BAYESIAN PROBABILISTIC REASONING APPLIED TO MATHEMATICAL EPIDEMIOLOGY FOR PREDICTIVE SPATIOTEMPORAL ANALYSIS OF INFECTIOUS DISEASES

Kaja Moinudeen Abbas

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

May 2006

APPROVED:

Armin Mikler, Major Professor
Samuel Atkinson, Committee Member
Yan Huang, Committee Member
Roy T. Jacob, Committee Member
Joseph Oppong, Committee Member
Krishna Kavi, Chair of the Department of Computer Sciences and Engineering
Oscar Garcia, Dean of the College of Engineering
Sandra L. Terrell, Dean of the Robert B. Toulouse School of Graduate Studies

Abbas, Kaja Moinudeen, Bayesian Probabilistic Reasoning Applied to Mathematical Epidemiology for Predictive Spatiotemporal Analysis of Infectious Diseases. Doctor of Philosophy (Computer Science), May 2006, 92 pp., 8 tables, 38 illustrations, bibliography/references/reference list, 76 titles.

Probabilistic reasoning under uncertainty suits well to analysis of disease dynamics. The stochastic nature of disease progression is modeled by applying the principles of Bayesian learning. Bayesian learning predicts the disease progression, including prevalence and incidence, for a geographic region and demographic composition. Public health resources, prioritized by the order of risk levels of the population, will efficiently minimize the disease spread and curtail the epidemic at the earliest. A Bayesian network representing the outbreak of influenza and pneumonia in a geographic region is ported to a newer region with different demographic composition. Upon analysis for the newer region, the corresponding prevalence of influenza and pneumonia among the different demographic subgroups is inferred for the newer region. Bayesian reasoning coupled with disease timeline is used to reverse engineer an influenza outbreak for a given geographic and demographic setting. The temporal flow of the epidemic among the different sections of the population is analyzed to identify the corresponding risk levels. In comparison to spread vaccination, prioritizing the limited vaccination resources to the higher risk groups results in relatively lower influenza prevalence. HIV incidence in Texas from 1989-2002 is analyzed using demographic based epidemic curves. Dynamic Bayesian networks are integrated with probability distributions of HIV surveillance data coupled with the census population data to estimate the proportion of HIV incidence among the different demographic subgroups. Demographic based risk analysis lends to observation of varied spectrum of HIV risk

among the different demographic subgroups. A methodology using hidden Markov models is introduced that enables to investigate the impact of social behavioral interactions in the incidence and prevalence of infectious diseases. The methodology is presented in the context of simulated disease outbreak data for influenza. Probabilistic reasoning analysis enhances the understanding of disease progression in order to identify the critical points of surveillance, control and prevention. Public health resources, prioritized by the order of risk levels of the population, will efficiently minimize the disease spread and curtail the epidemic at the earliest.

ACKNOWLEDGMENTS

I am profoundly thankful to Dr. Armin R. Mikler for his constant encouragement, discussion and critique during my masters and doctoral program at the University of North Texas. The discussions range from pure academics and research to trivial issues in life. This dissertation is a result of his exemplary role in advising and directing me in the research goals.

I wish to express my sincere thanks to the dissertation committee, including Dr. Armin R. Mikler, Dr. Sam Atkinson, Dr. Yan Huang, Dr. Tom Jacob and Dr. Joseph Oppong for their quality time in guiding and supporting my research. I am thankful to Dr. Farhad Shahrokhi and Dr. Ram Dantu for their invaluable support in enhancing my research credentials. I thank the Department of Computer Science and Engineering at the University of North Texas for providing the computing and financial support during my tenure as a student.

I am thankful to my fellow graduate students, Sangeeta Venkatachalam, Courtney Corley, Robert Gatti, Marty O'Neill II, Amir Ramezani, Sheena Menezes and Venkatesan Iyengar Prasanna in working well together on different research projects. Special thanks are due to Robert Gatti for his exemplary system administration of the computing infrastructure. I wish to thank Fabio G. Cozman for the use of his software package, JavaBayes, to analyze the Bayesian networks.

I worked well together with Dr. Mikler, Amir Ramezani and Sheena Menezes for the study on the Bayesian spatial correlation of influenza and pneumonia prevalence. Demographic risk analysis for influenza was carried out working with Dr. Mikler and Robert Gatti. The temporal analysis of HIV in Texas was done working with Courtney Corley, Dr. Mikler and Dr. Oppong. I wish to specially thank Dr. Oppong for letting the HIV dataset accessible for this study. Hidden Markov Model analysis of influenza infectivity was conducted working with Sangeeta

Venkatachalam and Dr. Mikler. It has been an exhilarating and rewarding experience to work collaboratively and learn from each other during the different phases of research.

I thank all the students and the faculty of the Computational Epidemiology Research Lab for the regular research discussions and seminars. The discussions sharpened our critical thinking and logical reasoning skills and enhanced the quality of our research and analysis. I am excited and proud of all the young and smart minds at play in the lab and wish them the very best in all their endeavors.

To think and rationalize beyond the norm has been the drive in this research endeavor. The doctoral program has been a vibrant learning experience in research as well as in teaching and grant writing skills. I am grateful for the unconditonal support of Dr. Mikler throughout the program. I had an wonderful time and lots of smiles to cherish forever.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1. Objective

The focus of this study is in the development of a predictive computing model for epidemiological sciences. The constrained resources of the public health departments have to be optimally allocated [37]. Epidemiological studies and models of the present and the past have primarily focussed on the analysis of past outbreaks as well as on the present diseases status parameters in a given population. This study will aid in the visualization of scenarios of disease outbreaks, that occur naturally as well as induced, as in the case of bio-terror attacks. Bayesian probabilistic reasoning of epidemiological data will facilitate the effective allocation of public health resources by identifying the high risk groups in the population, and enhance the decision making process under uncertainty.

Computational epidemiology forges the epidemiological and computational sciences in the study of diseases. Algorithms are used to develop computational tools to interface with large databases for the retrieval of pertinent information. Data mining techniques are applied on the disease data to reveal the correlation of the different parameters resulting in the varied epidemiological events. The predictive disease progression model will aid in understanding the implicit features from the explicit characteristics and identify the points of control, prevention and surveillance of diseases. The computational framework will facilitate by prioritizing the dissemination of public health resources to the high risk groups in the population, and aid in constructive public health policy making.

1.1.1. *Overview*

The remainder of the chapter illustrates the potential impact of computational epidemiology, motivation for this study, and the need for probabilistic reasoning for epidemiological analysis. Chapter 2 introduces the concepts of probabilistic reasoning, including Bayesian learning, dynamic Bayesian networks and hidden Markov models. Chapter 3 highlights the effective role of preventive medicine in curtailing infectious diseases' epidemics. Chapter 4 sketches on the history of past epidemics and mathematical models used in epidemiological studies and includes the related work in mathematical epidemiology. Chapter 5 elucidates the use of probabilistic reasoning for analysis of epidemiological data and highlights on the the methodologies of the later chapters. Chapter 6 identifies the risk ordering of demographics for an influenza outbreak in a population. Chapter 7 illustrates the spatial correlation of disease prevalence with respect to demographics for influenza and pneumonia in two different geographic regions using Bayesian learning. Chapter 8 undertakes a temporal study of HIV prevalence in Texas using dynamic Bayesian networks. Chapter 9 illustrates the use of hidden Markov models to quantify the social behavioral interactions; thereby, providing insight on the disease progression in a population. Chapter 10 summarizes the different phases of study in using probabilistic reasoning for epidemiological data analysis and includes notes for future work.

1.2. Towards Computational Epidemiology

Computational biology and bioinformatics have harnessed the computational power of today's cyber infrastructure to tackle the computational complexity of tasks associated with genomics research. There are other domains of life sciences in which scientists can take advantage of recent advances in computer science and scientific computing. One such domain is epidemiology with its multiplicity of sub-domains, including field epidemiology, epidemiological genomics, infectious disease epidemiology, social and behavioral epidemiology, and

surveillance [66, 67]. As opposed to the combinatorial complexity of computation associated with computational biology, epidemiology deals with data that are often sparse, widely distributed, incomplete (often due to confidentiality and other constraints), and frequently compromised by conflicting data that confound or disguise the evidence epidemiologists attempt to reveal. It is the ability to draw conclusions and make predictions from this type of information that identifies epidemiology. Today, the role of epidemiologists has become even more pronounced as the significance of public health has been recognized. Development of specific computational models and tools for epidemiology is needed to enhance the quality of information, facilitate prediction, and accelerate the generation of answers to specific questions. Ever increasing population diversity, the ability to travel long distances in a short time, increased globalization, social behavioral complexity and the associated exposure to new public health threats make it imperative to develop new models that take advantage of today's cyber infrastructures and facilitate disease tracking, analysis, surveillance, and control.

With new and re-emerging local disease outbreaks and the increased threat of bioterrorism, disease monitoring can not continue to be fragmentary and inadequate, focusing on small special domains [1]. Developing tools that will accelerate epidemiological research, disease tracking and surveillance is thus imperative. Computational models for the simulation of disease dynamics are required to facilitate adequate what-if analysis. This necessitates the collaboration of scientists from biology, medical geography, epidemiology, computer science, biostatistics, sociology, and environmental sciences to develop and implement computational models and tools in support of epidemiological research. Study of infectious diseases are in need of models pertinent to spatially delineated environments, such as a tuberculosis outbreak in a homeless shelter or factory setting, as well as non-delineated models for a geographic region, such as the progression of influenza in specific regions of the United States. In order to yield adequate precise information, the combined factors of geography, demographic composition, and social behavior must be analyzed.

The significance of computational epidemiology as a new field has been underscored by a special program at the Center for Discrete Mathematics and Theoretical Computer Science (DIMACS) [28], funded as an National Science Foundation (NSF) Technology Center. A 5-year program, consisting of working groups and short-courses focusing on computational and mathematical epidemiology began in the summer of 2002. It emphasizes the development and strengthening of collaborations and partnerships between mathematicians, computer scientists, biologists, sociologists, bio-statisticians, and epidemiologists.

1.3. Motivation

During the last century, life sciences has made tremendous progress in identifying, treating, or even eradicating many infectious diseases. This can primarily be attributed to the increased understanding of the etiology and pathogenesis of such diseases. Nevertheless, newly emerging or re-emerging infectious diseases continue to occur regularly [40]. Some diseases have changed their appearance, some have become resistant to drug treatment, while others are new that no previous outbreaks have ever been studied. It is ironic that epidemiologists have to take advantage of a disease outbreak in order to collect data necessary to formulate public health policy. Medical research has enhanced the understanding of disease characteristics in an individual. For example, the characteristic epidemiological stages of influenza as described by latent period, infectious period, and recovery period [15] as experienced by an individual are well known [15, 25]. So are the symptomatic stages of influenza (i.e., incubation period until symptoms occur) . The manifestation and the spread of many infectious diseases in the population remains elusive and is dependent on the socio-behavioral interaction patterns and population dynamics.

To gain insight into the intricacies of disease dynamics in a specific population, statistical and mathematical models of infectious disease epidemics have been developed. However, further understanding into the composition of an epidemic will facilitate better policy and and

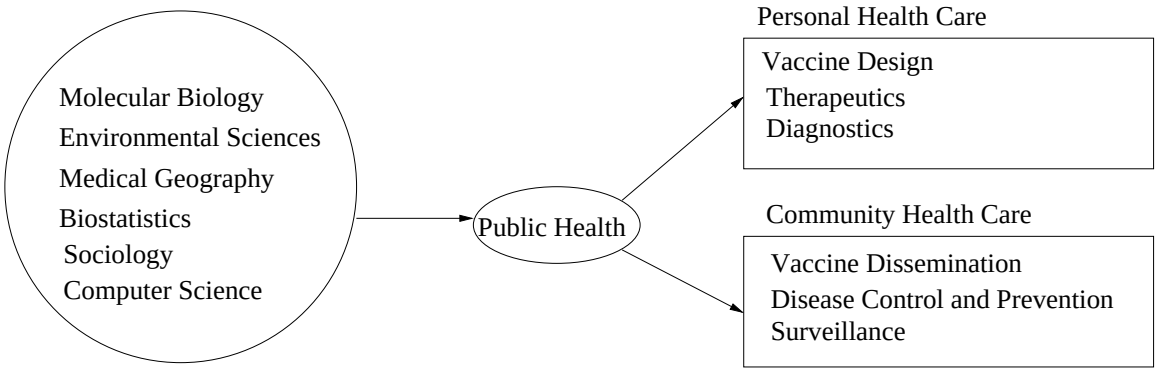


Figure 1.1. Public Health - Multi-disciplinary Domain

planning tool for the allocation of public health resources. Most models operate on the presumption of a closed population, assuming that the epidemic spreads rapidly enough that the changes brought in by births, deaths, migration and demographic changes are negligible [8]. Recently, some computational disease models have emerged, which facilitate the simulation and thus the investigation of different disease characteristics. These include models that exploit the susceptibles-infectives-removals paradigm, cellular automata methodology, agent-based modeling and Bayesian reasoning. This study focuses on the probabilistic analysis of the progression of infectious diseases in non-delineated environments.

1.3.1. *Interdisciplinary Domain of Public Health*

Public health domain in the modern world brings together diverse disciplines, including molecular biology, environmental sciences, medical geography, biostatistics, sociology and computer science (Fig. 1.1). The enhanced understanding of epidemiology and public health will increase the quality of individual health care, including vaccine design, therapeutics and diagnostics, as well as augment the community health care through better measures for vaccine dissemination, surveillance, disease control and prevention.

The study of progression of infectious diseases in different demographic and geographic populations is intrinsically correlated to the genetic makeup, population dynamics and environmental factors of the region (Fig. 1.2). The correlation of these factors will lend to

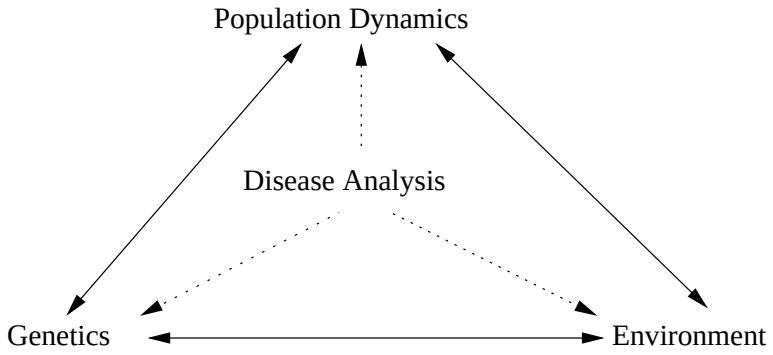


Figure 1.2. Population Dynamics, Genetics and Environment Correlate in the Study of Disease Analysis

a better understanding of epidemiology and public health. The focus is to realize a better understanding of the population dynamics of disease progression in a demographic and geographic region.

1.4. Probabilistic Reasoning for Epidemiology

Health data collection, surveillance and analysis involves information with a unknown degree of uncertainty. This is due to several factors. including the hierarchical nature of data propagation right from the field to the epidemiologists as well as from the local health departments to the national health databases. The modes of data collection include disease reports from health care personnel and hospitals, as well as from national health surveys. Constrained resources and integration complexity of collected disease data impose limitations at achieving a complete view of the health care system including delivery, surveillance and prevention measures. Hence, an optimal overview of the health care program will have to be inferred from a comprehensive analysis of the available data with the maximal plausible degree of confidence. Probabilistic reasoning provides an apt methodology for analyzing health care data in order to infer useful predictive measures and solutions.

Probabilistic reasoning overcomes the limitations of rule-based reasoning and is well suited to the dynamic nature of the real world. Disease incidence in an individual is deterministic, in the sense that he/she is either infected or non-infected. On the other hand, the risk

of an individual being infected is intrinsically stochastic. The interplay of several factors, such as social and community behavior, genetics, inherent or acquired immunity, travel, and other factors determine the risk of infection for an individual. In order to target a safer environment free from diseases and lower the disease prevalence in a given population, the focus is to minimize the risk level of every individual in being susceptible to a disease. Herd-immunity is the desired result, wherein a disease progressing through a population does not result in an epidemic.

Diseases are categorized as reportable and non-reportable. For reportable diseases, hospitals, clinics and physicians that diagnose a case are mandated to report to the local public health departments. The local public health departments report to higher level departments and a national database of statistics on reported diseases is created. In case of non-reportable diseases, such as influenza, mandatory reporting is unwarranted and no definite estimates or statistics is available.

In general, disease outbreak data is not readily available and do not include finer demographic details for comprehensive analysis. The number of reported disease cases determines the lower bound of the disease prevalence in a population. The true prevalence and the upper bound has to be extrapolated from the reported and available data. Also, non-reportable diseases add to the problem of lack of disease data. This necessitates the need for a disease outbreak simulator that can reflect the disease incidence in a given population. Also, such a simulator will fill the gaps in case data that are caused by incomplete or imprecise outbreak descriptions.

Complexity of disease modeling requires high performance computing to generate the outbreak data with all the requisite parameters for further analysis. Dynamic Bayesian methodology aids in the temporal analysis of the data to identify the critical points of control and surveillance of data. Identification of risk ordering among the different geographic and demographic groups will help in prioritizing the prevention measures for optimal decrease in

the disease incidence. Hidden Markov models can analyze the disease data at a meta level and correlate the social behavioral interactions of a population to the disease progression. This will further aid in the knowledge of diseases and design effective public health measures. The resultant predictive model for disease progression for newer geographic and demographic domains will facilitate the optimal deployment of public health resources ahead of disease outbreaks and curtail them at the earliest.

1.5. Contribution of the Study

Probabilistic reasoning under uncertainty suits well to analysis of disease dynamics. The stochastic nature of disease progression is modeled by applying the principles of Bayesian learning. Bayesian learning predicts the disease progression, including prevalence and incidence, for a geographic region and demographic composition. Public health resources, prioritized by the order of risk levels of the population, will efficiently minimize the disease spread and curtail the epidemic at the earliest.

A Bayesian network representing the outbreak of influenza and pneumonia in a geographic region is ported to a newer region with different demographic composition. Upon analysis for the newer region, the corresponding prevalence of influenza and pneumonia among the different demographic subgroups is inferred for the newer region.

Bayesian reasoning coupled with disease timeline is used to reverse engineer an influenza outbreak for a given geographic and demographic settings. The temporal flow of the epidemic among the different sections of the population is analyzed to identify the corresponding risk levels. In comparison to spread vaccination, prioritizing the limited vaccination resources to the higher risk groups results in relatively lower influenza prevalence.

HIV incidence in Texas from 1989-2002 is analyzed using demographic based epidemic curves. Dynamic Bayesian networks are integrated with probability distributions of HIV surveillance data coupled with the census population data to estimate the proportion of HIV incidence among the different demographic subgroups. Demographic based risk analysis lends to observation of varied spectrum of HIV risk among the different demographic subgroups.

A methodology using hidden Markov models is introduced that enables to investigate the impact of social behavioral interactions in the incidence and prevalence of infectious diseases. The methodology is presented in the context of simulated disease outbreak data for influenza.

Probabilistic reasoning analysis enhances the understanding of disease progression in order to identify the critical points of surveillance, control and prevention. Public health resources, prioritized by the order of risk levels of the population, will efficiently minimize the disease spread and curtail the epidemic at the earliest.

CHAPTER 2

PROBABILISTIC REASONING

Uncertainty in knowledge is inevitable when analyzing real world events but lacking full perspective of the real world domain. Decisions are made with partial knowledge through rational analysis. The public health domain often presents scenarios that necessitate decisions upon analysis of incomplete representation of disease incidence and prevalence over a given demographic and geographic region. Probabilistic reasoning is a powerful methodology to address such uncertainty in the public health domain. The principles of non-temporal and temporal probabilistic reasoning will be briefly reviewed in this chapter, and applications based on Bayesian learning are introduced. There are many good resources of study that comprehensively illustrate Bayesian probabilistic reasoning, including [45],[50],[55],[60].

2.1. Probability Overview

Probabilistic reasoning is based on the axioms and rules in probability theory. A random variable defines a set of domain values or events. Random variables can take discrete or continuous values. Probability values are in the interval $[0, 1]$. Prior or unconditional probability $P(a)$ defines the probability of the random variable a independent of other random variables. Posterior or conditional probability $P(a/b)$ defines the probability of the random variable a, given that the random variable b is observed with the different possible outcomes. Probability distribution $P(a)$ defines the set of probability values for all possible outcomes of a random variable a. Joint probability distribution $P(a, b)$ defines the set of probability values for all possible combinations of the outcomes of the random variables a and b. Given the joint probability distribution of a set of random variables, probabilistic inferences can be derived for queries on a subset of random variables.

2.1.1. *Bayes' Theorem*

The fundamental principle in probabilistic reasoning is Bayes' theorem. The theorem updates the existing belief in an hypothesis, given the evidence, by use of probability distributions. The prior probabilities of an hypothesis transform to posterior probabilities, taking into account the evidence.

The Bayes' theorem is useful to compute probabilities for a random variable given a certain evidence of other random variables. Given two random variables $\{a, b\}$, the theorem formulates,

$$P(a/b) = \frac{P(b/a)P(a)}{P(b)}$$

The theorem can be reformulated as,

$$P(a \wedge b) = P(a)P(b/a) = P(b)P(a/b)$$

In a simple scenario where the random variables $\{a, b\}$ take the values $\{true, false\}$, the theorem states that the probability of both random variables being true concurrently, (i.e.) $P(a = true, b = true)$ is the product of the probability of random variable a being true $P(a = true)$ and the probability of b being true given that a is true, (i.e.) $P(b = true/a = true)$. This probability is also the same as the product of the probability of random variable b being true $P(b = true)$ and the probability of a being true given that b is true, (i.e.) $P(a = true/b = true)$.

