included in all sets, since these edges form the only paths from diseases to genes. The performance lifts for each network dimension are shown in Figure 4.6.1, measured in terms of area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPR). For a given edge type $x$ and a subset of dimensions $S$ containing $x$, the lift was calculated as $\Lambda = \frac{\text{AUROC}(S)}{\text{AUROC}(S/x)}$, and similarly for AUPR. Thus, $\Lambda = 1$ is the neutral value indicating equivalent performance. A lift value is calculated for all sets $S : x \subset S$, for a total of 15 lift values per metric per dimension. In Figure 4.6.1, the subsets are partially ordered from least to most included dimensions.

First, we note the differences between AUROC versus AUPR, and explain why we evaluate using both metrics. While ROC and precision-recall curves are constructed from the same points, the area under the curves often differ widely, especially for imbalanced data [29, 40]. While AUROC is the more widely used and understood method, it tends to skew toward high values in imbalanced data despite many false positives. AUPR is heavily biased toward precision at the top of a rank list, which corresponds to strict avoidance of false positives. Thus, some researchers have argued that it provides a more informative representation of an algorithm’s real performance and room for improvement [29]. We consider both methods to be useful measures for different disease gene prediction tasks. For broad searches for candidate pairs among many diseases and genes, such as the one in this experiment,
Figure 4.5. **Lift in AUROC and AUPR attributable to each edge type.**

The lifts yielded by each edge type $x$ are in a separate bar graph. In the graph for type $x$, each bar indicates the lift in CW-MRLP performance yielded by adding the type $x$ dimension to a base network. Each base network is a combination of homogeneous dimensions from the disease-gene network. The x-axis indicates the set of dimensions comprising each base network. Thus, each bar graph shows the lift trends of an edge type $x$ across all combinations of homogeneous dimensions of the disease gene network. Blue bars were calculated based on AUROC performance and yellow bars were constructed from AUPR. A lift of 1 is neutral, indicating no change in AUC. Bars in the region above 1 indicate that adding dimension $x$ increased AUC and bars below 1 indicate decreased AUC. This figure is continued on the following page.
Figure 4.5. Continued
AUPR is undoubtedly a better measure of the utility of a top-down exploration of the ranked predictions. However, it is more realistic for a laboratory researcher to compare performance among a smaller predefined pool of genes. AUROC is a better measure of the likely value of any subset chosen from our comprehensive rank list.

In general, AUROC steadily increased with the number of edge types included in the network. This is a good sign, supporting both the potential of diverse evidence and the ability of the classifier to capture the added value. As more edge types are added, there are diminishing returns for each edge type, a normal and expected effect of redundancy between data sources. AUPR also increases with added dimensions in many cases, but sometimes suffers performance losses when comorbidity or expression relationships are incorporated. The simultaneous gains in AUROC and losses in AUPR for these three dimensions are due to broad coverage combined with weak evidence. These three networks are both denser and noisier than sequence similarity and protein protein interactions, which have more concrete biological meaning. Weak evidence is very valuable for the overall ranking though. The expression correlation dimensions had the most AUROC lift in all cases, while protein interactions and sequence similarity were essential for the highest AUPR values.

The relative value of comorbidity changes depending on the combination. As a lone predictor, these relationships result better AUROC than sequence similarity and protein interactions, but most of the value for both metrics is lost when expression correlation is introduced. We suspect that the information in the phenotypic observations is mostly redundant when any other source is available. However, phenotypic observations like comorbidity are inexpensive and highly complete, which may make them a very valuable resource when knowledge is limited.

Expression correlation shows a particularly interesting pattern of gains and
losses. In all cases where positive or negative correlation show a performance loss, the opposite type is not included. In some dimension subsets, for example (G,P,S), adding positive or negative expression causes performance loss, but adding both results in large gains, greater than the combined value of both types as individual predictors. This is a surprising result; the reason for this effect is still under investigation.

As individual predictors, the ordering of the individual edge dimensions is $E^+ > E^- > P > S > I$. The gene expression correlations, both positive and negative, are far superior to the other dimensions. In fact, positive gene expressions showed better performance than the combination of protein interactions, sequence similarity, and comorbidity. As an individual predictor, comorbidity was superior to sequence similarity and protein interactions. However, this advantage disappears when expression correlation is introduced, while the latter two information types have less redundancy and maintain more of their value. Alone, sequence similarity produces a better ranking than protein interactions, but they have very similar value in combination with any other dimensions.

The results presented in Figure 4.6.1 represent the feature construction case where overlapping edges are treated as independent paths rather than compound types. We ran an identical experiment on feature vectors which include compound types. The resulting performances were overall comparable or slightly worse on average. However, the compound edge types result in the best overall AUPR values for the top 5 highest performing sets, although the differences were fairly small. These results suggest that some overlapping types may have distinct value, but the overall effect is added noise. The noise isn’t surprising, since compound types can add multiple rare features with no value for most pairs.
4.7 Conclusion

The disease-disease network central to the previous chapter was used to examine the relationship between disease comorbidity and genetic association. We broadened our view of the disease domain by expanding our representation into a disease-gene network that explicitly represents disease-gene associations and incorporates protein interactions, protein sequence similarity, and gene expression correlation. We then further explored the task of prediction disease-gene associations.

Our MRLP method, a novel tool for multi-relational networks has shown promising performance relative to baseline methods. In this chapter we discussed reasons why the original formulation MRLP may not be well suited to the challenges of multi-relational data. Some studies have previously suggested that supervised learning is a better approach [76]. We further described why the link formation assumptions of unsupervised methods can be especially detrimental in multi-relational networks, where multiple distributions and complicated decision boundaries are more of a rule than an exception. In response to these issues, we adapted MRLP to fit into a flexible classification framework.

Finally, we apply the new classification weighted multi-relational link predictor (CW-MRLP) exhaustively to combinations of homogeneous dimensions from the disease-gene network. The results suggest that every type of biological relationship in the disease-gene network adds value. Introducing weaker evidence types like comorbidity and expression correlation introduces a trade-off, improving the overall ranking but also increasing noise at high ranks. Phenotypic evidence showed diminishing lift as more dimensions were added to the network, indicating possible redundancy of information with more reliable types. This is consistent with the observations in 3.8.1 that phenotypic evidence is weak and most valuable when other information is scarce. Finally, we found that the value of overlapping edge types
is comparable overall to that of the two edges considered separately. This is good news for computational methods, since independent objects are much easier to work with. However, there is evidence that a subset of the compound types may be useful. This motivates an interesting feature selection problem, which we reserve for future work.
CHAPTER 5

CARE: PREDICTING INDIVIDUAL PATIENT DISEASE RISK BASED ON MEDICAL HISTORY

Another challenge of “treating the disease” is recognizing the presence of the problem before the symptoms escalate beyond control. Research has shown many conditions to have recognizable indicators before onset or preventable risk factors. From these discoveries comes the idea of prospective medicine, aimed at determining and minimizing individual risk, as well as actively addressing conditions at the earliest indication. In theory, these practices reduce the number of conditions needing treatment and improve the effectiveness of necessary interventions.

Chapters 3 and 4 look at computational tools for understanding diseases from a genomic perspective. Indeed, technologies ranging from linkage equilibrium and candidate gene association studies to genome wide associations are filling out an extensive list of diseasegene associations. These studies are offering detailed information on mutations, SNPs, and associated risk estimates for specific disease phenotypes [23]. The underlying hypothesis behind this line of research is that once we catalogue all disease-related mutations, we will be able to predict the susceptibility of each individual to future diseases using various molecular biomarkers, ushering us into an era of predictive medicine. This line of research is very valuable, but clinical relevance is still limited. The feature space for genetic variation is enormous, and most disease associated SNPs or mutations associated with non-Mendelian diseases
offer weak signals which cannot offer stand-alone clinical value.\[78].

Does this mean that prospective approaches to health care will have to wait until the genomic approaches sufficiently mature? Our aim in this chapter is to show that network-centric methods can harness extra value in existing clinical records.

5.1 Motivation

In previous work, we proposed CARE, a Collaborative Assessment and Recommendation Engine, which uses patient medical history to generate a personalized risk profile for a patient \[26, 27\]. CARE is a comprehensive recommendation system that considers the experience of millions of patients to answer the question: What are my disease risks?

CARE uses collaborative filtering, the method is used by companies like Amazon.com to recommend items based on the preferences of other similar users. Collaborative filtering can be framed as a link prediction problem on a network of users and items in which links are used for prediction and all other features are ignored \[57\]. For our application, the users are patients and we predict diseases based on the diagnoses of patients who had a similar medical history. For each patient, the CARE system outputs a ranked list of diseases from the highest risk score to the lowest. This framework was shown to perform well on a large database of Medicare patients, capturing 41% of all future diagnoses in the top 20 ranks in experiments across 20,000 patients. In Section 5.3, we confirm that the CARE framework also performs well for patients of all ages.

One weakness of CARE was a lack of temporal influence on measuring similarity between patients. It did not take the order of or length between disease diagnoses into account when determining similarity among patients. Also, there are situations where it may be more valuable to weight patient similarity only within a certain
time span, such as the progression of cancer. In these cases, we want to put greatest weight on patients who experienced the most similar medical progression relative to the current situation, not just the greatest number of matching diagnoses. In Section 5.4, we describe an extension to CARE that locates and focuses on this “best match” time period. This strategy makes the CARE framework feasible for long-term, diverse data, such as public health records.

