Razan Alsehibani, May 2024 APPLIED MATHEMATICS

STATISTICAL MODELING OF THE IMPACT OF MULTIPLICITY POOL TESTING AND THE ESTIMATION OF INFECTION AND RECOVERY RATES OF PARTIALLY KNOWN NETWORKS USING HYBRID SAMPLING (122 pages)

Dissertation Advisor: Dr. Omar De La Cruz Cabrera

Detection and control of epidemic outbreaks require effective testing measures, identification of highly-connected members in social networks, as well as the estimation of important epidemic parameters. Pool testing have been proven to be an efficient testing approach to control epidemic spread by reducing the total number of tests. However, pool testing can also be used to improve the accuracy of the testing process. One objective of this thesis is to improve the accuracy of pool testing using the same number of tests as that of individual testing taking into consideration the probability of testing errors and pool multiplicity classification thresholds. Statistical models are developed to evaluate the impact of pool multiplicity classification thresholds on pool testing accuracy using the receiver operating characteristic (ROC) curve and the area under the curve (AUC). The findings indicate that under certain conditions, pool testing multiplicity yields superior testing accuracy compared to individual testing without additional cost.

Modelling the spread of epidemics requires the identification of well-connected nodes in partially known networks where network sampling can be leveraged to detect important nodes in these networks. This thesis extends prior research by developing a hybrid sampling method based on simple random sampling and network sampling to identify well-connected nodes in partially known networks. The performance of the proposed method is evaluated in terms of the Perron eigenvalue of the sampled subnetwork using simulation. The performance evaluation shows that the hybrid sampling method yields significantly superior performance compared to that of simple random sampling. The performance of the different levels of the partial combinations of the hybrid sampling is also evaluated where we find that the different hybrid levels give differing results under varying conditions. The findings reveal that by sampling only a small proportion of the individuals, the hybrid sampling very efficiently identifies well-connected ones.

Finally, recent developments in social networks research enabled researchers to model

the spread of infectious diseases using network structures. This thesis develops statistical models to estimate the infection rate and recovery rate in partially known networks. A joint sampling-infection process is implemented and its outcomes are fed as input to two back tracing algorithms to estimate the health status of individuals during the periods before they are sampled. The infection and recovery rates for partially known networks are then estimated. The findings reveal that the identification of well-connected nodes using the proposed hybrid sampling method leads to significantly lower total number of infections and lower infection peak rates. The results also indicate that one of the two fill-up methods performs better than the other but incurs extra computational time.

STATISTICAL MODELING OF THE IMPACT OF MULTIPLICITY POOL TESTING AND THE ESTIMATION OF INFECTION AND RECOVERY RATES OF PARTIALLY KNOWN NETWORKS USING HYBRID SAMPLING

A dissertation submitted
to Kent State University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

by
Razan Alsehibani

May 2024

Dissertation written by

Razan Alsehibani

B.S., Ursuline College, 2015

M.S., Cleveland State University, 2017

Ph.D., Kent State University, 2024

Approved by

Dr. Omar De La Cruz Cabrera, Chair, Doctoral Dissertation Committee

Dr. Oana Mocioalca, Members, Doctoral Dissertation Committee

Dr. Jun Li,

Dr. Xiang Lian,

Dr. Maxim Dzero,

Accepted by

Dr. Andrew M. Tonge, Chair, Department of Mathematical Sciences

Dr. Mandy Munro-Stasiuk, Dean, College of Arts and Sciences

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my thankfulness to our Creator the Almighty for all the blessings bestowed upon us. Then I would like to express my sincere gratitude to my advisor Professor Omar De La Cruz Cabrera for his valuable guidance and great support. My warmest gratitude is also extended to my parents, husband, children, family, friends, colleagues, and mentors. Words alone cannot express the sincere appreciation I owe them.

CHAPTER 1

Introduction

This dissertation consists of two applications of statistical modeling to important problems in the study of infectious diseases. The first application concerns the use of pool testing (when a screening test for an infectious disease is applied to samples from two or more people mixed together); we show how the combination of pooling with repeated testing can improve the cost and/or the accuracy of the tests. The second application concerns the modeling of the spread of an infectious disease by regarding the population as a network in which an infected individual can infect a susceptible one only if they are connected nodes in the network; crucial properties of the disease spread process (e.g., whether an outbreak is likely to die out or to become an epidemic) depend on properties of the network, and we analyze how these properties can be studied when only a small sample of the whole network is known.

An infectious disease is defined as an illness that can be transmitted by an agent from a sick to a healthy individual [18]. Infectious diseases can reach an epidemic state which means that the number of infected individuals is greater than what is expected in a specific population in a specific region [9, 47] resulting in huge negative impacts on plants, animals, humans, and economic stability. The emergence of COVID-19 for example resulted in severe levels of medical, social, psychological, and economic losses worldwide [6, 10, 32, 55]. Most countries enforced very strict and unimaginable lockdown or restriction measures. Another illustration of the severity of infectious diseases is the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) which is a fatal communicable disease that first appeared in Saudi Arabia in 2012 [88]. According to the World Health Organization (WHO), MERS-CoV is a viral respiratory disease caused by a novel coronavirus and is known to be a zoonotic virus, since it can be transmitted from animals to people [89]. In Saudi Arabia in September 2012, the first case of MERS-CoV was reported by

health officials. During the period between 2012 and 2021, the World Health Organization (WHO) received 2600 reports of MERS-CoV infections. More than 2193 infections resulting in 854 deaths occurred in Saudi Arabia, where 84% of the total cases have been reported [90]. Researchers have shown that the MERS-CoV can be transmitted from animals to animals, from animals to humans, and from humans to humans. Dromedary camels are considered the main reservoir of MERS-CoV where the virus can be transmitted to humans through direct or indirect contact with infected camels [16, 31]. More potent human-to-human transmission has spread in healthcare environments, where healthcare workers are considered to have a high risk for the infection due to their close and frequent contacts with the MERS-CoV's patients. The most extensive outbreak of the MERS-CoV outside the Middle East occurred in the Republic of Korea in 2015, where the 185 cases reported to the World Health Organization resulted in 38 deaths [90]. The MERS-CoV provoked an intense panic among the people in Saudi Arabia and South Korea, and the Ministers of Health in these two countries were fired as a result. Another example, in 2001 specifically in the United Kingdom, the foot-and-mouth disease led to the culling of over six million animals to control the outbreak [37].

This thesis explores several issues related to the detection and control of the spread of infectious diseases. First, pool testing has been utilized to reduce the cost of testing large populations. The remarkable spread of the COVID-19 virus has emphasized the paramount need to test millions of people quickly, efficiently, effectively, and repeatedly in order to curb the proliferation of the disease especially with the presence of asymptomatic cases. Pool testing has been used by practitioners and researchers to improve the efficiency of the testing process. Chapter 2 shows that multiplicity pool testing can also be used to improve the accuracy of testing without incurring extra cost. The chapter develops statistical models for the evaluation of the impact of multiplicity pool testing using the receiver operating characteristic (ROC) curve and the area under the curve (AUC) taking into consideration potential testing errors and pool multiplicity classification thresholds. The results show that pool testing multiplicity provides superior testing accuracy compared to that of individual testing without additional cost given certain conditions.

Second, prior research assumes that the network structure is fully known, however, in many situations the complete information about the network structure is often partially known and collecting the full network structure could be costly and time consuming [70]. Identifying highly-

connected nodes in partially known networks is essential for many applications. Network sampling can be efficiently utilized to identify important nodes in these networks. Based on the compartmental susceptible-infected (SI) virus propagation model in epidemiology, chapter 3 proposes an unknown-known (NK) compartmental model where individuals are sampled and are permanently moved from the unknown state to the known state. The chapter extends prior research by developing a hybrid sampling method based on a combination of simple random sampling and network sampling to identify highly-connected nodes in partially known networks. Simulation is used to evaluate the performance of the proposed sampling method with respect to the Perron eigenvalue of the sampled subgraph. The findings reveal that the hybrid sampling method yields significantly higher performance compared to that of simple random sampling.

The estimation of the infection rate and the recovery rate in partially known networks is essential for the control of epidemics especially among communities of undocumented immigrants. Statistical modeling of the spread of the infectious diseases can enable researchers and decision makers to have a better understanding of the dynamics of these diseases and to develop more practical control measures. Network based individual-level models (ILMs) have emerged as a realistic framework for modelling of the spread of infectious diseases since they account for the heterogeneity in the ability of individuals to infect others. Networks however are often partially known which adds to the complexity of the estimation process. chapter 4 develops statistical models to estimate the infection rate and the recovery rate in partially known networks where the hybrid sampling approach of chapter 3 is employed to identify unknown individuals in these networks. Based on the Susceptible-Infectious-Susceptible (SIS) virus propagation model, the chapter develops a joint infection-sampling process to estimate the health status of individuals in a partially known network. Based on the outcomes of the joint process, the recovery rate is calculated and two back tracing methods are presented to estimate the health status of individuals in the days before their sampling where the infection rate is estimated accordingly. The simulation results show that there are tradeoffs between the two back tracing methods in terms of computational time and estimation accuracy.

CHAPTER 2

Statistical Modeling and Evaluation of the Impact of Multiplicity Classification Thresholds on the COVID-19 Pool Testing Accuracy

2.1 Introduction

An important application of statistical modeling in the study of infectious diseases is improving the efficiency and the effectiveness of the testing process. Infectious disease testing is a costly process especially when there is a need to quickly test large numbers of people and testing also might need to be repeated frequently to monitor the spread of the disease. In this chapter, multiplicity pool testing where samples from every individual are assigned to several pools such that every two individuals are common in at most one pool, is modeled in order to identify conditions under which the cost as well as the accuracy of pool testing can be improved. The impact of the probability of testing errors, multiplicity pool testing classification thresholds, testing tool specifications, and batch size on the efficacy of the pool testing process is evaluated.

The emergence of COVID-19 resulted in growing severe levels of medical, social, psychological, and economic losses [6, 10, 32, 55]. The fast spread of the COVID-19 virus has emphasized the paramount need to test millions of people quickly, efficiently, and effectively in order to curb the proliferation of the disease. Unlike many other infectious diseases, one challenge facing the combat of COVID-19 is that the majority of cases are asymptomatic individuals who can be contagious [2, 27, 110]. Since many of the asymptomatic cases might not be aware of their infection, they need to be quickly identified before they infect others [85].

The economist Robert Dorfman [40] developed a novel pool testing algorithm with the objective of reducing the total number of tests where individual specimens are grouped into a pool to be tested using one test instead of conducting individual testing [40]. If the pool tests negative

then all individuals are declared healthy, otherwise a second round of testing is needed. Pool testing is used in several fields to identify "defective" subjects and there is an increasing need for better understanding of not only how to reduce the number of tests but also to increase the accuracy of the pool testing process.

With the emergence of COVID-19, several researchers and practitioners stressed the importance of utilizing pool testing in controlling the spread of the disease [73]. In February 2020, COVID-19 pool testing methods enabled Stanford University's researchers to quickly identify several positive infections [34]. Pool testing is useful also because negative results can be communicated faster to individuals, since this method reduces the time needed to analyze tests [48]. However, the lack of understanding of how to design an optimal pooling scheme to improve classification accuracy under budget constraints, is hindering screening efforts [15]. Since the Dorfman's pool testing proposal, researchers introduced several algorithms to implement variations of the original method [28, 35, 61, 66, 100, 101].

A main objective of prior research was to improve the efficiency of pool testing by minimizing the number of required tests which consequently reduces the cost of the testing process [19]. Bish et al.[20] develop a robust model based on the Dorfman pool testing method to determine optimal pool size assuming perfect test specificity with the objective of reducing the total number of tests. De Wolff et al. [36] and Verdun et al. [103] perform an evaluation of several pool testing methods to identify under what conditions certain algorithms improve testing efficiency. However, there is a need to improve the accuracy of pool testing to increase the effectiveness of the testing process which will not only curb the spread of epidemic disease but also to ultimately reduce the testing costs. The objective of this chapter is to complement prior research in pool testing by developing models to improve the pool testing accuracy without incurring extra cost, taking into consideration that probability of testing error and pool multiplicity classification threshold.

The contributions of this chapter to pool testing research is multi-fold. First, the relevant literature is reviewed to identify research gaps. Next, the multiplicity pool testing method of [101] is extended by including the probability of testing errors and classification thresholds into the modeling process with the objective of improving the pool testing accuracy. The impact of

several multiplicity classification thresholds on pool testing specificity and pool testing sensitivity is evaluated analytically and through simulation. The ROC and the AUC methods are employed to evaluate the performance of the proposed models. Then, the impact of batch sizes on pool testing accuracy for specific pool testing multiplicity levels is examined. Finally the effect of the manufacturer's test sensitivity on the pool testing accuracy is compared to that of the manufacturer test specificity. Thus, the proposed models extend prior research on pool testing (e.g., [15, 67, 76, 101]).

2.2 Classification of Pool Testing Methods

Pool testing methods are typically classified into hierarchical and non-hierarchical methods. In hierarchical methods, individuals are tested in non-overlapping pools at any specific stage of the testing process. The testing plan at any subsequent stage depends on the results of the tests in the previous stage. The Dorfman method is considered a two-stage hierarchical algorithm. Since the Dorfman's two stage pool testing proposal, several researchers developed extensions of the Dorfman's original method. These extensions include partitioning pools which test positive, into non-overlapping sub pools repeatedly, until all positive individuals are identified through individual tests. For example, Finucan [45] developed a three-stage pool testing method where initially a master pool that contains all individuals is tested, then sub pools are tested in the middle stage, and finally individual retesting is conducted in the final stage.

Hierarchical pool testing methods are typically called "adaptive" because the test is conducted in stages or rounds and the results of any stage depend on the results of previous stages. These testing methods require a first round of testing to test the pools and a second round of tests for individuals in positive pools. This second round might require extracting samples which could overload laboratories especially if samples are extracted manually. These methods might not be efficient particularly in situations where the results need to be delivered quickly. Although adaptive pool testing methods might require fewer tests, non-hierarchical or "non-adaptive" pool testing schemes; where overlapping pool testing is completed in a single step, allow for parallel testing and do not require extra samples be extracted, which improves the testing efficiency [101, 102].

The array pool testing approach is the most common type of non-hierarchical pool testing algorithms where individual specimens are arranged into rows and columns of an array. Row pools and column pools are simultaneously tested in parallel [92]. In two-dimensional array pool testing algorithms, every individual is typically a member of two pools: one row pool and one column pool such that a sample of each individual is located at the intersection of a unique pair of pools. In the first stage of the testing process, all row pools and all column pools are tested. All individuals who are at the intersection of a positive row pool and a positive column pool need to be retested individually. Under the assumption that tests are error-free, the decision is simple in that all individuals that are at the intersection of a positive row pool and a positive column pool are declared positive [67, 76]. However, under the more realistic assumption that tests might have errors, the decision is more complicated, since it is possible that a row pool tests positive with no column pool testing positive, and the other way around [58].

Typically, tests are subject to errors which can occur for many reasons such as an erroneous testing tool or an inadequate test implementation. Therefore, there is a need to account for these testing errors. Kim et al. [67] developed a two-dimensional pool testing method that takes into consideration testing errors where entire row pools or column pools might be retested. The authors also developed a three-stage pool testing method by adding a master pool and derived models for the expected number of tests for their pool array testing algorithms. Hudgens and Kim [61] analyze the impact of the pool size on the expected number of tests for square array pool testing without master pools and provide bounds for optimal pool sizes in case of homogeneous populations assuming error-free tests.

