OPTIMIZING NON-PHARMACEUTICAL INTERVENTIONS USING MULTI-COAFFILIATION NETWORKS

Olivia G. Loza

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

May 2013

APPROVED:

Armin R. Mikler, Major Professor
Subhash Aryal, Committee Member
Rada Mihalcea, Committee Member
Robert Renka, Committee Member
Chetan Tiwari, Committee Member
Barrett Bryant, Chair of the Department of Computer Science and Engineering
Mark Wardell, Dean of the Toulouse Graduate School

Loza, Olivia G. Optimizing Non-Pharmaceutical Interventions Using Multi-Coaffiliation Networks. Doctor of Philosophy (Computer Science), May 2013, 130 pp., 25 tables, 44 illustrations, 169 numbered references.

Computational modeling is of fundamental significance in mapping possible disease spread, and designing strategies for its mitigation. Conventional contact networks implement the simulation of interactions as random occurrences, presenting public health bodies with a difficult trade-off between a realistic model granularity and robust design of intervention strategies.

Recently, researchers have been investigating the use of agent-based models (ABMs) to embrace the complexity of real world interactions. At the same time, theoretical approaches provide epidemiologists with general optimization models in which demographics are intrinsically simplified. The emerging study of affiliation networks and co-affiliation networks provide an alternative to such trade off. Co-affiliation networks maintain the realism innate to ABMs while reducing the complexity of contact networks into distinctively smaller k-partite graphs, were each partition represents a dimension of the social model.

This dissertation studies the optimization of intervention strategies for infectious diseases, mainly distributed in school systems. First, concepts of synthetic populations and affiliation networks are extended to propose a modified algorithm for the synthetic reconstruction of populations. Second, the definition of multi-coaffiliation networks is presented as the main social model in which risk is quantified and evaluated, thereby obtaining vulnerability indications for each school in the system. Finally, maximization

of the mitigation coverage and minimization of the overall cost of intervention strategies are proposed and compared based on centrality measures.

ACKNOWLEDGMENTS

This dissertation would have not been possible without the love and support of my family. It is impossible to express in words how much the trust of my parents Grace and Eduardo mean to me, and how their own academic and work careers have inspired me to pursue a doctorate degree. I would like to acknowledge and profoundly thank the contributions made by Prof. Armin Mikler to this research and my academic work. His constant guidance, ubiquitous knowledge across diverse science fields, and furthermore innate mentoring skills have been invaluable. He has inspired me and countless researchers to pursue the advancement of science as a personal goal. I am also deeply grateful to Dr. Rada Mihalcea, Dr. Subhash Aryal, Dr. Robert Renka, and Dr. Chetan Tiwari for their extremely valuable feedback, insightful comments, and for serving as committee members. Their time, effort, and altruistic help provided a quality assurance for this research. Additionally, I would like to acknowledge Dr. Farhad Shahrokhi, not only for his incommensurate technical knowledge, from which I have benefited greatly, but also his ability to share that knowledge. I am so grateful: To Reynaldo, the love of my life, my inspiration, my motivation and overall, my reason to be.

To Christian, the guru and code master, I am blessed to have you as my brother, you are the light that showed me the path to follow.

To Daniela, Giovanna, Vanessa, Tomyo and Martino my beloved sisters and brothers for their support.

To my colleagues and CERL team mates, Iris, Jessica, Tamara, Tina, Angel, David, Erwin, Jedsada, Jorge, Marty, Phat, and Sarat.

To all the people who helped me though my academic career and to all the people who will find this work useful.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Computational epidemiology is the formal study of how technological tools can be applied to analyze, visualize, and understand the dynamics of disease. Its mathematical origin connects the elegance of differential equations, graph theory, and related areas with the reality of our world. Like all new fields pertaining to the area of modern computer science, it has flourished through the use of high performance computing and parallelization techniques. While this field shows great potential for public health applications, the formulation of models which, represent human societies accurately is still an open question.

Through years of research, computational epidemiologists have focused their efforts on two central problems. First, establishing models that represent networks of social relationships, as close to reality as possible. Second, constructing frameworks that allow the study of what-if scenarios to simulate the spread of disease and to analyze feasible ways to respond to threats. A key challenge in developing computational models is the validation of their correspondence with human behavior. In order to disperse, human diseases are constrained to people or vectors (i.e. mosquitoes) therefore, the more accurate the social network, the better understanding of disease dynamics. Partly in response to the challenge of validation, data collections coming from online communities have become the norm for sources of information. On the other hand, public health is a global concern. Furthermore, because globalization and air travel are becoming ever more frequent, preparing for pandemics and world-traveling infectious agents is a reality confronted through simulation of different scenarios and attainable responses.

The next section outlines the challenge of cost-efficient non-pharmaceutical interventions, including the difficulties of designing social models. Section 1.2 describes the state-of-the-art regarding mathematical and simulation models for optimization methodologies. Section 1.3 presents a novel algorithm to construct and validate networks of communities. Finally, the main contributions of this dissertation are detailed in Section 1.4, and disserta-

tion's structure is presented in Section 1.5.

1.1. The Problem of Optimizing Intervention Strategies

A fundamental reason for studying disease epidemic models is to improve implementation of disease control. As an interdisciplinary endeavor, mathematical and computational models have been fundamental tools applied to the evaluation of different intervention strategies. On one hand, mathematical approaches based on strong assumptions (i.e. homogeneous, closed populations) formulate the models in terms of exactly solvable problems to establish epidemic thresholds and the probability of a large outbreak. On the other hand, computational models rely highly on incorporating realism by the power of parallization and super-computing. This research is an effort to leverage the use of theoretical tools in combination with highly parameterized simulations to find an effective combination of these two approaches oriented to the optimization of disease intervention strategies.

Allocation and deployment of antiviral treatment and prophylaxis are inherently complex and highly technical problems because they require identifying groups of the population that should be prioritized. Public health bodies attempt to optimize the distribution of scarce or costly control mechanisms to maximize their impact on the outbreak dynamics. Risk identification has focused on schools and child-care centers mainly because they represent dense masses of highly immunologically naive hosts for the pathogens. It is believed that "interventions targeted at school-aged children, should be most effective in the early stages of an outbreak" [15]. Measures are significantly more valuable at the start of the pandemic when the incidence becomes comparable among children and adults [157]. For seasonal viruses like influenza, the U.S. Centers for Disease Control and Prevention (CDC) use a short nomenclature describing the type and subtype of the virus. Likewise, as a result of the SARS epidemic, the European Union created the European Centre for Disease Prevention and Control (ECDC) making their mandate to control communicable diseases[141]. Globally, the existent three types of influenza viruses are classified as: A, B, and C. Subtypes only exist for influenza A viruses and are further divided "on the basis of the two main surface glycoproteins hemagglutinin (HA) and neuraminidase (NA)" [60]. For example, an

"H1N1 virus" designates the influenza A subtype that has an HA 1 protein and an NA 1 protein.

Fig. 1.1 shows the estimated number of annual influenza-associated deaths from 1976 to 2010, by age range. In Europe H1N1 attack rates showed a substantial variation that depended on the socio-demographic structure, school calendars, mobility patterns and sociodemographic structures [110]. Schools closures have been found to be an important mitigation tool for infectious diseases. During the 2009 H1N1 pandemic, the CDC left the decision of school closings to local officials. Initially, the CDC recommended that schools should close for confirmed or suspected cases, but as the pandemic progressed, there was not a clear perspective on the effectiveness that the measure had on mitigating the spread of the disease. The decision to close schools varied considerably from community to community; as a result, school closings and the delivery of targeted vaccines were the most frequent mitigation response [139]. The global response to H1N1 shows the necessity of a decision support tool for targeted interventions. To facilitate the design of policies to mitigate regional epidemics, some models have considered schools as isolated entities, focusing on specific student populations to estimate the impact of an outbreak. Closing schools is a controversial decision mainly due to the hidden costs associated to epidemiological benefits.

Lowering the epidemic severity through reducing school-age contacts is an important component of the U.S. mitigation strategy [101] and it involves two types of costs. First, absenteeism of workers who stay at home while children are out of school; and second, the fraction of those workers that are health care providers themselves. These two aspects have a negative impact on the contingency tasks and their cost could be substantial [101]. In a pioneer study, the economic impact of closing schools in the United Kingdom has been estimated in terms of loss income of working parents.

From information taken in 2008, under a global threat, the cost of closing all schools

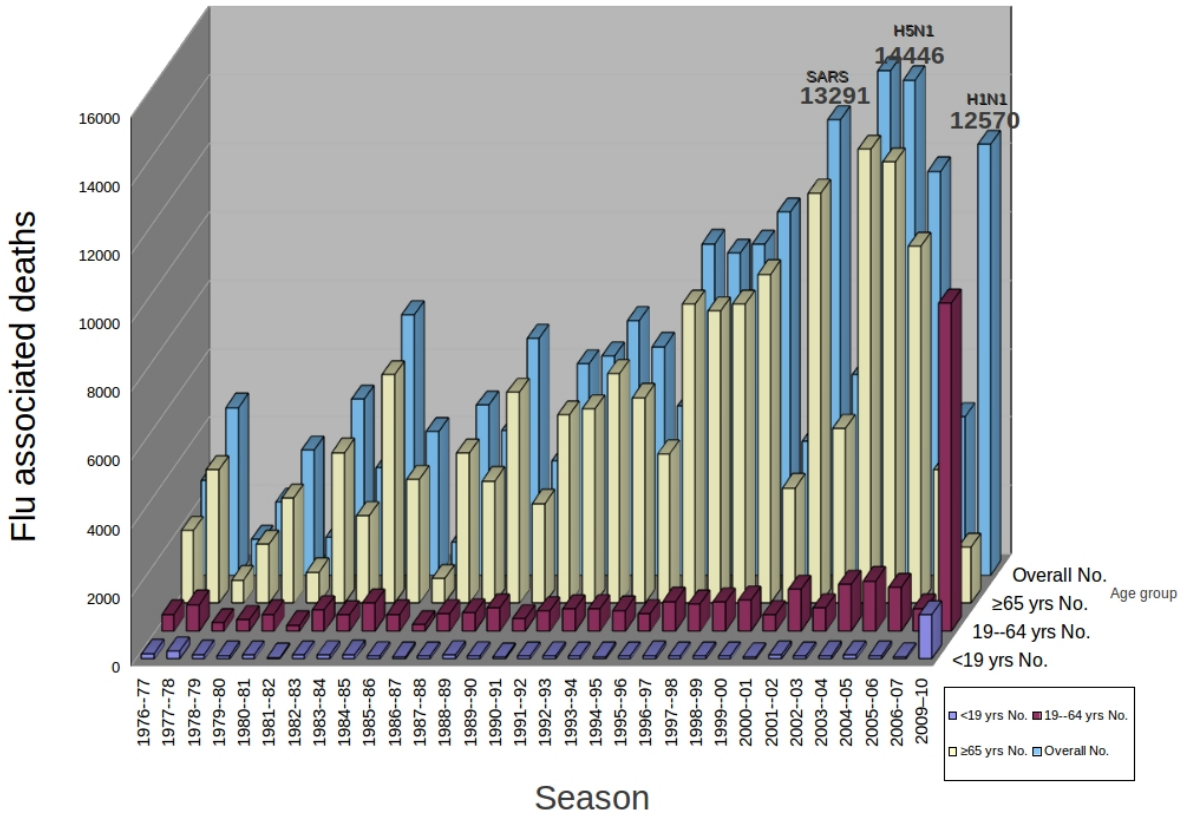


FIG. 1.1. Estimated number of annual influenza-associated deaths (1976-2010) [61]

for four weeks in the United Kingdom would be between 0.1% and 0.4% gross domestic product (GDP) [137]. By the time statistics were retrieved from the data source, 38% of the workforce had dependent children. According to Census 2010, American family households account for 66.4% of all households, and 49.6% of family households have children under 18 years, a percentage that has slowly but undoubtedly increased over the last 50 years, as shown in Fig. 1.3. The Center on Social and Economic Dynamics (CSED) estimated that closing all schools in the U.S. for four weeks could cost between $10 and $47 billion (0.1%-0.3% of GDP) and would lead to a reduction of 6% to 19% in key health care personnel (Fig. 1.2).

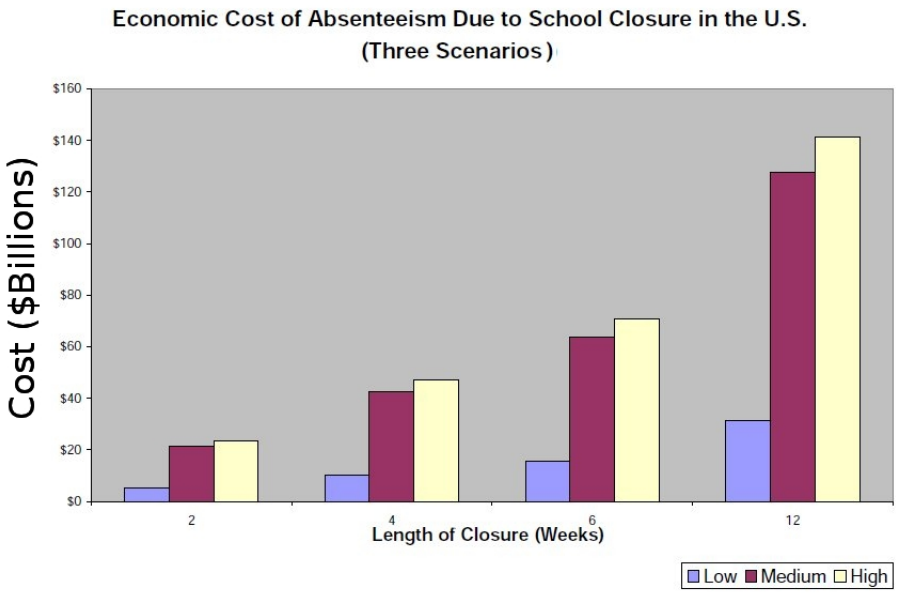


FIG. 1.2. Potential cost of the closure of all U.S. schools for four weeks [101]

The World Health Organization (WHO) declared the H1N1 pandemic officially over in August 10, 2010. Despite the measures taken, H1N1 reached pandemic status, showing that uncertainty is always present in epidemics. Dr. Margaret Chan, WHO Director-General at her Opening Intervention at the International Health Regulations Review Committee in Geneva, Switzerland 2010, stated that "the world was better prepared for a pandemic than at any time in history, but it was prepared for a different kind of event than what actually occurred" [33].

1.2. Current Approaches to Optimization

To achieve maximum effectiveness at minimal cost, when using one additional unit of prophylaxis or vaccine, key aspects of the structure of the community need to be considered. Since experimental investigation of disease dynamics is for the most part unfeasible, and mostly not considered ethical, researchers have turned their efforts to building simulated scenarios to study outbreak dynamics. Analysis of epidemic models is crucial and is mainly based on information reported after an outbreak has occurred. Because of the scarcity of complete information, results have a sensitive dependence on assumptions. The ultimate target of the analysis is to recognize which components of models have the most effect on

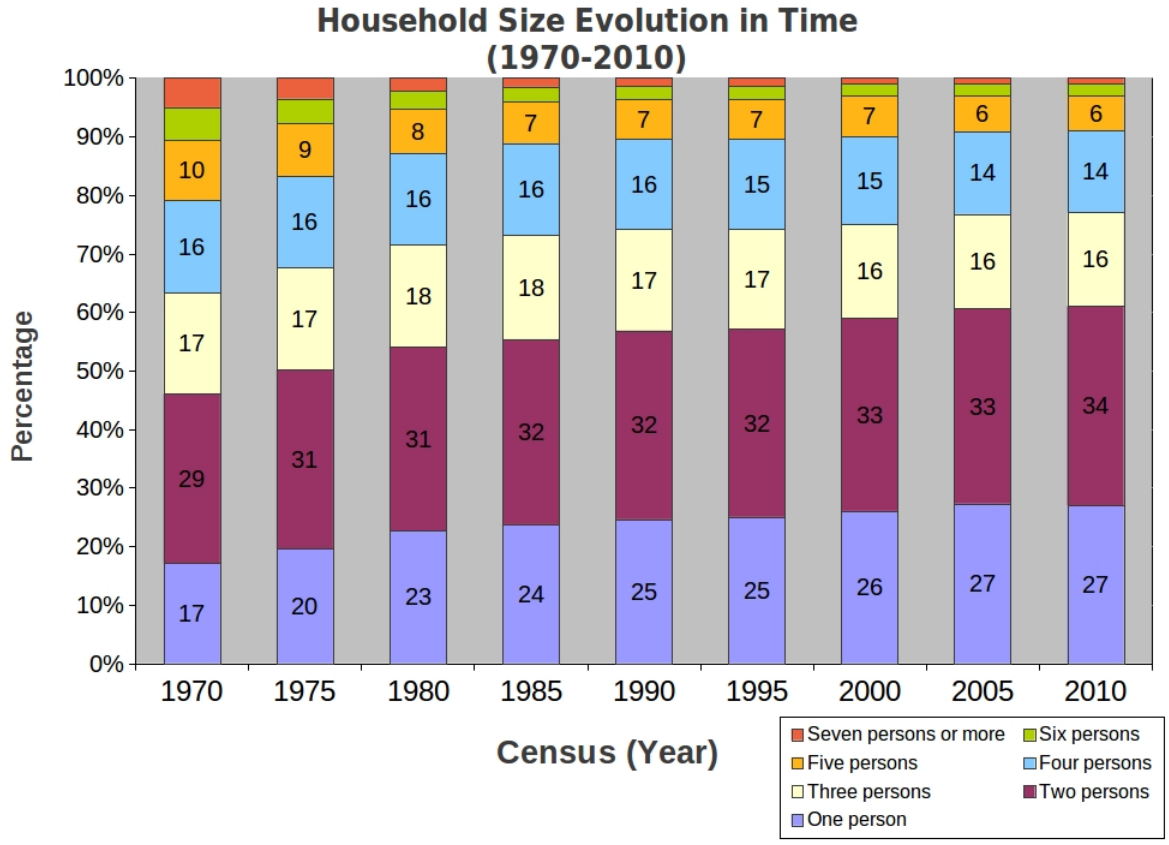


FIG. 1.3. Household size change in time 1970 - 2010, source [30]

the dynamics of the spread of disease [114].

On one hand, advancements on graph theory have incorporated a significant corpus of quantitative tools and mechanisms for describing networks that have epidemiological applications [92]. Diseases that are transmitted by person-to-person contact rely on the social contacts of the initial host to propagate. In graph theory and social network analysis (SNA), individuals are represented as nodes or vertices and connections among them are edges of a graph that depicts the social model. The nature of the connections has driven a great deal of attention because they are a key component of the construction of the network. Properties like symmetry, transitivity, and weight of the edges can be interpreted as social cohesion and are evaluated before defining what an edge really represent [92].

Determining the mixing network in full extend, requires complete knowledge of every individ-

ual in a population. In addition, complete recall of the person's relationships are needed to construct the network. These tasks seem unattainable and impractical even for small groups. To establish a workable framework, researchers initially applied known datasets from social networks based on random graphs with arbitrary degree distribution that have exactly solvable models [120], random graphs with tunable clustering [26], and stochastic process with global [54] and local contacts [10], [154], [111]. Exactly solvable models can only exist under the assumption that the vertices are homogeneous. In these models, stochastic or deterministic processes govern the creation of edges. From the theoretical point of view, homogeneity is a simplification that reduces the complexity of a problem where geography, demography, environment, and migration patterns cannot be accommodated. To address the problem of modeling realistic networks, one can gather field information by sampling a selected population and then by extrapolating the results into families of graphs calibrated to resemble the parameters found on the sample.

Most of the initial work on epidemic graphs have focused on the influence that topology of the social networks has on the dynamics of the disease. The social model is mapped into a graph or families of graphs where the interactions or contacts are represented as links. Random graph models consider each individual as member of a number of social structures or subgraphs G_i of a complete graph G, where a link exists with probability p_{G_i} [77]. Nevertheless, it is crucial to find realistic models for the social structure of the population.

Human networks and specifically contact networks have been proposed to exhibit a "strong community structure" [138]. Several algorithms have been designed to computationally generate networks with such property [40] and others like hierarchical structures [39]. Researchers have also looked at the corpora gathered by online communities and empirical networks. For instance, the "Facebook dataset" [150], a corpus that contains the friendship network of US universities on the social network website, has been used to generate the social contact networks with epidemiological applications [138]. The dataset also contains information about gender of the individual, the dormitory residence, and major. Advantages of using online communities to model face-to-face interactions include the accessibility of the

datasets, intermediate readiness for computational processing, and the general intuition that such interconnection networks are similar to real social networks, to some degree.

On the other hand, disease outbreaks studied on artificially built societies have been a milestone for epidemiology due to the realism that can be achieved through these models. Agent-based models (ABM) embrace the complexity of capturing large-scale social networks and the direct contacts of individuals in which a change in behavior can be modeled as well. ABMs massive simulation approaches, demonstrated a strong correlation between local demographic characteristics and pandemic severity. The methodology accounted for the simulation of the daily trips and activities of nearly 20 million synthetic individuals on their everyday movements, activities, and social interactions [143]. The algorithmic and structural properties of the contact network dataset produced by the massive simulation constitute a bipartite graph of people and places. Eubank et al. and Riley proposed to generate families of graphs with the same properties in near-linear time [56], [134]. More recently, Zhang et al. expanded the model to include community structures with intra-community hierarchy [168]. The magnitude of the contact networks is usually in the order of million of nodes, making the algorithms applied to the networks computationally expensive and in most cases, impractical without parallization techniques. Nevertheless, flexibility comes at a cost. To build a robust statistical portrait comparable to epidemic data, valuable for policy makers, models need to be executed thousands of times. Thereby, avoiding biases caused by a particular run of the model. In addition, limited tractability is innate to these more complex models, hence for the most part, general conclusions are elusive and difficult to draw [162].

1.3. School Affiliation Network Discovery (SAND) Algorithm

This research uses synthetic populations, a essential constituent of ABMs, to build a hierarchical affiliation network of households and schools. The hierarchical structure of school districts and the corresponding association to households depicts affiliation networks that could potentially host the propagation of infectious diseases. The characteristics of such networks are highly dependent on regional demographics and are persistent for rela-

tively long periods of time, compared to other relationships. In this sense, the affiliation of households to schools can be considered "cohesive" as families that have members attending the same school would seem more likely to share connections among themselves rather than with households outside this group. The resulting network is a weighted network for which generalized centrality algorithms are used in order to identify the most "relevant" node in the network. Generalized centrality measures are employed in order to accommodate the weights for nodes and edges. The relevance of the nodes is evaluated by the impact its removal causes on the overall outbreak, at a defined point in time. Final outbreak sizes and overall costs are compared in order to establish the most efficient mode of selection.

1.4. Contributions of the Dissertation

The most important contributions of this dissertation, can be stated as follow:

- This research proposes a characterization of the properties of co-affiliation networks constructed in the k-dimensional space using $k-partite$ graphs. Research on social networks, particularly affiliation networks, has been focused on the binary nature of relationships. By extending this concept to accommodate k non-overlapping relationships, the concept is generalized, allowing utilization of the algorithm for a k-dimensional space.
- This work proposes optimization of intervention strategies oriented to mitigate the spread of disease on the school system. Optimization of public health resources is a very active field of research. A cost optimization methodology is proposed and measured in terms of fixed and variable costs. Degree centrality of the graph and vertex cover are evaluated against the naive approach by comparing the general output of the epidemic in terms of duration, total number of infectious and maximum peak of the epidemic curve.
- Finally, this research extends the synthetic reconstruction model to address school attendance zones as intrinsic part of the generation process. The association of households and schools is analyzed. Previous algorithms have chosen a distance metric as the first and preferred way to associate households and schools (or other

gathering points). While this approximation is fast and computationally inexpensive, for some study areas, information with the actual mapping of attendance zones is available.

1.5. Dissertation Structure

This dissertation is composed by three parts: first, background on computational epidemiology, graph theory and the present challenges for the design and evaluation of intervention strategies; second, the proposal for modeling and constructing a new social network based on communities of households and third the proposal for measuring the impact, the cost, and the overall performance of intervention strategies based on the SAND algorithm. An overview of computational epidemiology is described in Chapter 2. Chapter 3 discusses the evolution and current trends on population reconstruction. Extension rules of the synthetic population construction are stated in order to accommodate school attendance zones. An example simulation is described and the standard error is calculated. Graph theoretical approaches to epidemiology are summarized and MCA generation algorithm are defined by Chapter 4, additionally analysis on connectivity of the MCA graphs and additional results are presented. Chapter 5 describes the generalization for centrality measures applied to the MCA graph and evaluation of the centrality measures (degree, closeness and betweenness) as minimization parameters. Chapter 6 evaluates the cost-efficiency og mitigation strategies based on the centrality measures and proposed two evaluation function based on different optimization objectives. Finally, conclusions and a summary for the proposed solutions are stated on Chapter 7, opinions on future opportunities and challenges for the computational design of intervention strategies are mentioned.

CHAPTER 2

BACKGROUND

> I simply wish that, in a matter which so closely concerns the well-being of the human race, no decision shall be made without all the knowledge which a little analysis and calculation can provide
>
> *Daniel Bernoulli, 1760*

This chapter gives an overview of the theoretical fields that serve as foundations for this research work. Rather than providing a tutorial on the diverse topics, the intent of this chapter is to give a framework, background, and terminology for the rest of this dissertation. Computational epidemiology is founded on the interdisciplinary collaboration of social and exact sciences. The theory applied in this dissertation comes from three main fields of science: epidemic models, agent-based simulations and graph theory (Fig. 2.1). First, approaches based on compartmental models to study disease spread are described in Section 2.1. Section 2.2 describes the methodology known as synthetic reconstruction in order to model populations and interactions. Section 2.3 describes graph theoretic models in computational epidemiology. Finally the outline of the problem in proposed in Section 2.4 and a brief summary closes the chapter.

The effect that voluntary inoculation has on the dynamics of diseases has raised controversy for centuries. During 1760, the famous mathematician Daniel Bernoulli (1700-1782) formulated the question of whether it would be beneficial for the general population to be vaccinated against smallpox. In his path to demonstrate that the benefits of voluntary vaccination outweighed the risks for the general population, he intrinsically connected mathematical models with health policy [19]. This chapter discusses mathematical models used to model disease spread as well as the new developments made to represent human contact networks. Additionally, different methods that have addressed the optimization of intervention strategies are reviewed.

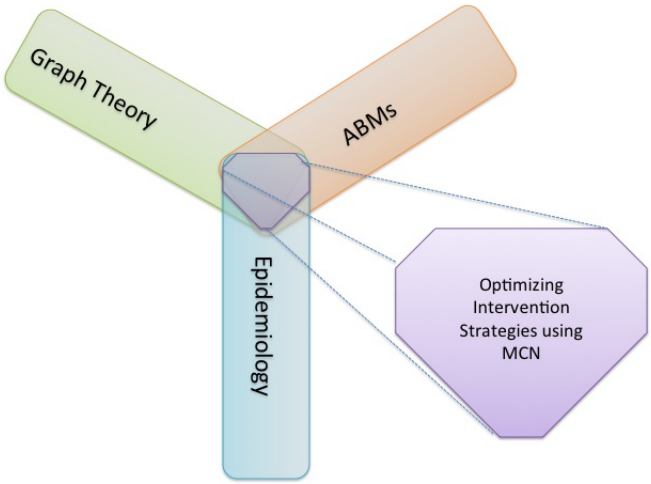


FIG. 2.1. Related fields

2.1. Models of the Spread of Disease

Understanding how diseases spread through the populations is complicated due to all the biology systems involved at the individual and social group level. In practice, approaches are based on simplified models that provide insights and guide the understanding of disease behavior under controlled conditions. Contact networks drive the spread of diseases that are communicated through the air, by touch or intercourse.

2.1.1. The SIR Model

To construct a mathematical representation, disease dynamics are reduced to changes between disease states. Individuals are assigned states and the changes between them are monitored following a timeline. The SI model stands for the simplest version in which just two states exist *susceptible* (S) and *infected* (I). A person in the susceptible state is someone who does not have the disease and is labeled as "S." Susceptible individuals can catch the disease if in contact with infected (I) individuals. The SI model can not be used for diseases that confer lifelong immunity such measles and chicken pox [5], [156], [10]. To study such diseases, mathematical models applied to explain and to predict outbreaks are based on the interaction principles between groups of susceptible (S), infective (I), and

recovered/removed (R) individuals [5]. The three possible stages: susceptible, infected and recovered are represented as an array of the differential equations shown in (1).

$$\begin{aligned} \frac{dS}{dt} &= bN - \beta IS - dS \\ \frac{dI}{dt} &= \beta IS - \gamma I - dI \\ \frac{dR}{dt} &= \gamma I - dR \end{aligned} \tag{1}$$

Birth rate b and a death rate d are considered in the model as well. Disease specific parameters in the model are: γ that represents the rate of recovery and β the force of infection, or in other words the rate at which one susceptible individuals becomes infected. The interactions among individuals are implicitly accounted for in this model. S, I, and R stand for the number of people in each state which is equal to N, the population size (2)

$$S + I + R = N \tag{2}$$

Equation (1) can be expressed in terms of fractions s, i, r and assuming a close system where birth and death rate are not considered. The new set of equations is presented next:

$$\begin{aligned} \frac{ds}{dt} &= -\beta si \\ \frac{di}{dt} &= \beta si - \gamma i \\ \frac{dr}{dt} &= \gamma i \end{aligned} \tag{3}$$

$$s + i + r = 1 \tag{4}$$

From (3) the mean time of infection can be derived. The variable τ stands for the length of time an individual is likely to remain at stage I before moving to stage R and $\delta\tau$ represents any time interval. Given γ, the probability p of recovering in any time interval $\delta\tau$ is shown in (5).

(5) $$p(\delta\tau) = gamma\delta\tau$$

Conversely the probability of remaining in stage I is $1 - \gamma\delta\tau$. Equation (6) expresses the probability that an individual stays at stage I after a total time τ.

(6) $$\lim_{\delta\tau\to 0}(1-\gamma\delta\tau)^{\frac{\tau}{\delta\tau}} = e^{-\gamma\tau}$$

The probability $p(\tau)d\tau$ that the individual remains at stage I and then recovers in the interval between τ and $\tau + d\tau$ is (6) multiplied by $\gamma d\tau$:

(7) $$p(\tau)d\tau = \gamma e^{-\gamma\tau}d\tau$$

The basic reproduction number denoted as R_0 is defined as the average number of additional people that an infectious person passes the disease onto in a complete susceptible population. R_0 is an important quantity in epidemics and can be derived from (3). If the infectious period is denoted by τ then the expected number of contacts during that time is $\beta\tau$, and the average number R_0 from (7):

(8) $$R_0 = \beta\tau\int_0^\infty \gamma e^{\gamma\tau}d\tau = \frac{\beta}{\gamma}$$

[121]

Then one can conclude that the epidemic transition or transition from an epidemic to a non-epidemic event happens when $\beta = \gamma$. Therefore, when $R_0 < 0$ the incident is considered a non-epidemic event, $R_0 > 1$ is considered an epidemic and $R_0 = 1$ is defined as an endemic.

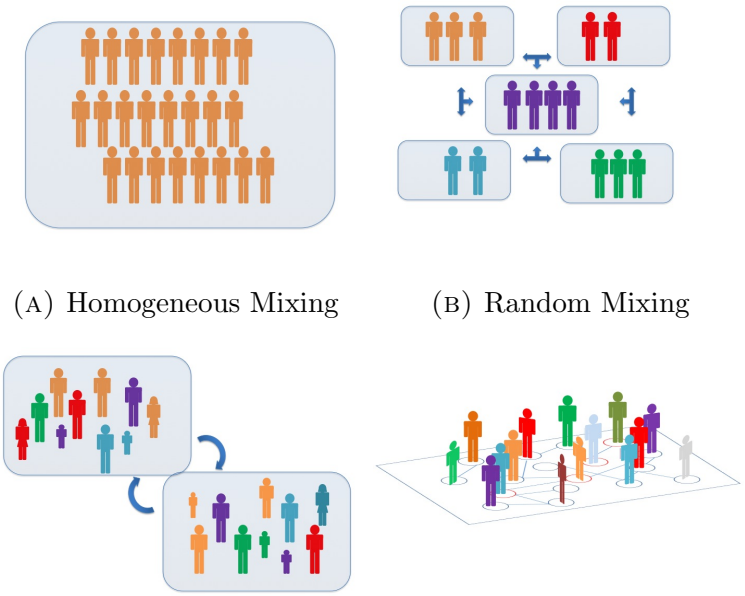
2.1.2. Homogeneous Mixing

Homogeneous mixing is probably the strongest and most questionable assumption of the model. As initially stated in (1), individuals mix at a fix rate and become infected under a general parameter β. Demographic and geographic details such family size, geographic location, incomes are omitted and all social processes are averaged out to represent the epidemiological process only in terms of rate of susceptibles, infectives, and recovered. Recent studies consider homogeneous mixing as a starting point that has given rise to several theoretic studies that highlight the relevance of disease disparities relevant to age, gender, contact rates, geographical placement, etc.

2.1.3. Random Mixing

Transmission matrices are tables that represent mixing patterns among different groups. The groups can be based on demographic characteristics as age, gender, sexual orientation, etc. In random mixing models the number of contacts effective per unit of time is continually changing as opposed to models based on networks for which the number of contacts is fixed. The basic model has been extended to accommodate heterogeneous populations, composed by subgroups with different mixing rates among themselves. These subgroups may represent demographic characteristics such age [78],[79], gender or partnership [2], [164], global and local contacts [154], [111]. Other studies analyzed sexually transmitted diseases (STD) and transmission on networks [14], [97]. Fig. 2.2b shows the population divided into five groups with different mixing patters.

The role of non-homogeneous mixing in population with geographical and social structure has provided the stage for the rise of metapopulation models where "homogeneous mixing holds within local contexts, and that these contexts are embedded in a nested hierarchy of successively larger domains" [162], models characterized by heterogeneous connectivity and mobility patterns [45] and urban networks [55]. Fig. 2.2c exemplifies a model that considers two populations. Finally, contact networks estimate mixing patterns through the construction of social networks where the contacts are analyzed based on mobility and simulated interactions [136]. Fig.2.2d presents a schema of contact networks.



