Social Approaches to Disease Prediction

by

Mehrdad Mansouri
B.Eng., Sadjad University Of Technology, 2011

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Computer Science

Social Approaches to Disease Prediction

by

Mehrdad Mansouri
B.Eng., Sadjad University Of Technology, 2011

Supervisory Committee

Dr. Ulrike Stege, Co-Supervisor
(Department of Computer Science)

Dr. Panajotis Agathoklis, Co-Supervisor
(Department of Electrical and Computer Engineering)

Supervisory Committee

Dr. Ulrike Stege, Co-Supervisor
(Department of Computer Science)

Dr. Panajotis Agathoklis, Co-Supervisor
(Department of Electrical and Computer Engineering)

ABSTRACT

Objective: This thesis focuses on design and evaluation of a disease prediction system that be able to detect hidden and upcoming diseases of an individual. Unlike previous works that has typically relied on precise medical examinations to extract symptoms and risk factors for computing probability of occurrence of a disease, the proposed disease prediction system is based on similar patterns of disease comorbidity in population and the individual to evaluate the risk of a disease.

Methods: We combine three machine learning algorithms to construct the prediction system: an item based recommendation system, an bayesian graphical model and a rule based recommender. We also propose multiple similarity measures for the recommendation system, each useful in a particular condition. We finally show how best values of parameters of the system can be derived from optimization of cost function and ROC curve.

Results: A permutation test is designed to evaluate accuracy of the prediction system accurately. Results showed considerable advantage of the proposed system in compare to an item based recommendation system and improvements of prediction if system is trained for each specific gender and race.

Conclusion: The proposed system has been shown to be a competent method in accurately identifying potential diseases in patients with multiple diseases, just based on their disease records. The procedure also contains novel soft computing and machine learning ideas that can be used in prediction problems. The proposed system has the possibility of using more complex datasets that include timeline of diseases, disease networks and social network. This makes it an even more capable platform

for disease prediction. Hence, this thesis contributes to improvement of the disease prediction field.

Keywords: Disease Prediction, Comorbidity, Machine Learning, Predictive Model

ACKNOWLEDGEMENTS

I would like to thank Prof. **Ulrike Stege** and Prof. **Pan Agathoklis** for their guidance in my research. In addition, I would also like to thank my family, **Maral**, **Gita** and **Bahman** for supporting me through my education.

A mind is fundamentally an anticipator, an expectation-generator. It mines the present for clues, which it refines with the help of the materials it has saved from the past, turning them into anticipations of the future. And then it acts, rationally, on the basis of those hard-won anticipation.

Daniel Dennett

Contents

List of Figures

List of Tables

Abbreviations

CP	**C**onditional **P**robability
DCPN	**D**isease **C**ontrol **P**riorities **N**etwork
DPS	**D**isease **P**rediction **S**ystem
ER	**E**xpectation **R**atio
FN	**F**alse **N**egative
FP	**F**alse **P**ositive
ICF	**I**tem-based **C**ollaborative **F**iltering
ICI	**I**ndependence of **C**ausal **I**nfluences
IG	**I**nformation **G**ain
JI	**J**accard **I**ndex
ODE	**O**rdinary **D**ifferential **E**quation
PCA	**P**rinciple **C**omponent **A**nalysis
PGM	**P**robabilistic **G**raphical **M**odel
RBF	**R**adial **B**asis kernel **F**unction
ROC	**R**eceiver **O**perating **C**haracteristic curve
RR	**R**elative **R**isk
RS	**R**ecommendation **S**ystem
SGM	**S**i**G**moid **F**unction
SMC	**S**imple **M**atch **C**oefficient
WHO	**W**orld **H**ealth **O**rganization

Symbols

n	$\in$	$\mathbb{Z}^+$	Number of all diseases
s	$\in$	$\mathbb{Z}^+$	Number of diseases backed by evidence
r	$\in$	$\mathbb{Z}^+$	Number of predicted diseases
D	$=$	$\{d_i\}_n$	Set of all possible diseases
D^*	$=$	$[0\ \ 1]_{n\times 1}$	Vector of actual state of diseases
e_i	$\in$	$[0\ \ 1]$	Prior evidence of disease i
E	$=$	$[e_i]_{n\times 1}$	Vector of prior evidence about diseases
P	$=$	$[p_i]_{n\times 1}$	Vector of probability of diseases
R	$=$	$\{d_i\}_r$	Set of predicted diseases
N	$\in$	$\mathbb{Z}^+$	Total prevalence of all diseases
N_i	$\in$	$\mathbb{Z}^+$	Prevalence of disease i
N_D	$=$	$[N_i]_{n\times 1}$	Vector of prevalence of diseases
N_{ij}	$\in$	$\mathbb{Z}^+$	Prevalence of disease i and j simultaneously
$\bar{N}_X$	$=$	$N - N_X$	Prevalence of complement of disease set X
sim_{ij}	$:$	$\{d_i, d_j\} \to \mathbb{R}$	Similarity between disease i and j
SIM	$=$	$[sim_{ij}]_{n\times n}$	Similarity matrix of all disease
pt_l	$\in$	$\mathbb{R}^+$	Necessary threshold for probability of a disease to be recommended
pt_h	$\in$	$\mathbb{R}^+$	Sufficient threshold for probability of a disease to be recommended
S	$\in$	$\mathbb{R}^+$	Cost function of the recommender
α	$\in$	$[0\ \ 1]$	Conservativeness factor of the recommender cost
L_0	$\in$	$\mathbb{R}^+$	Laplacian bias of conditional probability
w	$\in$	$\mathbb{R}^n$	Vector of weight parameters of ICI model
s	$\in$	$\mathbb{Z}^+$	Number of reported diseases
ED	$\in$	$\mathbb{Z}^+$	Number of recommended diseases
HD	$\in$	$\mathbb{Z}^+$	Number of hidden diseases

FP	$\in$	$[0\ \ 1]$	False positive rate of disease prediction
FN	$\in$	$[0\ \ 1]$	False negative rate of disease prediction
A	$\in$	$[0\ \ 1]$	Long-run prediction accuracy

Operators

X^T	Transpose of a matrix or a vector
$\sum_{i=X}^{Y} f_i$	Summation of elements of a series f from X to Y
$\prod_{i=X}^{Y} f_i$	Product of elements of a series f from X to Y
RPCa (X)	Reduced version of a matrix X by principal component analysis
SPd (X, Y)	Sparse Production of two matrices X and Y
Sum (X)	Sum of the elements of a matrix X along the first dimension (row)
$X \cap Y$	Intersection of two sets X and Y
$X \cup Y$	Union of two sets X and Y
$\in X$	A member of set X
$\propto X$	Is proportional to value X
Var (X)	Variance of a random variable
Cov (X)	Covariance of random variables X and Y
$\hat{X}$	Expected value of a random variable X
Exp (X)	Exponential function of variable X
Log (X)	Logarithmic function of variable X
Sgm (X)	Sigmoid function of variable X
Max_Y (X)	Y largest elements in a vector X
ASum (X)	Absolute sum of elements of vector X
RMS (X)	Root mean square of vector X
Sup (X)	Least upper bound of elements of vector X
$\parallel X \parallel_R$	R-Norm of of vector X

Chapter 1

Introduction

Imagine an automated system examines you and informs you that you will probably have a certain disease. It then advices you of the proper way of reducing its chance of occurrence and effects. This ability of predicting future or potential diseases of an individual and preventing it has always been one of the dreams of medical sciences, and is a crucial step for personalized medicine to revolutionize healthcare. The realization of such a mechanism will ultimately help us to protect ourselves from diseases more effectively and stop these main sources of human's suffering and death.

In order for this to become a reality, multiple layers of complex data analysis are needed to find a variety of reliable patterns from a vast amount of medical data. The goal of this thesis is to propose an automated mechanism for disease prediction based on individual records of previous diseases. Specifically, we define disease prediction as the capability of predicting upcoming diseases in an individual, based on the available information about her internal and external world. In this research we will investigate how patterns in disease records of an individual can be used to estimate risk of emergence of the individual's future diseases.

The co-occurrence of a set of diseases in different individuals is called comorbidity. Comorbidity can be caused by either causal effects or correlations, or a combination of both. Correlational comorbidity of diseases can be due to similarity in the individual's genetic roots, environmental factors or life style risk factors. Causal comorbidity of diseases takes place when one disease systematically produces another disease or increases its chance of occurrence by indirectly affecting the body. In this thesis we will consider both the causal comorbidity and the correlational comorbidity.

This thesis proceeds as follows. First, in Chapter 2 we describe the motivation for approaching the problem of disease prediction using social scale data. Chapter 3 reviews different disciplines in which this problem has been tackled, from epidemiological methods to data mining techniques, and addresses existing limitations.

Chapter 4 consists of stages of designing an automated disease prediction system. We first introduce a dataset of disease comorbidity extracted from patient records across the U.S. We then propose multiple similarity measures for our problem and finally design our three layers of our prediction system: (1) an item based recommendation system, (2) an ICI graphical model and (3) a rule based recommender.

In Chapter 5, we evaluated the quality of the prediction system and discuss how to set parameters of the proposed system. In this chapter we also compare the quality of our prediction system to a typical recommendation system and show the improvements of prediction if we use the system for a specific gender and race. Chapter 6 contains contributions and potential applications of this study and discusses possibilities for future works. Finally, Chapter 7 provides a summary of the thesis.

Chapter 2

Motivation

2.1 Introduction

Recent trends in public health studies suggest that, in order to achieve an improved quality of medical care of the individual, we need to look beyond conventional analysis of diseases based on only the individual's data, and study diseases in the context of society as a whole [1]. In the following sections we discuss different layers of reasoning for studying diseases based on social scale data. We first look at possible mechanisms of influence of social factors in disease patterns. Next, we show why this relation is statistically significant and worth studying. Then we argue why statistical models based on social variables may be good models for predicting diseases.

2.2 Social influence on diseases

There are many studies that propose a correlation between a socially related factor (such as income group, social class and residence) and an indicator of health quality (such as mortality rate) [2]. Many of these studies, however, lack a firm evaluation. As a result, it is not surprising that after a careful and unbiased experiment results of them are either unrepeatable or have insignificant magnitude): In a large portion of the remaining valid studies, unfortunately, the correlation is not causal and is susceptible to many statistical fallacies such as endogeneity [2]. This makes the result inaccurate as well as unfalsifiable. To make the matters worse, the underlying mechanisms that produce these relations are

often too complex, vague or unknown. However, in recent years there have been reliable studies that show causal and countable correlational relations between social structure and the medical well-being of individuals, in addition to models for the mechanisms that generate these relations [3]. Below we summarize these areas of achievements that are relevant to the thesis.

2.2.1 Social contagion

Since 2002, a series of interdisciplinary studies in the field of social network analysis have been done, on the propagation of traits in social networks, mainly by NA.Christakis and JH.Fowler. These studies show strong relationships between some of the traits and behaviors of individuals and the people who they are connected to in the social networks. This phenomenon, called social contagion, happens both as clustering of traits of individuals with similar global position in the network and clustering of traits in local neighborhoods [4–18].

Christakis and Fowler have reported contagion of a wide range of mental [4–6] and physical [7–9] health problems, and medically relevant conditions and behaviors [10–14]. They offered three explanations for the contagion: (1) textithomophily, which occurs when the subject has the tendency of associating with others exhibiting similar traits; (2) textitcovary, which occurs when the subject and its contacts are jointly influenced by an omitted variables or shared context; and (3) textitinduction, which occurs when the subject is influenced by its contacts [19–22]. Recent works claim that all three mechanisms may involve in medically related processes [3].

2.2.2 Human genetic clustering

Beyond the contagion of phonotypical attributes, there are hypotheses of correlation between genes of contacts in the social network [23, 24]. An important point about these recent results is that this correlation cannot be completely explained by the confounding effect (i.e., similarity as the result of a hidden variable or processes) as originally expected. There may exist causal factors in clustering of genes in populations (i.e., similarity as the result of a direct process between individuals) [3, 23].

This causality can be hypothesized in two ways. First, a bottom-up mechanism in which the tendency of locating of subjects in certain parts of the network is enforced by genes.

This can be explained by an evolutionary adaptation of individuals to certain configurations of the network. This idea itself is usually represented as hidden rules in the functional sociology, for example the emergence of "tit for tat" in the prisoners dilemma in local networks [25–27].
The second and maybe more important interpretation of the observed causality can be described by a top-down process in which the social network influences the local and global pattern formation in the social structure. To prevent misinterpretation, it is worth noting that the notion of causality here is not the basic physical determinism, but as a one-way statistical dependence between parameters of the social network model.

2.3 Importance of influences on disease

We already discussed how society may influence traits of individuals, from the small scale to the large scale. Now we show statistical evidences of significance of this effect on the well-being of individuals and argue why it is crucial to study diseases socially. We will argue that not only socially related diseases are the main source of preventable casualties around the world, but also they are becoming more challenging for traditional approaches to predict.

2.3.1 Scale of socially related diseases

Demographics show a strong impact of socially related diseases on the preventable causes of death. The World Health Organization (WHO) has provided a list of the leading causes of death in 2008, in which socially related preventable diseases were forming 34.7% of total worldwide deaths [28, 29]. Other studies approximate that half of 10.4 million deaths among children under age 5 in 2004 were due to four preventable and treatable communicable diseases [30]. Similar behaviors can be observed even more significantly in high income countries including US [30]. According to the disease control priorities network (DCPN), most of these causes of death can be declined dramatically by providing a good understanding of their behavior in large scale. These results demonstrate the importance and priority of studying preventable fatal diseases in a social scale.