The naive Bayes model [60] illustrates the use of Bayes' theorem to correlate the probability distribution of the real world cause and effect relationships. Given a cause c and a set of independent effects $\{e_1, e_2, ..., e_n\}$, the probability of concurrent observation of the cause and all the independent effects is the following:

$$P(c, e_1, e_2, ..., e_n) = P(c)\prod_i P(e_i/c)$$

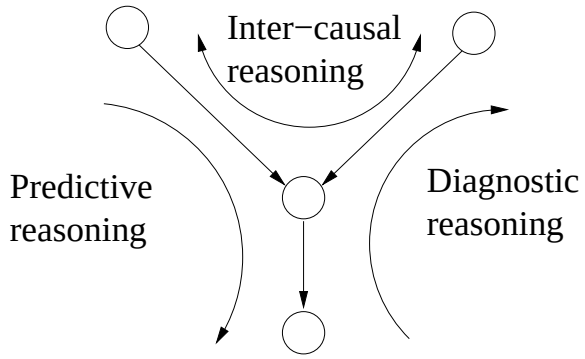


Figure 2.1. Reasoning Methodologies in Bayesian Networks

2.2. Bayesian Networks

Probabilistic reasoning and reasoning under uncertainty are the fundamental principles of Bayesian philosophy. The real world probabilistic relationships can be modeled using Bayesian networks.

A Bayesian network models the different parameters/random variables of the domain in the form of nodes, with directed links/edges between them reflecting their dependencies. The nodes are associated with the corresponding probability distributions for their beliefs. The network can be constructed from a data set that contains a collection of data records for the random variables.

The flow of information in a Bayesian network leads to different inference methodologies, as illustrated in Fig. 2.1. Diagnostic reasoning involves backward reasoning from the effects to the causes. Predictive reasoning uses the information of the causes leading up to the effects. Inter-causal reasoning is the role of mutual causes on a common effect. The above types of reasoning can be combined in any desired ways.

2.2.1. *Knowledge Representation using Bayesian Networks*

Bayesian networks represent the knowledge of both discrete and continuous random variables. The conditional relationships are defined by the directed edges between two random variables/nodes. As shown in Fig. 2.2, the directed edge $a \rightarrow c$ defines that random variable

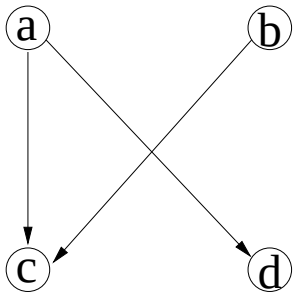


Figure 2.2. Bayesian Network Example for Four Random Variables {a, b, c, d}

a influences random variable c, (i.e.) c is conditionally dependent on a. Similar conditional dependent relationships exist for $a \rightarrow d$ and $b \rightarrow c$. While a prior or unconditional probability distribution quantify the random variables a and b, posterior or conditional probability distributions quantify the random variables c and d.

Using terminology of graph theory [74], a Bayesian network is a directed graph G of vertices/nodes V and directed edges/links E. $G(V, E)$, where $V = \{a, b, c, d\}$ and $E = \{(a, c), (b, c), (a, d)\}$ represents the Bayesian network example in Fig. 2.2. The parents of children nodes c and d are $\{a, b\}$ and $\{a\}$ respectively. The parent nodes define the conditional dependence/influence on the children nodes. The probability distribution for a node is represented by $P(node/Parents(node))$, (i.e.) the probability distribution quantifies and satisfies all the conditional dependencies. The Bayesian network represent the real world probabilistic dependencies through a directed acyclic graph, (i.e.) there should be no directed cycles in the Bayesian network. If any directed cycles exist, it will lead to a non-terminal probabilistic relationships/dependencies among the random variables; in fact, such a scenario represents a flaw in the design of the Bayesian network.

2.2.2. *Bayesian Network Design*

Given a set of random variables $\{a_1, a_2, ..., a_n\}$, the joint probability distribution [60] for a set of assigned values, (i.e.) $\{a_1 = a_1^{'}, a_2 = a_1^{'}, ..., a_n = a_n^{'}\}$ is given in Eq. 1 The formulation is further extended using the product and chains rule of probability theory in Eq. 2. These

formulations are the fundamental principles in the construction of Bayesian networks. Given the parents of a random variable or node in a Bayesian network, the node is independent of all other random variables.

$$(1) \qquad P(a_1' \wedge a_2' \wedge ... \wedge a_n') = P(a_1', a_2', ..., a_n') = \prod_i P(a_i'/Parents(a_i'))$$

$$P(a_1', a_2', ..., a_n') = P(a_n'/a_1', a_2', ..., a_{n-1}')P(a_1', a_2', ..., a_{n-1}')$$

$$(2) \qquad P(a_i/a_1, a_2, ..., a_{i-1}) = P(a_i/Parents(a_i))$$

2.2.3. *Inference in Bayesian Networks*

The random variables are of three types, namely, query, evidence and hidden variables. The query variable is the value of interest and the evidence variables contain the observed values. The hidden variables may contain any of the acceptable values; (i.e.) their values are not evident or observed. Inference in Bayesian networks is the process that quantifies the probability of the different values for a query variable, given the values of the evidence variables.

Exact inference is achieved by enumeration and variable elimination [60]. Equation 3 illustrates the computation of the query variable Q, given the evidence variables E, and hidden variables. H. The values of evidence variables $\{E_1, ..., E_m\}$ are $\{e_1, ..., e_m\}$ correspondingly. The summation takes into account all possible combinations of values of the hidden variables.

$$(3) \qquad P(Q/E) = \Sigma_H P(Q, \{e_1, ..., e_m\}, \{H\})$$

2.2.4. *Sampling Algorithms*

The time complexity of computing exact inference is linear in singly connected networks, while it is intractable in multiply connected networks. As shown in Fig. 2.3, for singly connected networks, there is only one viable path of reasoning between any pair of random variables. For example, there is only one path from random variable a to f, (i.e.) $a \rightarrow c \rightarrow d \rightarrow f$. For a multiply connected network, there is more than one path of reasoning between random

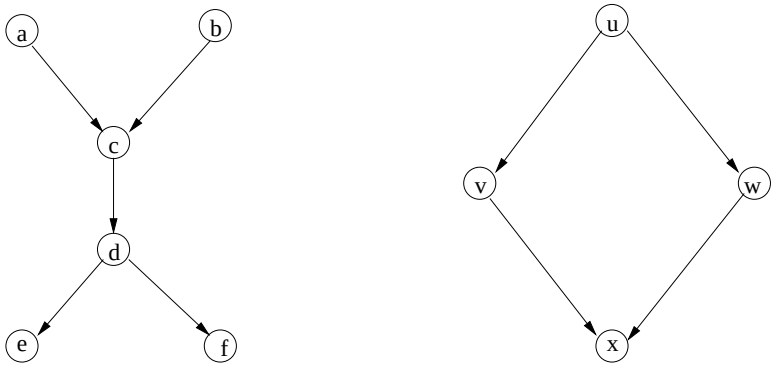


Figure 2.3. Network Complexity

variables. In the multiply connected network shown in the example, between random variables u and x, there are two paths, (i.e.) $u \rightarrow v \rightarrow x$ and $u \rightarrow w \rightarrow x$.

Approximate inference is implemented for intractable multiply connected networks using Monte Carlo random sampling algorithms [60]. Large set of samples are generated satisfying the prior and posterior probability conditions. Queries are answered based on the probabilistic observation and evidence of the query in the sample set, (i.e.) the proportion of samples that satisfy the query will determine its probability.

2.2.5. *Markov Chain Monte Carlo Algorithm*

Markov Chain Monte Carlo (MCMC) algorithm is a sampling algorithm that may be used to generate random samples satisfying a Bayesian network. MCMC generates a new random sample depending on the current sample. One hidden random variable of the current sample is randomly changed, depending upon its conditional probability distribution. Hence, a MCMC algorithm can be visualized as a random walk in state space where neighboring states are related by a change in only one random variable.

2.3. Dynamic Bayesian Networks

Bayesian networks are analytical models for probabilistic analysis of a given system, but lack the ability to express temporal dynamics. This is overcome by dynamic Bayesian networks

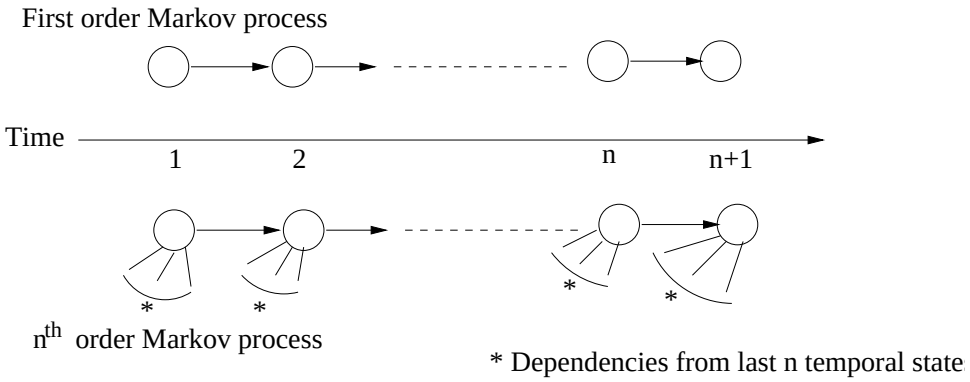


Figure 2.4. Markov Processes

that incorporate the temporal dynamics of the random variables. The temporal values of a random variable is a random walk in state space, which may include temporal dependencies.

Both dynamic Bayesian networks and hidden Markov models use the Markov assumption that the current state of a (Markov) process is dependent on a finite set of previous states [55]. The current state of a first order Markov process depends only on its most recent state. Figure 2.4 illustrates a first-order and n^{th} order Markov process. In a n^{th} order Markov process, the current state at time t_{n+1} is dependent on the previous n states over the time interval t_1 through t_n.

Figure 2.5 depicts a generic dynamic Bayesian network, using the first order Markov assumption. The intra-slice arcs/edges portray the probabilistic dependencies between the random variables, similar to a Bayesian network. The inter-slice arcs illustrate the temporal dependencies of the random variables. BN_{ij} represents the Bayesian network and probabilistic dependencies between the random variables for the time interval $[i, j]$. For example, BN_{01} illustrates the state of the system for the time interval $[0, 1]$.

2.4. Hidden Markov Models

Real world processes can be characterized as observable signals and are illustrated by developing corresponding signal models. The hidden signal source has to be understood without accessing the signal source [56]; since the observed signals are the effects of the

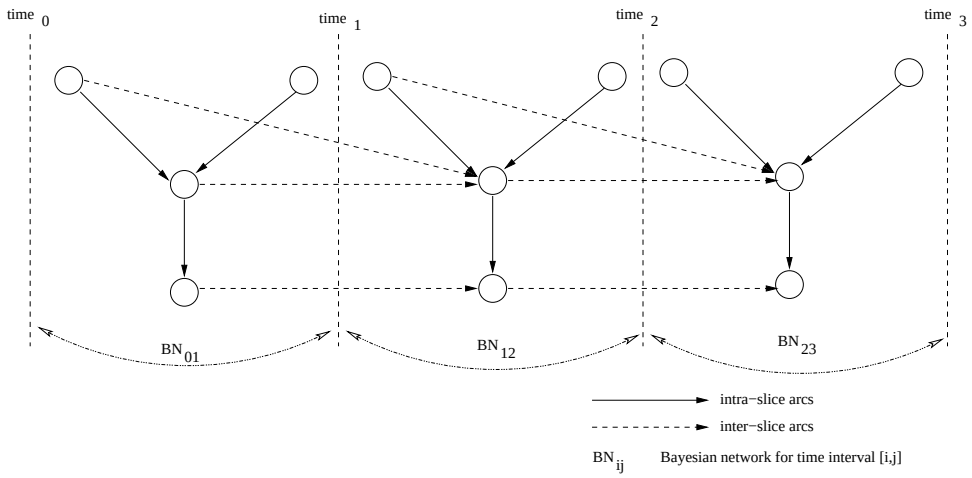


Figure 2.5. Generic Dynamic Bayesian Network

hidden signals. The causal effect relationship is depicted by the hidden signals causing the observed signals.

Random variables of a probabilistic system may be either observable or hidden, and follow the Markov assumption. The temporal values of the observed random variables are probabilistically correlated to the temporal change in values of the hidden random variables. Hidden Markov models (HMMs) are used to determine the hidden random variables from the observed random variables [42]. The forward algorithm, the Viterbi algorithm, and the forward-backward algorithm form the basic principles of HMMs [41, 56].

The different combination of values of the set of random variables defines the possible states of the Markov process. The hidden states define all possible combinations of values of the hidden random variables; correspondingly, the observed states for the observed variables. Figure 2.6 illustrates the correlation between a set of n hidden states and m observed states of a Markov process.

The initial hidden state at $time = 0$ is defined by an initialization vector π. The state transition matrix S defines the probability of all transitions between the hidden states. The confusion matrix C defines the probability of observable states for a given hidden state. Hence, a Markov process defined by the initialization vector π, state transition matrix and

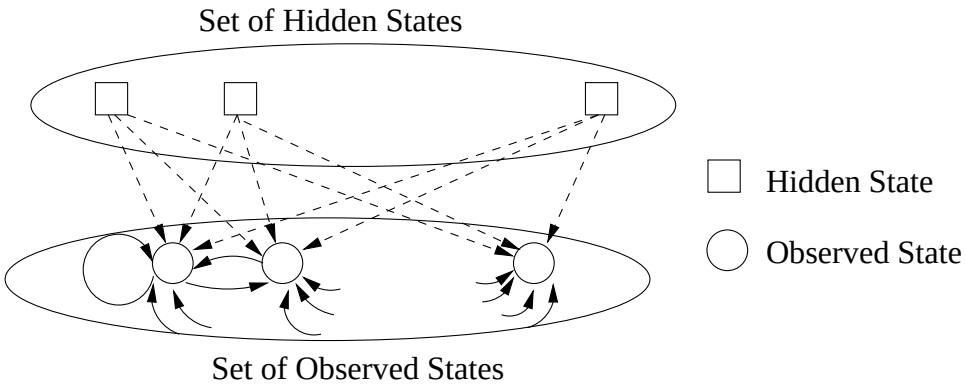


Figure 2.6. HMMs Correlate the Hidden States and the Observed States

a confusion matrix is a hidden Markov model $< \pi, S, C >$. The primary difference between a HMM and a dynamic Bayesian network is the additional set of hidden states and hidden random variables in a HMM.

2.4.1. *Forward Algorithm*

The forward algorithm evaluates the probability of an observed sequence for a given set of HMMs, and selects the best HMM that determines best for the observed sequence. The complexity of the problem is reduced by computing partial probabilities of the observed sequence. For a given observed sequence from $time = 1$ to $time = n$, the solution is built up iteratively starting from partial sequence of size 1 through n. A number of HMMs are tested on the observed sequence. The best-fit HMM will generate the highest probability of correlation to the observed sequence.

2.4.2. *Viterbi Algorithm*

The Viterbi algorithm decodes the probable sequence of hidden states that correlates to a given sequence of observed states for a given HMM. Similar to the forward algorithm, the complexity is minimized by computing best partial paths to an intermediate hidden state. The best-fit path of hidden states, that correlates to the observed states, will have the maximal probability among all possible paths through the hidden states.

2.4.3. *Forward-backward Algorithm*

The forward-backward algorithm learns the complete HMM system that correlates to a given sequence of observed states. The complete triple tuple $< \pi, S, C >$, (i.e.) the initialization vector, state transition matrix and the confusion matrix defining the HMM have to be computed. Beginning with a random assignment of the parameters, the process is repeated with continual refinement until an HMM is quantified within acceptable margins of confidence.

2.4.4. *Limitations of Hidden Markov Models*

The temporal independence of the state transition and the confusion matrices is a primary limitation, (i.e.) the matrices do not change over time. This inhibits the modeling of real world dynamics using hidden Markov models, yet useful in building good approximation models of complex systems. There are good tutorials are available that explain hidden Markov models [41, 56].

2.5. Applications

Bayesian learning has been successfully applied in the areas of medical diagnosis, weather forecasting, gaming and fault diagnosis in various domains. Pathfinder[39] is a Bayesian expert system used for lymph-node pathology diagnosis. Hailfinder[5] is a weather forecasting system for severe summer hail applied in northeastern Colorado. The chip producer, Intel, uses Bayesian networks for fault diagnosis of semi-conductor chips. Dynamic Bayesian networks have been used to detect ambulation and fall detection of people susceptible to falling down, in order to provide prompt medical assistance [76]. Hidden Markov models have been extensively used in applications for speech recognition and is a popular machine learning method in bioinformatics [42].

CHAPTER 3

INFECTIOUS DISEASES

Diseases are classified in several ways, including disease source, cause, chronic, genetic and modes of infection [67]. Infectious diseases are diseases that are communicable between members of the same species under normal circumstances, as well as transmissible from species to species. The transmission may involve an intermediate vector, such as in zoonosis. The infection source for infectious/communicable diseases are primarily viruses, bacteria and parasites.

3.1. Modes of Transmission

The modes of transmission for infectious diseases are airborne, intestinal, open sores, zoonosis and fomiteborne [67]. Airborne transmission diseases, including influenza and pneumonia, are spread through the dissemination of the disease causing pathogen from an infectious person. The infectious person releases the pathogen into the environment during the acts of coughing and sneezing. The pathogen cloud, when inhaled by susceptible individuals in the vicinity are vulnerable to infection. Intestinal transmission is coupled with water-borne transmission, such as in cholera. It occurs through secretion of the pathogen by infectious individuals and leading to scenarios of contaminated food intake of susceptibles. Open sores transmission occurs by open wounds that discharge the pathogen upon direct contact between infectives and susceptibles. Examples of open sores or lesion transmission are AIDS, anthrax and smallpox. Zoonosis transmission involves an intermediate vector, and the transmission is from the vector animal to humans. Examples of zoonosis include malaria wherein the vector is mosquitos, and lyme disease whose vector is ticks. Fomiteborne transmission involves an intermediate agent in the form of inanimate objects, that aid in transmission upon direct

contact with the susceptible individuals. Examples of such diseases include influenza and tuberculosis.

3.1.1. *Infection Timeline*

Infectious diseases are characterized by the time periods of incubation, latency, infectivity and recovery. Incubation period is the period from the onset of the pathogen entering the human body to the onset of symptoms. The latent period is the period from the onset of the pathogen entry until the infected individual is capable of disease transmission onto other individuals. After the infectious period, when the infected individual is no more capable of pathogen transmission, the recovery period starts. Upon full recovery, the individual may or may not acquire immunity to the disease. If lacking immunity, the recovered individual is susceptible to re-infection.

While the incubation period may end before or after the end of latent period, it is almost a necessary condition for the incubation period to terminate after the end of the latent period. This intermediary period between the termination of latent and incubation periods provides a timeline for infection spread by an infectious individual without the knowledge of being infected due to lack of any symptoms. In contrast, if the incubation period ends before latent period terminates; upon onset of symptoms, the infected individual is likely to quarantine and isolate himself. The infected individual will continue to restrict his behavioral interactions until recovery. This change in behavior of infected individuals will most likely prevent disease outbreaks. In case of epidemic outbreaks, it is imperative that the infectious period will probably start before the onset of symptoms.

3.2. Preventive Medicine

The best cure and remedy for disease outbreaks is to address the root of disease causes and primarily prevent the disease onset. Preventive medicine is the key to minimize the disease incidence and prevalence, as well as promote a healthy living environment. Health

care education and awareness, and enhancing the immunity by vaccinations and immunizations are the prime means of implementing preventive medicine in a community.

3.2.1. *Vaccine Trials*

The immunity level of an individual determines his/her susceptibility to diseases. Active or acquired immunity results when antibodies are produced by the individual in response to the entry of the disease pathogen in the body [67]. Acquired immunity is triggered in response to vaccinations, that contain antigens which stimulate the human immune system to produce antibodies. Passive immunity is due to the transmission of the mother's immunity to her child through placental transfer.

Pharmaceutical companies are regulated by the Food and Drug Administration (FDA) [33] governing body in the development of vaccines. There are strictly regulated four phases of trials that every vaccine has to undergo for approval of mass distribution.

Phase I trials are carried on a smaller group of people of around 20-80 [26]. The focus of this trial is to determine the safety and effects of the vaccine on the body. There is minimal interest on the therapeutics of the vaccine during phase I trials. Phase II trials are administered over a larger group of people of 100-300 to evaluate the safety and determine the acceptable dosage levels of the vaccine. Phase III trials are administered over larger groups of people of 1000-3000 to determine the therapeutics and efficacy of the vaccine. The safety levels of the vaccine and side effects are evaluated in relation to known treatment measures. Once vaccine pass successfully phase III trials, FDA approves the use of the vaccine and phase IV trials begin in a surveillance program on the long term effects on the vaccine on different demographic and geographic populations. The results of phase IV trials helps in evaluation of the benefits and risks of the vaccine, as well as expand the vaccine use to newer populations optimally.

3.2.2. *Biological Safety Levels*

In order to access the plausible impact and understand the biological mechanisms of infectious disease agents and develop potential vaccines, research is carried out in biological labs at different degrees of safety levels The labs are labeled at four different levels of safety, namely BSL-1, BSL-2, BSL-3 and BSL-4 [2]. Biological safety level-1 labs (BSL-1) handle experiments of disease pathogens that are not hazardous to the environment and minimal risk of infection to the investigating personnel. The infectious agents are not known to cause disease among healthy individuals. The labs do not need any special precautionary measures. Escherichia coli strain K12 and bacillus thuringiensis are examples of micro-organisms tested in BSL-1 labs. Viruses causing influenza and hepatitis-B are some of the infectious agents characterized in BSL-2 labs. Research scientists direct the laboratory work of the lab personnel and caution is applied to avoid contamination. The disease pathogen that causes tuberculosis is analyzed in BSL-3 labs. The pathogens that are researched in BSL-3 labs can cause serious ailment upon exposure. Training in working with lethal disease agents are accorded to the laboratory personnel. BSL-4 labs experiment in infectious agents that cause diseases of high mortality rate, such as Ebola hemorrhagic fever. This is the highest level of safety accorded in infectious diseases labs and implement strict safety procedures.

3.2.3. *Reportable and Non-reportable Diseases*

Diseases are categorized as reportable and non-reportable diseases. Reportable diseases have to be recorded by the public health department. Physicians and clinicians who diagnose newer cases of reportable diseases are required to report the same to the public health departments. Public health departments at different levels of hierarchy exchange and aggregate the recorded information. Statistical analysis of the collected data lend to measuring the disease prevalence and incidence status in the population as well as derive useful inferences. HIV is a reportable disease, while influenza is a non-reportable disease. Influenza being non-reportable compounds the problem of determining the corresponding prevalence and incidence

levels. This leads to difficulty in identifying the points of control that can thwart influenza outbreaks.