The CARE framework was also used to explore some additional studies that are relevant to a deployment situation or interesting from a clinical viewpoint. In Section 5.5, we analyze performance trends dependent on the amount of data known for a testing patient and the length of time between diagnoses. This information provides guidelines for efficient use in a practical setting. We also designed studies to answer very specific questions about the performance of CARE. Section 5.6.1 explores the predictability of individual disease codes. In Section 5.6.2, we measure how well time-sensitive ICARE can distinguish between repeating and non-repeating diagnoses. Section 5.6.3 takes a deeper look at the effect of patient age on predictive performance. In Section 5.6.4 focuses on how far into the future predictions can reliably be made. Like performance trends, these studies also provide guidelines for appropriate use and unexplored avenues for improvement.

5.2 Background and Previous Work

5.2.1 Related Work

The related work for the CARE framework includes the larger body of research on collaborative filtering (Section 2.4), studies from the medical community which further support the need for preventative medicine, and various interdisciplinary efforts which previously led to computer-aided medical prediction systems. While most of these systems are only loosely comparable to CARE, they are representative
of the same goals. We are not aware of any work which is directly comparable to CARE.

Early treatment [25], screening, lifestyle change [59], and other interventions [33, 68] are common themes in modern medical research, where early intervention is shown repeatedly to improve disease outcome and quality of life. Nonetheless, these proactive treatments are far from the norm in our largely reactive health care system. In [120] Snyderman and Williams provide an outstanding overview of the flaws of the current system and potential benefits of a prospective health care system. They suggest that data mining is a “central feature” of prospective health care. Glasgow et. al [39] support the feasibility of the preventive approach. They state that much of the chronic disease burden can be prevented, and further posit that existing management strategies can also be used to advance prevention.

Many proponents of prospective medicine emphasize genomic studies and other breakthrough research in human biology. It is undeniable that genomic research is rapidly advancing [23] and holds great promise for medicine. Unfortunately, applicability to the general public is still very limited [79]. Similarly, in [136] Weston et al. express excitement with advancements in systems biology and proteomics, but acknowledge that we still need to learn how to realistically translate discoveries into health benefits. Also, they recognize that there are still “enormous challenges” to overcome. Though low-tech in comparison, CARE demonstrates that existing data and technology can provide immediate advancement toward prospective medicine.

A wide variety of studies on disease comorbidity, i.e. the simultaneous occurrence of two or more distinct diseases, have shown that multiple risk factors cannot reliably be considered in isolation [121]. Co-occurring factors can have a synergistic effect, leading to unexpectedly high risk [78, 61]. In [130], van den Akker et al. mention that the incidence of comorbid diseases is increasing. They state that statistical
clustering of comorbid diseases was surprisingly strong, even among young subjects. This results implies likely interaction between many of the coinciding diseases.

Many different computer-aided methods have been developed for medical prediction. Most of these systems are designed to make predictions about a single disease or class of diseases. Usually, the predictions are generated from some combination of basic data such as demographic information and physical description with addition condition-specific test results or family history. One well-known system is Apache III [138], a prognostic scoring system for predicting inpatient mortality. Apache uses a combination of acute physiological measurements, age, and chronic health status. A wide variety of systems have been developed for predicting risk of individual diseases or complications, such as specific heart conditions [24], Alzheimer’s disease [77], and cancer [85]. While data mining has been widely used to explore medical problem, collaborative filtering has not been used. An exception is [60], which discusses the use of collaborative filtering into the CHORUS system for efficiently locating relevant radiology information. CHORUS is essentially a text-classification program and is not comparable to our work.

5.2.2 Data Format

The methods described in this chapter use individual patient medical histories in the form of a vector of ICD-9-CM diagnosis codes. More information about the ICD-9-CM hierarchy was previously provided in Section 3.3. The patients are first divided into disjoint training and testing sets. Patients in the training set are represented by their full code vector. For each testing patient, the code vector is split into two subvectors, one for input into the CARE system and one for evaluation. The codes in the vector seen by CARE should have been diagnosed at earlier dates than before those in the evaluation vector. Some of the experiments required additional
TABLE 5.1

SAMPLE PATIENT RECORD IN THE MEDICARE DATASET

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Vector of ICD-9-CM Disease Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>9142409</td>
<td>40291 57420 5301 5533 2780</td>
</tr>
<tr>
<td>9142409</td>
<td>29624 4019 2768 2780</td>
</tr>
<tr>
<td>9142409</td>
<td>2967</td>
</tr>
<tr>
<td>9142409</td>
<td>25090 7906 E9331 20300</td>
</tr>
<tr>
<td>9142409</td>
<td>25090 E9331 20300 4019</td>
</tr>
<tr>
<td>9142409</td>
<td>3101 20300 25001</td>
</tr>
</tbody>
</table>

information about patient age, race, or gender. The demographic information was used only for pre-processing and is not directly used by the CARE methods.

The database used in the previous CARE work comprises the Medicare records of 13,039,018 elderly patients in the United States with a total of 32,341,348 hospital visits. The data was originally compiled from raw claims data for beneficiaries who were at least 65 years old as of January 1993 [19]. Such Medicare records are highly complete and accurate, and they are frequently used for epidemiological and demographic research [72, 83]. A sample patient medical history is shown in Table 5.1; each line represents one hospital visit.

The hospital visits include timestamps and provide a partial time-ordering on the diagnoses. For testing patients, any split between two consecutive visits forms input and evaluation vectors. In previous work and all experiments in this chapter which use the Medicare data, we evaluate on all possible splits. Thus, each patient is tested multiple times with varying amounts of known and hidden data. Figure 5.1 provides a pictorial explanation of this process. All testing patients were required to have at least 5 visits.
Figure 5.1. **Construction of input and evaluation vectors for the Medicare testing patients.** The original patient records are provided by hospital visit. All splits between two time-consecutive visits form a valid input/evaluation vector pair. We evaluate on all possible splits.

In the Medicare database, the number of visits per patient ranges from 1 to 155, with a median of 2. Also, though up to ten diagnosis codes are permitted, the average is only 4.32 per visit, making this dataset very sparse. There are a total of 18,207 unique disease codes expressed in the database. However, only 169 diseases occur at 1% or more in the population (across visits for patients). Table 5.2 shows the 20 most prevalent diseases in our database.

5.2.3 The CARE Framework

In this section, we provide an abridged description of the CARE framework, focusing primarily on details necessary for understanding the extensions and results in the following sections. The system we describe here is the ICARE method in [28], which was consistently far superior to earlier versions of CARE during experimentation. We use the terms CARE and ICARE interchangeably; the original CARE method is obsolete and is not relevant to any results in this chapter.

A high-level preview of the entire CARE framework is provided in Figure 5.2. CARE takes as input a training set and a testing set, both comprised of vectors of ICD-9-CM code describing a patient’s diagnosis history (see Section 5.2.2). Predictions are generated for one testing patient at a time, referred to as the active testing patient. The training patients are clustered into groups if they meet a similarity con-
### TABLE 5.2

#### THE 20 MOST PREVALENT DISEASES

<table>
<thead>
<tr>
<th>Disease</th>
<th>Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>unspecified essential hypertension</td>
<td>33.64%</td>
</tr>
<tr>
<td>coronary atherosclerosis</td>
<td>21.16%</td>
</tr>
<tr>
<td>congestive heart failure</td>
<td>18.16%</td>
</tr>
<tr>
<td>urinary tract infection</td>
<td>16.67%</td>
</tr>
<tr>
<td>chronic airway obstruction</td>
<td>14.69%</td>
</tr>
<tr>
<td>atrial fibrillation</td>
<td>14.03%</td>
</tr>
<tr>
<td>volume depletion</td>
<td>11.90%</td>
</tr>
<tr>
<td>hypopotassemia</td>
<td>11.34%</td>
</tr>
<tr>
<td>diabetes uncomplicated type II</td>
<td>10.47%</td>
</tr>
<tr>
<td>pneumonia, organism unspecified</td>
<td>9.35%</td>
</tr>
<tr>
<td>angina, unstable</td>
<td>8.72%</td>
</tr>
<tr>
<td>hyposmolality and/or hyponatremia</td>
<td>8.47%</td>
</tr>
<tr>
<td>unspecified anemia</td>
<td>8.38%</td>
</tr>
<tr>
<td>acute posthemorrhagic anemia</td>
<td>8.14%</td>
</tr>
<tr>
<td>unspecified angina pectoris</td>
<td>7.90%</td>
</tr>
<tr>
<td>hyperplasia of prostate</td>
<td>6.54%</td>
</tr>
<tr>
<td>other specified cardiac dysrhythmias</td>
<td>5.61%</td>
</tr>
<tr>
<td>osteoarthrosis unspec gen/loc unspec site</td>
<td>5.20%</td>
</tr>
<tr>
<td>unspecified hypothyroidism</td>
<td>5.14%</td>
</tr>
<tr>
<td>unspecified chronic ischemic heart disease</td>
<td>5.13%</td>
</tr>
</tbody>
</table>
Similarity Constraint
(|J_a| disease constraints)
Collaborative Filtering
Inverse Frequency
Vector Similarity

Individual Medical History

Other Patients’ Medical Histories

ICARE Ensemble

Personal Risk Profile
Patient 156

Rank  Code  Disease
1  401.9  Hypertension
2  205   Diabetes
...

Ranked List of Disease Risks

Figure 5.2. A high-level overview of the CARE system.
straint with the active patient. This is done for multiple similarity constraints. The groups do not form a partition; each training patient may be no group, all groups, or any number in between. Collaborative filtering is then applied individually to each group. The output of each collaborative filtering run is a ranked list of disease risk scores for the active patient, ranked from highest risk to lowest. The resulting ranked lists are combined into an ensemble to produce a final ranked list of disease risk scores for the active testing patient.