Kim and Hudgens [66] analyze the performance of three-dimensional array pool testing under the assumption that the population is homogeneous. They find that three-dimensional array pool testing can reduce the expected number of tests compared to two-dimensional array pool testing. However, according to the method of Kim and Hudgens [66], individuals are arranged in three dimensional cubes and the pooling is performed along hyperplanes. This way, every individual becomes a member in three pools but any two hyperplanes will intersect in more than one individual rather than a single individual, which might negatively affect the performance of the algorithm. Mutesa et al. [82] propose an adaptive algorithm for pooling subsamples based on a hypercube

structure that, at low prevalence, accurately identifies individuals infected with SARS-CoV-2 using a small number of tests and few rounds of testing.

Haber et al. [53] has reviewed recent developments in pool testing research with a focus on Dorfman's algorithm for a homogeneous population using several case studies. The authors indicate that most prior research on pool testing focuses on minimizing the expected number of tests and they call for paying more attention to the benefits of pool testing in improving the accuracy of the testing process. The preprint Fargion et al. [43] indicates that for homogeneous populations, array pool testing might yield "mirror" false positives as a result of individuals who are healthy being located at the intersection of a positive row pool and a positive column pool. Yelin et al. [108] report that pool testing can detect COVID-19 infections in pools of up to 64 members. A recent study found that a pool size of five is cost-effective for monitoring the COVID-19 spread at Northeastern University [48].

Recent research analyzes non-adaptive pool testing methods where each individual is assigned to several pools. Hanel and Thurner [54] study the impact of test accuracy on the selection of the pool size with the objective of minimizing the number of tests. They propose to test replicas of the same pool to improve the accuracy on the expense of the efficiency in terms of the number of tests and indicate that no more than two replicas of the same pool improve the testing accuracy while the same pool-replicas of three are worthwhile only in the case of large pool sizes. Another line of research assigns every individual to several pools such that every two individuals are common in at most one pool assuming a homogeneous population and error-free tests [101]. The number of pools to which an individual is assigned is called *pooling multiplicity* where all individuals are assigned to an equal number of pools. A multiplicity of k, means every individual is a member in exactly k pools such that every individual is tested k times but in different pools. However, the assumption that tests are error-free is not realistic in many situations. Testing errors on the individual level happen, when an individual specimen who is sick (healthy) is incorrectly declared as negative (positive).

A popular method to detect sick individuals using non-adaptive pool testing is the combinatorial orthogonal matching pursuit (COMP) which is attributed to [63]. According to COMP,

any individual in a negative pool is declared definitely healthy while the remaining individuals are considered possibly sick. Since COMP is considered a noiseless pool testing method, hence it produces no 'false negatives' but might yield a high rate of 'false positives'.
The presence of testing errors can introduce false negatives which can be mitigated using the noisy COMP (NCOMP) algorithm. The NCOMP name is attributed to [4] and the basic concept has been introduced by [25, 26]. According to NCOMP, any individual who is a member in a certain minimum number of positive pools is declared sick, otherwise it is declared healthy.

The performance of both COMP and NCOMP is analyzed by [99] who indicate that COMP is a special case of NCOMP. Lets denote the number of pools in which the individual is a member as the membership size m. The authors state that imposing further conditions on the multipooling matrices; other than constant pool size, constant membership size, and dot product between columns of at most one, will not reduce the expected number of false positives in COMP and NCOMP. They also show that increasing the membership size decreases the pooling sensitivity but increases the pooling specificity [99]. A variant of the COMP algorithm is the Definite Defective (DD) algorithm [3] which performs better than COMP in terms of the number of tests in cases when the prevalence level is low [4]. The DD starts by using COMP to identify the definitely healthy. Next, any individual who is the only potentially sick in a positive pool is declared sick, while all other remaining individuals are declared healthy. Since the DD is noiseless, hence it might produce high false negative rates. To overcome the limitations of the DD, a noisy DD algorithm has been developed in [98] in which the test outcomes are based on some pre-specified threshold values. The Noisy DD has been shown to perform better than NCOMP in terms of the number of tests [4] as well as in probability of detection success. However, as indicated earlier, the COMP paradigm is a basic step in other pool testing algorithms and therefore our method is based on the noisy COMP.

Given the growing importance of improving the accuracy of pool testing, recently a smart pool testing software application has been developed based on the Tapestry hybrid pool testing [49] where the COMP is used as an initial stage in the testing process. According to this method, the COMP identifies definitely healthy individuals who are consequently excluded from further investigation. Ghosh et al. [49] excludes not only healthy individuals but also negative pools from further investigation and they analyze the performance of several compressed sensing methods as

a second stage of testing following the COMP stage to identify the health status of the remaining possibly sick individuals. In the Tapestry pool testing, each individual is assigned as a member into three pools and any two individuals are common in at most one pool. The test outcomes of individuals are then classified into three classes: sick, healthy, and unidentified [24] and therefore a second round of testing might be needed in rare cases. A hybrid approach is also applied in [91] where a compressed sensing algorithm is used as a second stage of testing after excluding the definitely healthy individuals identified using COMP in the first stage where based on the final testing outcome, individuals are classified as either healthy or sick.

A main difference between our method and Tapestry is that unlike Tapestry in which each individual contributes to exactly three pools, our method is more general since each individual can be a member in any number of pools, up to a certain maximum value as shown by [99], provided that any two individuals are members in exactly one pool. Another limitation of the Tapestry is that it is based on an algorithmically two-stage approach where in the first stage the COMP is applied and then the output of the COMP stage is fed as an input to the CS stage. However, in such two-stage method, errors committed in the first stage are irreversible in the second stage. For example, in cases when the manufacturer's sensitivity of the test is low then, if the COMP stage erroneously declares a specific individual to be negative, then this individual will not be considered into the second stage the CS stage.

Altman and Bland [13] developed two main measures of testing accuracy: the test sensitivity and the test specificity. The test sensitivity S_e is the proportion of the true positives that are classified correctly by the test while the test specificity S_p is the proportion of the true negatives that are classified correctly by the test [13]. During the pool testing process, the test might be applied on the same sample multiple times whether individually or as a member of a pool. Therefore, the test sensitivity S_e and the test specificity S_p as quoted by the manufacturer are not sufficient to estimate the probability of an individual being correctly diagnosed by the pool testing method. Consequently researchers developed other measures of testing accuracy for pool testing including pooling sensitivity and pooling specificity. The pooling sensitivity PS_e is defined as the probability that an individual is classified as positive by the pool testing algorithm, provided that the individual is sick. While, the pooling specificity PS_p is the probability that an individual is

classified as negative by the pool testing algorithm, provided that the individual is healthy [57].

Unlike prior research in pool testing that mainly attempts to minimize the number of tests, this thesis aims to improve pool testing accuracy using the same number of tests used by individual testing considering the probability of testing errors and pool multiplicity classification thresholds. This is accomplished by adopting a pooling multiplicity approach where every individual is assigned to several pools such that every two individuals are common in at most one pool. Statistical models are developed to evaluate the impact of pool multiplicity classification thresholds on pool testing accuracy using the receiver operating characteristic (ROC) curve and the area under the curve (AUC).

2.3 Statistical Models

Prior research developed several pools formation methods like the Shifted Transversal Pool Testing Design [102] which seeks to reduce the number of joint membership of individuals in any given pool, and at the same time generates pools that intersect in an equal number of locations. These two properties can improve the non-adaptive detection process significantly. A multipool matrix can be generated using the Shifted Transversal Design method, when the pool size is chosen to be a prime number and can be generated using the more general Reed- Solomon method [96] when the pool size is chosen to be a power of a prime (see [99] for detailed illustrations). Given these designs, Schumacher and Tauffer [99] define a multipool as a structure in which all pools are of equal size, every individual has the same membership size (the number of pools in which the individual is a member), and any two pools intersect in at most one location. The authors also prove that a multipool matrix exists if and only if the membership size has an upper bound for the case when the pool size is a prime or a power of a prime. They demonstrate that this upper bound is equal to the pool size plus one, given the pool size is the square root of the population size.

Our method is based on a multipool design [99, 101], where individuals are grouped into pools of size n such that every individual is a member in exactly n pools and such that any two individuals are common members in exactly one pool. Table 3 provides a list of our model parameters. The N individual samples can be arranged in an $n \times n$ square array with the number of

rows denoted as J and the number of columns denoted as K where $J = K$ for square arrays. Then the pools can be generated by partitioning individuals equally into J row pools and also partitioning individuals equally into K column pools. For example, Row-Pool(j) contains individuals who are located on row j. Individuals are marked by their coordinates or location where an individual who is located on the intersection of the j^{th} row and the k^{th} column is denoted by I_{jk}. This individual becomes a member in the Row-Pool (j) and also a member in Column-pool (k). In other words, every individual's sample is divided into k subsamples and assigned to k different equally-sized pools of size n where no two individuals are common members in more than one pool.

Note that any two pools do not intersect in more than one location if n is a prime or a power of a prime [99, 101]. The pools formation process starts by generating n patterns of n pools each. For example, one pattern could consist of the set of the J row pools and another pattern could consists of the set of the K column pools. Patterns also can be generated along diagonals where an additional pattern can consist of all the D main diagonal pools (running from the upper left corner to the lower-right corner), where $J = K = D$. More patterns can be generated along other types of diagonals as well [101].

In order to simplify the coding process during simulation, n patterns that consist of 5 diagonal patterns, rather than row patterns, column patterns and diagonal patterns, are developed. As an example, lets assume $N = 25$ individuals, hence, $n = J = K = \sqrt{N} = 5$. Table 1(a) shows the diagonal vertical 0-offset (column) pattern with 5 pools where every pool is marked by a distinct color. Compared to prior research, our method has the advantage of reducing the memory requirements significantly since the pool membership data is being calculated by the algorithm rather than storing them as a pooling matrix.

The diagonal patterns are formed by declaring a horizontal offset value (h_offset) and a vertical offset value (v_offset) for each pattern. We simplify the process by fixing $v\_offset = 1$ throughout the pattern and pool formation process. The 1st pool of the 2^{nd} pattern (the 1st diagonal pattern or the main diagonal pattern) is generated using $h\_offset = 1$; meaning that the first pool in this pattern starts with the upper-right corner individual (I_{00}) then horizontally we move right by ($h\_offset = 1$) location and vertically we move down by ($v\_offset = 1$) location, and so

on, until we include n individuals into this pool. The 2^{nd} pool in this pattern starts with individual (I_{01}) and the remaining members of this pool are generated similarly but using $h\_offset = 2$ and $v\_offset = 1$. In general, the i^{th} pool in this pattern will be generated starting with individual (I_{0i}) using $h\_offset = 1$ and $v\_offset = 1$. To avoid the "fall-off" (exceeding) the array boundaries, the arithmetic *modulo* n function can be used to wrap the pool generation process where the process starts from 0 whenever we reach $(n-1)$[101]. As in Tauffer (2020), lets represent the members of pool l that belongs to pattern m as the set $PP(l,m)$ where:

$$PP(l,m) = \{I_{j,(l+j\times m)(mod\ n)} : j = 0, 1, \cdots, n-1\},\ \forall l, m = 0, 1, \cdots, n-1.$$

Table 1(b) shows the 5 pools of the 1^{st} pattern (the 1^{st} diagonal pattern).

Likewise, the 1^{st} pool of the second pattern (the 2^{nd} diagonal pattern) is generated starting with individual (I_{00}) but with $v\_offset = 2$, and so on. In general, the i^{th} pool of the j^{th} diagonal pattern is generated starting with (I_{0i}) but with $h\_offset = j$. Table 1(c), Table 1(d), and Table 1(e) show the pool formation for the remaining patterns.

Table 1: An example of 5 patterns with 5 pools each. Every small table represents a pattern, and each color represents a pool in the pattern.

(a) Pattern 1

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25

(b) Pattern 2

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25

(c) Pattern 3

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25

(d) Pattern 4

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25

(e) Pattern 5

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25

Assume a population size of N individuals and a multiplicity level $n = \sqrt{N}$ where the N individuals can be arranged in an $n \times n$ square array. Each individual's sample is broken up into n sub-samples and the sub-samples are assigned to n different pools in n different diagonal patterns such that every two individuals are common in at most one pool. Consistent with prior research, it is assumed that:

- *Assumption 1. The true statuses of individuals are independent and identically distributed random variables with probability p of being sick.*
- *Assumption 2. Given that the true status of an individual I_{ij} who is a member of pool P_k is sick; i.e., $(Y_{ij} = 1)$, then pool P_k tests positive with probability S_e and testes negative (i.e. false negative) with a probability $1 - S_e$. This implies that the pool test sensitivity is independent of the pool size.*
- *Assumption 3. Given that all the individuals in pool P_k are healthy, then pool P_k tests positive (i.e. false positive) with a probability $1 - S_p$ and testes negative (i.e. true negative) with a probability S_p. This implies that the pool test specificity is independent of the pool size.*
- *Assumption 4. The test outcomes of intersecting pools are conditionally independent of each other.*
- *Assumption 5. The pool size n is a prime number.*

A homogeneous population is assumed and the prevalence is defined as $p = P(Y = 1)$, where Y, represents the true status of an individual. Similar to Kim et al. [67], McMahan et al. [76], Aprahamian et al. [14], and Hitt [57] we assume that the true statuses of individuals are mutually independent random variables. Let $X_{ij} = 1$ if the test outcome of the individual at the location ij is diagnosed positive; $X_{ij} = 0$ otherwise. Let $Y_{ij} = 1$ if the true status of the individual at the location ij, is sick; $Y_{ij} = 0$ otherwise.

Let the manufacturer-reported specificity and sensitivity be denoted by $S_p = P(X = 0|Y = 0)$ and $S_e = P(X = 1|Y = 1)$, respectively. Assume that S_e and S_p are known, diagnostic test dependent, independent of the individual's covariates, independent of the number of individuals per pool, i.e. no dilution. Two main types of testing approaches: individual testing, and pool testing, are compared. Let the individual testing specificity and sensitivity be denoted by IS_p and IS_e, respectively, and let the pool testing specificity and sensitivity be denoted by PS_p and PS_e, respectively.

Let P_k represent pool number k for $k = 1, \cdots, N$. The set of pools to which individual I_{ij} belongs is denoted as SP_{ij}, i.e.

$$SP_{ij} = \{P_k : I_{ij} \in P_k, k = 1, \cdots, N\}$$

for every pattern $l = 0, \cdots, n-1$,

$$k = ((j - (i \times l)) mod\ n) + 1 + l \times n$$

where $i = 0, \cdots, n-1$ is the row number and $j = 0, \cdots, n-1$ is the column number of the location of individual I_{ij}. Every individual I_{ij} belongs to exactly n pools. Pools are arranged in n patterns of n pools each. Rather than storing the pool information as a binary pooling matrix, our algorithm assigns individuals to pools at run-time as can be seen from Example 2 in Table 2 below. This feature has the advantage of saving memory considerably.

Table 2: An example of a pooling matrix generation process with n=25. Individual's positions are fixed in every colored matrix. The numerical value represents the pool number while the color represents the pattern number.