3.2.4. *Influenza Like Illness*

Influenza is an air-borne disease that affects the respiratory system. Symptoms include cough, sore threat, head ache, and fever. These symptoms are common to several respiratory diseases, including pneumonia. Many respiratory diseases, including influenza, are non-reportable. Due to lack of definite measures of influenza impact in a geographic region, surveillance programs are initiated to record the measure of Influenza Like Illness (ILI). Flu medication sales coupled with recorded cases of ILI, supplied by proactively participating medical personnel, help in determining the prevalence and incidence levels of ILI. When higher than normal incidence measures are observed, vaccination programs are triggered to curtail the spread of ILI diseases, including influenza. Monitoring ILI arises due to the need to infer the real levels of manifestation of influenza, including incidence and prevalence, and prevent influenza outbreaks.

3.2.5. *HIV Vaccine*

Vaccines are effective in prevention of epidemic outbreaks and eradication of diseases, such as the vaccinia vaccine that eradicated small pox globally [67]. Human Immuno-deficiency Virus (HIV) adversely affects the immune system and leads to AIDS. An effective vaccine for HIV is yet to be discovered. HIV is a retro virus and replicates with a higher mutation rate. Hence, the genetic makeup of the virus is continually changing. This rapidly diverging HIV genome and resulting variations in the encoded proteins make the task of producing an effective vaccine complex [44]. HIV transmission dynamics are well studied and the primary modes of transmission are through sexual practices, shared needles, blood transfusion and vertical transmission from infected mother to foetus. The lack of an effective vaccine makes it imperative to promote health care education and awareness programs promoting preventive measures to HIV infection.

3.2.6. *Health Care Education*

The primary foundation of preventive medicine is in the promotion of health care education and awareness programs in the community. The health care programs promote healthy living by raising the awareness of factors that cause disease spread and the corresponding prevention measures. Life style changes oriented towards enhancing the healthy well being of individuals and the community are campaigned by the public health personnel. In case of infectious diseases, awareness of symptoms and spread dynamics will raise the community vigilance in preventing the disease transmission from infectious individuals. Isolation and quarantine of infectious cases may be warranted in special cases to curtail the disease spread. Health care awareness campaign is an invaluable asset in promoting the prevention methodologies that curtail the transmission of infectious diseases.

CHAPTER 4

MATHEMATICAL AND COMPUTATIONAL EPIDEMIOLOGY

Epidemiology is the study of diseases' characteristics, their causes and progression, and control measures in a population [67]. The study of causes and progression of infectious diseases in different populations have been the prime focus of study in epidemiology; although chronic disease studies are becoming prominent in the modern era. The domain of epidemiology includes several sub-domains, such as field epidemiology, epidemiological genomics, infectious disease epidemiology, social and behavioral epidemiology, and surveillance.

Field epidemiology explores the process of collecting disease prevalence data from different demographic populations in a geographic region to provide useful insight on the disease causes in the population. Public health laws and legal issues are taken into account while administering the surveys and collecting the samples. Epidemiological genomics investigates the role of genetics in populations of different environments and geographic regions. The corresponding variations in incidence and prevalence of different diseases are analyzed. Infectious disease epidemiology studies the progression of infectious diseases and address prevention measures through vaccination strategies. The focus of this dissertation is on the design of computational models for analyzing infectious diseases. Social and behavioral epidemiology analyzes the risk behavior levels that enhance the disease incidence in different demographic populations. Surveillance is the process of data collection at different hierarchical levels, including public health departments at county, state and national levels as well as hospitals, physicians, and related health care information sources. Surveillance attempts to identify the critical points of disease control and prevention and enhance the public health decision making process.

4.1. Concepts in Epidemiology

Etiology is the study of diseases upon analysis of epidemiological data, and includes forging the expertise from diverse disciplines to gain insight into the causative factors. Pathogen are disease causing living organisms, including viruses and bacteria, that infect susceptible individuals. Susceptibles are individuals who lack disease immunity and are at risk of infection from the disease. Infectives are the individuals who are actively infected by the disease and are infectious, (i.e.) they are capable of transmitting the disease causing pathogen to other suscpetibles. Removals are individuals who are immune to the disease and may be in the process of recovery from the disease. The immunity may have been induced by either vaccination or by prior exposure to the disease. Although not formally defined, virulence is often used to describe the pathogen strength in causing the disease and the capacity to inflict severe illness. Vaccine efficacy is a measure of a vaccine in successfully reducing the risk of an individual in contracting the disease.

The average disease prevalence in a population or demographic sub-group illustrates the endemic scenario for the disease. Epidemic is a disease outbreak in a population with a relatively higher number of infectives than the normal expected levels. Pandemic is a large scale epidemic that crosses boundaries across nations, and a cooperative concerted effort from public health departments worldwide is needed to counter the disease spread.

Primary case is the first disease case in a population and acts as the source of infection for transmission to secondary cases. Index case is the first reported case to the public health department. Mortality is a measure of fatal cases or death due to the disease. Morbidity refers to the impact of illness in a given population and is measured by incidence, prevalence and attack rate. Incidence is the rate of newly infected cases in a population over a period of time. Attack rate refers to the different incidence rates for different sub-groups of a population. Hence, the cumulative measure of attack rates will determine the incidence. Prevalence gives

the complete measure of infected cases in a population. Quarantine is the process of isolation of infectives in order to curtail the spread of the disease to other susceptibles.

4.2. History

The human race has exacted astronomical casualties to epidemic outbreaks. The study of epidemics dates back to the 4th century B.C. era of Hippocrates, who speculated on the causes of disease outbreaks. The 14^{th} century Europe lost a quarter of its 100 million population to Black Death. The fall of the Aztecs empire in 1521 was due to smallpox that eradicated half of its $3\frac{1}{2}$ million population. Fracastorius proposed that disease transmission between people is due to a living contagion in 1546. This proved to be true and infectious diseases are spread by either viruses or bacteria from person to person. The 17^{th} century witnessed the commendable efforts of John Graunt and William Petty in using the London Bills of Mortality for biomedical statistics [12]. The Broad Street pump investigation by John Snow identified the source of water contamination that lead to a cholera outbreak at London in 1855. This was a pioneering progress in the use of spatiotemporal mapping of disease cases to identify the roots of the problem. The pandemic influenza of 1918 caused a fatality of 20 million in twelve months. More recently, the Severe Acute Respiratory Syndrome (SARS) outbreak of 2003 highlighted the rapid spread of an epidemic at the global level. The outbreak, emanating from a small Guangzhou province in China, spread around the world requiring a concerted response from public health administrations around the world and World Health Organization (WHO) to curtail the epidemic [40]. In recent times of the early 21^{st} century, the threat of avian influenza flu pandemic resulting in severe casualties has caught the prime attention of the world. Health organizations around the world, include WHO and Centers for Disease Control and Prevention (CDC) are proactively working on prevention plans to minimize the ill effects with the available health resources. In the 20^{th} century, the better standards of healthy living in conjunction with the rapid advances in biomedical sciences has resulted in diminishing rates of morbidity and mortality for the infectious diseases in the

advanced western countries; and chronic diseases are the major health care concern. But in the developing countries, the infectious diseases are still the prime diseases of concern and result in higher rates of morbidity and mortality. Health surveillance, data collection and analysis using the cyber infrastructure is possible in today's modern world. This facilitates the use of high performance computational models to draft informed decisions and public health policies.

4.3. Healthcare Levels and Modes of Prevention

Healthcare services are of three different types, namely, primary care, secondary care and tertiary care [60]. Primary care refers to the first time visits to a physician or clinic. Secondary care refers to the health care offered in hospital or clinical settings, and may include minor or simple surgeries. Tertiary care involves sophisticated health care technologies and includes complex surgeries as well as special intensive care.

Prevention modes are of three kinds, namely primary, secondary and tertiary preventions. The three prevention modes strive to lower the dependency on the corresponding health care services and optimize the available health care resources. The focus of primary prevention is to minimize newer infections, thereby halting their use of primary health care services. Secondary prevention targets to reduce the use and dependency on secondary health care; similarly, tertiary prevention lowers the use of tertiary health care. Primary prevention promotes the awareness of healthy living, nutrition and sanitation in order to improve the overall community health. Secondary prevention includes health screening programs targeted at identifying individuals at early stages of a disease. The infected individuals are referred to the appropriate health care. Disease detection at early stages of development is expected to speed up the medication and recovery process, thereby preventing the disease entering the severe stages of development. Tertiary prevention includes rehabilitation programs for disabled individuals and pave a recovery path from their disorders. The focus is to lead the individuals to lead independent lives, without continual dependence on health care resources.

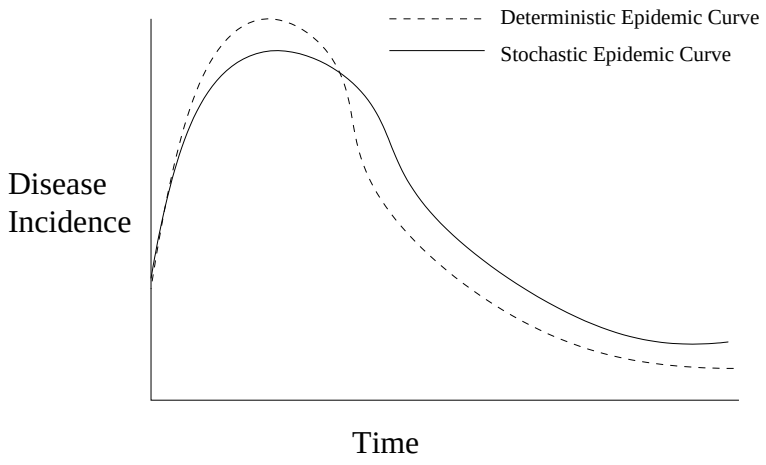


Figure 4.1. Epidemic Curves for Deterministic and Stochastic Models

While epidemiology identifies the factors for higher disease rates in different subgroups of a geographic region, preventive medicine is the key for better utilization of health care resources by reducing the incidence of diseases, thereby enhancing the public health.

4.4. Mathematical Epidemiology

The early 20^{th} century laid the foundations of the mathematical theory of epidemics [38],[59]. The initial work from 1900 to 1930 in epidemics had a deterministic character. The probability aspects of the different processes were not included. From 1930, binomial distributions were used for disease cases and stochastic modeling got its roots in the study of epidemics. Epidemiological triangle model [51, 67] for infectious diseases comprises of the host, environment, and agent. The inter-relationships between the above parameters are included in the evaluation and analysis of the diseases.

4.4.1. *Deterministic and Stochastic Models*

The deterministic model used basic differential equations to model the spread of an epidemic. The behavior of the stochastic model is similar to the deterministic model for larger number of susceptibles and infectives. The stochastic models differ from the deterministic models for epidemic studies of smaller granularities in area. Stochastic models provide a closer

real life reasoning and modeling of the spread of infectious diseases, by introducing the probability metrics into the deterministic model. The epidemic curve illustrates the incidence rate in a population of a given area over a period of time. Figure 4.1 shows the epidemic curves of both the models. The stochastic curve has a lower peak compared to the deterministic curve, and proves to be the better model for the study of infectious diseases' outbreaks in the real world setting. High performance computing is a requisite to realize the fuller potential of the stochastic models. The stochastic models and the analysis of the epidemic curves have proved to be primary mathematical framework of visualizing infectious diseases outbreaks, including the 21^{st} century [8].

4.5. Susceptibles-Infectives-Removals Model

Mathematical models of infectious diseases are based on the principles of *susceptibles, infectives, and removals*, namely the SIR model. *Susceptibles* are those individuals in a population who can be infected by the disease under study. *Infectives* are those individuals who have been infected by the disease and are infectious. *Removals* include all individuals that are incapable of transmitting the infection, and are either recovering, fully recovered, expired from the disease, or immune to the disease. In complex models, the removals who recover may revert to susceptibles. In case of influenza, a recovered individual can not be infected by the same influenza strain due to acquired immunity during the infection. Nevertheless, the recovered individual may remain susceptible to other influenza strains. Figure 4.2 shows the transient curves for the susceptibles, infectives and removals during the course of a disease epidemic in a given population.

The Kermack-McKendrick Threshold Theorem [12] is the basis for the SIR model. A continuous influx of susceptibles is a requisite for sustained infection in a population. This is the case of endemic diseases, such as tuberculosis, which prevail in a community at all times. The model is based on the presumption of a closed population, assuming that the

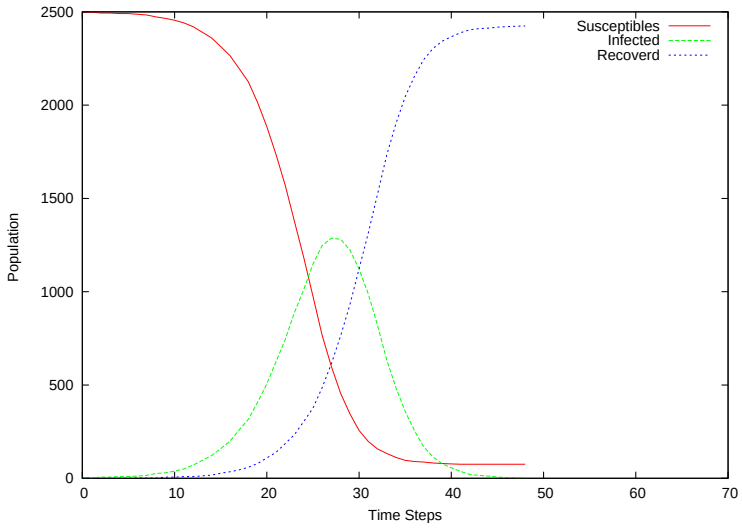


Figure 4.2. SIR Epidemic Curve for a Sample Population

epidemic spreads rapidly enough that the changes brought in by births, deaths, migration and demographic changes are negligible [8].

During the start of a disease epidemic, the total population comprises of susceptibles, excluding those that have inherent immunity to the disease. The *primary case* is the first infected individual and is the source of the infection. During the infectious period, the infection is passed on to some susceptibles, who interact with the primary case in close proximity to contract the infection. This triggers the cycle of infections spreading through the population. Once the infected individuals become non-infectious, they move over to the removals category. The underlying assumption is that the total number of susceptibles (S), infectives (I), and removals (R) is a constant (Eq. 4) and holds true during the course of the disease outbreak. The rising infection on reaching the peak starts to recede due to the decrease in the number of susceptibles, and diminishes eventually.

(4) $$S + I + R = constant$$

(5) $$\begin{aligned} \frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= +\beta SI - \gamma I \end{aligned}$$

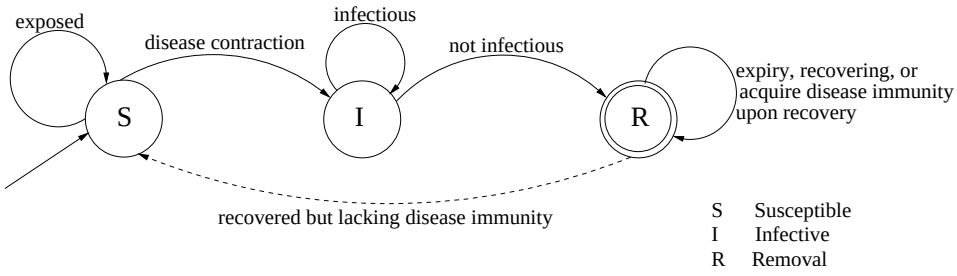


Figure 4.3. SIR/SIRS State Diagram

$$\frac{dR}{dt} = +\gamma I$$

The random mixing of susceptibles and infectives [8] is given by the multiplicative product, $S * I$. β defines the transmission coefficient [7] based on contact rate between susceptibles (S) and infectives (I), and infectivity of the disease. γ defines the rate of infectives (I) becoming non-infectious. Hence, the average duration of infectivity is given by $1/\gamma$ [8]. The set of differential equations used in classic SIR model for a closed population are shown in equation 5. The transfer rates of individuals from $S \rightarrow I$ and $I \rightarrow R$ are given by dS/dt and dR/dt respectively. The rate of change of infectives is given by dI/dt.

The SIR/SIRS state diagram (Fig. 4.3) illustrates the course of a disease in an individual. A susceptible individual may be exposed to a disease pathogen and continue to be in the susceptible state. A susceptible becomes an infective, once the susceptible is able to transmit the pathogen onto others. The recovery state begins once the ability to infect ceases. The individual continues the state of recovery from the disease, or may expire. On full recovery, the individual may acquire full immunity from disease, and hence is no more susceptible to the disease (SIR model). The individual reverts to a susceptible on full recovery when lacking disease immunity (SIRS model).

4.5.1. *Related Work*

In *susceptibles, infectives, and susceptibles* (SIS) model, the infectives upon recovery return to the susceptibles category. The SIS model for spread of infectious diseases has been

analyzed for non-uniform population densities, thereby accounting for mobility of individuals [18]. The *susceptibles, exposed, infectives, and removals* (SEIR) model includes the class of exposed individuals who have been exposed to the disease, but are in the latent stage, and hence are not capable of infecting the susceptibles. When the exposed becomes an infective, the individual is capable of transmitting the disease to the susceptibles. A 4-dimensional SEIR epidemic model has been considered and the criteria for stable equilibria is established in [30].

The spatial and temporal correlation of influenza epidemics in the United States, France, and Australia from 1972 to 1997 has been analyzed in [73]. The results indicate a high correlation between United States and France, but irregularity in the patterns between Australia and the other two countries. Demography is highlighted as one of the reasons that may be causing the discrepancies, and recommends mathematical modeling for further investigation.

Geographic-environmental re-infection modeling simulator (GERMS) [6] is a toolkit for modeling transmission of infectious diseases. The model is equipped to handle heterogeneous populations with varied socio-geographic characteristics, complex interactions among individuals, and infection specific features, such as transmission probabilities

4.6. Cellular Automata

Cellular automata have been used for several decades [35] in the domain of computational models. Infectious disease modeling uses the cellular automata paradigm to analyze the spatial progression and distribution of diseases [71, 72].

The basic unit of cellular automata is a cell and may represent an individual or a subpopulation. Each cell can be characterized with state and likelihood risks for exposure and contracting the disease. The spatial disease progression is modeled via the cell neighborhoods, wherein each infected individual may diffuse and spread the disease to the adjacent neighbors. Figure 4.4 illustrates the update of the center cell as a function of its adjacent eight neighbors for a two-dimensional cellular automata. The center cell is updated with respect to a Moore neighborhood model.

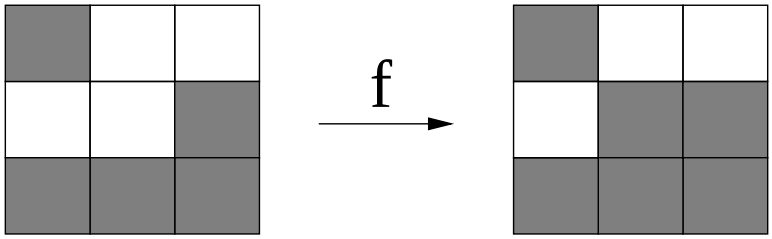


Figure 4.4. Cellular Automata Update from time step *t-1* to *t*

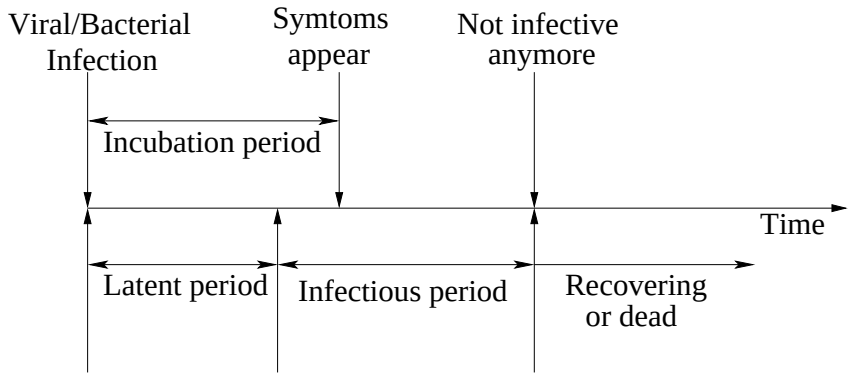


Figure 4.5. Infection Time-line

Similar to the Susceptibles-Infectives-Removals (SIR) model, state *S* for susceptible is defined as the state in which the cell is capable of contracting a disease from its neighbors. In the infectious state *I*, the cell is capable of transmitting the infection to its neighbors. In the recovery state *R*, the cell is neither capable of passing on the infection nor capable of contracting the infection. On full recovery and acquiring of disease immunity, the cell shall continue in the removal state (R). The time-line for infection is illustrated in Fig. 4.5.

The state of the center cell transitions to a state, which is in majority among the cells in the neighborhood and itself. The update rule determines the deterministic or stochastic behavior of cellular automata. Stochastic behavior is seen by probabilistic update rules in non-deterministic state transitions.

4.6.1. *Related Work*

The earliest example of use of cellular automata is Bailey's lattice model [13] for the spread of diseases from micro-level interactions. Schönfisch has analyzed varied cellular automata models to study the dynamics of epidemics [61]. Di Stefano *et al* [64] have developed a lattice gas cellular automata model to analyze the spread of epidemics of infectious diseases. The model is based on individuals who can change their state independent of others and can move from one cell to other. However, this approach does not consider the critical factor of the infection time-line. Fu has used stochastic cellular automata to model epidemic outbreaks that take into account the heterogeneous spatiality [34]. Situngkir has developed a dynamic model of spatial epidemiology to study avian influenza disease in Indonesia and uses cellular automata for computing analysis [62]. Bonabeau has studied the spatio-temporal characteristics of influenza outbreaks in France. The study infers that the global transportation systems of the modern world lends to propagation of influenza epidemics dominated by a global mixing process in comparison to local dynamic heterogenities [19]. Duryea has analyzed spatially detailed epidemic models using probabilistic cellular automata for heterogeneous population densities in a region [29]. Benyoussef has used a one-dimensional lattice model and a two-dimensional automata network model to illustrate the spatial spread of rabies among foxes [16]. Fuks describes a SIR epidemic in the spatio-temporal domain via a lattice gas cellular automaton for both human and animal populations. Vaccination strategies are incorporated and dynamics of the disease spread are investigated in relation to the spatial distribution of the vaccinated individuals [36]. Global stochastic cellular automata paradigm and field simulation is used to study the dynamics of infectious diseases' epidemics, and vaccination strategies in controlling them [49, 71, 72].