2.3.2 Dominance of complex diseases

The traditional approach where diseases are based on only the current status of the patient is insufficient for complex emerging diseases [3, 31]. Increase of life expectancy and improvements in health services in recent decades has shifted the major death factors from famine and bacterial epidemics, toward mutation in human genes and more complex viruses such as cancer and HIV [28, 32]. The emergence and evolution of these complex diseases has been highly dependent on interconnections in populations, and by increase of population and global communication, the role and complexity of social transmission of diseases will increase. Therefore, the policy of concentration on details of metabolic behavior of the patient and ignoring information about diseases in the society will be less and less effective, and the necessity of including the social behavior of the patient will be more vivid [3, 33].

2.4 Effectiveness of social models

In this section, we argue for effectiveness of predictive models of diseases based on relatively simple and social indicators of diseases such as comorbidity of diseases over a population. In the next sections, we suggest why with a small number of large scale indicators one can model a large amount of complex interactions of an individual and why this top-down approach should be the dominating strategy in modeling complex processes like disease progression.

2.4.1 Simplicity emergence

One of the most important properties of complex systems is the birth of new patterns from the interactions of their parts in the smaller scale, called emergence [34]. If the system patterns contain more information than the sum of its parts, it is called complexity emergence, and if it contains less information than the sum of its parts, it is called simplicity emergence [34]. Many studies have been done on complexity emergence, which are usually more frequent and noteworthy [35–37].

A good analogy for simplicity emergence in our social system is the thermal behavior of gas particles in response to heat. By heating, the complex microscopic movement of

massive amounts of particles that have rapid and interdependent dynamic can macroscopically be modeled by a set of linear equations [37].

The human social network is certainly a complex system and many researchers have studied its complex behaviors for many applications [38–40]. From a complex system perspective, what makes the statistical analysis of comorbidity so interesting is the use of simplicity emergence in moving from an individual scale to a social scale.

In other words, the microscopic biological processes that in reality cause the diseases are too complex to model, but prediction of diseases based on their realization in comorbidity across a population may be possible by using simple and elegant models. There is still a long way to achieve the full capacity of this methodology, but results up to now show that many social traits follow simple patterns [3].

2.4.2 Promises of Structuralism

In recent decades two opposite schools have been dominant in facing complex systems and scientific discipline as a whole. On the one side is the bottom-up attitude, mainly inspired by Skinnerian behaviorism [41] and discipline of artificial intelligence, which seeks to model systems as set of distributed self-adaptive agents that from raw random initial states obtain properties of the true complex system (such as self-management in chaotic environment) by simple reinforcement rules.

On the other side is the top-down attitude, based on structuralism and functionalism doctrines. It is defined in contrast to the first approach by arguing that if agents of a complex system (in our case, humans and diseases) were blank slates, environmental factors wrote on them, they would be impoverished systems [42]. In other words, the lack of stimulus due to the limited interactions, requires that the system has some prior mechanisms that present the existing enrichment of its dynamic.

Above yields the following hypothesis: the attributes and patterns are the result of unfolding of genetically determined programs and social structures [42]. It follows that the basic structure of behavior is simply determined by both the initial state of the system and the fundamental relation laws applied to certain large scale social patterns. Therefore, our task as scientists is to determine what are those laws and what are the fundamental principles behind them.

A systematic realization of this idea is the frequentist statistical analysis [43], in which the analysts find the stochastic rules based on the mathematical correlations in attributes

of subjects in a certain population.

Some unresolved theoretical criticism exists against structuralism, mainly under the post-structuralism ideas by Michael Foucault [44], Martin Heidegger [45] and Slavoj Zizek [46]. In practice, however, empirical results support the structuralism approach, and in various disciplines of sciences, applied sciences, medical sciences, social sciences and anthropology it is reemerging as a dominant methodology.

Methods with structuralism themes have been used by many recent top thinkers in various contexts, including Steven Pinker [47], Noam Chomsky [42], Daniel Dennett [48]. In addition, from the perspective of scientific evaluation, Occam's razor suggests that a theory that can predict results statistically from earlier stages is more interesting and useful than a model based on adaptive uncertain noisy chain of events with many sensitive parameters. In summary, from a demographical viewpoint, it seems that we are not adaptive agents of the Nash equilibrium (behaviorism) but genetic "inputs" to a social structure "function" (structuralism). It should be pointed out that the writer, similar to many, considers structuralism more as a methodology than a worldview. Jean Piaget puts it nicely that "there exists no structure without a construction, abstract or genetic" [49].

2.5 Summary

In this chapter we proposed multiple motivations for applying disease prediction systems in a social scale. We first introduced studies that show influences of society on diseases through social contagion and possibly human genetic clustering. We then illustrated the impact of these influence on diseases through their share in mortality rates around the world. Finally we argued why we think emergence property and structuralism viewpoint suggest that we can achieve valid prediction and control models of diseases of a population.

Chapter 3

Literature review

3.1 Introduction

Since the first systematic attempt to quantify causes of death in 1662 by John Graunt [50], the problem of assessment of risk of diseases has been tackled in various disciplines. Disease risk assessment is the systematic and quantitative evaluation of risk or time of a disease or symptom based on certain risk factors. Disease prediction is defined as the prediction of incoming diseases of an individual in a specific period of time. It is also worth mentioning that from this perspective disease prediction is an extension of disease risk assessment due to its potential need of assessment of risk of all possible diseases, although in practice, majority of techniques apply a full analysis on only potential diseases. Our focus will be on disease prediction, although some of the techniques can also be considered as disease risk assessment.

Due to explosion of medically related data and the steady increase of computational capacity, disease prediction has attracted increasing attention in the recent decades, from both medical science and computer science communities. These disease prediction studies try to extract underlying patterns of diseases in the environment, genes and lifestyle risk factors of individuals. In the following, we classify and review the various contexts in which researchers have tackled the problem of disease prediction, recent progresses and represent the gaps and possibilities in the area.

3.2 Epidemiology

Epidemiology is probably the oldest approach in predicting diseases. Although Epidemiological models are usually designed for estimating propagation of a transmittable disease in a population, results of such estimations can sometimes be used for estimation of risk of the disease transmission to a specific individual and ultimately evaluating whether risk of the disease for the individual is feasible in future or not.

For decades, ordinary differential equations (ODEs) were the standard model for describing phases of transmission a disease in a population [51]. This was natural since the propagation of a disease can be simplified as transition between sub-populations. In this simplified outlook, population is partitioned into different compartments, called states, each representing a specific stage of the epidemic. These states act as variables and interact over time by a fixed set of rules. These rules that govern the transition rates between states mathematically are best expressed as a set of ODEs [51].

Specifically, a more common form of this family is the SEIR model which is a set of bilinear ODEs compartmenting population into susceptible population (S), exposed population in latent period (E), infected population (I) and removed population immune to reinfection (R). This model has been improved by taking more factors into account such as natural birth and mortality rate [52], disease survival rate [53], post infection sates (carriers and resistant population) [54], vertical Transmission (maternally immune and inherited individuals) [55], controls (vaccination and isolation) [56], vectors (agents transmitting the pathogen) [55], lurker delay [57], fractional orders [55] and nonlinear elements [58]. Recent proposed structures for describing these transitions between states as usually discrete systems such as cellular automata [59], Kinetic Monte Carlo [60], Hidden Morkov models [61] and graph based models.

An epidemic in practice is not a smooth exponential rise and fall as is expressed in these models, but usually contains complex dynamics and small fluctuations. Moreover population in practice is not "well-mixed", and disease has complex transmission pathways that are dependent on the structure of the network [62]. To represent these irregularities, one should either use stochastic variations or more sophisticated compartmental structures that simulate the underlying events of the epidemics. The nature of stochastic models is Bayesian, and they tend to describe the probability distribution of diseases and phenotypes in presence of a stochastic mechanism of exposure. These models are

especially useful when the temporal and geographical fluctuation of transmission is important, as in small populations [63].

Compartment models are deterministic mathematical structures that either simulate or model transmission of the disease or phenotypes among agents or subpopulations. An example of these methods is the web model, that is a network between incidences of occurrences of a disease based in proximity of their time and location [64, 65]. These models have been successful in predicting and proposing controlling policies for SARS [66] and Influenza [67].

3.3 Disease network

A disease network is a network model of the relation between diseases and a factor. The factor demonstrates one aspect of the underlying mechanism that causes diseases and can be genes, protein interactions, enzyme mutation, clinical history of patients or other phenotypes. Various disease networks have been constructed from genetic networks [68, 69], proteomic networks [70, 71] and metabolic networks [72] to phenotypic networks [73]. Although, to the best of our knowledge, these studies have not yet been used directly for disease prediction of an individual, they give a new perspective about potential relations between specific diseases and factors, and have strong potentials for using them in the disease prediction process. We believe that a practical prediction model that use a combination of disease networks on decision making, can make a breakthrough in disease prediction field.

There are two representations to show the links between diseases and factors: (1) the binomial network in which both the elements of the factor and diseases are nodes and the edges show the sheer existence of the relation between them, (2) the diseasome in which the nodes correspond to the diseases and edges represent elements of the factor, which are shared between every pair of diseases. For the purpose of risk assessment, binomial networks are needed to be reduced to the diseasome. This is because (1) connections of factor are redundant and only increase the complexity and (2) the patterns of clusters and paths among diseases are not clear in the binomial tree.

There are multiple potential applications for disease networks: (1) visualizing medical health records, (2) studying the disease evolution of patients, (3) identifying key diseases (e.g. highly connected, precedent) for health care policy, (4) integrating phenotypic data

with genetic and proteomic data to better elucidate disease etiology, (5) risk assessment of diseases for patients, and (6) determining whether differences in the Comorbidity patterns expressed in different populations indicate differences in biological processes, environmental risk factors, or health care quality provided for each population.

3.4 Data Mining

Like many other areas of research, disease prediction has been changed by data mining and machine learning systems and approaches. By shifting longitudinal studies and censuses sophisticated data gatherings to extraction of the explosion of data available through online social networks [74], websites and interactive applications [75], data mining methods are becoming a key approach to both statistical and clinical medicine. Some implementations have even claimed to be able to compete with medical doctors in both precision and coverage.

Although there is a long history of risk assessment of disease, only recent researches have attempted to predict diseases in the context of data of the population. Some researchers applied the Support Vector Machine (SVM) [76, 77] and Associative Classification [78, 79] as a high dimensional classifier to the medical record data. There are also studies that have used the null space class selection ability of recommender systems [80]; while the majority of papers have tried to find nonlinear patterns in the data using heuristic and AI methods such as neural networks [81, 82] and genetic algorithms [83].

It should be pointed out that we separated the subfield of relational learning methods such as probabilistic graphical models and graph based algorithms form the general machine learning techniques. We did this to be able to zoom in and give a better perspective to these methods for the sheer importance of their network structure in disease prediction.

3.5 Social network analysis

Many researchers have tried to apply ideas of social network analysis to the context of medical conditions. Some of the most outstanding results are from the studies of N.A. Christakis and J.H. Fowler. They studied contagion of a wide range of phenomenon not just by applying traditional egocentric metrics but offering Socio-centric and global

Social network analysis methods. Their results ranges from physical risk factors and diseases such as obesity [7, 9], smoking [10], food consumption [11], influenza [8], alcohol consumption [14], drug use [12] to mental phenotypes such as happiness [5], loneliness [4], depression [6], divorce [18] and sleep loss [12]. A summary of their results has been published in their book "Connected" in 2009 [84] and two reviews of their works in 2008 and 2012 [3, 85].

The most influential point they made is the idea of "3 Degree of Influence" in longitudinal social networks, which in a nutshell, claims based on various empirical cases that on average, there is a statistically significant and substantively meaningful relationship (correlation in traits) between the ego and alters up to Geodesic distance (i.e., the number of steps taken through the network) of three, before it could plausibly be explained as a chance occurrence. Using this Idea they tried to explain the correlations of traits among close individuals by "Droplets of Epidemic" (distributed clusters).

3.6 Graphical Models

A probabilistic graphical model (PGM) is an extension of machine learning methods that are specifically designed to deal with relations and dependences between variables. Variable dependencies are modeled on a graph called dependency graph. Mathematically, given the dependency graph and conditional probability function of each node given its contacts, one can compute the probability of every even in every node.

PGM can be classified into two types based on the constraints they put on the dependency graph. If the dependency graph is limited to directed acyclic graphs, it is called Bayesian network. Otherwise if the network can have cycles but is limited to undirected graphs, it is Markovian network. Various Markovian and Bayesian models have been proposed to tackle more general graphs and also to reduce the complexity of the algorithms.

In 2004 [86] proposed the relational Markov networks to model cross dependencies. Two years later the Markov logic networks [87] and relational dependency networks [88] improved the learning method in various aspects. The successful use of the PGM model for disease prediction can be traced back to [89], yet later attempts are still limited to theoretical and analysis and have not been used widely in medical institutions.

The PGM researchers lack the intension to accumulate other prior data rather than the

social network data, for instance known disease genes or symptoms which seem to be best coded in disease networks. Studies also lack the statistical spirit of dealing with complex systems and have took a more deterministic record learning approach which can cause the over-fitting problem and may not be able to generalize to new cases and environments.

3.7 Practical Limitations

One basic epistemological question that might be raised is that if such essential need and theoretical potential exists for this topic, why it has not been studied comprehensively yet. In this part some explanations are offered:

3.7.1 Capacity of computation of social data

Sociometric and network science studies show that in the last decade, the raise of "social sensors" that produce social data automatically and in large scale, and also increase of the computational limits that has reached to the edge of providing realistic simulations of social network, can promise computer based studies of social systems [74, 90, 91].

3.7.2 Era of scientific social sciences

In the recent decades, a new wave of analytical and empirical researchers started to implement quantitative theories and statistical models for social and medically related behaviors and shift the health information paradigm from the ideological hypothesizing to rational investigations [90]. This work tries to play its role in expanding this approach.