(A) Homogeneous Mixing (B) Random Mixing

(C) Meta-population (D) Contact Networks

FIG. 2.2. Schematized mixing patterns used in computational and mathematical models

2.2. ABMs - Synthetic Reconstruction

As computational power grows, models have shifted from a simplification of real processes that emphasize first principles or "strategic models" to models that represent real situation as close as possible or "tactical models"[43]. Artificially built societies or agent-based micro-simulation models are snapshots of the entire population of a selected study area. Synthetic populations initially were created to estimate traffic needs and development [16] but were quickly adopted by epidemiology researches to demonstrate a "strong correlation between local demographic characteristics and pandemic severity" [143]. At the most general level, the methodology involves two steps: taking data from the census and adjusting multi-way tables with joint totals to their up-to-date equivalents and selecting households to fit those totals, optionally assigning them to geographic areas [116] and multi-level controls among household and person-levels [7]. Other approaches have used random graph models and link probability models [108]. For validation, a sample from the generated households

is taken and compared with the original values of the joint totals.

The primary tool used to complete the multi-way table for each census tract is iterative proportional fitting (IPF). The next step is to to use the probabilities obtained by the IPF technique to select households from the sample in order to construct the synthetic population. The number of households of each type used in the population is determined by each census track.

In addition to the synthesizer, it is important to select a physical place to bind the generated population to a geographical location. Spatial patters may reveal high risk communities, problem areas, or even possible causality that can overlooked using other approaches[100].Complex spatial simulations require a variety of modules, for example Epigrass [43] is a tool used in the construction, simulation and analysis of disease-spread scenarios using standard GIS file formats and requires the interconnection of multiple databases and application tools.The methodology has been applied to diverse sources of information for different geographies: the city of Portland, Oregon [56], New York [107], 50 states and the District of Columbia [166], Switzerland [67], Belgium [32], Singapore [168]. Comparisons among methodologies has also been done by [116], and [146]. The use of meta population models and virtual populations has been extensive during the past years for the exploration on theoretical models applied to epidemiology [102] as well as risk assessment for different communicable [44], non-communicable [102] diseases, and computational frameworks for data modeling over large scale networks [59].

Human mobility networks are an intrinsic part of agent-based models. Meta-population models assume stochastic movements of agents that can be either local or global. A step further for these models is the discovery of individual mobility networks, made possible through the collection of mobility data [130] and census information [8]. It has been proposed that when individual mobility is included in the construction of the meta-population model, if the flux of individuals between populations is sufficiently high, outbreaks are not only unavoidable but global [17]. While countless applications for meta-population models have been suggested, this research focuses on the contributions made to capture the dynamic behavior

between families and schools, and the social interactions among school-age children. Using standard concepts and techniques from physical systems acquaintances in a social environment can be reproduced. Gonzales et al. proposed a system based on colliding particles to reproduce intra-school interactions and friendships. Such system comprehends N particles (or individuals) of radius r, moving continuously in a bi-dimensional space shaped as a square of size L. Collisions of the particles are interpreted as interactions, allowing this model to reproduce empirical data from school friendship networks [74] and reported sexual interaction networks [75] with a higher accuracy than previous models.

2.3. Graph Theory and Social Networks

A social network is a diagram that describes a social structure of units or "actors" and their relationships through interactions. "Actors and their actions are viewed as interdependent rather than independent" and the relationship linkages between actors are channels that allow the transfer of resources or "flow" [160]. Network model's structure is considered to be the lasting patterns that are meaningful in some context and subject to study. Different computational frameworks have been proposed to model social networks [47]. Even though the most disease models use compartmentalization of individuals following their disease status [92], other modifications center on biological observations have been accommodated in this basic framework to address heterogeneous mixing, core groups, communities [138] and both unclustered [133], clustered networks [145] and tunable clustered networks [26]. The graphs representing real-world phenomena are generally large, sparse and complex [9]. Spatial location and spatial dependence are important on social and contact networks since topology alone may not contain all the information relevant to the model [95]. How spatial constraints may have an effect on the structure and properties of networks and disease dynamics has been studied through empirical observations [12]. Spatial constraints can also be recognized on the diffusion processes that take place on secluded spaces [41].

Research has also been focused on multi-layered random graphs, [99] proposed an algorithm for modeling the World Wide Web (WWW) graph as an ensemble of slices generated by independent stochastic processes. Another type of networks are those that allow

multiple type of edges or multiplex networks[27].

Moreover, the study of networks has been developed from two main knowledge fields: social sciences and graph theory. In recent years SNA has received the attention from behavioral and social science disciplines [159], in particular epidemiology [114] because SNA studies regular patterns of relations connecting a set of entities built into a macro-social context. SNA is based mainly on the quantitative mapping of networks and aims to measure their formal properties [81]. Network analysis measure and represent structural relations among entities and attempts to explain both why they occur and what their consequences are [81]. Although the major criticism has been that SNA is "merely descriptive" as opposed to being a theoretical tool, social phenomena in a variety of disciplines has been explained in terms of this field [25]. Some events, like a disease outbreaks, occur due to the existence of a wider network that individuals are not well aware of. In addition, social networks are not static entities but they have been found to change and evolve in time. The Dynamic Social Network Analysis (DSNA) is the study of the evolution of social networks. DSNA has been used to identify facilitators on illegal social networks [90], cohesive reading groups [66], biological networks [70], and the spread of obesity[37]. Observation of this behavior for a period of time may also produce changes on the structure of the network and the community.

Communities in the environment of social networks and graph theory are an active research field. Researchers have studied the impact that a change on the structure of the social graph would have on the dynamics of the disease. Results from [157] suggest that "the key risk factors for infection should be used to define a population structure". The community structure present in dynamic networks has also been studied. Dynamic networks can maintain the stable community structure that has been observed in many social and biological systems [28].

Clustering coefficient of a graph is defined as a the probability that two neighbors of a randomly selected node are neighbors as well. It can be interpreted as a measure of connectivity. In homogeneous networks with a given degree distribution and average transmissibility, clustering is a dominant factor for controlling growth rate of an epidemic [113].

Another theoretical tool in studying disease dynamics comes from control theory. Control theory provides mathematical tools for modifying the dynamics of engineered and natural systems towards a desired state. Models that describe the interactions that take place on social networks give rise to complex network structures that have proven extremely difficult to understand. Social processes can be described as dynamic systems, where key properties determining the system outcome [142]. Scale-free degree distributions, frequently used for modeling social processes, demonstrated to have better controllability properties than uncorrelated networks [118].

Complex self-organized systems represent a completely different challenge. On arbitrary networks, identifying the minimum number of driver nodes, can guide the system's dynamics. Yet driver nodes tend to avoid high-degree nodes [106]. This result can be translated to the study of disease dynamics. Although much work has tried to demonstrate the correlation between "relevance of a node in the network" and it's degree, new definitions of "relevance" or centrality have appeared to focus not only on degree buy on different measures to try to capture the set of nodes that have a grater impact on the final state of a network.

2.3.1. Modeling Contacts

Online social networks like Facebook and LiveJournal provide a corpus of relationships that have been data-mined to extrapolate the dynamics of how real-life individual contacts are created, evolve and disappear in time. Social networking sites seem to be ubiquitous among the U.S. and everywhere where internet is available. Such potential has been used to draw conclusions on how relationships among individuals in the population evolve in time. The Facebook dataset [150] gave rise to epidemiological studies that made use of the community structures discovered in the dataset to design intervention strategies.

Friendship relations created on Facebook and "pokes" (messages to attract attention of another user) were studied to devise contact patters and behavior at large scale. School affiliation was found to be correlated with online friendship providing a partial geographic boundary [73] for online relationships. LiveJournal dataset has been data-mined

in the search for methodologies to accurately predict and classify friendships. Self-reported attributes (e.g., interests, communities) have been shown to be a reliable predictor for the existence of edges representing relationships on such graphs [84]. Social networks have also been used to study relationships that are neither fraternal nor friendly, but rather reflect the process of influence. Given a network, each individual or agent formulates an opinion and revises it gradually based on the surrounding environment. Therefore, a division of real social networks into groups or communities of individuals with similar opinions is formed. Whether individuals evolve to be alike minded or they develop more network connections because they are like-minded is object of research. In an attempt to solve this question physics-based models, namely assortative mixing or homophily have been used. After simulating the effect of both processes alone and together it has been suggested that opinion formation is indeed a combination of the two [83]. Regarding groups' conflicting interests, it has been suggested that uninformed individuals have an important role achieving democratic consensus [48].

The Ising model [20] examines the majority rule dynamics, where one agent may change its state or "opinion" according to the majority opinion of the neighbors. It has been shown that the difference on polarization between two loosely connected Barabasi-Albert networks depends on density of internetwork links [144]. A similar phenomenon has been observed on clustered epidemic networks, where once the threshold of inter-cluster connectivity has been reached the system behaves as a single network [45]. It can be hypothesized that the environmental triggers like events, or opinions in the proximity of an agent have a measurable influence on its behavior. To test this hypothesis in humans, technology-based approaches have been taken by surrounding test subjects with video cameras and sensors. The analysis of visual attention and the probability of pedestrian adoption a certain behavior has been tracked in urban scenarios when confronted with weak stimuli. The results revealed that visual interactions among pedestrians occur within a two-meter range. Additionally spatial features, social context, and sex of the stimulus affect the overall tendency

to respond [68].

Modeling peer-to-peer relations between individuals helps to understand the role that spatial heterogeneity has on disease dynamics without the need for large-scale computer simulations. When the network is considered to be homogeneous, then its structure can be mostly characterized in terms of the average number of neighbors and their interconnections [91]. Random networks of interacting agents can then be modeled in a simpler fashion to study dynamic behavior and cascades. Cascades are shocks that traverse the network on a time-step fashion. They have been used to model spatial diffusion of diseases on geographically constraint area. Watts [161] proposed an experiment, in which the action of the agents was only determined by the actions of their neighbors. Several simulations determined that on highly connected networks cascade propagation is limited by local stability.

2.3.2. Communities in Networks

While the definition for *communities* is domain dependent, one can take the notion of a collection of similar entities that interact with unusual frequency. Nevertheless, representation of human relationships is challenging because relationships change over time, people appear and disappear and there are ever changing cycles occurring. By exploring network communities one can:

- Define the network organization in order to be able to reproduce or modify it [98].
- Understand dynamic processes that may be affected by the modular structure of the graph.
- Uncover relationships between the nodes that are not evident by inspection.

The first step to study the effect that community structure has on the diffusion process, whether one focuses on diseases or information, is to obtain or generate the network. Blogs and online communities have been subjects of study due to the availability of information already in electronic format. It is this characteristic that makes online communities a focus of interest for epidemiology [148]. If communities are identified in complex networks, then it is possible to leverage this property to modify the dynamics of the network. The pur-

pose of designing algorithms to generate networks with community structure is to facilitate the study of real-world complex social structures. Some algorithms create static networks; others focus on network models that evolve having a community structure. Notably, preferential attachment mechanisms both inner and intra-community would yield the construction of an evolving network with community structure [104].

Once the information relevant to the network is at hand, either by generation, by data-mining processes or by other method, the next issue to concentrate on is the identification of communities. A wealth of algorithms for identifying communities on a graph have been proposed. The algorithm used to detect community structure in a network is hierarchical clustering, but others methods based on betweenness centrality [70], combinatorial optimization [158], and spectral clustering [132] have been developed in the last few years. Girvan et al. [70] proposed a method for detecting communities, using edge betweenness. Edge betweenness is defined as the number of shortest paths from any pair of vertices that include a particular edge. The community bridge finder (CBF) [138] is an algorithm that attempts to mitigate the spread of disease by identifying communities' bridges (or nodes that connect multiple group of clustered nodes). Nevertheless, for large networks most community finding algorithms make heavy demands on computational resources, running in $O(m^2n)$ time for an arbitrary network of n nodes and m edges, or $O(n^3)$ time on a sparse graph. A modification of the algorithm that runs in $O(n \log^2 n)$ was proposed by [40] for hierarchical-structured networks, making the observation communities may be found at many scales. Later this approach was extended as a "general technique for inferring hierarchical structure from network data and the existence of hierarchy" [39]. A comparative analysis of community identification algorithms is presented by [98].

The identification of community bridges in networks yields several strategic opportunities. Under the assumption of homogeneity, people who act as intermediaries or bridges between distinct groups may have access not only to more diverse information but to positional advantage [94]. Furthermore, if the network is one that describes disease spreading,

bridges may represent the opportunity for its expansion, making bridge identification an interesting endeavor to accomplish when planning intervention strategies.

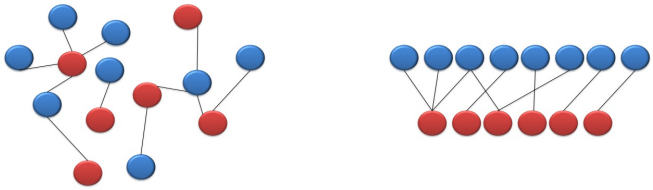
Online communities also provide helpful insights for early detection. Monitoring health-seeking behavior as online queries has been used to predict ILI percentages during influenza season [69], when large populations of web search users inhabit the study area. Nevertheless the challenge is even greater for emerging diseases, where information may not be available or even exist. Using visual pattern-recognition, epidemiological surveillance has moved into the technological era by using well trained "epidemiology watchers." The goal is to develop unsupervised methods to monitor and rise alert when unexpected events happen [103].

Identifying at-risk communities is a key aspect when developing containment strategies. During 1997-1998, the state of California implemented a strategy designed to contain syphilis cases coming from Mexico. Prevention not only played a key factor but also averted reestablishment of ongoing transmission. Containment procedures were implemented once clusters were identified. In addition, surveillance caused near-elimination of syphilis in the area [76].

2.3.3. Affiliation Networks

Affiliation networks represent binary relations between members of two sets of items. [22] On this context, the term affiliation is used to acknowledge participation or membership of people in events, projects, or groups. Affiliation networks are traditionally represented as bipartite graphs [121]. Complex evolving networks can be studied using the results and methodologies developed by using affiliation networks [11]. A result that can be derived from of affiliation networks are co-affiliation networks. Co-affiliation is defined as a tie between two members that belong to the same set of nodes and have in common a member of another set of nodes. There are several techniques to measure and normalize the weight of co-affiliation, depending on whether they represent opportunity or are taken as indicators [22].

It has been stated that affiliation networks tend to model the social structures more exactly than simple networks [120]. Food-web networks represented as graphs, have a de-



(A) Simple Graph (B) Bipartite Graph

FIG. 2.3. Graph comparison

gree distributions that do not display a universal functional form, although their topology is consistent with patterns resembling small-world and scale-free graphs [51]. For complex ecosystems, bipartite networks represent structures that allow depicting mutualistic interactions accurately. Mutualistic networks are the combination of two classes of networks biologically different. To characterize these networks is compulsory to describe the degree distribution. It has been found that by maximum entropy could potentially be used as a model for the degree distributions in bipartite ecological networks [167].

Affiliation networks and co-affiliation networks are depicted as simple graphs where the existence of an edge is representative of an ongoing relation among nodes. Other networks represent edges and nodes including a value or weight that can be contextualized according to the problem being schematized. Weighted graphs have been used to represent individual popularity within groups of students [85]. Students become nodes of the network and edge weights represent the depth of favor or disfavor between them. The evolution of friendships can then be analyzed through the evolution of the system for which the change of weights is guided by random encounters. In this model, the final edge weights are mostly influenced by "the first impression" (initial weight assigned to the relationship when setting up the initial adjacency matrix) between any two nodes [85].

2.4. Optimization of Intervention Strategies

Measuring and targeting intervention strategies is a highly complex problem. Research on novel infections showed that targeting intervention measures is more effective if

the group with the highest risk can be identified [157]. Disease control measures comprehend vaccine, antiviral treatment, prophylaxis, and non-pharmaceutical interventions (NPI) such quarantine, isolation, school closure, and social distancing [78]. Timing is a key factor for interventions, scheduled vaccination is a part of almost any nation's public health policy while social distancing occur once cases of a particular disease have been identified.

Intervention strategies and their optimization depend on finding the right people to target. It has been hypothesized that in the social network context, disease traverses through targets that belong to shortest paths of the network in order to propagate. Finding the shortest paths is directly related to the type of network. To identify shortest paths, search algorithms benefit form local information about the target more than the naive approach based on high degree [1]. While some of the network characteristics may always remain unknown for the modeler, information of the network structure in combination with real-life heuristics represents the best actual knowledge when designing disease mitigation strategies. In reality, when it comes to evaluate risk of real-life scenarios, primary schools constitute an important risk group [46] even when their social patterns may remain unknown.

For many infections diseases, vaccinations are the most effective means of control and their objective is to immunize a sufficiently high proportion of the individuals in the community to prevent epidemics [15]. After a transversal comparison of several interventions for influenza, it was discovered that early detection and initiation of measures and school closure play important roles in reducing influenza transmission [78]. Through the identification of communities and community bridges in the social model, targeted immunizations could be more effective [138]. This concept has also been studied at the granularity of households and transmission within and between households the household for isolation, quarantine, vaccination or prophylactic treatment [65]. At the household level, social distancing limits the transmission that happens between households.

Given a social network represented as a graph, finding a set of relevant nodes is an optimization problem. Depending on the type of network and the problem at hand the opti-

mization could potentially be formulated as a minimization or maximization. Maximization of influence is an NP-hard problem but several approximations have been proposed [93]. Domingos and Richarson [50] proposed modeling markets as social networks where spread of influence is maximized. The overall objective was to asses the net value of customers in the network through data-mining. In most recent studies, Even-Dar and Shapira [57] used the same methodology to study how to identify the most influential individuals to maximize the adoption of a new technology using the voter model. They found that as a special case, the naive solution of targeting the nodes with highest degree first yield the optimal solution. Intervention strategies depend on the social contact networks and the how people interact within them. Through surveys, the dynamic behavior of children and teenagers has been model to try to quantify local transmission of influenza. Although students, their groups an activities in public places represent a great challenge to model their importance regarding the next pandemic is utterly considerable. It is believed that high-schools may play a potential backbone for disease transmission in the next pandemic [71]. Early detection and containment are the most desirable scenarios for well known as well of emergent diseases. Disease spread in rural areas has different dynamics due to population density and transportation than cities. It has been suggested that a highly effective strategy for rural areas is to target vaccination campaigns at popular location is more effective that random vaccination at the same places [140].

2.4.1. Discovering Relevance

Developing a feasible method to identify the most important set of nodes in a network is an open question. The evaluation of relevance in a network is in direct connection with how the network is constructed. Only on complete networks all nodes are connected with their peers, otherwise either a random or well defined criterion is used to create links among nodes. The concept of proximity considers the characteristics of the individuals pairwise in order to create a link among them. A widely used methodology for real-life network data is to define a measure of proximity also called "social distance." The existence of links between two individuals is both cause and consequence of similarities shared in certain respects of

their intrinsic characteristics or personal choices. Although there exist different basis on how to define social distance, loyalty and affiliation are considered appropriate measures for the existence of links and are based on the overlap of interpersonal environments [3].

When the underlying network is unknown, identifying well-connected, central nodes has become a research question for different domains. Spreading of disease, emergence of fads or even diffusion of memes have been proven to shown a similar behavior as they all make use of a social structure to disseminate. For example, the size of the fad is in direct proportion to the degree of the first adopter, regardless of the underlying network [6]. Likewise, the degree of an index case or initial patient of a population is positively correlated with the probability of an outbreak [88].

Centrality and how to measure it, is one of the most researched concepts in SNA particularly for continuously growing networks. Several measures to determine the most central nodes of a network have been proposed and developed, including degree, closeness, information and betweenness centrality. In order to understand centrality, one must first define the flow of a network. Flow can be conceptualized as the paths that exist among nodes. In general, centrality measures can be regarded as measures that generate expected values for some sets of nodes given implicit models of how traffic flows in the network [23]. Centrality measures have been used to study the structure of online communities to identify hate groups [34], in combination with game theory, to find optimization of bounded budgets [35], and to calculate risk of infection [38], [88] among others. Centrality measures have also demonstrated a high correlation with self-assessed relationship strength among HIV-positive patients and their role in the HIV diffusion network [135].

Betweeness centrality has been used in the study of ego networks. Ego networks consist of actors or "egos" as a focal point and the nodes that egos are connected to, referred as "alters." Betweeness centrality is used as a measure of the density of the ego network as well as the centrality of each actor in the network [58].

Information centrality is a similar to betweenness centrality in that is a network connectivity metric as well. This method accounts for indirect as well as shortest or geodesic

paths among actors considering their degrees. Information centrality of the complex social networks of long-tailed manakin (Chiroxiphia linearis) has been used to calculate the odds of males that will socially rise to be alpha-males [109].

2.4.2. Optimal Mitigation Strategies

Finding optimal vaccination strategies has been a very prolific field of research during the last decade. Vaccination is an important disease control method not only for human diseases but also in natural resources management. Optimization can be defined in two ways: the social cost of disease and the monetary cost related to vaccination programs. Regardless of the disease model used or the disease itself, the optimization problem is defined as either finding the strategy with the minimal cost provided an public health aim or given a restriction in the cost find the most effective strategy [115].

Disease modeling is the first challenge one faces in the search for optimal vaccination strategies. The mathematical or computational representation of the disease plays the most important role in the optimization process along with the parameters used by the model. In the most general form, the optimization function is a minimization of the number of vaccines, with the additional component of the structure of the population.

However uncertainty of conditions, disease or population parameters have also been studied. Under the stochastic programming framework, [147] presented the optimal vaccination policy for disease epidemics with parameter uncertainty. The objective function is defined in terms of vaccine coverage, two additional restriction functions balance the proportions of the vaccination among diverse population and bring the reproduction number below zero.

Voluntary vaccination presents the general population with a decision-making challenge, particularly for parents. Two main risks should be considered: the probability of becoming infected and morbidity from vaccination. When the portion of the population that is immune prevents the spread of disease and gives a protection level for the individuals unvaccinated then herd immunity has been reached. If a population reaches herd immunity then the risk associated with vaccination outweighs the risk of infection. Under these

assumptions, the decision of getting vaccinated can be modeled from the game theoretic perspective. If the perceived risk associated with vaccination increases, vaccine uptake declines. In addition, if a vaccination scare happens, after it has ended and perception of risk is reduced, levels of coverage are more difficult to restore [13]. Indeed, for some initial conditions of the system, minimizing the prevalence of a disease for the entire population disagrees with the individual evaluation of the risk [117]. When considering vaccination strategies, it is important to include not only high-risk groups of individuals but those groups that are likely to provide paths for disease spread. Under certain conditions increasing vaccination resources to high-activity groups instead of high-risk increases herd immunity for the entire population [149].

2.5. Summary

Computational epidemiology can be considered a "young" field of science where researchers are faced with the challenge of handling interdisciplinary knowledge. Mathematical and computational models in combination with physic and modern biology theories are combined to give rise to modern epidemiology. Diseases are ubiquitous in the animal kingdom and while their understanding and consequent control could potentially represent a wealth of improvements for the quality of life, this endeavor has proven to be one of the most complex problems. If the problem was completely stripped out of the social and economical complications, the interaction between the biological systems involved at the the macroscopic level such communities and contacts, intra-person level, such the immune system, and the microscopic level like to the biology of pathogen itself, do not reduce the complexity of the problem. For this reason, models tend to take a simplification of the reality in order to better understand the effects that small changes in a particular set of parameters have in the system behavior. Compartmental models simplify disease dynamics allowing only a finite set of states. In the model, the particular stages of the disease become states and differential equations for which the assumption of homogeneous or heterogeneous mixing drive the outcome and the general result. Models that come from computational expensive methodologies such ABMs model populations as close to reality as possible, whereas theoretic approaches

provide robust solutions to simplified models. Finally, to optimize mitigation methodologies, researchers look into both methodologies to identify what constitute high-risk targets bounded by the structure of the social model.

CHAPTER 3

POPULATION RECONSTRUCTION

> Maps answer the question: where? They can reveal spatial patterns not previously recognized or suspected from the examination of a table of statistics
>
> *Lawson, A., 2001*

In this chapter the concept of synthetic population is introduced as the methodology for the recreation of the population of Denton County. Three questions are addressed: (1) how to reconstruct the records of an entire population, for which demographic characteristics and exact locations of people and households are available, (2) how to build the affiliation network among households and schools, and (3) the relevance that affiliation networks have when designing intervention strategies. The last question will help to determine the relevance of this research in the field of computational epidemiology (Chapter 6), whereas the answer of the first two questions will define the methodology proposed by this research on how to construct multi co-affiliation networks (Chapter 4 and Chapter 5). In the next section, the infrastructure and methodology used for synthetic reconstruction is described and the sources of information are listed. Section 3.2 defines the probabilistic function used in the selection of households from the general sample and exemplifies the general procedure using multi-control levels. Section 3.3 describes the quantitative evaluation of the synthetic population compared with the initial information and forecast year. The chapter is closed in Section 3.4 with a discussion on the methodology presented and compared with the most dominant developments in the field of Agent-Based Models (ABMs).

3.1. Experimental Infrastructure and Methodology

In order to predict the future state of a system, micro-simulation models are applied to emulate the behavior and movement of people and their actions in their physical environment through agents. Initially, the model requires the creation of a set of agents and their relationships, hereafter called "synthetic population." Additionally, the process includes the creation of establishments that associates agents together such as households, schools, work-

ing or shopping places, and other gathering locations where interactions occur. The base methodology was proposed by Beckman et al. [16] as a technique for the reconstruction of populations to estimate traffic demand. While the granularity of the model allows a wide range of semantic meanings for agent (i.e. vehicles [16], individuals [7], [131]) this research has focused on the creation of synthetic populations composed by persons, households, and schools. Synthetic reconstruction is used in this research as a methodology to simulate interactions that take place in the school system. The main challenge to be addressed is how to design intervention strategies oriented to pinpoint critical locations in the school system. In order to approach the problem, a three-layer methodology was proposed composed by: en epidemic simulation layer, a reformulated affiliation networks layer, and a synthetic reconstruction layer. Fig. 3.1 presents an overview of the structure of model.

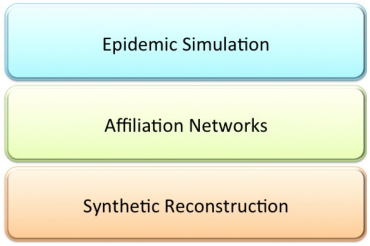


FIG. 3.1. The model has three different layers: epidemic simulation, affiliation networks, and synthetic reconstruction.The epidemic simulation contains the whole model and all the objects representing the synthetic population and the affiliation networks.

3.1.1. Population Generation Architecture

ABMs use detailed information describing characteristics and behavior of agents. On one hand, due to privacy and feasibility issues, such detailed records pertaining to individuals and housing are not available for public use. On the other hand, aggregated information of population demographic descriptors is almost ubiquitous among nations and geographies. The synthetic generation algorithm requires both types of information, detailed individual

records and general counts. Demand for synthetic populations emerges from the necessity of having the micro-information (also called micro-data) about agents, their characteristics, and interactions at an atomic level. The synthetic reconstruction methodology is essentially composed by:

- Input: datasets from the information available through public databases.
- Generation process: synthetic reconstruction.
- Output: synthetic population distributed over a geographical space with the corresponding schools.

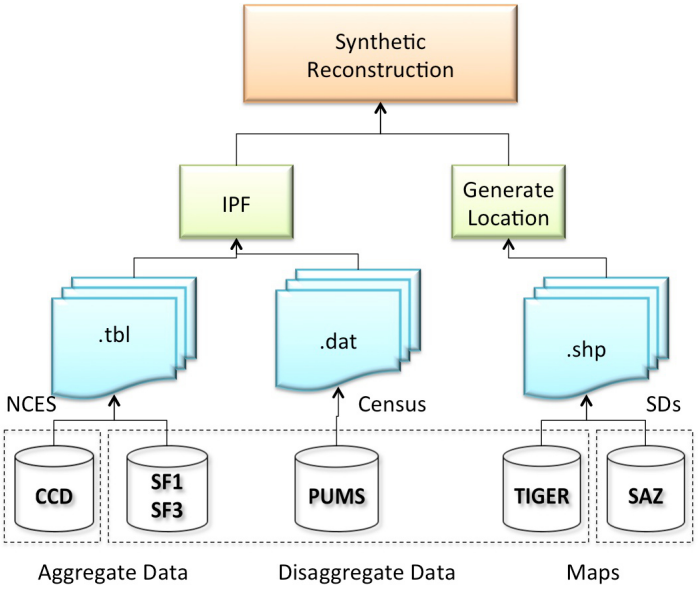


FIG. 3.2. Architecture of the synthetic population generator POPSYN

Fig. 3.2 depicts the architecture used by the population synthesizer (POPSYN). The major components are the input database, the Iterative Proportional Fitting (IPF) engine, and the synthetic reconstruction module.

3.1.2. Input Database

Collecting the input database is a data intensive exercise because of the size and diversity of the information required. Two main types of datasets are needed: counts (i.e. people,

households, schools) and maps to locate these objects. The main sources of information in order to perform a synthetic reconstruction are:

- US Census Bureau Website (`http://www.census.gov/`)
  (1) Summary Files (SFs) 1,2,3,4
  (2) Public Use Micro data Sample (PUMS)
- National Center for Education Statistics (NCES) (`http://nces.ed.gov/`)
- School District (SD) Review Program (`http://www.census.gov/geo/www/schdist/sch_dist.html`)

Table 3.1 is an example of an aggregate dataset representing the counts of people according to gender and age that can be found in the SF1 and SF3. Table 3.2 exemplifies information found on an individual record of the PUMS.

3.2. Synthetic Reconstruction

3.2.1. Creating Synthetic Households and Household Members

Synthetic reconstruction is a process that creates data records describing socio-demographic features of households (HH) and household's members (P) residing in the area of study. The generation process requires an *aggregate dataset* that contains the marginal distribution of the variables estimated for the year of study and a *disaggregate dataset* or a sample of records with complete information about Ps and HHs in the population. The aggregate datasets are a set of cross tabulations that describe one, two, or multiple way counts of socio-demographic attributes or control variables, at different spatial resolutions i.e. the smallest component for all census geography is the block [31].

The spatial units called target areas are units for which the aggregate distribution information is available. Disaggregate distribution information is available for areas (seed areas) typically larger than the target areas. Once the aggregate and disaggregate datasets are obtained, the next step is to generate population records, which is dobe by selecting sample records from the disaggregate dataset and matching the aggregate dataset's marginal distribution. The process outputs data records with the selected demographic descriptors

TABLE 3.1. Two-way table corresponding to gender and age for two census tracks of Denton County, Texas [30]

Denton County, Texas				
	Census Tract 201.03		Census Tract 208	
Age Range	Male	Female	Male	Female
Under 5 years	261	415	61	105
5 to 9 years	500	299	0	38
10 to 14 years	527	285	90	32
15 to 17 years	173	239	165	0
18 and 19 years	51	111	81	338
20 years	73	48	201	349
21 years	34	48	454	509
22 to 24 years	140	71	489	375
25 to 29 years	361	240	146	175
30 to 34 years	426	385	86	60
35 to 39 years	401	236	28	15
40 to 44 years	352	429	83	126
45 to 49 years	410	357	65	17
50 to 54 years	293	321	172	214
55 to 59 years	158	287	18	95
60 and 61 years	87	121	33	70
62 to 64 years	115	104	7	17
65 and 66 years	85	91	18	0
67 to 69 years	133	126	59	43
70 to 74 years	80	132	19	39
75 to 79 years	106	206	0	7
80 to 84 years	123	32	12	12
85 years and over	29	58	20	10

TABLE 3.2. Micro data sample depicting housing and personal records

(9) $\underbrace{\overbrace{H}^{\text{RecType}_1}\ \overbrace{0002599}^{\text{Serial}_{2-8}}5483700100\ \overbrace{48090}^{\text{PUMS}_{19-23}}\ 9999799979\ldots000950\ldots9600}_{267}$

(10) $\underbrace{\overbrace{P}^{\text{RecType}_1}\ \overbrace{0002599}^{\text{Serial}_{2-8}}\ \overbrace{01}^{\text{Num.Person}_{9-10}}\ \ldots\ \overbrace{2}^{\text{Sex}_{23}}\ 0\ \overbrace{15}^{Age_{25-26}}\ 00000000000022324200\ldots0000}_{316}$

updated to the corresponding year of study. Fig. 3.3 shows a simplified flow chart for the process.

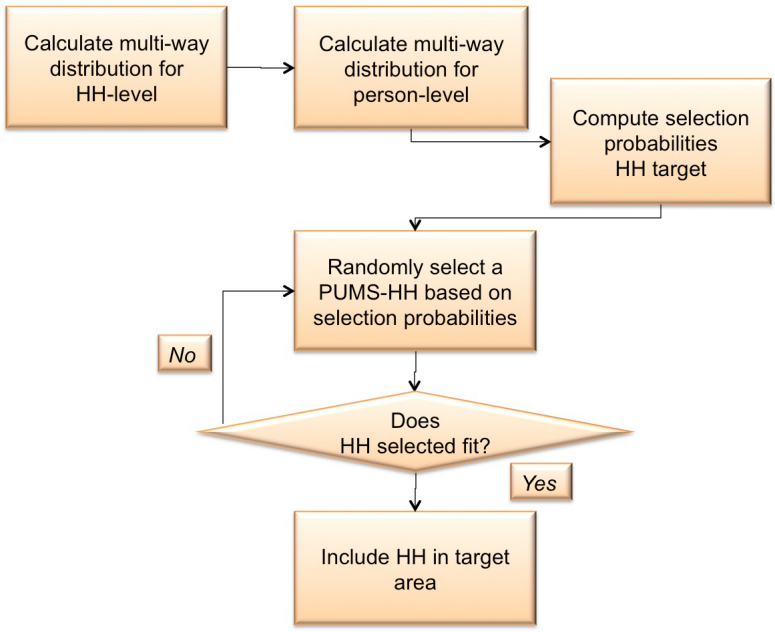


FIG. 3.3. Methodology schematic overview

The methodology presented in [16] and [131] requires US Census Bureau summary files 1 and 3 [30] and PUMS at 5%. Information is publicly available and accessible though the US Census Bureau Website [30] for Census 2000, and partially available for Census 2010. In both cases, the SFs contain aggregated population and housing characteristics collected from a 1 in 6 household sample and weighted to represent the total population. The micro data

sample contains disaggregated data records representing a sample of 5% of the occupied and vacant housing units in the United States, and the attributes of people living in the occupied units. PUMS files for 2010 Census are scheduled for release on December 2012 through April 2013, making Census 2000 the latest complete information available at the time data collection for this work was conducted. This research focused on information relative to Denton County demarcated by the Public Use of Microdata Area (PUMA) code 48090, as shown in Fig. 3.4.