(a) Pattern 1

1	2	3	4	5
1	2	3	4	5
1	2	3	4	5
1	2	3	4	5
1	2	3	4	5

(b) Pattern 2

1	2	3	4	5
5	1	2	3	4
4	5	1	2	3
3	4	5	1	2
2	3	4	5	1

(c) Pattern 3

1	2	3	4	5
4	5	1	2	3
2	3	4	5	1
5	1	2	3	4
3	4	5	1	2

(d) Pattern 4

1	2	3	4	5
3	4	5	1	2
5	1	2	3	4
2	3	4	5	1
4	5	1	2	3

(e) Pattern 5

1	2	3	4	5
2	3	4	5	1
3	4	5	1	2
4	5	1	2	3
5	1	2	3	4

Throughout this thesis, the terms "positive" and "negative" are used to refer to the test outcomes (i.e., to indicate the presence or absence of the disease based on the test outcomes, respectively), while we use the terms "sick" and "healthy" to indicate the "true status" of an individual. To simplify the presentation, the term "individual" is used to refer both to the individual and to the sample taken from the individual.

In the multiplicity pool testing method ($MPTM$) we define n different classification protocols to identify positive individuals. Each protocol identifies positive individuals based on a minimum threshold value representing the number of positive pools in which that individual is a

Table 3: The multiplicity pool testing parameters.

Parameter	Description
N	The population size
n	The pool size, $n = \sqrt{N}$
p	The incidence of the infection
I_{ij}	The individual who is in the i^{th} row and the j^{th} column
X_{ij}	The test outcome of individual I_{ij}
Y_{ij}	The true status of Individual I_{ij}
P_k	The pool number k, for $k = 1, \cdots, N$
XP_k	The test outcome of pool P_k
SP_{ij}	The set of pools to which individual I_{ij} belongs
$0 \leq S_e \leq 1$	Test sensitivity
$0 \leq S_p \leq 1$	Test specificity
IS_p	Individual testing specificity
IS_e	Individual testing sensitivity
$PS_p(\mathcal{T})$	Pool testing specificity for a threshold of $\mathcal{T} = 1, \cdots, n$
$PS_e(\mathcal{T})$	Pool testing sensitivity for a threshold of $\mathcal{T} = 1, \cdots, n$

member. In particular, protocol i indicates that an individual will be declared positive if the test outcome of at least i of its pools turn positive (classification threshold value of i). The multiplicity pool testing sensitivity with a threshold of $\mathcal{T}$ assuming a homogeneous population has been derived by [99] as follows:

Let $n = \sqrt{N}$, where n is a prime number, be the multiplicity level and let $\mathcal{T}$ be the classification threshold, where $\mathcal{T} = 1, \cdots, n$, then the multiplicity pool testing sensitivity $PS_e(\mathcal{T})$ can be expressed as

$$PS_e(\mathcal{T}) = \sum_{i=\mathcal{T}}^{n} \binom{n}{i} \Big(1 - (1 - S_e)\big(S_p(1 - pS_e)^{n-1}\big)\Big)^{i} \Big((1 - S_e)\big(S_p(1 - pS_e)^{n-1}\big)\Big)^{n-i} \tag{2.1}$$

for any values of S_e, n, and $\mathcal{T}$.

According to the proposed multiplicity pool testing method, an individual who is in the i^{th} row an the j^{th} column is declared positive (i.e. X_{ij} =1) if at least $\mathcal{T}$ of its pools test positive for any specific classification threshold $\mathcal{T} = 1, \cdots, n$.
More formally, for individual I_{ij},

$$\text{if } \big(\textstyle\sum_{k=1:P_k \in SP_{ij}}^{N} XP_k\big) \geq \mathcal{T}), \text{ then } X_{ij} = 1,$$

where $i = 0, \cdots, n-1$; $j = 0, \cdots, n-1$; and $k = 1, \cdots, N$,

The multiplicity pool testing specificity with a threshold of $\mathcal{T}$ assuming a homogeneous population has also been derived by [99] as follows:

Let $n = \sqrt{N}$, where n is a prime number, be the multiplicity level and let $\mathcal{T}$ be the classification threshold, where $\mathcal{T} = 1, \cdots, n$, then the Multiplicity pool testing specificity $PS_p(\mathcal{T})$ can be expressed as

$$PS_p(\mathcal{T}) = 1 - \sum_{i=\mathcal{T}}^{n} \binom{n}{i} \left(1 - \left(S_p(1-pS_e)^{n-1}\right)\right)^i \left(S_p(1-pS_e)^{n-1}\right)^{n-i} \tag{2.2}$$

for any values of p, S_e, and S_p, n, and $\mathcal{T}$.

The outline of the multiplicity pool testing algorithm is presented in Algorithm 1 below. The R code implementation of the algorithm is available at https://github.com/ralsehib/Multiplicity-Pool-Testing.git. Our R code implementation has the advantage of being concise as well as supporting parallelism. The code generates pools at run-time rather than storing the pool information as a binary matrix which saves memory significantly. The R software package used is RStudio version 1.1.419.

Algorithm 1 Multiplicity Pool Testing

for every individual **do**
    Generate infection status based on the prevalence level p
end for
for every pattern **do**
    Form n unique pools that belong to the current pattern
    Identify pools that have any sick individual, mark these pools as sick
    Based on S_e and S_p, simulate the test of every pool
    if the pool is sick, **then**
        The pool will test positive with a probability equal to S_e
    else
        The pool will test negative with a probability equal to $1 - S_p$
    end if
    for every individual, **do**
        cumulatively, identify the number of its positive pools
    end for
end for
for every individual, **do**
    estimate the individual status, based on the given threshold
end for
Finally, the pooling sensitivity and specificity are estimated based on the true status of individual and the multiplicity pool testing outcome of the individual

2.3.1 The Area Under the ROC Curve (AUC)

The impact of the pool testing conditions on the joint accuracy measures (pool testing sensitivity and pool testing specificity) of classification in diagnostic settings can be analyzed using the receiver operating characteristic (*ROC*) curve which is a commonly used visual illustration. ROC curves display the true positive rates versus the false-positive rates for a range of classification threshold values. The ROC curve describes the ability of the test to identify sick from healthy individuals and it can also be used in identifying the threshold value that gives the optimal testing accuracy [46]. The ROC curve is a plot of *sensitivity* versus $(1 - \textit{specificity})$ for a range of possible classification threshold values and it represents a trade-off between sensitivity and specificity.

An ROC curve starts at the $(0, 0)$ coordinate, corresponding to the case where all test results are negative and ends at the $(1, 1)$ coordinate, corresponding to the case where all test results are positive. The typical lower limit of the ROC curve is a diagonal line that connects the

lower left and the upper right corners of the graph with an area under the curve of 0.5. In other words, the diagonal line that connects the $(0,0)$ and $(1,1)$ points represents the ROC curve of a random test that does not distinguish sick from healthy individuals. ROC curves that lie above this diagonal has some diagnostic ability where the farther the ROC curve from the diagonal (the closer to the upper left-hand corner), the better the diagnostic accuracy of the test [46, 87].

A popular measure of test accuracy is the area under the ROC curve, denoted as (AUC) [87]. The AUC is calculated using the trapezoidal rule. Denote the coordinate of the curve given the threshold i as $(x_i, y_i) \quad \forall i = 1, \cdots, n$. Let the initial coordinate of the ROC curves be always $(0,0)$. Note that,

$$x_i = 1 - PS_p(i) \quad \forall i = 1, \cdots, n$$

and,

$$y_i = PS_e(i) \quad \forall i = 1, \cdots, n$$

Hence, the total area under the curve $(TAUC)$ can be expressed as:

$$TAUC = \sum_{i=1}^{n} (((1 - PS_p(i+1)) - (1 - PS_p(i))) \times PS_e(i)) + (\frac{1}{2} \times ((1 - PS_p(i+1)) - (1 - PS_p(i))) \times (PS_e(i) - PS_e(i+1)))$$

2.4 Results and Discussion

The performance of the multiplicity pool testing is evaluated and the overall testing accuracy is estimated through simulation using the R software package. The simulation code is efficiently developed by considering n diagonal patterns, rather than row, column, and diagonal patterns. The true status of individuals will be randomly generated based on a Bernoulli distribution with the prevalence level of the disease p as a given probability parameter. The simulation of our method generates a "sick" true status with a probability of p and generates a "healthy" true status with probability $1-p$. Individual test outcomes are estimated based on a Bernoulli distribution with the manufacturer testing sensitivity S_e or the manufacturer testing specificity S_p as given probability parameters.

After estimating the test outcomes of all individuals through pool testing, the values of the accuracy measures are calculated in a way that is similar to that of individual testing explained above. For both individual testing and pool testing simulations, we run 1000 independent repetitions to take variability into consideration where averages across the 1000 repetitions are reported.

2.4.1 Accuracy Measures vs. Prevalence

Assume a population of $N = 25$, then 25 pools are formed where each pool contains $\sqrt{N} = 5$ members. Every individual will be a member in exactly 5 different pools in five different patterns i.e. a zero step-based diagonal pool (column pool), a one step-based diagonal pool, a two step-based diagonal pool, a three step-based diagonal pool, and a four step-based diagonal pool.

Let's assume 5 different levels of prevalence ranging between 0.005 and 0.20, as well as 3 different values of S_p and S_e ranging from 0.90 to 0.99. The pool testing multiplicity level is assumed constant with a value of 5 throughout the first stage of the simulation. Comparison of the multiplicity pool testing and the individual test accuracy measures: specificity and sensitivity, versus different values of prevalence between 0.005 and 0.20 for individual testing and pool testing are shown in Fig 1 and Fig 2. The left figure presents the testing specificity and the right figure presents the testing sensitivity. The solid line represents the test accuracy measures. Different colors of curves represent different pool testing classification thresholds. The black color represents individual testing. The red color represents pool testing with a threshold of 5. The green color represents pool testing with a threshold of 4. The blue color represents pool testing with a threshold of 3. The light-blue color represents pool testing with a threshold of 2. The pink color represents pool testing with a threshold of 1.

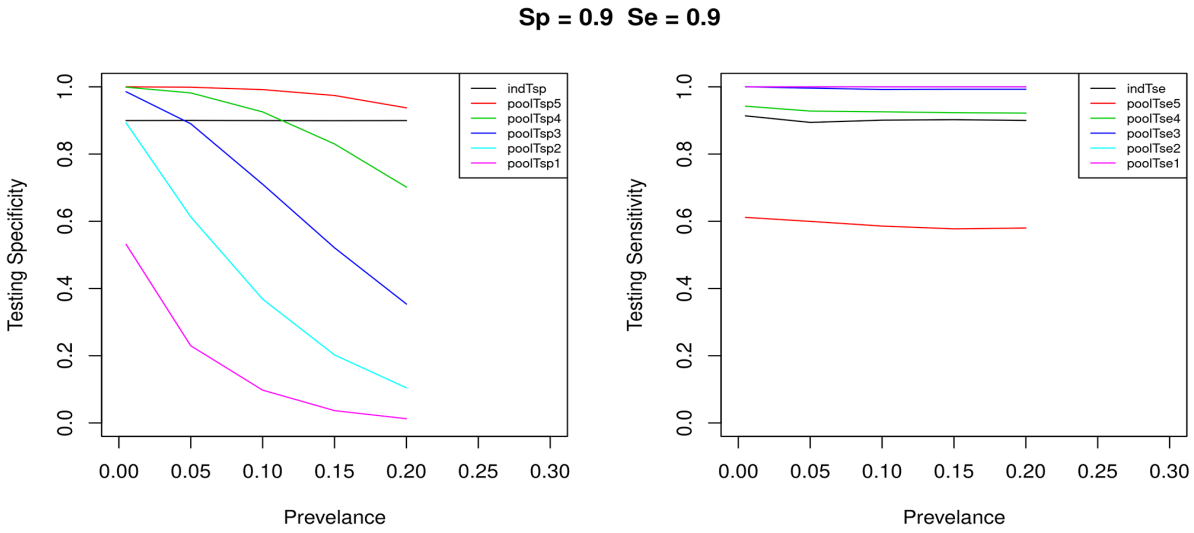


Figure 1: Comparison of the test accuracy measures given $S_p = 0.90$ and $S_e = 0.90$ for individual testing and pool testing.

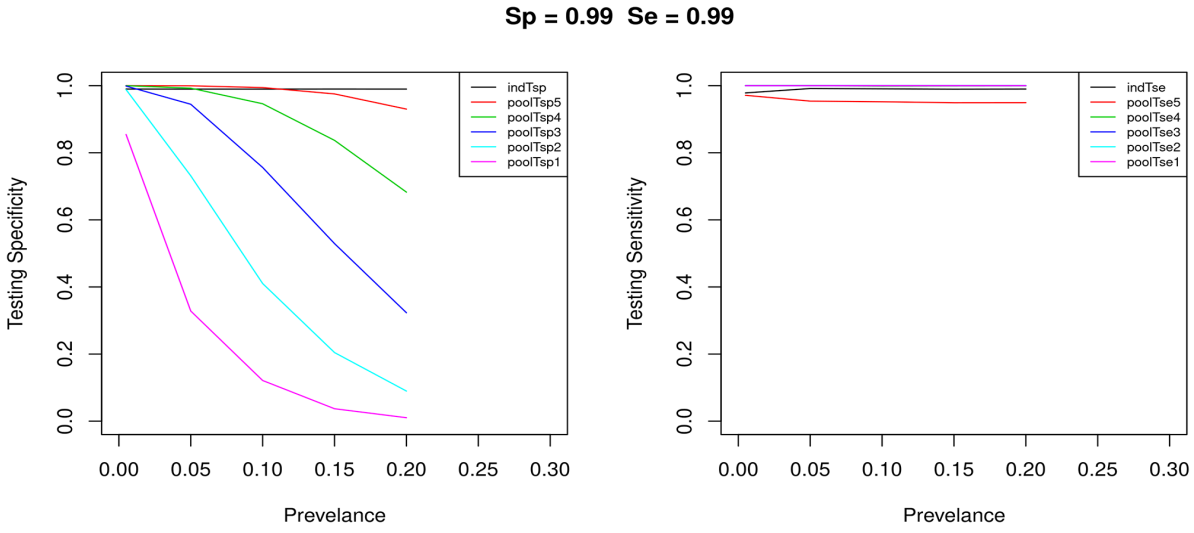


Figure 2: Comparison of the test accuracy measures given $S_p = 0.99$ and $S_e = 0.99$ for individual testing and pool testing.

From the experimental results, it can be concluded that under certain conditions, the, multiplicity pool testing method gives higher accuracy compared to individual testing without additional cost. For example, when the prevalence level is low; e.g, $p \leq 0.1$, classification threshold 4 gives higher pool testing sensitivity and higher pool testing specificity compared to individual

testing (manufacture reported sensitivity and specificity). This is particularly true for the case when the manufacture reported sensitivity and the manufacturer reported specificity are low; i.e. $S_e = 0.9$ and $S_p = 0.9$. Even for the case when the manufacture reported sensitivity and the manufacturer reported specificity are high; i.e. $S_e = 0.99$ and $S_p = 0.99$, classification threshold 4 gives higher pool testing sensitivity and higher pool testing specificity compared to individual testing, but only when the prevalence level is ≤ 0.05.