Our collaborative filtering technique is derived from the vector similarity algorithm presented by [13], which we modify to fit the application. We begin with some notation. The entire training set of users is defined as $I$, and $I_j$ is the subset of users who have been diagnosed with a disease $j$. Similarly, $J$ is the set of all disease and $J_i$ is the set of diseases diagnosed to patient $i$. The variable $v_{i,j}$ is a patient’s “vote” for a disease; $v_{i,j} = 1$ if patient $i$ was diagnosed with disease $j$ and $v_{i,j} = 0$ otherwise. The collaborative filtering method produces a prediction score $p(a, j)$ for the active patient $a$ and a disease $j$ based on similarity weight $w(a, i)$ between $a$ and all training patients $i \in I_j$. Specifically, the prediction score is

$$p(a, j) = \frac{\sum_{i \in I_j} w(a, i)}{\sum_{i \in I} w(a, i)}$$

(5.1)

This is essentially a weighted calculation of the percentage of training patients who have been diagnosed with $j$. In previous work, an additional term introducing the population prevalence of disease $j$ was also included [28], but we have since found the algorithm to work better without it [32].

The similarity $w(a, i)$ between users $a$ and $i$ is calculated by vector similarity (also called cosine similarity) with inverse frequency. Formally,

$$w(a, i) = \sum_j \frac{f_j v_{a,j}}{\sqrt{\sum_{k \in J_a} f_k^2 v_{a,k}^2}} \frac{f_j v_{i,j}}{\sqrt{\sum_{k \in J_i} f_k^2 v_{i,k}^2}}$$

(5.2)
where

\[ f_j = \log \frac{|I|}{|I_j|} \]  

(5.3)

is the inverse frequency of disease \( j \). The inverse frequency term gives lower weights to very common diseases in the training set, based on the intuition that sharing a rare disease has more impact on similarity than sharing a common disease. This is particularly influential in the Medicare database, where extremely common codes such as \textit{essential hypertension} indicate little about patient similarity.

In early experiments, we found that inverse frequency was not sufficient to control the dominance of common diseases. We still observed that common diseases can overwhelm less common diagnoses since they account for the majority of the patients in the cluster. Ideally, we wanted to capture the effect of each individual disease with minimal noise from other diseases, but without the loss of information due to removing them. This is the purpose of the similarity constraints and the ICARE ensemble.

For each disease \( c \in J_a \) diagnosed for the test patient \( a \), collaborative filtering is applied separately to the group of training patients with disease \( c \) and a ranked prediction list is produced. Note that the variable \( c \) indicates the constraining disease, as opposed to the prediction target \( j \). While a single disease is used to constrain group membership, the collaborative filtering algorithm still uses the entire disease vector of the patients to determine the similarity score. Within each group, equation 5.1 is now as weighted calculation of the percentage of training patients with the constraining disease \( c \) who have also been diagnosed with \( j \). Thus, the constraining disease \( c \) has strong individual impact, but the other diagnoses determine whether a patient \( a \) is at higher or lower risk relative to other patients with \( c \). The ensembles
are combined by taking the maximum prediction score for each disease; that is

\[ p(a, j) = \max_{c \in I_a} \left( \frac{\sum_{i \in I_{j,c}} w(a, i)}{\sum_{i \in I_c} w(a, i)} \right) \]  \hspace{1cm} (5.4)\]

where \( I_{j,c} \) is the set of training patients with both \( j \) and \( c \). We choose the maximum since diseases are generally not protective against each other, with few exceptions.

In order to reduce the number of predictions and the runtime of the ensembles, we only generate a prediction for disease \( j \) if it is significantly more prevalent in the constrained group than in the entire training population. Otherwise, we assume that the constraining disease \( c \) does not influence the risk of \( j \). We determine significance using a two-proportion z-test.

5.2.4 Evaluation, Validation, and Practical Details

We evaluate the performance of CARE using recall in top 20 rank positions, referred to as coverage in previous work [28]. This sets a static threshold on the number of “positives”, rather than setting a threshold on the prediction score. Avoiding a threshold on prediction score is valuable since highest risk scores for one patient might be relatively low for another patient with more obvious concerns. The top 20 threshold is intended to represent a reasonable number for a medical practitioner to consider, condense, and potentially take action on. For some experiments, we also include recall in the top 100 rank positions, which is simply a broader version of the same concept. This is perhaps a reasonable length for an in-depth study of a specific patient.

We used a 2-fold cross-validation scheme for all experiments. For all methods, individual predictions are independent, making it easy to run experiments in a distributed fashion. Also, static calculations such as determining individual disease groups, random expectations, and inverse frequencies are preprocessed to avoid repetition.
5.3 Performance on Regional Outpatient Data

Our previous CARE experiments were all evaluated on the Medicare dataset, which is fairly age-homogeneous. To further validate the predictive performance, ICARE has also been tested on regional outpatient dataset, which spans all ages over a longer time period.

The regional dataset comprises outpatient records for approximately 250,000 patients within a large regional health system. Each patient is represented by an anonymized ID and associated ICD-9-CM diagnosis codes. The codes are provided in chronological order, with the length of time between diagnoses provided. This data spans 12 years, from 1997-2009, for patients of all ages. Basic demographics including gender and age range were included.

First, ICARE was applied to the regional dataset using the experimental setup described in 5.2.4. Table 5.3 presents the results for both the Medicare dataset and regional dataset. CARE shows better performance on the regional dataset. Thus, the age diversity and longer time span do not negatively impact the performance.

We also recreated the demographic experiments previously performed on the Medicare data in [27]. For these experiments, we partitioned the patient set into demographically homogeneous groups and evaluated performance separately for each

<table>
<thead>
<tr>
<th></th>
<th>Top 20</th>
<th>Top 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regional</td>
<td>0.448</td>
<td>0.662</td>
</tr>
<tr>
<td>Medicare</td>
<td>0.412</td>
<td>0.605</td>
</tr>
</tbody>
</table>

TABLE 5.3
ICARE PERFORMANCE ON THE MEDICARE AND REGIONAL DATASETS
ICARE PERFORMANCE ON DEMOGRAPHIC PARTITIONS OF THE REGIONAL DATASET

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Heterogeneous</th>
<th>Homogeneous</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-34</td>
<td>0.446</td>
<td><strong>0.459</strong></td>
</tr>
<tr>
<td>35-49</td>
<td><strong>0.436</strong></td>
<td>0.426</td>
</tr>
<tr>
<td>50-64</td>
<td><strong>0.455</strong></td>
<td>0.431</td>
</tr>
<tr>
<td>65-79</td>
<td><strong>0.472</strong></td>
<td>0.469</td>
</tr>
<tr>
<td>80+</td>
<td>0.434</td>
<td><strong>0.441</strong></td>
</tr>
<tr>
<td>Male</td>
<td><strong>0.459</strong></td>
<td>0.457</td>
</tr>
<tr>
<td>Female</td>
<td><strong>0.441</strong></td>
<td>0.433</td>
</tr>
</tbody>
</table>

The performance is compared for two training schemes. In the heterogeneous scheme, ICARE was trained on the full set of training patients then later demographically partitioned and evaluated separately for each group. In the homogeneous scheme, the training and testing sets were pre-partitioned and ICARE was trained separately for each demographic group using a homogeneous training set. The demographics available for this experiment were 15-year age categories and patient gender. Ages less than 20 were omitted due to low representation. The results are provided in Table 5.4. The evaluation metric is recall in the top 20 ranks. The bold values indicate the better performance between the heterogeneous and homogeneous training schemes. As previously observed in [27], the homogeneous training scheme degrades performance in most cases.

5.4 Time-Sensitive ICARE

ICARE does not take the order of or length between disease diagnoses into account when generating vector similarity among patients. However, a patient should
be considered more similar to another if their shared diseases follow a similar tem-
poral pattern, as well. Similarly, matching with two diseases which occurred many
years apart may not be relevant. For this reason, we modified our methods to in-
corporate the length of time between medical events (in our case, hospital visits).
In addition to more realistic similarity weights, using temporal information allows
our framework to extend to broad, general datasets with more complete medical
history. A limitation of our dataset is that the disease onset can only be identified
with the hospital admission, which might not accurately reflect the time the disease
was developed. Our dataset is also limited to a scope of four years, another disad-
vantage. Nevertheless, the goal of our system is to have the capability to incorporate
temporality, which has distinct advantages.

As shown in Equation 5.2, ICARE determines the similarity of the active patient
\( a \) and a training patient \( i \) as the vector similarity between the disease vector of \( a \) and
the entire disease vector of \( i \). The prediction score \( p(a, j) \) for every disease \( j \) in the
training vector will be weighted by this similarity. This implementation is blind to
the order of disease occurrence in the training patient; a common disease between the
active patient and training visit 5 will increase prediction scores for diseases which
occurred in training visit 1. This captures correlation, but it misses any causality
effects or natural ordering of disease occurrence. We are only interested in predicting
the future, so an overlapping disease should ideally only increase prediction scores
for diseases occurring in later training visits. However, this is too simplistic. In
most cases, \( a \) and \( i \) will have multiple overlaps in different visits, and considering
them individually would lose complex or synergistic effects.