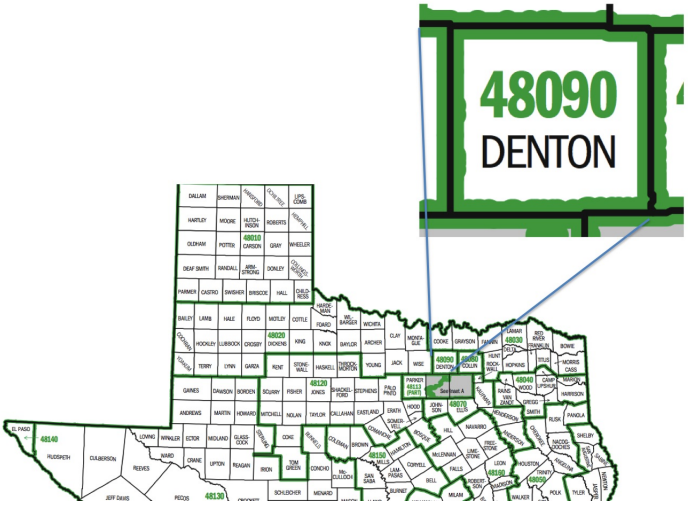


FIG. 3.4. Denton County PUMA code and delimitation [29]

Super PUMA Code	PUMA Code	Area Name
48090	02202	Denton city (part)

3.2.2. List of Control Variables

Table 3.3 shows the number of categories for selected demographics corresponding to the aggregate datasets. Table 3.4 contains a list of control variables used for the synthetic reconstruction of Ps and HHs. Table 3.11 contains the list of control variables used for the allocation of the resulting population P into schools through school enrollment.

TABLE 3.3. Features and subcategories selected at the person-level

Feature	Subcategories Examples	T. # of Subcat.
Age	5 to 9 years, 85 years and over	18
Gender	Female, Male	2
Race	White, Black	7
Ethnicity	Hispanic, Non Hispanic	2

TABLE 3.4. Variables and correspondent columns in SF1

Control	Description	SF1
P-Age	Age of the individual	P12
P-Race	Race of the individual	P7
P-Gender	Gender of the individual	P12
HH-Fam	Family or non-family household	P26
HH-Size	Size of the household	P26
HH-Type	Type of the household	P20
HH-Child	Households by presence of people under 18 years old	P19

3.2.3. Forecasting Population

Beckman et al. initially proposed the calculation of the selection probability of HHs using a simple ratio:

$$Pr_{i,L} = \frac{HH_i}{\sum_{j,i \neq j} HH_j} \tag{11}$$

Where $Pr_{i,L}$ is the probability of selecting household i of type L, HH_i is the weight for household i and HH_j represent the remaining households in the subregion L [16]. In order to select HHs with and additional control level this work considered the extension

methodology proposed by Auld and Mohammadian [7] and presented in (12). The new probability function is derived using Bayes theorem based on conditional probability, in order to select households of the sample using multi-level control.

$$Pr_{i,L} = \frac{HH_i \prod_j^{HH_i(size)} \frac{MWAY_P^* \times (p_{i,j}(1), p_{i,j}(2), \ldots, p_{i,j}(n))}{N_{remain}}}{\sum_k^{N_L} (HH_i \prod_j^{N} \frac{MWAY_P^* \times (p_{i,j}(1), p_{i,j}(2), \ldots, p_{i,j}(n)}{N_{remain}})} \tag{12}$$

Where:

- $Pr_{i,L}$ is the probability of selecting household i of type L
- HH_i is the weight for household i
- $HH_i(size)$ represents the size of household i
- $p_{i,j}(k)$ index of control variable k of person j of household i
- $Nremain$ represents the number of remaining persons in the sub-region
- N_L represents the remaining households in the sub-region L
- $MWAY_P^* \times (p_{i,j}(1), p_{i,j}(2), \ldots, p_{i,j}(n))$ represents the remaining cell frequency in zonal personal-level joint distribution

The following example (Tables 3.5 - 3.10) illustrates how multi-level control is used in the recreation of the synthetic population.

Table 3.5 shows the initial information found in the public micro-sample data and Tables 3.6, and 3.7 show the household-level joint distribution and person-level joint distribution, respectively. The probability for each type of house calculated using (11) is $Pr(HH_i) = 0.25$, since in the sample number of houses containing individuals with the desired characteristics is one per type of house HH_i. Therefore, in order to fulfill the required number of households $HH = 50$, 12.5 of each type of HH_i should be selected. On the other hand by using multi-level controls (12), the selection changes to better reflect person-level distribution. Table 3.9 shows the reformulated probability, using (12) and new number per HH_i. Tables 3.8 and 3.10 show that both joint distributions are preserved.

TABLE 3.5. Example of calculation of multi-level control selection of HHs

Micro-data Sample

Type of HHs	Sample (HH_i)	Demographics $p_{i,j}$	$Pr(HH_i)$[16]
HH_1	1	1 male 5 to 9 years, 1 female 5 to 9 years	0.25
HH_2	1	1 male 10 to 14 years, 1 female 10 to 14 years	0.25
HH_3	1	1 male 5 to 9 years, 1 female 10 to 14 years	0.25
HH_1	1	1 male 10 to 14 years, 1 female 5 to 9 years	0.25

TABLE 3.6. HHs joint distribution

HH size(2)	50

TABLE 3.7. P-Level joint distribution

Gender $p_{i,j}(1)$-Age$p_{i,j}(2)$	5-9	10-14	Total Rows
Male	20	30	50
Female	25	25	50
Columns Total	45	55	100

From (12), the probability of selecting a HH from PUMS is calculated for each HH_i:

$$Pr_{HH_1} = \frac{(1) \times \frac{20/100}{25/100}}{(1) \times \frac{20/100}{25/100} + (1) \times \frac{30/100}{25/100} + (1) \times \frac{20/100}{25/100} + (1) \times \frac{30/100}{25/100}} = 0.2 \tag{13}$$

$$Pr_{HH_2} = \frac{(1) \times \frac{30/100}{25/100}}{(1) \times \frac{20/100}{25/100} + (1) \times \frac{30/100}{25/100} + (1) \times \frac{20/100}{25/100} + (1) \times \frac{30/100}{25/100}} = 0.3 \tag{14}$$

$$Pr_{HH_3} = \frac{(1) \times \frac{20/100}{25/100}}{(1) \times \frac{20/100}{25/100} + (1) \times \frac{30/100}{25/100} + (1) \times \frac{20/100}{25/100} + (1) \times \frac{30/100}{25/100}} = 0.2 \tag{15}$$

$$Pr_{HH_4} = \frac{(1) \times \frac{30/100}{25/100}}{(1) \times \frac{20/100}{25/100} + (1) \times \frac{30/100}{25/100} + (1) \times \frac{20/100}{25/100} + (1) \times \frac{30/100}{25/100}} = 0.3 \tag{16}$$

Table 3.8. HHs

HHs	$Pr(HH_i)[7]$	Total
HH_1	0.2	10
HH_2	0.3	15
HH_3	0.2	10
HH_4	0.3	15

Table 3.9. Final selection of HHs

Type of HHs	Demographics $p_{i,j}$
HH_1	10 males 5 to 9 years, 10 females 5 to 9 years
HH_2	15 males 10 to 14 years, 15 females 10 to 14 years
HH_3	10 males 5 to 9 years, 10 females 10 to 14 years
HH_1	15 males 10 to 14 years, 15 females 5 to 9 years

Table 3.10. Synthetic Ps

Gender -Age	5-9 years	10-14 years	Total Rows
Male	20	30	50
Female	25	25	50
Total Columns	45	55	100

3.2.4. Assigning Children to Schools

The National Center for Education Statistics (NCES) and the Core Common Data (CCD) have made available information regarding total school enrollment, grades, and location for SDs and their schools through their website [122]. Total enrollment, number of schools and additional demographic counts by state and by SD are also available. Complete information with several multi-way tables is available at the NCES website [63] from 1986 to present.

School Attendance Zones (SAZ) are areas that surround public schools defining the set of HHs eligible to attend each school. The U.S. Census Bureau provides information regarding the School District Review Program that includes SDs boundaries and registers their changes over time. Nevertheless, SDs and Independent School Districts (ISDs) are not required to report information regarding SAZs boundaries. SDs make SAZs information available for

TABLE 3.11. School information used

Control	Description
S-Type	School type
S-Enrollment	Student enrollment of school
S-LocationX	Location of the school longitude component
S-LocationY	Location of the school latitude component
S-HighLevel	Highest level of the school
S-LowLever	Lowest level of the school

the public on their websites using different formats such PDFs and interactive maps. Since information of attendance zones is essential not only for research studies but for other commercial uses, independent companies offer such information and additional services. Partial information on school boundaries can be found at SABINS (School Attendance Boundary Information System) [123]. Non-public nation-wide information on SAZs is available through Maponics (`www.maponics.com/`). Fig. 3.5 exemplifies the SAZ map for middle schools in the Denton ISD. The datasets listed above were, to the best of our knowledge, the only data sources available at the time this research was conducted.

In order to extend the synthetic reconstruction methodology, is necessary to include schools associated to households through the use of total school enrollment, and attendance zones. The second part of the process is to create the households-schools affiliations when appropriate. The procedure shown in Fig. 3.6 explains how the association works. At this step only households with eligible school children (persons from age 3-18 and enrolled in public schools) are taken into account. Once a household with an eligible individual is encountered, the algorithm searches for the nearest school with the appropriate level. If the school has not reached its enrollment level then the household is assigned to that school;

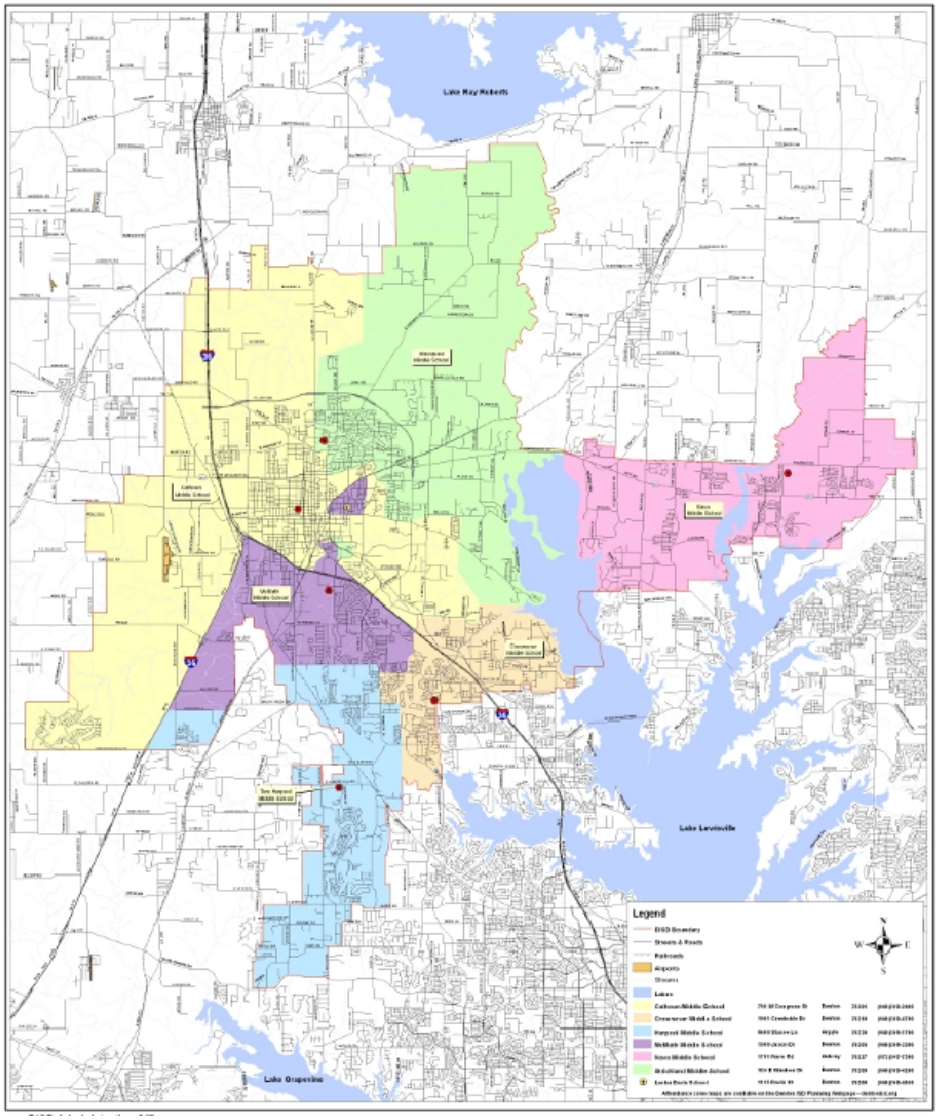


FIG. 3.5. SAZ coding showing the correspondence street to middle-school for the Denton ISD. The different colored area corresponds to the catchment area assigned to each school, retrieved from the Denton ISD website [86]

otherwise, the selection procedure goes to the next closest school.

3.3. Results

To validate the simulator POPSYN, the joined distribution at the person-level of the synthetic population generated is compared with the census person-joint distribution for the entire area and the forecast joint distributions for the target year (2009). The selected

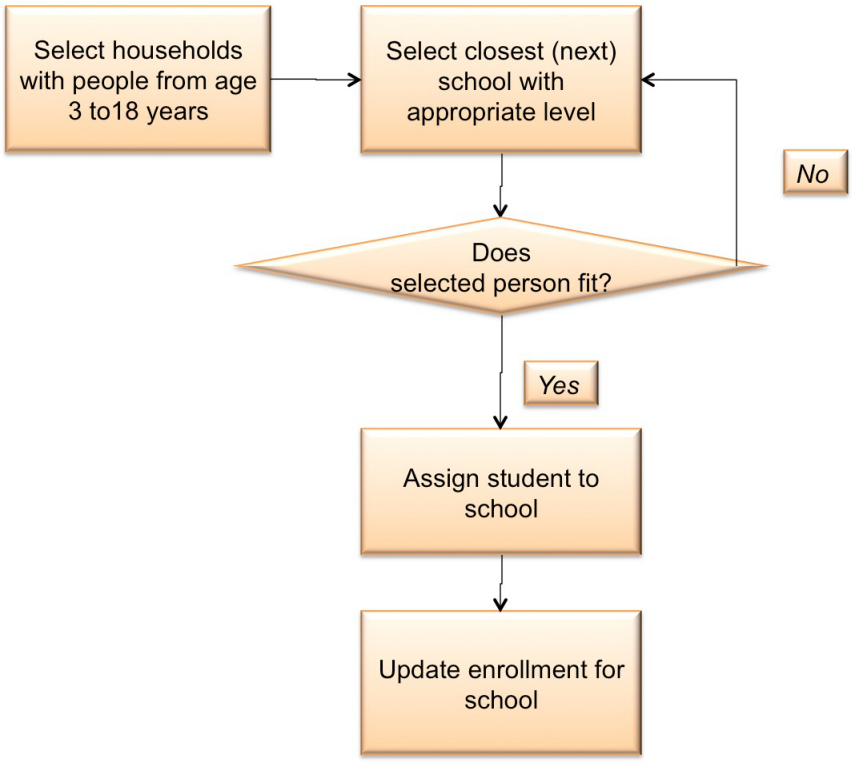
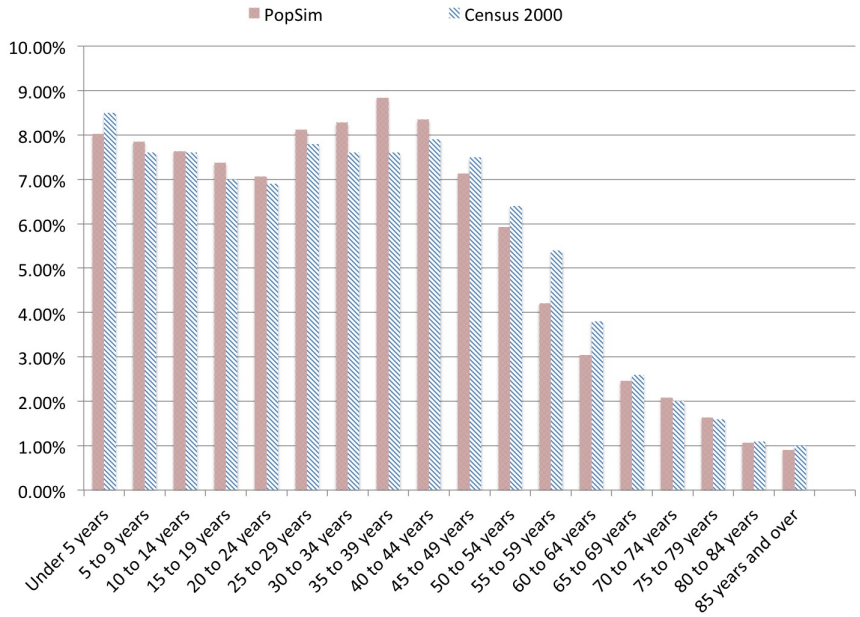


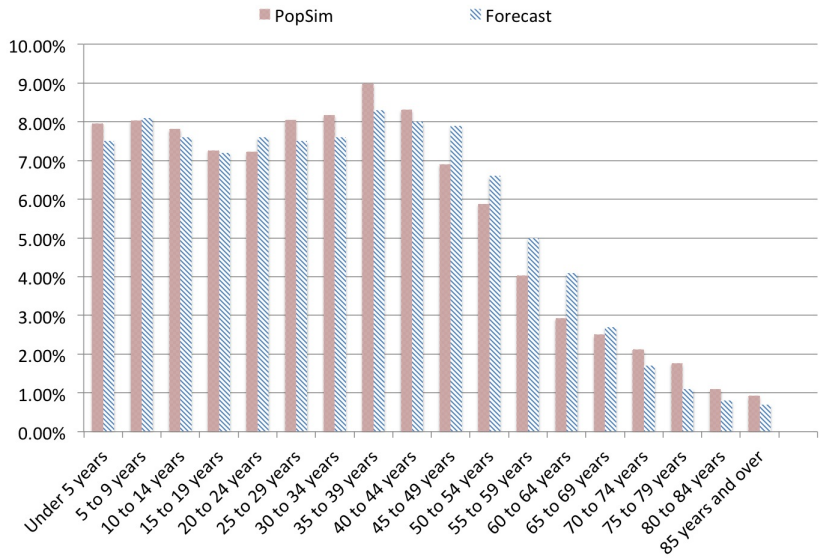
FIG. 3.6. Second part of the methodology, used to assign households to school attendance zones.

categories for comparison are age (Fig 3.7), gender (Fig. 3.8) and race (Fig. 3.9).

The largest difference between the expected and observed person-marginal distribution totals is found at the age feature. The difference of sub category "60-64 years" is equal to 1.18%, Fig. 3.7. In this case, the synthetic reconstruction is compared with the base year distributions. For the gender category, the largest difference is 0.51% and also occurs with the difference between the base year and the synthetic population Fig. 3.8. In the case of the feature "Race", the largest difference is 6.7% found at the "White" subcategory, comparing the synthetic reconstruction with the forecast year. The actual plot for the synthetic population mapped into Denton County's map is shown in Figs. 3.11 and 3.10.



(A) Expected and generated marginal distributions for person-level base year

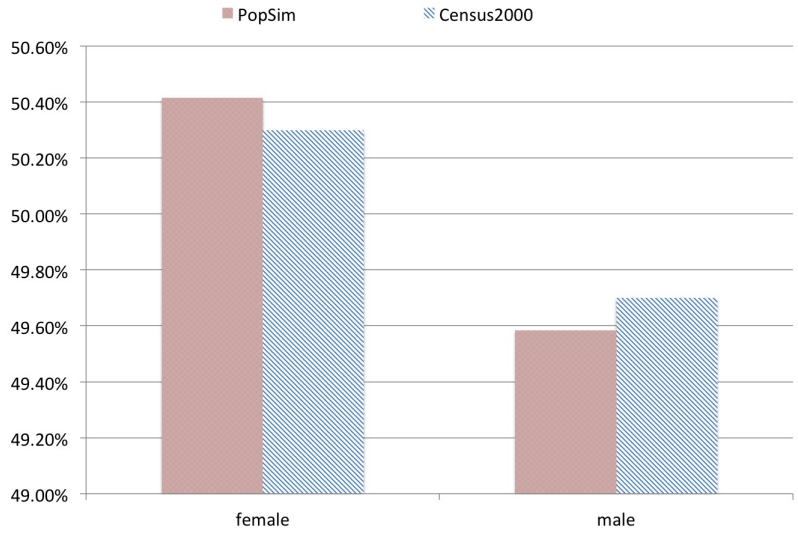


(B) Expected and generated marginal distributions for person-level forecast year

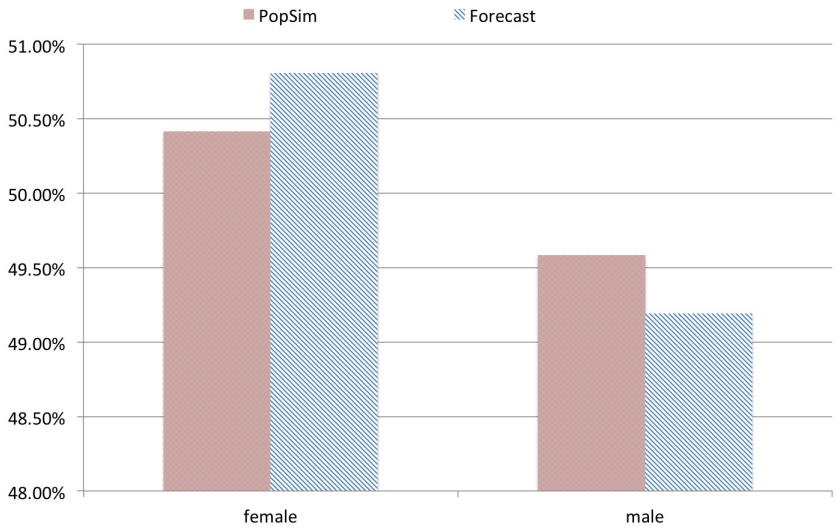
FIG. 3.7. Individual-level control variable "Age"

3.3.1. Standard Deviation and Standard Error

In order to estimate an accurate size for the synthetic population, the standard error and standard deviation were calculated for selected categories. Populations of sizes of a



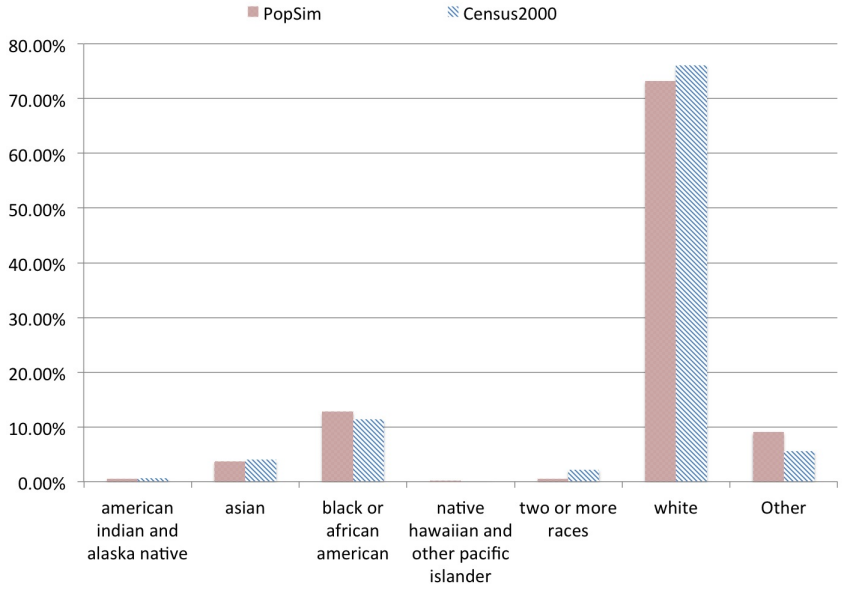
(A) Expected and generated marginal distributions for person-level base year



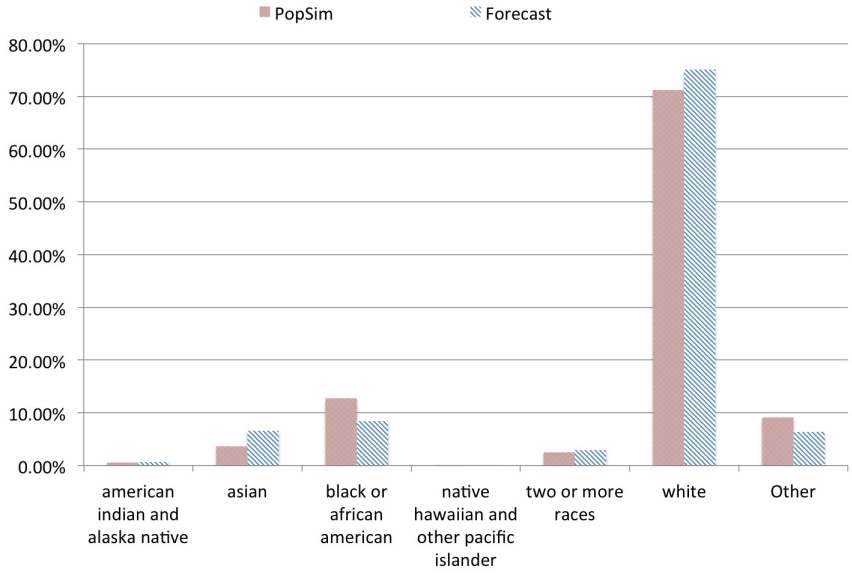
(B) Expected and generated marginal distributions for person-level forecast year

FIG. 3.8. Individual-level control variable "Gender"

thousand, ten thousand, and one hundred thousand were simulated fifty times in order to calculate both standard error and standard deviation. Standard deviation is calculated using



(A) Expected and generated marginal distributions for person-level base year



(B) Expected and generated marginal distributions for person-level forecast year

FIG. 3.9. Individual-level control variable "Race"

(17) and the standard error is calculated using (18). The results are presented in Table 3.12.

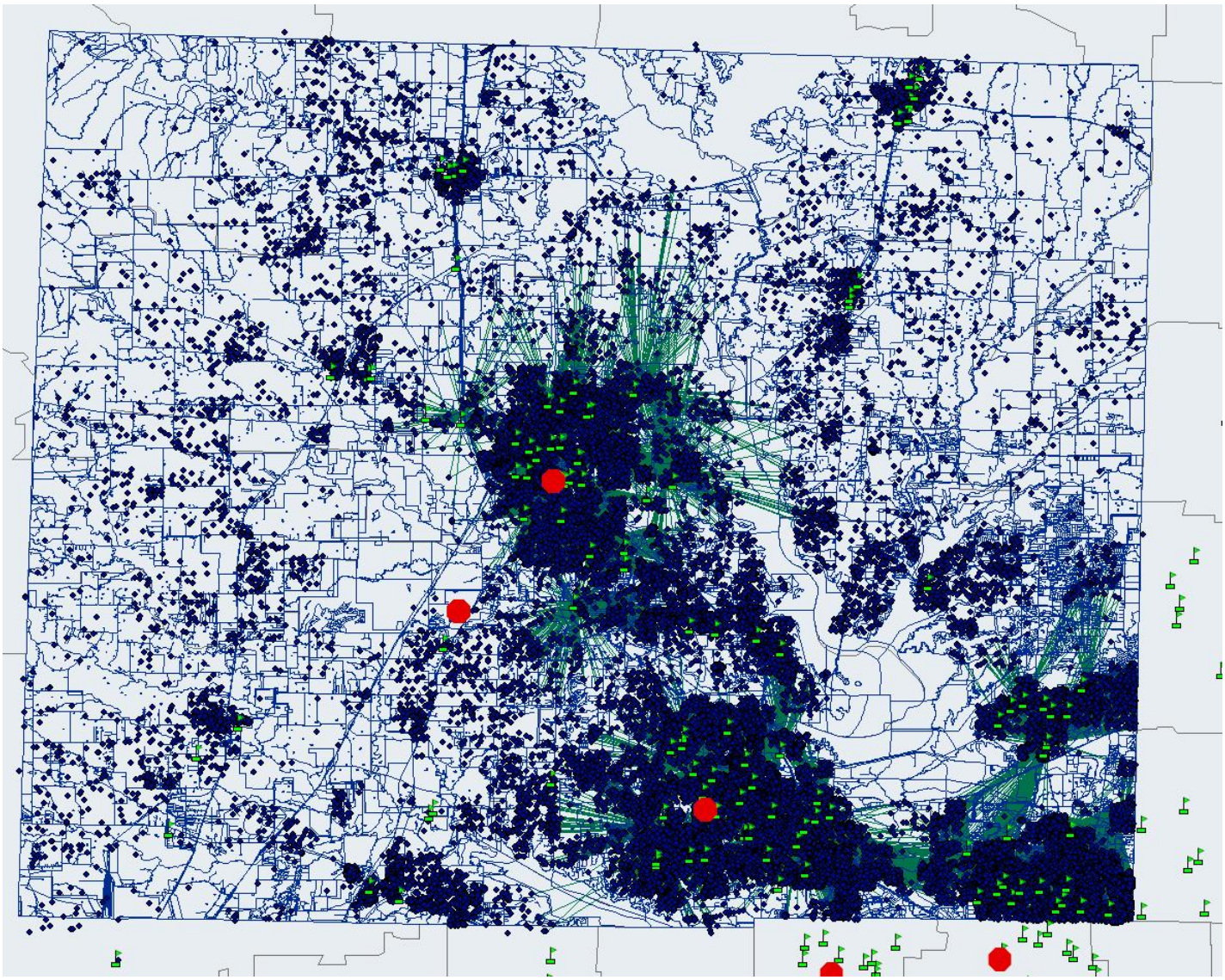


FIG. 3.10. Plot of households distributed over the Denton County map

$$STD = \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{17}$$

$$STDERROR^2 = \mathrm{var}(\bar{x}) = \frac{\sigma^2}{n} \tag{18}$$

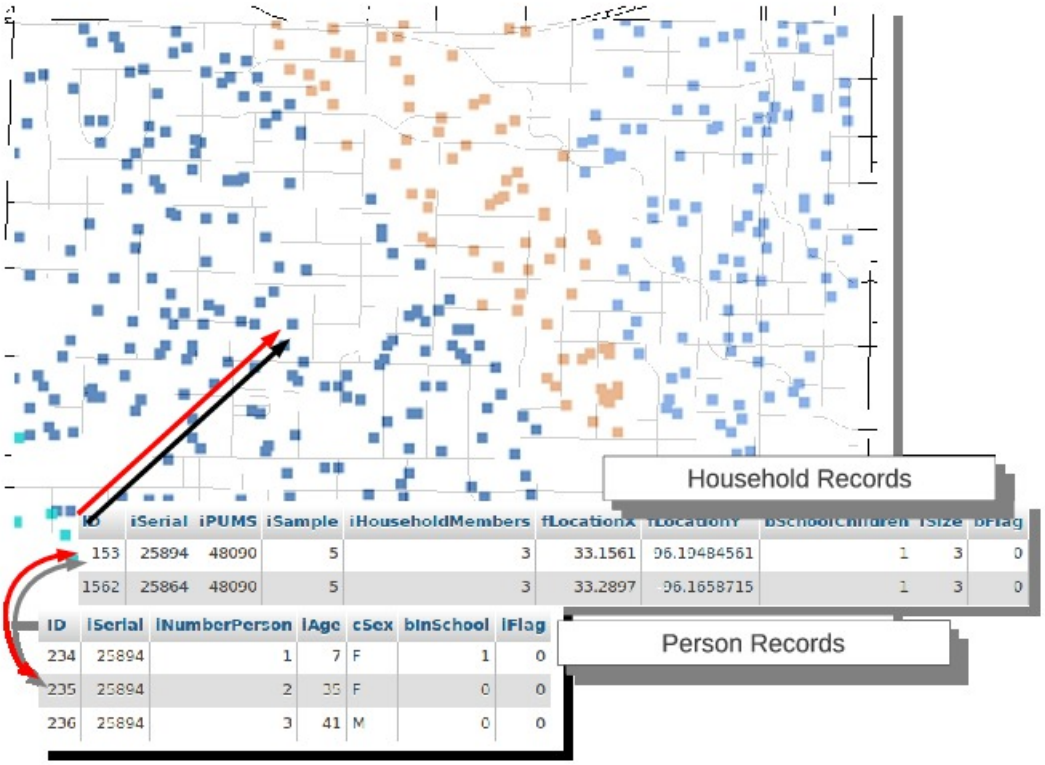


FIG. 3.11. Plot of households distributed a geographic region. Tables at the household and personal-level are shown for a selected household

3.4. Discussion

There has been a notable theoretical interest in synthetic reconstruction as the base methodology used to obtain records at personal and household level oriented towards various objectives. Synthetic reconstructions are needed in order to map the interactions that may occur in the schools system. Although the methodology was designed initially to estimate traffic demand, it has been widely employed in the field of computational epidemiology as a framework for the study of disease dynamics. The extension of the methodology using multi-control levels [7] allows a more accurate selection of the sample households included in the PUMA. In the case of this study, Denton County's area was entirely enclosed in the

TABLE 3.12. Standard deviation and standard error for different synthetic population sizes

TABLE 3.13. STD ‰

	Synthetic Population Size		
	1000	10000	100000
Age	6.829‰	1.967‰	0.670‰
Gender	19.200‰	3.644	1.15‰
Race	2.478‰	0.725‰	0.2373‰
Ethnicity	11.778‰	5.453‰	1.178‰

TABLE 3.14. STDERROR ‰

	Synthetic Population Size		
	1000	10000	100000
Age	1.247‰	0.743‰	0.335‰
Gender	3.505‰	1.377‰	0.576‰
Race	0.452‰	0.274	‰0.119‰
Ethnicity	2.150‰	2.061‰	0.589‰

super PUMA 48090, but for other geographies this is not the case. Other counties may belong to one or more super PUMAs and the process for the household selection changes. Wheaton et al. addressed modifications that were made in the case the target area includes more than one super-PUMA [166]. The extension of the algorithm to allocate children attending schools was initially proposed by [166], but as more information is made available regarding SAZs, the approximation function for enrollment of students could potentially be made more accurate. Nevertheless, the allocation of households in any methodology still follows a random pattern.

3.5. Summary

The synthetic reconstruction methodology with multiple control levels is presented as the methodology used in order to reconstruct the population of Denton County and the school population of the Denton ISD. First, the general methodology has been described, including the list of control variables used in this research. Next, the modification to include multi-control levels has been presented with an illustrative example. Finally, the results of the reconstruction are presented, compared and contrasted with the initial distribution and the distribution estimated for the forecast year.