3.8 Summary

This chapter covered areas of research in the literature that have the potential to model diseases of a population. We first reviewed how epidemiological models like SEIR can model the propagation of a disease in a population and what various add-ons to these models represent. Disease network models were introduced that capture the interconnected relationships diseases and an aspect of its realization such as genes, proteins and

phenotypes. We showed groups of data mining techniques that have been used to model risk of a disease in an individual and a stream of social network analysis researchers that have attempted to show and model contagion of a disease in social network. We finally summarized how probabilistic graphical models were used to find patterns of variables in a network and the possibility of using them for predicting diseases in a population.

Chapter 4

Disease Predictor

4.1 Introduction

In this chapter, we propose a multi-layer disease prediction system based on the comorbidity of diseases in a population and discuss the technical possibilities of such system. In Section 4.2, we review a disease dataset from [73] that includes comorbidity of diseases in a population and study statistical properties of the data and its limitations. In Section 4.3, we define the problem of disease prediction for individuals. In the remaining sections, we introduce a machine learning algorithm that applies three stages of analysis over the co-occurrence records of diseases (see Figure 4.1): In Section 4.4, we design the first block of the prediction system: a recommendation system that generates recommendation probability of diseases based on the disease dataset and the disease record of the patient. In Section 4.5, we design the third block of the prediction system; a threshold based recommender that outputs diseases that have prediction probability above an appropriate threshold. In Section 4.6, we insert a probabilistic graphical model between the recommendation system and the threshold recommender that maps the recommendation probability to prediction probability, in order to enhance the prediction accuracy. Basic statistical inferences and terminologies of this chapter are from the book Complex Social Networks, by F. Vega Redondo [92].

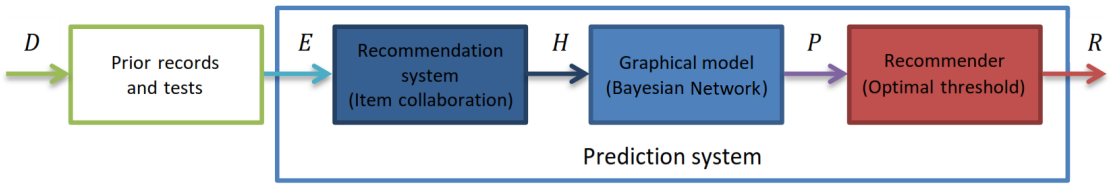

FIGURE 4.1: Stages of disease prediction algorithm

4.2 Data

4.2.1 Source and structure of data

In 2009, Hidalgo et al.[73] studied comorbidity, association and progression of diseases. For this they collected a clinical history of phenotypes of illnesses of 13,039,018 hospital inpatients from MedPAR records.

The original data used in Hidalgo et al.[73] consisted of 32,341,347 medical records of 96% of Americans of age at least 65 in the period of 1990 to 1993, totalling to 32 million individuals. Each record, in addition to the date of the visit and a primary diagnosis, included up to 9 secondary diagnoses, coded using ICD9-CM (available at `http://www.icd9data.com/` [93]). In this coding, diseases are defined as specified sets of phenotypes that affect physiological systems. Each disease is represented by 5 digits. The first three digits code the 657 main categories of the diseases, while the two remaining digits code 16,459 sub categories with more specific information about the diseases. The authors from Hidalgo et al. [73] constructed a phenotypic disease network and did further analysis on the data. They derived prevalence and two-by-two co-prevalence of diseases from the records. Results are classified by race and gender. The data and its detailed description are publicly available at `http://hudine.neu.edu/` [94]. It is important to note that the prevalence measures the number of people with the condition rather than the incidences of that condition. Hence the prevalence is immune to the bias of multiple sampling.

4.2.2 Defects and Limitations of Data

We emphasize that in some of the records there may exist ambiguities in classification of some phenotypes and errors in the diagnosis of diseases. However because of robustness of our recommendation system, such noise typically does not change the statistical properties of the designed system.

In addition to above-65 age bias, the data exhibits a bias of gender and race with 58.28% female and 90.08% white patients. Because of this and biological differences between genders and races we designed a prediction system for every combination of race and gender, and compared the results across different races and genders (Chapter 5).

4.2.3 Frequency Representation

A dataset of clinical history usually has diseases of every patient as a separate instance. HuDiNe dataset however, only has prevalence as the number of occurrences of each disease in population.

$$d_i = [\quad ... \quad \{\in \mathbb{R}\} \quad ... \quad]_{N\times 1} \rightarrow N_i, N_{ij} \tag{4.1}$$

A miniature example of this compression with three diseases and six patients can be seen below:

$$E = \begin{bmatrix} 0.1 & \mathbf{0.8} & 0.0 & \mathbf{1.0} & 0.2 & \mathbf{1.0} \\ \mathbf{1.0} & \mathbf{0.7} & 0.1 & 0.3 & 0.0 & 0.0 \\ \mathbf{1.0} & 0.0 & \mathbf{1.0} & \mathbf{1.0} & 0.0 & \mathbf{1.0} \end{bmatrix} \rightarrow N = 6\,, \quad \begin{bmatrix} N_1 = 3 \\ N_2 = 2 \\ N_3 = 4 \end{bmatrix}, \quad \begin{bmatrix} N_{12} = 1 \\ N_{13} = 2 \\ N_{23} = 1 \end{bmatrix} \tag{4.2}$$

This compression of a vector of state of diseases to a prevalence number has two benefits: (1) It simplifies process of preparation of inputs and interpretation of results, especially for large scale analysis over a population. (2) Size of the data is reduced and more importantly does not scale with sample size, i.e. number of patients. But these benefits come with a cost. Because there is no instance (in our case patient) to make inference, available methods become very limited.

4.2.4 Statistical Properties of Data

As briefly noted in the original paper by Hidalgo et al. [73], both the disease prevalence and the disease co-prevalence express a power law distribution with an exponent of 2.45 and 2.41, respectively (Figure 4.2).

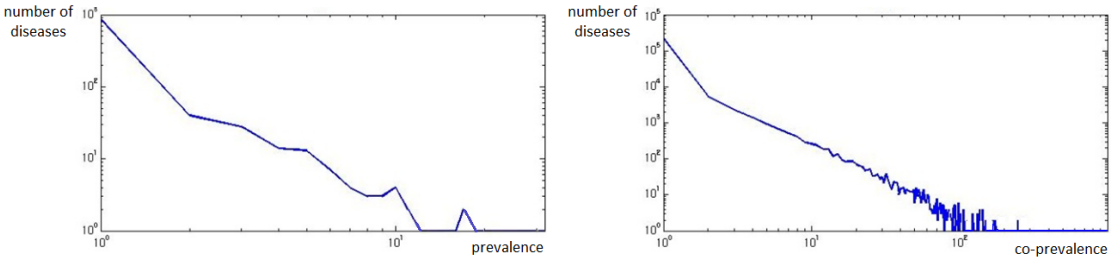

FIGURE 4.2: LEFT: Log-log plot of distribution of the number of disease prevalences. RIGHT: Log-log plot of the distribution of the number of disease co-prevalences. Both graphs show a close to linear pattern, which indicates a power law distribution.

The fact that both disease prevalence and disease co-prevalence behave according to power law distributions with an exponent of less than three has important consequences for our prediction:

(1) The distribution is heavy-tailed, hence only the weak version of the law of large numbers holds, meaning that with increasing number of samples, disease prevalence and disease co-prevalence converge to their average values at a slow rate [92].

(2) The variance of prevalence of diseases is divergent, the central limit theorem does not hold, and hence models that are based on the normal distribution of error are not valid.

(3) The distribution is scale invariant, meaning that by splitting each disease into specific sub-diseases or by merging diseases to a general disease family, the distribution and its properties remain the same.

In the rest of the chapter we will introduce a disease prediction system that can be applied to our prevalence and co-prevalence data.

4.3 Disease Prediction

Our objective is to design a disease prediction system (DPS) that can predict hidden or future diseases, based on the patient's known medical phenotypes and demographics

such as gender and race. Specifically, the task of the DPS for each patient is to find probable unknown diseases R from the set of possible diseases D, given the evidences E for the patient's real diseases D^*.

$$R = DPS\,(D) \qquad R \subset D \qquad E \leq D^* \tag{4.3}$$

In the following we propose our disease prediction system consisting of three stages (Figure 4.1): the recommendation system, the probabilistic graphical model and the recommender.

4.4 Recommendation System

Given the disease co-prevalence data, a recommendation system (RS) is a suitable choice for extracting probable diseases for two reasons:
(1) It learns which diseases to use and hence reduces the size of the sparse but high dimensional problem.
(2) It has been shown to be effective in filtering out noise, uncertainty and complex relations that are the main source of error in disease risk assessment [95].
To satisfy these objectives, we propose a recommendation system based on collaborative filtering.

4.4.1 Item-Based Collaborative Filtering

Item-based collaborative filtering (ICF) is one of the most successful families of RS. ICF initially models the similarity between items (in our problem, co-prevalence of diseases) [96]. It then assesses the probability of occurrence of every possible disease based on its weighted linear association with the existing diseases of the patient. ICF can be formulated as:

$$H_j = \frac{1}{\sum_{i=1}^{n} sim_{ij}} \sum_{i=1}^{n} e_i \; sim_{ij} \qquad \forall\, j = 1, \ldots n \tag{4.4}$$

In the above equation, H_j denotes the probability of recommending disease j, sim_{ij} is the similarity between diseases i and j (extracted from the data by a process described in the next section) and e_i denotes the prior evidence of disease i. Specifically, e_i is a value between zero and one, representing the certainty about the presence of disease d_i

and is determined by the user based on the knowledge of the patients medical condition. e_i is set to one if d_i is already detected with certainty, and e_i is set to zero if there is no evidence of disease i.

The ICF has some key advantages in mining disease comorbidity data over other recommenders and machine learning algorithms as pointed out in [95, 96]:

- Instead of analyzing the massive, complex and uncertain data of patients, only the comorbidity and association between diseases is required.
- Unconventional patients, who are not rare in medical systems, will get better recommendations [96].
- ICF is extendable to more complex similarity measures that take the joint distribution of sets of diseases into account.
- ICF generally has higher performance than user based recommenders on data that have many items (in our case diseases) [96]. There are other available RSs that can be effective for diagnosis, such as latent (regression) collaborative filtering methods [97], but these methods cannot be applied to our data since they require patient's records in order to be trained.

4.4.2 Compressed Model

In this section we discuss some additional mathematical tools that can make the training process of the recommendation system more efficient. We can generate a vector of recommendation probabilities of all of the diseases by representing equation 4.2 with a matrix equation (see graphic presentation in Figure 4.3):

$$H\left(E\right) = \frac{1}{Sum\left(SIM\right)} E^T \, SIM \tag{4.5}$$

Here, H is the vector of all H_j and denotes the probabilities of recommending diseases, SIM is the matrix of similarities of the diseases, $Sum(\cdot)$ denotes the sum of the elements of its input matrix along each row, E is the prior evidence vector with each row as a possible disease. Now we are able to apply principle component analysis (PCA) to

reduce the dimension and thus computation time and storage required for our RS:

$$H(E) = \frac{1}{Sum\left(RPCa\left(SIM\right)\right)} SProd\left(E,\ RPCa\left(SIM\right)\right) \tag{4.6}$$

Here, $RPCa\left(\cdot\right)$ denotes the compressed map of its input similarity matrix SIM using principal component analysis [98] and finally, $SProd\left(\cdot\right)$ gives the sparse product of its two inputs E and $RPCa\left(SIM\right)$ [99].

In addition to reducing computation, the compressed format also reduces the number of parameters of the model. This will reduce the risk of over-fitting, which is a common problem for complex models and hence helps dealing with cases that do not exactly fit the patterns of the disease dataset, as well as marginalizing the effect of rare patterns over the more common ones.

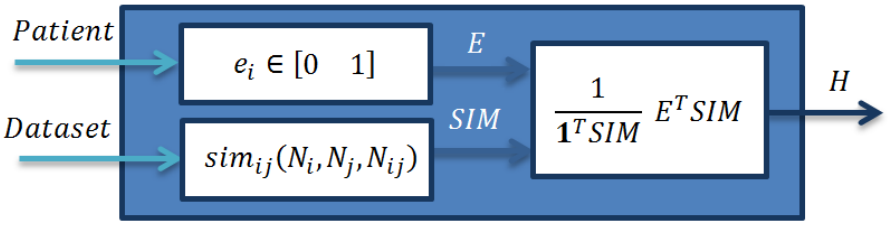

FIGURE 4.3: Block digram of variables and operations of the Recommendation system

4.5 Similarity Measures

Measure of similarity between diseases SIM plays a central role in a RS. In the context of the DPS, similarity also represents the association between diseases and hence it is crucial to define a metric that can properly represent characteristics of similarity between prior evidences of diseases E.

From this point on, we limit ourselves only to data of prevalence, co-prevalence and certainty in the diagnosis of diseases. Hence for every disease, instead of a set of numbers that has state of that disease in all patients, we have only one number (prevalence). This assumption also implies that instead of a continuous spectrum of values that show the degree of development of the disease, E can only have two possible states, one for the diagnosis of the disease with certainty and zero for otherwise. This Compression of information of vector of state of a disease in patients to a prevalence number:

(1) Simplifies the process of importing records and interpretation of results.
(2) Allows us to modify sophisticated similarity and correlation metrics in order to compute them for every set of diseases, using only their joint prevalences.
To find the similarity measures for our problem, one needs to map records of that each contain sets of continuous variables, to nominal variables that are based on prevalence. Let us define N_i as the prevalence of disease D_i for all i, and N as the total prevalence of all diseases in our list of records. Further N_{ij} is the number of co-prevalence of disease i and j.
It is worth mentioning that N_i is the prevalence of disease i (i.e., number of people with disease i), which is not necessarily equal to the incidences of disease i. Hence, if one would want to use incidences instead of prevalences, she would need the information that identifies patients. In an ideal world, the data include all patients IDs. If patients IDs are not available, number of incidences of a disease can be approximated by multiplying the prevalence of the disease by the ratio of occurrence of the disease per patient.
In the rest of Section 4.5, we study concepts from the literature [100, 101] and discuss how they are used as similarity measure sim_{ij} for recommendation (see Equations 4.2 and 4.3). Further, we modify introduced similarity measures for our problem and also propose a new similarity measure.
Standard similarity measures require set of records of all patients. Hence to be applicable to our data, one has to map sets of continuous variables to nominal variables that are based on prevalence. For this, similarity measures are required to satisfy two conditions:
(1) The ability of computing similarity of binary inputs (this was done by extending the metrics to nominal variables).
(2) Evaluating distance with the prevalence rather than the complete information about all patient records.