4.7. Agent-based Models

In agent-based modeling for simulating the spread of infectious diseases, people are represented by agents. A mobile agent is an autonomous entity with an independent agenda and

are capable of movement similar to normal people. The actions of the agents are driven by their desires, and correspondingly the desires will define the agents' mobility patterns. Similar to the Susceptibles-Infectives-Removals (SIR) model, the mobile agents can be in either of the three states, susceptible S, infective I or removal R. When an infectious agent is introduced into the population, the disease progression is plausible when a infectious agent comes in contact with a susceptible agent. The mobility patterns of the agents correspond to the everyday mobility patterns of real individuals.

Agents can be further modeled to analyze the diffusion of disease pathogens. In addition to people being modeled as agents, disease pathogens are also modeled as agents. Infectious person-agents will be disseminating the pathogen-agents in the environment. A susceptible agent in close proximity to the clouds of pathogen-agents risks the contraction of the infection.

Hence, interactions and disease transmissions are primarily of two kinds, namely, direct and indirect. Direct transmission is from an infectious person-agent to a susceptible person-agent; while indirect transmission is from infectious person-agent to environment and environment to susceptible person-agent. Agent based modeling relates well to the visualization of real world scenarios, yet the fidelity and the scalability of the simulations in representing the real world complexity is to be investigated.

4.7.1. *Related Work*

AIDS/HIV epidemics are studied by use of agent-based models to analyze the disease spread in the population as well as study the immune levels in an individual in response to the infection [20]. Epidemic simulation models have been surveyed [11] and agent based models were implemented via agents following rigid rules of interaction. BioWar is a agent-based system that analyzes the disease spread, treatment, and recovery, by porting principles of interactions from social, knowledge and work networks [17]. A graph theoretical flow model

is developed to simulate the disease spread and analysis in a large urban population. The contact patterns are derived from the transportation systems [31].

4.8. Bayesian Analytic Models

Bayesian learning and probabilistic reasoning is applied to public health data to infer the health parameters of significance. A Bayesian network can be constructed to study the stochastic dependencies among the modeled demographic and health indicator variables. The network provides a compact representation of the collected health data for probabilistic data analysis and predictions. In addition, inferences learned from a disease outbreak model of a geographic region can be ported to newer geographic regions to estimate the expected levels of disease prevalence. The stochastic nature of probabilistic dependencies among the studied variables in a Bayesian network is well suited for analysis of the real world dynamics of disease progression.

Bayesian networks lack the temporal flow of information. Dynamic Bayesian networks overcome this deficiency and aid in understanding the temporal characteristics of disease progression for a given population. Bayesian networks are built for continual time periods and correlation between Bayesian networks are derived using the Markov assumption. Network complexity increases for higher order Markov process and time intervals of finer granularity. The choice of the time intervals needs to be optimal to gain higher fidelity in the temporal analysis of disease progression while maintaining minimal complexity of the network.

Disease progression is intrinsically coupled with the social behavioral interactions in a population. Hidden Markov Models can be used to gain a better understanding of the social behavioral interactions. This will lend to a better insight on the hidden dynamics of disease spread for a given population.

4.8.1. *Related Work*

Microfilariae had been studied in the Amazonian focus of onchocerciasis (river blindness) to identify the communities that need priority ivermectin treatment[22]. A Bayesian hierarchical model for human onchocerciasis was developed to investigate the role of individual and community characteristics in the infection. The model aids in research and control planning for the public health department as well as in its policy decision making. Bayesian analysis for health technology assessment has been investigated[63], and highlights the practical advantages of the Bayesian approach in handling complex interrelated problems. Bayesian classifiers are used in the Real-time Outbreak and Disease Surveillance (RODS) system[68], a computer based public health surveillance system that detects disease outbreaks. RODS had been used in the 2002 Winter Olympics. Pennsylvania and Utah currently use RODS for public health surveillance.

In social science hypothesis testing, the increase in independent variables for regression models leads to misleading errors, while Bayesian approximation reduces the uncertainty in error[57]. Bayesian concepts are used to calculate the risks of leukemia following chemotherapy for hodgkinks disease, based on case-control studies[9]. Bayesian monitoring of critical factors in cancer related clinical trials, such as toxicity and quality of life measures, led to higher accuracy[32].

An epidemiological model using Bayesian analysis has been developed for the disease, plasmodium falciparum malaria, in Ndiop, Senegal[21]. The incidence of cancer in multiple cities has been collected from the survey data for the state of Sao Paulo, Brazil[10]. A correlation analysis using Bayesian methods between the multiple cancer sites estimated the cancer rate in a given area. The results had a better precision compared to the prevalent methods. Bayesian learning is used to infer the dependency of the demographics on the incidence of diseases in different geographic regions [4].

Hidden Markov models with an exponential-Gaussian mixture have been used for automated detection of influenza epidemics [58]. A Monte Carlo simulation using a Markov model is implemented to study the infection models, that occur naturally, such as influenza whose viral pathogen spreads through a susceptible community, or induced deliberately, as in the case of bio-terror attacks [52].

CHAPTER 5

PROBABILISTIC ANALYSIS OF EPIDEMIOLOGICAL DATA

Bio-safety gains importance in a world, where new and re-emerging local diseases threaten all mankind. Disease monitoring can not continue to be fragmentary and inadequate, focusing on small spatial domains [1]. As the recent outbreak of SARS showed, effective surveillance is critical to an effective defense against disease threats, and requires consideration of huge volumes of data from other parts of the world. Mathematical and computational models for the study of global disease dynamics will facilitate adequate what-if analysis, and use the power of high performance computing infrastructure [14, 46]. Probabilistic reasoning using Bayesian learning is presented to design predictive disease models. The models will analyze disease progression in varied demographic and geographic settings, and act as a test platform to analyze the different prevention strategies.

5.1. Bayesian Networks

The inter-relationships of demography and the prevalence of symptoms and diseases in a geographic area is illustrated through a Bayesian network (Fig. 5.1). The Bayesian network

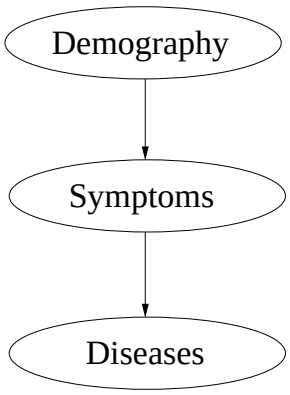


Figure 5.1. Bayesian Network Illustrating the Relationships between Demography, Symptoms and Diseases

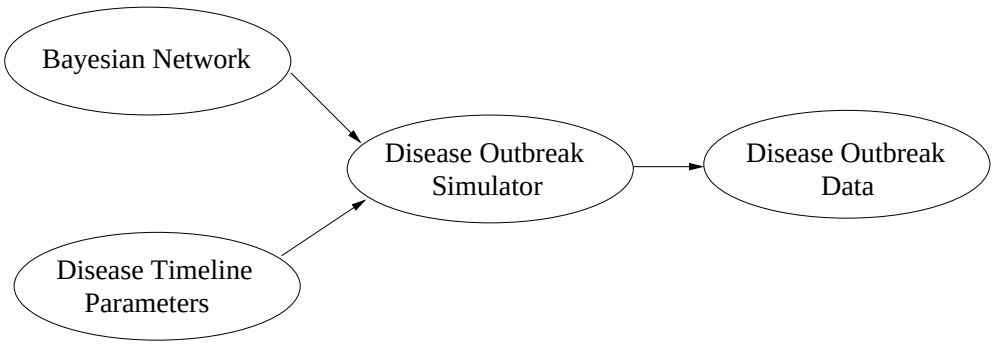


Figure 5.2. Disease Outbreak Simulator

can be learned from real disease case data, or can be developed to model a synthetic disease outbreak. If real data is available for an epidemic outbreak, the learning can be ported onto a Bayesian network. It can be further developed into a dynamic Bayesian network, that can include the temporal flow of the disease progression. In most cases. real disease data is not readily available, and when available, the data is sporadic and incomplete. This necessitates the need for an disease outbreak simulator. The Bayesian network acts as a platform to embed various outbreak scenarios for different diseases. In general, an abstract disease can be modeled and studied for its spread and progression in different demographic and geographic settings.

5.2. Disease Outbreak Simulator

The Bayesian network and the disease timeline characteristics are input parameters to the disease outbreak simulator. Disease time-line parameters include incubation, latent, infectious, and recovery periods. The simulator processes the Bayesian network coupled with the temporal features of the disease, and generates the epidemic outbreak data, as shown in Fig. 5.2.

The outbreak simulator is an useful tool to model the progression of infectious diseases. The diseases can be real or hypothetically developed synthetic diseases. This enables a what-if analysis for the public health professionals and enhances the understanding of disease dynamics. The deficiency of complete real disease data can be overcome by the reverse

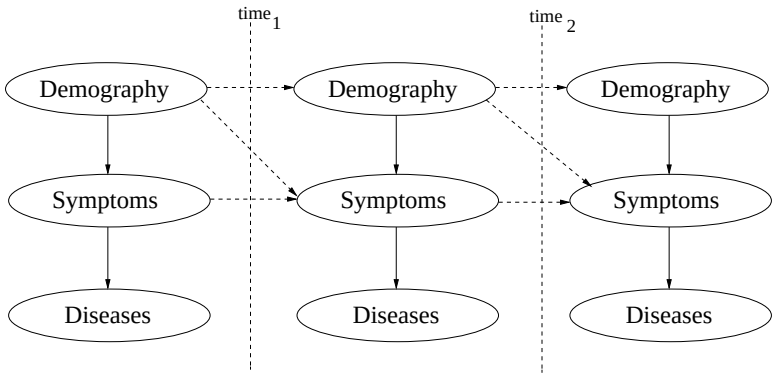


Figure 5.3. Dynamic Bayesian Network Analysis of Disease Progression

engineer processing of the outbreak simulator, that results in data for disease progression in both the spatial and temporal domains.

5.3. Probabilistic Inferences

The outbreak data incorporates the temporal disease incidence and prevalence for a given geography and demographic composition. In order to decipher the temporal probabilistic inferences, dynamic Bayesian networks are developed to analyze the outbreak data. Further, a Hidden Markov Model can be used to infer the hidden epidemic properties of the analyzed disease.

5.3.1. *Dynamic Bayesian Networks*

Dynamic Bayesian networks import the temporal feature into Bayesian networks, thereby addressing the issue of temporal analysis and progression of dynamic data. The outbreak data will be analyzed in time slices and the temporal dependency in the disease progression are illustrated through continual time slices.

Figure 5.3 shows a dynamic Bayesian network for disease progression analysis for a specific geography and demography over a period of time. The Bayesian network of each time slice illustrates the disease dynamics during that time period. The temporal flow of diseases is studied by correlation of the Bayesian network for a time interval with respect to the Bayesian network in adjacent time intervals.

The analysis of the temporal disease progression through the dynamic Bayesian network will aid in the identification of the probable paths of disease diffusion through the demographic space. This will lead to identification of key demographic control points to recede the disease spread. Public health resources, addressed to the prioritized control points, will have the maximal effect in the reduction of disease incidence and prevalence.

5.3.2. *Hidden Markov Models*

Disease incidence and prevalence differ for different geographic and demographic compositions. The epidemic properties of a disease are determined by its rate of transmission, virulence, infectivity and the time periods of incubation, latency, infectiousness and recovery. While the epidemic properties are not readily evident, they determine the disease impact for a given geographic area and demographic mix of the population. Hence, it is imperative to decipher these epidemic properties. Hidden Markov Models (HMMs) hold the key to identify the hidden epidemic disease properties through probabilistic analysis of the disease data. Once the epidemic properties are well understood, the task of predicting the disease spread for newer demographic and geographic areas is feasible.

The epidemic properties are the hidden parameters to be identified by the machine learning approach of HMMs. The accuracy in determining the epidemic properties is dependent on the data quality. Epidemic properties learned from varied geographic and demographic scenarios will enable in the development of a robust HMM, that will effectively predict outbreak scenarios for newer demographic and geographic settings.

5.3.3. *Disease Progression Predictive Models*

Predictive disease modeling will aid in efficient allocation of public health resources and deployment of preventive measures in order to control the disease spread. The predictive model provides a framework to identify the critical points of control, prevention and surveillance of disease progression. The pubic health measures can be optimally deployed ahead of disease outbreaks and efficiently curtail the disease outbreak.

CHAPTER 6

DEMOGRAPHIC RISK ANALYSIS FOR INFLUENZA

[1] A Bayesian network is developed to embed the probabilistic reasoning dependencies of the demographics on the incidence of infectious diseases. Influenza epidemics occur every year in both hemispheres during the winter. The Bayesian learning paradigm is used to create synthetic data sets that simulate an outbreak of influenza for a geographic area. The Bayesian prior and posterior probabilities can be altered to represent an outbreak for various demographics in different geographic regions. Epidemic curves are generated, via time series analysis of the data sets, for the temporal flow of influenza on different variants of the demographics. The analysis of the demographic-based epidemic curves facilitates in the identification of the risk levels among the different demographic sections. Spread vaccination lowers the impact of the epidemic, depending on the efficacy of the vaccine. The model is equipped to analyze the effects of spread vaccination and design vaccination strategies, that optimize the use of public health resources, by identifying high-risk demographic groups. The

[1] This chapter is reprinted from: K. Abbas, A. Mikler, and R. Gatti. Temporal Analysis of Infectious Diseases: Influenza. Proceedings of the ACM Symposium on Applied Computing, SAC'05, Sante Fe, NM, March, 2005.

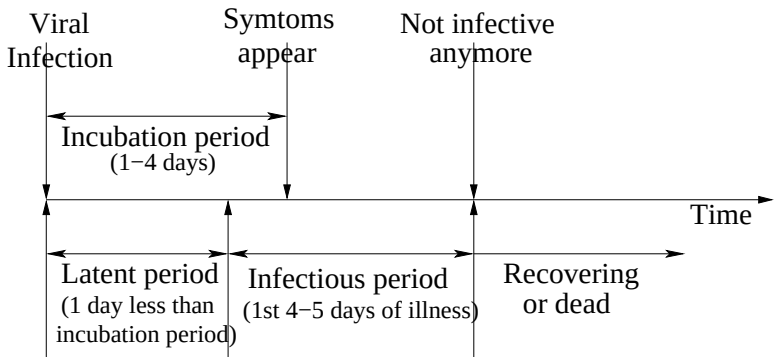


Figure 6.1. Infection Time-line for Influenza

results show that application of the vaccine in the order of risk levels will further lower the epidemic impact as compared to uniform spread vaccination.

6.1. Influenza

Influenza is an infectious disease, which is more commonly known as the flu. Symptoms include fever, respiratory symptoms, nasal discharges, cough, headache and sore throat. *Incubation period* is the time period between the start of infection and the onset of symptoms. Influenza has an incubation period of 1-4 days [65]. The *latent period* is the time between being infected and becoming infectious, that is, capable of passing on the infection to others. Influenza's latent period is a day less than the incubation period. The latent period period is followed by 4-5 days of *infectious period* [65] . Once the infectious period ends, the *recovery period* starts, during which the infected individual is no more transmitting the infection to others. Figure 6.1 illustrates the infection time-line for influenza.

The 1918 pandemic, caused by the *H1N1 Spanish* strain, is a historical event, resulting in 20 million deaths in its first year alone [65]. World Health Organization (WHO) [75] and Centers for Disease Control and Prevention (CDC) [23] are involved in influenza surveillance around the world and in the design of effective vaccination programs.

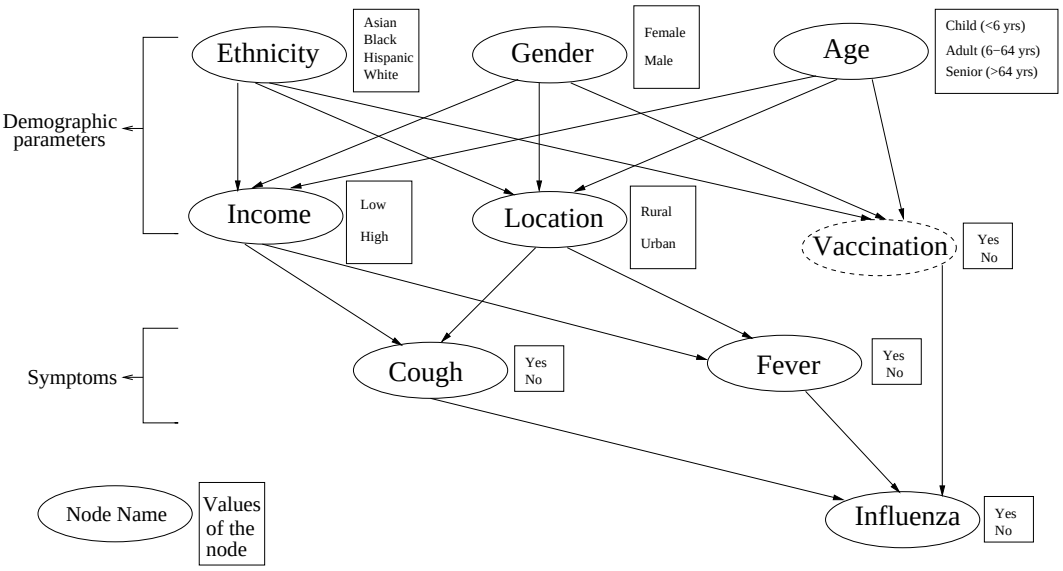


Figure 6.2. Bayesian Network

6.2. Epidemic curves

An epidemic of an infectious disease can be studied and analyzed by use of epidemic curves, which plot the incidence of the disease over time. In the ensuing analysis, demographic based epidemic curves for a geographic region are generated. The temporal analysis of these curves will aid in the identification of the critical groups of the population, that require primary attention in curtailing the spread of the disease.

6.2.0.1. *Bayesian Network.* We develop a Bayesian network illustrating the probabilistic dependencies of the demographics on the incidence of influenza in a geographic region. The Bayesian network is used to generate synthetic data sets, that reverse engineer an influenza outbreak embedded in a temporal domain.

For the study, the demographics and the symptoms considered are not comprehensive to keep the Bayesian model simple. The demographic parameters are ethnicity, gender, age, income, and location; and the symptoms are cough and fever. Figure 6.2 shows the Bayesian network illustrating the dependencies between demographics, symptoms, and influenza incidence. The analysis includes the absence and presence of vaccination.

Table 6.1. Symbols for Parameters and Parameter Values

Parameter	Ethnicity	Gender	Age	Income	Location	Vaccination	Cough	Fever	Influenza
Symbol	E	G	A	I	L	V	C	F	IN

Parameter Value	Asian	Black	Hispanic	White	Female	Male	Child	Adult	Senior
Symbol	As	Bl	Hi	Wh	Fe	Ma	Ch	Ad	Se

Parameter Value	Low	High	Rural	Urban	Yes	No
Symbol	Lo	Hi	Ru	Ur	Ye	No

Table 6.2. Probability Distributions: Ethnicity, Gender, Age, Cough, Fever, & Influenza

E	P(E)	G	P(G)	A	P(A)	I	L	P(C/I,L)	P(F/I,L)	C	F	P(I/C,F)
As	0.15	Fe	0.52	Ch	0.10	Lo	Ru	0.45	0.10	Ye	Ye	0.60
Bl	0.20	Ma	0.48	Ad	0.70	Lo	Ur	0.50	0.20	Ye	No	0.25
Hi	0.30			Se	0.20	Hi	Ru	0.20	0.05	No	Ye	0.40
Wh	0.35					Hi	Ur	0.25	0.15	No	Ye	0.05

6.3. Artificial Data Sets

The Bayesian network (Fig. 6.2), coupled with the influenza infection time-line (Fig. 6.1) are used to generate synthetic data sets for a population size of 100,000. The lack of available real data sets for epidemic outbreaks with finer demographic details necessitates the development of approaches to synthetically replicate epidemic data sets. The probability distributions for the the random variables in the Bayesian network are shown in Table 7.2 and Table 7.4 [3]. A contact rate of 10 is used, which implies that each infected individual contacts an average of 10 other people during the infectious period of 4-5 days. A contact is considered successful if the disease is transmitted. The primary case can be visualized as the root of a tree, with the infection spreading along the branches of the tree until the leaves being unable to make successful contacts with susceptibles.

Table 6.3. Probability Distributions: Income & Location

E	G	A	P(I=Lo/E,G,A)	P(L=Ru/E,G,A)	E	G	A	P(I=Lo/E,G,A)	P(L=Ru/E,G,A)
As	Fe	Ch	0.95	0.20	Hi	Fe	Ch	0.98	0.10
As	Fe	Ad	0.60	0.18	Hi	Fe	Ad	0.65	0.08
As	Fe	Se	0.40	0.25	Hi	Fe	Se	0.70	0.18
As	Ma	Ch	0.96	0.22	Hi	Ma	Ch	0.97	0.12
As	Ma	Ad	0.65	0.15	Hi	Ma	Ad	0.70	0.09
As	Ma	Se	0.50	0.27	Hi	Ma	Se	0.75	0.20
Bl	Fe	Ch	0.96	0.30	Wh	Fe	Ch	0.97	0.60
Bl	Fe	Ad	0.55	0.35	Wh	Fe	Ad	0.40	0.55
Bl	Fe	Se	0.50	0.40	Wh	Fe	Se	0.30	0.75
Bl	Ma	Ch	0.95	0.33	Wh	Ma	Ch	0.96	0.55
Bl	Ma	Ad	0.50	0.36	Wh	Ma	Ad	0.35	0.58
Bl	Ma	Se	0.40	0.42	Wh	Ma	Se	0.25	0.70

6.4. Demographic-based Epidemic Curves

An epidemic curve visualizes the incidence (rate) that traces the the number of newly infected individuals over time. Data mining is applied to synthetic data to extract pertinent information of the influenza epidemic, including demographic-based ethnic curves. The role of vaccination in curtailing the impact of the epidemic is investigated.

Epidemic curves (Fig. 6.3-6.6) are generated for the entire population, as well as for various demographic categories. These figures use cubic spline curves, to connect consecutive data points.