Our method is a compromise. First, we developed an algorithm to find the
subset of consecutive training visits of \( i \) with the best vector match to the active
patient \( a \). We define \( sub_{s,z} \) to be the consecutive set of visits from visit \( s \) to visit \( z \).
Algorithm 1 Pseudocode for finding the best match subset of training visits

Algorithm best_match(a, visits)

1: maxsofar = 0
2: maxstart = 0
3: maxend = 0
4: currentmax = 0
5: currentstart = 0
6: for all m in visits do
7:   if \( w(a, sub_{m,m}) \geq w(a, sub_{currentstart,m}) \) then
8:     currentmax = \( w(a, sub_{m,m}) \)
9:     currentstart = m
10:  else
11:     currentmax = \( w(a, sub_{currentstart,m}) \)
12:  end if
13:  if currentmax \geq maxsofar then
14:     maxsofar = currentmax
15:     maxstart = currentstart
16:     maxend = m
17:  end if
18: end for
19: Return maxsofar, maxstart, maxend
For training patient $i$ with $n$ visits, the best match $\text{best}(a, i)$ to active patient $a$ is $\text{sub}_{s,z}$ such that

$$\max(w(a, \text{sub}_{s,z})), 1 \leq s \leq z \leq n$$

(5.5)

Similar to other maximum subsequence problems, $\text{best}(a, i)$ can be found in linear time. Our pseudocode is shown in Algorithm 1. While our algorithm is heuristic, it will be accurate for nearly all sequences that would realistically occur. As the algorithm scans, it tests whether the current visit yields a higher vector similarity in conjunction with the preceding best sequence or standing alone. If the current visit $v$ performs better without the preceding set of visits, then those visits will not be beneficial to any sequence containing $v$. Conversely, if similarity is higher when $v$ is combined with earlier visits, then those earlier visits will continue to be beneficial in any subset containing $v$. The concept is similar, though not identical, to fraction multiplication; note that for any two vectors, $0 < w(a, i) < 1$. Starting with the largest fraction will always be better, regardless of future values.

Intuitively, $\text{best}(a, i)$ is the time period when training patient $i$ was having the most similar medical experience to the active patient $a$, and the visits immediately following should have the most relevant information to the prognosis of $a$. Thus, we modified the general equation so the “best match” vector similarity only adds prediction weight for diseases which occur in visits after the “best match” time frame. Assuming that $\text{best}(a, i) = \text{sub}_{s,z}$ and $i$ has $n$ visits, then $Z_j$ is the set of patients $i$ such that $j \in \text{sub}_{z+1,n}$. The time-sensitive general equation is then

$$p(a, j) = \frac{\sum_{i \in Z_j} w(a, \text{best}(a, i))}{\sum_{i \in I} w(a, \text{best}(a, i))}$$

(5.6)

If the “best match” includes the last visit for $i$, we assume that $i$ provides no knowledge about the future of $a$.

The time sensitive extension can be used with or without ICARE ensembles.
TABLE 5.5

EVALUATION OF PERFORMANCE ICARE AND TIME-SENSITIVE ICARE COMPARED WITH THE BASELINE RANKING

<table>
<thead>
<tr>
<th></th>
<th>Top 20</th>
<th>Top 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.321</td>
<td>0.585</td>
</tr>
<tr>
<td>ICARE</td>
<td>0.412</td>
<td>0.605</td>
</tr>
<tr>
<td>time ICARE</td>
<td>0.385</td>
<td>0.594</td>
</tr>
</tbody>
</table>

When using a constraint $c$, we simply require that the best match algorithm skip subvectors which do not contain $c$. The prediction score is slightly modified to

$$p(a, j) = \frac{\sum_{i \in Z_{c,j}} w(a, \text{best}(a, i))}{\sum_{i \in I_c} w(a, \text{best}(a, i))}$$

(5.7)

where $Z_{c,j}$ is the set of patients $i$ such that $c \in \text{sub}_{s,z}$ and $j \in \text{sub}_{s+1,n}$.

Finding a best match subset of visits incorporates the ordering of diseases and resolves the problem of “predicting the past”, while still preserving multiple-disease interactions. Additionally, this strategy makes the CARE framework feasible for long-term, diverse data, such as public health records. Over a lifetime, people may go through many different medical experiences and phases, and very few people will have the same experience over a period of many years. However, similarity within a short window may be very strong. The best match is able to isolate the most relevant time periods without all of the noise generated by the rest of the medical record. A simple cutoff mechanism could easily be used to limit the breadth of the training patients’ ‘future’ influencing the predictions, as well.
5.4.1 Performance Results

Table 5.5 displays the experimental results. The metrics are also applied to the baseline ranking, which is a list of the diseases ranked in order from highest baseline prevalence to lowest.

Time-sensitive ICARE shows a small loss of performance, but still outperforms the baseline significantly. This loss is easily explainable, since ICARE has a stronger bias for ranking chronic diseases, which are very prevalent among senior citizens. Since the ensemble method looks at each diagnosis individually, any repeat diseases will create a group around themselves with $v_{j,c} = 1$. This leads to a perfect ranking of chronic diseases. The time-sensitive method does not carry that bias, since only visits occurring after the “best match” affect $v_{j,c}$, which may or may not contain a repeat code. The time-sensitive version will predict chronic diseases based on their likelihood of repeat visits rather than assuming 100% likelihood. In the Medicare data, the dominance of chronic diseases causes ICARE’s assumption to be beneficial, but would likely be less influential in a more general setting. In fact, we find that the time-sensitive version performs slightly better when considering non-repeat (not previously diagnosed to the active patient) diseases. Specifically, we see 19.9% recall in the top 20 ranks versus 16.9% with unmodified ICARE. Finally, it is worth noting that the time-sensitive methods are necessary for computational efficiency and noise control when applied to a long-term general database. We posit that a minor drop in performance is acceptable in light of these more practical concerns.

5.5 Performance Trends

We also are interested in how performance changes with respect to the amount of data known about the testing patient. This analysis provides insight into optimal deployment of such a system in a practical setting. It provides guidelines
for the minimum amount of information needed for meaningful (better than baseline) results and a threshold for good results without overcomplicated computation. Specifically, we looked at the number of visits known by CARE about the testing patient (Figure 5.3A), the total number of unique diseases known about the patient (Figure 5.3B), and the length of time in days between the patient’s last known visit and the following unknown, and thus predictable, visit (Figure 5.3C). We look at the recall within the top 20 ranks, which we believe is our most practical measure of performance. This experiment uses the Medicare dataset.

The visit and diseases trends show that performance continually increases as more information is known about the patient. The results suggest that ICARE on a single visit is sufficient to outperform the baseline, though 5.3B shows that the visit should have at least 3 diseases. The benefit of additional diseases flattens around 25 unique diagnoses. The data for patients with more than 35 diseases is too sparse for further conclusions, but can be expected to continue in a flat line near 57% recall.

Unsurprisingly, as the length of time since the last visit increases, a modest drop in performance can be observed. The intuition here is that older diagnoses are less relevant to immediate concerns, on average. Despite the downward trend, ICARE still outperforms the baseline after gaps of more than 2-3 years. A more long-term study of this effect would be interesting, but we are limited by the scope of our data.

Since we have mentioned multiple times the dominance of common diseases, we examined our method’s ability to control this effect. To do this, we looked at the distribution of disease prevalence of the patients’ actual future diseases compared to the predictions from the baseline and ICARE. This analysis is shown in table 5.6. The first column is the percent prevalence of a disease in the patient population, which is equivalent to the random expectation $\overline{v}_j$. The second column shows what percentage of the actual diagnoses fall within each prevalence range. While there are
Figure 5.3. Recall trends with respect to known testing patient data. (A) Performance with respect to the number of visits known by CARE about the testing patient. (B) Performance with respect to the total number of unique diseases known about the patient. (C) Performance with respect to the length of time in days between the patient’s last training visit and first testing visit.
TABLE 5.6


<table>
<thead>
<tr>
<th>% Disease Prevalence</th>
<th>Actual Diagnoses</th>
<th>Baseline Top 20</th>
<th>ICARE Top 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-5</td>
<td>53%</td>
<td>1%</td>
<td>24%</td>
</tr>
<tr>
<td>5-10</td>
<td>20%</td>
<td>6%</td>
<td>22%</td>
</tr>
<tr>
<td>10-15</td>
<td>5%</td>
<td>29%</td>
<td>10%</td>
</tr>
<tr>
<td>15-20</td>
<td>3%</td>
<td>13%</td>
<td>8%</td>
</tr>
<tr>
<td>20-25</td>
<td>4%</td>
<td>18%</td>
<td>10%</td>
</tr>
<tr>
<td>25-30</td>
<td>4%</td>
<td>13%</td>
<td>8%</td>
</tr>
<tr>
<td>30-35</td>
<td>2%</td>
<td>4%</td>
<td>5%</td>
</tr>
<tr>
<td>35-40</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>40-45</td>
<td>4%</td>
<td>8%</td>
<td>8%</td>
</tr>
<tr>
<td>45-50</td>
<td>2%</td>
<td>4%</td>
<td>2%</td>
</tr>
<tr>
<td>50-55</td>
<td>2%</td>
<td>4%</td>
<td>3%</td>
</tr>
</tbody>
</table>
a few diseases which are very common, there are many different uncommon diseases which account for most of the actual diagnoses. In fact, 53% of actual diagnoses are diseases which are less than 5% prevalent in the whole patient population. The final two columns shows the percentage of the top 20 predictions which fall within each prevalence range; the third column uses the baseline ranking and the fourth column uses ICARE. The baseline results clearly show that extra controls are needed to avoid skew toward common diseases. Also, the results show that ICARE does well at limiting very common diseases to a realistic percentage of the strongest predictions. Also, most of the top predictions by ICARE are low-prevalence diseases, which is the case in reality.