4.5.1 Conditional Probability

The conditional probability (CP) of a disease i with respect to another disease j can be used as a similarity measure between two diseases, and is formally described as follows:

$$CP_{ij} = P\left(d_i \mid d_j\right) \tag{4.7}$$

CP represents similarity well because it is equal to one if all the cases of the disease under condition also express the conditioning disease and is equal to zero if there is no case to express both diseases at the same time. Further, CP increases by increase of the similarity monotonically.

Lemma 1

CP of two diseases can be computed as the ratio between the co-prevalence of both diseases divided by the prevalence of the conditioning disease:

$$CP_{ij} = \frac{N_{ij}}{N_i} \tag{4.8}$$

Proof

- CP of a disease i with respect to another disease j is equal to the ratio between marginal probability of d_j and joint probability of d_i and d_j.

$$P\left(d_i \mid d_j\right) = \frac{P\left(d_i, d_j\right)}{P\left(d_i\right)} \tag{4.9}$$

- If the total number of records N is sufficiently large, the marginal probabilities and joint probabilities of diseases can be approximated using the portion of prevalence and co-prevalence of the diseases in all records respectively.

$$P\left(d_i\right) = \frac{N_i}{N} \qquad P\left(d_i, d_j\right) = \frac{N_{ij}}{N} \tag{4.10}$$

- Hence we can derive the CP for d_i and d_j as:

$$CP_{ij} = P\left(d_i \mid d_j\right) = \frac{P\left(d_i, d_j\right)}{P\left(d_i\right)} = \frac{N_{ij}}{N_i} \quad \blacksquare \tag{4.11}$$

CP benefits from various advantages:
(1) It is one of the simplest similarity formulas that preserve the distribution of prevalence and hence preserves the statistical properties of prevalence.
(2) It linearly maps the prevalence to the standard interval of [0 1].
(3) It can be computed very efficiently.

(4) It is superior to more complex models that have a similar level of performance (Occam's razor).

CP also suffers from multiple downsides. Most importantly it has a high sensitivity to noise of N_{ij}. Sensitivity with respect to noise of N_{ij} is important because N_{ij} is much smaller than N_i, N_j and N and even one wrongly recorded patient can shift its value drastically. Sensitivity of CP_{ij} respect to noise of N_{ij} is high because the only element in its nominator is N_{ij} and there is no element in denominator to compensate the effect of its possible noise.

Another issue worth mentioning is the asymmetric form of CP_{ij} with respect to its arguments i and j. This causes serious problems in formulation and accuracy of recommendation algorithm, though it can be easily avoided by considering the arithmetic, geometric or harmonic mean of the two versions of similarity as its final value.

$$\begin{cases} CP_{ij} \leftarrow Mean_x\left(CP_{ij}, CP_{ji}\right) \\ CP_{ji} \leftarrow Mean_x\left(CP_{ij}, CP_{ji}\right) \end{cases} \qquad Mean_x(A,B) = \begin{cases} \frac{1}{2}(A+B), & \text{if } x = \text{Arithmetic} \\ (A.B)^{\frac{1}{2}}, & \text{if } x = \text{Geometric} \\ 2\left(A^{-1}+B^{-1}\right)^{-1} & \text{if } x = \text{Harmonic} \end{cases} \tag{4.12}$$

4.5.2 Jaccard Index

Jaccard Index (JI) is a common similarity metric. It measures the cardinality of the intersection sets of two diseases, d_i and d_j, divided by the cardinality of their union:

$$JI_{ij} = \frac{|\, d_i \cap d_j \,|}{|\, d_i \cup d_j \,|} \tag{4.13}$$

Where $|\cdot|$ denoted the set cardinality.

Lemma 2

JI of categorical variables can be computed as:

$$JI_{ij} = \frac{N_{ij}}{N_i + N_j - N_{ij}} \tag{4.14}$$

Proof

- The cardinality of union of the two sets can be represented by the sum of cardinality of the original sets minus cardinality of their intersection:

$$| d_i \cup d_j | = | d_i | + | d_j | - | d_i \cap d_j | \tag{4.15}$$

- For categorical variables, the cardinality of the sets is equivalent to their prevalences:

$$| d_i | = N_i \qquad | d_j | = N_j \qquad | d_i \cap d_j | = N_{ij} \tag{4.16}$$

- By combining these equations JI can be computed as:

$$JI_{ij} = \frac{N_{ij}}{N_i + N_j - N_{ij}} \quad \blacksquare \tag{4.17}$$

The fact that the key factor in both nominator and denominator is N_{ij}, which is usually much smaller than N_i and N_j, causes an underestimation of the importance of the intersection of the two sets. This can be avoided by multiplying the value of N_{ij} by a constant larger than one, say α. To apply the real effect of multiplying the intersection size, we must separate the subset that are under the effect of the intersection from the subsets that are not:

$$| d_i \cup d_j | = | d_i - d_j | + | d_j - d_i | + \alpha | d_i \cap d_j | = | d_i | + | d_j | - (2 - \alpha) | d_i \cap d_j | \tag{4.18}$$

Here, α is an extension factor with default value of one that can magnify the effect of co-prevalence. Finally, the extension of JI can be formulated as:

$$JI_{ij} = \frac{\alpha | d_i \cap d_j |}{| d_i \cup d_j |} = \frac{\alpha N_{ij}}{N_i + N_j - (2 - \alpha) N_{ij}} \tag{4.19}$$

JI does not suffer from most of the issues of CP, such as being asymmetric. Moreover, it contains most of the advantages of CP, such as simplicity and being limited between 0 and 1. But it suffers from a fundamental problem that is ignoring all null sets, hence $\bar{N_i}$, $\bar{N_j}$ and $\bar{N_{ij}}$ cannot take into account information of cases where diseases do not exist although evidences point that they should.

4.5.3 Simple Match Coefficient

The simple match coefficient (SMC) is a variation of the city block distance for categorical variables. It can be derived by inversing the normalized distance of cardinality of the two diseases.

$$SMC_{ij} = 1 - \frac{\| d_i - d_j \|_1}{\max_{i,j} \| d_j - d_i \|_1} \tag{4.20}$$

Lemma 3

SMC similarity for categorical variables can be determined as follows:

$$SMC_{ij} = \frac{N + 2N_{ij} - N_i - N_j}{N} \tag{4.21}$$

Proof

- First-norm (city block or Manhattan) distance between two variables is a common of computing their difference that can be computed by the absolute sum of the difference between the two variables in all dimensions:

$$\| d_i - d_j \|_1 = \sum_{k=1}^{n} | d_i(k) - d_j(k) | \tag{4.22}$$

- By normalizing this metric to its maximum, that is the largest possible value for the variables, it can be bounded between zero and one. By inversing the range linearly, i.e. subtracting the inverse of the normalized distance from one, SMC is emerged as the similarity between two vectors:

$$SMC_{ij} = 1 - \frac{\| d_i - d_j \|_1}{\| d \|_1} \tag{4.23}$$

- For categorical inputs, the number of all recorded diseases N is the largest possible value for variables. Using set combinations SMC can be simplified as:

$$\begin{aligned} SMC_{ij} &= 1 - \frac{\| d_i - d_j \|_1}{\| d \|_1} = 1 - \frac{(N_i - N_{ij}) + (N_j - N_{ij})}{N} \\ &= 1 - \frac{N_{ij} + \bar{N_{ij}}}{N} = \frac{N + 2N_{ij} - N_i - N_j}{N} \quad \blacksquare \end{aligned} \tag{4.24}$$

Similar to JI, SMC includes most of pros and excludes most of cons of CP yet does not directly model the important effect of marginal probabilities (i.e., N_i and N_j) and also underestimates similarity between rare and common diseases.

4.5.4 Relative Risk

Relative Risk (RR) is a common risk assessment metric that is used also for as a similarity measure. RR is the comparison of the probability of occurrence of a disease d_i given another disease d_j and the probability of occurrence of d_j in the null model.

$$RR_{ij} = \frac{P\left(d_j \mid d_i\right)}{P\left(d_j\right)} \tag{4.25}$$

Lemma 4

RR of categorical variables can be computed based on the prevalence and co-prevalence as [73]:

$$RR_{ij} = \frac{N_{ij}\, N}{N_i\, N_j} \tag{4.26}$$

Proof

- As for CP, we can approximate the marginal and joint probabilities with the prevalence of diseases in the dataset:

$$P\left(d_j\right) \cong \frac{N_j}{N} \qquad\qquad P\left(d_i, d_j\right) \cong \frac{N_{ij}}{N} \tag{4.27}$$

- The conditional probability of disease j with respect to disease i can be defined as ratio between the joint probability of disease j and i and the marginal probability of disease i:

$$P\left(d_j \mid d_i\right) = \frac{P\left(d_i, d_j\right)}{P\left(d_i\right)} \tag{4.28}$$

- Hence RR can be modeled as:

$$RR_{ij} = \frac{P\left(d_j \mid d_i\right)}{P\left(d_j\right)} = \frac{P\left(d_i, d_j\right)}{P\left(d_i\right)\, P\left(d_j\right)} = \frac{\frac{N_{ij}}{N}}{\frac{N_i}{N}\,\frac{N_j}{N}} = \frac{N_{ij}\, N}{N_i\, N_j} \quad \blacksquare \tag{4.29}$$

RRs relation with the concept of posterior chance of disease can be used to import more qualitative information of medical diagnoses in records. But RR has high sensitivity to noise of N_{ij}, overestimates similarities involving infrequent diseases and underestimates similarities involving frequent ones. Moreover, RRs limit is not defined and varies by data characteristics in the range of $[\frac{N_{ij}}{N_i\, N_j}\ \ \frac{N}{N_{max}}]$ (where RR expected by chance is 1) and hence needs an extra step of mapping to the range of $[-1\ 1]$.

4.5.5 Pearson Correlation

Pearson correlation (φ) is a common measure of linear dependency between two variables and is widely used in various fields of engineering and science. Correlation of two diseases is defined as the cross expectation of standardization of their prevalence.

$$\varphi_{ij} = E\left[\frac{d_i - avg\,(d_i)}{\sqrt{var\,(d_i)}}\ \frac{d_j - avg\,(d_j)}{\sqrt{var\,(d_j)}}\right] \tag{4.30}$$

Based on the result from [73] we derived the prevalence for categorical variables.

Lemma 5

The correlation for categorical variables can be derived as:

$$\varphi_{ij} = \frac{N_{ij}\, N - N_i\, N_j}{\sqrt{N_i\,(N - N_i)\, N_j\,(N - N_j)}} \tag{4.31}$$

Proof

- Correlation of two random variables can be simplified as their covariance, normalized by root of their variances:

$$\varphi_{ij} = \frac{N_{ij}\, N - N_i\, N_j}{\sqrt{N_i\,(N - N_i)\, N_j\,(N - N_j)}} = \frac{cov\,(d_i, d_j)}{\sqrt{var\,(d_i)\, var\,(d_j)}} \tag{4.32}$$

- Since the existence of a disease is a Bernoulli random variable, it is either zero or one, and $d_i^2 = d_i$. Therefore, we can map the variance and covariance of prevalence

of a disease i and j as:

$$\begin{aligned} var\left(d_i\right) =& E\left[d_i^2\right] - E\left[d_i\right]^2 = P\left(d_i\right) - P\left(d_i\right)^2 \\ =& \frac{N_i}{N} - \left(\frac{N_i}{N}\right)^2 = \frac{N_i\left(N - N_i\right)}{N^2} \\ cov\left(d_i, d_j\right) =& E\left[d_i, d_j\right] - E\left[d_i\right] E\left[d_j\right] = P\left(d_i, d_j\right) - P\left(d_i\right) P\left(d_j\right) \\ =& \frac{N_{ij}}{N} - \frac{N_i}{N}\frac{N_j}{N} = \frac{N_{ij}\, N - N_i\, N_j}{N^2} \end{aligned} \tag{4.33}$$

- Hence the correlation for categorical variables can be derived as:

$$\begin{aligned} \varphi_{ij} = \frac{cov\left(d_i, d_j\right)}{\sqrt{var\left(d_i\right) var\left(d_j\right)}} &= \frac{\frac{N_{ij}\, N - N_i\, N_j}{N^2}}{\sqrt{\frac{N_i\left(N - N_i\right)}{N^2}\,\frac{N_j\left(N - N_j\right)}{N^2}}} \\ &= \frac{N_{ij}\, N - N_i\, N_j}{\sqrt{N_i\left(N - N_i\right) N_j\left(N - N_j\right)}} \quad \blacksquare \end{aligned} \tag{4.34}$$

Correlation is one of the best similarity metrics that uses all information of disease sets i and j. Yet similar to SMC it underestimates similarity when the prevalences of the two diseases are very different. It should be pointed out that, although correlation is defined between -1 and 1, for every given set of diseases the possible range shrinks with the square root of ratio between most frequent and least frequent diseases:

$$\varphi_{ij} \in \sqrt{\frac{N_{min}}{N_{max}}}\,[-1\ 1] \tag{4.35}$$

Other measures that should not be neglected here are extensions of common distance measures, discussed below.