The benefit gained in accuracy is higher for the case when the prevalence level is low and the manufacturer reported specificity and sensitivity are low. For example, for $p = 0.050$, from Fig 1, when $S_e = 0.90$ and $S_p = 0.90$ a threshold of 1 yields an improvement gain in testing sensitivity of pool testing over individual testing of about 11% compared to an improvement gain of about 1% when $S_e = 0.99$ and $S_p = 0.99$ as shown in Fig 2. The simulation results show that when the prevalence is high and the test tool manufacturer's reported accuracy is high then there is no need to use pool testing to improve accuracy because under these conditions the individual accuracy is higher than the pool testing accuracy. To validate the models, we compared the performance of the simulation models to that of the theoretical models given in Equation (2.1) and Equation (2.2). Figure 3 and Figure 4 present the comparison results of pool testing specificity and pool testing sensitivity for the simulation compared to that of the theoretical models. The results show that results of the simulation closely match those of the theoretical models.

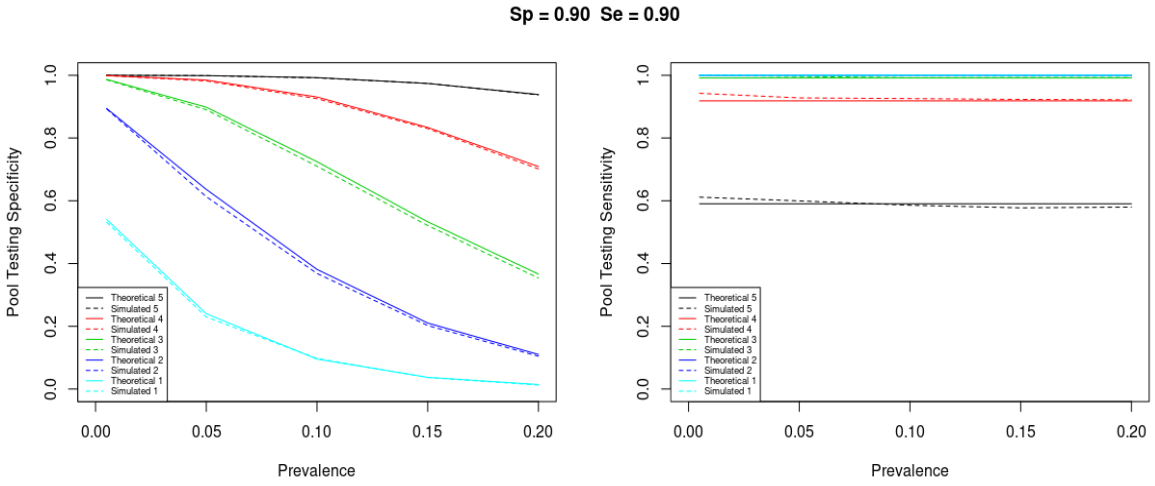


Figure 3: Comparison of the performance of the simulation and theoretical models given Sp=0.90 and Se=0.90.

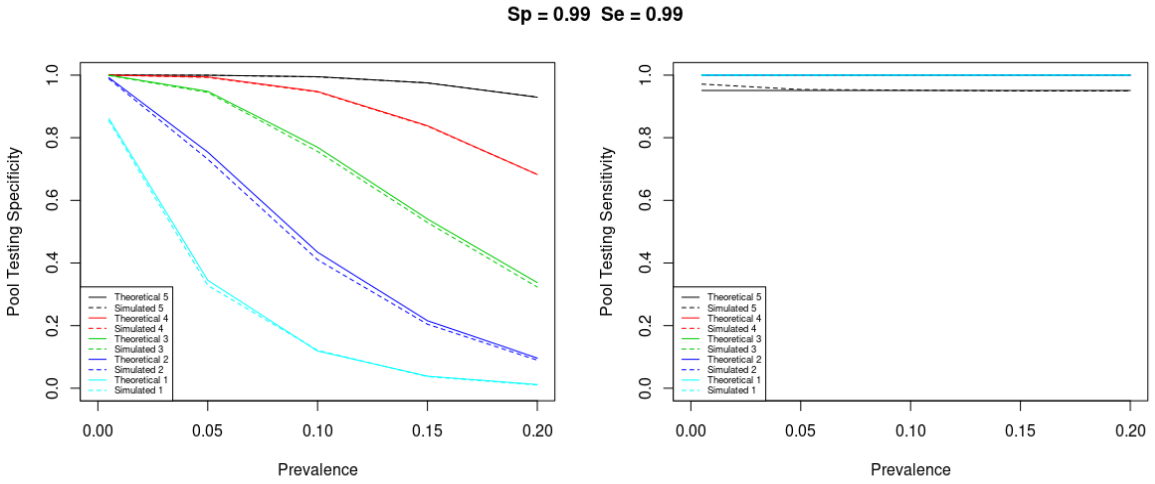


Figure 4: Comparison of the performance of the simulation and theoretical models given Sp=0.99 and Se=0.99.

Typically, false negatives might lead to significant risky consequences compared to false positives since false positives could be subject to further verification testing [93, 106]. These consequences include worsening medical complications of the infected individual and the continuous spread of the disease, especially if the individual has many contacts. Therefore, there is a paramount

need to develop testing methods that mainly reduce the probability of false negatives as a main objective and at the same time reduce the probability of false positives as a secondary objective. The probability of false negative where a small number of sick individuals are missed, is associated with a high value of test sensitivity [50, 74]. In other words, the probability of false negative is inversely proportional to the test sensitivity. The results show that different classification thresholds give different levels of pool testing accuracy depending on the pool testing conditions. For example, Fig 4 shows that, for prevalence level of $p = 0.005$, classification threshold 4 gives higher pool testing sensitivity and higher pool testing specificity compared to individual testing. However, if perfect pool testing sensitivity; i.e. PS_e of 1 is required, then classification threshold 3 could be chosen even if its pool testing specificity is less than that of threshold 4, since it still gives higher pool testing specificity compared to individual testing.

2.4.2 Classification Accuracy

In the ROC curve we plot the $(1 - \textit{specificity})$ on the x-axis and the sensitivity on the y-axis where each line on the plot represents a different prevalence level p. The performance of the pool testing method is simulated for a population of 25 individuals with a multiplicity level of 5 using different threshold values and different testing conditions. To examine the impact of different levels of prevalence on the classification accuracy, we let $p = 0.005, 0.05, 0.1, 0.15,$ and 0.2. For each p, we experiment with different values of the manufacturer-reported specificity S_p and sensitivity S_e, where we let the values of S_p and $S_e = 0.90, 0.95,$ and 0.99 resulting in 9 different combinations of testing accuracy measures based on five classification threshold values $\mathcal{T} = 1, 2, \cdots, 5$. Given the values of p, S_p, and S_e, the ROC curves enable us to identify the classification threshold value that should be selected to get the optimal testing accuracy (the highest true positive rate and at the same time the lowest false positive rate). Fig 5 shows the ROC curves under several testing conditions, given $S_p = 0.90$ and $S_e = 0.90$.

The experimental results show that different pool testing conditions (e.g. prevalence, S_e, and S_p) might require different classification thresholds to obtain the best pool testing accuracy. For example, in the case of a population of 25, a manufacturer reported specificity ($S_p = 0.90$), manufacturer reported sensitivity ($S_e = 0.90$), and a prevalence ($p = 0.005$), pool testing sensitivity

of 1 can be achieved for several threshold values. From Fig 5, when the prevalence for example is 0.005, we can see that the false positive rate in pool testing for the threshold value of 3 is approximately 2% while the false positive rate is equal to 47% for the threshold value of 1, where in both cases, the pool testing sensitivity is 1. This example shows that the classification threshold should be selected cleverly to obtain the highest testing accuracy.

For a batch size of 25, from the ROCs in Fig 5 we observe that as the prevalence value decreases the pool testing performance in terms of accuracy increases, as expected. Also, from this figure, we observe that as the manufacturer reported accuracy increases, the accuracy of the pool testing method improves, as measured by the ROCs, for the different levels of prevalence. Also, we observe that different pool testing methods yield different testing accuracy levels depending on the testing conditions i.e. prevalence, manufacturer reported specificity, manufacturer reported sensitivity and the threshold value. Therefore, there is a need to develop a software tool or an application to associate the different threshold values with the testing conditions in order to identify the classification thresholds that give the highest performance in terms of accuracy.

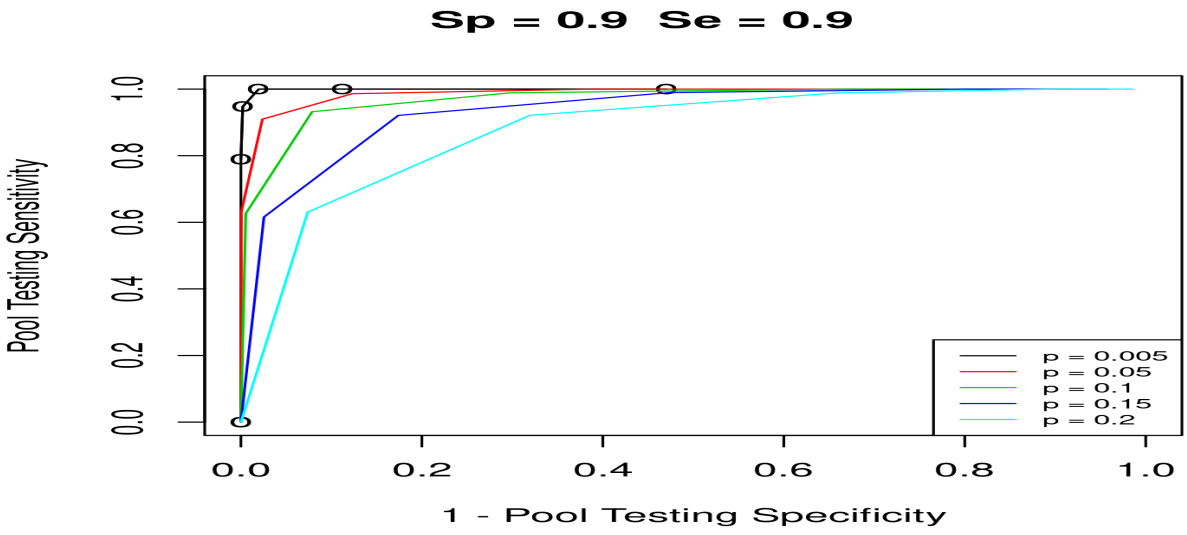


Figure 5: ROC curve for several prevalence levels given $S_p = 0.90$ and $S_e = 0.90$.

2.4.3 Impact of the Manufacturer's Sensitivity and Specificity on the AUC

The AUC for full multiplicity pool testing using nine tests with different values of manufacturer's test sensitivity and specificity for a prevalence of 0.05 is shown on Fig 6. The figure shows that for each specific value of manufacturer's test sensitivity, the AUCs of the different tests are almost similar to each other. Typically, higher manufacturer's test sensitivity and manufacturer's test specificity incurs higher cost. The findings show that significant cost savings can be earned through multiplicity pool testing using low–cost tests. For example, Fig 7 shows the improvement in the pool testing accuracy, measured by the AUC, in the case of prevalence level of 0.05. From the figure, it is clear that using a low-cost test yields accuracy that is comparable to a high cost-test, based on multiplicity pool testing.

In other words, using a test of low manufacturer's specificity might incur lower testing costs and at the same time gives comparable pool testing accuracy to other higher-cost tests. For example, from Fig 6, for a prevalence level of $p = 0.05$, using a test of manufacturer's sensitivity and specificity of $S_e = 0.90$ and $S_p = 0.90$, respectively gives an AUC of 0.980 while using a test of manufacturer's sensitivity and specificity of $S_e = 0.9$ and $S_p = 0.99$, respectively gives an AUC of 0.988 Note that the percentage of gain in accuracy is less than 0.82% On the other hand, from Fig 6, for a prevalence level of $p = 0.05$, using a test of manufacturer's sensitivity and specificity of $S_e = 0.90$ and $S_p = 0.90$, respectively gives an AUC of 0.980 while using a test of manufacturer's sensitivity and specificity of $S_e = 0.99$ and $S_p = 0.90$, respectively gives an AUC of 0.998 Note that the percentage of gain in accuracy is about 1.8%. Therefore, as can be seen from the figure, a low-cost test leads to accuracy that is comparable to a high cost-test. However, the manufacturer's test sensitivity has more significant impact on the accuracy of pool testing compared to that of manufacturer's test specificity. In other words, from multiplicity pool testing perspective, if the test cost is a critical factor in selecting a certain type of test (among tests of the same manufacturer's test sensitivity), then a test of lower manufacturer's test specificity might be an optimal option.

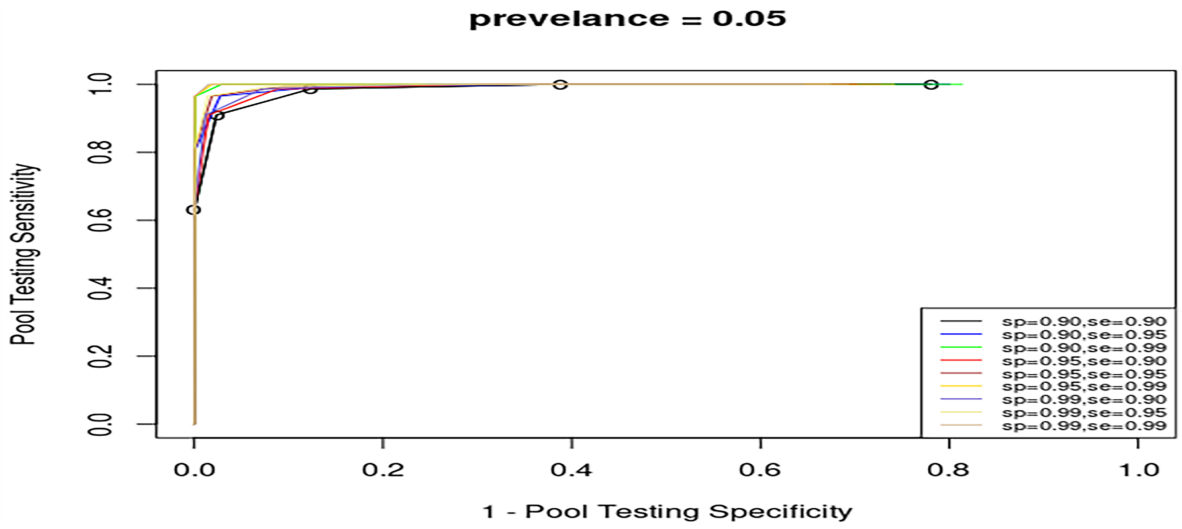


Figure 6: ROC curve for several manufacturer testing specificity and sensitivity levels given $p = 0.05$.