Figure 6.3 shows the dispersion of the epidemic among the different ethnic sub-groups. The level of impacts are ordered <hispanics, whites, blacks, asians>, thereby, placing larger number of hispanics at higher risk. Figure 6.4 shows the normalized version of the epidemic spread among the ethnic sub-groups, wherein the newly affected cases are proportioned over the total population of the specific sub-group. This helps in visualizing the outbreak among

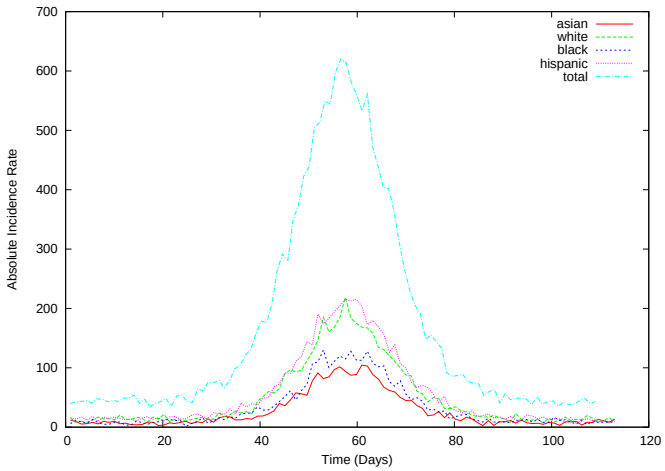


Figure 6.3. Ethnicity

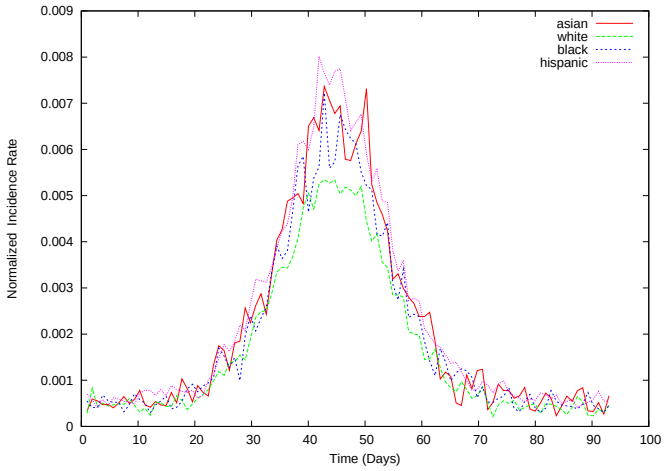


Figure 6.4. Ethnicity-Normalized

the sub-groups on a balanced platform. The ordered list of epidemic impact is <hispanics, asians, blacks, whites>. The asians are categorized correctly at the second highest risk level, while the earlier ordering placed asians at the lowest risk. Hence, normalized curves help in perceiving the true risk groups, based on proportions, rather than epidemic curves based on absolute number of disease incidence.

Figure 6.5 depicts the normalized epidemic spread among the genders, female and male. The females are observed at a higher risk level compared to males. Influenza spread does not

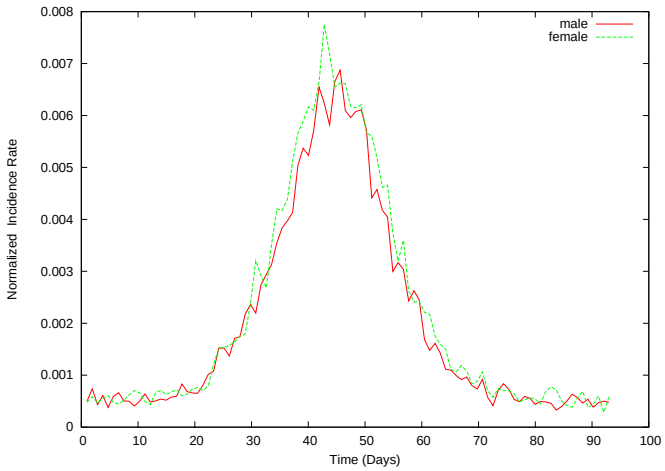


Figure 6.5. Gender-Normalized

have a gender bias. The observed result may be attributed to the complex social interaction patterns in the geographic region under study.

The age groups have been categorized as child ($\leq$5 years), adult (6-64 years), and senior ($\geq$65 years). Fig. 6.6 shows the normalized dispersion of the epidemic among the three age groups. Child subgroup are at the highest risk. Adults and seniors are observed at lower risk levels, closely interspersed with each other. CDC strongly recommends vaccination for seniors, since 90% of influenza related mortality occurs in the senior group. The observed lower level for the seniors in our experiments may be attributed to the specifics of the synthetic data.

6.4.0.2. *Vaccination.* Vaccination is the key preventive measure used by public health departments in curtailing the annual influenza epidemics. An uniform random distribution of the vaccines among the population is deployed by use of spread vaccination. Limitations on public health resources prohibit the goal of herd immunity. Consequently, resources must be applied optimally in order to curtail the epidemic.

Vaccines have associated effective rates of success. An ideal vaccine will have a 100% success in protecting a vaccinated individual from influenza. Nevertheless, influenza vaccines have an efficacy of 50%-80%. Figure 6.7 shows the epidemic curves for 10% of the population

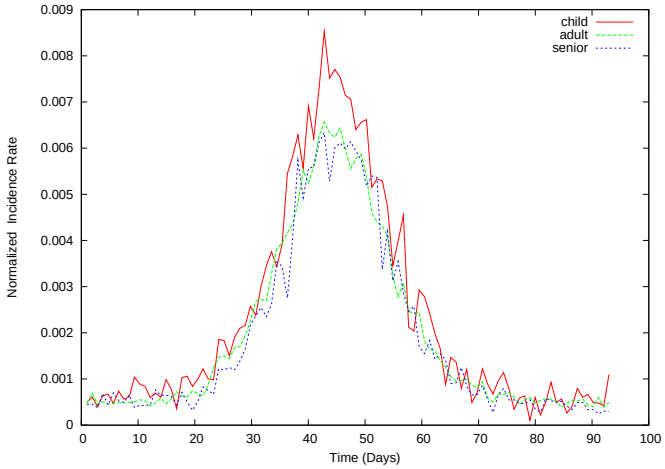


Figure 6.6. Age-Normalized

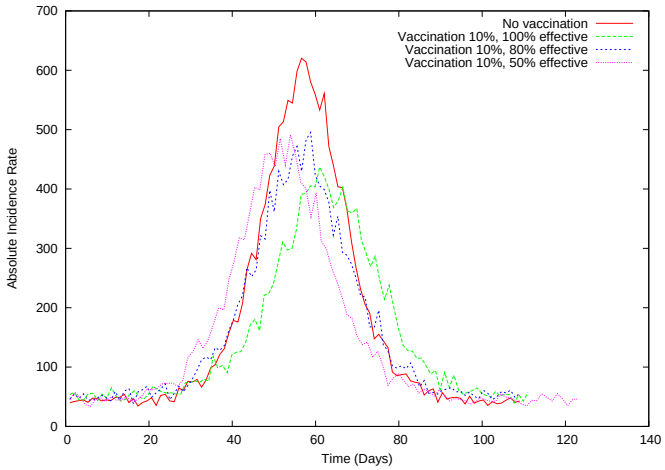


Figure 6.7. Varied Spread Vaccination Efficacy Rates

being spread vaccinated with 100%, 80%, and 50% success rates respectively, in comparison to a non-vaccinated population. The epidemic impacts for vaccine scenarios are all toned down compared to no vaccination scenario, reflecting on the success of the vaccination programs. As expected, 100% vaccine efficacy yields the best result.

Figure 6.8 illustrates the normalized bezier epidemic curves for ethnicity, gender and age groups. The analysis of the demographic graphs on a normalized scale aids in creation of a hierarchical list of demographic sub-groups ordered by their respective levels of risk. The

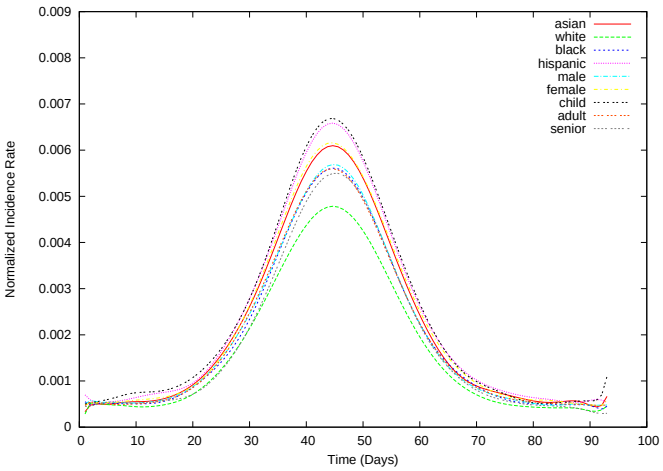


Figure 6.8. {Ethnicity, Gender, Age} Normalized

ordered list in the decreasing level of risk is <child, hispanic, asian, female, male, black, adult, senior, white>.

The experiments incorporate the assumption that vaccines are available for 10% of the population. In the first case, these vaccines are equally deployed among the three high-risk groups <child, hispanic, asian>. In the second case, these vaccines are applied equally among the three low-risk groups <adult, senior, white>. Figure 6.9 depicts the epidemic curves of both the cases, along with the curve for simple spread vaccination. The vaccine efficacy of 80% is tested in the experiments. Vaccination, prioritized on the high-risk groups, yield the best results in curtailing the epidemic, while vaccination of low-risk groups increases the epidemic impact, as compared to naive spread vaccination. Hence, demographic based high-risk group vaccinations should be accorded higher priority to curtail the influenza epidemic. This necessitates a better understanding of disease progression in a given demographic domain.

6.5. Inferences

Bayesian probabilistic reasoning is used in developing synthetic data sets, portraying the outbreak of an influenza epidemic in a geographic region. Influenza epidemic due to a single strain prevails for around eight weeks, and all the epidemic curves generated from the synthetic data also depict similar epidemic behavior. The epidemic curve for the total population

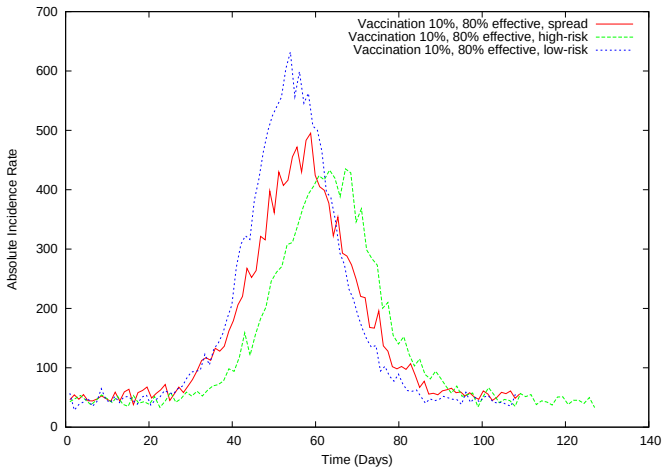


Figure 6.9. Vaccination based on Risk-groups

is dispersed into demographic-based epidemic curves. The temporal flow of the epidemic among the different sections of the population is analyzed. The different population sections are ordered in a hierarchical list, based on risk levels. The higher end risk spectrum of the ordered list need prioritized health care, in comparison to the lower end risk spectrum, to curtail the epidemic.

Spread vaccination is applied uniformly across a proportion of the population, and is witnessed to scale down the epidemic. The higher the efficacy of the vaccine in thwarting the disease, the lower is the impact of the epidemic. Vaccination resources are alternatively applied in two different ways. In the first scenario, the high risk groups are targeted, while in the second case, the lower risk groups are targeted. As per expectations, the epidemic is reduced extensively when high-risk groups have been vaccinated. On the other hand, when applied to low-risk groups, the desired effects of our vaccination effort are less, as compared to naive spread vaccination. This validates the reasoning in identification of high risk demographic sections of the population, and prioritizing them in the allocation of the limited public health resources.

CHAPTER 7

SPATIAL CORRELATION OF DISEASE PREVALENCE FOR INFLUENZA AND PNEUMONIA

[1] Disease monitoring plays a crucial role in the implementation of public health measures. The demographic profiles of the people and the disease prevalence in a geographic region are analyzed for inter-causal relationships. Bayesian analysis of the data identifies the pertinent characteristics of the disease under study. The vital components of control and prevention of the disease spread are identified by Bayesian learning for the efficient utilization of the limited public health resources. Bayesian computing, layered with epidemiological expertise, provides the public health personnel to utilize their available resources optimally to minimize the prevalence of the disease. Bayesian analysis is implemented using synthetic data for two different demographic and geographic scenarios for pneumonia and influenza, that exhibit similar symptoms. The analysis infers results on the effects of the demographic parameters, namely ethnicity, gender, age, and income levels, on the evidence of the prevalence of the diseases. Bayesian learning brings in the probabilistic reasoning capabilities to port the inferences derived from one region to another.

7.1. Bayesian Analysis

The Bayesian network (Fig. 7.1) analyzes the effects of the demographic parameters on the incidence of symptoms and the related diseases in a geographic area. The demographic

[1]This chapter is reprinted from: Advances in Bioinformatics and its Applications, Proceedings of the International Conference, Nova Southeastern University, Fort Lauderdale, Florida, USA 16 - 19 December 2004. K. Abbas, A. Mikler, A. Ramezani and S. Menezes, Computational Epidemiology: Bayesian Disease Surveillance, pp. 95-106, 2004, with permission from World Scientific Publishing Co. Pte. Ltd, Singapore.

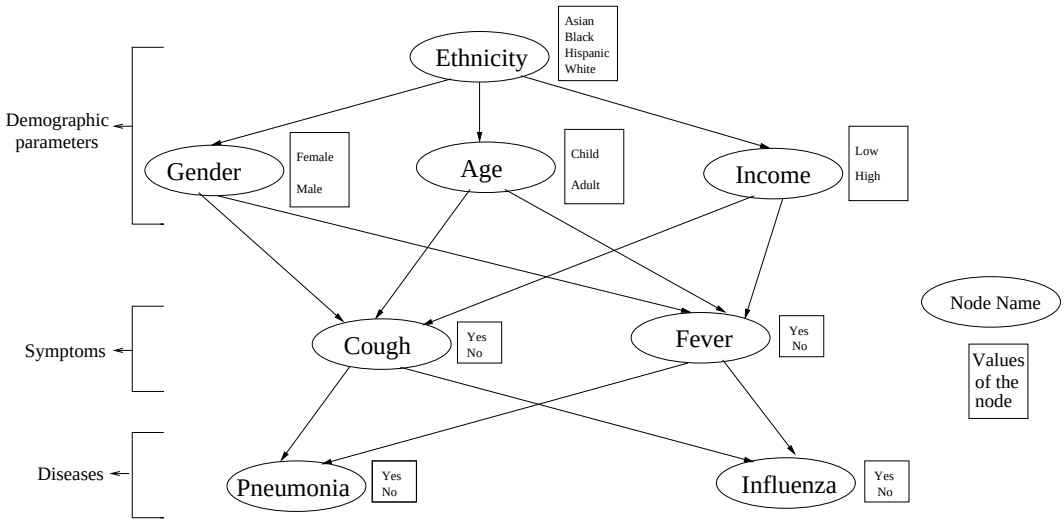


Figure 7.1. Bayesian Network for Demographic Analysis of Diseases

parameters are ethnicity, gender, age, and income; symptoms are cough and fever; and diseases are pneumonia and influenza. The Bayesian network illustrates the predictive reasoning of the demographics on the prevalence of diseases. Table 7.1 defines the list of symbols used for the parameters and their corresponding values. Severe disease outbreaks in two smaller geographic areas are considered[2].

7.2. Scenario I

Table 7.2 includes the beliefs for the demographic parameters in geographic area I. The probability distributions for the symptoms and diseases are given in Table 7.3. For example, P(F/G,A,I) refers to the conditional probability of fever, given the evidence of gender, age, and income. The Bayesian network is analyzed to derive useful inferences on the prevalence of pneumonia and influenza. The population affected by pneumonia and influenza are 11.21% and 8.84% respectively.

The values of each of the demographic parameters are ordered in their levels of significance on the outcome of the two diseases (Fig. 7.2). The relative levels of significance within

[2]The artificial data has been synthetically generated and is not reflective of any real demographic and geographic settings.

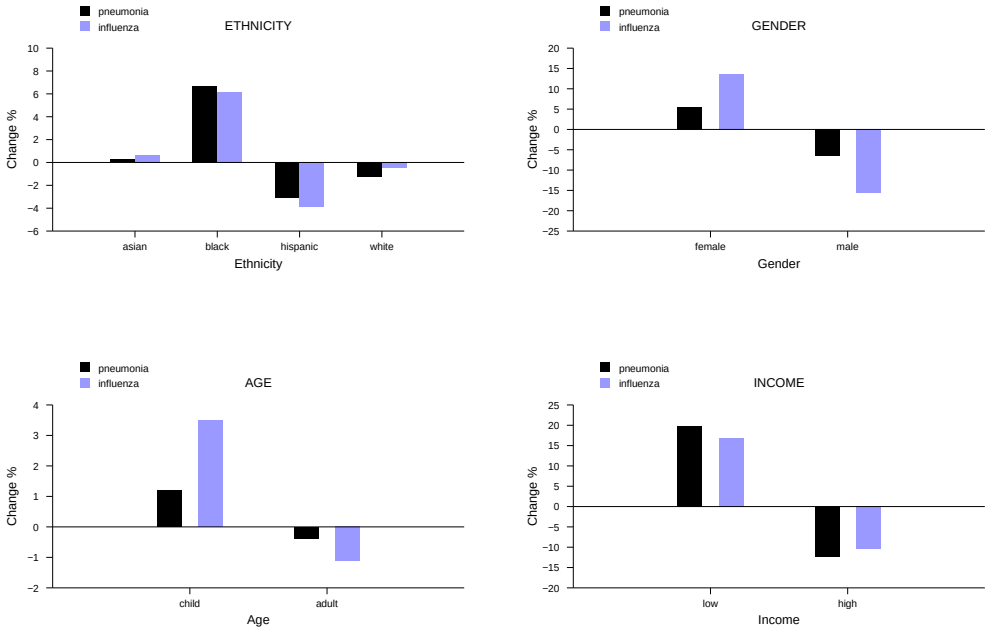


Figure 7.2. Results of Bayesian Analysis for Geographic Area I

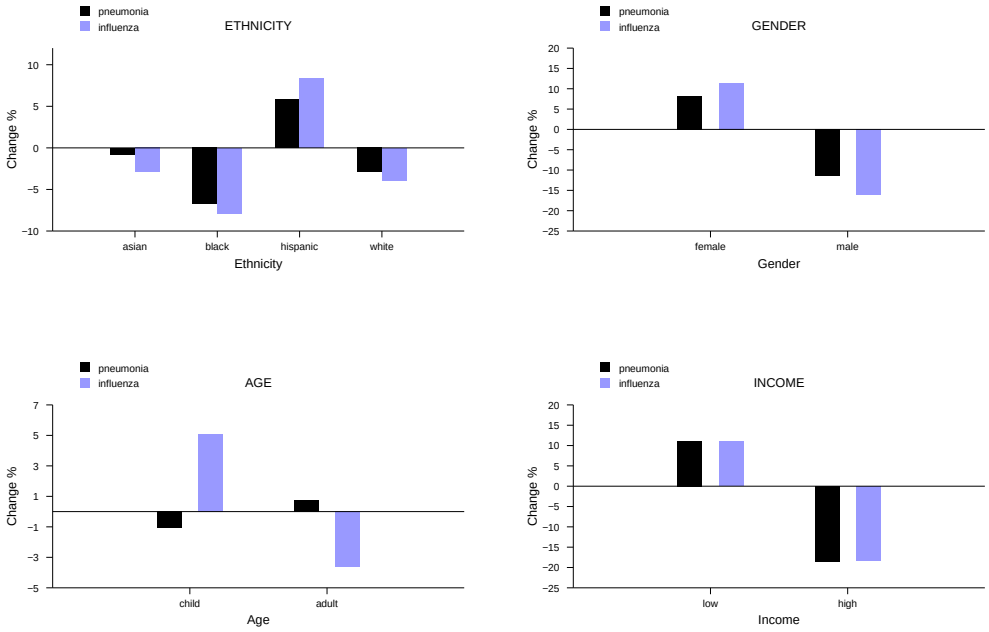


Figure 7.3. Results of Bayesian Analysis for Geographic Area II

Table 7.1. Symbols for Parameters and Parameter Values

Parameter	Symbol	Parameter Value	Symbol
Ethnicity	E	Asian	As
Gender	G	Black	Bl
Age	A	Hispanic	Hi
Income	I	White	Wh
Cough	C	Female	Fe
Fever	F	Male	Ma
Pneumonia	P	Child	Ch
Influenza	IN	Adult	Ad
		Low	Lo
		High	Hi
		Yes	Ye
		No	No

each demographic parameter are derived by setting each individual value to absolute unity, that is (P(value) = 1). All other parameters' beliefs are kept the same as before, and the proportional changes in the people affected by the diseases are inferred. Hence, by raising the whole community to be hispanics, a decrease in pneumonia (-3.08%) and influenza (-3.89%) prevalence is observed.

Among the different ethnicities, the black ethnic subgroup is observed to be at most risk for both pneumonia and influenza, followed by asians, whites and hispanics. A similar analysis for the other demographic parameters gives a suit of significant results. In case of gender, females are at a relatively higher risk in comparison to males for both pneumonia and influenza. For age, children exhibited a higher risk in comparison to adults for both the

Table 7.2. Probability Distributions of Demographics for Scenario I

E	P(E)	G	E	P(G/E)	A	E	P(A/E)	I	E	P(I/E)
As	0.15	Fe	As	0.55	Ch	As	0.25	Lo	As	0.40
Bl	0.20	Fe	Bl	0.60	Ch	Bl	0.15	Lo	Bl	0.55
Hi	0.30	Fe	Hi	0.46	Ch	Hi	0.30	Lo	Hi	0.30
Wh	0.35	Fe	Wh	0.56	Ch	Wh	0.23	Lo	Wh	0.35
		Ma	As	0.45	Ad	As	0.75	Hi	As	0.60
		Ma	Bl	0.40	Ad	Bl	0.85	Hi	Bl	0.45
		Ma	Hi	0.54	Ad	Hi	0.70	Hi	Hi	0.70
		Ma	Wh	0.44	Ad	Wh	0.77	Hi	Wh	0.65

diseases. The lower income people are inferred to be more likely infected by both diseases, compared to the higher income people.