5.6 Focused Experiments for Data Exploration

We also did a more in-depth exploration of the regional dataset introduced in Section 5.3 by applying CARE as a research tool to focused groups of diseases and patients. Through samples of the population controlled by disease type, patient age, or record timespan, we observed various factors that influence the measured performance or practical value of the CARE predictions.

5.6.1 Focus on Disease Predictability

While CARE has shown good overall performance, it seems inevitable that some disease categories are less predictable than others. Also, some situations may have very clear-cut association with a single ICD-9 code, while others may be susceptible to variation. We evaluated predictive performance of specific diseases with varying levels of hierarchical granularity.

We designed an measure to determine how well CARE performs at predicting each disease code for the right patients. Performance is evaluated separately for each unique disease code. For a disease $d$, each patient was an instance, the prediction
Figure 5.4. **Calculating disease predictability: a toy example.** (A) A feature vector is constructed for each disease \( d \). Each patient is an instance with two features: the rank of \( d \) for the patient and a class indicating whether the patient later develops \( d \). (B) Precision and recall are calculated for each rank threshold. (C) The AUPR is used as a measure of the predictability of disease \( d \).

Score was the rank position of \( d \) (\( \infty \) if \( d \) is not ranked), and the class was 1 if the patient later develops \( d \) and 0 otherwise. Patients were excluded if previously diagnosed with the code, regardless of whether the diagnosis was repeated later. We then calculated precision and recall for each rank threshold from 1-50, creating a Precision-Recall curve. Finally, we used the area under the curve (AUPR) as a measure of the predictability of disease \( d \). This procedure essentially rewards high rankings for the right patients and penalizes high rankings for the wrong patients. An small-scale toy demonstration of this process for 10 patients over the top 5 ranks is provided in Figure 5.4.

The experiment described above is performed separately for each level of the code hierarchy. That is, we first compared the predictability of 5-digit codes, then 4-digit codes, and finally, 3-digit codes (see Section 3.3). When going to less specific level, each ICD-9-CM code is also counted as an instance its hierarchical parent.
For example, *Cerebral thrombosis with cerebral infarction* (434.01) would count as a *Cerebral thrombosis* (434.0) at the 4-digit level and *Occlusion of cerebral arteries* (434) at the 3-digit level. Thus, if the patient actually develops *Cerebral thrombosis without cerebral infarction* (434.00), incorrectly predicting the presence of a cerebral infarction is only an error at the 5-digit level, and the prediction of cerebral thrombosis would still be considered correct. This way, instead of allowing only exact matches, we measured the level of granularity to which each predicted code is correct.

The top 10 most predictable and least predictable diseases at each level of granularity are shown in Table 5.7. The most predictable disease codes correspond to the highest AUPR values, and least predictable codes are the most prevalent diseases that are never highly ranked for the correct patients.
### TABLE 5.7

THE MOST PREDICTABLE AND LEAST PREDICTABLE ICD-9-CM DISEASES CODES

<table>
<thead>
<tr>
<th>3-Digit Codes</th>
<th>4-Digit Codes</th>
<th>5-Digit Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Top 10 Most Predictable Codes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>401 Hypertension</td>
<td>2826 Sickle-cell disease</td>
<td>25002 Diabetes w/out complication, uncontrolled</td>
</tr>
<tr>
<td>272 Disorders of lipid metabolism</td>
<td>4019 Unspec. hypertension</td>
<td>60001 Hypertrophy of prostate w/ obstruction</td>
</tr>
<tr>
<td>250 Diabetes</td>
<td>2500 Diabetes w/out complication</td>
<td>60000 Hypertrophy of prostate w/out obstruction</td>
</tr>
<tr>
<td>786 Symptoms, resp. system</td>
<td>6000 Hypertrophy of prostate</td>
<td>78000 Nodular prostate w/out obstruction</td>
</tr>
<tr>
<td>780 General symptoms</td>
<td>7194 Pain in joint</td>
<td>78001 Nodular prostate with obstruction</td>
</tr>
<tr>
<td>600 Hyperplasia of prostate</td>
<td>4660 Acute bronchitis</td>
<td>58381 Nephritis and nephropathy non-chronic</td>
</tr>
<tr>
<td>719 Unspec. disorders of joint</td>
<td>2724 Unspec. hyperlipidemia</td>
<td>78079 Malaise and fatigue</td>
</tr>
<tr>
<td>466 Acute bronchitis and bronchiolitis</td>
<td>4619 Acute sinusitis unspec.</td>
<td>34500 Nonconvulsive epilepsy</td>
</tr>
<tr>
<td>461 Acute sinusitis</td>
<td>1891 Malignant neoplasm of renal pelvis</td>
<td>78650 Unspec. chest pain</td>
</tr>
<tr>
<td>733 Other disorders of bone/cartilage</td>
<td>5838 Nephritis and nephropathy w/ lesion</td>
<td>73300 Osteoporosis</td>
</tr>
<tr>
<td><strong>Top 10 Least Predictable Codes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>553 hernia of abdominal cavity</td>
<td>7886 Abnormality of urination</td>
<td>72887 Muscle weakness</td>
</tr>
<tr>
<td>389 Hearing loss</td>
<td>5246 TMJ disorders</td>
<td>78909 Abdominal pain other specified site</td>
</tr>
<tr>
<td>560 Intestinal obstruction w/out hernia</td>
<td>3899 Unspec. hearing loss</td>
<td>78863 Urgency of urination</td>
</tr>
<tr>
<td>536 Disorders of stomach function</td>
<td>5368 Dyspepsia</td>
<td>68110 Cellulitis/abscess of toe</td>
</tr>
<tr>
<td>410 Acute myocardial infarction</td>
<td>6010 Acute prostatitis</td>
<td>80700 Closed fracture of rib(s)</td>
</tr>
<tr>
<td>368 Visual disturbances</td>
<td>0093 Diarrhea of infectious origin</td>
<td>46619 Acute bronchiolitis of infectious origin</td>
</tr>
<tr>
<td>524 Dentofacial anomalies</td>
<td>4109 Acute myocardial infarction unspec.</td>
<td>38830 Tinnitus unspec.</td>
</tr>
<tr>
<td>482 Other bacterial pneumonia</td>
<td>6825 Cellulitis/abscess of buttock</td>
<td>52460 TMJ disorders unspec.</td>
</tr>
<tr>
<td>235 Neoplasm of digestive/resp. systems</td>
<td>7842 Swelling lump in head/neck</td>
<td>37272 Conjunctival hemorrhage</td>
</tr>
<tr>
<td>214 Lipoma</td>
<td>4661 Acute bronchiolitis</td>
<td>68102 Onychia and paronychia of finger</td>
</tr>
</tbody>
</table>
The most predictable diseases consist primarily of very common diagnoses that occur in many elderly patients (hypertension, diabetes, osteoporosis), symptoms of existing problems (fatigue, chest pain, joint pain), and acute respiratory conditions (bronchitis, sinusitis) that often develop during other illnesses. A few are clearly permanent chronic conditions that are simply not included earlier in the patient’s record (sickle-cell disease, epilepsy), but likely show obvious signs. It is not surprising that the most predictable diseases are also some of the most obvious. It is a good sign that CARE can capture obvious omissions or common high-probability events, but for useful deployment, additional filters or stronger measures for the “interestingness” of a diagnosis would be beneficial.

The least predictable diseases include many injuries (hernia, fracture of ribs, swelling lump on head, etc.) and infections (diarrhea, bronchiolitis, pneumonia, prostatitis, abscesses, onychia of finger, etc.) which are inherently difficult or impossible to predict. The code “acute myocardial infarction” stands out as a disappointing entry on this list. While predicting the exact time of a heart attack is obviously difficult, there are known risk factors. However, the most well-known risk factors, such as smoking, obesity, and high cholesterol, are extremely under-reported in our data. The Centers for Disease Control and Prevention (CDC) reported a 26.3-29.7% obesity rate for the region corresponding to our dataset in 2003, the middle of the data collection span [18]. In drastic contrast, the reported prevalence of adult obesity in our dataset is 1.3%. Similarly, a study by Kuklina et al. determined that the prevalence of high cholesterol in the U.S. ranged from 31.5% in 1999 to 21.2% in 2006 [71], while our reported prevalence is 4.3%. Only 13 patients total are reported as tobacco users. This is a reminder that accurate computational medical prediction requires consistent, complete records, including non-acute risk factors that are often overlooked.
5.6.2 Focus on Chronic vs Non-chronic Diseases

We explored how well CARE can distinguish which diseases will repeat in later visits versus those which will not. Using time-sensitive ICARE, we scored and ranked only diseases which were included in the diagnoses given to the algorithm. We then evaluated how often repeating diseases are ranked above non-repeating diseases using AUROC, treating repeats as the positive class and non-repeats as the negative class. The resulting AUROC was .7889, where repeating diseases had an average rank of 13.4693 and non-repeats had an average rank of 43.606.