CHAPTER 4

DEFINITION OF MULTI-COAFFILIATION NETWORKS (MCN)

[124]

> Perhaps even more than to the contact between mankind and nature, graph theory owes to the contact of human beings between each other
>
> ---
>
> *Dénes König,1936*

In this chapter the concept of Multi Co-affiliation Networks (MCNs) [124] is formalized. MCNs are the result of the execution of the School Affiliation Network Discovery (SAND) algorithm [125] over a synthetic reconstruction. The chapter starts with the formal definition of graphs and continues revising concepts used to characterize the structure of MCNs. Graph theory is the area of the mathematics that studies graphs and their properties. In the next section, the mathematical background for bi- and k-partite graphs is addressed. Section 4.2 defines the methodology for construction of MCNs based in the synthetic reconstruction discussed in Chapter 3. MCNs are weighted graphs that depict the strength of affiliation between a community formed by households and schools. Weighted networks are seldom used to represent social graphs, although recent studies have found them to be of special interest in order to represent hierarchical organization and connectivity of nodes [169]. Section 4.3 exposes the result of constructing the MCN from the synthetic population of Denton County and the Denton Independent School District (ISD). The chapter concludes with a discussion of the result and the implications the methodology proposed to generate MCNs.

Definition 4.1. Definition of Graph [165] A graph is a tuple $G = (V, L)$, V represents the set of vertices and L represents the set of edges.

A graph with undirected edges and free of self-loops is called a simple graph. In epidemiology, graphs have become an instrumental representation of the underlying social structure that diseases use to propagate. Graphs that depict biological processes or associa-

tions are generally sparse, complex, large, and have global topological properties in common [9]. Nodes are not isolated, on the contrary, they are reachable within a small number of connections one from another. Disease, influence, and information propagation in such networks, are a combination of the internal state of a node and the state of the immediate environment [83]. Extensive analysis of online networks' structures has found a strong correlation between the attributes of a node and the preferential attachment to other nodes. The likelihood of the existence of a link between nodes is driven by the sharing of a particular set of attributes. Attributes of nodes that belong to online communities have been found to have corresponding demographics based in real life characteristics [84]. In the search of a network that strongly resembles real life connections and acquaintances, this research has focused on the communities that are conformed by schools and household. In this chapter, the concept of contact networks is addressed from a new perspective. Long-term affiliations formed in localized communities are studied to analyze their structure. Households are represent as nodes while links are defined as the association between households and schools.

4.1. Bipartite and $K-$partite Graphs

DEFINITION 4.2. Bipartite Networks [165] A graph G is bipartite if $V(G)$ is the union of two disjoint (possibly empty) independent sets called partite sets of G.

$G = (V, E)$ is bipartite when:

$$\begin{aligned} V &= V_1 \cup V_2 \\ V_1 \cap V_2 &= \emptyset \\ E &= V_1 \times V_2 \end{aligned} \tag{19}$$

A contact network model is a graph $G = (V, E)$ in which the nodes represent individuals (V) and edges (E) represent possible contacts between nodes. In [56] the network construction is motivated by simulated contacts and their estimated positions and activities on a step-by-step basis. In the general model, the social contact network is represented as a bipartite graph, G_{PL}, of people (P) and location (L) that can be in the order of the millions of nodes

[55]. Edges represent possible contacts of two people attending the same location L, at overlapping times. To draw an edge, the duration of contact needs to be greater or equal to a set minimum threshold i.e. two hours. Two projections of G_{PL} can be obtained by drawing an edge between all pairs of vertices at a distance of two from each other. A graph: G_P, containing only people vertices, and G_L, containing only locations. This approach considers all the locations to be of a similar category.

More recent models have extended the concept by including three or more categories such as workplaces and shopping places [107], and hierarchy inside the locations [168] (i.e. grades inside a school). These models focus on the personal level and in all cases require either supercomputing or parallelization techniques. Contact networks are instrumental tools because they allow studying how the structure of the network impacts the spread of the pathogen. By focusing not necessarily on the "who" (i.e. a particular individual) but rather on the "where" and "how", general conclusions can be drawn with respect to intervention strategies. Affiliation networks go one step forward, by modifying the selection of random contacts into long-term affiliations. For communicable diseases like influenza, the frequency and length of the contact period between a susceptible individual and an infectious one, is a determinant factor for the spread of the disease. Furthermore, the presence of a child in the household has been linked with the increase of individual-risk level for 2009 AH1N1 [134]. Finally, the spatial distribution of H1N1 in Mexico has been linked with the school cycles in the country, suggesting that school closures and similar mitigation measures have a remarkable potential to control future epidemics.

The proposed model focuses on the long-term affiliations that schools and households form during an academic year. In order to study the school system as the main strategic point for intervention measures, each school is considered to be a community of households. School districts demarcate the set of houses that are assigned to the schools. A school attendance zone is the area surrounding a particular school that defines the houses that are assigned to such school.

4.2. School Affiliation Network Discovery (SAND) Algorithm [125]

The *School Affiliation Network Discovery (SAND)* algorithm generates MCNs based on the synthetic reconstruction discussed in the previous chapter. The model is based on the following observations.

- Schools districts pre-designate attendance zones. There exists a one-to-one correspondence of schools and households for every School Attendance Zone (SAZ).
- The affiliation relation "child-school" is persistent for long periods of time (i.e. duration of a school year).
- Small children are highly limited in self-movement. Therefore, their location is constrained to their houses and childcare or school locations.
- SDs are composed by hierarchies of by elementary (E), middle (M) and high schools (H).

4.2.1. Model Assumptions

The SAND algorithm makes the following assumptions regarding people attending schools.

- Each household with school-age children is affiliated with the closest school by calculating the Euclidean distance from hh_i, s_j.
- If the household has more than one member attending a specific school type (i.e. elementary school s_i), then those members are grouped and assigned to the same school s_i. Therefore, each household has assigned only one school for each type.

The euclidean distance has been used in previous research papers to assign agents to schools and workplaces [166]. The second assumption of the model, although restrictive, reflects the reality of the majority of U.S. households, where SAZs are predefined by the school districts.

4.2.2. Generation Algorithm for Affiliation Networks, Graph A

Affiliation networks describe a binary relationship between members of two distinct set of items. A classic example of this type of networks is the data-set collected by Davis et

al. in which the attendance of women to social events on a small southern town was recorded [49]. The dataset of women and events has become an example of affiliation network [22] and subject to analysis for clustering and centrality algorithms [21], [24]. Similarly, this research constructs affiliation networks of households and schools. First, a study area is selected; in this research the study area is Denton ISD. Utilizing the synthetic reconstruction algorithm defined in Chapter 3, the synthetic households are distributed over the study area. The set of nodes is formed by the set of schools of corresponding to the SD. The resulting network, renamed as graph A, is defined as the affiliation network of households and schools.

DEFINITION 4.3. Affiliation Network [125] A simple undirected bipartite graph $A = (V, L)$, $V = (S \cup HH)$ is constructed to represent the affiliation network formed by schools S and households HH, an edge $(hh_i, s_j) \in L$ represents a child belonging to household $hh_i \in H$ attending school $s_j \in S$.

The indicator function $\mathcal{A}(hh_i, s_j)$ is a function that identifies the membership of a hh_i to s_j.

$$\mathcal{A}(hh_i, s_j) = \begin{cases} 1 & \text{if } (hh_i, s_j) \in L \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

The definition of indicator function $\mathcal{A}$ could potentially vary from region to region. In the United States, SDs are responsible for delimiting the school attendance zones and making the information available to the public. Maps can be publicly accessed in diverse formats, regularly through SDs websites. The attendance zones may also be updated as the population changes and new establishments are created. Although there is not a defined standard for how the information is presented to the public, in the United States, it can be obtained in PDF format, interactive maps or a listing of streets.

Nevertheless, lack of information in other geographic regions may become an additional constraint for the model depending on the availability of data. For other geographies, the selection process for attending public schools may be mainly driven by density of the population, religion preference or purchasing power. By defining the preferential attachment

of schools and households as a function, the model is open to allocate other definitions for "affiliations."

Following, the algorithm that generated affiliations is described. Initially, a geographical unit $\mathcal{L}$ for which the synthetic reconstructions has been generated is selected, where:

$$\mathcal{L} = \begin{cases} HH = \{hh_1, h_2, \ldots, hh_i, \ldots, hh_n\} \\ P = \{p_{i1}, p_{i2}, p_{i3}, \ldots\} \\ S = \{s_1, s_2, \ldots, s_k, \ldots, s_m\} \end{cases} \tag{21}$$

$HH = \{hh_1, hh_2, \ldots, hh_i, \ldots, hh_n\}$ is the set of households hh_i located within the area of $\mathcal{L}$. Each household is defined as a set of attributes: $hh_j(k) = \{[serial], [number\_people], ...\}$ characteristics extracted from the of the synthetic reconstruction.
$P = \{p_{i1}, p_{i2}, \ldots, p_{ij}, \ldots\}$ is the set of j people p_{ij} that belongs to household i. The largest household size $max(|hh_j|)$ registered in Texas at five percent sample file SF1, is thirteen.
Each person also represents a set of attributes:
$p_{i,j}(k)$ = {[age],[gender],[attends_school],[household_serial],...} .
$S = \{s_1, s_2, \ldots, s_k, \ldots, s_m\}$, is the set of schools s_k, that belong to the SD located in the area of $\mathcal{L}$ $s_k = \{[type], [lower\_level], [higher\_level], [locationX], [locationY], ...\}$

Detailed attributes for P, HHs and S come directly from those used in the synthetic reconstruction and were discussed in the previous chapter. The following algorithm is used to link schools and households.

Algorithm 1: Affiliation Network: Generation of Graph A [1]

Require: HH, P, S

1: **for** $i = 1$ TO $|HH|$ **do**
2: **for** $j = 1$ TO $hh_i[num\_people\_in\_HH]$ **do**
3: **if** $p_{(i,j)}.attends\_schl$ is TRUE **then**
4: **for** $k = 1$ TO $|S|$ **do**
5: **if** $p_{(i,j)}.level \geq s_k.lower\_level$ AND $p_{(i,j)}.level \leq s_k.upper\_level$ **then**
6: $psbl\_schls_{(i,k)} \leftarrow dist(hh_i.loc, s_k.loc)$
7: **end if**
8: **end for**
9: $sschl \leftarrow min(psbl\_schls)$
10: $lnk\_schlsHH(i, sschl) \leftarrow crt\_lnk(i, sschl)$
11: **end if**
12: **end for**
13: **end for**

Algorithm 1 generates Graph A as follows. Lines 1-7 review all households and select the ones with students, calculate the distance between each eligible household and all the schools with appropriate level. Lines 9 and 10 select the minimum distance among all schools and create the link school-household. Lines 1,2, and 4 show that the algorithm is executed in $O(S \times P)$ and constructs all $HH - S$ affiliations.

LEMMA 4.4. *Graph A is a bipartite graph with partitions HH and S*

PROOF. Suffices to observe that function $\mathcal{A}$ as defined in (20), can only generate edges

[1]This figure is reproduced from [124], with permission from Global Science and Technology Forum `http://www.globalstf.org`

joining one element of HH and one element of S. Since edges are only created joining one element of HH to an element form S the graph A is bipartite with partitions HH and S $\square$

4.2.3. Multi Co-affiliation Networks (MCNs), Graph B [124]

As mentioned in the previous section, affiliation networks demarcate the relation of two different set of objects. In some cases the purpose of construction affiliation data is not to understand the links between two disjoint sets but to understand the patterns within only one set [22]. The term co-affiliation refers to the relation that arises form joint affiliations shared among two similar objects. Co-affiliation networks can be understood as the realization of a relation that may not be discernible at first sight.

In this research, the school system is the main focus of analysis. The main purpose is to set a methodology that allows to quantify the relevance of each school in the system of a particular location $\mathcal{L}$. Starting from the affiliation network represented by graph A, this research proposes the construction of the Multi Co-affiliation Network (MCN) or graph B of schools based on the affiliations that schools share through common households.

Definition 4.5. Definition $k-$partite Graph [165]

By generalization of definition 4.2, if $G = (V, E)$ is k-partite, then:

$$
\begin{aligned}
& V = \cup_1^n V_i, i = 1, 2, 3 \ldots n \\
& \cap_{(i,j), i \neq j} V_i, V_j = \emptyset \\
& E = V_i \times V_j, \{\forall (i,j), i \neq j, i = 1, 2, 3 \ldots n\}
\end{aligned} \tag{22}
$$

Graph A is transformed into a weighted graph $B = (S, EE)$, $S = (E, M, H)$:

$$
e_k = (s_i, s_j) \in EE = \forall_{hh_k \in HH} \mathcal{A}(hh_k, s_i) \times \mathcal{A}(hh_k, s_j) \tag{23}
$$

The function $W(e_k) \geq 0$ represents the weight of the edge e_k and is defined as follows:

$$
W(e_k) = \sum_{hh_k \in HH} \{(hh_k, s_i), (hh_k, s_j)\} \in A \tag{24}
$$

The construction of graph B is schematized in Fig. 4.1.

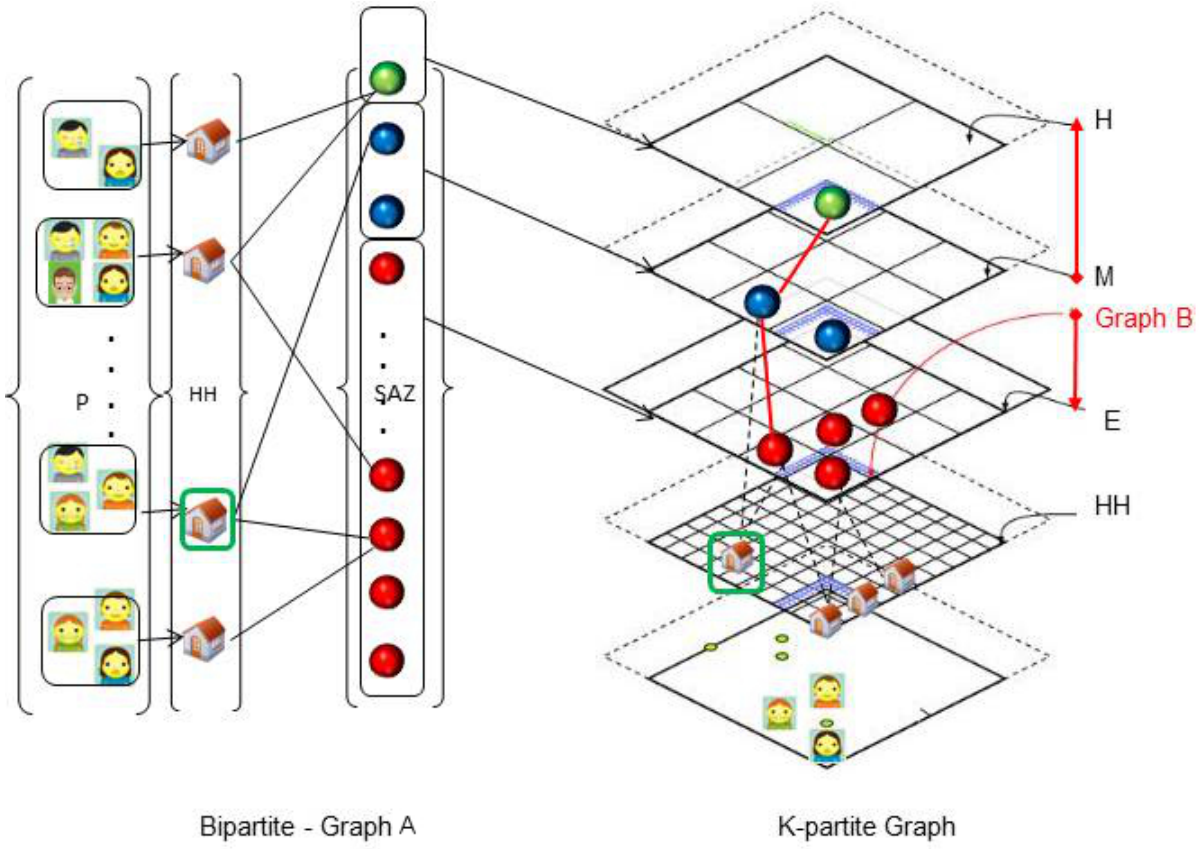


FIG. 4.1. Construction of the k-partite graph. Starting from the affiliation network represented by graph A, edges and nodes of graph B are derived. [2]

The graphic on the left side corresponds to graph A, the affiliation network between schools and households constructed using the synthetic reconstruction. Each household represent a group of people that live at location (x, y). The graphic on the right side demonstrates how links between schools are formed by stretching common households between schools into links. Each plane corresponds to a specific type of school and is denoted by its initial (i.e. "Elementary" = E). In the multidimensional space, the first plane corresponds to the map in which all households are located.

Algorithm 2 constructs graph B as follows. Lines 1-7 describe the loop that reviews

[2]This figure is reproduced from [124], with permission from Global Science and Technology Forum `http://www.globalstf.org`

Algorithm 2: Graph B: Generation of MCNs [3]

Require: $S, lnk\_schlsHH$

1: **for** $k = 1$ TO $|S|$ **do**
2: $list\_HH_{[k]} \leftarrow select(k, lnk\_schlsHH)$
3: $lnk\_schls_{[k]} \leftarrow select(list\_HH_{[k]}, lnk\_schlsHH)$
4: **for** $i = 1$ to $|lnk\_schls_{[k]}|$ **do**
5: **if** $(lnk\_schls[i] \neq s[k])$ **then**
6: $lnk\_schls\_schls \leftarrow crt\_lnk[(k, lnk\_schls[i])]$
7: **end if**
8: **end for**
9: **end for**

all households linked to schools and links schools together when a common household is found. Lines 1 and 4 are executed $|S|$ and $|HH|$ times respectively, therefore the run-time of the algorithm in the worst case scenario is $O(|S| \times |HH|)$ to construct all $S-S$ co-affiliations.

THEOREM 4.6. *Graph $B = (S, EE)$ is a $k-$partite graph, with $k = 3, S = E, M, H$.*

PROOF. If B is not $k =$ *partite* then there must be an edge $e_k = (s_i, s_j)$,such $s_i, s_j \in E$ or $s_i, s_j \in M$ or $s_i, s_j \in H$.
By (23) if $e_k = (s_i, s_j) \in EE = \mathcal{A}(hh_k, s_i) \times \mathcal{A}(hh_k, s_j)$.
Then hh_k has two members that have the same corresponding school type either E, M, H attending different schools, which is a contradiction to the initial assumptions of the model. Therefore, for each $e_k = (s_i, s_j)$ if $s_i \in (E|M|H)$ then $s_j \in (\bar{E}|\bar{M}|\bar{H})$

□

Finally, it can be concluded that MCNs are k-partite, with $k = 3$.

[3]This figure is reproduced from [124], with permission from Global Science and Technology Forum `http://www.globalstf.org`

4.2.4. Selection of Affiliation Function $\mathcal{A}$

In the United States, SDs are responsible for delimiting SAZs and making the information available to the public. The selection of $\mathcal{A}$ may be described by one of following the cases:

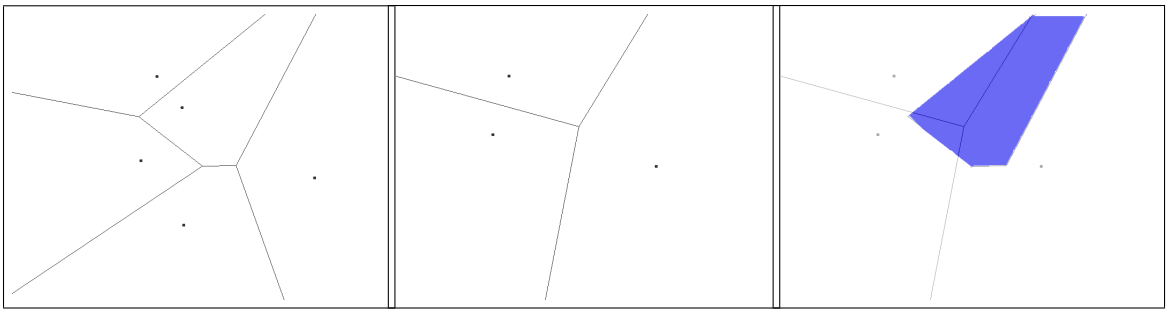
- Utilization of maps or SAZs.
- A proximity function by measuring household-school distance.
- School information regarding the list of students attending schools.

TABLE 4.1. Selection of $\mathcal{A}$

$\mathcal{A}$	Sources of Information	Format
Street location	SABINS MAPONICS	Shapefiles, digital information available
Distance function	Voronoi Diagrams	Automatically generated
Attendance list	Schools in the system	Diverse

The SDs' areas can be conceptualized as planes in space. A useful definition on how to partition the plane is the concept of Voronoi diagrams. A Voronoi diagram is a "partition of the plane with n points into convex polygons such that each polygon contains exactly one point and every point in a given polygon is closer to its generating point than to any other" [163]. Each school type divides the area in a similar fashion than a Voronoi diagram although in the actual SAZs the polygons are neither convex nor continuous. Therefore, the entire school system would be represented by as many planes as school types, in this case three. Fig. 4.2 exemplifies the intersection that is produced by overlapping divisions of two Voronoi diagrams.

Function $\mathcal{A}$ may change for environments for which information is not available or when the catchment area is defined by other measures. The function $D(\mathcal{A})$ can be modified to accommodate appropriately the corresponding sources of information. In this research, D is defined as Euclidean distance (25) which was used to draw the links between S and HHs.



(A) Voronoi diagram 1 (B) Voronoi diagram 2 (C) Intersection

FIG. 4.2. Intersection of two Voronoi Diagrams 1 and 2, showing how one cell from diagram 1 intersects multiple cells from diagram 2

$$D = \sqrt{(x(h_i) - x(s_j))^2 + (y(h_i) - y(s_j))^2} \tag{25}$$

In (25) the functions $x(p)$, and $y(p)$ retrieve the latitude and longitude of a given point p.

The resulting division of the plane will be equivalent to a Voronoi diagram for which centroids are the locations of a schools. Each school type corresponds to a plane, in which the number of Voronoi cells are equal to the number of schools. The function $\mathcal{A}$ defines the division of the plane corresponding to each type of school. A different way of constructing MCNs is to look into the intersections caused by the overlap of the planes representing each type of school.

4.3. Experimental Results

4.3.1. Application Example

Maps identifying the attendance zones for the Denton ISD are available on their website. Herein codes defined in Table 4.2 are used.

By the time the present research was carried on, the NCES [62] had a record of thirty eight schools corresponding to the Denton ISD. Although, the school attendance zone maps for each type of school posted on the ISD website identify thirty SAZs. Coding, name and

TABLE 4.2. Denton ISD application coding

Type	Assigned Code	Total Number of Schools
Elementary	A	21
Middle	B	6
High	C	3

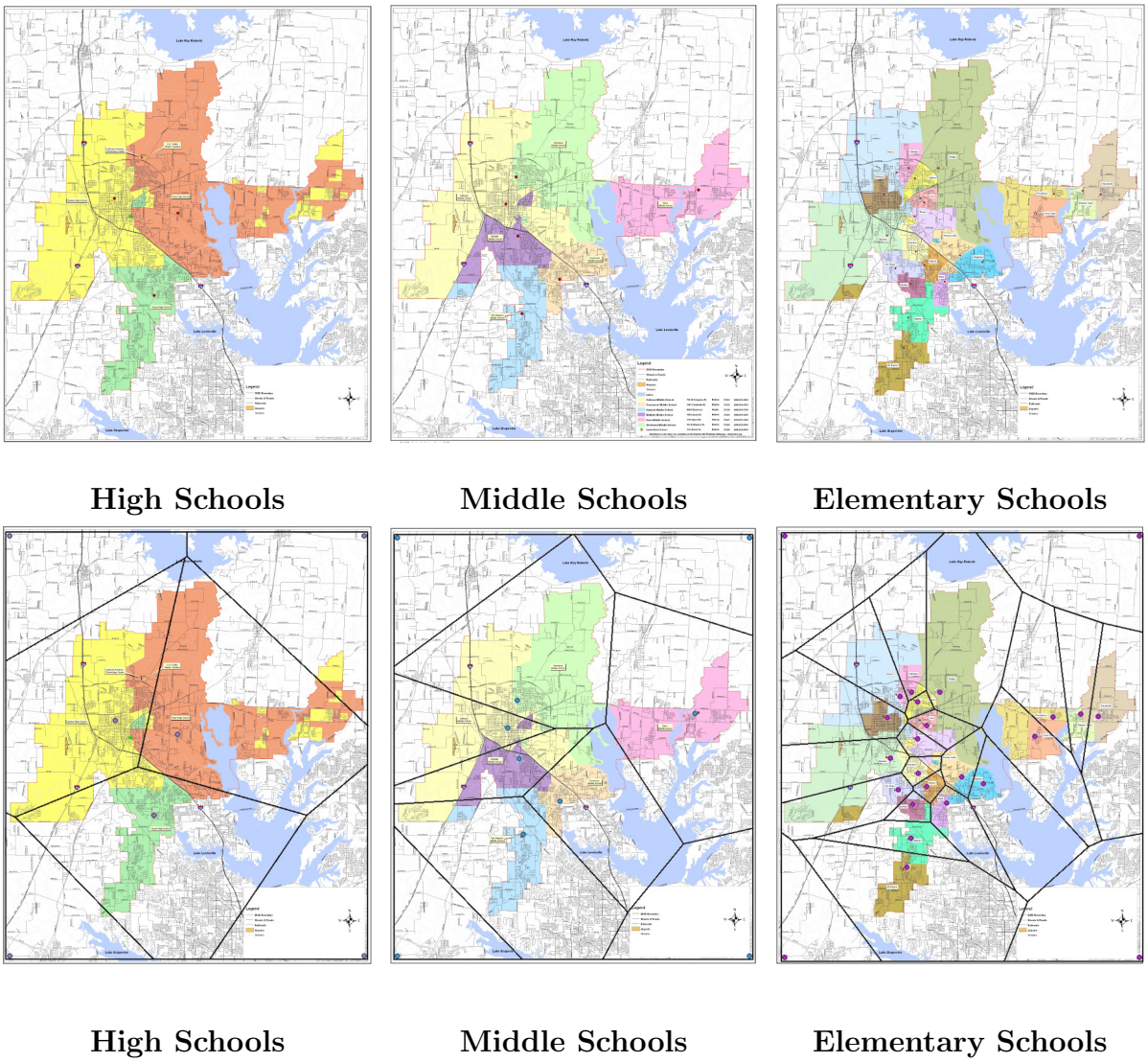


FIG. 4.3. Voronoi tessellation of Denton ISD attendance zones

corresponding streets for each school are described on Appendix A. The accuracy could potentially be estimated by comparing known SAZs maps and the Voronoi diagram. A comparison based on Denton ISD is shown in Fig. 4.3. The network of schools or MCN is the weighted graph represented in Fig. 4.4

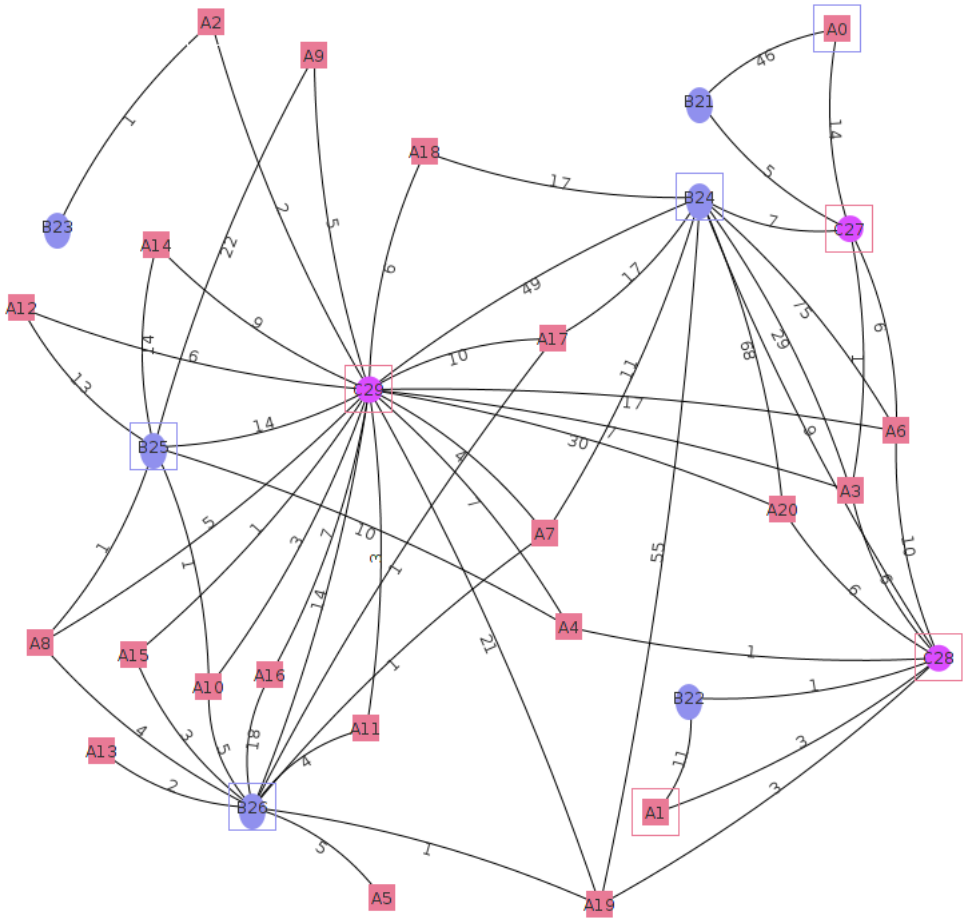


FIG. 4.4. Denton ISD, school type: A(Elementary), B(Middle), and C(High)

An example of the general output the algorithm is shown in Fig. 4.5. The left image depicts location of households, schools and their affiliation. The image on the right exposes the resulting school network. The output of the SAND algorithm is a graph weighted on

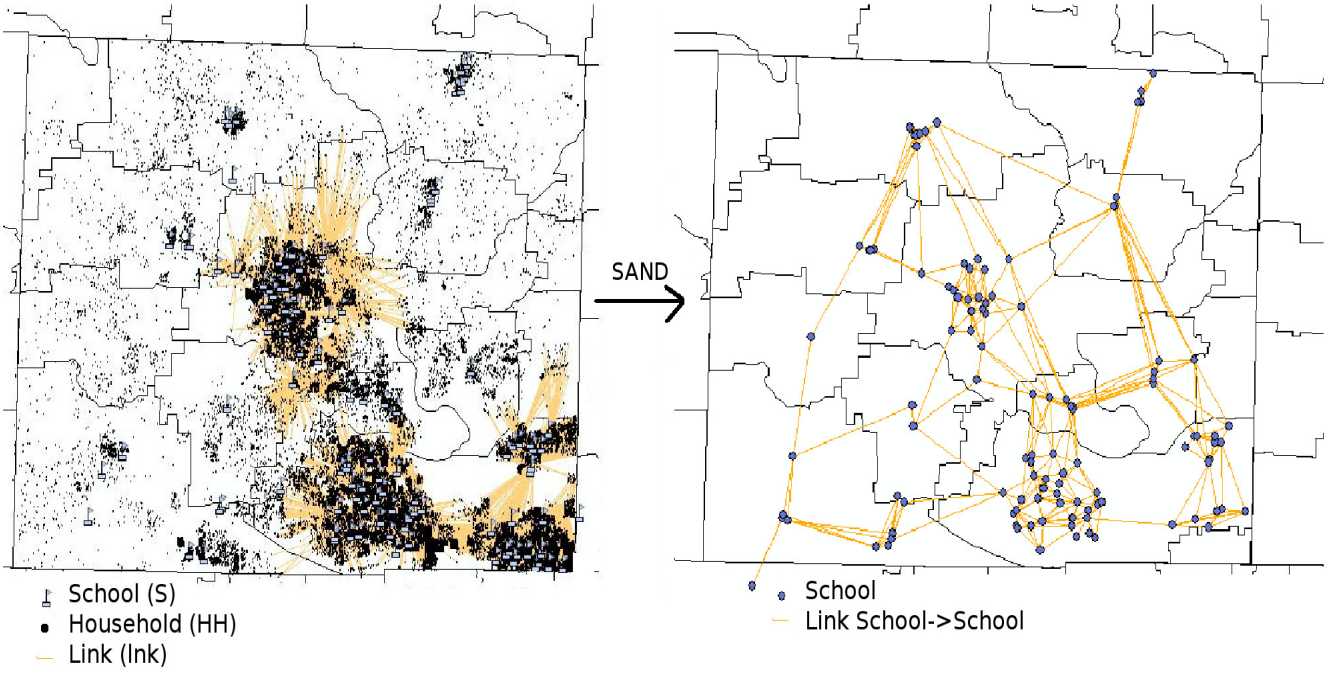


FIG. 4.5. An example of the synthetic reconstruction and the MCN derived from it through the SAND algorithm [4]

the vertices and edges. The weight of the vertices can be considered as an intra-community measure, whereas the weight of the outgoing edges could be considered as an inter-community measure. Graphs with a small number of nodes such those representing school districts are in the order of hundreds or even few thousands nodes (table 4.3). The run-time of standard algorithms such the shortest path calculation and other more complex such the vertex cover are reduced on MCNs.

4.3.2. Reduction of Complexity

It is important to notice the scale difference between a contact network and a resulting MCN. While the general population of a study area could reach the order of millions of agents, the cardinality of the graphs that are produced by the algorithm presented in this chapter is in the order of hundreds. The complexity reduction yields on the potential application of $NP-hard$ algorithms to analyze the structure of the resulting graph. Table 4.3 shows some

[3]This figure is reproduced from [124], with permission from Global Science and Technology Forum `http://www.globalstf.org`

examples of such reduction.

The reduction in complexity of the final graph representing the school network has a direct impact on the execution time of the algorithms applied to optimize mitigation strategies. Smaller graphs not only imply a shorter processing time, but also allow the application of approximation algorithms of NP-hard problems such as the vertex cover of a graph. The reduction in size of the social model for the state of Texas is shown in Table 4.4. Denton County and Denton ISD are shown in Table 4.5 and Fig. 4.4.

TABLE 4.3. U.S. general population and information on student population and number of schools on the independent districts [64], [128], [31]

Location	Population	Num. of HHs	Num. of Students	Num. of Schools
New York, NY	8175133	3047249	148980	305
Los Angeles, CA	3792621	1314198	667273	932
Chicago, IL	2695598	1033022	405644	642
Houston, TX	2099451	764758	204245	307
Philadelphia, PA	1526006	574488	166233	274
Phoenix, AZ	1445632	515701	34144	39
San Antonio, TX	1327407	461139	55116	101
San Diego, CA	1307402	474906	131785	225
Dallas, TX	1197816	449597	157162	242
Denton, TX	113383	39060	23994	38
Baltimore, MD	620961	238392	83800	196

4.4. Summary

MCNs constitute a notion of social networks and differ from other biology inspired networks in that they are not sparse. Additionally, dimension of MCNs is severely reduced

TABLE 4.4. State of Texas housing information, data source: US Census Bureau and Texas Public School Directory

Variable	Description	Value
P	Total Population	25,145,561
HH	Number of Households	8,922,933
S	Number of Public Schools	8,317

TABLE 4.5. Denton County Statistics, data source: US Census Bureau and Texas Public School Directory

Variable	Description	Value
P	Total Population	662,614
HH	Number of Households	240,289
S	Number of Public Schools	30
	Total Students	22,825
S	Number of Nodes Graph B	30
EE	Number of Edges Graph B	60
	Is Graph B Connected?	True

compared to contact networks. MCNs depict long-term relationships rather than probabilistic associations as opposed to large scale-semantic associations or food-networks. In the next chapter, the structure of MCNs is discussed.