The demographic parameters can be further combined and ordered in their levels of importance in the spread of diseases. Based on the artificial data for area I, lower income black female children are at the higher end of the risk spectrum, while higher income hispanic male adults are at the lower end of the spectrum for pneumonia. The spread of influenza also exhibited similar results in geographic area I.

7.3. Scenario II

Table 7.4 defines the demographic probability distributions of geographic area II. Similar to area I, area II is experiencing an epidemic of pneumonia and influenza. The prevalence of the two diseases, pneumonia and influenza, are currently unknown. The posterior probability distributions for the symptoms and the diseases of area I are ported into the Bayesian learning process for area II. The developed Bayesian network can then be analyzed for the role of demographics on the two diseases.

Table 7.3. Probability Distributions of Symptoms and Diseases

G	A	I	P(C/G,A,I)	G	A	I	P(F/G,A,I)
Fe	Ch	Lo	0.35	Fe	Ch	Lo	0.75
Fe	Ch	Hi	0.25	Fe	Ch	Hi	0.40
Fe	Ad	Lo	0.88	Fe	Ad	Lo	0.44
Fe	Ad	Hi	0.05	Fe	Ad	Hi	0.85
Ma	Ch	Lo	0.15	Ma	Ch	Lo	0.54
Ma	Ch	Hi	0.85	Ma	Ch	Hi	0.64
Ma	Ad	Lo	0.54	Ma	Ad	Lo	0.27
Ma	Ad	Hi	0.64	Ma	Ad	Hi	0.20
	C	F	P(P/C,F)		C	F	P(IN/C,F)
	Ye	Ye	0.30		Ye	Ye	0.25
	Ye	No	0.10		Ye	No	0.05
	No	Ye	0.10		No	Ye	0.10
	No	No	0.001		No	No	0.00

The infected population of pneumonia and influenza are 12.06% and 9.67% respectively. Figure 9.3 shows the relative levels of significance of the values of each demographic parameter. Considering ethnicity for both the diseases, hispanics exhibit the highest risk, followed by asians and whites, while blacks have the least risk. In case of gender, females show higher risk to both diseases in comparison to males. Children exhibited lower risk to pneumonia compared to adults, while adults have a lower risk to influenza in comparison to children. The lower income groups are observed to be more prone to both diseases in relation to the higher income groups.

Table 7.4. Probability Distributions of Demographics for Scenario II

E	P(E)	G	E	P(G/E)	A	E	P(A/E)	I	E	P(I/E)
As	0.20	Fe	As	0.60	Ch	As	0.15	Lo	As	0.55
Bl	0.15	Fe	Bl	0.55	Ch	Bl	0.25	Lo	Bl	0.40
Hi	0.35	Fe	Hi	0.75	Ch	Hi	0.60	Lo	Hi	0.75
Wh	0.30	Fe	Wh	0.40	Ch	Wh	0.46	Lo	Wh	0.64
		Ma	As	0.40	Ad	As	0.85	Hi	As	0.45
		Ma	Bl	0.45	Ad	Bl	0.75	Hi	Bl	0.60
		Ma	Hi	0.25	Ad	Hi	0.40	Hi	Hi	0.25
		Ma	Wh	0.60	Ad	Wh	0.54	Hi	Wh	0.36

On analysis of the synthetic data for area II, lower income hispanic female adults have a higher risk of pneumonia infections, while higher income black male children have a lower risk. For influenza, lower income hispanic female children have a higher risk, while higher income black male adults are at a lower risk. Although the data is hypothetical, the differences in the critical risk groups of the two areas indicate the significance of analyzing the demographic parameters. For instance, if the surveillance and preventive measures developed for area I had been applied to area II, the more critical groups of area II would have been less addressed.

7.4. Inferences

The probabilistic reasoning Bayesian methodology aids in the identification of the critical points for control, prevention and surveillance of diseases. The limited resources of the public health department can be aptly used in the order of the identified high risk groups to derive the best gains. The Bayesian network learned from a specific demographic and geographic settings for a disease outbreak can be transferred to different demographic and geographic

settings. The analysis of the adapted network helps in the identification of the control points for different demographic and geographic settings.

The critical groups identified at higher levels of risk are different for the two geographic regions. This underlines the significance of analyzing the demographics to discover the higher risk spectrum of the population. The high risk groups are to be accorded prime attention to curtail the epidemic. Consequently, it is imperative to develop more tools that allow epidemiologists to extrapolate the findings across multiple geographic regions, thereby allocating the public resources efficiently.

CHAPTER 8

TEMPORAL ANALYSIS OF HIV IN TEXAS

The inherent complexity of global, national and regional influences in Human Immunodeficiency Virus (HIV) transmission dynamics over the spatio-temporal domain warrants multiple scales of analysis. An analysis of HIV incidence in Texas from 1989-2002 using probabilistic reasoning is implemented. Bayesian networks are developed representing the HIV surveillance in time intervals of two years from 1989 to 2002. The Bayesian networks are correlated temporally to synthesize the dynamic Bayesian network. Demographic-based epidemic curves for HIV incidence are inferred by the model and aids in identification of risk levels among the different demographic subgroups. The analysis enhances the understanding of the temporal dynamics of HIV transmission for different demographic subgroups of age, gender and ethnicity. Whilst our results reveal the complexity in quantifying the HIV spread in a heterogeneous population, identification of high-risk demographic subgroups enables efficient utilization of constrained public health resources.

8.1. HIV Surveillance

Infectious diseases of the past century and emerging infectious diseases are a growing concern for the public health professionals and the whole community in general. Mathematical and computational models illustrating the disease dynamics are important in understanding the transmission of pathogens and disease control measures in different demographic and geographic settings [8, 12]. Human Immunodeficieny Virus (HIV) is the disease causing virus/pathogen of AIDS in individuals. The global effects of HIV for the past quarter of a century signifies the complexities in the transmission dynamics. History, economics, politics, socio-cultural issues and medical infrastructure play an integral role in the spread of HIV [43].

Health surveillance is a valuable asset to analyze and identify high risk demographic subgroups of a population. The Texas Department of State Health Services (DSHS) [69] maintains a surveillance database of the individuals infected by the HIV virus in the state of Texas. We analyze this dataset using probabilistic reasoning and Bayesian analysis. Dynamic Bayesian networks are constructed to quantify the HIV incidence, that is the rate of newly infected individuals, from 1989 until 2002. The model identifies the high-risk demographic subgroups for HIV and will supplement the health surveillance process.

8.2. Human Immuno-deficiency Virus

The socio-economic and political perspectives of the modern world are threatened by the HIV spread at both epidemic and pandemic levels. HIV infections are on a rise in North America, with 45,000 new infections every year [44]. HIV in Texas reflects the contrasting nature of HIV incidence and progression in the western world, upon comparison to HIV spread in developing countries, including sub-Saharan Africa [53]. The characteristics of HIV cases in Texas highlights the differences in vulnerability and survival with respect to different demographics [54].

The primary modes of transmission for HIV are via sexual contact, intravenous drug use, blood transfusion and vertical pediatric transmission from mother to the unborn child. The behavioral change of HIV infected individuals upon knowledge of being HIV positive is a complex issue [43]. The access to modern medical infrastructure and the response of the infected individual to the retro-viral drugs are key factors in diminishing the HIV infection to AIDS. In the Bayesian analysis, the focus is on the incidence of newly infected individuals with HIV over the demographic space.

HIV and AIDS in Africa has been extensively investigated from both the biomedical and epidemiological perspectives [43]. The transmission dynamics in sub-Saharan Africa is characterized as a complex and regionally specific phenomenon that is rooted in local economics, deepening poverty, migration, gender war. global economies and cultural politics. HIV virus

was identified in 1984, and since then over 60 clinical trials for a HIV vaccine have been conducted [48]. Ethical concerns over phase III testing on healthy individuals impedes the discovery of an effective vaccine. Human sexual contacts exhibit a small world phenomena and scale free network due to large degree of clustering [47]. The small average path length between two individuals necessitates strategic campaigns that efficiently curtail the progression of sexually transmitted diseases, including HIV. Human Papilloma Virus (HPV) is a sexually transmitted disease, similar to HIV, that leads to cervical cancer. An online accessible HPV model has been developed to predict HPV prevalence in varied demographic settings and quantify different vaccination policies [24, 27].

8.3. Bayesian Analysis of HIV Incidence in Texas

8.3.1. *Dynamic Bayesian Network Model*

The probabilistic reasoning capabilities under uncertainty are integral to Bayesian learning and analysis [55]. The principle of Bayes' theorem is to update the beliefs of a hypothesis, given evidence. A Bayesian network (see Fig. 8.1) is developed to illustrate the probabilistic dependencies of the demographics on the incidence of HIV in Texas. The demographic parameters in the study are age, gender and ethnicity.

Dynamic Bayesian networks enhance the capability of Bayesian networks by embedding the temporal relationships among the studied variables [45]. In the analysis, the dynamic Bayesian network will correlate the temporal flow of HIV incidence for every two years (see Fig. 8.2). The analyzed HIV dataset is from 1989 to 2002, and time intervals of two years are chosen. For every Bayesian network corresponding to the two year intervals between 1989 to 2002, a Bayesian probabilistic prediction is implemented for years 2000 and 2002 (see Fig. 8.3).

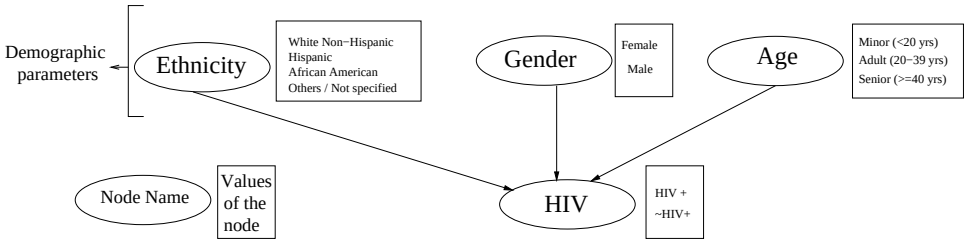


Figure 8.1. Bayesian Network to Analyze HIV Incidence

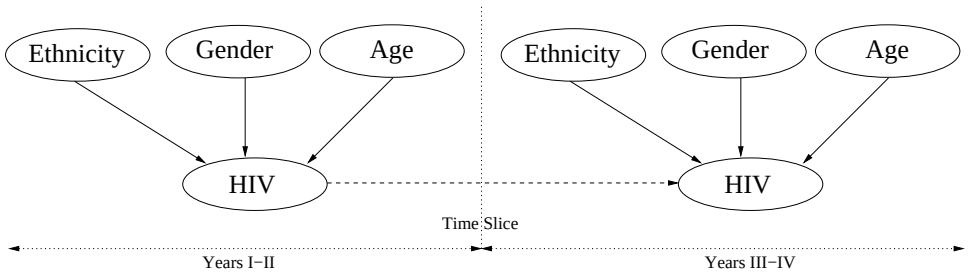


Figure 8.2. Dynamic Bayesian Network Analysis for HIV Incidence

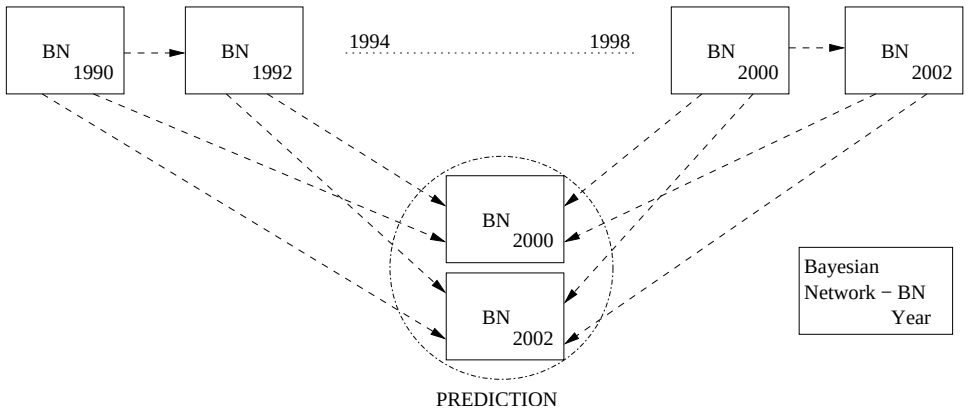


Figure 8.3. Probabilistic Analysis of HIV Incidence

8.3.2. *HIV Surveillance Dataset*

Texas Department of State Health Services [69] maintains a HIV surveillance database of the HIV infected individuals in the state of Texas. 82,842 cases recorded from 1989 to 2002 are analyzed. The analysis is focused to the demographic sub-groups based on age, gender and ethnicity. The age subgroups are {Minor - 0 to 19 years}, {Adult - 20 to 39 years}, {Senior - 40 years and above} while male and female are the two gender subgroups. Ethnicity subgroups

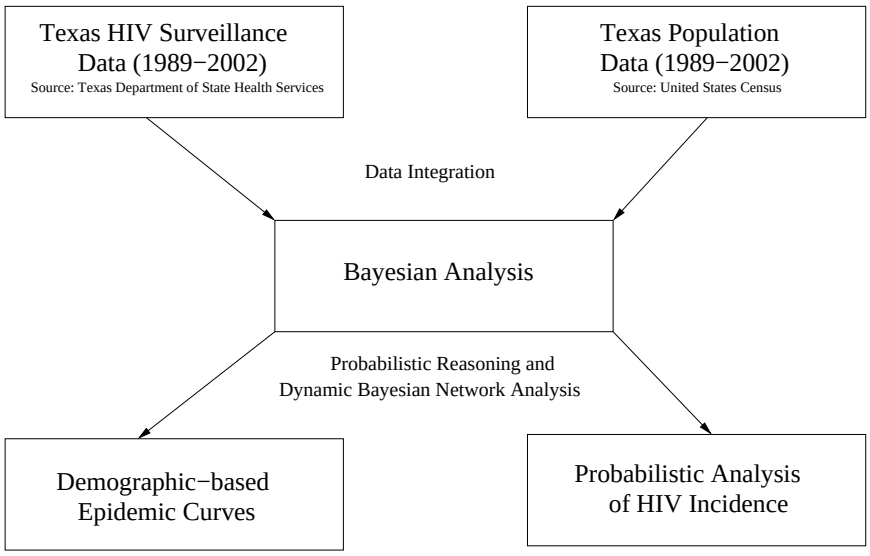


Figure 8.4. Framework for HIV Incidence Data Analysis

are White non-Hispanic, African American, Hispanic and other/non-specified. These classes of subgroups enable to analyze the HIV incidence based on 24 (3*2*4) different demographic subgroups (refer Fig. 8.1).

8.3.3. *Bayesian Analysis*

The HIV surveillance dataset is integrated with the population estimates of the different subgroups, obtained from the United States Census Bureau [70]. This enables to quantify the proportions of HIV incidence in each demographic subgroup. This resultant proportions and census population estimates of the subgroups are integrated into the Bayesian networks. The temporal flow of HIV incidence is learnt upon analysis of the dynamic Bayesian networks and thereby deriving demographic-based epidemic curves. Figure 8.4 depicts the schematic flow of our probabilistic analysis methodology.

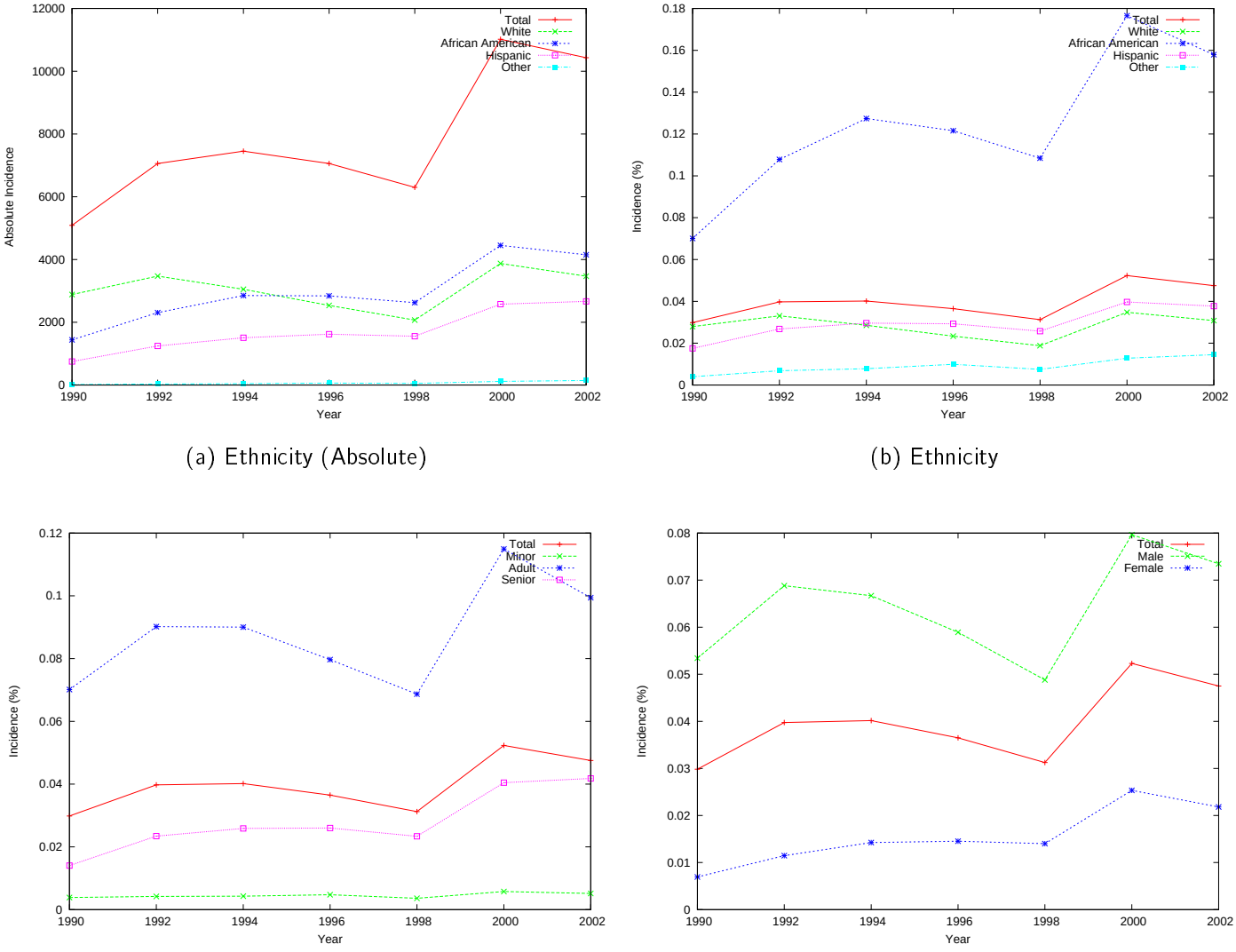
8.3.4. *Demographic-based Epidemic Curves*

An epidemic curve visualizes the incidence (rate) that traces the number of newly infected individuals over time [3]. Time intervals of two years are used; for example, 1990 denotes the time interval 1989-1990. Figure 5(a) illustrates the absolute incidence of HIV for Texas from

1989 to 2002 as well as among the different ethnic subgroups. The ordered list of incidence impact is <white, african-american, hispanic, other> from 1989-1994 while it is <african-american, white, hispanic, other> from 1994-2002. Figure 5(b) enhances the information value by depicting the proportions of each subgroup that are affected. The proportion of african american population are the most affected, well above the proportion of HIV incidence among the total population. Also, the proportion of hispanic subgroup that are affected is more than the proportion of white (non-hispanic) subgroup from 1994-2002. This reaffirms the value of integrating the HIV surveillance data with the population census data, which enables the computation of the proportional values for HIV incidence.

Figure 5(c) depicts the proportion of HIV incidence among the three age groups. The proportion of newly infected adults comprise the highly infected age demographic subgroup, while incidence is lowest among the minors. Hence, the ordered list of HIV incidence impact for the age demographic subgroups is <adult, senior, minor> for all years from 1989 to 2002. Upon analysis of the gender demographic epidemic curves (see Fig 5(d)), the proportion of HIV incidence among male subgroup is witnessed to be 4 to 5 times higher than that of the female subgroup for all years.

The population estimates of the different demographic subgroups are available from the census for all years from 1989 to 2002. After analyzing 1989-1990 interval, the conditional probability distributions of HIV incidence is maintained the same while the probability distributions of age, gender and ethnicity are swapped by the values of 1999-2000 group. This enables us to predict a measure of HIV incidence for 1999-2000 from the witnessed HIV incidence relationships of 1989-1990. This analysis is repeated for every two year interval from 1989-2002 against 1999-2000. The whole process is further executed for expected values for 2001-2002 against all intervals between 1989-2002. The results of this analysis is illustrated in Fig. 8.6. The predicted probabilistic measures of HIV incidence do not correlate well with the observed HIV incidence in the time intervals 1999-2000 and 2001-2002. This inability to



(a) Ethnicity (Absolute)

(b) Ethnicity

(c) Age

(d) Gender

Figure 8.5. Demographic-based Epidemic Curves for HIV Incidence

measure a predictable course of HIV incidence is a consequence of the complexity involved in quantifying socio-behavioral interaction patterns and population dynamics.