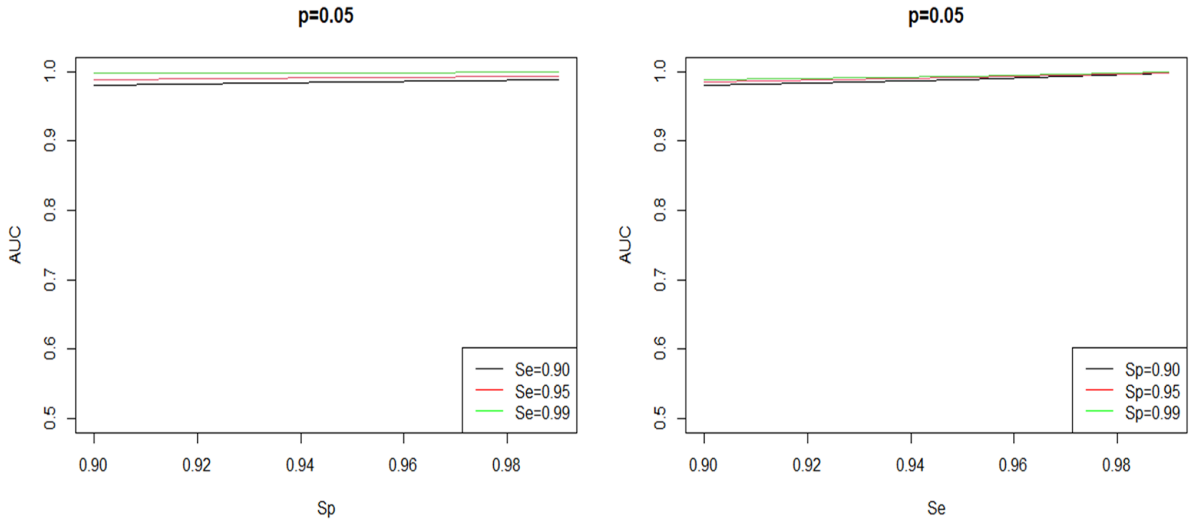


Figure 7: The improvement in the pool testing accuracy, measured by the AUC as a function in S_e and S_p.

2.4.4 The Impact of the Batch Size on the AUC

A set of N individuals can be partitioned into batches of different sizes before applying pool testing. For example, a set of 100 individuals can be divided into 4 batches of size of 25 individuals each or

can be divided into 25 batches of size of 4 individuals each, where pool testing can be conducted on each batch. We analyze the impact of different batch sizes on the pool testing specificity, by considering different batch sizes and different prevalence levels.

The performance of the diagnostic test can be evaluated by estimating the area under the ROC curve (AUC). The AUC takes values between 0 and 1 and AUCs that have values close to 1 indicate high testing accuracy. Once the ROC curves are generated, the AUC for every curve can be estimated using either the Trapezoidal rule or the Simpson's rule. In this thesis, we use the Trapezoidal rule since the generated curves are not smooth curves because they are developed mainly by connecting several points with straight lines. The estimated AUCs are visually displayed using color-coded heat maps to represent the pool testing accuracy given different prevalence levels and batch sizes. Fig 8 through Fig 16 show the heat maps of the AUC for each combination of the manufacturer's reported sensitivity of 0.90, 0.95, and 0.99 and the manufacturer's reported specificity of 0.90, 0.95, and 0.99. Observe that for example from Fig 8 there is a banana-shaped pattern representing the performance of different batch sizes under different levels of prevalence. As can be seen from Fig 8, for low prevalence levels, pool testing using large batch sizes has higher accuracy than pool testing using small batch sizes. While for high prevalence levels, pool testing using small batch sizes performs better than pool testing using large batch sizes. For every combination of the manufacturer's reported sensitivity and the manufacturer's reported specificity, the AUC results can be used as a guide for selecting the recommended batch size for every given prevalence value. A future research direction is to develop a software tool or an application to associate the different batch sizes with the testing conditions in order to identify the batch size that gives the highest performance in terms of accuracy.

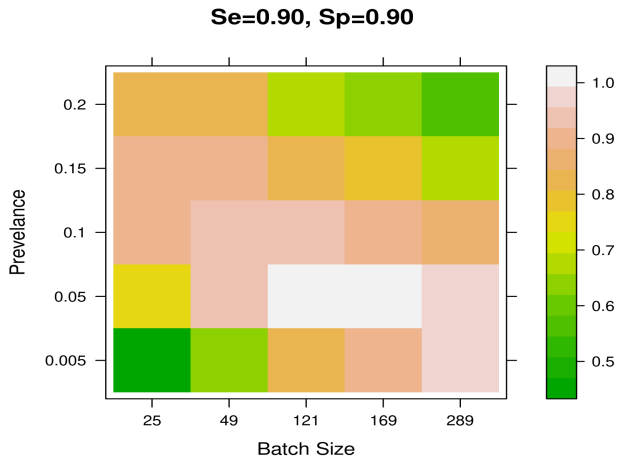


Figure 8: AUC Heat Map for $Se = 0.90$ and $Sp = 0.90$.

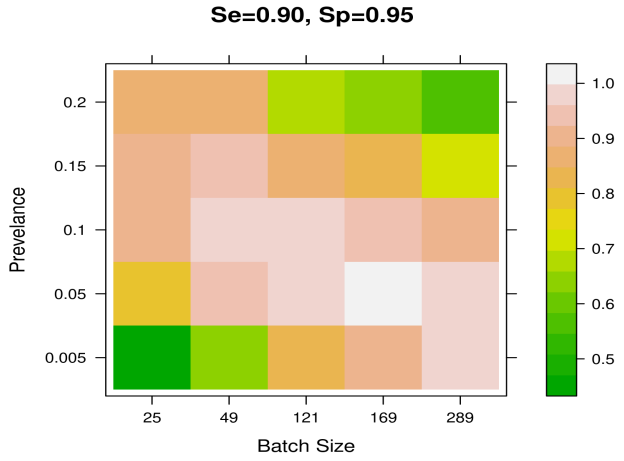


Figure 9: AUC Heat Map for $Se = 0.90$ and $Sp = 0.95$.

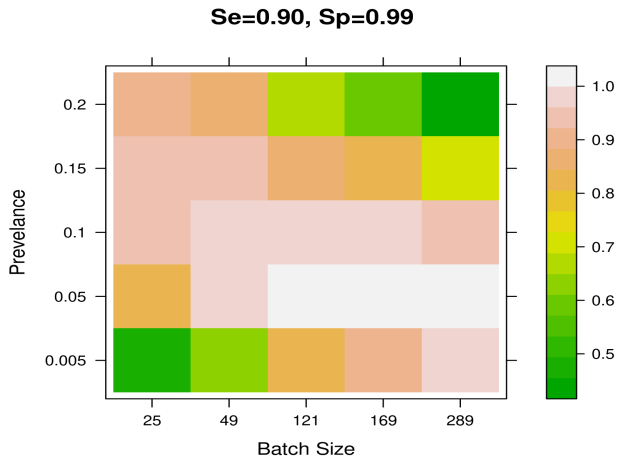


Figure 10: AUC Heat Map for $Se = 0.90$ and $Sp = 0.99$.

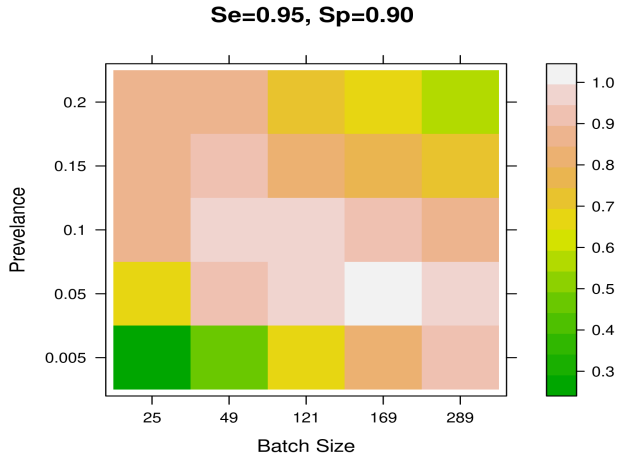


Figure 11: AUC Heat Map for $Se = 0.95$ and $Sp = 0.90$.

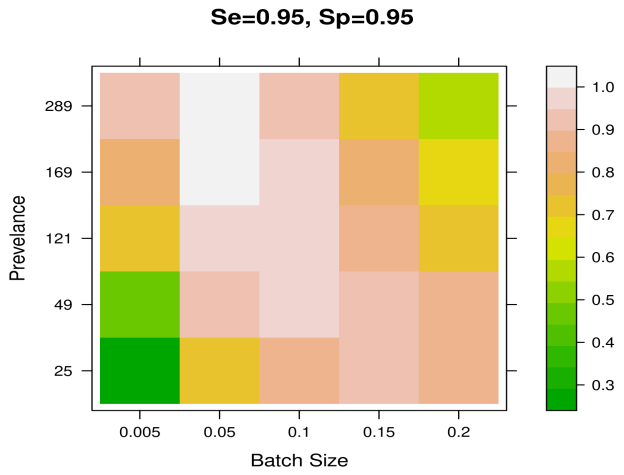


Figure 12: AUC Heat Map for $Se = 0.95$ and $Sp = 0.95$.

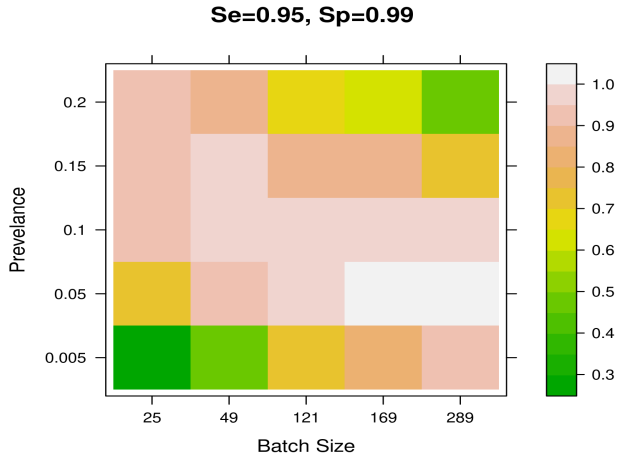


Figure 13: AUC Heat Map for $Se = 0.95$ and $Sp = 0.99$.

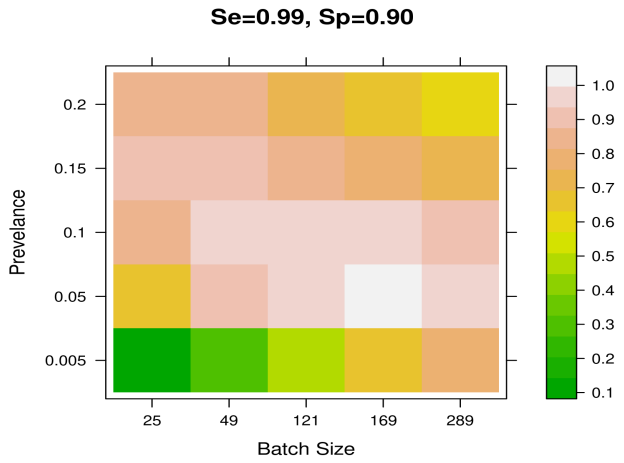


Figure 14: AUC Heat Map for $Se = 0.99$ and $Sp = 0.90$.

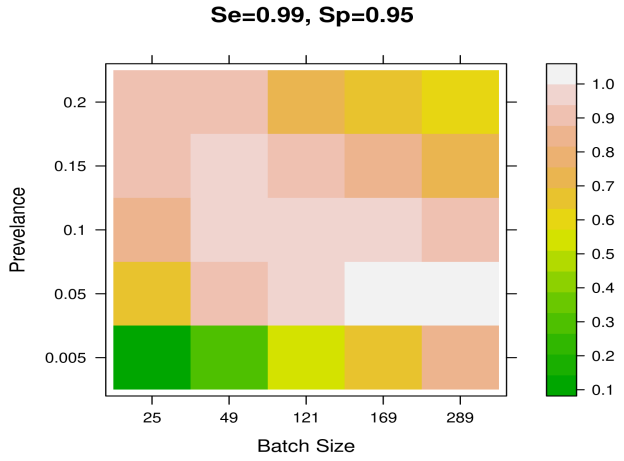


Figure 15: AUC Heat Map for $Se = 0.90$ and $Sp = 0.95$.

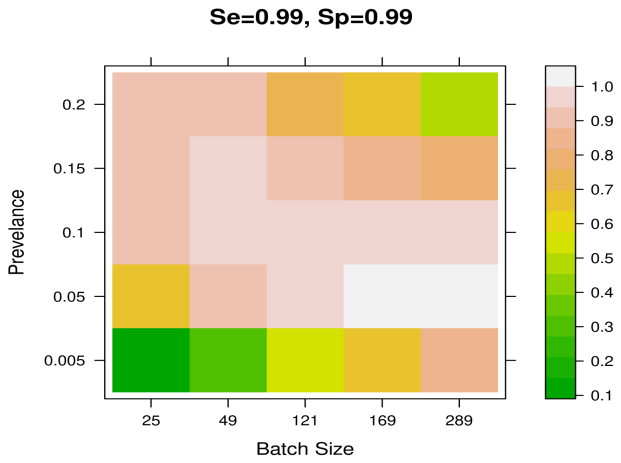


Figure 16: AUC Heat Map for $Se = 0.99$ and $Sp = 0.99$.

2.5 Implications

This research has demonstrated that pool testing can be used to improve the testing accuracy; i.e. testing sensitivity as well as testing specificity. In particular, it is demonstrated that under certain conditions, the multiplicity pool testing performs better compared to individual testing in terms of testing accuracy without the need for extra tests. Furthermore, the impact of several classification threshold values on the testing accuracy is analyzed. For instance, a pool testing algorithm might use a threshold value of 1 in all different conditions whereas the proposed approach has the advantage that it enables decision makers to identify under what conditions to get higher testing sensitivity (higher true positive rate) and at the same time higher testing specificity (lower false positive rate). For example, for a population of size 25, $S_p = 0.90$, $S_e = 0.90$, and the prevalence $p = 0.005$, Fig 5 shows that a simple approach might choose a threshold value of 1, which would give the highest pool testing sensitivity $PS_e = 1$, but would give a false positive rate $(FPR) = 47\%$, while a smarter approach would recommend a threshold value of 3 which will give us the same pooling sensitivity $PS_e = 1$, but with much lower FPR of 2%. Additionally, the results indicate that different batch sizes can be used intelligently, depending on the prevalence level of the disease, to improve the performance of the pool testing method.

2.6 Future Work

The independence assumptions simplify the modelling process but might not be realistic in real testing situations since contamination due to handling errors before pooling might increase the false positive rate. Therefore, future research can relax the independence assumption since lab handling errors might affect several pools concurrently. A growing line of research has started to investigate the impact of dilution on pool testing accuracy especially for large pool sizes since dilution might increase the rate of false negatives. For example, when applying Tapestry to real testing situations, the authors in [24] conducted three real experiments. In the first one, they accounted for dilution by increasing the pooled amount of a positive sample. They indicated however, that the impact of dilution was not significant and therefore in the other two experiments, they pooled equal amounts of samples regardless of whether the individual was healthy or sick.

A recent study reported a pooling sensitivity of 93%, 91%, and 81% for pools of size, 5, 10, and 50 respectively, using a PCR test with 99% manufacturer reported sensitivity for individual tests. The authors suggest that pool testing could be used mainly for the screening of asymptomatic individuals [17]. Another study which used pool testing for the screening of 7400 healthcare workers, revealed that in situations of low prevalence levels, dilution as a result of pooling did not yield significant loss in testing sensitivity [42]. A contemporary study proposes to use swab pooling in which pools are formed at the time of sample collection. Under this scheme, two swabs are collected from every individual such that the first is stored in an individual tube and the other is inserted in a pool with a size of up to 16 different samples, collected individually within a period of one hour. The study focused on asymptomatic individuals in a low-prevalence setting where authors report that the dilution impact was insignificant since swab pooling and individual testing delivered highly similar performance in terms of diagnostic accuracy [30]. Therefore, there is a need for future research to analyze the impact of dilution on the multiplicity pool testing process.