CHAPTER 5

STRUCTURE AND CHARACTERISTICS OF MCNS

> However, not everything that can be counted counts, and not everything that counts can be counted
>
> *William Bruce Cameron, 1963*

In this chapter, the structure of Multi Co-affiliation Networks (MCNs) is analyzed. Additionally, the definition of centrality measures is stated and extended for weighted networks. The centrality measures of the MCN of Denton Independent School District (ISD) are calculated for betweenness, closeness and degree in Section 5.2. Section 5.3 states the known theorems for k-partite graphs as a framework for MCNs. The chapter concludes with a summary of the experimental results obtained.

5.1. Structure of MNCs

5.1.1. Number of Edges

In the most general form, the relationship between the number of edges and the number of vertices of a simple graph $G = (V, E)$ can be stated as follows:

$$|E| \leq \frac{|V| \times (|V| - 1)}{2} \tag{26}$$

Equation (26) bounds the maximum number of edges that can be contained in a simple graph. If equality occurs, the graph is called a complete graph.

In the case of bipartite graphs such as $G = ((V_1 \cup V_2), E)$, the maximum number of edges is represented by the relationship in (27)

$$|E| = |V_1| \times |V_2| \tag{27}$$

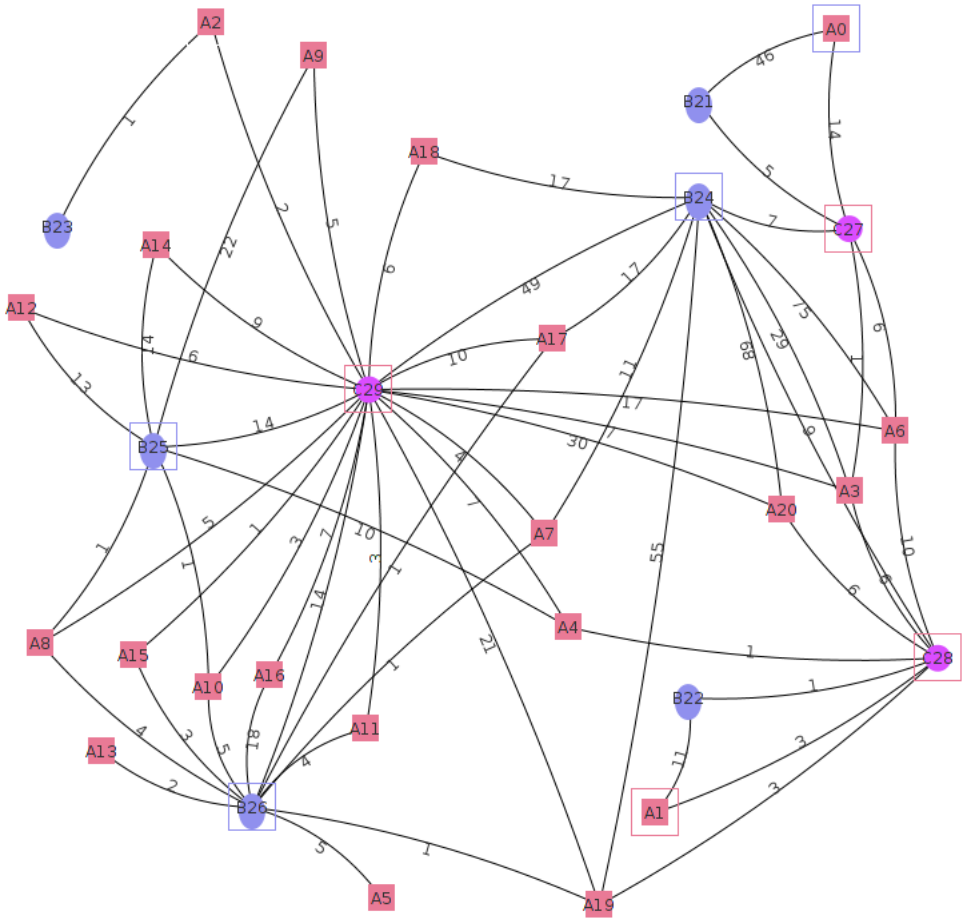


FIG. 5.1. Realization of a MCN

The generalization of (27) for complete $k-$partite graphs $G = (\cup_1^n V_i, i = 1, 2, 3 \ldots k, E)$ is stated in (28). Where V_i represents the ith partition of vertices, all partitions are of the same size $|V_i|$, and value of k represents the number of partitions

(28) $$|E| = \frac{|V_i|^2 \times (k)(k-1)}{2}$$

The maximum number of edges in a graph $B = ((E \cup M \cup H), EE) \in MNCs$ can then be directly derived from (28). Considering $|E| > |M| > |S|$.

(29) $$|EE| \leq \frac{|E|^2 \times (k)(k-1)}{2}; k = 3$$

It is necessary to observe that in real-life examples equality does not occur.

THEOREM 5.1. *For any graph* $B = ((E \cup M \cup H), EE)$ *generated following Algorithm 2* $|EE| < |E|^2 \times 3$*, with* $|E| > |M| > |H|$.

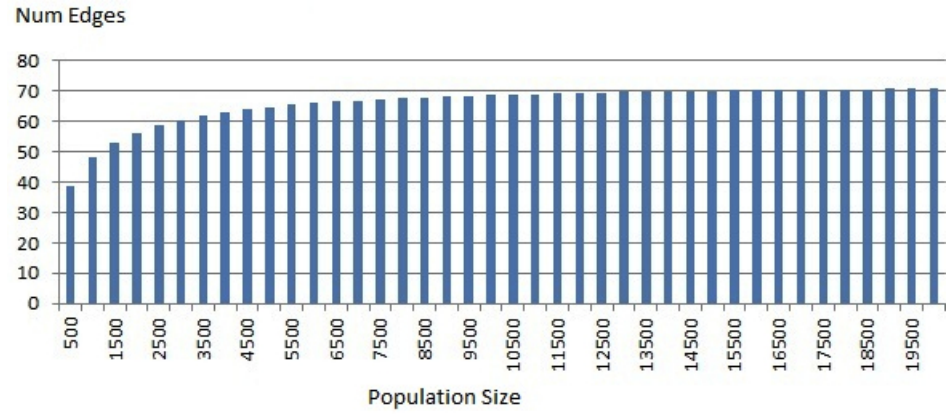
PROOF. The School Attendance Zones (SAZs) are defined in the two-dimensional space that represents the map associated with location $\mathcal{L}$ discussed in Chapter 3. Let e_i be the area representing the SAZ of an elementary school. e_i must intersect with at least one $m_j \in M$ and at least one $h_k \in H$ since the sets E, M, H are defined inside the same physical area. In these intersections, edges between schools are formed due to the existence of at least one hh_l that is affiliated to both SAZs. Indeed, the equality $|EE| = |E| \times |M| \times |H|$ holds if and only if each $e_j \in E$ intersects all $m_k \in M$ and all $h_l \in H$. Such intersection can only take place in a three-dimensional space if $|E| > 1, |M| > 1, |H| > 1$. Therefore, if $|H| = 1$, then maximum number of intersections can be represented as a bi-dimensional matrix with $|E| \times |M|$. □

5.1.2. Maximum and Minimum Degree

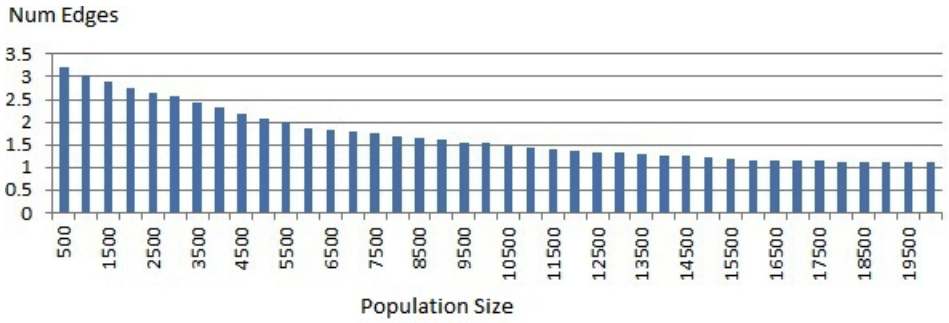
The adjacency matrix X of a graph $G = (V, E)$ is defined as a $|X| \times |X|$ matrix where the value of cell x_{ij} is defined as 1 if node v_i is connected to node v_j, and 0 otherwise. Weighted networks are denoted by W where w_{ij} represents the weight of the link between nodes v_i and v_j. If $w_{ij} \neq 0$ then node v_i is connected to node v_j, and the value represents the weight of the tie [126].

DEFINITION 5.2. The degree of a vertex $v_i \in V$, denoted by $D(v_i)$, represents the number of incident edges to v_i.

(30) $$D(v_i) = \sum_{j, j \neq i}^{|V|} x_{ij}$$



(A) Average over a 1000 simulations



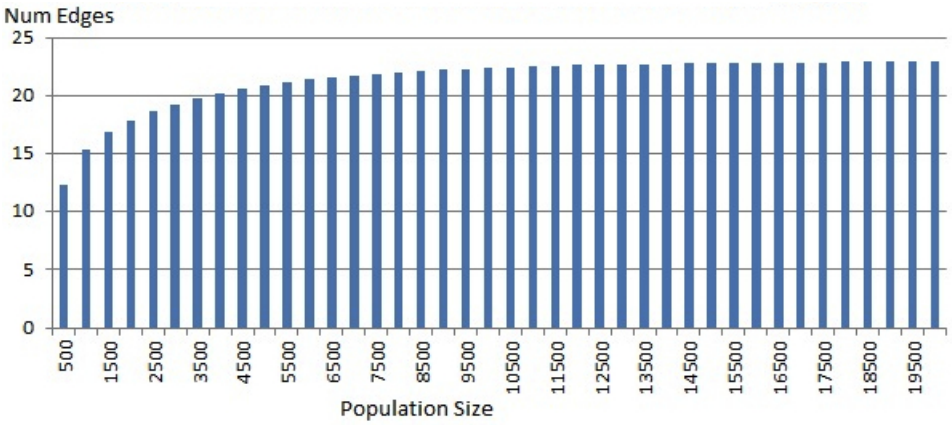
(B) STD over a 1000 simulations

FIG. 5.2. Number of households in sample and number of distinct edges of graph B.

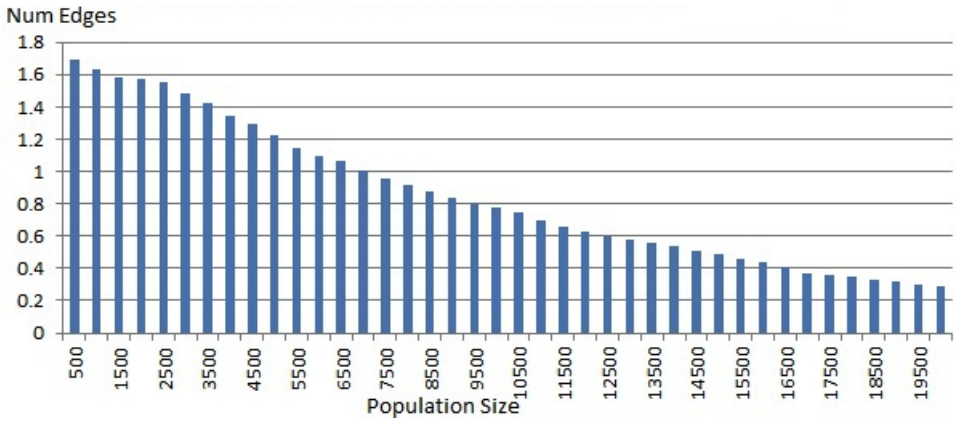
Likewise, on weighted networks the degree of a vertex is defined in (31).

$$Dw(v_i) = \sum_{j,j\neq i}^{|V|} w_{ij} \tag{31}$$

The maximum degree denoted by Δ stands for the vertex with the highest number of incident edges. On simple graphs Δ is bounded by $(|V|-1)$. Likewise, the minimum degree on simple graphs, denoted by δ stands for the vertex with the lowest (possibly 0) number of incident edges. In order to study the influence of population size over the properties of graph B, Algorithm 2 was executed varying the total number of households starting from 500 and ending in 20,000; with increasing steps of 500. At each step, 1000 simulations were conducted, making a total number of 40,000 executions. The number of edges, maximum



(A) Average over 1000 simulations



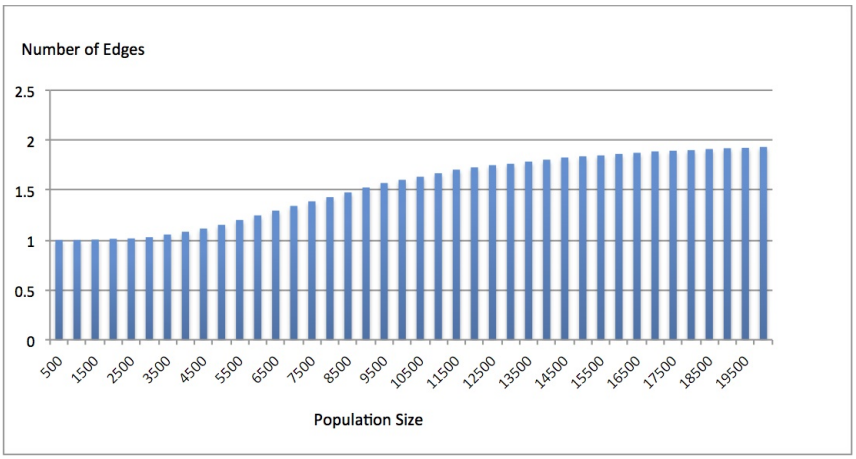
(B) STD over a 1000 simulations

FIG. 5.3. Number of households in sample and maximum degree of graph B.

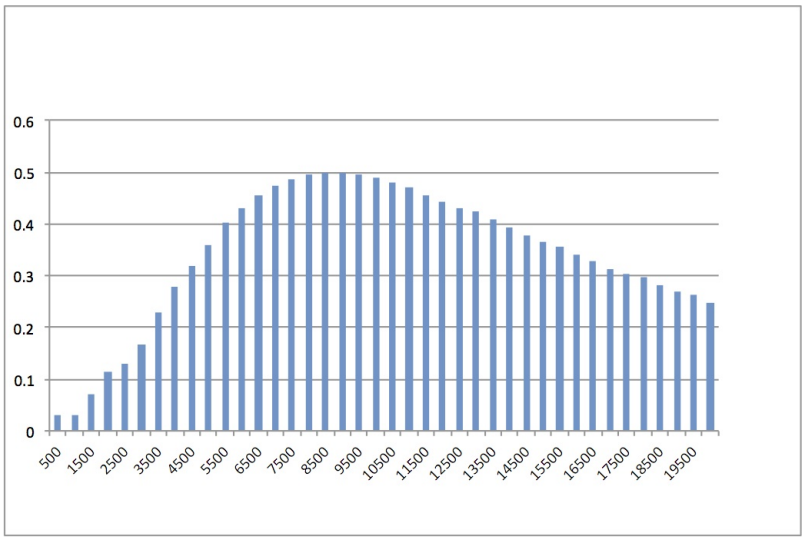
and minimum degrees were recorded. Fig. 5.2 shows the average and standard deviation (STD) of the results obtained for the number of distinct edges of graph B, changing the size of the population. Fig. 5.3 shows the average and standard deviation of the results for the maximum degree of graph B. Fig. 5.4 shows the average and standard deviation of the results obtained for the minimum degree of graph B, changing the size of the population.

5.2. Generalization of Centrality Measures for Weighted Networks

Centrality measures have interested the social network research community for its intrinsic linkage with quantifying relevance of nodes in a network. When first stated, three key characteristics were identified in order to consider a node important: number of edges, shortest paths traversing the node, and a small number of steps to reach other nodes. Freeman



(A) Average over a 1000 simulations



(B) STD over a 1000 simulations

FIG. 5.4. Number of households in sample and minimum degree of graph B.

[66] presented key characteristics of centrality and defined three different measures: degree, closeness and betweenness, defined for simple undirected unweighted graphs. Furthermore, numerous attempts have been made to extend the centrality definitions to a broader family of graphs. In [151], [152], and [126] new definitions for degree, closeness, and betweenness centrality were proposed. Centrality measures have a natural relevance for social network analysis. They have been used to identify and classify risk behaviors relevant to AIDS [135], to identify epidemic potential and vaccination effectiveness [88], traffic flows [23], strategic network formation [35], and even species extinction [4]. The applicability of centrality

measures to biology questions seems limitless.

5.2.1. Degree Centrality

DEFINITION 5.3. Degree centrality, denoted by C_D is the number of nodes to which a node is connected to. On the graph $G = (V, L)$ the degree centrality of a node v_i is calculated by (32).

$$C_D(v_i) = D(v_i) = \sum_{j;j\neq i}^{|V|} x_{ij} \tag{32}$$

In the case of weighted network, several definitions have been proposed to calculate C_D. In (33) a straight forward definition is proposed by using W instead of X [151].

$$C_{Dw}(v_i) = Dw(v_i) = \sum_{j;j\neq i}^{|V|} w_{ij} \tag{33}$$

Opsahl [126], proposed a new definition to accommodate the strength of a node given by the diversity of its neighbors. Equation (34) states the new definition for C_{Dw}.

$$\begin{aligned} C_{Dw}^{\alpha}(v_i) &= C_{Dw}(v_i) \times (\tfrac{C_D(v_i)}{C_{Dw}(v_i)})^{\alpha} \\ &= C_{Dw}(v_i)^{(1-\alpha)} \times C_D^{\alpha} \end{aligned} \tag{34}$$

The value α is a non negative tuning parameter $(0 - 1)$. It is important to notice that when $\alpha = 0$, C_{Dw}^{α} is equal to C_D, the degree centrality of the unweighted version of the graph. Simplicity and straight forward calculation are advantages of this centrality measure. Additionally, only local information relative to the node is required, which makes this measure specially usable when only partial information of the network is known.

5.2.2. Betweenness Centrality

A path in a graph $G = (V, L)$ between two vertices v_i, v_j is a sequence of pairwise vertices such $(v_i, v_h) \ldots (v_k, v_j)$. For a simple graph $G = (V, L)$ the shortest path between two vertices v_i, v_j is known as the geodesic path (GP). Let $\mathcal{P}(\S_{\rangle\langle})$ represents a path between nodes (i, h). Therefore, $GP(v_i, v_j) = min(\mathcal{P}(\S_{\rangle\langle})) + \ldots + min(\mathcal{P}(\S_{|||}))$, where h, k stand for

TABLE 5.1. Comparison of degree centrality values evaluated for graph B presented in Fig. 5.5

Num	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1.0$
0	B24=337.0	B24=139.869	C29=66.332	C29=36.423	C29=20.0
1	C29=220.0	C29=120.802	B24=58.052	B24=24.094	B26=11.0
2	A6=108.0	A6=47.379	B26=25.259	B26=16.669	B24=10.0
3	A20=104.0	A20=42.86	B25=22.913	B25=12.665	C28=8.0
4	A19=80.0	B25=41.454	A6=20.785	C28=11.887	B25=7.0
5	B25=75.0	B26=38.275	A19=17.889	A6=9.118	C27=5.0
6	A0=60.0	A19=37.83	C28=17.664	A19=8.459	A3=4.0
7	B26=58.0	C28=26.246	A20=17.664	C27=8.014	A6=4.0
8	B21=51.0	A0=25.637	A3=13.115	A20=7.279	A19=4.0
9	A3=43.0	A3=23.747	C27=12.845	A3=7.243	A4=3.0
10	C28=39.0	B21=22.695	A0=10.954	A17=5.244	A8=3.0
11	C27=33.0	C27=20.589	B21=10.1	A4=4.695	A7=3.0
12	A17=28.0	A17=16.019	A17=9.165	A0=4.681	A20=3.0
13	A9=27.0	A9=14.086	A4=7.348	A7=4.559	A10=3.0
14	A16=25.0	A16=13.296	A9=7.348	B21=4.494	A17=3.0
15	A18=23.0	A18=12.49	A16=7.071	A8=4.054	A2=2.0
16	A14=23.0	A14=12.49	A7=6.928	A10=3.948	A1=2.0
17	A12=19.0	A4=11.501	A18=6.782	A9=3.834	A9=2.0
18	A4=18.0	A12=10.822	A14=6.782	A16=3.761	B22=2.0
19	A7=16.0	A7=10.529	A12=6.164	A18=3.683	B21=2.0
20	A1=14.0	A1=8.607	A8=5.477	A14=3.683	A0=2.0
21	B22=12.0	B22=7.667	A1=5.292	A12=3.511	A18=2.0
22	A8=10.0	A8=7.401	A10=5.196	A1=3.253	A11=2.0

possible intermediate nodes (i.e. $h = k$, $GP(v_i, v_j) = 2$). $GP_i(v_j, v_k)$ denotes all the geodesic paths from v_j to v_k that have v_i as intermediate node.

TABLE 5.2. Comparison of betweenness centrality values obtained from graph B presented in Fig. 5.5

Num	*Summatory*	$\alpha = 0$	$\alpha = 1$
0	C29=279.0	C29=0.649	C29=0.642
1	B26=55.0	B26=0.131	B24=0.624
2	C27=54.0	C28=0.127	B26=0.364
3	C28=54.0	C27=0.126	B25=0.134
4	A3=52.0	A3=0.12	C28=0.127
5	A19=50.0	A19=0.12	C27=0.126
6	B24=29.0	B24=0.069	A0=0.067
7	A2=28.0	A2=0.064	A1=0.065
8	A4=3.0	A4=0.007	A2=0.064
9	A1=0.0	A1=0.0	A4=0.0
10	A6=0.0	A6=0.0	A3=0.0
11	A5=0.0	A5=0.0	A6=0.0
12	A8=0.0	A8=0.0	A5=0.0
13	A7=0.0	A7=0.0	A8=0.0
14	A9=0.0	A9=0.0	A7=0.0
15	B23=0.0	B23=0.0	A9=0.0
16	B22=0.0	B22=0.0	B23=0.0
17	B21=0.0	B21=0.0	B22=0.0
18	B25=0.0	B25=0.0	B21=0.0
19	A0=0.0	A0=0.0	A20=0.0
20	A20=0.0	A20=0.0	A19=0.0

DEFINITION 5.4. Simple graph betweenness centrality, denoted by C_B is the number of geodesic paths a node is part of. On the graph $G = (V, L)$, the betweenness centrality of a node v_i is calculated by (35).

(35) $$C_B(v_i) = \frac{GP_i(v_j,v_k)}{GP(v_j,v_k)}; \forall j, \forall k$$

For weighted networks, [119], and [151] have proposed a modified geodesic definition, expressed on (36).

(36) $$GP_w(v_i) = min(\frac{1}{W(v_i,v_h)} + \ldots + \frac{1}{W(v_k,v_i)})$$

Therefore, the length of the shortest path between two nodes is expressed on (37).

(37) $$GP_w^{\alpha}(v_i, v_j) = min(\frac{1}{(W(v_i,v_h))^{\alpha}} + \ldots + \frac{1}{(W(v_k,v_i)^{\alpha}})$$

The value α is a non negative tuning parameter $(0 - 1)$.

Definition 5.5. Betweeness centrality of a weighted network can be defined in terms of GP_w, based on the combination of the number of intermediate nodes in geodesic paths and the edges weights, as described in (38).

(38) $$C_Bw(v_i) = \frac{GP_w^{\alpha}i(v_j,v_k)}{GP_w^{\alpha}(v_j,v_k)}$$

5.2.3. Closeness Centrality

Definition 5.6. Simple graph closeness centrality, denoted by C_C, is the length of the paths from a node to all other nodes in the network. For simple graphs, (39) defines the C_c of v_i.

(39) $$C_C(v_i) = (\sum_{j;j\neq i}^{|V|} GP(v_i, v_j))^{-1}$$

For weighted networks, the definition proposed by [126] is described in (40).

(40) $$C_{Cw}(v_i) = (\sum_{j;j\neq i}^{|V|} GP_w^{\alpha}(v_i, v_j))^{-1}$$

TABLE 5.3. Comparison of closeness centrality values calculated for graph B presented in Fig. 5.5

Num	$\alpha = 0$	$\alpha = 1$
0	C29=0.141	C29=0.024
1	B24=0.14	B24=0.02
2	A6=0.135	A3=0.018
3	A20=0.133	A6=0.018
4	A19=0.131	B26=0.018
5	A3=0.124	A19=0.018
6	B26=0.122	A7=0.017
7	B25=0.117	A20=0.017
8	A18=0.114	A17=0.017
9	A17=0.114	A4=0.016
10	A7=0.103	B25=0.016
11	C28=0.102	A18=0.016
12	A9=0.102	A2=0.015
13	A14=0.102	A8=0.015
14	A16=0.102	C28=0.015
15	C27=0.095	A11=0.015
16	A4=0.094	A10=0.015
17	A12=0.093	A15=0.015
18	A0=0.081	A16=0.015
19	A8=0.08	C27=0.014
20	B21=0.077	A9=0.014

5.3. Connectivity on Graphs

Connectivity of a graph $G = (V, L)$ is defined as having a path among any two vertices $v_i, v_j \in L$ [165]. There are two types of connectivity, vertex and edge. Vertex connectivity, represented by κ, and edge connectivity represented by λ stand for the minimum number of

vertices or edges whose removal produces a disconnected graph.

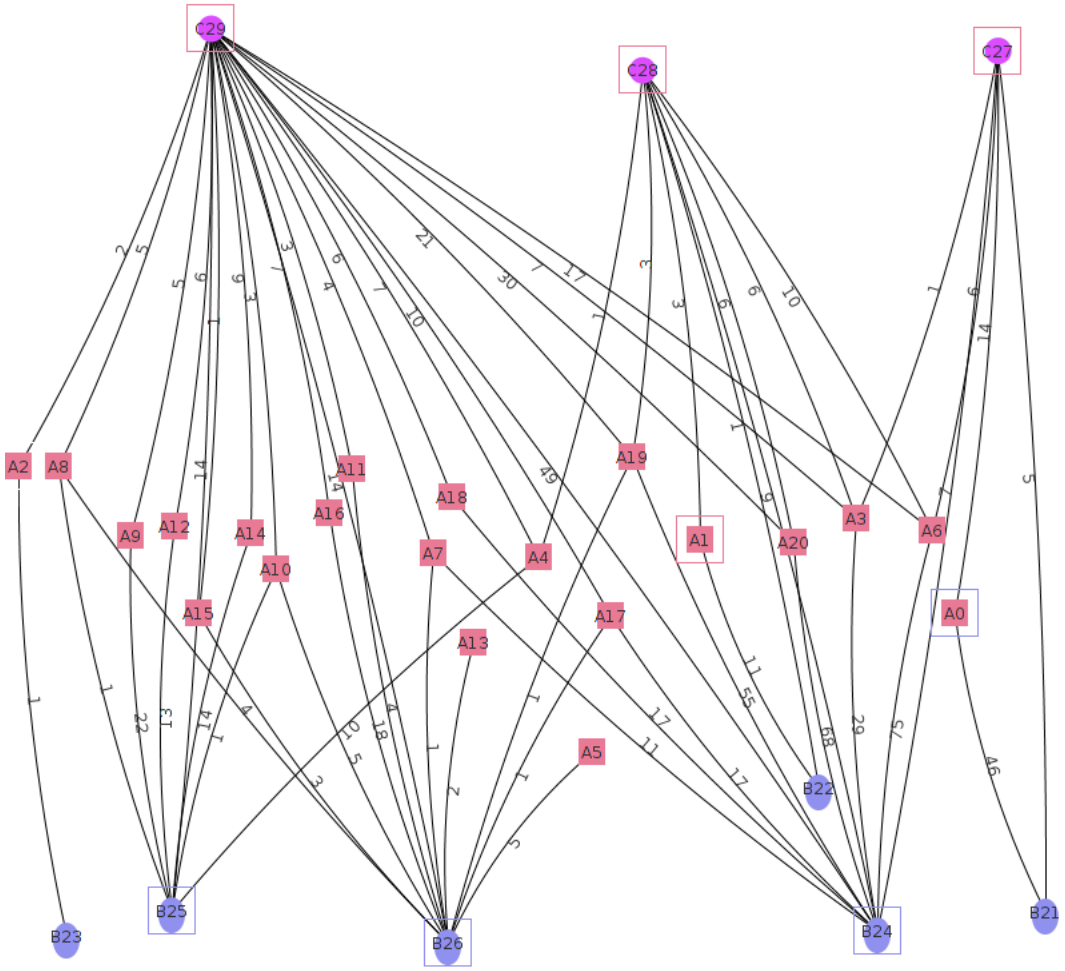


FIG. 5.5. Realization of a MCN, arranged as a $k-Partite$ graph

THEOREM 5.7 (Edge-Connectivity Version of Menger's Theorem). [165] *Let $G = V, L$ be a simple graph and v_i and v_j two distinct vertices. Then the minimum number of edges whose removal disconnects v_i and v_j is equal to the maximum number of pairwise edge-independent paths from v_i and v_j.*

LEMMA 5.8. *The maximum number of edges $|EE|$ that graph B can have without being connected is $|EE| = |E-1| \times |M| \times |H|$*

THEOREM 5.9. *(Edge)Connectivity* $\lambda(G)$

The maximum (edge) connectivity of G [80] *is defined as:*

$$\lambda(G) = \begin{cases} 0 & \textit{when } |E(G)| < |V(G)| - 1 \\ \textit{(2—E(G)—) /V(G)} & \textit{when } |E(G)| \geq |V(G)| - 1 \end{cases}$$

Nevertheless, for graph B, $|EE| >>> |S|$, and for highly populated zones $\delta >>> |S|$, $\delta(G)$ being the minimum degree of graph B.

THEOREM 5.10 (Edge-connectivity for k-partite Graphs). [155] *The edge connectivity* $\lambda(G)$ *of a* $k-$*partite graphs is*

$$\lambda(G) = \begin{cases} \delta(G) & \textit{If } |V| \leq \frac{(2k\delta)}{(k-1)} - 2 \\ & \textit{or in special cases} |V| \leq \frac{(2k\delta)}{(k-1)} - 1 \end{cases}$$

It is evident that when $\delta \leq |S|$ then number of components for graph B is exactly $|H|$. This case exemplifies an interesting structure for the school system. Under a realistic environment, it implies that every attendance zone for a lower lever school is completely contained in the corresponding next level school attending zone. Hence, while this case may exist under some circumstances, the sampling has shown that the contrary is the rule. Attendance zones for elementary schools are rarely completely contained in middle school attendance zones, and middle schools attendance zones are rarely contained in those of high schools. In particular, the set of households that belong to the intersections of the SAZs seem more interesting for the analysis as they provide bridges among different school levels.

5.4. Summary

MCNs are described in terms of number of different edges, maximum, and minimum degree. Additionally, centrality measures for MCNs are defined, following the definition for centrality measures on weighted graphs. Finally, the connectivity of MCNs is described.

CHAPTER 6

OPTIMIZATION OF INTERVENTION STRATEGIES

> In retrospect, it appears obvious that social network theory is a natural paradigm for understanding infectious disease transmission
>
> *Richard B. Rothenberg, 1995*

Mathematical and Computational Models (MCM) have been widely acknowledged in epidemiology. In the absence of data or when information is not reliable, MCMs have provided public health officials with insights and decision support scenarios impossible to obtain otherwise. MCMs construct a rational framework for the analysis of the economic impact of infectious diseases control measures [82]. Influenza pandemics have been widely studied through the use of MCMs [153], in communities [105], and schools [78]. In this chapter the efficiency of different intervention strategies applied to control the spread of disease in a school system is studied. The rational behind the evaluation of risk is based on centrality and network measures of the multi co-affiliation network of households and schools.

In the next section a detailed explanation of the implementation of the framework is addressed. The main sources of information for the disease parameters are also mentioned. Section 6.2 describes the baseline used for comparison of the strategies; additionally, the optimization considerations are listed. Section 6.3 describes the methodology used to compare the efficiency of the strategies proposed by this research. The chapter concludes with a summary of the results and the conclusion.

6.1. Implementation

The implementation of simulation scenarios requires the interconnection of three modules. First, a module that simulates disease spreading over a sample of the synthetic population of Denton County generated using the synthetic reconstruction. Second, an intervention module that allows the application of different strategies over the precise same population to obtain comparable outputs. Finally, a classification module that classifies the output of

the simulation module.

In order to establish the efficacy of intervention strategies the initial step is to define an outbreak or epidemic. One can define the concept of epidemic as a form of a self-sustainable disease contagion among a population. For public health officials, an outbreak is an incident that is significantly above a baseline or above a specific threshold [79]. To mathematically define an epidemic, the concept of R_0 discussed in previous chapters becomes crucial. By allowing the simulation parameters to be probability distributions as opposed to fixed values, computational models become a powerful tool in the estimation of R_0. Under these assumptions, identical initial parameters previously used to define R_0 may yield different results (i.e. an incident resulting in an epidemic vs no outbreak) caused by the stochasticity of the process. Therefore, R_0 is calculated as an average of several executions of a particular algorithm to ensure consistency of the results.

6.1.1. Simulation Parameters

The SIRS compartmental model discussed in Chapter 2 was used to simulate the spread of disease. The simulation parameters are described next. The assumption that initial parameters are fixed values is expanded; and therefore, a subset of the parameters are described in terms of a probability distribution. This extension of mathematical programming is refereed to as stochastic programming [147].

- Transmission rate ($\mathcal{T}$) is a value between [0 - 1] that represents the probability that given a contact, the disease is passed on from an household (HH) currently infectious (HH with status="I") to an HH in susceptible state (HH with status="S").
- Radius (∇) or neighborhood size is the ratio between the size of the area of action of a household (HH) compared to the total size of the study area. To build the social network, the neighbors of a selected HH are chosen based on the radius of action of a HH.
- Duration of I ($\mathcal{DI}$) in days. $\mathcal{DI}$ is a discrete distribution of period lengths and probability related to the duration of a HH in I state

- Duration of R ($\mathcal{DR}$) in days. $\mathcal{DR}$) is a discrete distribution of period lengths and probability related to the duration of a HH in R state
- Intervention size (ι) is the percentage of HHs removed at the beginning of the simulation as part of the mitigation strategy.
- P value is a value between [0 - 1] that represents the proportion of neighbors of each HH selected from the same school system.
- α is a non-negative tuning parameter $(0 - 1)$ for centrality measures in weighted graphs.