Of course, many of these highly ranked repeating diseases are actually chronic diseases rather than recurring conditions. Chronic diseases are the easiest to predict but also the least interesting, since they persist by definition. For most of the highly ranked repeating disease, the strongest indicative factor was an earlier diagnosis of the same disease. Conversely, many of the “errors” in ranking are chronic diseases which were not noted during an unrelated medical visit, but nonetheless probably persisted in reality. The easiest way to eliminate these uninteresting predictions and deceptive “errors” is to eliminate previously diagnosed diseases from the ranking. However, there are cases where the repetitions of a normal non-chronic disease are independent, or maybe point to another underlying condition. Our results show some examples of conditions do not strongly indicate their own repetition, but can be correctly repeated due to other factors. For example, hypocalcemia never correctly indicates its own repeat, but is predicted around hypoparathyroidism with 64% precision. Thus, by checking the strongest indicative factor, we can not only distinguish repeating diseases versus non-repeating diseases, and also chronic diseases versus recurrent events. This will allow us to filter chronic diseases into a separate list from actual risk scores, so that information about chronic diseases is available but not represented as something to be treated as a potential future risk.
5.6.3 Focus on Age

When initially testing CARE, we experimented to see if patient age could be used to improve the predictive performance of the CARE framework. We hypothesized that using training sets homogenized with respect to age category might be beneficial. However, we found that this was not the case in the Medicare data; segmenting the data into age-homogeneous groups did not significantly improve performance, and in some cases even degraded performance [27]. This also holds for the regional data used in these studies as well, as previously shown is Section 5.3, Table 5.4. We conjectured that because we choose groups of similar training patients based on each individual disease in the history of the patient (See ICARE with ensembles,[27]), the appropriate demographic distribution for each disease emerges automatically for most patients. However, there are still patients for whom the homogeneous training sets were beneficial. The following experiment explores the differences between patients who benefit from one training set over the other.

We focused on how the age distribution of a patient’s diseases affect the performance of the age-heterogeneous vs. the age-homogeneous runs. For this experiment, we defined the age distribution of a disease $d$ as the percentage of all patients diagnosed with disease $d$ that fall into each age category. Thus, each distribution had five values, one for each age category, which sum to 100. These values correspond to the proportion of patients from each age category within the cluster formed around disease $d$ when applying ICARE. For each age group, we averaged the age distributions of all diseases which occur within the group to get a baseline distribution. Similarly, we averaged the distributions of each patient’s past diagnoses and future diagnoses, respectively. Finally, we examined how the patient-specific distributions for patients in a given age category compare to the corresponding baselines. Specifically, we were interested in the differences between patients whose predic-
Figure 5.5. **Comparison of diagnoses for patients who benefit from homogeneous vs heterogeneous training sets.** Comparison of past and future diagnosis age distributions for patients in the age category 20-35, compared to the baseline distribution of diagnoses for all patients in the age category. **(A)** Averaged age distributions for patients in the 20-35 age category for whom performance is better with age-homogeneous training sets. **(B)** Averaged age distributions for patients in the 20-35 age category for whom performance is better with age-heterogeneous training sets.

Tions benefit from age-heterogeneous training sets versus those that benefit from age-homogeneous data.

Figure 5.5 shows an example of the trends for age category 20-34. Across patients who benefit for age-homogeneous training sets, we observed that the patients’ past diagnoses shift away from their own age category, while their future diagnoses shift towards the age category. Across those who benefit from the age-heterogeneous training set, the trend shows that both past and future diagnoses shift away from the age category. All of these shifts seem slight, but it only takes one influential and highly correlated disease to shift the predictions. Furthermore, the trends are generalizable to all of the age categories studied.

Since the age distribution of past diagnoses for both group of patients shift in the same direction, there is no clear way to choose the better training set based only
on the known diagnoses for a patient. Incorporating the age of the patient into the collaborative filtering weighting scheme is probably a better approach. This would allow higher weights within the same age category without losing information about diseases that are more common in another age group. Of course, this approach has many of its own challenges, making it an interesting avenue for future work.

5.6.4 Focus on Timespan

The 12 year span of the regional data provides an opportunity to look at how far into the future predictions can reliably be made. Specifically, we partitioned the active patient’s diagnoses by years and evaluate performance within the next year through 10 years in 1 year increments. Since recall improves with the amount of data known [28], we looked at the year-by-year performance with respect to the number of years of patient history given to the CARE process. The trends for recall in the top 20 ranks are shown in Figure 5.6.

As previously observed, each additional year of information improves the performance of CARE, as demonstrated by the vertical separation of the lines in Figure 5.6A. The horizontal trend shows that CARE has highest recall in the immediately following year, which then declines for each additional year into the future. Nonetheless, Figure 5.6B shows that CARE continues to outperform the baseline even at 9 years into the future, although the baseline performance declines more slowly. The exception is when only one year of medical history is provided, which only outperforms the baseline up through the first four years; performance then degrades to approximately the same as baseline.

5.7 Conclusion

In this chapter, we described an extension to incorporate temporal information into the CARE framework, further validated the promising performance on a re-
Figure 5.6. **Year-by-year performance of CARE.** (A) Recall of diagnoses occurring within each future year, categorized by the number of years of previous medical history. (B) Increased recall of CARE over the baseline, year-by-year.
gional outpatient dataset, and demonstrated the use of CARE as a tool to explore focused questions about prediction trends and performance. The addition of time-sensitivity makes the CARE framework more feasible on large, diverse datasets spanning many years, such as the comprehensive records mentioned above. By making our framework more time sensitive, we reap multiple practical benefits. The time-sensitive approach also adds some limited ability to differentiate between chronic disease and lone occurrence. The focused studies provided interesting insight into which diseases are predictable, the characteristics of challenging patients, and the time span for which predictions are valuable. All of these observations can be used to improve the system in future work.
CHAPTER 6

NETWORK ANALYSIS OF A NICU TEAM STRUCTURE

Understanding biological disease mechanisms and clinical co-morbidities are certainly valuable strategies for treating diseases, recognizing risk, and informing clinical decisions. Still, some avoidable medical events are strongly or entirely attributable to environmental factors, behavior patterns, or human error. Physical injury and infectious diseases are often symptoms of circumstance rather than biology. This does not mean we are helpless against them; on the contrary, modifying behavior and environment is generally far easier, cheaper, and less dangerous than drug or genetic therapy. In this chapter, we demonstrate that networks modeling human behavior patterns can be powerful tools for preventing harmful events.

6.1 Motivation and Background

Despite the incredible advances in neonatal intensive care in the past 3 decades, neonatal intensive care unit (NICU) patients remain at substantial risk of preventable medical harm, including adverse drug events, nosocomial infections, and morbidities that are unique to premature infants [44, 126, 113]. In addition to intrinsic patient susceptibility, the increased risk results in part from the complexity and long duration of NICU stays. Cohesive, well-functioning, multidisciplinary teams must assemble at the patient level to manage information flow accurately and to provide seamless care during extended, complex hospitalizations. However, maintaining
effective, well-functioning teams is a daunting challenge during hospital stays that may include up to 300 nursing transitions/handoffs at shift changes. Identification of solutions has been hampered by the lack of robust methods for quantification of the nature of team interactions, to identify patterns that are potentially associated with adverse events [43, 11]. The availability of network analytic tools to apply to complex systems such as those involved in NICU care provides an opportunity to enhance the research on effective team structure and function.

Across a diverse spectrum of system types, it has become clear that the performance and fault tolerance of networks or teams are related to both network topological features and the ways in which they form over time [137, 73]. Furthermore, a large body of evidence demonstrates the importance of organizational characteristics of care as predictors of performance across multiple domains [117, 116, 115, 114]. These characteristics are predictors not only of clinical outcomes but also of patient and provider satisfaction. Although the term “team” is often applied to the groups of providers who come together to care for a patient, it is not always clear that these groups truly are teams. Grumbach and Bodenheimer [48] suggest that 2 elements, appropriate composition and successful communication, are crucial if these groups are to function successfully.

This project was designed as a demonstration of the use of network analysis to describe the structure of NICU nursing care teams. Using data from the electronic health record (EHR) of a large, tertiary care NICU, we explored the topological properties of nursing handoff networks and examined the relationship between network characteristics and one aspect of quality, namely, family satisfaction with care. Section 6.2 provides details and basic statistics about the EHR data used for this study. In Section 6.3, we describe the experimental methods used in this study, including construction of the nursing handoff networks (6.3.1), calculation of topo-
logical properties (6.3.2), evaluation of family satisfaction with care (6.3.3), and statistical analysis (6.3.4). The results are presented in Section 6.4, followed by a discussion of the findings and limitations in Section 6.5.

6.2 Data Description

This study included all eligible infants admitted to the NICU at the Beth Israel Deaconess Medical Center between January 1, 2002 and December 31, 2007 and the nurses who cared for them. The primary patient eligibility requirement was that at least one nursing care handoff(s) occurred during the infant’s NICU stay. A nursing handoff occurs when responsibility for an infant’s care is transferred from one nurse to another, usually during a nursing shift change. Demographic characteristics, including birth weight and gestational age, were recorded. During the study period, 168 nurses provided care to the 3891 eligible infants. The median length of stay was 9 days (18 handoffs), with the largest number of handoffs being 311. Characteristics of this group of infants are presented in Table 6.1. This group of patients was used to explore the properties of the nursing team structure.