8.4. Inferences

Probabilistic analysis of HIV incidence in Texas from 1989-2002 using demographic-based epidemic curves enables to quantify the HIV incidence risk levels among the different demographic subgroups. An ordered hierarchy of risk levels among the demographic subgroups is an asset in prioritizing the limited public health resources. The integration of HIV surveillance

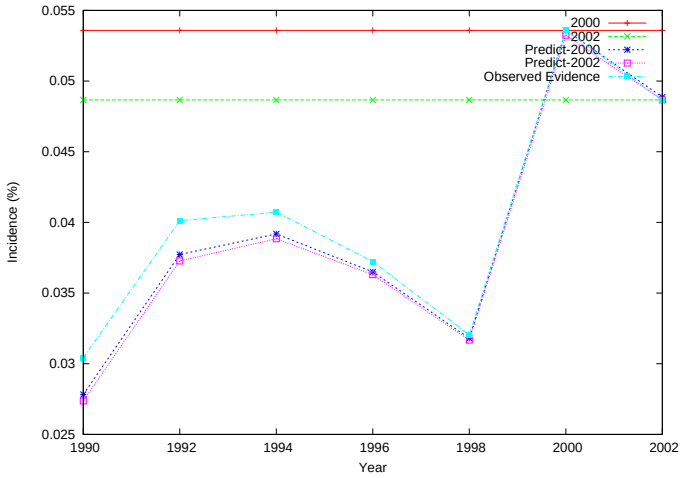


Figure 8.6. Probabilistic Analysis of HIV Incidence

data and the census population estimates lets us to analyze the proportion of HIV incidence among the different demographic subgroups. The risk of HIV incidence among the different subgroups based on age, gender and ethnicity have a varied spectrum, thereby highlighting the utility of demographic based risk analysis. Probabilistic prediction of the course of HIV spread in a heterogeneous population has complexities involving behavioral patterns and population dynamics and is beyond the scope of available analytical models.

CHAPTER 9

SOCIO-BEHAVIORAL ANALYSIS OF INFLUENZA OUTBREAKS

Prevalence, incidence, morbidity and mortality rates are used to evaluate the impact of an infectious disease outbreak in a population. Knowledge of the social behavioral interactions among the people will complement the understanding of the progression of air-borne diseases such as influenza. A methodology using Hidden Markov models (HMMs) is presented to gain a better understanding of the social behavioral interactions that directly relate to the observed prevalence and incidence of an infectious disease.

9.1. Influenza Outbreak Data Simulator

Influenza outbreak data is generated using the simulator developed by S. Venkatachalam and A. Mikler [49, 71, 72]. The simulator is based on the principle of global stochastic field simulation (GSFS) and incorporates geographic and demographic interactions. The interactions are based on the geographic information systems (GIS) gravity model. The GSFS model is oriented for heterogeneous population, and can incorporate interactions based on geography, demography, environment and migration patterns.

Spatial distribution of the population is represented as cells, similar to the traditional cellular automata paradigm. Each cell represents an individual or a sub-population. Each cell is characterized with state and likelihood risks for exposure and disease contraction. To simulate the disease spread in such an environment, contacts need to be established between cells. A cell is capable of interaction with any other cell in the environment. The probability of contact is based on an interaction coefficient that takes into account the distance, population, demographics and socio-economic factors.

In the presented analysis, the outbreak data includes the set of all contacts during an influenza outbreak in a population. Disease progression for the primary infected case and the spread through the population is contained in the data set. For each contact between two cells, the proportion of infectious people in each cell is given; also, whether the contact resulted in an infection is included.

9.2. Hidden Markov Models

Hidden Markov models are represented by a tuple $< \pi, S, C >$. π represents the vector of the initial hidden state probabilities. S is the state transition matrix and C is the confusion matrix. In this analysis, the random variables investigated are the proportions of infectious people of each cell for a contact, and the result of disease transmission. The two hidden states are defined as infected (I^+) and not infected (I^-). At the start of simulation, there are no infected people, the initial vector will have null probability for I^+, as shown in Eq. 6.

$$\pi = \begin{array}{c} \begin{array}{cc} I^+ & I^- \end{array} \\ \begin{pmatrix} 0 & 1 \end{pmatrix} \end{array} \tag{6}$$

The state transition matrix S defines the probability distribution for temporal transition between these two states for the Markov process, as shown in Eq. 7. S_{ij} defines the probability of transition from the hidden state corresponding to row i to the hidden state corresponding to column j. The probabilities for each row sum up to 1.

$$S = \begin{array}{c} \begin{array}{cc} I^+ & I^- \end{array} \\ \begin{array}{c} I^+ \\ I^- \end{array} \begin{pmatrix} & \\ & \end{pmatrix} \end{array} \tag{7}$$

The observed states relate to the proportion of infectious people in the two cells during a contact. For a given proportion of infectious people in a cell, threshold value μ is set. If both the cells have infectious people proportion less than μ for a contact, the observed state is defined by U_1. In a contact, if only one of the cells have infectious people proportion less

than μ, the observed state is defined by U_2. If both the cells have infectious people greater than μ, the state is defined by U_3. The confusion matrix C illustrates the probabilities of the observed states given the hidden state, as shown in Eq. 8. The sum of probabilities for each row will add to 1.

(8)
$$C = \begin{array}{c} \\ I^+ \\ I^- \end{array} \begin{array}{c} \begin{array}{ccc} U_1 & U_2 & U_3 \end{array} \\ \begin{pmatrix} & & \\ & & \end{pmatrix} \end{array}$$

9.2.1. *Purpose*

Incidence, prevalence, and rates of morbidity and mortality are measures of the impact of a disease outbreak in a population. They do not quantify the risk associated with the social behavioral interactions that lead to disease transmission and progression. In order to gain insight on the different levels of contact associated with disease spread as well as those contacts that do not result in disease transmission, hidden Markov models is used.

The state transition matrix S is an indicator of the social behavioral interactions in the population over the temporal domain. The matrix S visualizes the outbreak data as an event flow model. It quantifies the chance and probability of encountering successive contacts of all four possible events. The four possible event pairs are {infection $\rightarrow$ infection, infection $\rightarrow$ no infection, no infection $\rightarrow$ infection, no infection $\rightarrow$ no infection}. The state transition matrix shown in Eq. 7, illustrates the four state transitions, { $I^+ \rightarrow I^+$, $I^+ \rightarrow I^-$, $I^- \rightarrow I^+$, $I^- \rightarrow I^-$ }.

9.3. Analysis

Six different simulated influenza outbreak data sets are analyzed. Hidden Markov models (HMMs) are developed for each of the data sets. The HMM for a data set is evaluated against the same data set as well as against the other data sets.

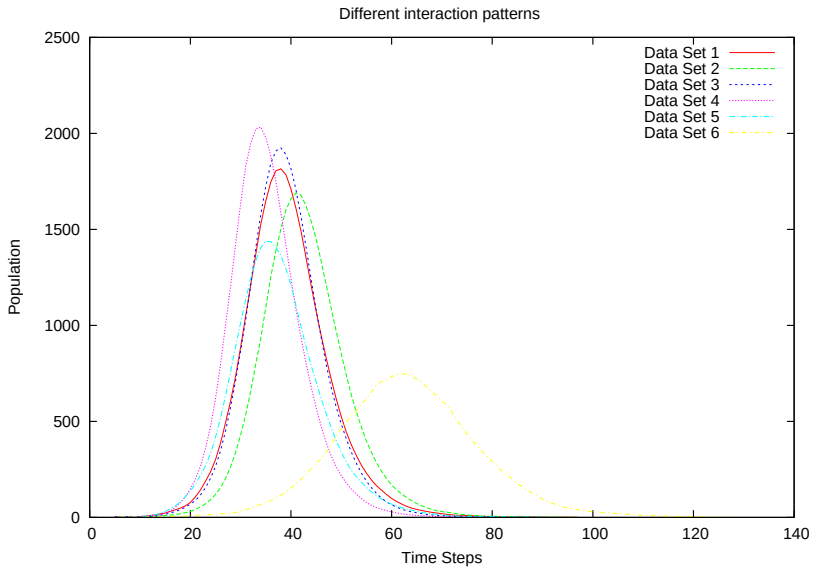


Figure 9.1. Epidemic Curves for Six Different Influenza Outbreaks

9.3.1. *Simulated Data for Influenza Outbreak*

The influenza outbreak data sets contain the complete list of contacts for a demographic population in a geographic region. The region is spatially divided into 2000 cells over a grid of size $20 * 100$. Global stochastic field simulation is used to generate the contacts [49, 71, 72]. Each contact may or may not lead to transmission of infection. The result of the transmission is also included in the data set. For each contact between two cells, the proportion of infectious people in the two cells are also contained in the data set. Hence, for every contact, the data set contains the tuple $<$ proportion of infectious people in first cell, proportion of infectious people in second cell, transmission result $>$.

Epidemic curves illustrate the influenza incidence rate, or the rate of newly infected individuals in a population. The epidemic curves for the six data sets are generated, as shown in Fig. 9.1. Contacts are initiated between individuals of any two cells for each day during the course of the disease outbreak. The infectivity levels are different for the data sets and are in the range [0.25, 0.4]. Infectivity determines the probability of disease transmission for a contact involving an infectious individual and a susceptible individual. Also, the contact

Table 9.1. Simulated Data Set Parameters

Data Set	Contact Rate	Infectivity	Contacts per Day
1	8	0.4	92506
2	8	0.35	124298
3	8.25	0.4	109767
4	8.25	0.4	89942
5	8.25	0.25	104956
6	8.25	0.03	70040

rate is altered between the different simulated data sets. Table 9.1 illustrates the rates of contact, infectivity and the number of contacts per day for the different data sets.

9.3.2. *Learning the Hidden Markov Model*

Hidden Markov models are developed for each of the data sets. The primary case is introduced into the population at $time = 1$. Hence, the initialization vector at $time = 0$ will be $< I^+ = 0, I^- = 1 >$. Each contact contained in the data set is processed in temporal order to compute the state transition matrix S and the confusion matrix C.

For any two consecutive contacts in the temporal domain, the results of infection will be either of the four possibilities, $\{ I^+ \rightarrow I^+, I^+ \rightarrow I^-, I^- \rightarrow I^+, I^- \rightarrow I^- \}$. Depending upon the change in hidden state for a contact, that may or may not result in an infection, the state transition matrix is correspondingly updated. The pair of cells during a contact have a proportion of infectious individuals, that is dynamically changing during the disease outbreak. The threshold value μ classifies the contacts into three categories, thereby each contact is associated with an observed state from $\{U_1, U_2, U_3\}$. The result of disease transmission and the observed state is known, thereby the confusion matrix can be updated. Here, learning

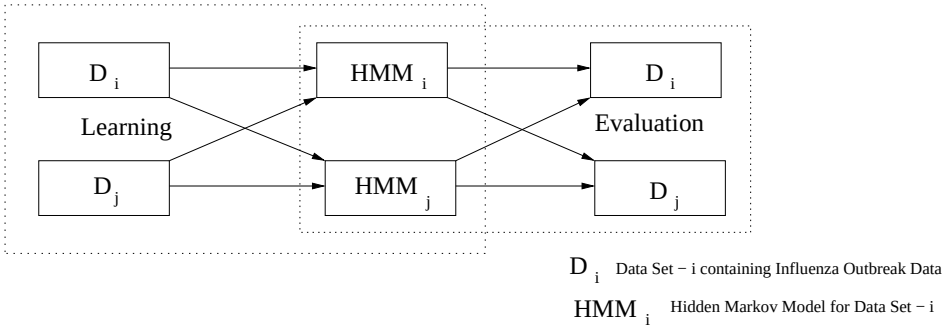


Figure 9.2. Hidden Markov Model Evaluation

is synonymous with developing the HMM from the complete data set of hidden and observed values. This is conceptually different from learning the HMM, given only the temporal observed states.

9.3.3. *Evaluation*

The developed hidden Markov models of the different data sets are evaluated against the corresponding influenza outbreak data set as well as against the other data sets, as shown in Fig. 9.2. The hypothesis is that the HMM for a given data set will correlate higher in comparison to its correlation to other data sets.

Each data set is processed to compute the probability of the observed sequence for a HMM. The tuple $< \pi, S, C >$ is known for the HMM. Since the hidden state for an observation is unknown, the probability of an observed state is dependent on the probability of both the hidden states. This correlation is consistent through out the disease outbreak and is given by the confusion matrix C. In addition, each hidden state for the current time step correlates to the hidden state in the previous time step. This temporal correlation is given by the state transition matrix S.

Equation 9 illustrates the computation of partial probability of an observed state at the current time step t. The partial probabilities $p_t(I^+)$ and $p_t(I^-)$ are summed up together over the temporal domain of the complete data set. $p_t(I^+)$ is the partial probability of hidden state I^+ at $time = t$, while $p_t(I^-)$ is the partial probability of hidden state I^- at $time = t$.

A higher measure indicates a higher order of correlation. U_t refers to the observed state at $time = t$.

$$p_t(I^+) = P(I^+ \rightarrow U_t) * (p_{t-1}(I^+) * P(I^+ \rightarrow I^+) + p_{t-1}(I^-) * P(I^- \rightarrow I^+))$$

$$p_t(I^-) = P(I^- \rightarrow U_t) * (p_{t-1}(I^+) * P(I^+ \rightarrow I^-) + p_{t-1}(I^-) * P(I^- \rightarrow I^-)) \quad (9)$$

The data sets contain both the kind the contacts, those resulting in disease transmission and otherwise. Logical reasoning leads one to expect the highest correlation between a data set and the HMM developed from the same data set. This is not necessarily the case, as witnessed by the results.

A threshold value of (μ=0.01) or 1% is used in the generation of all the HMMs. The observed state U_1 results for every contact between two cells, wherein both cells have less than 1% infectious population. The observed state U_3 is witnessed for contact between cells, wherein both cells have more than 1% infectious population. For all other contacts, the observed state will be U_2.

For each observed state, the partial probability of a HMM generating that observed state is given by Eq. 9. The partial probabilities are added together to derive the evaluation of a HMM over a given data set. There are six data sets and correspondingly six HMMs. Figure 9.3 illustrates the evaluation of each of the data sets against each of the six HMMs.

Let the data sets be referred as $\{dataset-1, dataset-2, dataset-3, dataset-4, dataset-5, dataset-6\}$ and the corresponding HMMs be $\{HMM-1, HMM-2, HMM-3, HMM-4, HMM-5, HMM-6\}$. $HMM-6$ has the best evaluation score for $dataset-1$, ahead of $HMM-1$. $Dataset-2$ shows similar results upon evaluation by $HMM-2$, wherein $HMM-1$ evaluates best for $dataset-2$ followed by $HMM-2$. While the epidemic curves for $dataset-3$ and $dataset-4$ depict similar disease progression, as shown in Fig. 9.1, the comparison of evaluations for $dataset-3$ and $dataset-4$ by the HMMs depict similar

disease progression and behavioral interaction patterns. In contrast, irrespective of similar evaluations for $dataset-5$ and $dataset-6$ by the HMMs, the epidemic curves illustrate differences in the disease progression pattern over the temporal domain.

In order for a HMM developed for a dataset to evaluate and correlate higher to the same dataset in comparison to other datasets, multiple threshold values can be used, thereby creating newer observed states. Higher number of observed states is expected to produce well learned HMMs that illustrate the social behavioral interactions and disease progression at finer granularity and fidelity but at the expense of higher complexity.

9.3.4. *Inferences*

In epidemic theory, the reproduction number R_0 quantifies the number of susceptible individuals who get infected by one infectious person. In order for an epidemic to take place, the value of R_0 should be greater than 1. For endemic diseases, the value of R_0 is 1 and a steady state prevalence is witnessed at all times. The hidden Markov models learn the socio-behavioral dynamics in a population during the influenza outbreak. They quantify both the interactions that result in disease transmission as well as the passive contacts with null disease spread. This feature of quantifying both the types of contacts is unique to the use of HMMs. In addition to quantifying the epidemic impact by incidence and prevalence, the enhanced understanding of social behavioral interactions will help to identify control points that aid or discourage the influenza progression.

The observed states used in the study can be supplemented with multiple thresholds, thereby adding newer observed states. Also, in order to address the behavioral changes in the population in reaction to an outbreak, use of dynamic hidden Markov models will enhance the analysis by developing different HMMs for different temporal phases of disease progression.

9.3.4.1. *Dynamic Hidden Markov Models.* Hidden Markov models (HMMs) are valuable in analyzing the temporal dynamics during a disease progression in a population. The state

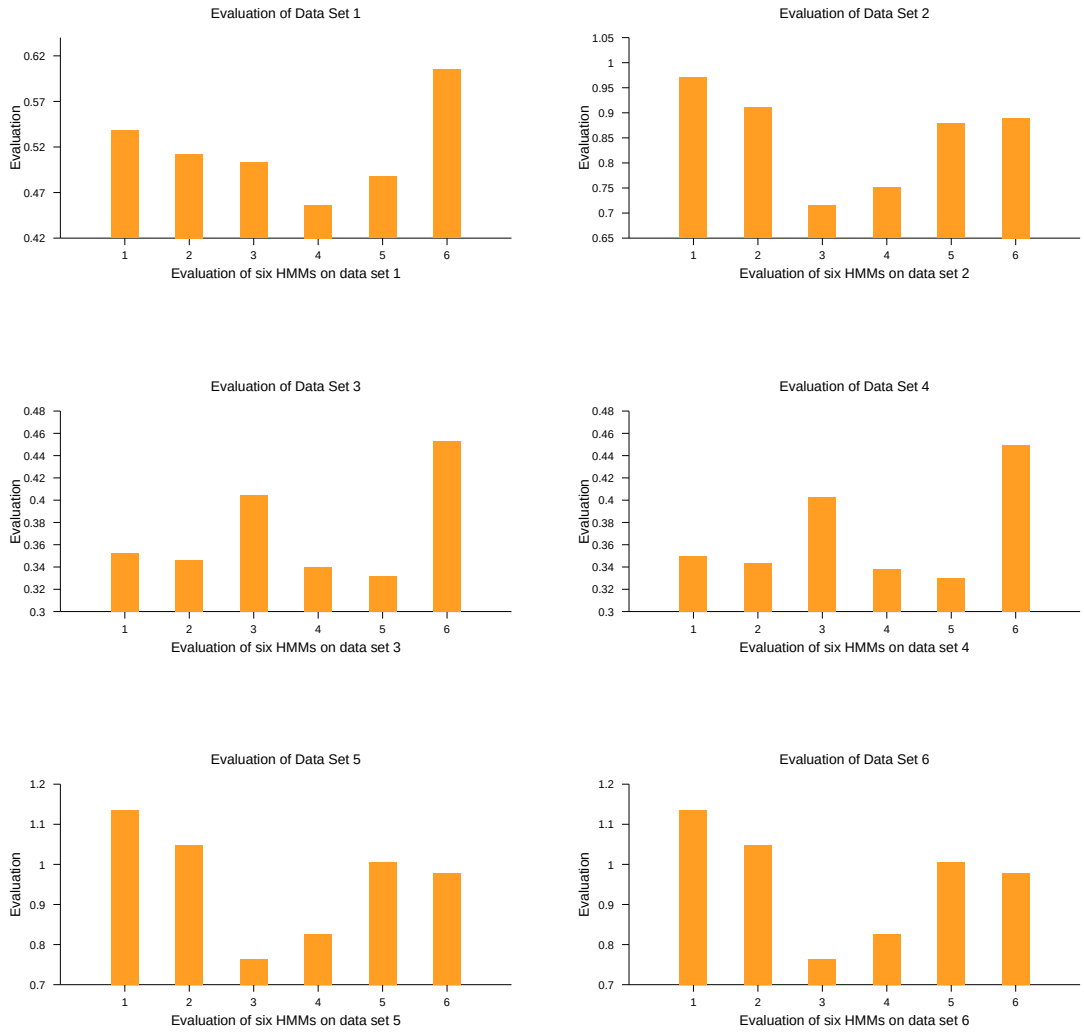


Figure 9.3. Comparative Analysis of Social Behavioral Interactions

transition matrix portrays the social behavioral interactions during the course of the full outbreak. Nevertheless, it is independent of time and is a prime limitation.

The social behavioral interactions will undergo changes during a disease outbreak in the real world. In order to capture the evolving behavior in response to an outbreak, dynamic hidden Markov models are proposed that extend the HMMs; similar to the extension of Bayesian networks to dynamic Bayesian networks,

The disease outbreak timeline can be split into periodic time intervals. For each time interval, the corresponding HMM can be developed. Applying first order Markov process,

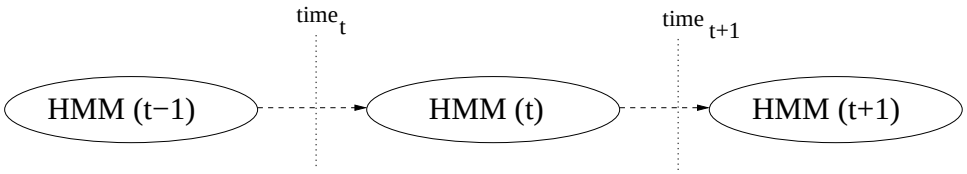


Figure 9.4. Dynamic Hidden Markov Models

HMMs, each pair of consecutive time intervals can be correlated, as shown in Fig. 9.4. In a general scenario, for a n^{th} order Markov process, HMMs of n continuous time intervals will have dependency relationships associated with them.

In case of severe outbreaks, quarantine of infectious individuals will be executed. Also, the susceptible individuals will orient towards behavioral interactions that have minimal risk of contracting the infection. This change in behavior can be illustrated by a different HMM, compared to the initial HMM portraying the behavioral interactions under normal circumstances. Dynamic HMMs will enhance the precision of analysis of social behavioral interactions during a disease outbreak in a population.

CHAPTER 10

SUMMARY

The focus of this study is to use probabilistic reasoning in spatiotemporal analysis of infectious diseases. Analyses of public health data involve a degree of uncertainty due to incomplete data sets. Probabilistic reasoning complements the existing inferential statistics methodologies to derive analytical results with a degree of confidence. Public health resource allocation and policy making stands to gain by use of computational methodologies to execute what-if analysis of epidemic scenarios. Modeling and simulation provides a platform to test different hypotheses and public health strategies, that can not be readily tested in the field due to ethical concerns.

10.1. Complexity of Disease Analysis

Genetics, environment, and population dynamics play an inter-dependent pivotal role on the disease manifestation in an individual as well as in the disease progression in a population. Infectious diseases are continually emerging due to the evolving changes in the environment, immune response systems, and mobility patterns of people. The viral disease strains adapt to become resistant to prevalent drugs; thereby lending to development of newer vaccines to protect against the newer strains. Vaccine dissemination among the different demographic sub-groups in a population and the timeline for dissemination will have to be effective to attain herd immunity. Disease analyses from the perspective of genetics to population dynamics present complex issues that require good understanding of the fundamental principles that aid in the evolution of viral strains as well as the disease progression in a population.