4.5.6 Distance Measure Extensions

Distance measures [100, 102] are an important class of similarity metrics. First consider the Minkowski distance, the r-norm of difference between prevalence of two diseases, as the initial distance between the diseases.

$$\delta_{ij}\left(r\right) = \left\|\Delta d_{ij}\right\|_r = \left(\sum_{k=1}^{n}\left|\Delta d_{ij}\left(k\right)\right|^r\right)^{1/r} \tag{4.36}$$

Here $|\Delta d_{ij}(k)|$ is the absolute of difference between the corresponding elements of two vectors $d_i(k) - d_j(k)$ and r is a real number representing the dimension of distance. The

common value used for r and their interpretations are listed below:

$$d_r = \begin{cases} AbSum\,(\Delta d) = \sum_{k=1}^{n} |\Delta d(k)| & for\ r = 1 \quad Manhattan\ Distance \\ RMS\,(\Delta d) = \sqrt{\Delta d^T \Delta d} & for\ r = 2 \quad Euclidean\ Distance \\ Sup\,(\Delta d) = \max_k \Delta d(k) & for\ r \to \infty \quad Chebyshev\ Distance \end{cases} \tag{4.37}$$

Lemma 6

RMS (Euclidean distance) can be expressed for categorical variables as:

$$\delta_{ij}^2 = \frac{(N - N_i - N_j)\,(N_i + N_j) - 2\,N_{ij}\,N}{N^2} \tag{4.38}$$

Proof

- We reformulate the second norm ($r = 2$) as:

$$\begin{aligned} \delta_{ij} &= \left(\sum_{k=1}^{n} |d_i(k) - d_j(k)|^2 \right)^{1/2} \\ &= \sqrt{\sum_{k=1}^{n} (d_i(k))^2 + \sum_{k=1}^{n} (d_j(k))^2 - 2 \sum_{k=1}^{n} d_i(k)\, d_j(k)} \end{aligned} \tag{4.39}$$

- Sums inside the square root can be constructed using a non-normalized version of variance and covariance of diseases i and j. We modified the expected value, variance and covariance for categorical prevalence in previous sections and hence can simplify the second norm as:

$$\begin{aligned} \delta_{ij} &\cong \sqrt{N\,(var\,(d_i) + var\,(d_j) - 2\,cov\,(d_i, d_j))} \\ &= \sqrt{N \frac{N_i\,(N - N_i) + N_j\,(N - N_j) + 2\,N_i\,N_j - 2\,N_{ij} N}{N^2}} \end{aligned} \tag{4.40}$$

- To make the metric unit-less and independent of size of the samples, we remove the scaling N from the equation and simplify the distance as:

$$\delta_{ij}^2 = \frac{(N - N_i - N_j)\,(N_i + N_j) - 2\,N_{ij}\,N}{N^2} \quad \blacksquare \tag{4.41}$$

Radial Basis Kernel Function

The kernel method, in addition to being used in various data mining and statistical applications, has shown to be a reliable similarity measure. Radial basis kernel function (RBF) is one of the most simple and most successful kernel models available. We combine our categorical distance measure with the idea of using the RBF kernel as similarity to construct a new similarity measure, which fits the analysis of categorical data:

$$K_{RBF}\left(\delta_{ij}\right) = exp\left(\frac{-1}{2\,\sigma^2}\,\delta_{ij}^2\right) \tag{4.42}$$

Sigmoid Function

Sigmoid Functions (SGM) describe another commonly used family of equations in machine learning, which can be extended to a similarity measure by applying our modified distance measure as their input and linear reverse. We can apply our mapped distance measure to the SGM function and produce a new distance measure. We have chosen two functions from the SGM family that fit our application of categorical distance to represent the similarity measure, namly the hyperbolic tangent and the algebraic sigmoid functions.

$$\begin{aligned} SGM_{Hyperbolic\,Tangent}\left(\delta_{ij}\right) &= \frac{1}{1 + exp\left(-\delta_{ij}\right)} \\ SGM_{Algebraic}\left(\delta_{ij}\right) &= \frac{\delta_{ij}}{\sqrt{1 + {\delta_{ij}}^2}} \end{aligned} \tag{4.43}$$

4.5.7 Information Gain

Entropy (H) is one of the most widely used concepts and functions in science and is commonly applied in computer science and more specifically in machine learning as an information gain (IG) measure. We introduce a new IG measure based on the concept of entropy that can appropriately evaluate similarity of categorical variables by their prevalence. Let us first introduce entropy as:

$$H(R) = -R\,\log_2 R \tag{4.44}$$

H is maximum when the variable is uncertain, i.e. its input probability is 1/2. Hence, if we define the information function (I) as the entropy of cumulative ratios between two random variables N_1 and N_2, it will be maximum if N_1 and N_2 have similar statistical patterns.

$$I\left[N_1,\ N_2\right] = H\left(\frac{N_1}{N_1+N_2}\right) + H\left(\frac{N_2}{N_1+N_2}\right) \tag{4.45}$$

Finally we create a new similarity measure formed by weighted summation of information function for $\{N_i,\ N_{ij}\}$ and $\{N_j,\ N_{ij}\}$.

$$IG = \frac{N_i}{N_i+N_j}\ I\left[N_i,\ N_{ij}\right] + \frac{N_j}{N_i+N_j}\ I\left[N_j,\ N_{ij}\right] \tag{4.46}$$

Although these sophisticated distance based and entropy based metrics are statistically superior to their simpler alternatives and have many advantages such as higher noise resistance, they might cause distortions that are hard to diagnose and have a high degree of mathematical complexity that might cause overestimation of the model [102]. Hence they should be avoided if there exist simpler measures with acceptable performances.

4.5.8 Expectation Ratio

We introduce expectation ratio (ER) as a new metric for computing the similarity between categorical variables, especially useful for diseases. ER is defined as the ratio between expectation of co-prevalence of two diseases and square root of multiplied expectation of each:

$$ER = \frac{E\left[d_i\ d_j\right]}{\sqrt{E\left[d_i\right]\ E\left[d_j\right]}} = \frac{N_{ij}}{\sqrt{N_i\ N_j}} \tag{4.47}$$

Although ER is a simple and symmetric model, it does not have most of the mentioned disadvantages, such as strong overestimations and underestimations of high frequent and low frequent diseases, uses all information of disease sets ($N_i,\ N_j\ and\ N_{ij}$), is bounded between 0 and 1, increases linearly with respect to N_{ij}, and is the only metric that is not directly dependent on the total sample size (which is certainly a positive point in disease prediction). Like any other statistical method, ER has some potential downsides, most vividly, the sensitivity to noise of N_{ij}.

We can use one of these similarity measures to construct disease similarity matrix SIM from prevalence N_i and co-prevalence of diseases N_ij, which can then be used as input of recommendation system. There exist other available methods such as Bayesian metrics

[103] and a neighbor joining algorithm [104], which we will not address here, since they require more information than prevalence only.

4.6 Recommender

After describing RS and similarity measures, we introduce the third stage of prediction system (Figure 4.1). The task of this stage is to predict the hidden and high risk diseases, given the probability of diseases generated in the previous stage.

4.6.1 Rule based recommender

After a careful analysis, we decided to use a rule based selection mechanism that consists of the union of two threshold layers: the *necessary layer* and the *sufficient layer*. In the necessary layer, the recommender predicts the ED of most probable diseases if their probabilities are sufficiently large, i.e. more than a threshold pt_l. In the sufficient layer, the system recommends diseases that are highly probable, i.e. their probabilities are more than a threshold pt_h. Since passing a disease from sufficient layer is regardless of probability of other diseases, pt_h should be larger than pt_l.

$$\textbf{Recommend } d_i \textbf{ IF } \{H_i \in MD_{ED} \textbf{ AND } H_i > pt_l\} \textbf{ OR } \{H_i > pt_h\} \tag{4.48}$$

Here MD_{ED} is the set of the ED most probable diseases to be recommended.

Learning Thresholds

After defining the decision making threshold for the recommender, the remaining questions are: What are the right thresholds pt_l and pt_h? What is the number of recommended diseases ED? ED should be evaluated based on the level of wellbeing of the person and can be set either by the user or can be based on the number of already known diseases.

Let us define a cost function S as the weighted sum of false positive (FP) and false

negative (FN) rates of disease prediction:

$$\begin{aligned} \min_{pt} S &= \alpha\, FP + (1-\alpha)\, FN \\ FP &= D^{*T}\,(1-\hat{D}_{pt}) \\ FN &= (1-D^{*})^{T}\,\hat{D}_{pt} \end{aligned} \tag{4.49}$$

Here pt is the set of recommender parameters $\{pt_l, pt_h\}$, α is the conservativeness factor, $\hat{D}_{pt}$ is the vector of probability of diseases computed by RS, FP is the probability of a real disease D^* to be undetected, FN is the probability of a predicted disease $\hat{D}_{pt}$ to be nonexistent and subscript pt stands for dependency of the predicted disease to thresholds.

Theorem 1

Assuming that pt_h is the major criteria of recommender (in compare to pt_l), proportion of diseases in real incidents and training records having the same statistics and that $E\left[\hat{D}_{pt}\right] >> \mathbf{1}$ and given the thresholds pt_l and pt_h and parameter α, the expected value of the defined cost function S can be approximated from the dataset as:

$$E[S_\alpha] \cong \left(\frac{N_D}{N} - (1-\alpha)\,\mathbf{1}\right)^{T} \left[\frac{N_D}{N}\, SIM > pt_h\right] \tag{4.50}$$

Here $\mathbf{1}$ stands for a vector of ones of size n.

Proof

- Let us assume that SIM is the similarity recommendation matrix, trained from the data, N_D is the vector of number of real cases of disease in the test dataset, $\hat{N}_D$ is the vector of number of reported cases of each disease and N is the total number of cases in the dataset. Hence the expectation of a disease to be reported is related to the chance that the sum of the chance that other diseases be reported multiplied by the similarity between them and the disease be larger than the threshold. By combining the expectation for all diseases we can have compact formula:

$$E\left[D^*\right] = \frac{N_D}{N} \qquad E\left[\hat{D}_{pt}\right] \geq \left[\frac{\hat{N}_D}{\hat{N}}\, SIM > pt_h\right] \tag{4.51}$$

- To compute the expected value of S as a function of threshold, we need to make a fundamental assumption that the ratio of reports on a disease is proportional to the real number of cases of that disease regardless of the disease; i.e.:

$$\frac{N_D}{N} \propto \frac{\hat{N}_D}{\hat{N}} \tag{4.52}$$

  This provides us the possibility of estimating expected errors in absence of statistics of test population N_D by approximating it with corresponding values in training dataset $\hat{N}_D$. This assumption is false in an absolute sense since the proportion of records of some diseases in training dataset can be different from their reports in test population. This is due to the fact that there are diseases that are harder to detect or get reported less. But the assumption is still a good approximation of reality for multiple reasons:

  (1) This phenomenon is not a dominant pattern in distribution of diseases because otherwise the whole medical evaluation system would be und-ermined.

  (2) Effect is distributed relatively uniformly among disease groups.

  (3) Item based recommendation model is resistant to distortion of prior distribution.

  (4) If the rate for a specific disease is far off then the medical diagnosis system itself is undermined.

- By this assumption and substituting the above equations to cost function, the expectation of the cost function can be simplified as:

$$\begin{aligned} E[S] &= \alpha\, FP + (1-\alpha)\, FN \\ &= \alpha\, E\left[D^{*T}\,(1-\hat{D}_{pt})\right] + (1-\alpha)\, E\left[(1-D^{*})^T\,\hat{D}_{pt}\right] \\ &\cong (N_D - (1-\alpha)\, N\, \mathbf{1})^T\, \hat{D} - \alpha\, N_D^T\, \mathbf{1} \\ &\cong \left(\frac{N_D}{N} - (1-\alpha)\,\mathbf{1}\right)^T \left[\frac{N_D}{N}\, SIM > pt_h\right] \quad \blacksquare \end{aligned} \tag{4.53}$$

Given the SIM matrix, vector N_D and parameter α, $S\,(pt_h)$ can be minimized for using heuristic search techniques such as primal dual path following algorithm or genetic algorithm. There is an alternative strategy for finding thresholds pt_h and pt_l by analyzing the Receiver Operating Characteristic (ROC) curves of the disease prediction respect to pt_h or pt_l. We will discuss this technique and its issues in Chapter 5.

4.7 Probabilistic Graphical Model

To improve the performance of the prediction system, we include a layer of probabilistic graphical model between the RS and threshold recommender. A probabilistic graphical model (PGM) is a network of conditional dependencies among a set of random variables, used in machine learning. Applying a suitable PGM can enhance the recommendations result by:

(1) Leveraging relations between diseases and prior chances of diseases.

(2) representing a sparse joint distribution of the RS

(3) combining probabilities of many weakly relevant diseases [105].

A sparse Bayesian network can provide all of the above properties. Best candidates for such a model are naive Bayes and independence of causal influences (ICI). In the following, both naive Bayes and ICI will be represented in the context of our prediction system but naive Bayes will only be used in the evaluation stage.

4.7.1 Naive Bayes

Naive Bayes is one of the simplest forms of directed PGMs. It assumes conditional independence of causing factors of an outcome, which in our problem means that in order to predict the occurrence of a disease using co-prevalences with other diseases, one can ignore the relation of those diseases with each other.