The simulation parameters are provided as a flat file representing the pair $[parameter, value]$.

TABLE 6.1. Baseline disease parameters

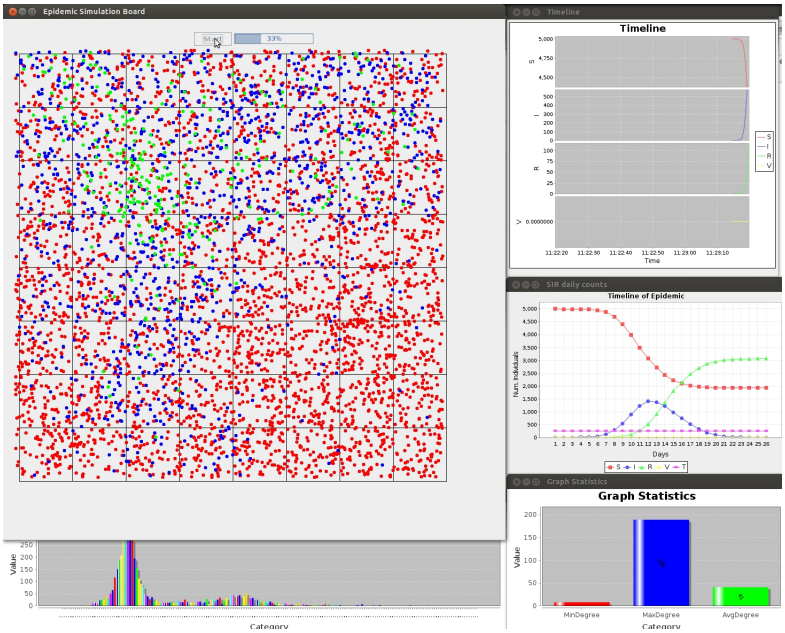
Variable	Value
Trans. rate ($\mathcal{T}$)	0.2
Radius (∇)	1.18%

	(I) Period (τ) [53]				(R) Period (γ)		
Number of day	3	4	5	6	224	225	226
Probability (all HHs)	0.3	0.4	0.2	0.1	0.2	0.3	0.5
Mean days	4.1		225.3				

The simulator uses a SIRS model. During the simulation process, individuals may return to the susceptible state after a recovery period. It is possible to model different disease models (i.e. SI, SIR) using the architecture and implementation of the simulator. Each incident may be executed for up to 5000 timesteps, which allows endemic events to be identified. Additionally, a recovery period of 225.3 timesteps in average, allows the case of seasonal diseases in which an individual may become ill the next year.

6.1.2. Simulation Output

The Java implementation of the simulation module includes two possible outputs; a graphical representation of the possible outbreak, and a text file that contains the statistics of the incident. A sample of the graphical interface is shown in Fig. 6.1.



(A) Simulation module graphical output

FIG. 6.1. Simulation module output

6.1.3. Outbreak Classification Algorithm (OCA)

The infection process in the SIR model may not always produce an outbreak. Therefore, scenarios for which outbreaks occur have to be identified, in order to compare the efficiency of intervention strategies. The outbreak classification algorithm (OCA) is an unsupervised classification algorithm for the output of the simulation module. Its objective is to identify epidemic events from incidents in which the epidemic threshold was not reached. The execution of the algorithm produces the immediate and unsupervised classification of the results of multiple different simulation scenarios.

The parameters for the classification are set at the beginning of an scenario and then several hundred runs are executed for the same parameters. Scenarios that are classified as outbreaks are then evaluated to estimate the efficacy of the different intervention strategies. The OCA requires the definition of an outbreak. The threshold could be set to zero but in general the threshold is defined to match the public health goal of intervention. The attack rate of H1N1 influenza during 2009 was calculated to be between 5% and 15% [127]

therefore, an outbreak might be expected if community attack rate of 15-25% is reached[79]. The evaluation of all intervention measures are sensitive to this definition and consequent classification of what constitutes an epidemic. Algorithm 3 is used in order to classify each output of the simulation scenario. To set the parameters for the OCA, a disease similar to influenza is considered. The CDC has used 10% as an informal marker for the beginning and end of the flu season [60], and therefore, the threshold is set to 10%. Reported cases of Influenza Like Illness (ILI) are monitored throughout year around, but in the United States the flu season typically begins to increase in late December or early January and peaks in February most commonly [60]. The algorithm takes timesteps that are equivalent to one day and therefore, the window of time for monitoring is the entire year or 365 days.

Algorithm 3: Outbreak Classification Algorithm (OCA)

Data: SIR DUR, TOT_I, TOTAL_POPULATION

Result: OutbreakClassification = [Epidemic, Endemic, NoOutbreak]

initialization;

read input;

OutbreakClassification = Epidemic;

if $DUR > 365 \wedge TOT\_I < 0.15$ **then**

    OutbreakClassification = Endemic;

else

    if $TOT\_I < 0.1 \times TOT\_POPULATION$ **then**

        OutbreakClassification = NoOutbreak;

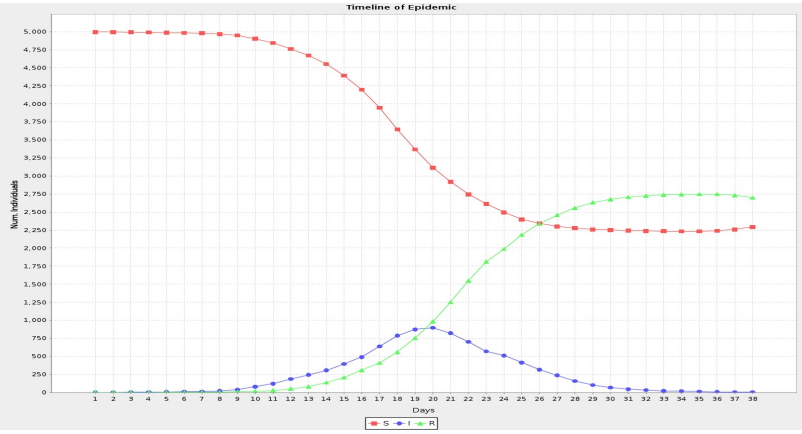
    else

        OutbreakClassification = Epidemic;

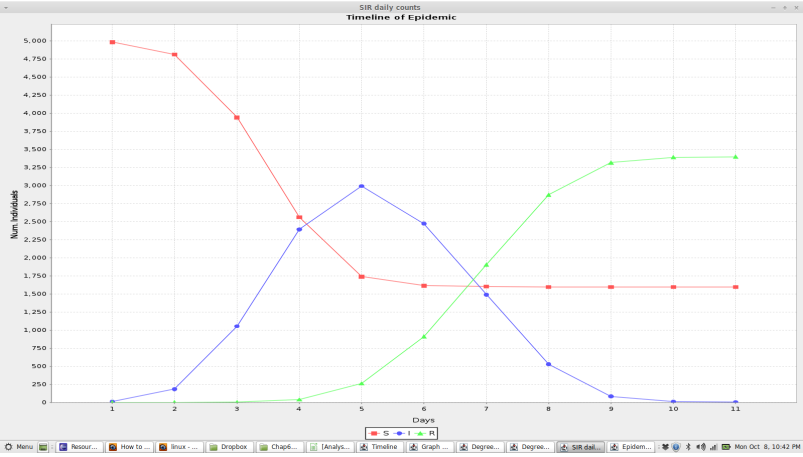
    ;

Events are represented as plots of the total count for individuals in states S, I, and R. Classification of these curves are used to illustrate how the algorithm identifies epidemic events. Incidents classified as endemic are shown in Fig. 6.2, epidemic events Fig. 6.3, and

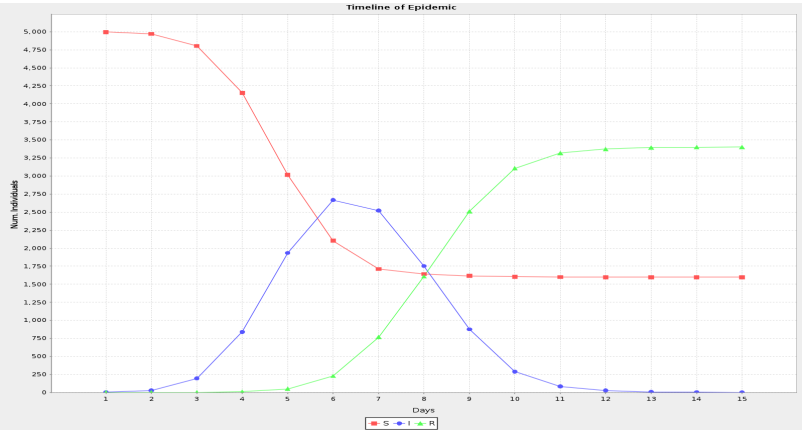
no outbreak Fig. 6.4.



(A) $\mathcal{T} = 0.04$, $\nabla = 1/20$, $\mathcal{DR} = \{24(0.2), 26(0.6), 27(0.2)\}$, SimID = -951589224

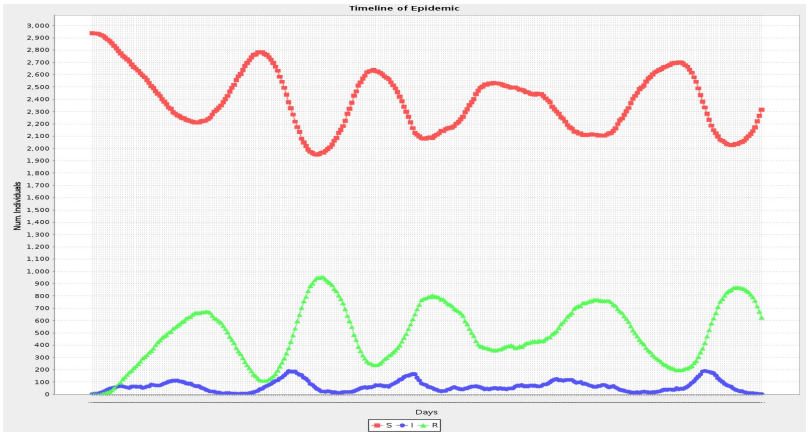


(B) $\mathcal{T} = 0.1$, $\nabla = 1/8$, $\mathcal{DR} = \{24(0.2), 26(0.6), 27(0.2)\}$, SimID = -1638067850

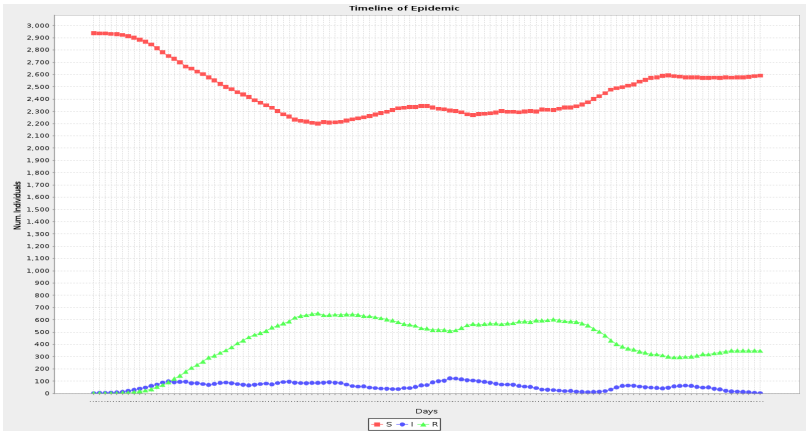


(C) $\mathcal{T} = 0.07$, $\nabla = 7/80$, $\mathcal{DR} = \{24(0.2), 26(0.6), 27(0.2)\}$, SimID = 2092624704

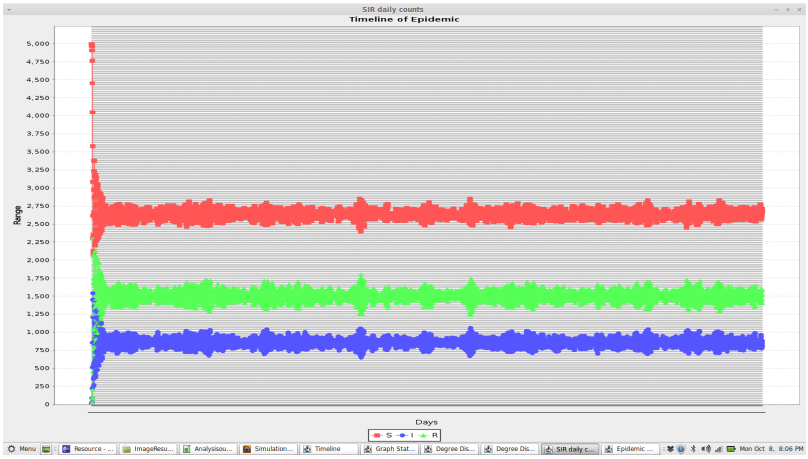
FIG. 6.2. Outbreaks classified as *epidemic*



(A) $\mathcal{T} = 0.07$, $\nabla = 7/80$, $\mathcal{DR} = \{24(0.2), 26(0.6), 27(0.2)\}$, SimID = 2092624704

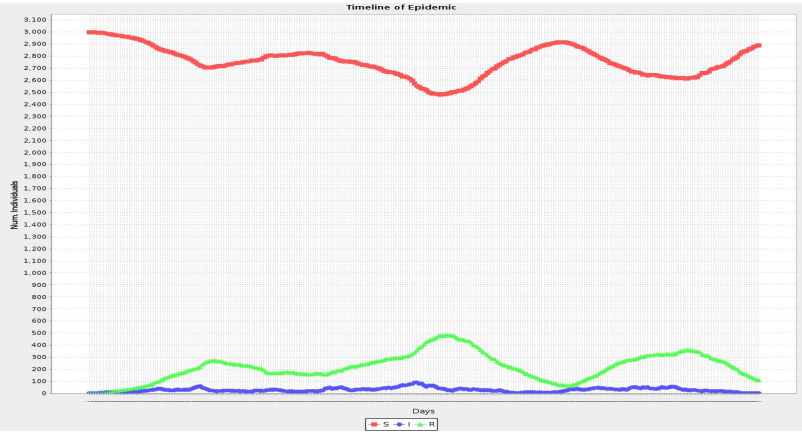


(B) $\mathcal{T} = 0.04$, $\nabla = 1/8$, $\mathcal{DR} = \{24(0.2), 26(0.6), 27(0.2)\}$, $\iota = 0.02$, SimID = -1193959466

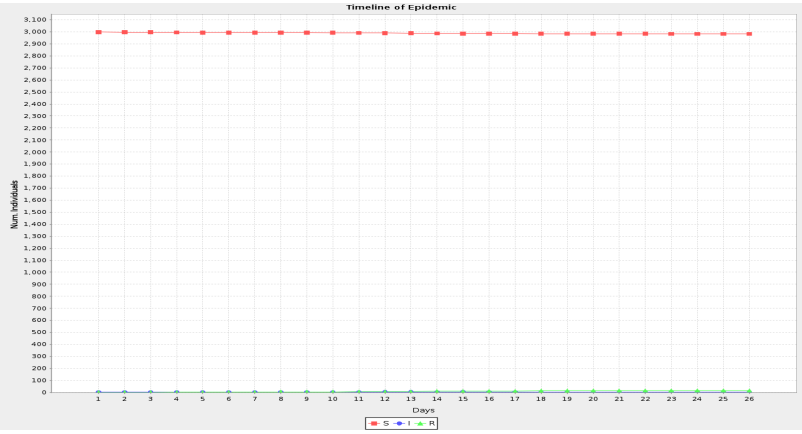


(C) $\mathcal{T} = 0.04$, $\nabla = 1/8$, $\mathcal{DR} = \{4(0.2), 6(0.6), 7(0.2), \}$, SimID = -951589224

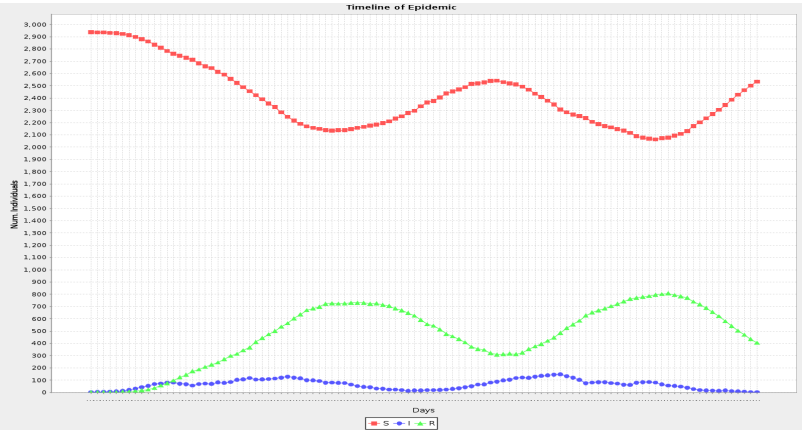
FIG. 6.3. Outbreaks classified as *endemic*



(A) $\mathcal{T} = 0.03$, $\nabla = 1/20$, $\mathcal{DR} = \{24(0.2), 26(0.6), 27(0.2), \}$, SimID = 2134557307



(B) $\mathcal{T} = 0.025$, $\nabla = 1/20$, $\mathcal{DR} = \{24(0.2), 26(0.6), 27(0.2), \}$, SimID = 2134557307



(C) $\mathcal{T} = 0.04$, $\nabla = 1/20$, $\mathcal{DR} = \{24(0.2), 26(0.6), 27(0.2)\}$, SimID = -1429538713

FIG. 6.4. Outbreaks classified as *noOutbreak*

Table 6.2 lists output variables calculated for each simulation in the scenario. R_0^{exp} and R_*^{exp} are defined as follows:

$$R_0^{exp} = I_{t=2} - I_{t=1} \tag{41}$$

R_0^{exp} represents the average number of secondary infections caused by a initial infectious household in a totally susceptible population.

$$R_*^{exp} = \frac{\sum_{t=1}^{n}(I_{t+1} - I_t)}{n}; iff I_{n+1} \leq I_t \tag{42}$$

R_*^{exp} represents the average number of secondary infections per time-step caused by the infectious population of households of the previous time-step during the exponential phase of the incident.

TABLE 6.2. SIR Output Parameters

Acronym	Variable	Description
DUR	Duration (days)	Number of timesteps with $\|I\| > 0$
TOT_I	Total I (number)	Total number of "I" in the incident
MAX_I	Maximum I (number)	Max. number of "I" per timestep over the entire incident
AVG_I	Average I (number)	Avg. number of "I" per timestep over the entire outbreak
THR_I	Threshold I (percentage)	Min. % of "I" population to be classified epidemic
R_0^{exp}	Basic reproduction number	Equation (41)
R_*^{exp}	Reproduction number	Equation (42)

6.2. Baseline Analysis

Before executing any simulation, tuning of the simulator is required in order to perform a basic mapping of the search space produced by the combination of the simulation

parameters. For the simulation module, the parameters that determine the type of incident are transmission rate ($\mathcal{T}$) and radius (∇). Initially, β represented the average rate of contacts between individuals. In the proposed model $\nabla \times |HH|$ is the average number of neighbors of one household or contact rate, approach similar to [88]. Fig. 6.5 depicts the combination of contact rate and $\mathcal{T}$ for which incidents classified as epidemic occur.

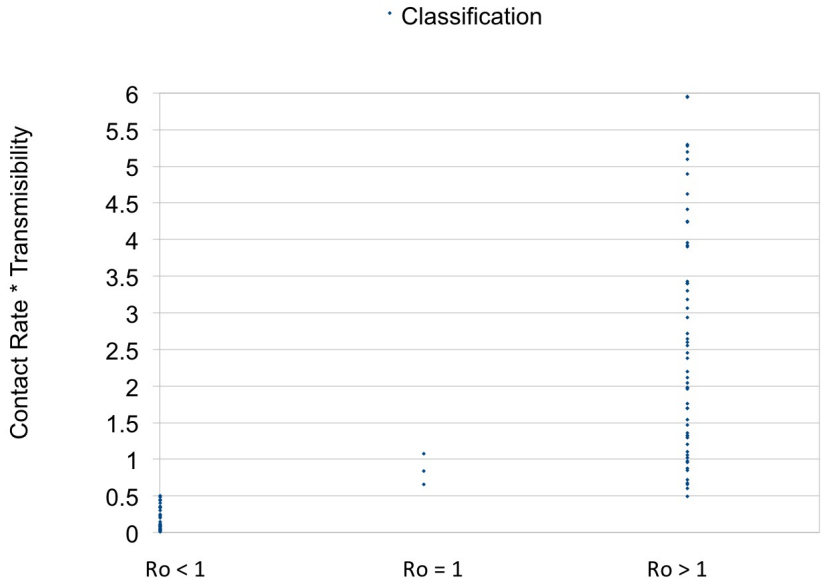


FIG. 6.5. Transmission and contact rate combination and incident classification type

In order to maximize the probability of producing an outbreak, the simulator is tuned for the ranges $\nabla = [0.20\% - 1.77\%]$ and $\mathcal{T} = [0.2 - 0.6]$. This range coincides with infection attack rates estimated from age-structured population models and the 1957 pandemic [72].

6.2.1. Intervention Strategies

Mitigation strategies can be implemented following different assumptions of how a specific disease propagates. For instance, ring vaccination was used to eradicate smallpox [96]. Ring vaccination effectiveness was based on the fact that in close contact rings of the population the probability of transmission is five times higher than casual contact ring. In

the case of most infectious diseases, physical proximity is necessary for propagation. Initial interventions have a different impact on the disease dynamics than later measures. On a completely susceptible population, a disease may grow exponentially if it survives early extinction [96]. Intervention strategies are studied in order to estimate their potential to reduce infectious diseases' morbidity and mortality. With the use of synthetic populations, pharmaceutical interventions consisting of antiviral treatment and prophylaxis can be studied and experimented with. Non-pharmaceutical interventions such as household isolation, quarantining of household contacts, closure of schools, and social distancing in the workplace and the community may be explored without any risk for the actual population [78]. School closures have been an effective method for ILI-like diseases containment [112], [36], and [52]. This research studies school closure optimization based on the cost associated with closing schools and its epidemic prevention potential (EPP) [78]. The EPP is defined in (43):

$$EPP = 1 - \frac{Pr(e)_A}{Pr(e)_B} \tag{43}$$

$Pr(e)_A$ represents the probability of an epidemic utilizing strategy A (i.e. centrality-based selection) and $Pr(e)_B$ is the baseline scenario (i.e. random vaccination). One quantitative way to compare intervention strategies is to calculate their EPP values compared to the baseline of complete absence of intervention measures. A different approach is to evaluate the impact that the intervention strategy has on the dynamics of the disease. To quantify the effect that a particular mitigation has on the disease dynamics, the total number of infections and duration of the outbreak are compared.

6.2.2. Optimization Considerations

State Departments of Health and other organizations define policies and control guidelines in order to assist school district staff members in their efforts to preserve and protect the health of both students and employees. Local health officers are encouraged to take whatever action deemed necessary to control or eliminate the spread of the disease. Those actions include the following [18]:

- Close the affected school(s).
- Close other schools in the local health officers jurisdiction.
- Cause the cessation of selected school activities or functions.
- Exclude from school attendance any students, staff, and volunteers who are infected with or at higher risk to contract the disease

Schools are not isolated entities, but rather form a network and a community in conjunction with households. If an event that can be considered an emergency occurs, several paths of actions can be taken with respect to closing schools:

- Close all the schools at the beginning of the event, which represents minimal risk, but at the same time is the most cost-effective measure.
- Close schools when suspected or confirmed case is found.
- Close schools following some other heuristic approach.
- Close schools by proximity with those with confirmed cases.

To provide a decision support tool, it is imperative to compare different scenarios and their estimated result in a quantitatively way. Additionally, the optimization methodology parameters should be measurable and based on real world restrictions such cost.

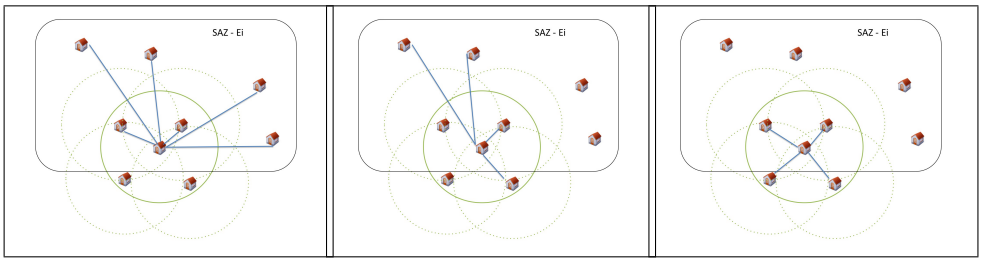
Optimization of intervention strategies in contact networks may come from two main scientific fields. SNA methodologies involve defining and applying centrality measures of contact networks in order to identify high-risk individuals [36], [92], [135] or groups [38], [46], [140] in the network. The most central nodes or nodes with the highest centrality measure are discovered through the use of algorithms that can be executed for networks with millions or tens of millions of nodes. Complete graph theoretical approaches to optimization with respect to the graph structure have been almost non-existent. Instead, graph theoretical models are used to mathematically derive the final number of infections, total duration of an outbreak for different types of graphs [87], [91], clustered networks [113], and the reproduction number [77], [129].

MCNs represent an interesting hybrid approach because they allow to leverage optimization methods found in both fields. Whereas the disease simulation may account for

all eligible households in a particular region, the MCN representing the risk network can be considered small; for example, in the order of tens of nodes in the case of Denton County, only few thousands in the case of New York city. The problem is to try to identify the minimum number of schools to close in order to stop or at least delay the spread of disease. In terms of graph theory, this problem can be stated as given a graph $G = (V, L)$, find the minimum number of vertices $C \in V$ such that the disease can not traverse the network. One possible solution to the problem could be to identify a subset vertices, $C \subseteq V$, such that each edge $e \in L$ has at least one endpoint in C and remove those from the network. The set of nodes or vertices such that each edge of the graph is incident to at least one vertex in the set is known as a Vertex Cover (VC) of the graph. The problem of finding the minimum vertex cover is a classical optimization problem and its complexity has been shown to be NP-complete [89]. Approximation algorithms are used to find near-optimal solutions. To addressed the problem one strategy used in this research work is based on the vertex cover of the MCN.

6.3. Methodology

The SIRS simulation takes place over a household sample of size $N = 5000$ taken from synthetic reconstruction of Denton County.



(A) $p = 1$ (B) $p = 0.5$ (C) $p = 0$

FIG. 6.6. Values of parameter p and neighbourhood assignment

Households are assigned a neighborhood of a predetermined radius and corresponding schools when eligible. Each household is assigned an initial state of S. The contact network is formed by houses in the neighborhood (NEI) of each household and households that are

assigned to the same school (SCH). The value p represents the percentage of neighbors selected from SCH and $1-p$ represents of neighbors selected form NEI. Fig. 6.6 exemplifies the selection process for the neighborhood of a HH, the green circle of radius ∇ represents the houses in the set NEI, the black rectangle demarcates the SAZ and houses assigned to the same school SCH. Houses that belong to NEI and SCH have an increased probability of being selected. The neighborhood relation is considered to be reciprocal.

6.4. Results

6.4.1. Proactive Approach

A proactive approach is defined as the estimation of the optimal locations to intervene before an event has occurred. Once the initial case is identified at $t = 1$ set of households is removed and set to state R. The possible selection scenarios are detailed in Table 6.3.

Table 6.3. SIR Output Parameters

Strategy	Description
Complete System	All schools are closed and households moved to state R
Betweenness Centrality	Only a set of schools are closed, based on their betweenness centrality value
Closeness Centrality	Only a set of schools are closed, based on their closeness centrality value
Degree Centrality	Only a set of schools are closed, based on their degree centrality value
Set Cover	Only the schools belonging to the set cover are closed
Naive	Select households randomly and set their state to R
None	No household is ever moved to state R

The null strategy or "None" provides a baseline for comparison. The naive approach as a baseline allows to understand the difference between having a targeted intervention versus random selection of households for removal. Fig. 6.7 summarizes the overall results

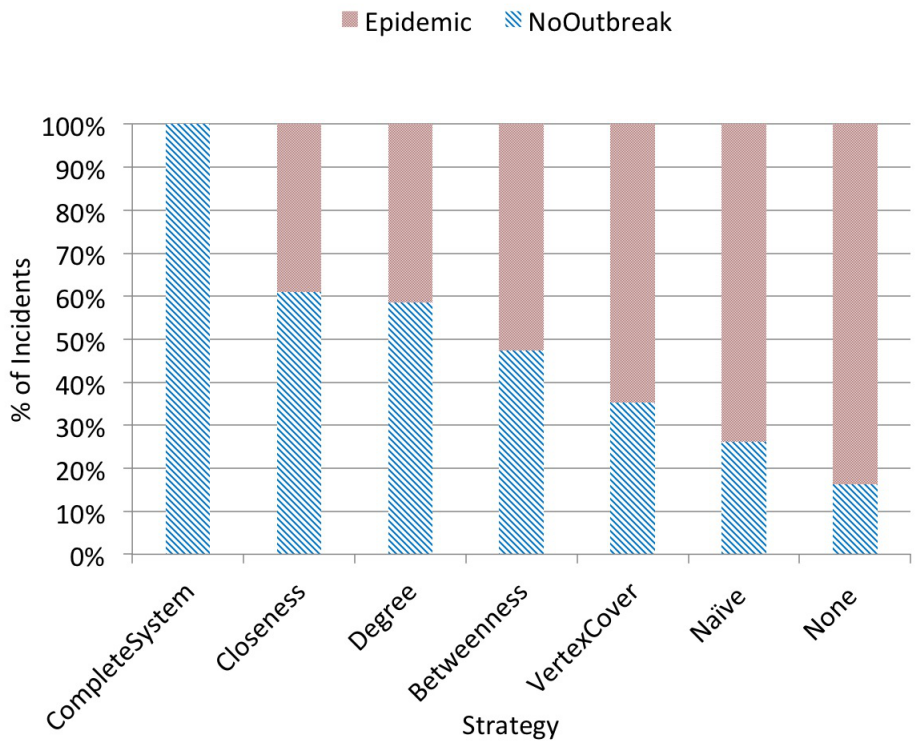
of all strategies preventing incidents (Fig. 6.7a) and modifying disease dynamics (Fig. 6.7b). The following results are obtained by analyzing a scenario in which the disease primarily disperses over the school system and not outside of it. The complete system intervention, the closure of all schools at the beginning of an incident, achieves 100% of efficacy but this is the most expensive intervention.

Fig. 6.7b compares the potential of each strategy to reduce duration of the epidemic and the average number of infectious per day.
Fig. 6.8 shows the strategies based on MCNs compared to different baselines, in Fig. 6.8a the baseline is the null intervention and in Fig. 6.8b the baseline is the average of random interventions.

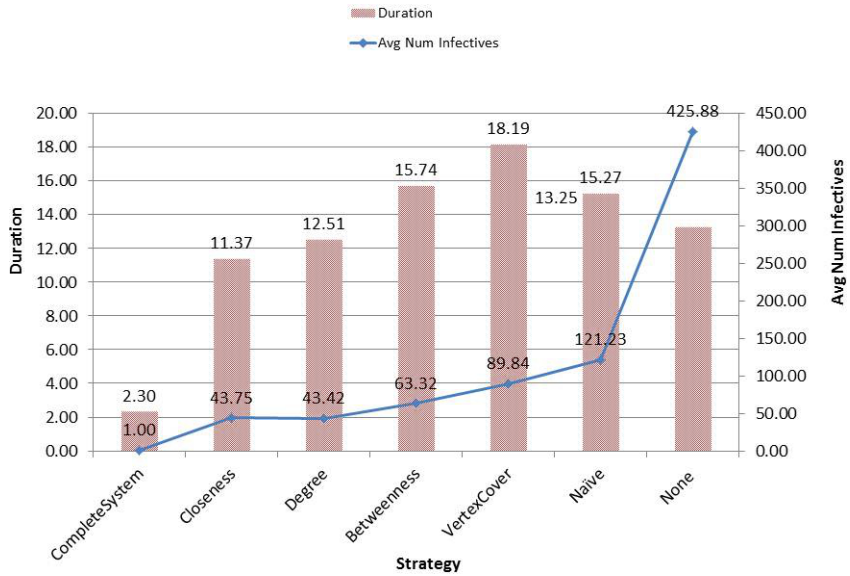
Note that in a network all nodes are assigned a centrality value. For all centrality-based interventions, the selection of schools varies with the particular realization of the MCN. Therefore, to make mitigation-efficacy comparable, the same number of schools is intervened in each scenario. The size of the vertex cover is used to establish the number of selected nodes considered for each intervention to make possible the comparison between all strategies.

Fig. 6.9 shows the results of applying the strategy based on a vertex cover of the MCNs. In addition to modifying the parameter p, the change in the parameter α is also studied. Fig. 6.10 depicts the results obtained by applying betweenness centrality, and the application of closeness centrality is shown in Fig. 6.11. Degree centrality is shown in Fig. 6.12.