10.1.1. *Computational Epidemiology*

The primary aim of computational epidemiology is to apply computational science paradigms to the field of public health, thereby providing methodologies and tools for epidemiologists and scientists in the public health domain. These novel methods will aid in the prediction and analysis of disease manifestation and spread in a given population through modeling, simulation, and visualization, thereby enable epidemiologists to conduct focused what-if-analyses that facilitate the allocation of public health resources.

The ability to predict how a disease might manifest itself in the population at large is essential for identifying disease monitoring and control strategies. Epidemiologists are traditionally relying on data that have been collected during previous outbreaks. However, for newly emerging or re-emerging infectious diseases, such data are often unavailable or outdated. Changes in population composition and dynamics require the design of models that bring together knowledge of the specific infectious diseases with the demographics and geography of the region under investigation. Expertise from diverse domains are forged together to develop new scientific methods that will enhance the understanding of the complexities of disease dynamics in a population.

10.2. Epidemic Analysis using Probabilistic Reasoning

Spatiotemporal analysis of infectious diseases using probabilistic reasoning enhances the planning and development of policies for optimal allocation of public health resources. Probabilistic reasoning under uncertainty suits well to analysis of infectious disease dynamics. The principles of Bayesian probabilistic reasoning is used to learn the stochastic dynamics of the progression of infectious diseases. Quantitative and qualitative analyses of the disease progression, including the incidence and prevalence, will aid in predicting the spatiotemporal outbreak patterns for a demographic population in a specific geographic region.

Enhanced understanding of disease progression will facilitate in identifying the critical points of surveillance, control and prevention. Public health resources are to be prioritized to

the higher risk groups of the population in order to curtail the disease spread and epidemic outbreaks as early as possible. In accordance with the mission of the upcoming field of computational epidemiology to research and develop newer computational models that facilitate prediction, prevention, and control of epidemics, there are four different phases of analysis presented in this study.

10.2.1. *Demographic Risk Analysis*

Artificial data sets for influenza outbreaks are generated using a Bayesian network and the influenza infection timeline. The Bayesian network represents the probabilistic dependencies among the demographics, symptoms and influenza prevalence. Demographic based epidemic curves are generated and used to rate the risk levels associated with the different demographic subgroups to influenza infection. Temporal analysis of the epidemic curves lends to developing a priority list of the demographic sub-groups, based on their risk levels.

Both cases of vaccination and null vaccination are executed to analyze the extent of decline in influenza prevalence by vaccination of sections of a population. While a cent per cent supply of vaccines for the entire population is the ideal scenario of protecting the full population, vaccines are in short supply and limited in the real world situations. Public health policy making with constrained resources is implemented on the influenza outbreak data sets. A hypothetical scenario of vaccines limited to only 10% of the population is tested. Spread vaccination is used in the first case, wherein the limited vaccines are uniformly disseminated among all the demographic sub-groups. The second case tests the deployment of the vaccine uniformly among the three lowest risk demographic sub-groups. The resulting epidemic is observed to be higher in comparison to spread vaccination. The final case tests the vaccine dissemination uniformly among the three highest demographic sub-groups. The resulting epidemic is observed to be less severe in comparison to spread vaccination. This result justifies the identification of high risk demographic sub-groups, and prioritize the limited public health resources among the high risk groups.

10.2.2. *Spatial Correlation of Disease Prevalence*

Bayesian networks are used to represent the prevalence of influenza and pneumonia in different geographic regions with corresponding different demographic compositions. Influenza and pneumonia outbreak in a geographic region is illustrated by a Bayesian network, that includes the probabilistic dependencies of demographics, symptoms and prevalence of influenza and pneumonia. The demographic composition of a different geographic region is available and the task is to predict the prevalence for influenza and pneumonia for the new region. The probabilistic dependencies between demographics and prevalence of the known geographic region is correlated with the demographic composition of the new geographic region to predict the prevalence among the different sub-groups. The results indicate the risk order levels associated with the different demographic sub-groups are different for the two geographic regions. This exemplifies the worthy use of probabilistic reasoning to predict the most likely disease manifestation among the different demographic sub-groups. This analysis can be extended to learn the probabilistic dependencies among the demographics and disease progression from multiple geographic regions and predict the prevalence for an outbreak in a new geographic region.

10.2.3. *Temporal Analysis*

Until 1998, the process of reporting new cases of HIV to the public health departments was voluntary. Physicians and clinicians who diagnose HIV among their patients were reluctant to report these cases. This led to under reporting and non-reliability of the data collected until 1998. Ever since the mandate was passed for compulsory reporting of new HIV cases to the public health department, both reliability and completeness associated with the measure of HIV incidence and prevalence has improved. Nevertheless, a predictive analysis is needed to account for HIV positive cases that are yet to be tested or diagnosed. In general, analyses on public health data have problems of incomplete data, missing values, as well as lack of some

significant parameters. A robust analysis has to take into account these shortcomings and be fault tolerant to infer results with the highest possible degree of confidence.

Bayesian networks are developed for time intervals of two years from 1989-2002. Bayesian network are developed by integrating the HIV surveillance dataset and the demographic composition of Texas obtained from the latest census data. The proportions of HIV incidence among the different demographic sub-groups are used to compute the conditional probabilities of the Bayesian network. In addition, Bayesian networks of two consecutive time intervals are correlated using a first order Markov process, thereby developing a dynamic Bayesian network to analyze the HIV incidence from 1989-2002. Demographic based epidemic curves are derived to analyze HIV incidence. Unreliable data until 1998 led to a conservative analysis. Demographic based risk analysis allows for the observation of a varied spectrum of HIV risk among the different demographic subgroups. A predictive analysis of HIV incidence in 2000 and 2002 from earlier Bayesian networks representing surveillance from 1989-1998 led to inconclusive results. The analyzed parameters are deficient to warrant a probabilistic prediction of the disease progression in the upcoming years.

10.2.4. *Socio-behavioral Analysis*

A methodology using hidden Markov models (HMMs) is introduced that facilitates the investigation of the impact of social behavioral interactions in the incidence and prevalence of infectious diseases. The methodology is presented in the context of simulated disease outbreak data for influenza. The extent of calamity of an infectious disease in a population is measured by incidence, prevalence, rates of morbidity and mortality. These parameters represent the ill-effects of the disease on the population but do not quantify the risk involved for each contact between any two individuals.

The analysis of social behavioral interactions using HMMs quantifies the active contacts resulting in influenza transmission as well as the passive contacts that do not lead to any

transmission. This helps to understand the risk level associated in a random contact. Reproduction number R_0 represents the effective number of infections inflicted by an infectious individual. In addition to this information provided by R_0, HMMs also account for the associated number of contacts made by an infectious individual without infection transmission.

10.3. Future Work

Bayesian networks from different demographic and geographic settings can be correlated to predict the spatial flow of the disease. Further, correlation can be applied to the temporal flow of infectious disease epidemics across different geographic regions with varied demographics. The understanding of the spatial-temporal flow of the disease, coupled with the disease pertinent characteristics, will aid in the identification of disease epidemic properties. To this end, the analysis can be strengthened to include different interaction patterns among demographic sub-groups as well as ethnicity specific behaviors and characteristics.

The study can be applied to real data collected from past infectious diseases outbreaks. This will help in quantifying the limitations and validity of probabilistic analysis for epidemic data. Although, the analysis has focused on epidemic disease outbreaks, it can be adapted to study the steady state prevalence of endemic diseases.

The limitations of hidden Markov models (HMMs) in representing the social behavioral interactions in a population can be alleviated by use of dynamic HMMs. HMMs for different time intervals can be developed and temporally adjacent HMMs can be correlated either as a first or higher order Markov process. Such an analysis is expected to incorporate the knowledge of changes in social behavioral interactions during different phases of a disease outbreak.

10.4. Multi-disciplinary Collaboration

In the quest to improve the community health, public health of today brings together hitherto disparate fields, including epidemiology, medical geography, environmental sciences, molecular biology, biostatistics, sociology, mathematics and computer science. The diverse

fields complement each other and collaborate on concepts and methodologies to facilitate the public health decision and policy making process. The intrinsic complexity of modeling a suite of known and unknown parameters affecting an infectious disease outbreak make it imperative to take advantage of today's high performance computing infrastructure. High performance computing facilitates hypothesis testing and what-if analyses of scenarios that do not readily lend themselves to field testing. These include epidemic analysis, vaccination strategies and different resource allocation policies. The upcoming field of computational epidemiology will aid in the prediction and analysis of disease manifestation and spread in a given population through modeling, simulation and visualization; thereby leading to effective public health policy making.

10.5. Final Remarks

Human population dynamics are continually increasing in complexity, especially in today's modern world with rapid mobility and interaction patterns. The disease pathogens of molecular size and invisible to our naked eye co-exist, adapt continually, and challenge the human existence at the top of the food chain. The biochemical path of the pathogen within species and across species, and the immune response systems in fighting the pathogen add another layer of complexity. All living organisms are continually evolving to the changes in the co-habiting environment. Disease progression analysis couples together the understanding of viral genetics, environment and population dynamics, to infer the best prevention strategies in thwarting disease outbreaks. The road to discovery in this realm of science will fascinate and challenge our human mind and spirit for times to come.

BIBLIOGRAPHY

[1] "Guidelines for Preventing the Transmission of Mycobacterium Tuberculosis in Health-Care Facilities," Centers for Disease Control and Prevention," MMWR, October 1994.

[2] *Biosafety in Microbiological and Biomedical Laboratories*, 4th ed. Centers for Disease Control and Prevention and National Insitutes of Health, April 1999. [Online]. Available: http://bmbl.od.nih.gov/

[3] K. Abbas, A. Mikler, and R. Gatti, "Temporal Analysis of Infectious Diseases: Influenza," in *Proc. of the* 20^{th} *Annual ACM Symposium on Applied Computing*, Sante Fe, NM, March 2005.

[4] K. Abbas, A. Mikler, A. Ramezani, and S. Menezes, "Computational Epidemiology: Bayesian Disease Surveillance," in *Proc. of the International Conference on Bioinformatics and its Applications*, FL, USA, December 2004.

[5] B. Abramson, J. Brown, W. Edwards, A. Murphy, and R. Winkler, "Hailfinder: A Bayesian System for Forecasting Severe Weather," *International Journal of Forecasting*, vol. 12, no. 1, pp. 57–72, 1996.

[6] A. Adams, J. Koopman, S. Chick, and P. Yu, "GERMS: An Epidemiologic Simulation Tool for Studying Geographic and Social Effects on Infection Transmission," in *Proc. of the* 31^{st} *conference on Winter simulation: Simulation - a bridge to the future*, vol. 2, 1999, pp. 1549–1556.

[7] E. Allman and J. Rhodes, *Mathematical Models in Biology An Introduction*. Cambridge University Press, 2004.

[8] J. Aron, *Mathematical Modeling: The Dynamics of Infection*. Gaithersburg, MD: Aspen Publishers, 2000, ch. 6.

[9] D. Ashby, J. Hutton, and M. McGee, "Simple Bayesian Analysis for Case-control Studies in Cancer Epidemiology," *The Statistician*, vol. 42, pp. 385–397, 1993.

[10] R. Assuncao and M. Castro, "Multiple Cancer Sites Incidence Rates Estimation using a Multivariate Bayesian Model," *International Journal of Epidemiology*, 2004.

[11] R. Bagni, R. Berchi, and P. Cariello, "A Comparison of Simulation Models Applied to Epidemics," *Journal of Artificial Societies and Social Simulation*, vol. 5, no. 3, 2002.

[12] N. Bailey, *The Mathematical Theory of Epidemics*. NY, USA: Hafner Publishing Company, 1957.

[13] ——, "The Simulation of Stochastic Epidemics in Two Dimensions," in *Proc. of the* 5^{th} *Berkeley Symposium on Mathematics and Statistics*, vol. 4. Berkeley and Los Angeles, CA: University of California, 1967.

[14] S. Bankes, "Agent-based Modeling: A Revolution?" *Proc Natl Acad Sci*, vol. 99, pp. 7199–7200, 2002.

[15] A. Benenson, Ed., *Control of Communicable Diseases Manual*. American Public Health Association, 1995.

[16] A. Benyoussef, N. Boccara, H. Chakib, and H. Ez-Zahraouy, "Lattice Three-species Models of the Spatial Spread of Rabies among Foxes," *International Journal of Modern Physics C*, vol. 10, pp. 1025–1038, 1999.

[17] (2004) BIOWAR: Simulating Disease Outbreaks using Social Networks. [Online]. Available: www.casos.ece.cmu.edu/projects/BioWar/biowar.doc

[18] N. Boccara and K. Cheong, "Critical Behavior of a Probabilistic Automata Network SIS Model for the Spread of an Infectious Disease in a Population of Moving Individuals," *Journal of Physics A: Mathematical and General*, vol. 26, no. 5, pp. 3707–3717, 1993.

[19] E. Bonabeau, L. Toubiana, and A. Flahault, "Evidence for Global Mixing in Real Influenza Epidemics," *Journal of Physics A: Mathematical and General*, vol. 31, pp. L361–L365, 1998.

[20] A. Callaghan, "Agent-Based Modelling applied to HIV/AIDS," *ERCIM News*, January 2005.

[21] N. Cancre, A. Tall, C. Rogier, J. Faye, O. Sarr, J. Trape, A. Spiegel, and F. Bois, "Bayesian Analysis of an Epidemiologic Model of Plasmodium Falciparum Malaria Infection in Ndiop, Senegal," *American Journal of Epidemiology*, vol. 152, no. 8, pp. 760–770, 2000.

[22] H. Carabin, M. Escalona, C. Marshall, S. Vivas-Martinez, C. Botto, L. Joseph, and M. Basanez, "Prediction of Community Prevalence of Human Onchocerciasis in the Amazonian Onchocerciasis Focus: Bayesian Approach," *Bulletin of the World Health Organization*, vol. 81, no. 7, pp. 473–550, 2003.

[23] (2004) The CDC Influenza Web Page. [Online]. Available: http://www.cdc.gov/flu

[24] (2004) Computational Epidemiology Research Lab (CERL). [Online]. Available: http://www.cerl.unt.edu

[25] J. Chin, Ed., *Control of Communicable Diseases Manual*, 17th ed. American Public Health Association, 2000.

[26] (2006) Clinical Trials. [Online]. Available: http://www.clinicaltrials.gov/

[27] C. Corley and A. Mikler, "Predicting Human Papilloma Virus Prevalence and Vaccine Policy Effectiveness," in *Proc. of the IEEE 5th Symposium on Bioinformatics and Bioengineering, BIBE05*, Minneapolis, MN, October 2005.

[28] (2004) Center for Discrete Mathematics and Theoritical Computer Science (DIMACS). [Online]. Available: http://dimacs.rutgers.edu

[29] M. Duryea, T. Caraco, G. Gardner, W. Maniatty, and B. Szymanski, "Population Dispersion and Equilibrium Infection Frequency in a Spatial Epidemic," *Physica D*, vol. 132, pp. 511–519, 1999.

[30] M. El-Sheikh and S. El-Marouf, "On Stability and Bifurcation of Solutions of an SEIR Epidemic Model with Vertical Transmission," *International Journal of Mathematics and Mathematical Sciences*, 2004.

[31] S. Eubank, "Scalable, Efficient Epidemiological Simulation," in *Proc. of the* 17^{th} *Annual ACM Symposium on Applied Computing (SAC'02)*, Madrid, Spain, 2002.

[32] P. Fayers, D. Ashby, and M. Parmar, "Tutorial in Biostatistics: Bayesian Data Monitoring in Clinical Trials," *Statistics in Medicine*, vol. 16, pp. 1413–1430, 1997.

[33] (2006) Food and Drug Administration. [Online]. Available: http://www.fda.gov/

[34] S. Fu, "Modelling Epidemic Spread through Cellular Automata," Master's thesis, The University of Western Australia, 2002.

[35] S. Fu and G. Milne, "Epidemic Modelling Using Cellular Automata," in *Proc. of the Australian Conference on Artificial Life*, 2003.

[36] H. Fukś and A. Lawniczak, "Individual-based Lattice Model for Spatial Spread of Epidemics," *Discrete Dynamics in Nature and Society*, vol. 6, pp. 191–200, 2001.

[37] L. Gordis, *Epidemiology*. W.B. Saunders Company, 2000.

[38] H. Hamer, "Epidemic Disease in England," *Lancet*, vol. 1, pp. 733–739, 1906.

[39] D. Heckerman, *Probabilistic Similarity Networks*. Cambridge, MA: MIT, 1991.

[40] D. Heymann and G. Rodier, "Global Surveillance, National Surveillance, and SARS," *Emerging Infectious Diseases*, vol. 10, no. 2, February 2004.

[41] (2004) Hidden Markov Models Tutorial Website. [Online]. Available: http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html_dev/main.html

[42] N. Jones and P. Pevzner, *An Introduction to Bioinformatics Algorithms*. MIT Press, 2004, ch. 11.

[43] E. Kalipeni, S. Craddock, J. Oppong, and J. Ghosh, *HIV and AIDS in Africa: Beyond Epidemiology*. Oxford UK: Blackwell Publishing, 2004.

[44] S. Klug and M. Cummings, *Essentials of Genetics*, 5th ed. Pearson Education Inc., 2005.

[45] K. Korb and A. Nicholson, *Bayesian Artificial Intelligence*. CRC Press, 2004.

[46] S. Levin, B. Grenfell, A. Hastings, and A. Perelson, "Mathematical and Computational Challenges in Population Biology and Ecosystems Science," *Science*, vol. 275, pp. 334–343, 1997.

[47] F. Liljeros, C. Edling, L. Amaral, H. Stanley, and Y. Aberg, "The Web of Human Sexual Contacts," *Nature*, vol. 411, pp. 907–908, 2001.

[48] H. Markel, "The Search for Effective HIV Vaccines," *New England Journal of Medicine*, vol. 353, no. 8, pp. 753–757, August 2005.

[49] A. Mikler, S. Venkatachalam, and K. Abbas, "Modeling Infectious Diseases using Global Stochastic Cellular Automata," *Journal of Biological Systems*, vol. 13, no. 4, pp. 421–439, December 2005.

[50] R. Neapolitan, *Learning Bayesian Networks*. Pearson Prentice Hall Series, 2004.

[51] K. Nelson, C. Williams, and N. Graham, *Infectious Disease Epidemiology*. Aspen, 2001.

[52] D. O'Leary, "Models of Infection: Person to Person," *Computing in Science & Engineering*, vol. 6, no. 1, Jan-Feb 2004.

[53] J. Oppong and S. Arbona, "HIV-AIDS in Texas," in *Proc. of the IXth International Symposium in Medical Geography*, Montreal, Canada, July 2000.

[54] J. Oppong and S. Ramisetty-Mikler, "Race and HIV-AIDS in Texas," in *Proc. of the Race, Ethnicity and Place Conference*, Washington DC, September 2004.

[55] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann, 1988.

[56] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, February 1989.

[57] A. Raftery, "Bayesian Model Selection in Social Research," *Sociological Methodology*, 1995.

[58] T. Rath, M. Carrerar, and P. Sebastiani, "Automated Detection of Influenza Epidemics Using Hidden Markov Models," in *Proc. of the* 5^{th} *International Symposium on Intelligent Data Analysis*, Berlin, Germany, August 2003.

[59] R. Ross, "An Application of the Theory of Probabilities to the Study of A Priori Pathometry, I," *Proc. Roy. Soc.*, vol. A, no. 92, pp. 204–230, 1916.

[60] S. Russell and P. Norvig, *Artificial Intelligence A Modern Approach*, 2nd ed. New Jersey: Prentice Hall, 2003.

[61] B. Schönfisch, "Zelluäre Automaten und Modelle für Epidemien," Ph.D. dissertation, University of Tübingen, 1993.

[62] H. Situngkir, "Epidemiology Through Cellular Automata," Bandung Fe Institute, Tech. Rep., March 2004.

[63] D. Spiegelhalter, J. Myles, D. Jones, and K. Abrams, "An Introduction to Bayesian Methods in Health Technology Assessment," *BMJ*, vol. 319, pp. 508–512, August 1999.

[64] D. Stefano, H. Fukś, and A. Lawniczak, "Object-oriented Implementation of CA/LGCA Modeling Applied to the Spread of Epidemics," in *Canadian Conference on Electrical and Computer Engineering*. Halifax: IEEE, 2000, pp. 26–31.

[65] M. Steinhoff, *Epidemiology and Prevention of Influenza*. MD: Infectious Disease Epidemiology, Aspen Publishers, 2000, ch. 16.

[66] J. Thomas and D. Weber, *Epidemiologic Methods for the Study of Infectious Diseases*. Oxford Press, 2001.

[67] T. Timmreck, *An Introduction to Epidemiology*. Boston: Jones and Bartlett, 2002.

[68] F. Tsui, J. Espino, V. Dato, P. Gesteland, J. Hutman, and M. Wagner, "A Real-time Public Health Surveillance System," *Journal of the American Medical Informatics Association*, vol. 10, no. 5, pp. 399–408, September 2003.

[69] (2004) Texas Department of State Health Services. [Online]. Available: http://www.dshs.state.tx.us/

[70] (2000) United States Census Bureau. [Online]. Available: http://www.census.gov/

[71] S. Venkatachalam and A. Mikler, "An Infectious Outbreak Simulator based on the Cellular Automata Paradigm," in *Proc. of the International Conference on Innovative Internet Community Systems*, Guadalajara, Mexico, June 2004.

[72] ——, "Towards Computational Epidemiology: Using Stochastic Cellular Automata in Modeling Spread of Diseases," in *Proc. of the* 4^{th} *Annual International Conference on Statistics*, January 2005.

[73] C. Viboud, P. Bolle, K. Pakdaman, F. Carrat, A. Valleron, and A. Flahault, "Influenza Epidemics in the United States, France, and Australia, 1972-1997," *Emerging Infectious Diseases*, vol. 10, January 2004.

[74] D. West, *Introduction to Graph Theory*, 2nd ed. Prentice Hall, 2001.

[75] (2004) WHO Influenza Web Page. [Online]. Available: http://www.who.int/csr/disease/influenza/en/

[76] D. Willis, "Ambulation Monitoring and Fall Detection System using Dynamic Belief Networks ," *Bachelors Thesis, Monash University*, 2000.