CHAPTER 3

Neighbor Voting Hybrid Sampling for the Identification of Highly Connected Nodes in Partially Known Networks

3.1 Introduction

A central application of statistical modeling in the study of infectious diseases when the population is represented as a network is analyzing how the properties of the network affect the spread of the disease when only a sample of the network is known. An important factor that we focus on is identifying important individuals in a network through sampling in order to improve the control of the spread of the disease. In this chapter, we study how hybrid sampling where network sampling is jointly implemented with simple random sampling can be developed to identify highly-connected nodes. Measures can be taken to move sampled individuals from a high risk category to a low risk category in order to control the spread of infectious diseases.

The COVID-19 epidemic has caused a significant level of damage to the medical, economic, and social aspects of life worldwide. Most countries enforced very strict and unimaginable lockdown or restriction measures. One factor that has contributed to the fast spread of the disease was the fact that the social networks among populations are only partially-known. In Saudi Arabia for example, the government observed that the early cases of the COVID-19 were brought to the country by citizens who returned back from abroad but did not disclose that they have visited infected countries, since some of these infected countries were under a travel ban. Therefore, in the early days of the epidemic, and in order to isolate the suspected cases, the government announced that citizens who are returning to the country must disclose the names of all countries they have visited including banned ones and assured travelers that disclosure will not lead to any penalties [11, 8]. Another observation was that the disease was spreading rapidly among undoc-

umented immigrants communities in Saudi Arabia [7, 77]. Therefore, the government announced that any undocumented immigrant will receive free COVID-19 testing and free treatment if necessary without fearing deportation or any other legal consequences [12]. The main outcome of these two decisions was the improved knowledge about these partially-known networks which assisted in avoiding uncontrollable outbreaks of the disease. Jointly with other precautions, these two measures led Saudi Arabia to relatively suffer few hospitalizations and fatalities compared to other countries, where the death rate (per 100,000 population) was 25 compared to the international average of 94, as of 14 November 2021 [23].

A new line of research has lately emerged that takes advantage of developments in network theory where relations among individuals can be accurately and efficiently represented using a network structure. Social, professional, spatial, and temporal networks among individuals have received increasing attention since new technologies have enabled the identification of relations among individuals as well as the representation of these relations. Graphs and their corresponding adjacency networks can be used to represent virtual and physical relations where nodes represent individuals and links represent the interactions among them like for example whether they live or work together at a certain place. Complex systems like the models of infectious disease spread can be represented as a graph that consists of a set of nodes connected through a set of edges since a graph structure can capture many aspects of the sophisticated behavior of the proliferation of the disease in a community. Early studies in epidemic modelling assume that the probability of contact between any two individuals in the population is the same, i.e. an individual contact is uniform with the entire population. In real life however, certain groups of individuals typically have a set of regular contacts, where for example, co-workers are highly likely to have a high level of contacts among themselves compared to their contacts with the remaining population. Therefore, the assumption of contacts homogeneity is not realistic. The contacts between individuals can be represented as a network, where the nodes represent the individuals and the edges represent the contact between the individuals. Furthermore, individuals typically differ in their characteristics and behavior; and therefore researchers started to use network-based individual-level modelling (ILM) to analyze many processes including the spread of infectious diseases. To model the spread of infectious disease, possible contacts between individuals can be represented as a network where

under this model, individuals can infect each other, if and only if they share a common link in the network. A degree of a node in a contact network is the number of links entering or leaving a node from other nodes, which indicates the number of possible contacts among these nodes [79].

3.2 Partially Known Networks

Statistical modeling of infectious diseases can enable researchers and decision makers to have a better understanding of the spread of infectious diseases and to develop more effective control measures to contain an outbreak. Identification of highly-connected individuals is essential to the success of infectious disease control campaigns since these individuals can be given priority for immunization strategies especially when resources are limited [71]. Also, monitoring the health status of well-connected individuals can be a powerful tool in estimating the level of spread of the disease. Inoculating well-connected individuals can significantly enable us to avoid drastic lockdown measures.

Prior research assumes that the network structure is fully known, but in many situations the network structure is only partially known and collecting information to build a full network could be often costly and time consuming [70]. Another challenge that researchers face in developing full contact networks of epidemics spread from real information is that some of the basic information for creating the contact network is considered private i.e. individuals are not always comfortable sharing this information. Furthermore, individuals may have many contacts, and therefore they may not recall these contacts easily or accurately [64]. For the case of diseases that can be transmitted from animals to humans, the contact network can represent animals trading trends i.e. if herd X trades off with herd Y then a link exists between X and Y. Some animal owners might try to hide the relationships between their farms and other infected farms to avoid the culling of animals or prevention from selling animal's products. Data about the network structure might not be fully known also because many social networks permit only restricted access to the network data [109, 104].

Therefore, sampling has been used as an effective tool to estimate properties of large or unknown networks as highlighted in a comprehensive survey by [60]. Zhou, et al. [109] develop

two algorithms; named the Circulated Neighbors and the Grouped Neighbors Random Walk sampling algorithms. Both algorithms set a specific sampled node as the current node where the first algorithm considers the neighbors of the current node one by one for sampling uniformly without replacement in a circular fashion until the list of the neighbors of the current node is exhausted. The other algorithm on the other hand, groups the neighbors of the current node based on a specific criterion, then groups are selected uniformly circularly, and one node from the selected group is uniformly sampled. Li, et al. [104] propose a network sampling algorithm that leverages the information about the candidate node and the latest sampled node (current node). The algorithm samples the candidate node with a probability that is directly proportional to the degree of the candidate node, but inversely proportional to the number of common neighbors between the candidate node and the current node. The authors also present two variants of the proposed algorithms where the first variant includes the concept of non-backtracking random walks, while the other variant considers more than one visited nodes as current nodes. Xu and Lee [107] propose a framework for a hybrid sampling approach in which crawling-based sampling is combined with random-jump sampling. The crawling-based method typically samples the neighbors of the current node while the random-jump sampling method selects the next node independently of the already sampled ones.

Almugahwi et al. [1] present a novel sampling algorithm where the probability of sampling a node is proportional to the number of its neighbors who have been previously sampled. We call this method the partial Neighbor Voting Sampling algorithm. A review of recent research in network sampling is provided in [104]. Prior research indicates that there are significant differences in the behavior of known and uknown populations [51, 77]. Therefore, it is of a paramount importance to develop fast and effective methods to sample partially-known networks to identify high risk (highly-connected) nodes during the early stages of the spread of infectious diseases. To model the spread of a disease in partially known networks, there is a need first to identify individuals as well as their contacts. Unlike the hybrid sampling method of [107] which is based on random jump sampling and random crawling sampling, our study develops a hybrid sampling method based on simple random sampling and partial neighbor voting network sampling to identify well-connected nodes in partially known networks. Using simulation, performance of the proposed method is evaluated

in terms of the largest eigenvalue of the sampled subgraph. In this thesis, the terms nodes and individuals are used interchangeably and similarly, the terms links, edges, and contacts are used interchangeably.

3.2.1 The Unknown-Known Compartmental Model

Leveraging knowledge from neighboring nodes can significantly improve the identification of well-connected nodes. According to the friendship paradox, [44], neighbors of a randomly sampled individual typically have higher degree than that of the randomly sampled one. Cohen et al, [33] builds on the friendship paradox observation to develop a sampling strategy that improves the epidemic threshold for an SIR virus propagation model. Christakis, N. A., Fowler, J. H. [29] demonstrate that neighbors of randomly selected individuals can act as sensors that are useful for the early detection of outbreaks. Based on the friendship paradox, [70] develop two network sampling strategies named the local strategy and the global strategy to sample high degree nodes in a partially-known network. The local method uniformly samples a random seed node and then randomly select one of its neighbors and adds it to the sampled list while the global method randomly selects one or more of the neighbors of the seed node and adds them to the sampled list. The authors demonstrate that these two methods perform better than a simple random sampling method in terms of yielding higher degree nodes. Novick, Y., Bar-Noy, A [86] call the above method; where rather than sampling a random node, a neighbor of a randomly sampled node is selected, they call it the Random Neighbor Sampling method and develop a cost model to analyze its cost.

The focus of this thesis is on partially known networks where we use sampling to identify new individuals. Benefiting from the research outcomes in the field of compartmental virus propagation models (VPM) of infectious diseases, where the Susceptible-Infectious (SI) model is the basic VPM model [83], we formulate the sampling process using two compartments: Known (K) and Unknown (N). Similar to the SI concept, for the sampling process, we assume that any individual who becomes known remains known as time $t \to \infty$ [83]. Note that the S and I states are considered mutually exclusive. Likewise, the K and N compartments are considered mutually exclusive too. The sampling process is formulated as a Markov Chain process based on the NK compartmental concept. For the sampling process, we developed a network-based ILM model which is modeled

in terms of the NK compartmental frameworks, such that at any given time, each individual can be in any of the two states: N or K. The unknown state represents individuals who have not been sampled yet, but they can be sampled if they have contacts with a sampled individual. The known state represents individuals who have been sampled where individuals move permanently from the N state to the K state; $N \rightarrow K$, during the sampling process. Identification of unknown individuals requires sampling and contact tracing. Two commonly used sampling methods are the simple random sampling (SRS) and network sampling (NS). Simple random sampling implies identifying individuals in a random manner repeatedly until the end of the sampling process regardless of the network structure. The network sampling method on the other hand benefits from the contact tracing concept [62], where any sampled individual is asked to list his adjacency information, then the next nodes are repeatedly sampled based on the evolving sampled subnetwork.

3.2.2 Preliminaries

Complex systems like the models of infectious disease spread can be represented as a graph that consists of set of nodes connected through a set of edges since a graph structure can capture many aspects of the sophisticated behavior of the proliferation of the disease in a community. A graph is mathematically described by its adjacency matrix that consists of a set of vertices (or nodes) $V = \{v_1, v_2, \cdots, v_n\}$ and a set of edges (or links) $E = \{e_1, e_2, \cdots, e_m\}$. In an undirected graph, each element of E is an unordered pair $e_k = (v_i, v_j)$ of elements of V, which connects v_i and v_j (and vice versa). Any two nodes that are linked to each other are called neighbors or adjacent nodes. In unidercted graph, the normal contact degree d_i for individual i is defined as the total number of contacts between individual i and all other individuals. A graph of n nodes can be described by its $n \times n$ adjacency matrix $A = [A_{ij}]$, where $A_{ij} = 1$ if there is a link that connects nodes v_i and v_j , and $A_{ij} = 0$ otherwise. The entries of the diagonal elements are set to 0 to represent the no self-loop assumption. The nondiagonal elements are binary indicating that the network is unweighted where all links, when they exist, are equally important. We assume that the underlying network is symmetric, undirected, and that multiple links between any pair of nodes are not allowed.

Given the infection rate β and the recovery rate γ, the epidemic threshold (τ) is a critical value above which outbreaks can lead to epidemics [59] such that when $\tau > \frac{\beta}{\gamma}$, then an outbreak dies

out, otherwise, when $\tau < \frac{\beta}{\gamma}$, the outbreak becomes an epidemic [105, 84]. The Perron eigenvalue (largest eigenvalue) λ of the adjacency matrix A provides important insights about the communicability of the different nodes in the network. The epidemic threshold has been shown to be the reciprocal of the Perron eigenvalue [105, 84] and therefore, computing an estimate of the Perron eigenvalue of the adjacency matrix A has a great importance. In this chapter, we present a method to approximate λ if A is only partially known. Typically, sampled nodes are moved from a high-risk to a low-risk category where they can be vaccinated or isolated for instance. This step reduces the overall risk in the full network that contains both known and unknown nodes. Computations can be accomplished more efficiently by using an approximated smaller matrix. Almugahwi et al. [1] provide a review of recent developments in research in matrix approximation. When A is only partially known, we assume that a set of full columns of A have been already identified through sampling. The sampled submatrix represents an approximation of the full matrix.

Table 4: The parameters of the hybrid sampling model.

Parameter	Description
n	Number of individuals
t	The simulation period (in days)
w	Vector of sampled individuals
e_i	A one-hot vector; a binary vector with zero everywhere except for the candidate individual in the i^{th} position
edge	Number of edges to be add in each time step in Barabási–Albert model
power	The preferential attachment factor (1: linear) in Barabási–Albert model
prob	The probability that two nodes being connected in Erdős–Rényi model
common	Number of nodes that connect the two clusters
blobN	The cluster size
S	Susceptible state
I	Infectious state
N	Unknown class
K	Known class
$Zeta(Z)$	Probability of sampling by random
δ	Probability of sampling through a network
β	Infection transmission rate
γ	Recovery rate
A	Known symmetric contact adjacency matrix
$Y_{i,t}$	The sampled status of individual i at time t
n_u	The total number of unsampled individuals at iteration u

3.3 Statistical Models

Our network sampling is based on sampling without replacement where only unsampled nodes are considered for sampling in every iteration. Our model utilizes network locality [75], where the status of neighbors affects the probability of unknown individual i becoming known. We develop a method to move individuals from the unknown status to the known status using network sampling. Table 4 provides a list of our model parameters.

3.3.1 The Neighbor Voting Sampling Method

A novel network sampling concept is presented in [1], where a node is sampled proportional to the number of its sampled neighbors. We call this method the partial neighbor voting sampling (NVS) method. Since the partial neighbor voting sampling (NVS) algorithm samples individuals according to the votes of their neighbors, hence, one limitation of the neighbor voting algorithm is that it might get stuck in one cluster in case we have multi-cluster network especially if the network is disconnected or weakly connected. Therefore, we extend the method proposed by [1] using a hybrid sampling (HS) method consisting of a combination of simple random sampling and neighbor voting sampling. We implement the HS method by introducing a weight factor Z to represent the weight of the random sampling component. We define Z such that $0 \leq Z \leq 1$, where $Z = 0$ denotes pure network sampling and $Z = 1$ denotes pure random sampling.

Definition 1. The sampled contact degree m_i for an individual i, is the total number of contacts between individual i and all other *sampled* individuals. It is calculated as

$$m_i = w^T A e_i.$$

3.3.2 Hybrid Sampling Method

We propose a hybrid sampling (HS) method that leverages both the simple random sampling (SRS) method and the partial neighbor voting (NVS) network sampling method. In this thesis, simple random sampling refers to uniform random sampling without replacement.

Definition 2. Let, Z be given where $0 \leq Z \leq 1$. The hybrid sampling probability (P_{HS}) is defined as a linear combination of the uniform random sampling probability (P_{SRS}) and the neighbor voting network sampling probability (P_{NVS}) as follows

$$P_{HS} = Z \times P_{SRS} + (1 - Z)P_{NVS}.$$

Definition 3. Let the SFS be the so-far sampled set. Then, in any iteration u of the sampling process, the number of unsampled individuals n_u can be calculated as

$$n_u = \sum_{j=1}^{n} \mathbb{1}(j)$$

where $\mathbb{1}$ is the Kronecker delta such that

$$\mathbb{1}(j) = \begin{cases} 1 & \text{if } j \in SFS^c. \\ 0 & \text{otherwise.} \end{cases}$$

The following theorem presents the probability of sampling an unknown individual i using hybrid sampling.