Weighted betweenness centrality is tuned by the non-negative factor α. For $\alpha = 0$ the definition of betweenness is the same as the unweighted version. Note that when $\alpha > 0$, the sorted classification on nodes is the same as the one obtained when $\alpha = 1$. Similarly, weighted closeness centrality is tuned by the non-negative factor α. For $\alpha = 0$ the definition of closeness centrality is the same as the unweighted version. As defined by [126], when $\alpha > 0$ the sorted classification on nodes is the same as the one obtained when $\alpha = 1$. Finally, in the case of degree centrality, the selection of nodes varies depending on the value of α. The

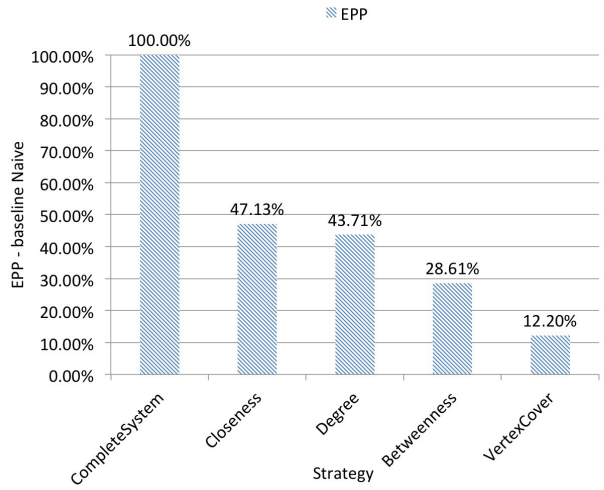


(A) Comparison of strategy type and its effect in the percentage of incidences

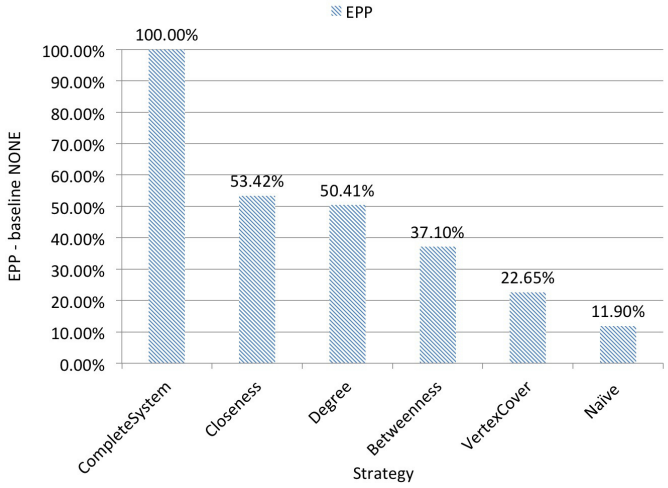


(B) Average duration and average number of infectious per day across all strategies

FIG. 6.7. Overall performance of centrality measure-based interventions



(A) Comparison of strategy type and EPP having as baseline incidents occurring in completely susceptible communities with no intervention strategies



(B) Comparison of strategy type and EPP having as baseline incidents occurring in communities for which interventions are done by selecting households randomly

FIG. 6.8. Strategy type vs. EPP comparison for different baselines

value of α changes in steps of 0.25 to evaluate the impact of α in the percentages of incidents.

6.5. Cost and Efficiency Evaluation

At the beginning of this chapter two challenges were proposed: first, to find the optimal cost-effective strategy to minimize the possibility of an outbreak when there exists

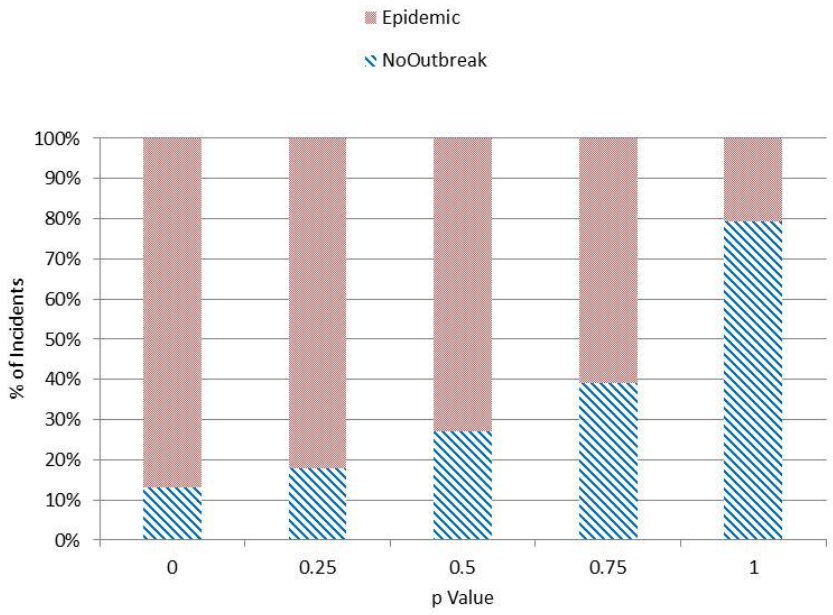
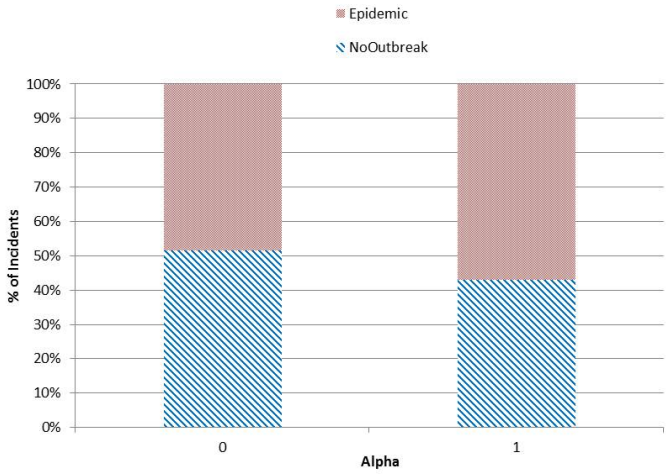


FIG. 6.9. Performance of vertex cover for different values of p

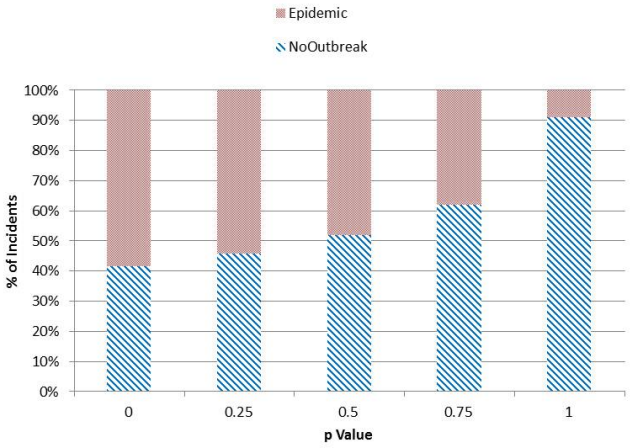
an efficacy threshold, and second, provided a level of immunization based on external constraints, find the strategy that would be most effective. The cost function related to the overall cost of the application of one or another intervention strategy is divided into two components, fixed costs and variable costs. Fixed costs include those incurred when closing any size of school. Variable costs depend on the size of the intervention. Then the total cost of strategy i (Y_i) can be calculated by (44). Note that by employing the naive approach the fixed costs equal to zero and variable costs depend solely on the size of the intervention.

$$\mathcal{Y}_\rangle = InterventionSize + VC \times SingleSchoolCost \tag{44}$$

Fig. 6.13 exemplifies the evaluation of cost function of intervention strategies based on centrality measures. It can be observed that the overall cost is directly influenced by the cost of a single school closure. The cost of closing a single school may vary greatly among



(A) Performance of betweenness centrality for different values of tuning parameter α

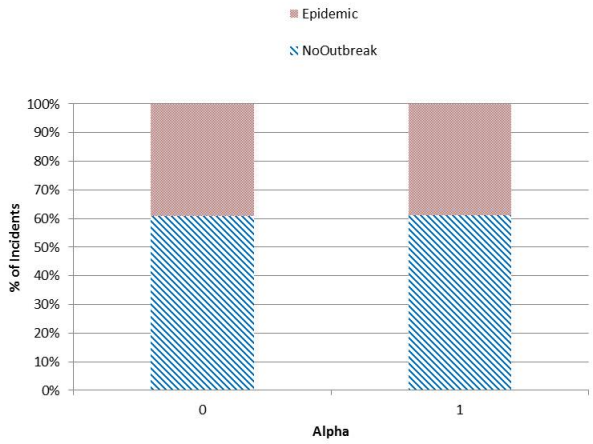


(B) Performance of betweenness centrality for different values of p

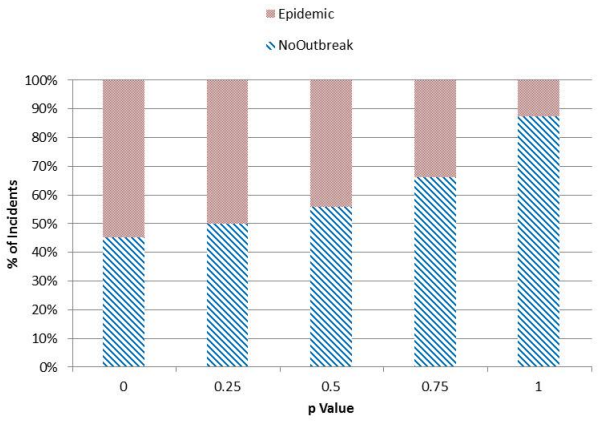
FIG. 6.10. Overall performance of betweenness centrality for different values of p and α compared to the percentage of incidents

geographies. Three main decision zones can be identified:

Z_1 Complete intervention. The cost of a complete closure of the school system is not significantly different from that of a partial closure.



(A) Performance of closeness centrality for different values of tuning parameter α

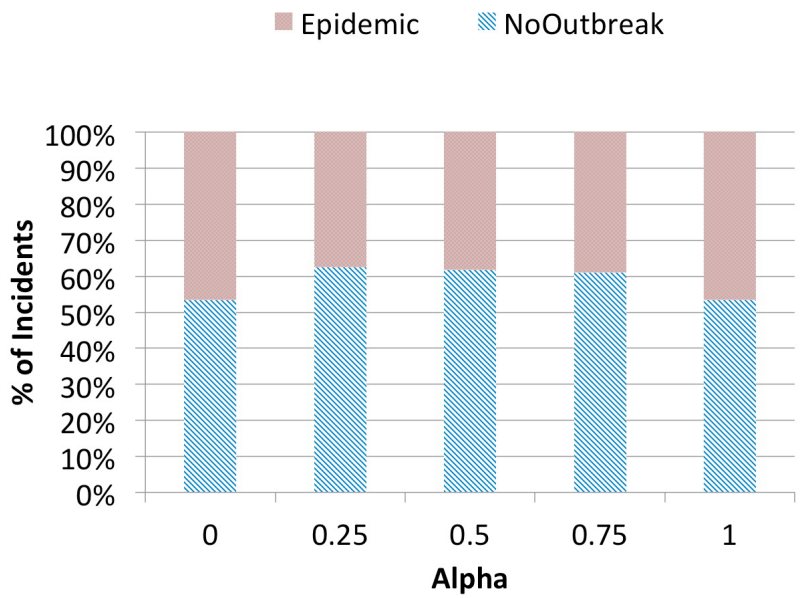


(B) Performance of closeness centrality for different values of p

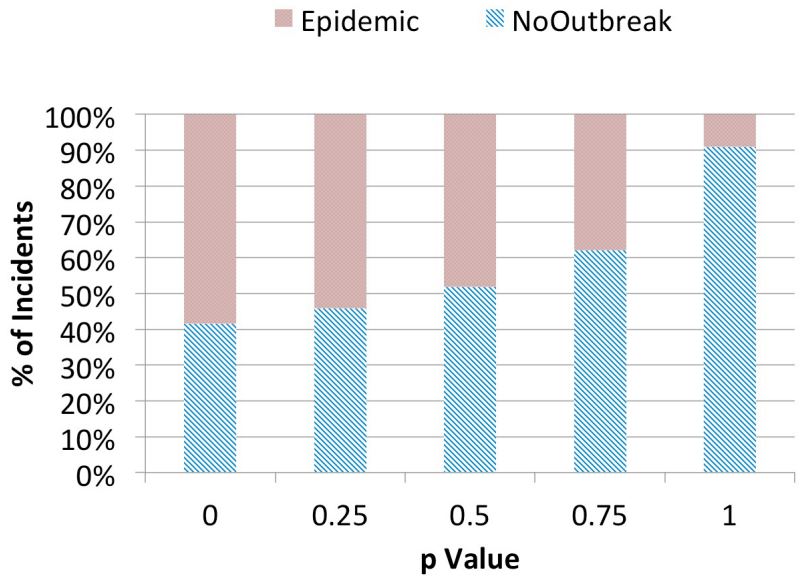
FIG. 6.11. Overall performance of closeness centrality for different values of p and α compared to the percentage of incidents

Z_2 Centrality-based intervention. The cost of a partial closure is significantly different from a complete closure and less than the naive approach.

Z_3 Any strategy. The overall cost difference between centrality-based interventions and the naive approach is not significant, therefore the selection of a strategy does not



(A) Performance of degree centrality for different values of tuning parameter α

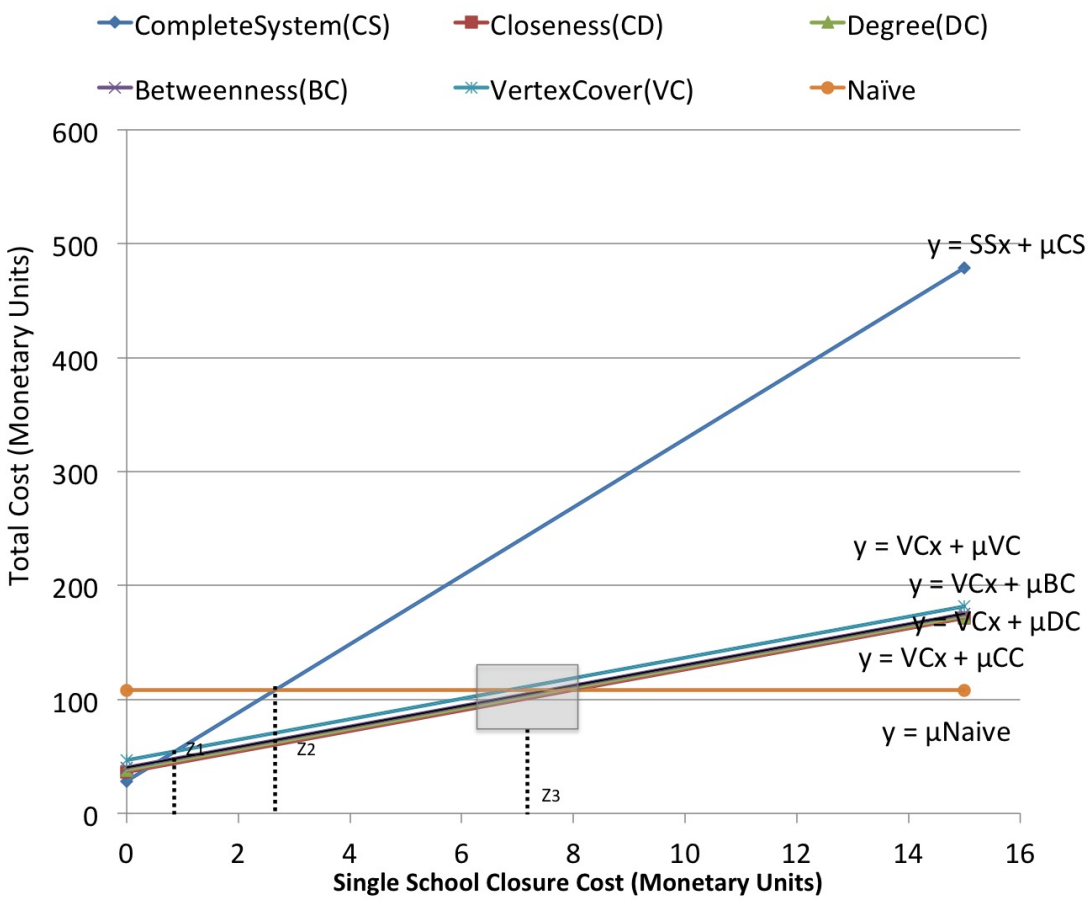


(B) Performance of degree centrality for different values of p

FIG. 6.12. Overall performance of degree centrality for different values of p and α compared to the percentage of incidents

influence the overall result.

Fig. 6.14 shows the analysis of cost-constraint strategies. The efficacy can then be stated as the relationship of the cost compared to the epidemic prevention potential. For example, when the percentage of interventions is equal to 50%, then the most cost-effective



(A) Cost evaluation of centrality-based intervention strategies

FIG. 6.13. Generic cost function for the application of interventions based on centrality measures

strategy is the rightmost point under the shadow rectangle, in this case a strategy based on SC, the naive approach may also satisfy the constraint but in terms of EPP, VC is more effective because it achieves a higher prevention potential. The most cost-effective measure is the complete system strategy.

6.6. Conclusion

This research underscores the importance of MCNS in order to meet the increasing public health challenges in the U.S. Two optimization problems have been considered: to find the strategy with minimal costs provided a required efficacy threshold and to find a strategy that maximizes EPP when an intervention cost restriction exists. The literature

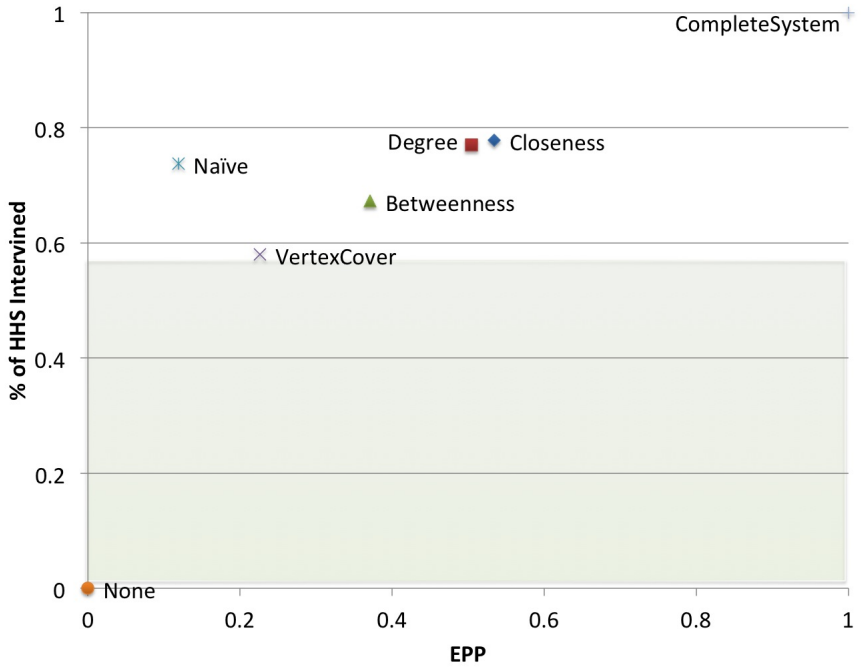


FIG. 6.14. Comparison of EPP and intervention size

about strategies for controlling the spread of disease widely focuses on how to identify high-risk individuals or groups. Previous research has identified households and schools as groups with high potential for the transmission of infectious disease and particularly influenza, as opposed to individual public activity events [71]. The description of the high-risk individuals normally includes demographic characteristics, super-spreaders [71], and groups classified as "high-activity group" [149]. However, by simply counting the number of contacts or the measure of the length of paths joining a pair of individuals the social aspects of their network context is not completely described. In order to characterize individuals, a holistic approach needs to be taken and affiliation networks respond to that challenge by addressing relations that last in time. Recent research has found that high-school students may play an important role in the next pandemic, acting like a "local transmission backbone" for disease [71]. This work has a possible explanation of this phenomenon in terms of the school system. Centrality measures for high schools make them high-risk points in all evaluation for centrality and vertex cover. Closing high-schools and keeping students at home during a

pandemic would reduce the possibility of an outbreak more than any other type of school. In addition, all centrality-based intervention outperformed the random intervention by at least 12.20%. The results presented suggest a new implementation of a highly effective strategy for targeted interventions through the use of MCNs. In addition, as other studies suggest, degree centrality outperforms betweenness centrality in assessing at-risk individuals [38]. Predicting the likely success of an strategy is vitally important in public health and particularly for schools. By modeling the correlations between schools, the risk of each school can be estimated without the need for large-scale computer simulations. This work presents a natural methodology for modeling communities of households and schools in a fixed network. This formulation is applied to the estimation of risk through the structure of the network to determine optimal intervention strategies. In addition to these results, MCNs may represent an appropriate tool for decision-makers in rural places. Targeted interventions of frequently visited locations have been found to be relevant when designing different mitigation strategies in rural places [140]. Under the most natural mechanism for identifying high-risk individuals, when the population is divided into categories, the resulting selection is based on these categories and not individuals in particular. In the case of MCNs, due to the realism of the model, not only high-risk school types are identified, but also specific schools may be pin pointed. Finally, previous research has suggested that in order to effectively combat pandemics in the future, resource-rich countries will be called upon to share vaccines and antivirals with other countries constrained in resources [42]. This research provides an alternative that can be used in any geography for which information is available, optimize schools closures.

6.7. Summary

This research investigates the relation of multiple measures of network centrality to the risk of infection after a emergence of an infectious disease in the school system. The methodology presented in this chapter utilizes a reconstruction of the population of Denton county for the year 2009. MCNs are used as the networks that the diseases use to propagate, in which schools are linked to each other through the households they have in common. Most

network-based studies analyze the impact that variations in network degree and clustering have on the dynamics of disease. In this chapter weighted centrality measures are used as the main strategic proposition to stop the spread of a disease.

CHAPTER 7

SUMMARY AND CONCLUSIONS

> The time has come, it would seem, to stop, take stock and try to make some sense of the concept of centrality and the range and limits of its potential for application
>
> *Linton Freeman, 1978*

In this chapter, dissertation results are summarized first in Section 7.1. Then directions for further research and reflections on computational epidemiology are outlined in Section 7.2.

7.1. Dissertation Summary

7.1.1. Synthetic Reconstruction

Is it possible to design a model with a compromise between parsimony and realism that would intrinsically incorporate various aspects of demographics and the interaction structure among individuals?

This research started by outlining a synthetic reconstruction methodology extended to accommodate school attendance zones information. Denton County and Denton ISD statistics were used as input information for the simulator that generated the synthetic population reconstruction of the county. The methodology utilized a multi-level control algorithm to better reflect household and person joint distributions. Allocation of children in schools was performed next, and was leveraged to define the affiliations between households and schools. The problem of computational feasibility for ABMs was addressed from a novel perspective. On one hand this research attempted to preserve the synthetic populations as close to reality as possible by utilizing census information instead of datasets from online communities. On the other hand, a theoretic approach that raised from formulating the problem in terms of a newly defined graph to utilize a more theoretic approach as opposed to simulation of stochastic contacts.

7.1.2. Affiliation Networks

Can the concept of affiliation networks be extended by the use of k-partite networks to accommodate the school system's hierarchical structure?

The school affiliation network discovery (SAND) algorithm is used to associate children and schools. The input for the algorithms is a synthetic reconstruction and the output is a bipartite graph A that represents the affiliation network of schools and households. In the methodology, the function $\mathcal{A}$ reflects what represents an affiliation for the particular study area. From graph A the co-affiliation network is formulated by stretching nodes representing households into links. The resulting $k-partite$ graph only contains schools and it is defined as multi co-affiliation networks or MCNs.

MCNs represent the initial social network and differ from other biology inspired networks in that they are not sparse. MCNs are simple, undirected, weighted, k-partite graphs that represent the hierarchical structure of the school system. Compared to the initial synthetic reconstruction, dimension of MCNs is drastically reduced. MCNs depict long-term relationships rather than random encounters that are the basic metric for contact networks.

7.1.3. Optimization of Intervention Strategies

Can a nested hierarchy of successively larger domains be used to quantify the effectiveness of an intervention method to either prevent transmission of an infectious disease or at least to keep it below a pre-defined limit in the school system?

This research seeks to identify how MCNs could constitute a tool for the design of intervention strategies. In order to meet the increasing public health challenges in the U.S., two optimization problems are considered: finding a strategy with minimal costs provided a required efficacy threshold and finding a strategy that maximizes EPP provided that a cost restriction exists for interventions.

Through the comparison of simulated scenarios, the cost-efficacy of different intervention strategies is established. To find the strategy at minimal cost given a efficacy thresholds depends directly of what constitutes fixed and variable cost of the intervention. In the case of school closures, the variable costs are directly associated with the size of the school.

Compared with the null strategy and the naive strategy, all centrality-based intervention outperformed the random intervention by at least 12.20%. In cases for which the efficacy threshold is less that 100%, the best strategy is based on closeness-centrality for the school system.

The second question is answered by comparing the EPP of each strategy and the percentage of the population accounted for in the intervention, which represents the costs. Note that at some cost levels, no strategy would accomplish prevention. Although at some intervention levels, more than one strategy could be applied and the most effective would be the strategy with rightmost point in Fig. 6.14. Finally, it can be concluded that by modeling the correlations between schools, the risk of each school can be estimated without the need for large-scale computer simulations. This work presents a natural methodology for modeling communities of households and schools in a fixed network.

7.2. Future Work

Intervention strategies focused on school children may have substantial benefits to society. Seasonality, which characterizes many infectious diseases, presents an opportunity for strategies specifically designed for school systems because it allows to estimate risks before an event occurs. To design strategies this research proposes the use of long-term affiliation networks in lieu of large-scale simulations. The constructions of MCNs is not restricted to schools, but the model is capable to accommodate other dimensions without increasing computation complexity. Places such as shopping centers, houses of worship, and other locations, which are defined by affiliations can be included in order to study different diseases.

The primary challenge of optimization is supporting what-if scenarios in order to evaluate and contrast different strategies. There is a compelling argument for the implementation of large-scale simulations to accomplish this objective. However, new methodologies formulated from hybrid approaches, such as MCNs, may represent a truly useful new set of paradigms. This dissertation has aimed to take steps in that direction, but the challenge extends far beyond the problem of evaluating centrality-based intervention strategies.

This research has focused on the proactive approach for mitigation strategies. Future work should include reactive approaches that are closer, in reality, how public health bodies address the problem of disease mitigation. The constructions of MCNs is not restricted to schools, but the model is capable to accommodate other dimensions. Different long-term affiliations could potentially be modeled to study infectious diseases that have a different contact network such as sexually transmitted diseases. Finally, pharmaceutical interventions such as prophylaxis and antiviral treatment can be optimized by including in the model the efficiency of the treatment and probability of contracting the disease through the vaccination itself.

APPENDIX A

DENTON ISD SCHOOL ATTENDANCE ZONE MAPS

A.1. Denton ISD SAZ Maps

(A) Denton ISD High School Attendance Zones

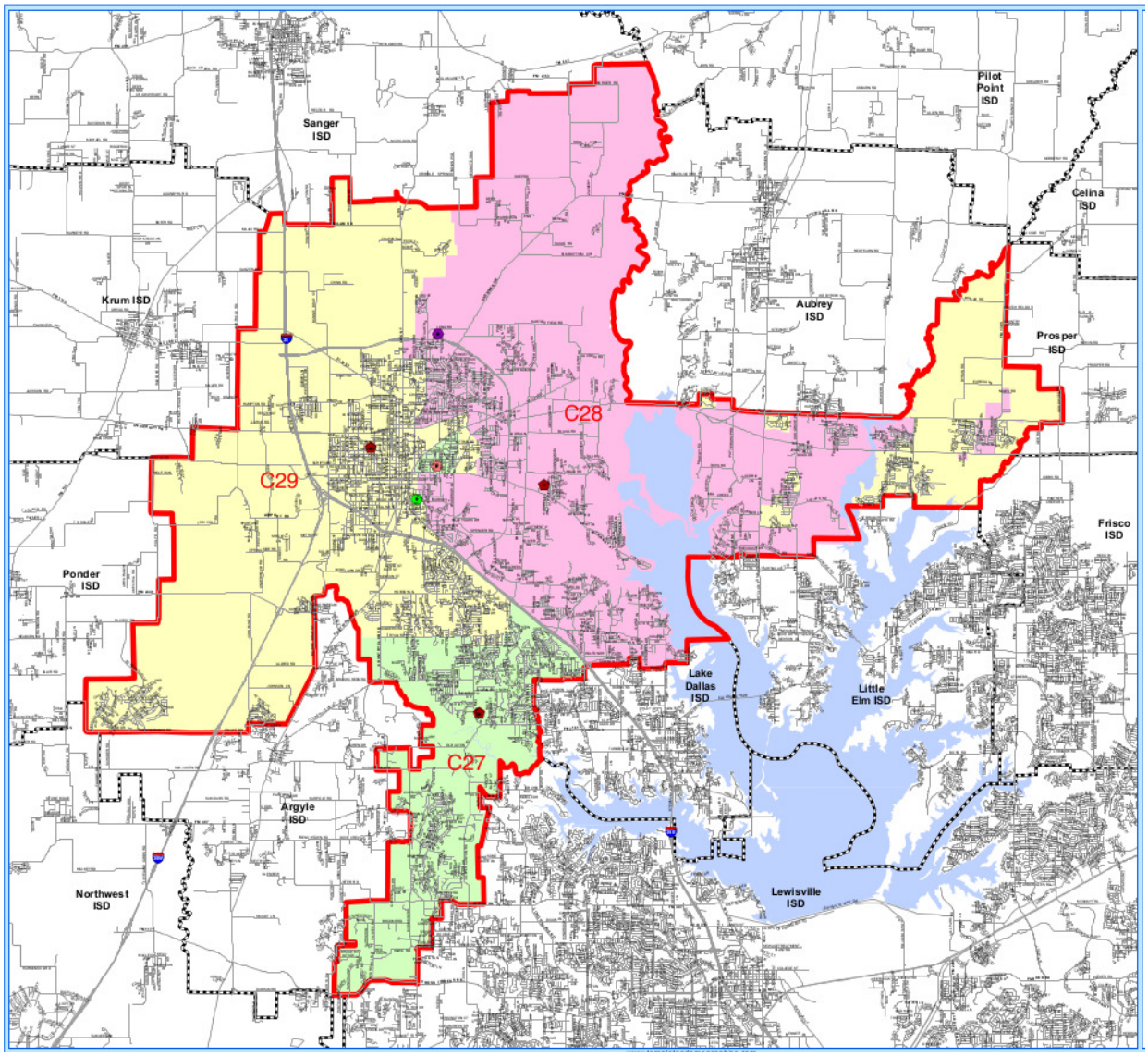


FIG. A.1.1. High School Codes for Denton ISD

(A) Denton ISD Middle School Attendance Zones

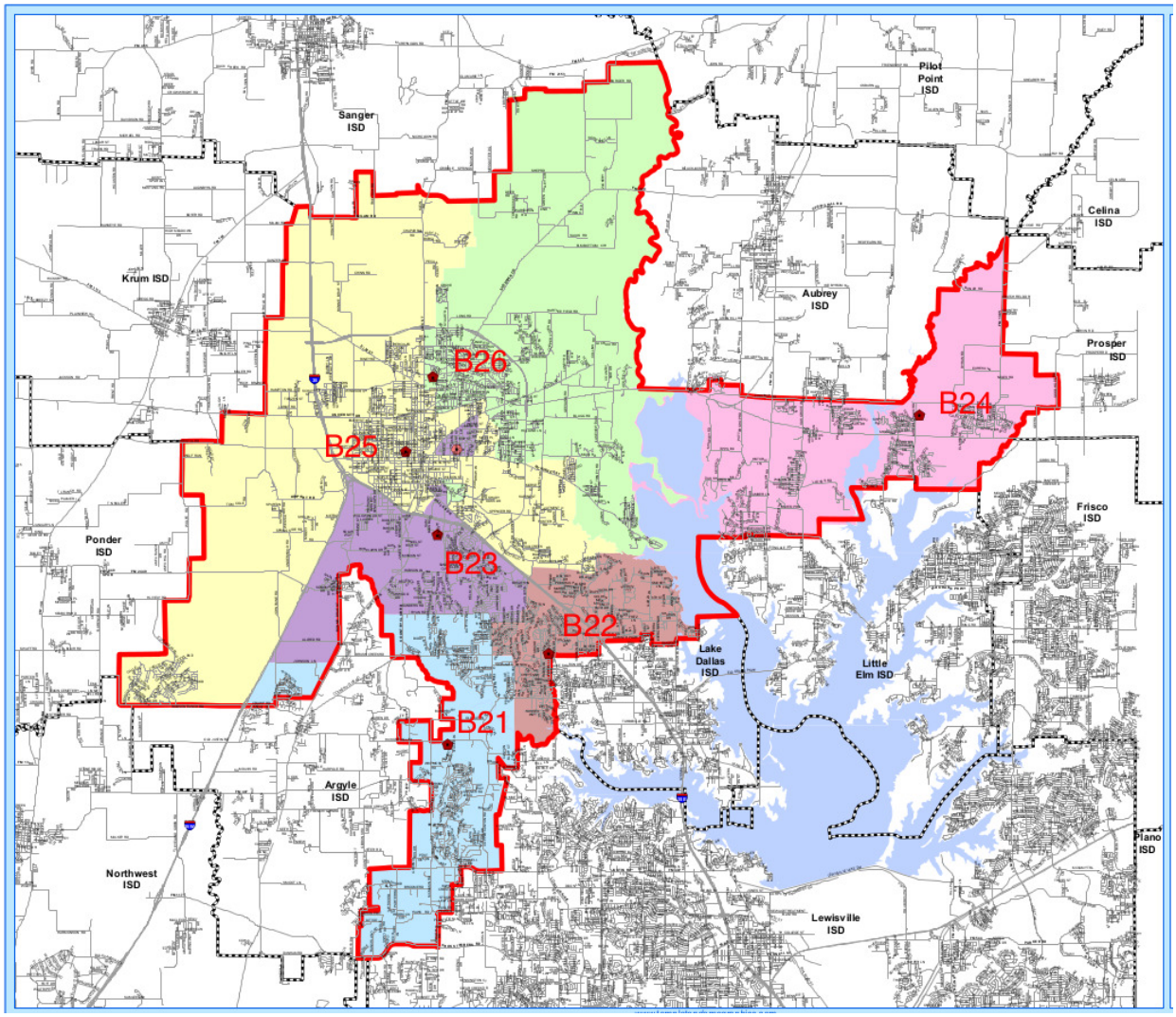


FIG. A.1.2. Middle Schools Codes for Denton ISD

(A) Denton ISD Elementary School Attendance Zones

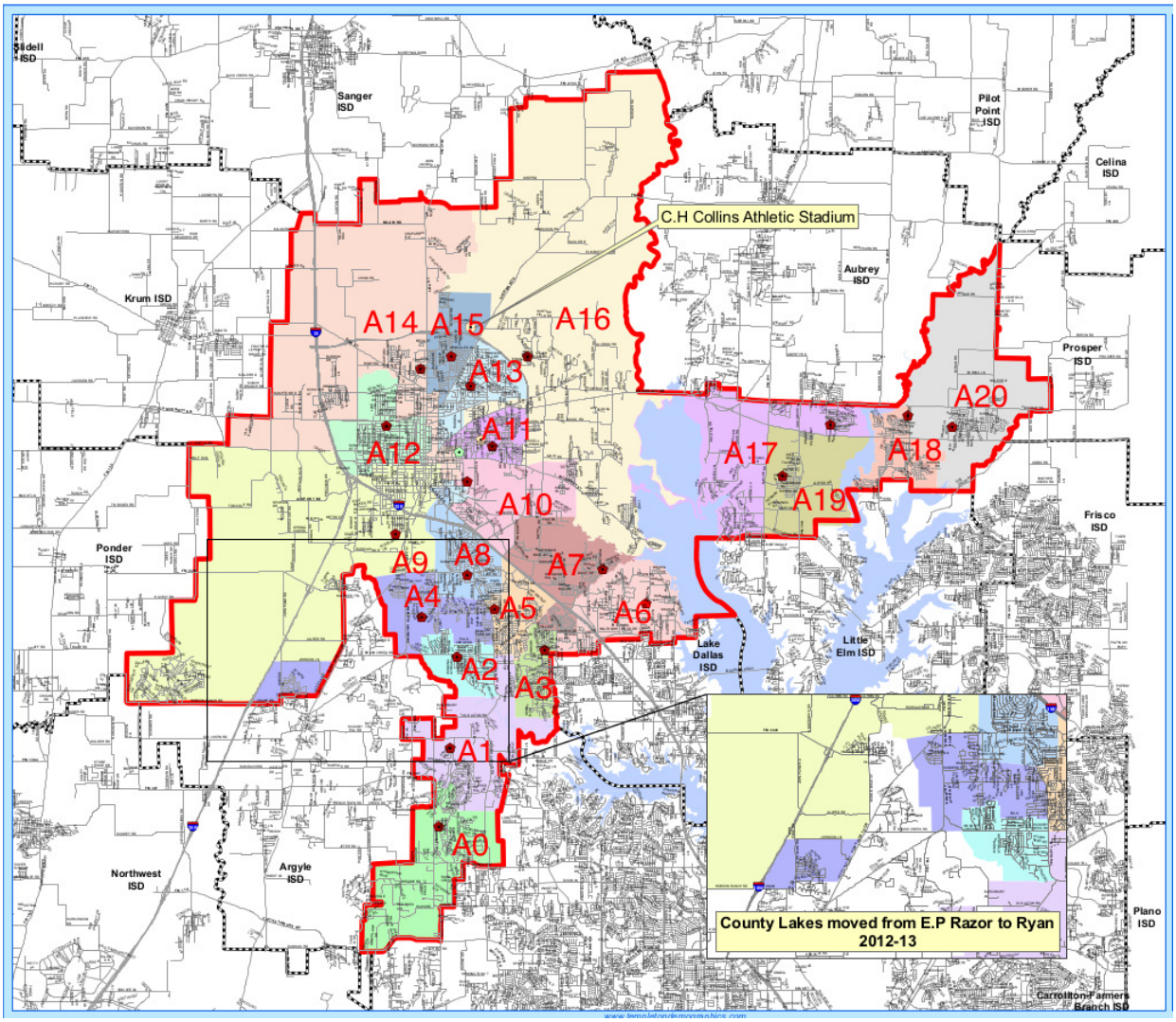


FIG. A.1.3. Elementary School Codes for Denton ISD

BIBLIOGRAPHY

[1] Lada Adamic and Eytan Adar, *How to search a social network*, Social Networks 27 (2005), 187–203.