A subgroup of these patients was used to explore the relationship between nursing team structure and quality of care. This analysis had the additional requirement that satisfaction survey results were returned and available for the infant. The survey results were a requirement of our experimental design and will be described in more detail in Section 6.3.3. Surveys were available for 93 (57%) of 165 otherwise eligible infants born in 2006 who were discharged to home. Responding families were similar to nonrespondents in terms of their infants’ length of stay, birth weight, gestational age, and nursing team size. On this basis, we assume that the pool of eligible infants sufficiently represent the NICU population.
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>n(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gestational Age</strong></td>
<td></td>
</tr>
<tr>
<td>≤ 28 wk</td>
<td>367 (9.4)</td>
</tr>
<tr>
<td>29 – 32 wk</td>
<td>820 (21.1)</td>
</tr>
<tr>
<td>33 – 36 wk</td>
<td>1667 (42.9)</td>
</tr>
<tr>
<td>≥ 37 wk</td>
<td>938 (24.1)</td>
</tr>
<tr>
<td>Data not available</td>
<td>99 (2.5)</td>
</tr>
<tr>
<td><strong>Birth Weight</strong></td>
<td></td>
</tr>
<tr>
<td>≤ 999 g</td>
<td>259 (6.7)</td>
</tr>
<tr>
<td>1000 – 1499 g</td>
<td>488 (12.5)</td>
</tr>
<tr>
<td>1500 – 2499 g</td>
<td>1652 (42.5)</td>
</tr>
<tr>
<td>≥ 2500 g</td>
<td>1388 (35.6)</td>
</tr>
<tr>
<td>Data not available</td>
<td>104 (2.7)</td>
</tr>
<tr>
<td><strong>Length of Stay</strong></td>
<td></td>
</tr>
<tr>
<td>≤ 7 d</td>
<td>1759 (45.2)</td>
</tr>
<tr>
<td>8 – 14 d</td>
<td>740 (19.0)</td>
</tr>
<tr>
<td>15 – 21 d</td>
<td>399 (10.3)</td>
</tr>
<tr>
<td>22 – 28 d</td>
<td>206 (5.3)</td>
</tr>
<tr>
<td>28 – 60 d</td>
<td>502 (12.9)</td>
</tr>
<tr>
<td>&gt; 60 d</td>
<td>285 (7.3)</td>
</tr>
</tbody>
</table>
6.3 Experimental Methods

6.3.1 Patient Handoff Networks

The NICU EHR was queried to identify all nurse-patient interactions during the study period. A nurse was considered to have cared for a patient if a progress note from the nurse was contained within the patient’s record. During each nursing shift (generally 12 hours in length), at least one note is written by the nurse caring for a patient, according to unit policy. Notes written by patient care assistants and registered nurses working in other roles, such as case management, were not considered. If more than one note was written by an individual nurse during a shift, then the second note was excluded. Timestamps were used to identify the sequence of nurse participation. For each eligible infant (eligibility requirements are described in Section 6.2), we extracted an individual patient handoff network, representing the handoffs that occurred between nurses during that patient’s care. Specifically, each nurse was represented as a network node, and each transfer of responsibility for the infant’s care forms a directed edge between two nurses. The edges are labeled such that the sequence of handoffs is preserved, and multi-edges do occur.

6.3.2 Measuring Team Structure

For each patient, network size was calculated as the number of unique nurses (nodes) contained in the patient’s handoff network. In addition, we also examined effective network diameter and average betweenness centrality [133] at the individual patient level. These measures were intended to capture the ease of information flow through the networks. The effective network diameter represents the maximum number of intermediate nurses necessary for the transfer of information from one nurse to another, while average betweenness centrality indicates the degree to which the network has a center of influential nurses who are positioned to control
information flow.

For each nursing shift, a nurse was considered a newcomer if he or she had not previously been part of the patient’s handoff network. The proportion of newcomer shifts in a patient’s nursing care team was calculated as the number of shifts provided by newcomer nurses divided by the total number of nursing shifts contained in the network.

To quantify the continuity of nursing care provided, we developed another network metric based on the concept of network cycles [133]. The mean repeat caregiver interval (MRCI) was calculated as follows. For each patient, beginning at the first nursing shift, we counted the number of shifts until the first care provider who had already cared for that patient was encountered. This was considered the repeat caregiver interval (RCI). Beginning at the end of the previous RCI, this process was repeated iteratively to determine subsequent RCIs for the index patient. RCIs were then summed per patient and divided by the number of RCIs encountered, to provide a MRCI for the patient. This process is diagrammed in Figure 6.1.

6.3.3 Measuring Quality of Care

The Picker Institute Neonatal Intensive Care Unit Family Satisfaction Survey [45, 22] was used to assess parental perceptions of the care provided to their family. This 74-item written questionnaire was administered as part of routine quality-improvement activities and was sent one month after discharge to families with a child who remained in the NICU for two or more weeks. Surveys were not sent to families of infants who died or were discharged to chronic care facilities.

Responses to each question were tabulated to form a problem score, representing the proportion of questions within a dimension that elicited a problem response. For example, one item asks, How would you rate the way the NICU staff worked...
Figure 6.1. **Example patient handoff network and calculation of mean repeat caregiver interval (MRCI).** (A) Each network represents the nursing care team of a single infant over the entire NICU stay. Each unique nurse (RN) participating in the care team is represented as a node in the network, and directed edges indicate the sequence in which responsibility for the patient is transferred between nurses. (B) Traversing the network in order starting from the first caregiver (RN A), 5 handoffs occur before a caregiver is repeated (RN C). Thus, the first repeat caregiver interval (RCI 1) is of length 5. The second RCI begins with the last nurse in RCI 2 (RN C), and 3 handoffs occur before another repeat. There are no additional handoffs, so this network has a total of 2 RCIs. The mean repeat caregiver interval (MRCI) is 4, the average of RCI 1 and RCI 2.
together? Possible responses to all items were {poor, fair, good, very good, excellent}. For this study, “poor” and “fair” were considered to be problem responses. Provider-specific problem scores for nurses, physicians, respiratory therapists, and social workers were calculated by tabulating responses for questions relevant to each discipline. For twins and other multiple gestations, a single infant was chosen randomly for inclusion in these analyses.

6.3.4 Statistical Analysis

All of the team structure measures were dichotomized as low or high relative to the median value observed for the cohort. We then explored the difference between high or low measurements with respect to family satisfaction for the relevant patient subpopulation. Univariate comparisons of problem scores from the satisfaction survey were performed by using Student’s t test. Multivariate testing to control for potential confounders was performed by using logistic regression analysis.

6.4 Results

6.4.1 Patient Handoff Networks

The nursing handoff networks varied from linear structures to more-densely connected networks with connections existing between many of the possible nurse dyads. Figure 6.2 presents three such networks and their associated topological features. The networks in Figure 6.2 A and B are reminiscent of the game telephone, in which information is passed from one participant to another, with limited possibilities for previous participants to verify that the retelling of information has been accurate. In contrast, the network in Figure 6.2 C contains multiple opportunities for previous participants to review information and to correct inaccuracies that may result from transmission errors. Therefore, we hypothesize that the handoff networks in Figure 6.2 A and B seem less likely to foster continuity of care and adequate, accurate,
Figure 6.2. *Nursing handoff networks and associated measures for 3 NICU patients*. Patient (A) was cared for by 20 nurses over a stay of 21 nursing shifts, indicating that only one nurse cared for the infant more than once. Patient (B) was hospitalized for 36 nursing shifts and cared for by 26 nurses, with some long repeat caregiver intervals. Finally, patient (C) was cared for by 19 nurses during 42 nursing shifts, with more frequent repeat caregivers. Based on the observed measurements, we hypothesize that (A) and (B) are less likely to result in effective, efficient communication than network (C).

6.4.2 Team Size and Frequency of Newcomers to Care Team

The size of an infant’s nursing team increased with longer lengths of stay (Figure 6.3). For example, for infants who remained in the NICU for 14 nursing shifts, the median number of nurses who cared for the infant was 9.7 nurses (interquartile range: 811 nurses). The rate of increase in team size remained high throughout the range of lengths of stay seen. When infants were grouped according to birth weight
Figure 6.3. **Nursing team size.** The number of nurses involved in a patient’s care team as a function of the length of stay.

category (<1500, 1500-2499, or >2500 g), the changes in team size according to duration of hospitalization were similar, which suggests that similar approaches to incorporation of newcomers were used across birth weight categories.

Viewed from a slightly different perspective, the data on team composition could be used to identify the proportion of team newcomers providing care at individual points during a NICU stay (Figure 6.4). Given our definition of team newcomers, care in all first and second shifts was provided by newcomers. A rapid decrease in the proportion of newcomers was seen over the first 14 nursing shifts (corresponding to 7 days of age, in most cases). Even by the 14th shift, however, 53% of shifts still involved newcomers. Only after the 56th shift were >1 of 3 shifts provided by a nurse who had not previously cared for the infant.

Although nursing teams for less-mature infants were larger and included more newcomers, those teams showed higher levels of centralization (ie, the presence of
Figure 6.4. **Percentage of newcomer nurses.** The proportion of shifts provided by newcomer nurses relative to the length of stay.

...a central core of nurses within the nursing team), compared with teams for more-mature infants (Figure 6.5). Similar trends were seen when infants were grouped according to birth weight. It is interesting to note that there was relatively greater variation in network centralization than in team size among infants with normal birth weights.

6.4.3 Relationship Between Team Structure and Family Satisfaction

For the 9 survey items related to nursing care, most families reported 0 (34.4% of respondents) or 1 (18.3% of respondents) problem responses. The number of problem responses did not differ according to birth weight category or length of stay.

We saw no significant difference in nursing dimension problem scores for patients with small versus large nursing teams (20% vs 27%; P > .05). Similarly, there
Figure 6.5. **Team centralization and size.** The relationships between gestational age category, NICU nursing team centralization index, and team size. The charted values are medians for each measure, with error bars representing interquartile ranges.

were no differences in nursing problem scores for patients with large versus small proportions of newcomer nurses during the hospital stay (23% vs 24%; P > .05), large versus small nursing network diameters (23% vs 23%; P > .05), or high versus low values of group betweenness (15% vs 16%; P > .05).

Finally, we examined the MRCI as a measure of continuity within the nursing care team. Higher MRCI values represent longer latency between repeat caregivers within a patient’s care network. The median MRCI value was 5.2 nursing shifts (interquartile range: 4.75.8 nursing shifts). Over the years studied, increases in MRCI values were seen among infants born weighing >1500 g (4.8 ± 0.77 nursing shifts in 2002 and 5.1 ± 0.82 nursing shifts in 2007). However, similar trends were not seen for very low birth weight infants.