Theorem 1. *Let, $Z, n_u, \delta, w, A, e,$ and m be given. Then the probability of an $N \to K$ transition for individual i in a time unit t based on the hybrid sampling is*

$$P(Y_{i,t} = K | Y_{i,t-1} = N) = [Z\frac{1}{n_u} + (1 - Z)(\frac{1 - e^{-(\delta w^T A e_i)}}{\sum_{j=1:j\in SFS^c}^{n} 1 - e^{-(\delta w^T A e_j)}})].$$

Proof. Similar to the compartmental SI virus propagation model in epidemiology [38, 65], we devlop a network-based sampling method based on a Poisson distribution with a mean of δm. The probability of an unknown individual being sampled through network is equal to the probability that the individual becomes known given its contacts with sampled individuals. In other words, given the Poisson process

$$P(z) = \frac{e^{-\delta}\delta^z}{z!}$$

where, δ is the sampling rate and z is the number of sampling events within the time interval (t, t+1] such that, z $\sim$ Poisson(δ), the probability that an unknown individual i is being sampled given a population size m within the time interval (t, t+1] is

$$\begin{aligned} P(Y_{i,t} = K|Y_{i,t-1} = N) &= 1 - P(i \text{ is not sampled in } (t, t+1]|i \text{ has not been sampled in } (-\infty, t-1]) \\ &= 1 - P(0) \\ &= 1 - e^{-\delta m} \end{aligned}$$

However, for the network-based sampling process and according to the network locality assumption,

$$P(Y_{i,t} = K|Y_{i,t-1} = N) = 1 - e^{-\delta m_i}.$$

According to Definition 2,

$$P(Y_{i,t} = K|Y_{i,t-1} = N) = 1 - e^{-\delta w^T A e_i}. \tag{3.1}$$

Note that $Y_{i,t} = N$ and $Y_{i,t} = K$ are complementary to each other. Thus,

$$P(Y_{i,t} = N|Y_{i,t-1} = N) = e^{-\delta w^T A e_i}.$$

On the other hand, using the uniform random sampling method and based on Definition 3, the probability that an unknown individual i moves to the known state is

$$P(Y_{i,t} = K|Y_{i,t-1} = N) = \frac{1}{n_u}. \tag{3.2}$$

Therefore, for the hybrid sampling and based on Equation 3.1, Equation 3.2, and Definition 2, the probability that an individual i will be sampled is

$$P(Y_{i,t} = K|Y_{i,t-1} = N) = [\frac{Z}{n_u} + (1 - Z)(\frac{1 - e^{-(\delta m_i)}}{\sum_{j=1:j\in SFS^c}^{n} 1 - e^{-(\delta m_j)}})]. \tag{3.3}$$

□

Table 5 displays the transition matrix for the hybrid sampling. Note that, based on our assumptions the probability of a $K \to N$ transition for individual i in a time unit t is

$$P(Y_{i,t} = N | Y_{i,t-1} = K) = 0. \tag{3.4}$$

Since $Y_{i,t} = N$ and $Y_{i,t} = K$ are complementary to each other. Hence,

$$P(Y_{i,t} = N | Y_{i,t-1} = N) = 1 - ([\frac{Z}{n_u} + (1 - Z)(\frac{1 - e^{-(\delta m_i)}}{\sum_{j=1:j \in SFS^c}^{n} 1 - e^{-(\delta m_j)}})]). \tag{3.5}$$

Note also that

$$P(Y_{i,t} = K | Y_{i,t-1} = K) = 1. \tag{3.6}$$

Table 5: Transition matrix of individual i for the hybrid sampling process.

	N	K
N	$1 - ([\frac{Z}{n_u} + (1 - Z)(\frac{1-e^{-(\delta m_i)}}{\sum_{j=1:j \in SFS^c}^{n} 1-e^{-(\delta m_j)}})])$	$[\frac{Z}{n_u} + (1 - (1 - Z)(\frac{1-e^{-(\delta me_i)}}{\sum_{j=1:j \in SFS^c}^{n} 1-e^{-(\delta m_j)}})]$
K	0	1

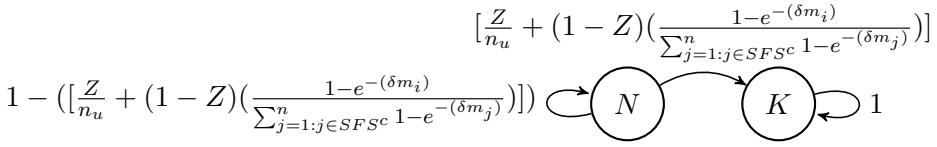


Figure 17: State diagram of individual i for the hybrid sampling process.

3.4 The Implementation of the Sampling Algorithms

We assume that the number of individuals n is known, and does not change throughout the sampling process and we also assume that the adjacency matrix is binary, symmetric, and static. We define an $n \times t$ non square matrix B whose columns are the adjacency lists of the sampled nodes. As more nodes are sampled, the adjacency list of the sampled nodes are added gradually as new columns to matrix B. Since the adjacency matrix A is assumed to be symmetric, hence B^T represents the

rows of the sampled nodes. The largest eigenvalue of the $n \times t$ sampled submatrix B is estimated as the square root of the largest eigenvalue of the B^tB square matrix [78].

We develop a network sampling method to move individuals from the unknwon status to the known status. Our network sampling is based on sampling without replacement where only unsampled nodes are considered for sampling in every iteration. A novel network sampling method is presented in [1], where a node is sampled proportional to the number of its sampled neighbors. We call this method the partial neighbor voting sampling (NVS) method and we implement it to find the largest eigenvalue λ of a partially known network. The scheme is simulated where individuals are sampled one by one to build a sub matrix B of the adjacency matrix A such that B includes the adjacency information of the so-far sampled individuals. In particular, as any individual is sampled, then the column of the adjacency matrix A corresponding to this individual is added to the sub matrix B. Note that at any stage t, matrix B has a size of $n \times t$ where $n \geq t$. Since the sub matrix B is not a square matrix unless all individuals are sampled, hence, a new square matrix C is built by taking the matrix cross product B^tB. Note that we could take the cross product BB^t, however, according to [78, page 555] the n eigenvalues of BB^t are the t eigenvalues of B^tB with the remaining $n-t$ eigenvalues are equal to 0. Therefore, we consider the cross product B^tB because it has a smaller size compared to BB^t especially in the early days. Note that the size of the matrix C is evolving as we sample more individuals. Lets denote the number of sampled individuals, at any stage t as $n_s(t)$, then the size of the matrix C is $n_s(t) \times n_s(t)$. In every iteration, the probability of next unknown individual to be sampled is proportional to this individual's number of its so-far sampled neighbors where when we sample the first individual, the size of the B sub-matrix will be $n \times 1$ and the size of the C sub-matrix will be 1×1.

Any diagonal element of the sub matrix C, i.e. the element in location (i, i) represents the degree of the node sampled on day i, while the off-diagonal element at any location (i, j) where $i \neq j$, represents the number of common neighbors between the node sampled on day i and the node sampled on day j. Let λ_A denote the largest eigenvalue of the adjacency matrix A, and λ_C denotes the largest eigenvalue of the sub-matrix C. Note that as we sample more nodes, the highest degree of the so-far sampled sub-network is monotonically increasing and therefore the largest eigenvalue of the so-far sampled sub-network is monotonically increasing too. As the size of the sub-matrix

C increases, we observe that λ_C approaches λ_A^2. This scheme provides an approximation of the largest eigenvalue in a partially known adjacency matrix.

Let

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

Lets assume that the number of the so-far sampled individuals $n_s = v$

$$B = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1v} \\ a_{21} & a_{22} & \cdots & a_{2v} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nv} \end{bmatrix}$$

Then, the sub-matrix C is

$$C = B^t \times B$$

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n_s} \\ c_{21} & c_{22} & \cdots & c_{2n_s} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n_s1} & c_{n_s2} & \cdots & c_{n_sn_s} \end{bmatrix}$$

where, $\forall i, j = 1, ..., n_s$

$$C[i,j] = \sum_{k=1}^{n} B^T[i,k] \times B[k,j]$$

Note that, the special case of the diagonal elements

$$C[i,i] = \sum_{k=1}^{n} B^T[i,k] \times B[k,i]$$

Since A is symmetric and B is a submatrix of A, then

$$B^T[i,k] = B[k,i]$$

Note that B is a binary matrix hence,

$$C[i,i] = \sum_{k=1}^{n} (B^T[i,k])^2 = \sum_{k=1}^{n} B^T[i,k]$$

Therefore, the diagonal element $C[i,i]$ represents the degree of the node sampled on day i and the off-diagonal element $C[i,j]$ represents the number of common neighbors between the node sampled on day i and the node sampled on day j.

Also, note that

$$C[i,j] \leq C[i,i], \forall i,j.$$

In other words, the degree of any node sampled on day i is larger than or equal to the number of common neighbors between node i and any other node. Let the set of the so-far sampled individuals be denoted as SFS. We assume that exactly one individual is sampled every time unit. Denote the size of the set SFS as $|SFS| = n_s$. On the first day, all individuals are equally likely to be sampled. The probability that an individual v is sampled at day $(t+1)$ depends on the adjacency structure of the so-far sampled individuals. In other words, the probability that a certain unknown individual is sampled on the next day is proportional to the number of known neighbors of that individual. Let X_v be a random variable that denote whether individual v is sampled or not; where

$$X_v = \begin{cases} 1 & \text{if individual } v \text{ is sampled} \\ 0 & \text{otherwise} \end{cases}$$

Recall that

$$A[u,v] = \begin{cases} 1 & \text{if there is a link between node } u \text{ and node } v \\ 0 & \text{otherwise} \end{cases}$$

Let $m(i)$ denote the partial degree of node i based on subsampled network; i.e. $m(i)$ denotes the number of the so-far sampled neighbors of node i. More precisely, m is a vector such that

$$m[i] = \sum_{(u:u \in SFS)} A[i,u], \forall i = 1, \cdots, n$$

We implemented the NVS concept using two alternative algorithms: NVS-A and NVS-B. In the NVS-A method, the probability weight for selecting node i is simply a linear combination of the votes of the so far sampled neighbors of the node. The probability weight based on NVS-A is expressed as

$$P(Y_{i,t} = K|Y_{i,t-1} = N) = \frac{m[i]}{\sum_{k=1}^{n} m[k]},$$

while for the NVS-B method the probability weight for selecting node i is based on Poisson distribution, where the probability weight is expressed in Equation 3.1. We compared the performance of the two alternatives and found that that both have a comparable performance in terms of the largest eigenvalue of the subsampled network for several types of graphs. Therefore, we focus on one of these alternatives which is the NVS-B, i.e, the neighbor voting sampling component of the hybrid methods is based on the NVS-B alternative. For clarity, the following example will be based on the algorithm NVS-A where the probability that individual v is sampled on the next time unit can be expressed as

$$P(X_{v,t} = 1|X_{v,t-1} = 0) = \frac{m[v]}{\sum_{k=1}^{n} m[k]}$$

This process is repeated $\forall v \notin SFS$ where after individual v is being sampled, then node v is added to SFS. To ensure sampling without replacement, $\forall v \in SFS$ we set the associated probability weight to 0.

3.4.1 Example

For example, let A be the adjacency matrix.

$$\mathbf{A} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

We sample exactly one individual per day. Assume that in day 3 ($n_s = 3$), the set of individuals who have been sampled so-far are

$$SFS = \{3, 6, 2\}$$

So matrix B becomes

$$\mathbf{B} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{array} \begin{array}{c} \begin{array}{ccc} 3 & 6 & 2 \end{array} \\ \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

Since the adjacency matrix B of the sampled individuals is a non-square matrix, and since the corresponding matrix C is a real symmetric matrix with non-negative values and finite dimensions, hence we estimate the eigenvalues of matrix B by taking the square roots of the eigenvalues of matrix C which is

$$C = B^t \times B$$

Then, for the example above we get

$$\mathbf{C} = \begin{array}{c} \\ 3 \\ 6 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 3 & 6 & 2 \end{array} \\ \begin{bmatrix} 7 & 3 & 6 \\ 3 & 5 & 4 \\ 6 & 4 & 9 \end{bmatrix} \end{array}$$

Next, we plot the largest eigenvalue of matrix C.

Vector m is updated to estimate the probability that a specific unknown individual is sampled the next day. Notice that any individual who have been already sampled will never be sampled again, and therefore its corresponding value in m vector will remain 0 until the end of the simulation. For example, note that individuals 3,6, and 2 have been already sampled and

therefore their corresponding value in the m vector will remain 0. Therefore, any individual except for individuals 3,6, and 2 can be sampled on the next stage. Hence the m vector becomes,

$$m = [2, 0, 0, 3, 3, 0, 2, 2, 2, 1]$$

Consequently, $P(X_1 = 1) = P(X_7 = 1) = P(X_8 = 1) = P(X_9 = 1) = \frac{2}{15}$, $P(X_4 = 1) = P(X_5 = 1) = \frac{3}{15}$, and $P(X_{10} = 1) = \frac{1}{15}$. We sample one of the 7 individuals for the next stage, and these steps are repeated until we sample all the individuals.

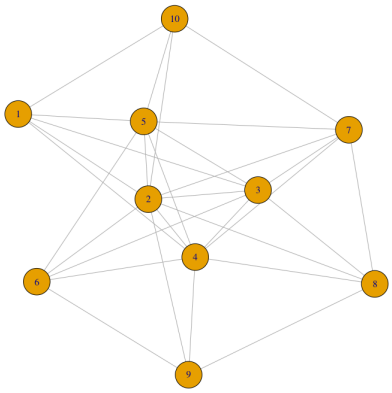


Figure 18: The Graph Structure of Example 1.

3.5 Types of Graphs

To study the impact of different network generating algorithms on the accuracy of the estimation process, we generate adjacency matrices based on two very well-known network generating models; The Barabási–Albert model (BA) and the Erdős–Rényi (ER) model. The Barabási–Albert algorithm is a simple stochastic algorithm for building a network. It is a discrete time step algorithm and in each time step a single node is added. Then the algorithm adds one node in each time step and the new node initiates some edges to old nodes. To generate a graph based on Barabási–Albert model that is composed of one or two clusters, we have set $edge = 5$, which means that 5 edges are added in each time step, and the parameter power is set to 1, which means that there is a linear preferential attachment. Also to generate Erdős–Rényi model that is composed of one or two

clusters, every two nodes have a specific probability of sharing an edge, and we set this probability to be equal to 0.50.

Figure 19 shows an example of a BA graph and an ER graph that is composed of one cluster each, whereas Figure 20 shows a BA graph and an ER graph that is composed of two clusters each. Note that there are two nodes in common between the two clusters for each algorithm.