Theorem 2

A Laplaceian naive Bayes predictor predicts the probability of a disease i, given the disease evidences $\hat{D}$:

$$P\left(d_i \mid \hat{D}\right) = \left(1 + \left(\frac{N}{N_i + 0.2} - 1\right)^{2-n} \prod_{j=1}^{n} \frac{N_j - N_{ij} + 0.1}{N_{ij} + 0.1}\right)^{-1} \tag{4.54}$$

Proof

By assuming the conditional independence of diseases, conditional probability distribution of diseases and hence the probability of each disease can be simplified:

$$P\left(d_i \mid \hat{D}\right) = \frac{P\left(d_i\right) \prod_{j \in n^*} P\left(\hat{d_j} \mid d_i\right)}{P\left(d_i\right) \prod_{j \in n^*} P\left(\hat{d_j} \mid d_i\right) + P\left(\bar{d_i}\right) \prod_{j \in n^*} P\left(\hat{d_j} \mid \bar{d_i}\right)} \tag{4.55}$$

Here, n^* is the set of reported diseases, $P(d_i)$ is the prior probability of disease i, $P(\hat{d_j}|d_i)$ is the likelihood of evidence of disease j given that disease i exists, and denominator of the fraction is the marginal probability of evidence of all other diseases under existence and nonexistence of the disease i and the outcome $P\left(d_i \mid \hat{D}\right)$ is the posterior probability of disease i given all evidences.

By directly applying the co-prevalences of diseases to the Laplaceian naive Bayes predictor, we have:

$$\begin{aligned} P\left(d_i\right) &= \frac{N_i + L_0}{N + 2\,L_0} \\ P\left(\hat{d_j} \mid d_i\right) &= \frac{N_{ij} + L_0}{N_i + 2\,L_0} \end{aligned} \tag{4.56}$$

$$\begin{aligned} P\left(d_i\right) P\left(\hat{d_j} \mid d_i\right) &= \frac{N_i + L_0}{N + 2\,L_0} \prod_{j=1}^{n} \frac{N_{ij} + L_0}{N_i + 2\,L_0} \\ P\left(\bar{d_i}\right) P\left(\hat{d_j} \mid \bar{d_i}\right) &= \frac{N - N_i + L_0}{N + 2\,L_0} \prod_{j=1}^{n} \frac{N_j - N_{ij} + L_0}{N - N_i + 2\,L_0} \end{aligned} \tag{4.57}$$

$$\begin{aligned} Num &= (N_i + L_0) \prod_{j=1}^{n} (N - N_i + 2\,L_0)(N_{ij} + L_0) \\ Den &= (N - N_i + L_0) \prod_{j=1}^{n} (N_j - N_{ij} + L_0)(N_i + 2\,L_0) \\ P\left(d_i \mid \hat{D}\right) &= \frac{Num}{Num + Den} \\ &\cong \left(1 + (\frac{N}{N_i + 0.2} - 1)^{2-n} \prod_{j=1}^{n} \frac{N_j - N_{ij} + 0.1}{N_{ij} + 0.1}\right)^{-1} \quad \blacksquare \end{aligned} \tag{4.58}$$

In the above equation, L_0 is the Laplacian constant. To compute the likelihood of evidence of disease j given disease i, we used the co-prevalence of disease i and j normalized by prevalence of disease i:

$$P\left(\hat{d_j} \mid d_i\right) \propto \frac{N_{ij}}{N_i} \tag{4.59}$$

Instead of the direct way as above, we can use the recommendation system's output $H_j\left(d_i\right)$ as the conditional probability $P\left(\hat{d_j} \mid d_i\right)$ to achieve a more reliable predictor (see equation 4.2). To see this consider that (see equation 4.2):
(1) RS can indicate the likelihood of disease j as a hidden disease, given each possible disease comorbidity that include disease i:

$$P\left(d_j|d_i, d_k\right) \triangleq H_j\left(D = d_i, d_k\right) \tag{4.60}$$

(2) Joint probability of each of those disease sets $P(D = d_i, d_k)$ is equal to the probability of their occurrence as input evidence of RS.
Hence, the weighted average of joint probabilities of disease i multiplied by recommendation probability of disease j provides an estimate of conditional probability of disease j given disease i that considers comorbidity of disease k for all k. In other words, we generate patients with a certain disease i and another possible disease k with probability N_{ik} as input and compute the probability of disease j as our output and consider it as the $P\left(\hat{d_j} \mid d_i\right)$.

$$P\left(\hat{d_j} \mid d_i\right) = \frac{1}{P\left(d_i\right)} \sum_{\forall k} P\left(d_i, d_k\right) H_j\left(D = d_i, d_k\right) \tag{4.61}$$

We repeat the algorithm for each disease index j until it converges. A similar process can be applied to compute $P\left(\hat{d_j} \mid d_i\right)$, with the only difference that, instead of N_{ik}, $(N_k - N_{ik})$ is used in the similarity measure formula. The values for $P\left(d_i\right)$ and $P\left(\bar{d_i}\right)$ remain the same in the Bayes equation. Hence we have:

$$\begin{aligned} P\left(d_i \mid \hat{D}\right) &= \left(1 + \frac{N - N_i}{N} \prod_{j \in n^*} \frac{P\left(\hat{d_j} \mid \bar{d_i}\right)}{P\left(\hat{d_j} \mid d_i\right)}\right)^{-1} \\ \frac{P\left(\hat{d_j} \mid \bar{d_i}\right)}{P\left(\hat{d_j} \mid d_i\right)} &= \frac{P\left(d_i\right) \sum_{\forall k} P\left(\bar{d_i}, d_k\right) H_j\left(D \neq d_i, d_k\right)}{\left(1 - P\left(d_i\right)\right) \sum_{\forall k} P\left(d_i, d_k\right) H_j\left(D = d_i, d_k\right)} \end{aligned} \tag{4.62}$$

4.7.2 Sigmoid Independence of Causal Influences

We can better mimic the behavior of disease co-occurrence, if we propose a model that captures the intermediate factors that causes both diseases and yet maintains the sparsity of the Bayesian network. Moreover, in a typical scenario, there may be multiple independent or uncorrelated factors that are causing multiple diseases; hence we also need a part of model to combine prospects about diseases [105].

This can be done by replacing the Bayesian network with an independence of causal influences (ICI) network with a sigmoid rule that is an extension of both noisy-OR and naive Bayes models (Figure 4.4) [106]:

$$P\left(d_i \mid \hat{D}\right) = sgm\left(z\right) \qquad z = w_0 + \sum_{j=1}^{n} w_j \, P\left(d_j \mid d_i\right) \tag{4.63}$$

Here, z is the intermediate variable, weights $w_0, ...w_n$ are the model parameters and represent the comorbidity between disease and disease of the index. Finally, the sigmoid function sgm $(\cdot)$ behaves as an influence aggregator (Mixer) and can have multiple realizations (e.g., $1/(1+e^{-z})$ or $(z+\sqrt{1+z^2})/(2\sqrt{1+z^2})$).

In ICI, the intermediate variable z also behaves as an uncertainty filter that adds robustness to noise and solves the zero frequency problem of naive Bayes.

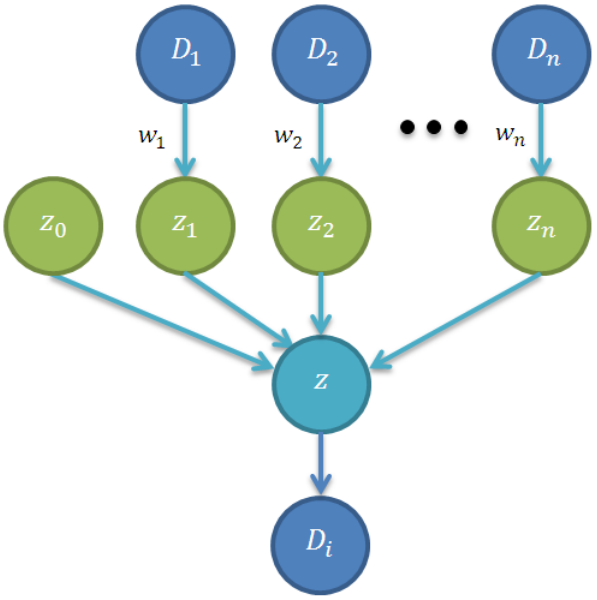

FIGURE 4.4: Block diagram of variables and operations of the ICI Network

It should be pointed out that as a result of placing PGM after the RS, instead of RS output (H_j in equation 4.2), PGM output ($P\left(d_i \mid \hat{D}\right)$ in equation 4.59 or 4.60) is the input to the recommender.

4.8 Summary

Focus of this chapter has been on designing a disease prediction system that detects hidden diseases and predicts upcoming diseases based on the disease records of an individual. As the first step, we introduced a disease comorbidity dataset and summarized its structure, statistical properties and defects. We then proposed a three-layer disease prediction system. As the first layer, we used an Item-based recommendation system to infer the initial probability of diseases and modified it into a compressed model that makes the computation more efficient. We then introduced nine disease similarity measures for the recommendation system, each with its own advantages and disadvantages and modified them to fit the limitations of our dataset and problem. As the last layer of the prediction system, we designed a rule based recommender with two thresholds that makes the final decision about which diseases to report to user. As the final block of the system, we introduced two probabilistic graphical models, naive Bayes and sigmoid independence of causal influences, and proposed them as the second layer that maps the recommendation system's output to the recommender's input.

Chapter 5

Evaluation

5.1 Introduction

Our objective in this chapter is to evaluate the proposed disease predictor. To do so, first we use permutation technique to generate the evaluation data. Second, to test how well a typical recommendation system works under the condition of our problem, a standard item based collaborative filtering is used to predict some of the hidden diseases in the evaluation data. Third, we apply our disease predictor to the evaluation data and compare the results with the naive Bayes predictor. Finally, we test the performance of the disease predictor when it is specialized with the demographic information.

5.2 Generation of evaluation data by permutation

Since the data we used did not have direct records of the individual patients, we could not use one part of the records for training and the other part for testing. To be able to evaluate and compare the prediction power of different systems, we created a permutation test consisting of a population of 100000 virtual patients to see how accurate the system can predict probable combinations of diseases. The permutation population shows the same statistical properties of the original population:

(1) The number of occurrences of a disease in the test population is equal to prevalence of the disease in the original database.

(2) The number of co-occurrences of two diseases in the test population is equal to

number of co-prevalence of the two diseases in the original database.

(3) The probability distribution of number of diseases of a patient in the permutation population is equal to the distribution of disease per patient of the original data.

To achieve these, for each virtual patient, the number of diseases (HD) is randomly selected from the distribution of number of diseases per patient in the dataset (see: Figure 5.1). To determine the primary disease, the chance that disease i is selected is $N_i / \sum_{j=1}^{n} N_j$, for $i = 1$ to n. For other diseases (if $HD >= 2$), the chance that disease j is selected is $N_{ij} / \sum_{k=1}^{n} N_{ik}$ where i is one of the previously selected diseases and for $j = 1$ to n, $j \neq i$.

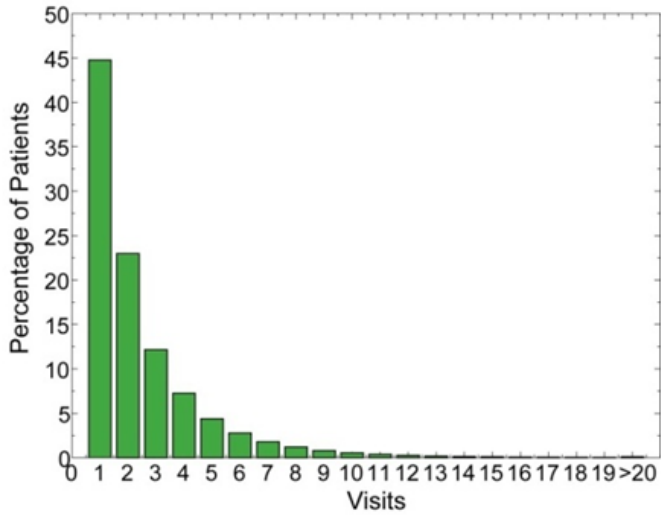

FIGURE 5.1: Frequency of number of diagnoses per patient in the dataset from [73].

5.3 Evaluation of the performance of standard recommendation system

'As the first step, we show the limitations of using only a standard item-based RS. We computed the vectors of sorted probabilities of diseases $Sort\,(H_j(d_i))$ (see equation 4.2) for 100000 permuted patients. In Figure 5.2, the average of the 100000 vectors is plotted. This gives an approximation of the probability distribution of a hidden disease in a patient.

It can be seen from Figure 5.2 that by only applying the recommendation system, the

probability distribution is too flat for the user to be able to choose a meaningful cut off value as the thresholds.

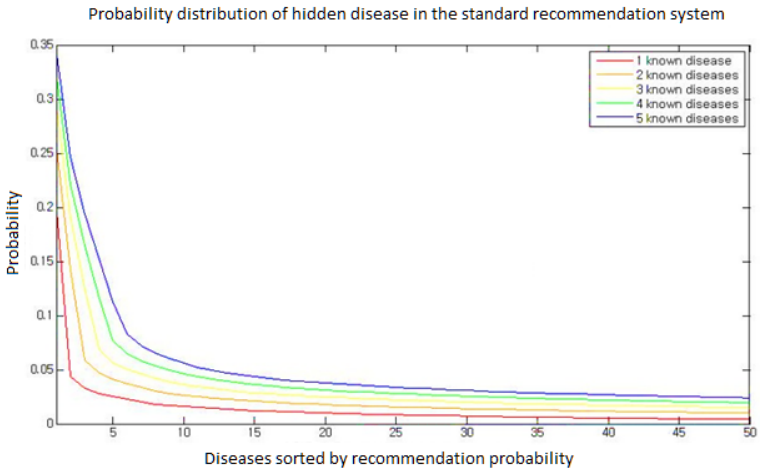

FIGURE 5.2: Probabilities assigned by the recommendation system to the most probable disease, given the different number of known diseases in the patient.

5.4 Evaluation of the proposed prediction system

To evaluate the prediction accuracy of proposed prediction system, we applied a permutation test to the complete system consisting of (1) RS, (2) PGM predictor and (3) recommender. We used a common configuration of prediction system settings to see the typical behavior of the system. Specifically, among the introduced similarity measures, we used Pearson correlation (see Equation 4.29) as a common choice. we used the compressed form (see Equation 4.4) for computing the RS to have a compact representation of the prediction system and speed up the computation process. We then used a naive Bayes with Laplaceian of 0.1 as a typical choice of PGM.