Nursing dimension problem scores were significantly higher (worse) for infants with longer intervals between repeat nursing caregivers (ie, higher MRCI values;
29.6% vs 17.6%; P < .05). As shown in Figure 6.6, for 8 of the 9 items constituting the nursing dimension, the rate of problem responses was higher for infants with high MRCI values, compared with their peers with low scores. These differences reached statistical significance for 8 of 9 questions, as well as the summary nursing dimension problem scores (P < .05). Even after controlling for birth weight, length of stay, and team size, we found that higher MRCI values remained a significant predictor of problem responses for the 3 questions found to be significant in univariate analyses. Of note, although physician (33% vs 24%) and neonatal nurse practitioner (29% vs 24%) problem scores were higher among patients with higher MRCI values, these differences did not reach statistical significance. This suggests that the MRCI value for nursing handoff networks specifically measures an aspect of nursing care that is important to families’ perception of nursing care.

6.5 Discussion

To our knowledge, these analyses represent the first large-scale application of network analytic methods to quantitative assessment of health care teams at the patient level. Many of these metrics have immediate face validity for clinicians, and we showed that one, the MRCI score, has a strong relationship with family satisfaction with care, which is an important measure of care quality. In addition, we demonstrated that information drawn from an EHR can be used for efficient collection of data needed for these analyses.

Our analyses suggest that the pattern of nursing assignments for an individual patient is strongly associated with parental satisfaction. Families that experience shorter intervals between care by familiar faces express higher levels of satisfaction with nursing care. Certainly, our results regarding the large size of teams encountered by NICU families and the persistent introduction of new nursing faces over the
Figure 6.6. **Nursing dimension problem scores.** The relationships between MRCI values and rates of problem responses for each of the 9 questions contained in the nursing dimension of the Picker NICU Family Satisfaction Survey. Item 10, denoted as the RN problem score, is the combined count of problem scores from the 9 nursing dimension questions. Patients with MRCI values that exceeded the median were considered to have high values.
course of NICU hospitalization are of interest and likely to be of concern to clinicians and families. However, it is somewhat reassuring that the teams for more immature infants continued to demonstrate relative centralization, with a core group of nurses providing care, and that team size and proportion of newcomers were not related to family satisfaction in the study cohort. In our analyses, the structure of the nursing team was more important than simply its size. Therefore, interventions designed to optimize the structure and timing of nurse participation in a patient’s care, such as with a core nursing group, would be beneficial.

The lack of relationship between parental satisfaction and measures such as team size and proportion of newcomers should not be taken as evidence that these measures are unimportant. These metrics have been found to be valuable predictors of success in other types of collaborative teams [49, 137], and such characteristics may play a role in determining the quality of care in domains other than that measured in this study. For example, these metrics may influence the quality of handoffs between clinicians. Given the role of handoffs in the genesis of preventable harm, further quantitative investigation into the relationship between care team structure and preventable harm is warranted.

It is important to note that these initial analyses were performed in a single, highly specialized, care area and examined care networks consisting of only one professional discipline. Although these results are likely to be generalizable to other locales and disciplines, examination of these methods across the continuum of care is needed to test this hypothesis and to identify modifications in approach that may be needed in other areas. In addition, these analyses used data drawn from only one source within the EHR, the NICU electronic documentation system. Including other sources, such as computerized provider order entry, medication administration records, ePrescribing, and picture archiving and communication systems would
expand the ability to characterize the care team across many disciplines and might prove valuable.

6.6 Conclusion

Taken together, our results suggest that simple measures such as team size are not sufficient to characterize health care teams at the patient level. More complex constructs, drawn from the field of network analysis, that assess both team structure and the sequence of participation for individual team members provide more robust characterization of health care teams. Data from EHRs can support the application of such methods in health care settings. The combined use of network science methods and EHR data is likely to enhance efforts to monitor and to improve team function and may be a valuable tool in other types of quality-improvement endeavors.
CHAPTER 7

CONCLUSION

In this section, we summarize the contributions of this dissertation and describe some interesting directions for future work to improve or expand the value of our tools and experiments.

7.1 Summary

We began with an introduction to terminology and relevant tools in Chapter 2. In Chapter 3, we built and analyzed a multi-relational network of disease connected by genetic similarity and patient comorbidity. We used existing network tools to make a number of observations about the dependency between shared genetics and clinical observations. We introduced MRLP, a novel multi-relational link prediction method, and attempted to predict additional disease pairs with an unobserved genetic relationship. We found that the patient data added value to the inference task, especially for diseases with relatively few known genetic associations.

In Chapter 4, we extended the disease-disease network into a disease-gene network and added relationships representing protein-protein interactions, protein sequence similarity, and genetic expression correlation. We also modified MRLP into a supervised classification problem to better handle the variance, imbalance, and complexity of link prediction in our network. Using this network, we exhaustively explored the value of each type of biological data for disease-gene candidate de-
tection. Adding additional edge types was beneficial in all combinations. Gene expression correlation was by far the most valuable data. With all data types combined, phenotypic information was the least valuable, but still improved performance. Finally, we observed that overlapping edges contain similar information whether considered separately or as a compound type.

Chapter 5 is an extension of previous work on CARE, which generates individual disease risk profiles using patient medical histories. We incorporated temporality in the collaborative filtering framework, making the system feasible for long term data. We also performed a number of focused studies which provide insight into the effectiveness of CARE under various conditions. These studies also yielded information about the properties of diseases and patients that does not capture well, which can be used to guide future improvements.

In Chapter 6, we demonstrated the value of network methods for answering real clinical questions. Our goal was to explore the relationship between the organization of the nursing team and the perceived quality of care in a large tertiary neonatal intensive care unit. We constructed individual networks describing the care team of each patient. We found that many commonly measured topological properties were not correlated with quality of care, but a novel metric describing the average loop length was significantly correlated with the perceived care. This observation is directly applicable to improving the experience of families with children in the NICU, and may be extensible to improving patient outcomes as well.

7.2 Discussion

Throughout this dissertation, we have demonstrated that the intersection of a network representation and data mining tools is beneficial across a wide spectrum of medical applications. Through our studies, we have made a number of contributions
to each of the fields in this intersection. Although developed around domain-specific
problems, many of the methodologies are generalizable to other applications both
within and beyond biology and medicine. MRLP and CW-MRLP are applicable
to any multi-relational networks, and can also be used for related inference tasks
like node label prediction by reversing the role of nodes and edges in the equations.
The length-to-repeat property captured by MRCI may be a useful measure in other
networks with time-ordered edges. The best-match mechanism we built into vector
similarity is useful for weighting pairs in any problem with long term data covering
multiple distinct phases.

To biology, we have contributed a number of observations and measurements
about important interplay between diverse data sources which have often been stud-
ied in isolation. CW-MRLP shows promise for the task of disease-gene candidate
detection. Directly incorporating patient data into the disease-gene studies was also
quite novel, and we found that this dimension does add useful data to the inference
task. On the clinical side, CARE is a novel proof-of-concept of the value of collabor-
orative filtering for preventative medicine. Finally, our work in Chapter 6 yielded
actionable observations about the relationship between the nursing schedule and the
perceived patient experience.

7.3 Future Work

The studies presented in this dissertation reflect just a tiny piece of the poten-
tial for network-centric data mining to benefit biomedical discovery, improve clinical
safety and efficacy, or explore other medical questions. Medically relevant knowl-
edge and observations data are constantly being expanded and revised. The power
of inference tools increases with the quality of the data, but both intellectual and
computational challenges increase with the size and complexity of the data. The
avenues for potential work are myriad. We limit our discussion to three broad categories: expanding the breadth and depth of the data, methodological improvements, and additional performance evaluation.

The disease-gene network is much complete than the disease-disease network, which is in turn far more expressive than the homogeneous representations of data. Regardless, two object types and a handful of diverse biological measurements are a drop in a bucket compared to the real molecular system underlying disease. The data representation can be incrementally improved and expanded to be as expressive as possible while still manageable. Similarly, the model should also continue to evolve. In particular, incorporating node and edge features may greatly increase the expressiveness of CW-MRLP.

Link prediction in the disease-gene network is not inherently limited to gene-disease association. MRLP or related methods have interesting potential for inference on other incomplete dimensions of biomedical data such as protein interactions or gene function.

As presented in Chapter 5, CARE is limited to ICD-9 data and temporal data, but the underlying collaborative framework has no such limitation. While it is an advantage that our system doesn’t require test results or special information, it would be naive to ignore these advanced results when they are available. CARE could exploit this information through similarity metrics which are appropriately modified for more complex representations of medical history. Also, the current formulation of CARE is unsupervised, so more flexible collaborative filtering methods should be explored.

Using the temporal data exploited by our time sensitive approach, CARE could be extended to predict the time of expected disease diagnosis in addition to the likelihood of occurrence. Such a mechanism is not well suited to our data, since
inpatient visits are fairly sporadic and may include diagnoses which do not relate to the timing of the hospitalization. However, in a database providing a more complete medical picture, this functionality could be an additional guide for scheduling of future checkups, screening, and tests.

The most obvious next step for our NICU study is to perform similar experiments on patient outcomes data. If significant trends could be isolated in this data, there is real potential for improved patient safety through a more principled scheduling approach. Also, models describing the dynamics of hospital employees may have applicability to other clinical problems like infection transmission.
BIBLIOGRAPHY


[18] Centers for Disease Control and Prevention: National diabetes surveillance system.  