[2] A.L. Adams, J.S. Koopman, S.E. Chick, and P.J. Yu, *Germs: an epidemiologic simulation tool for studying geographic and social effects on infection transmission*, Simulation Conference Proceedings, 1999 Winter, vol. 2, 1999, pp. 1549 –1556 vol.2.

[3] R. D. Alba and C. Kadushin, *The intersection of social circles: A new measure of social proximity in networks*, Sociological Methods & Research 5 (1976), no. 1, 77–102.

[4] Stefano Allesina and Mercedes Pascual, *Googling food webs: can an eigenvector measure species' importance for coextinctions?*, PLoS computational biology 5 (2009), no. 9, e1000494.

[5] May R.M. Anderson R.M., *Infectious diseases of humans*, Oxford University Press, 1992.

[6] Petter Holme Andreas Gronlund, *A network-based threshold model for the spreading of fads in society and markets*, Networks (2005).

[7] Joshua Auld and Abolfazl Mohammadian, *An efficient methodology for generating synthetic populations with multiple control levels*, Publications - Civil and Materials Engineering (2010).

[8] Duygu Balcan, Vittoria Colizza, Bruno Gonc, and Hao Hu, *Multiscale mobility networks and the spatial spreading of infectious diseases*, PNAS 106 (2009), no. 51.

[9] Michael E Bales and Stephen B Johnson, *Graph theoretic modeling of large-scale semantic networks.*, Journal of biomedical informatics 39 (2006), no. 4, 451–64.

[10] Frank Ball and Peter Neal, *Network epidemic models with two levels of mixing.*, Mathematical biosciences 212 (2008), no. 1, 69–87.

[11] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek, *Evolution of the social network of scientific collaborations*, 2002.

[12] Marc Barthelemy, *Spatial Networks*, Physics Reports 499 (2011), no. 1-3, 1–101.

[13] Chris T Bauch and David J D Earn, *Vaccination and the theory of games.*, Proceedings of the National Academy of Sciences of the United States of America 101 (2004), no. 36, 13391–4.

[14] Peter S. Bearman, James Moody, and Katherine Stovel, *Chains of affection: The structure of adolescent romantic and sexual networks*, American Journal of Sociology 110 (2002), 44–91.

[15] N Becker, *Optimal vaccination strategies for a community of households*, Mathematical Biosciences 139 (1997), no. 2, 117–132.

[16] Richard J Beckman, Keith A Baggerly, and Michael D McKay, *Creating Synthetic Baseline Populations*, Transportation Research Part A: Policy and Practice 30 (1996), 415–429.

[17] Vitaly Belik, Theo Geisel, and Dirk Brockmann, *Natural human mobility patterns and spatial spread of infectious diseases*, Physical Review X 1 (2011), no. 1, 1–5.

[18] Terry Bergeson, *Infectious disease control guide for school staff*, State of Washington Office of Superintendent of Public Instruction Department of Health, 2004.

[19] Daniel Bernoulli and Sally Blower, *An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it*, Reviews in Medical Virology 14 (2004), no. 5, 275–288.

[20] K. Binder, *Finite size scaling analysis of ising model block distribution functions*, Zeitschrift fr Physik B Condensed Matter 43 (1981), 119–140 (English).

[21] S Borgatti and M Everett, *A Graph-theoretic perspective on centrality*, Social Networks 28 (2006), no. 4, 466–484.

[22] S.P. Borgatti and D. Halgin, *Analyzing Affiliation Networks*, The Sage Handbook of Social Network Analysis (2011).

[23] Stephen P Borgatti, *Centrality and Network Flow*, Social Networks 27 (2005), no. 1, 55—-71.

[24] ———, *Centrality and Network Flow*, Social Networks 27 (2005), no. 1, 55—-71.

[25] Stephen P. Borgatti, Ajay Mehra, Daniel J. Brass, and Giuseppe Labianca, *Network analysis in the social sciences*, Science 323 (2009), no. 5916, 892–895.

[26] Tom Britton, Maria Deijfen, Andreas N Lager, and Mathias Lindholm, *Epidemics on random graphs with tunable clustering*, (2007), no. 1980.

[27] Charles D Brummitt and Kyu-min Lee, *Multiplexity-facilitated cascades in networks*, 1 (2012), no. Layer 2, 1–5.

[28] J. Bryden, S. Funk, N. Geard, S. Bullock, and V. a. a. Jansen, *Stability in flux: community structure in dynamic networks*, Journal of The Royal Society Interface 8 (2010), no. 60, 1031–1040.

[29] United States Census Bureau, *Public-use microdata samples (pums) texas map*, February 2012.

[30] ———, *Summary files*, February 2012.

[31] Technical Documentationprepared by the U.S. Census Bureau, *Technical documentation: Census 2000 summary file 1*, 2001.

[32] Cirillo C., Cornelis E., and Toint P., *Model of weekly working participation for a belgian synthetic population*, Proceedings of the European Transport. Conference (ETC) 2007 (2007).

[33] Margaret Chan, *Opening intervention at the international health regulations review committee*, September 2010.

[34] M Chau and J Xu, *Mining communities and their relationships in blogs: A study of online hate groups*, International Journal of Human-Computer Studies 65 (2007), no. 1, 57–70.

[35] Wei Chen, Shang-hua Teng, and Jiajie Zhu, *The betweenness centrality game for strategic network formations*, Microsoft Research (2008), no. Tech Report, 1–30.

[36] Gerardo Chowell, Hiroshi Nishiura, and Luís M a Bettencourt, *Comparative estimation of the reproduction number for pandemic influenza from daily case notification data.*, Journal of the Royal Society, Interface / the Royal Society 4 (2007), no. 12, 155–66.

[37] Nicholas A Christakis and James H Fowler, *The spread of obesity in a large social network over 32 years*, Network (2007).

[38] R M Christley, G L Pinchbeck, R G Bowers, D Clancy, N P French, R Bennett, and J Turner, *Infection in social networks: using network analysis to identify high-risk individuals.*, American journal of epidemiology 162 (2005), no. 10, 1024–31.

[39] Aaron Clauset, Cristopher Moore, and M E J Newman, *Hierarchical structure and the prediction of missing links in networks.*, Nature 453 (2008), no. 7191, 98–101.

[40] Aaron Clauset, Cristopher Moore, and World Wide Web, *Finding community structure in very large networks*, 066111 (2004), 1–6.

[41] A.D. Cliff, *The study of spatial difussion*, Spatial diffusion: an historical geography of epidemics in an island community, Cambridge University Press, 1981, pp. 6 – 35.

[42] Brian J Coburn, Bradley G Wagner, and Sally Blower, *Modeling influenza epidemics and pandemics: insights into the future of swine flu (H1N1).*, BMC medicine 7 (2009), 30.

[43] Flavio C Coelho, Oswaldo G Cruz, and Claudia T Codeco, *Epigrass: a tool to study disease spread in complex networks.*, Source code for biology and medicine 3 (2008), no. 1, 3.

[44] Vittoria Colizza, Alain Barrat, Marc Barthelemy, Alain-jacques Valleron, and Alessandro Vespignani, *Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions*, PLoS Medicine 4 (2007), no. 1.

[45] Vittoria Colizza and Alessandro Vespignani, *Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: theory and simulations.*, Journal of theoretical biology 251 (2008), no. 3, 450–67.

[46] A J K Conlan, K T D Eames, J A Gage, J C von Kirchbach, J V Ross, R A Saenz, and J R Gog, *Measuring social networks in British primary schools through scientific engagement.*, Proceedings. Biological sciences / The Royal Society 278 (2011), no. 1711, 1467–1475.

[47] Courtney D. Corley and Armin R. Mikler, *A computational framework to study public*

health epidemiology, 2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing (2009), 360–363.

[48] Iain D Couzin, Christos C Ioannou, Güven Demirel, Thilo Gross, Colin J Torney, Andrew Hartnett, Larissa Conradt, Simon a Levin, and Naomi E Leonard, *Uninformed individuals promote democratic consensus in animal groups.*, Science (New York, N.Y.) 334 (2011), no. 6062, 1578–80.

[49] A. Davis, B. B. Gardner, and M. R. Gardner, *Deep south; a social anthropological study of caste and class*, Chicago, IL, US: University of Chicago Press., 1941.

[50] Pedro Domingos and Matt Richardson, *Mining the network value of customers*, Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (New York, NY, USA), KDD '01, ACM, 2001, pp. 57–66.

[51] Jennifer a Dunne, Richard J Williams, and Neo D Martinez, *Food-web structure and network theory: The role of connectance and size.*, Proceedings of the National Academy of Sciences of the United States of America 99 (2002), no. 20, 12917–22.

[52] LILA ELVEBACK, EUGENE ACKERMAN, LAL GATEWOOD, and JOHN P. FOX, *Stochastic two-agent epidemic simulation models for a community of families*, American Journal of Epidemiology 93 (1971), no. 4, 267–280.

[53] LILA R ELVEBACK, JOHN P. FOX, EUGENE ACKERMAN, ALICE LANGWORTHY, MARY BOYD, and LAEL GATEWOOD, *An influmza simulation model for immunization studies*, American Journal of Epidemiology 103 (1976), no. 2, 152–165.

[54] Joshua M. Epstein, D. Michael Goedecke, Feng Yu, Robert J. Morris, Diane K. Wagener, and Georgiy V. Bobashev, *Controlling pandemic flu: The value of international air travel restrictions*, PLoS ONE 2 (2007), no. 5, e401.

[55] Stephen Eubank, Hasan Guclu, V S Anil Kumar, Madhav V Marathe, Aravind Srinivasan, Zoltán Toroczkai, and Nan Wang, *Modelling disease outbreaks in realistic urban social networks.*, Nature 429 (2004), no. 6988, 180–4.

[56] Stephen Eubank, V. S. Anil Kumar, Madhav V. Marathe, Aravind Srinivasan, and Nan Wang, *Structural and algorithmic aspects of massive social networks*, Proceedings of

the fifteenth annual ACM-SIAM symposium on Discrete algorithms (Philadelphia, PA, USA), SODA '04, Society for Industrial and Applied Mathematics, 2004, pp. 718–727.

[57] Eyal Even-dar and Asaf Shapira, *A Note on Maximizing the Spread of In uence in Social Networks*, Network.

[58] M Everett and S Borgatti, *Ego network betweenness*, Social Networks 27 (2005), no. 1, 31–38.

[59] Rasmus H Fogh, Wayne Boucher, Wim F Vranken, Anne Pajon, Tim J Stevens, T N Bhat, John Westbrook, John M C Ionides, and Ernest D Laue, *A framework for scientific data modeling and automated software development.*, Bioinformatics (Oxford, England) 21 (2005), no. 8, 1678–84.

[60] Centers for Disease Control and Prevention, *Influenza viruses*, November 2005.

[61] ———, *Estimates of Deaths Associated with Seasonal Influenza — United States, 1976–2007*, August 2010.

[62] National Center for Education Statistics, *Sumary files*, February 2012.

[63] National Center for Education Statistics-Common Core of Data, *Public elementary/secondary school universe survey data*, February 2012.

[64] National Center for Educational Statistics (NCES), *Public school districts information*, June 2010.

[65] Christophe Fraser, Medical Research, Council Centre, Outbreak Analysis, Infectious Disease Epidemiology, United Kingdom, Estimating Individual, Household Reproduction Numbers, Emerging Epidemic, and Plos One, *Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic*, America (2007), no. 8.

[66] Linton C. Freeman, *Finding social groups: A meta-analysis of the southern women data*, Dynamic Social Network Modeling and Analysis. The National Academies, Press, 2003, pp. 39–97.

[67] M. Frick and K.W. Axhausen, *Generating synthetic populations using ipf and monte carlo techniques*, 4th Swiss Transport Research Conference (2004).

[68] Andrew C Gallup, Joseph J Hale, David J T Sumpter, Simon Garnier, Alex Kacelnik,

John R Krebs, and Iain D Couzin, *Visual attention and the acquisition of information in human crowds.*, Proceedings of the National Academy of Sciences of the United States of America 109 (2012), no. 19, 7245–50.

[69] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant, *Detecting influenza epidemics using search engine query data.*, Nature 457 (2009), no. 7232, 1012–4.

[70] M. Girvan and M. E. J. Newman, *Community structure in social and biological networks*, Proceedings of the National Academy of Sciences 99 (2002), no. 12, 7821–7826.

[71] Laura M Glass and Robert J Glass, *Social contact networks for the spread of pandemic influenza in children and teenagers.*, BMC public health 8 (2008), 61.

[72] J. Glasser, D. Taneri, and Z. et al. Feng, *Evaluation of targeted influenza vaccination strategies via population modeling*, PLoS ONE (2010).

[73] Scott A. Golder, Dennis M. Wilkinson, and Bernardo A. Huberman, *Rhythms of social interaction: Messaging within a massive online network*, Proc. 3rd Intl. Conf. on Communities and Technologies, 2007.

[74] M. C. González, P. G. Lind, and H. J. Herrmann, *Model of mobile agents for sexual interactions networks*, The European Physical Journal B 49 (2006), no. 3, 371–376.

[75] Marta González, Pedro Lind, and Hans Herrmann, *System of Mobile Agents to Model Social Networks*, Physical Review Letters 96 (2006), no. 8, 2–5.

[76] R a Gunn, S L Harper, D E Borntrager, P E Gonzales, and M E St Louis, *Implementing a syphilis elimination and importation control strategy in a low-incidence urban area: San diego county, california, 1997-1998.*, American journal of public health 90 (2000), no. 10, 1540–4.

[77] Hakan and Andersson, *Epidemics in a population with social structures*, Mathematical Biosciences 140 (1997), no. 2, 79 – 84.

[78] M Elizabeth Halloran, Neil M Ferguson, Stephen Eubank, Ira M Longini, Derek A T Cummings, Bryan Lewis, Shufu Xu, Christophe Fraser, Anil Vullikanti, Timothy C Germann, Diane Wagener, Richard Beckman, Kai Kadau, Chris Barrett, Catherine A

Macken, Donald S Burke, and Philip Cooley, *Modeling targeted layered containment of an influenza pandemic in the united states.*, Proceedings of the National Academy of Sciences of the United States of America 105 (2008), no. 12, 4639–44.

[79] M. Elizabeth Halloran, Ira M. Longini, David M. Cowart, and Azhar Nizam, *Community interventions and the epidemic prevention potential*, Vaccine 20 (2002), no. 27-28, 3254 – 3262.

[80] Frank Harary, *The Maximum Connectivity of a Graph*, Proceedings of the National Academy of Science 48 (1962), no. 7, 1142–1146.

[81] S. Heath, a. Fuller, and B. Johnston, *Chasing shadows: defining network boundaries in qualitative social network analysis*, Qualitative Research 9 (2009), no. 5, 645–661.

[82] Gilles Hejblum, Michel Setbon, Laura Temime, Sophie Lesieur, and Alain-Jacques Valleron, *Modelers' perception of mathematical modeling in epidemiology: A web-based survey*, PLoS ONE 6 (2011), no. 1, e16531.

[83] Petter Holme and M. E. J. Newman, *Nonequilibrium phase transition in the coevolution of networks and opinions*, Phys. Rev. E 74 (2006), 056108.

[84] William H Hsu, Joseph Lancaster, Martin S R Paradesi, and Tim Weninger, *Structural Link Analysis from User Profiles and Friends Networks: A Feature Construction Approach*, Ratio (2007).

[85] Bo Hu, Xin-Yu Jiang, Jun-Feng Ding, Yan-Bo Xie, and Bing-Hong Wang, *A weighted network model for interpersonal relationship evolution*, Physica A: Statistical Mechanics and its Applications 353 (2005), 576–594.

[86] Denton Independent School District (ISD), *Attendance zone maps*, February 2012.

[87] Smith B. Islam MN, O'Shaughnessy CD, *A random graph model for the final-size distribution of household infections.*, Stat Med. 15 (1996), no. 7-9, 837–843.

[88] Tina V. Johnson, *The influence of social network graph structure on disease dynamics in a simulated environment.*, Denton, Texas. UNT Digital Library, 2010.

[89] RichardM. Karp, *Reducibility among combinatorial problems*, 50 Years of Integer Programming 1958-2008 (Michael Jnger, Thomas M. Liebling, Denis Naddef, George L.

Nemhauser, William R. Pulleyblank, Gerhard Reinelt, Giovanni Rinaldi, and Laurence A. Wolsey, eds.), Springer Berlin Heidelberg, 2010, pp. 219–241 (English).

[90] S. Kaza, Daning Hu, and Hsinchun Chen, *Dynamic social network analysis of a dark network: Identifying significant facilitators*, Intelligence and Security Informatics, 2007 IEEE, may 2007, pp. 40 –46.

[91] M J Keeling, *The effects of local spatial structure on epidemiological invasions.*, Proceedings. Biological sciences / The Royal Society 266 (1999), no. 1421, 859–67.

[92] Matt J Keeling and Ken T D Eames, *Networks and epidemic models.*, Journal of the Royal Society, Interface / the Royal Society 2 (2005), no. 4, 295–307.

[93] D Kempe, J M Kleinberg, and E Tardos, *Maximizing the spread of in uence through a social network*, In The Nineth International Conference on Knowledge scovery and Data Mining (KDD, 137–146.

[94] Jon Kleinberg, Siddharth Suri, Éva Tardos, and Tom Wexler, *Strategic network formation with structural holes*, SIGecom Exch. 7 (2008), no. 3, 11:1–11:4.

[95] Tom Koch, *Cartographies of disease: maps, mapping, and medicine.*, Redlands, Calif: ESRI Press., 2005.

[96] M Kretzschmar, S Van Den Hof, J Wallinga, and van Wijngaarden J, *Ring vaccination and smallpox control*, Emerg Infect Dis 10 (2004), no. 5, 832 – 841.

[97] Mirjam Kretzschmar and Anton J. Severijnen, *Modeling prevention strategies for gonorrhea and chlamydia using stochastic network simulations*, Am. J. Epidemiol 114 (1996), 306–317.

[98] Andrea Lancichinetti, Mikko Kivelä, Jari Saramäki, and Santo Fortunato, *Characterizing the community structure of complex networks.*, PloS one 5 (2010), no. 8, e11976.

[99] L. Laura, S. Leonardi, G. Caldarelli, P. De, and P. De Los Rios, *A multi-layer model for the web graph*, In On-line proceedings of the 2nd International Workshop on Web Dynamics, 2002.

[100] A. B. Lawson, F. L. R. Williams, and F. Williams, *An introductory guide to disease mapping*, John Wiley, 2001.

[101] Howard Lempel, Ross A Hammond, and Joshua M Epstein, *Center on social and economic dynamics working paper no.55*, Health Care (2009), no. 55.

[102] Reilly M Leu M, Czene K, *Population lab: The creation of virtual populations for genetic epidemiology research. epidemiology*, Epidemiology 18 (2007), 433–440.

[103] Pierre P Lévy and Alain-Jacques Valleron, *Toward unsupervised outbreak detection through visual perception of new patterns.*, BMC public health 9 (2009), 179.

[104] Chunguang Li and Philip K Maini, *An evolving network model with community structure*, Physica A 38 (2005), 9741–9749.

[105] Marc Lipsitch, Ted Cohen, Megan Murray, and Bruce R Levin, *Antiviral resistance and the control of pandemic influenza*, PLoS Med 4 (2007), no. 1, e15.

[106] Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási, *Controllability of complex networks.*, Nature 473 (2011), no. 7346, 167–73.

[107] Gonzalo Martín, Maria-Cristina Marinescu, David E Singh, and Jesús Carretero, *Leveraging social networks for understanding the evolution of epidemics*, BMC Systems Biology 5 (2011), no. Suppl 3, S14.

[108] Major Ian A Mcculloh, Cadet Joshua A Lospinoso, and Kathleen M Carley, *Network Simulation Models*, Proceedings of the Army Science Conference (26th) Held (Orlando, Florida), 2008.

[109] David B. McDonald, *Predicting fate from early connectivity in a social network*, Proceedings of the National Academy of Sciences 104 (2007), no. 26, 10910–10914.

[110] Stefano Merler, Marco Ajelli, Andrea Pugliese, and Neil M. Ferguson, *Determinants of the spatiotemporal dynamics of the 2009 h1n1 pandemic in europe: Implications for real-time modelling*, PLoS Comput Biol 7 (2011), no. 9, e1002205.

[111] A.R. Mikler, A. Bravo-Salgado, and C.D. Corley, *Global stochastic contact modeling of infectious diseases*, Bioinformatics, Systems Biology and Intelligent Computing, 2009. IJCBS '09. International Joint Conference on, aug. 2009, pp. 327 –330.

[112] R T Mikolajczyk, M K Akmatov, S Rastin, and M Kretzschmar, *Social contacts of*

school children and the transmission of respiratory-spread pathogens., Epidemiology and infection 136 (2008), no. 6, 813–22.

[113] Joel C. Miller, *Spread of infectious disease through clustered populations*, Journal of The Royal Society Interface 6 (2009), no. 41, 1121–1134.

[114] Denis Mollison, *The Structure of Epidemic Models*, Analysis (1995), no. Section 3, 17–33.

[115] J. Müller, *Optimal vaccination patterns in age-structured populations: Endemic case*, Mathematical and Computer Modelling 31 (2000), no. 4-5, 149–160.

[116] K. Müller, K.W. Axhausen, K.W. Axhausen, and K.W. Axhausen, *Population synthesis for microsimulation: state of the art*, ETH Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen- und Eisenbahnbau (IVT), 2010.

[117] Johannes Mller, *Optimal vaccination strategiesfor whom?*, Mathematical Biosciences 139 (1997), no. 2, 133 – 154.

[118] Tamás Nepusz and Tamás Vicsek, *Controlling edge dynamics in complex networks*, Nature Physics 8 (2012), 568—-573.

[119] M. E. J. Newman, *Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality*, Physical Review 64 (2001), 016132.

[120] M E J Newman, D J Watts, S H Strogatz, and Empirical Data, *Random graph models of social networks*, Fortune 99 (2002).

[121] M.E.J. Newman, *Networks: an introduction*, Oxford University Press, 2010.

[122] NCES Common Core of Data, *School districts information*, February 2012.

[123] The College of William, Mary, and the Minnesota Population Center, *School attendance boundary information system (sabins) version 1.0. minneapolis, mn: University of minnesota 2011.*, November 2010.

[124] Iris Gomez-Lopez Olivia Loza and Armin Mikler, *Multi-coaffiliation networks and public health applications*, GSTF Journal of BioSciences 2 (2012), no. 1.

[125] Iris Gomez-Lopez Olivia Loza and Armin R. Mikler, *Sand: School affiliation network*

discovery algorithm for public health advancement, Annual Global Healthcare Conference Proceedings, GHC 2012, 2012, pp. 99 – 105.

[126] Tore Opsahl, Filip Agneessens, and John Skvoretz, *Node centrality in weighted networks: Generalizing degree and shortest paths*, Social Networks 32 (2010), no. 3, 245–251.

[127] World Health Organization, *Assessing the severity of an influenza pandemic*, May 2009.

[128] Inc. Pearson Education, *Top 50 cities in the u.s. by population and rank*, June 2010.

[129] L. Pellis, N. M. Ferguson, and C. Fraser, *Threshold parameters for a model of epidemic spread among households and workplaces*, Journal of The Royal Society Interface 6 (2009), no. 40, 979–987.

[130] S. Phithakkitnukoon, Z. Smoreda, and P. Olivier, *Socio-geography of human mobility: A study using longitudinal mobile phone data*, PLoS ONE 7 (2012), no. 6, e39253.

[131] A. R. Pinjari, N. Eluru, R. B. Copperman, I. N. Sener, J. Y. Guo, S. Srinivasan, and C. R. Bhat, *Activity-based travel-demand analysis for metropolitan areas in texas: Cemdap models, framework, software architecture and application results*, Research Report, 40808, Texas Department of Transportation, Department of Civil, Architectural and Environmental Engineering, University of Texas Austin, Austin, 2006.

[132] Delgado J. Pujol J.M., Bjar J., *Clustering algorithm for determining community structure in large networks.*, Physical review. E, Statistical, nonlinear, and soft matter physics 74 (1 Pt 2) (2006), 016107.

[133] Jonathan M Read and Matt J Keeling, *Disease evolution on networks: the role of contact structure.*, Proceedings. Biological sciences / The Royal Society 270 (2003), no. 1516, 699–708.

[134] Steven Riley, *Large-scale spatial-transmission models of infectious disease.*, Science (New York, N.Y.) 316 (2007), no. 5829, 1298–301.

[135] Richard B. Rothenberg, John J. Potterat, Donald E. Woodhouse, William W. Darrow, Stephen Q. Muth, and Alden S. Klovdahl, *Choosing a centrality measure: Epi-*

demiologic correlates in the colorado springs study of social networks, Social Networks 17 (1995), no. 34, 273 – 297, ¡ce:title¿Social networks and infectious disease: HIV/AIDS¡/ce:title¿.

[136] Del Valle S., J. Hyman, H. Hethcote, and S. Eubank, *Mixing patterns between age groups in social networks*, Social Networks 29 (2007), no. 4, 539–554.

[137] Z Sadique, Elisabeth J Adams, and William J Edmunds, *Estimating the costs of school closure for mitigating an influenza pandemic*, BMC Public Health 7 (2008), 1–7.

[138] Marcel Salathé and James H Jones, *Dynamics and control of diseases in networks with community structure.*, PLoS computational biology 6 (2010), no. 4, e1000736.

[139] Joanna Schaffhausen, *Swine flu and schools closings: Faq*, May 2009.

[140] Caterina Scoglio, Walter Schumm, Phillip Schumm, Todd Easton, Sohini Roy Chowdhury, Ali Sydney, and Mina Youssef, *Efficient mitigation strategies for epidemics in rural regions.*, PloS one 5 (2010), no. 7, e11569.

[141] Jan C Semenza, Franklin Apfel, Tamsin Rose, and Johan Giesecke, *A network strategy to advance public health in Europe*, Commentary 18 (2008), no. 5, 441–447.

[142] Eduardo D. Sontag, *Kalmans controllability rank condition: from linear to nonlinear*, Mathematical System Theory (1991), 453–462.

[143] Phillip Stroud, Sara Del Valle, Stephen Sydoriak, Jane Riese, and Susan Mniszewski, *Spatial dynamics of pandemic influenza in a massive artificial society*, Journal of Artificial Societies and Social Simulation 10 (2007), no. 4, 9.

[144] Krzysztof Suchecki and Janusz a Hoyst, *Ising model on two connected Barabasi-Albert networks.*, Physical review. E, Statistical, nonlinear, and soft matter physics 74 (2006), no. 1 Pt 1, 011122.

[145] Balázs Szendroi and Gábor Csányi, *Polynomial epidemics and clustering in contact networks.*, Proceedings. Biological sciences / The Royal Society 271 Suppl 5 (2004), no. 1978, S364–6.

[146] Melhusin T., Blake M., and Day S., *An evaluation of synthetic household populations*

for census collection districts created using spatial microsimulation techniques, National Center for Social and Economic Modelling, University of Canberra, Australia (2002).

[147] Matthew W. Tanner, Lisa Sattenspiel, and Lewis Ntaimo, *Finding optimal vaccination strategies under parameter uncertainty using stochastic programming*, Mathematical Biosciences 215 (2008), no. 2, 144 – 151.

[148] Chayant Tantipathananandh, Tanya Berger-Wolf, and David Kempe, *A framework for community identification in dynamic social networks*, Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07 (2007), 717.

[149] Stephen Tennenbaum, *Simple criteria for finding (nearly) optimal vaccination strategies*, Journal of Theoretical Biology 250 (2008), no. 4, 673 – 683.

[150] Amanda L. Traud, Eric D. Kelsic, Peter J. Mucha, and Mason A. Porter, *Comparing community structure to characteristics in online collegiate social networks*, SIAM Review, in press (arXiv:0809.0960), 2010.

[151] Brandes Ulrik, *A faster algorithm for betweenness centrality.*, Journal of Mathematical Sociology 25 (2001), 163 – 177.

[152] ———, *On variants of shortest-path betweenness centrality and their generic computation*, Social Networks 30 (2008), 136 – 145.

[153] Maria D. Van Kerkhove, Tommi Asikainen, Niels G. Becker, Steven Bjorge, Jean-Claude Desenclos, Thais dos Santos, Christophe Fraser, Gabriel M. Leung, Marc Lipsitch, Ira M. Longini, Jr, Emma S. McBryde, Cathy E. Roth, David K. Shay, Derek J. Smith, Jacco Wallinga, Peter J. White, Neil M. Ferguson, Steven Riley, and for the WHO Informal Network for Mathematical Modelling for Pandemic Influenza H1N1 2009 (Working Group on Data Needs), *Studies needed to address public health challenges of the 2009 h1n1 influenza pandemic: Insights from modeling*, PLoS Med 7 (2010), no. 6, e1000275.

[154] S. Venkatachalam and A.R. Mikler, *Modeling infectious diseases using global stochastic*

field simulation, Granular Computing, 2006 IEEE International Conference on, may 2006, pp. 750 – 753.

[155] L. Volkmann, *Edge-connectivity in p-partite graphs*, Journal of Graph Theory 13 (1989), 1–6.

[156] A. G. McKendrick W. O. Kermack, *A Contribution to the Mathematical Theory of Epidemics*, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 115 (1927), no. 772, 700–21.

[157] Jacco Wallinga, Michiel van Boven, and Marc Lipsitch, *Optimizing infectious disease interventions during an emerging epidemic*, Proceedings of the National Academy of Sciences 107 (2010), no. 2, 923–928.

[158] Rui-Sheng Wang, Yong Wang, Xiang-Sun Zhang, and Luonan Chen, *Detecting community structure in complex networks by optimal rearrangement clustering*, Proceedings of the 2007 international conference on Emerging technologies in knowledge discovery and data mining (Berlin, Heidelberg), PAKDD'07, Springer-Verlag, 2007, pp. 119–130.

[159] Stanley Wasserman and Joseph Galaskiewicz, *Advances in social network analysis: research in the social and behavioral sciences*, SAGE, 1994.

[160] Faust K. Wasserman S., *Social network analysis*, Cambridge University Press, 1994.

[161] Duncan J Watts, *A simple model of global cascades on random networks*, Management 99 (2002), no. 9.

[162] Medina D C Watts D.J. Muhamad R. and Dodds P.S., *Multiscale, resurgent epidemics in a hierarchical metapopulation model*, PNAS 102(32) (2005), 11157—-11162.

[163] Eric W. Weisstein, *Voronoi diagram*, January 2013.

[164] G. Welch, S. E. Chick, and J. Koopman, *Effect of concurrent partnerships and sex-act rate on gonorrhea prevalence*, SIMULATION 71 (1998), no. 4, 242–249.

[165] Douglas B. West, *Introduction to Graph Theory (2nd Edition)*, Prentice Hall, August 2000.

[166] Williamd. Wheaton, James C. Cajka, Bernardette M. Chasteen, Diane Kr Wagene,

Philip Cooley, and Ganapathi, *Synthesized Population Databases: A US Geospatial Database for Agent-Based Models*, Database (2009), no. May.

[167] Richard J Williams, *Biology, methodology or chance? The degree distributions of bipartite ecological networks.*, PloS one 6 (2011), no. 3, e17645.

[168] Tianyou Zhang, Soon Hong Soh, Xiuju Fu, Kee Khoon Lee, Limsoon Wong, Stefan Ma, Gaoxi Xiao, and Chee Keong Kwoh, *Hpcgen a fast generator of contact networks of large urban cities for epidemiological studies*, Proceedings of the 2009 International Conference on Computational Intelligence, Modelling and Simulation (Washington, DC, USA), CSSIM '09, IEEE Computer Society, 2009, pp. 198–203.

[169] Yan-Bo Zhou, Shi-Min Cai, Wen-Xu Wang, and Pei-Ling Zhou, *Age-based model for weighted network with general assortative mixing*, Physica A: Statistical Mechanics and its Applications 388 (2009), no. 6, 999–1006.