Now the only remaining piece of the prediction system is finding a set of suitable values for pt_h and pt_l. In the following, instead of computing thresholds using optimization of cost function S (which we introduced in chapter 4), we use ROC curve analysis as the alternative. It should be pointed out that there is possibility of manual adjustment of parameters using trial and error, either as an alternative to optimization and ROC or as an additional adjustment after applying them.

Finding recommender thresholds using ROC curve

Receiver Operating Characteristic (ROC) curve is a plot that illustrates the intertwined changes of true positive and false positive of a prediction system as a result of variation in a parameter of the prediction system, called ROC parameter. To construct an ROC curve, one should compute the expected number of true positives (TP) and false positives (FP) of outcome predictions of a test population over different values of the ROC parameter and plot the trend of parameter on a plot with TP as vertical axis and FP as and horizontal axis.

For our disease prediction, our objective is to plot the ROC curve for disease prediction with respect to threshold and then choose values for them that gives the best TP and FP. However, since ROC analysis requires disease records of a set of test patients, it is out of the range of our direct analysis. But in order to demonstrate how the ROC curve should be and to have an initial estimation of suitable parameters for the evaluation of proposed system, we used our designed permutation test to generate virtual patients from predicting hidden diseases from those patients as indicators of TP and FP and each of pt_h and pt_l once as the ROC parameter. To achieve a recommendation with more confidence (i.e., training the RS so that it has higher chance that the predicted diseases in patient were existent) one should choose the parameter values which correspond to the left side of ROC. Similarly, for more support (i.e., higher probability of predicting existent diseases) the parameter should be chosen from the top of the ROC plane.

The result is the ROC curve for variation in parameter pt_h is represented in Figure 5.3. This strategy of computing ROC curve from permutation test with different thresholds can be used to find the best thresholds. We decided that recommender thresholds of $pt_l = 0.02$ and $pt_h = 0.08$ which is equivalent of TP=0.78 and FP=0.34 gives the most satisfying result as it is the closest point (in Euclidian sense) to the top-left corner of ROC curve. It should be pointed out that it is statistically unreliable to use a parameter (in our case thresholds) for both training and test stages. Hence we do not advice this strategy for the ultimate applications.

To analyze the errors of the model in a more standard way, we let the system give two predictions for each hidden disease ($ED = 2\,HD$). We introduced an accuracy measure $A_{HD,s}$ as the cumulative ratio of true positive (TP) predictions to false positive (FP) predictions and evaluated this measure for multiple values of number of reported

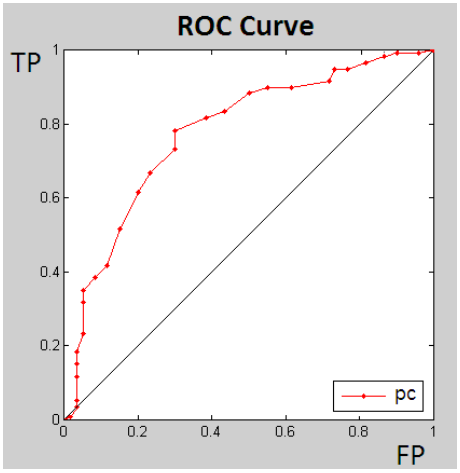

FIGURE 5.3: ROC curve of TP and FP values of disease prediction based on threshold of recommender. pt_h is the variable of the curve and changes from 0.02 to 12 logarithmically.

diseases (s), and number of assumed hidden diseases (HD):

$$A_{HD,s} = \frac{TP}{TP + FP} \tag{5.1}$$

Results of the accuracy measure for different values of number of existing diseases s and number of hidden diseases HD are provided (see, Table 5.1). The predictive power of the proposed system is shown by comparing its accuracy measure with the standard naive Bayes predictor based on the raw prevalence values. The naive Bayes predictor is equivalent to our system when no evidence is provided (i.e. $s \triangleq 1$) represented in the first column of Table 5.1. This is due to the fact that without any information about patients, the prediction power of the system will be completely bypassed and the result will only include the most probable diseases in the population.

HD/s	1	2	3	4	5	6	7	8
1	0.007	0.575	0.732	0.803	0.850	0.880	0.902	0.917
2	-	-	0.443	0.606	0.701	0.758	0.799	0.832
3	-	-	-	0.379	0.532	0.631	0.697	0.744
4	-	-	-	-	0.339	0.489	0.587	0.656
5	-	-	-	-	-	0.313	0.459	0.558

TABLE 5.1: Accuracy of the system with respect to different number of reported diseases and hidden diseases. Model's specifications were the compressed Pearson RS followed by the Laplacian NB and $pt_l = 0.02$ and $pt_h = 0.08$ for recommender thresholds.

5.5 Evaluation of the proposed prediction system for different demographic groups

The initial database had a relatively good variety of gender (58% female) and ethnicity (7.54% Black and 2.38% Hispanic, Asian, Native American). Because of this, Hidalgo et al. [73] provided separate information on the disease co-prevalences for different race and gender combinations (see, Table 5.2).

Patients	Male	Female	All
White	4910362 (37.66%)	6835054 (52.42%)	11745416 (90.08%)
Black	386663 (2.97%)	596432 (4.57%)	983095 (7.54%)
Total	5440490 (41.72%)	7598529 (58.28%)	13039018 (100%)

TABLE 5.2: Size of datasets separated by gender and race; i.e. Male, Female, Black and White.

Based on this data, we applied the proposed prediction system to datasets of different combinations of races and gender. These system need to be trained only using prevalence data of patients with specific gender and race in order to be able to predict disease of a patient with that combination of race and gender. We distinguish these systems from the original system that is trained and tested on all of the population by calling them "demographic prediction systems". After testing these systems under different configurations, we identified the following notable results:

(1) Demographic prediction systems trained on different genders and races led to radically different values for similarity matrices. This is due to the fact that patients with different race and gender have different physiology and lifestyle and hence different disease patterns.

(2) Difference in comorbidity data and similarity values between demographic groups had close to no effect on prediction quality of the demographic prediction systems. Prediction accuracy remained almost equal across different gender and race combinations (see Table 5.3). This shows the robustness of accuracy of the algorithm with respect to different data sources.

(3) Demographic prediction systems had significantly higher prediction accuracy than the original system (see Table 5.3). This was expected from an accurate prediction system, since the extra information on gender and race narrows the range of patterns.

This evaluation provides significant evidence that the system remains robust and consistent given different demographics and datasets.

A	Male	Female	All
White	0.726	0.728	0.701
Black	0.726	0.731	0.718
Total	0.693	0.722	0.606

TABLE 5.3: Accuracy of the system with respect to different dataset from Male-Female and Black-White combinations. Number of diseases in each patient is $s = 4$ and number of hidden diseases is $HD = 2$. Compressed Pearson RS and naive Bayes are used as the RS and PGM respectively. Parameters Laplacian, pt_l and pt_h are set based on the condition of each model.

5.6 Summary

In this chapter we presented multiple steps of test and analysis to evaluate accuracy of the proposed prediction system. We proposed a permutation test that generates virtual patients based on prevalence data, a strategy for finding threshold parameters and an accuracy measure that evaluate the performance of a prediction system under different configurations of the permutation test. By completion of design of evaluation process, we tested the Standard RS and showed how the slow ramp of recommendation probability distribution makes the Standard RS ineffective in dealing with the disease prediction problem. We then moved to evaluation of our proposed prediction system for different numbers of diseases and hidden diseases. The proposed prediction system performance under the permutation test showed that given enough number of recorded diseases, the prediction system can be a reliable source of detecting hidden and upcoming diseases. For comparison, the naive Bayes predictor was used, which showed drastically lower accuracy in compare to the proposed prediction system, illustrating the necessity of combining RS, PGM and recommender. We finally introduced demographic prediction systems based on prevalence data of specific race and genders and observed considerable improve in performance.

Chapter 6

Contributions

6.1 Conclusion

In this thesis we proposed a machine learning algorithm to predict future and hidden diseases of an individual, based on reported diseases of him/her. The final algorithm suggests the unknown diseases that their co-prevalences with reported diseases have strong patterns in its training dataset.

In chapter one the roadmap of thesis was introduced. In chapter two the scale of disease prediction problem were discussed and in chapter three related studies were reviewed. In chapter four we laid down the sections of our proposed disease prediction system.

As the first phase, we proposed HuDiNe [94] as the suitable choice for dataset and analyzed the statistical properties and limitations of it. For the second phase, we trained the disease prediction system. The goal of training was finding patterns of co-prevalence of diseases with limited data of occurrence of diseases in population.

In the third phase, the goal of the system is to find the undiagnosed diseases of a patient who is not a part of the database, given the patients disease history. The unknown diseases reported by the system are the ones that have strongest patterns in the database with respect to the patients disease history.

The system successfully achieved these goals by applying multiple layers of analysis over the dataset and patients data, including similarity measure, item-based collaborative filtering, probabilistic graphical model and threshold based recommender.

In the fourth phase, we evaluated our system. We designed a permutation test, consisting of virtual patients with the same statistical distribution of diseases as the database

population. The results showed that, with a good adjustment of parameters, the system can predict even multiple hidden diseases with high accuracy, especially when the patient report multiple relevant diseases.
We also designed a system for each specific race and gender. We observed consistency of performance of the algorithm across different races and genders, which supports that the algorithm will likely achieve the same good performance for other databases and populations.

6.2 Potential Applications

There are various potential applications for our predictive system. The first and arguably the most important one is risk assessment of diseases for individual patients. This, as a part of the personalized medicine system, can offer the diet, exercises and medical treatments that are necessary or helpful for a preventive healthcare and a better life quality overall.
A second benefit of the system is that it provides a new opportunity for studying the evolution and comorbidity of diseases in patients and perhaps their underlying causes. Thirdly, visualizing the system using a directed disease network or a follow-up risk table can give a better understanding of the condition to both doctors and patients.
Applying this methodology on specific populations and disease families can help determining whether differences in the comorbidity patterns of populations indicate differences in biological processes, environmental factors, or health care quality of each population. Finally, the system can be employed for identifying the key diseases that lead to many diseases with numerous patients. It can shift priorities for an effective health care policy by targeting those diseases.

6.3 Future Works

One of the key features of these types of analysis is that they have the capability to be extended or modified based on new datasets. First, by accessing a medical record that has the patients as separate instances, the system can take into account the co-prevalences of more than two diseases to produce a more complex conditional probability structure, that itself can be used for more sophisticated graphical models than naive Bayes or ICI.

The information of the patients can also be taken into account for prediction by computing the correlation between different demographics and diseases.

Second, by accessing social ties between individuals, the system can model the contagion and clusters of the diseases. By taking the social network into account in the graphical model, the system indirectly applies social network analysis on the disease patterns. The social network data can be very powerful for medical prediction. The social network data can provide evidence about the underlying genetic, environmental and lifestyle similarities between contacts.

Thirdly, we can use various disease networks, such as the phenotypic network, genetic network (Diseasome), metabolic network, protein interaction network, and the traditional systemic network. Each disease network has a potential predictive power, since the links between diseases in each of these networks takes into account a particular aspect of the similarity between the diseases. These predictive powers can be combined using an expert system that can be trained to weight the distances between diseases in each network and combine them in one single number representing the similarity.

Forth, using timeline in analysis not only can provide the capacity of modeling the dynamic of emergence and progression of diseases but also helps to identify whether the comorbidity of two diseases is due to a causal relation between them or just a correlation resulting from mutual risk factors. Although this time analysis can increase accuracy and shed light on progression of diseases, it can increase the complexity of the model, in both the number of parameters and nonlinearity, and produce multiple challenges in computational limits, theoretical analysis and over-fitting rate.

Fifth, a large scale mortality risk assessment can be applied to study risk of death of certain life quality reduction effects rather than the general disease. This can give a better perspective to the real mortality rate of diseases.

There are also extensions available beyond the database type and extraction of its patterns. An already built it extension to the current prediction system is using prior evidences e_i that are between zero and one rather than either zero or one. By this, instead of zero representing non-existence of disease and one representing the existence of disease, we can use a value between zero and one that represents the degree existence of the disease. For example, to represent the high blood pressure, one can map the normal blood pressure (120, 80) to zero and stage two high blood pressure (160, 100) to one and linearly map in between numbers to zero to one interval.

Bibliography

[1] Lance A Waller and Carol A Gotway. *Applied spatial statistics for public health data*, volume 368. John Wiley & Sons, 2004.

[2] John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

[3] Nicholas A Christakis and James H Fowler. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in Medicine*, 32(4):556–577, 2013.

[4] John T Cacioppo, James H Fowler, and Nicholas A Christakis. Alone in the crowd: the structure and spread of loneliness in a large social network. *Journal of personality and social psychology*, 97(6):977, 2009.

[5] James H Fowler and Nicholas A Christakis. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *BMJ: British medical journal*, 337, 2008.

[6] J Niels Rosenquist, James H Fowler, and Nicholas A Christakis. Social network determinants of depression. *Molecular psychiatry*, 16(3):273–281, 2010.

[7] Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379, 2007.

[8] Nicholas A Christakis and James H Fowler. Social network sensors for early detection of contagious outbreaks. *PloS one*, 5(9):e12948, 2010.

[9] James H Fowler and Nicholas A Christakis. Estimating peer effects on health in social networks: A response to cohen-cole and fletcher; trogdon, nonnemaker, pais. *Journal of health economics*, 27(5):1400, 2008.

[10] Nicholas A Christakis and James H Fowler. The collective dynamics of smoking in a large social network. *New England journal of medicine*, 358(21):2249–2258, 2008.

[11] Mark A Pachucki, Paul F Jacques, and Nicholas A Christakis. Social network concordance in food choice among spouses, friends, and siblings. *American Journal of Public Health*, 101(11):2170, 2011.

[12] Sara C Mednick, Nicholas A Christakis, and James H Fowler. The spread of sleep loss influences drug use in adolescent social networks. *PloS one*, 5(3):e9775, 2010.

[13] Nancy L Keating, A James O'Malley, Joanne M Murabito, Kirsten P Smith, and Nicholas A Christakis. Minimal social network effects evident in cancer screening behavior. *Cancer*, 117(13):3045–3052, 2011.

[14] J Niels Rosenquist, Joanne Murabito, James H Fowler, and Nicholas A Christakis. The spread of alcohol consumption behavior in a large social network. *Annals of Internal Medicine*, 152(7):426–433, 2010.

[15] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 2012.

[16] James H Fowler and Nicholas A Christakis. Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences*, 107 (12):5334–5338, 2010.

[17] David G Rand, Samuel Arbesman, and Nicholas A Christakis. Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences*, 108(48):19193–19198, 2011.

[18] Rose McDermott, James Fowler, and Nicholas Christakis. Breaking up is hard to do, unless everyone else is doing it too: social network effects on divorce in a longitudinal sample followed for 32 years. *Unless Everyone Else is Doing it Too: Social Network Effects on Divorce in a Longitudinal Sample Followed for*, 32, 2009.

[19] J Miller McPherson, Pamela A Popielarz, and Sonja Drobnic. Social networks and organizational dynamics. *American Sociological Review*, pages 153–170, 1992.

[20] Timothy La Fond and Jennifer Neville. Randomization tests for distinguishing social influence and homophily effects. In *Proceedings of the 19th international conference on World wide web*, pages 601–610. ACM, 2010.

[21] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

[22] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[23] James H Fowler, Christopher T Dawes, and Nicholas A Christakis. Model of genetic variation in human social networks. *Proceedings of the National Academy of Sciences*, 106(6):1720–1724, 2009.

[24] James H Fowler, Jaime E Settle, and Nicholas A Christakis. Correlated genotypes in friendship networks. *Proceedings of the National Academy of Sciences*, 108(5): 1993–1997, 2011.

[25] Eric A Fischer. Simultaneous hermaphroditism, tit-for-tat, and the evolutionary stability of social systems. *Ethology and Sociobiology*, 9(2):119–136, 1988.

[26] Martin A Nowak and Karl Sigmund. Evolutionary dynamics of biological games. *science*, 303(5659):793–799, 2004.

[27] Robert M Axelrod. *The evolution of cooperation*. Basic books, 2006.

[28] Alan D Lopez, Colin D Mathers, Majid Ezzati, Dean T Jamison, and Christopher JL Murray. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *The Lancet*, 367(9524):1747–1757, 2006.

[29] Colin D Mathers, Ties Boerma, and Doris Ma Fat. Global and regional causes of death. *British medical bulletin*, 92(1):7–32, 2009.

[30] Goodarz Danaei, Eric L Ding, Dariush Mozaffarian, Ben Taylor, Jürgen Rehm, Christopher JL Murray, and Majid Ezzati. The preventable causes of death in the united states: comparative risk assessment of dietary, lifestyle, and metabolic risk factors. *PLoS medicine*, 6(4):e1000058, 2009.

[31] Jan-Emmanuel De Neve, Nicholas A Christakis, James H Fowler, and Bruno S Frey. Genes, economics, and happiness. *Journal of Neuroscience, Psychology, and Economics*, 5(4):193, 2012.

[32] Steven Pinker and Arthur Morey. *The better angels of our nature: Why violence has declined*, volume 75. Viking New York, 2011.

[33] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[34] Bar-Yam Yaneer. Dynamics of complex systems. *Studies in Nonlinearity, Westview Press*, 1997.

[35] Markus Christen and Laura Rebecca Franklin. The concept of emergence in complexity science: Finding coherence between theory and practice. *Proceedings of the Complex Systems Summer School*, 4, 2002.

[36] Ingrid Lobo. Biological complexity and integrative levels of organisation. *Nature Education*, 1(1), 2008.

[37] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[38] Albert-Laszlo Barabâsi, Hawoong Jeong, Zoltan Néda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3):590–614, 2002.

[39] Yong-Yeol Ahn, James P Bagrow, and Sune Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010.

[40] Stanley Wasserman. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[41] Burrhus Frederic Skinner. *About behaviorism*. Random House LLC, 2011.

[42] Noam Chomsky. A review of bf skinner's verbal behavior. *Language*, 35(1):26–58, 1959.

[43] J Martin Bland and Douglas G Altman. Bayesians and frequentists. *Bmj*, 317 (7166):1151–1160, 1998.

[44] Hubert L Dreyfus. Michel foucault, beyond structuralism and hermeneutics. 1983.

[45] Martin Heidegger. Basic writings: from being and time (1927) to the task of thinking (1964). 1977.

[46] Judith Butler, Ernesto Laclau, and Slavoj Žižek. *Contingency, hegemony, universality: contemporary dialogues on the left*. Verso, 2000.

[47] Steven Pinker. *The blank slate: The modern denial of human nature*. Penguin, 2003.

[48] Danile C Dennett. *Kinds of minds: Toward an understanding of consciousness*. Basic Books, 2008.

[49] Terence Turner. Piaget's structuralism. genetic epistemology. jean piaget: Le structuralisme. jean piaget. *American Anthropologist*, 75(2):351–373, 1973.

[50] John Graunt. *Natural and political observations mentioned in a following index, and made upon the bills of mortality*. Springer, 1977.

[51] Fred Brauer and Carlos Castillo-Chavez. *Mathematical models in population biology and epidemiology*. Springer, 2011.

[52] Michael Y Li and James S Muldowney. Global stability for the seir model in epidemiology. *Mathematical Biosciences*, 125(2):155–164, 1995.

[53] David JD Earn, Pejman Rohani, Benjamin M Bolker, and Bryan T Grenfell. A simple model for complex dynamical transitions in epidemics. *Science*, 287(5453): 667–670, 2000.

[54] Juan Zhang and Zhien Ma. Global dynamics of an seir epidemic model with saturating contact rate. *Mathematical Biosciences*, 185(1):15–32, 2003.

[55] Nuri Özalp and Elif Demirci. A fractional order seir model with vertical transmission. *Mathematical and Computer Modelling*, 54(1):1–6, 2011.

[56] Urszula Ledzewicz and Heinz Schättler. On optimal singular controls for a general sir-model with vaccination and treatment. *Discrete and Continuous Dynamical Systems*, pages 981–990, 2011.

[57] TK Kar and Prasanta Kumar Mondal. Global dynamics and bifurcation in delayed sir epidemic model. *Nonlinear Analysis: Real World Applications*, 12(4):2058–2068, 2011.

[58] Dongmei Li, Chunyu Gui, and Xuefeng Luo. Impulsive vaccination seir model with nonlinear incidence rate and time delay. *Mathematical Problems in Engineering*, 2013, 2013.

[59] MA Fuentes and MN Kuperman. Cellular automata and epidemiological models with spatial dependence. *Physica A: Statistical Mechanics and its Applications*, 267(3):471–486, 1999.

[60] Alexander Shlyakhter, Leonid Mirny, Alexander Vlasov, and Richard Wilson. Monte carlo modeling of epidemiological studies. *Human and Ecological Risk Assessment*, 2(4):920–938, 1996.

[61] Ghassan Hamra, Richard MacLehose, and David Richardson. Markov chain monte carlo: an introduction for epidemiologists. *International journal of epidemiology*, 42(2):627–634, 2013.

[62] Michael Small. Infectious agents in heterogeneous systems: When friends matter. *IEEE Circuits and Systems Magazine*, 14(1):58–74, 2014.

[63] Helen Trottier and Pierre Philippe. Deterministic modeling of infectious diseases: theory and methods. *The Internet Journal of Infectious Diseases*, 1(2):3, 2001.

[64] Xinchu Fu, Michael Small, and Guanrong Chen. *Propagation dynamics on complex networks: models, methods and stability analysis*. Wiley Higher Education Press, 2014.

[65] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

[66] Michael Small and CK Tse. Clustering model for transmission of the sars virus: application to epidemic control and risk assessment. *Physica A: Statistical Mechanics and its Applications*, 351(2):499–511, 2005.

[67] Michael Small, David M Walker, and Chi Kong Tse. Scale-free distribution of avian influenza outbreaks. *Physical review letters*, 99(18):188702, 2007.

[68] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-Laszlo Barabasi. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[69] Igor Feldman, Andrey Rzhetsky, and Dennis Vitkup. Network properties of genes harboring inherited disease mutations. *Proceedings of the National Academy of Sciences*, 105(11):4323–4328, 2008.

[70] Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, et al. Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

[71] Ulrich Stelzl, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, Anke Schoenherr, Susanne Koeppen, et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, 2005.

[72] D-S Lee, J Park, KA Kay, NA Christakis, ZN Oltvai, and A-L Barabási. The implications of human metabolic network topology for disease comorbidity. *Proceedings of the National Academy of Sciences*, 105(29):9880–9885, 2008.

[73] César A Hidalgo, Nicholas Blumm, Albert-László Barabási, and Nicholas A Christakis. A dynamic network approach for the study of human phenotypes. *PLoS computational biology*, 5(4):e1000353, 2009.

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

[75] Kiran K Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J Rentfrow, Chris Longworth, and Andrius Aucinas. Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 281–290. ACM, 2010.

[76] Ms RR Ade, Dhanashree S Medhekar, and Mayur P Bote. Heart disease prediction system using svm and naive bayes. *IJESRT 2(5):277-9655*, 2013.

[77] T Mythili, Dev Mukherji, Nikita Padalia, and Abhiram Naidu. A heart disease prediction model using svm-decision trees-logistic regression (sdl). *International Journal of Computer Applications*, 68(16):11–15, 2013.

[78] M Akhil Jabbar, Bulusu Lakshmana Deekshatulu, and Priti Chandra. Heart disease prediction system using associative classification and genetic algorithm. *arXiv preprint arXiv:1303.5919*, 2013.

[79] Sellappan Palaniappan and Rafiah Awang. Intelligent heart disease prediction system using data mining techniques. In *Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on*, pages 108–115. IEEE, 2008.

[80] Francesco Folino and Clara Pizzuti. A comorbidity-based recommendation engine for disease prediction. In *Computer-Based Medical Systems (CBMS), 2010 IEEE 23rd International Symposium on*, pages 6–12. IEEE, 2010.

[81] Austin H Chen, Shu-Yi Huang, Pei-Shan Hong, Chieh-Hao Cheng, and En-Ju Lin. Hdps: Heart disease prediction system. In *Computing in Cardiology, 2011*, pages 557–560. IEEE, 2011.

[82] R Chitra and V Seenivasagam. Heart disease prediction system using supervised learning classifier. *Bonfring International Journal of Software Engineering and Soft Computing*, 3(1):01–07, 2013.

[83] Syed Umar Amin, Kavita Agarwal, and Rizwan Beg. Genetic neural network based data mining in prediction of heart disease using risk factors. In *Information & Communication Technologies (ICT), 2013 IEEE Conference on*, pages 1227–1231. IEEE, 2013.

[84] Nicholas A Christakis and James H Fowler. *Connected: The surprising power of our social networks and how they shape our lives.* Hachette Digital, Inc., 2009.

[85] Kirsten P Smith and Nicholas A Christakis. Social networks and health. *Annu. Rev. Sociol*, 34:405–429, 2008.

[86] Razvan Bunescu and Raymond J Mooney. Collective information extraction with relational markov networks. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 438. Association for Computational Linguistics, 2004.

[87] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62(1-2):107–136, 2006.

[88] Jennifer Neville and David Jensen. Relational dependency networks. *The Journal of Machine Learning Research*, 8:653–692, 2007.

[89] David E Heckerman and Bharat N Nathwani. An evaluation of the diagnostic accuracy of pathfinder. *Computers and Biomedical Research*, 25(1):56–74, 1992.

[90] Everett M Rogers. Progress, problems and prospects for network research: Investigating relationships in the age of electronic communication technologies. *Social Networks*, 9(4):285–310, 1987.

[91] Nicole B Ellison et al. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007.

[92] Fernando Vega-Redondo. *Complex social networks*. Number 44. Cambridge University Press, 2007.

[93] ICD9. Medical coding reference, July 2014. URL `http://www.icd9data.com/`.

[94] HuDiNe. the human disease network, July 2014. URL `http://hudine.neu.edu/`.

[95] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.

[96] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[97] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 39–46. ACM, 2010.

[98] Imola K Fodor. A survey of dimension reduction techniques, 2002.

[99] Raphael Yuster and Uri Zwick. Fast sparse matrix multiplication. *ACM Transactions on Algorithms (TALG)*, 1(1):2–13, 2005.

[100] Simone Santini and Ramesh Jain. Similarity measures. *Pattern analysis and machine intelligence, IEEE transactions on*, 21(9):871–883, 1999.

[101] Seung-Seok Choi, Sung-Hyuk Cha, and Charles C Tappert. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48, 2010.

[102] Manabu Ichino and Hiroyuki Yaguchi. Generalized minkowski metrics for mixed feature-type data analysis. *Systems, Man and Cybernetics, IEEE Transactions on*, 24(4):698–708, 1994.

[103] Pierre A Devijver. The bayesian distance: A new concept in statistical decision theory. In *Decision and Control, 1972 and 11th Symposium on Adaptive Processes. Proceedings of the 1972 IEEE Conference on*, volume 11, pages 543–544. IEEE, 1972.

[104] Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987.

[105] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[106] Adam Zagorecki and Marek J Druzdzel. Probabilistic independence of causal influences. In *Probabilistic Graphical Models*, pages 325–332. Citeseer, 2006.