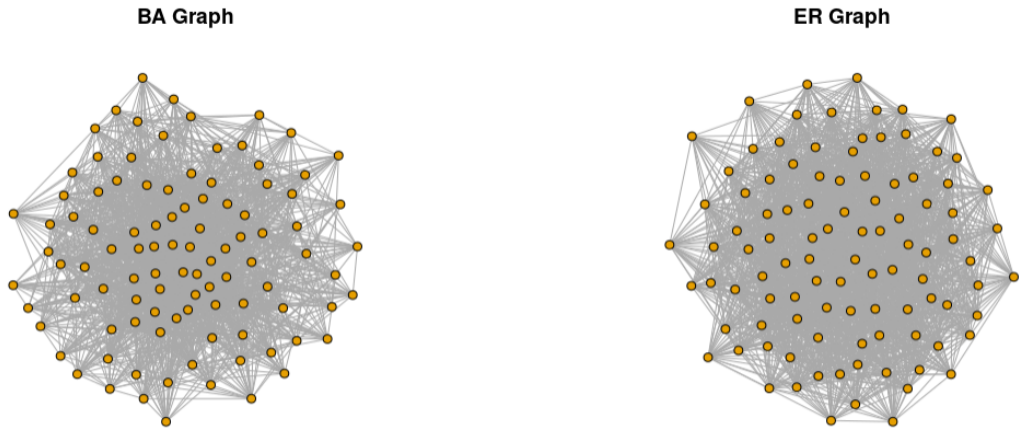


Figure 19: The BA graph and the ER graph composed of one cluster each.

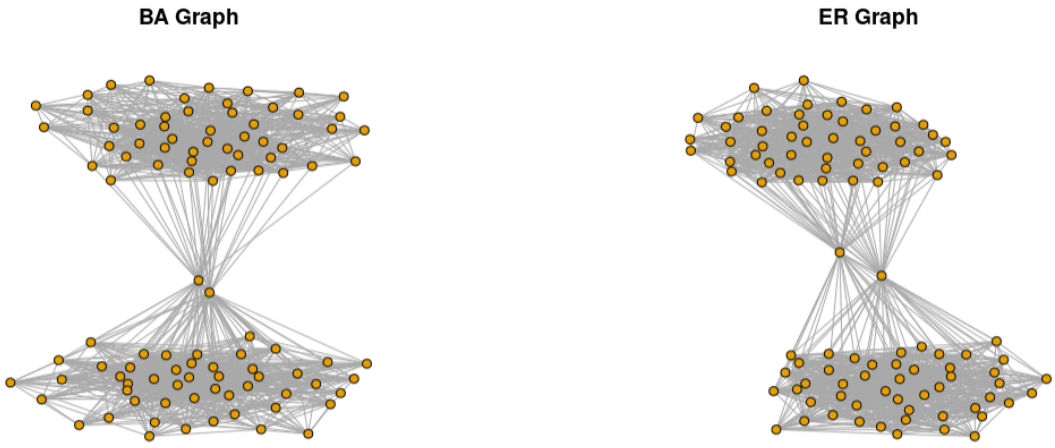


Figure 20: The BA graph and the ER graph composed of two clusters each.

Figure 21 and Figure 22, show the histograms of the node degree distribution for each graph, composed of one cluster and two clusters respectively. We can see for the ER graph composed of either one cluster or two clusters that the nodes degrees is approximately normally distributed, in other words, it can be considered a homogeneous network in terms of the degree distribution. While for the BA graph we can see that the histogram is skewed to the right, which indicates that the mean of the nodes degrees is greater than the median. In other words, many nodes have small degrees with a minimum degree of 20 for the one cluster graph and with a minimum degree of 15 for the two clusters graph, which indicates that the BA algorithm generates a heterogeneous network in terms of the degree distribution.

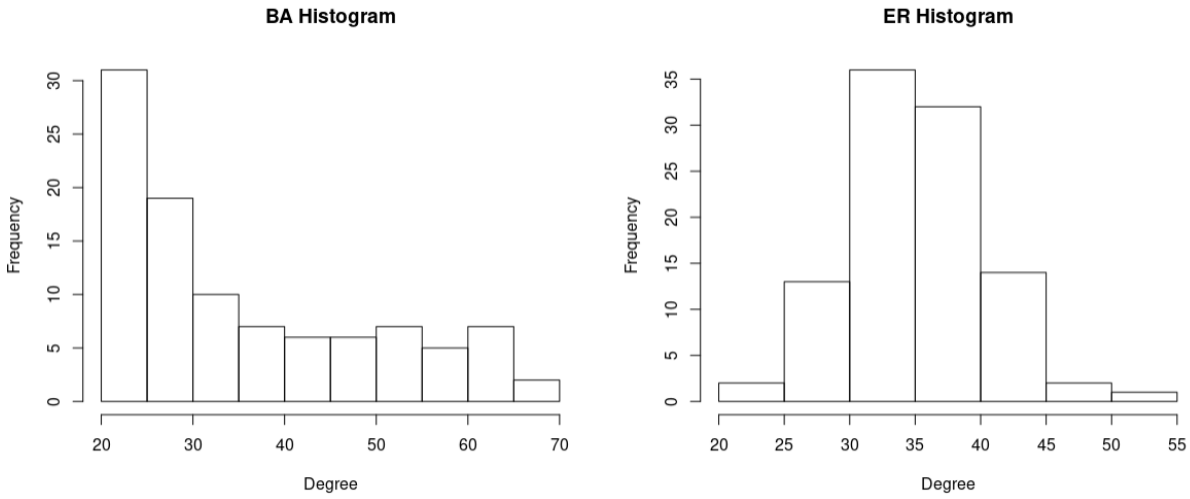


Figure 21: The BA histogram and the ER histogram composed of one cluster each.

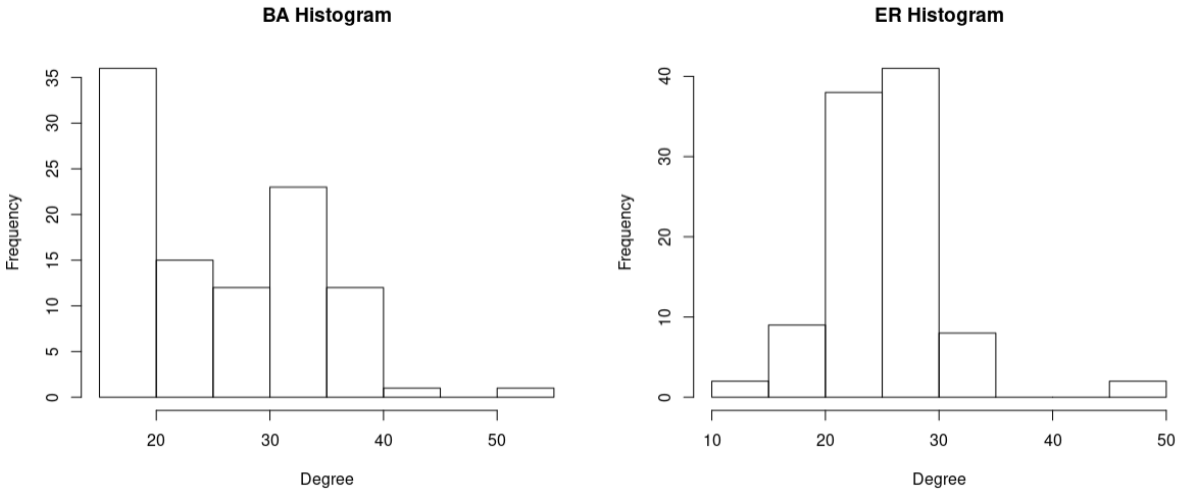


Figure 22: The BA histogram and the ER histogram composed of two clusters each.

3.6 Simulation

The performance of proposed method is evaluated through simulation using the R software package. The R code implementations of the algorithms are available at https://github.com/ralsehib/Hybrid-Sampling-Method.git. The R software package used is RStudio version 1.1.419. We run 100 independent repetitions to take variability into consideration where averages across the 100 repetitions are reported. Algorithm 2 provides the pseudocode for the SRS sampling method, while Algorithm 3 provides the pseudocode for the NVS sampling method.

Algorithm 2 Computing an Approximation of the Largest Eigenvalue of a Partially Known Adjacency Matrix based on a Sampled (Partial) Subgraph Using SRS Method

Generate A^{nxn} matrix
for $rep = 1, ..., repetitions$ **do**
    Denote the set of the so-far sampled individuals as SFS
    Initialize the probability weights vector: P_{SRS}
    for $i = 1, ..., n$ **do**
        Sample a random node v based on the probability weight vectors and add v to SFS
            (Calculate the largest eigenvalue of the sampled sub-graph)
        $e_{SRS}[i, rep] = eigen(crossprod(A[, SFS_{SRS}]))$
        The next node will be sampled without replacement as follows:
            9 SRS: Unsampled nodes are equally likely to be sampled
            (Update the probability weights vector)
        Set the probability of selecting node v to be 0 to ensure sampling without replacement
            $P_{SRS}[SFS_{SRS}] = 0$
        Update the probability weight vector for the unsampled nodes
            $P_{SRS} = \frac{P_{SRS}}{sum(P_{SRS})}$
    end for
end for

Algorithm 3 Computing an Approximation of the Largest Eigenvalue of a Partially Known Adjacency Matrix based on a Sampled (Partial) Subgraph Using the NVS Method

Generate A^{nxn} matrix
for $rep = 1, ..., repetitions$ **do**
    Denote the set of the so-far sampled individuals as SFS
    Initialize k and the probability weights vector: P_{NVS}
    for $i = 1, ..., n$ **do**
        Sample a random node v based on the probability weight vectors and add v to SFS
            (Calculate the largest eigenvalue of the sampled sub-graph)
        $e_{NVS}[i, rep] = eigen(crossprod(A[, SFS_{NVS}]))$
        $P_{NVS} = (!k) \times (1 - e^{(-\delta Ak)})$
            (Update the probability weights vector)
        Set the probability of selecting node v to be 0 to ensure sampling without replacement
            $P_{NVS}[SFS_{NVS}] = 0$
        Update the probability weight vector for the unsampled nodes
            $P_{NVS} = \frac{P_{NVS}}{sum(P_{NVS})}$
    end for
end for

Algorithm 4 provides the pseudocode that implements the HS method. We design the HS algorithm to be general such that 0 or more individuals can be sampled per time unit. We include a parameter PNK that denotes the probability of meeting new unknown individual(s) per unit time and an integer parameter η that denotes the number of sampled individuals per unit time.

Algorithm 4 Computing an Approximation of the Largest Eigenvalue of of a Partially Known Adjacency Matrix based on the Hybrid Sampling Process

Generate A^{nxn}
Pick a seed node at random, mark it as *known*
for $rep = 1, ..., repetitions$ **do**
    Denote the set of the so-far sampled individuals as SFS_{HS}
    Initialize k
        (Simulate the sampling process)
    $i = 0$
    while ($\exists$ unsampled nodes) **do**
        (Update the probability weights for network sampling)
        $pnet = (!k) \times (1 - e^{(-\delta Ak)})$
        $pnet = \frac{pnet}{sum(pnet)}$
        (Update the probability weights for random sampling)
        $prand =!k$
        $prand = \frac{prand}{sum(prand)}$
        (Update the probability weights for hybrid sampling)
        $P = Z \times prand + (1 - Z) \times pnet$
        Let $flagNewK$ denote whether we identify new unknown individual(s) per unit time
        Let PNK denote probability of meeting new unknown individual(s) per unit time
        Let η denote the number of sampled individuals per unit time
        $flagNewK = sample(x = 0 : 1, prob = c(1 - PNK, PNK))$
        if $((\sum_{k_i=1}^{n} k_i < (n - \eta)) \wedge$ flagNewK =1) **then**
            $indivSampl = sample(x = 1 : n, size = eta, prob = P)$
            add $indivSamp$ to SFS
            for $v = 1, ..., \eta$ **do**
                $tmp = indivSampl[v]$
                $k[tmp] = 1$
                $i = i + eta$
            end for
        end if
        $e_{HS}[i, rep] = eigen(crossprod(A[, SFS_{HS}]))$
    end while
end for

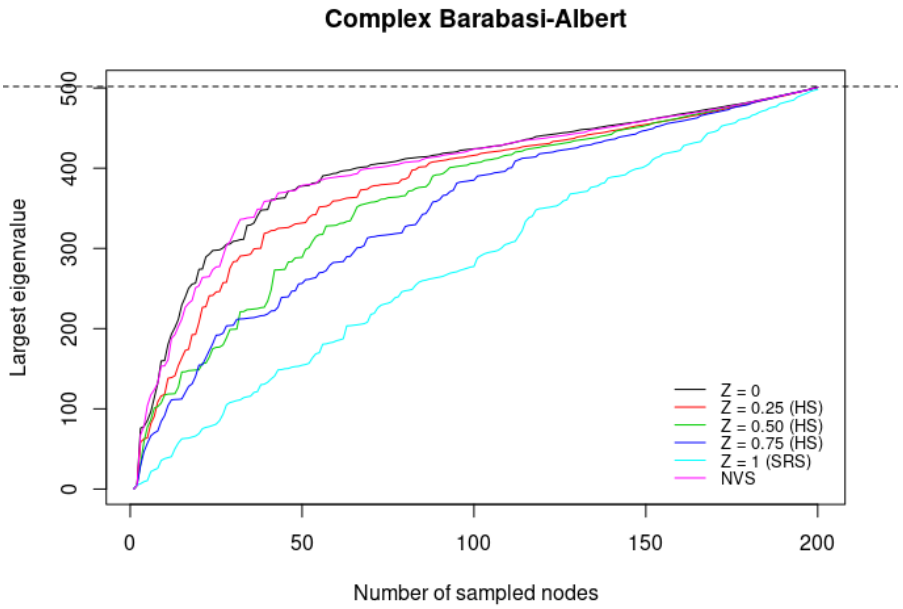


Figure 23: Comparison of the performance of sampling methods for strongly connected complex Barabási–Albert graph.

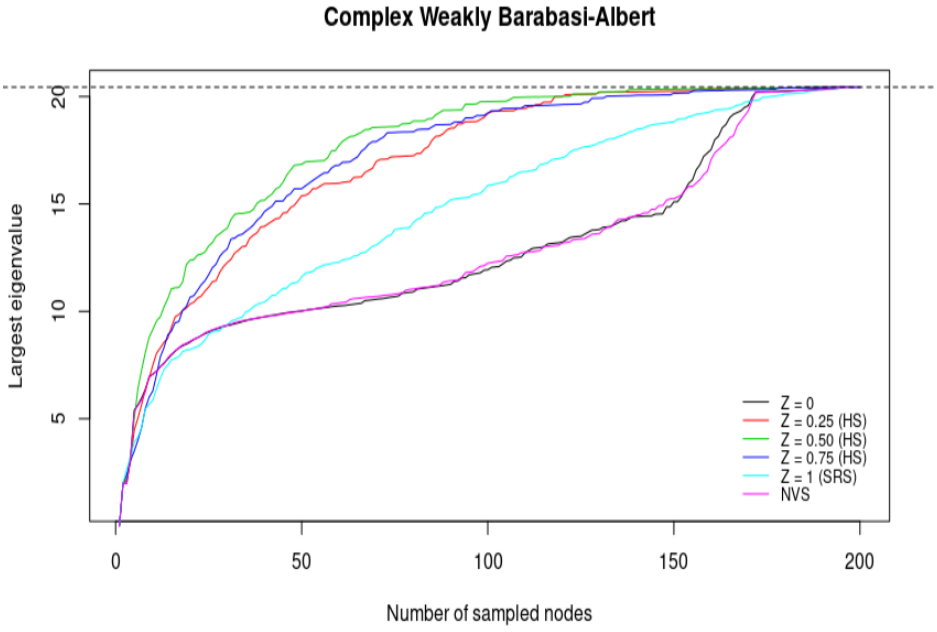


Figure 24: Comparison of the performance of sampling methods for weakly connected complex Barabási–Albert graph.

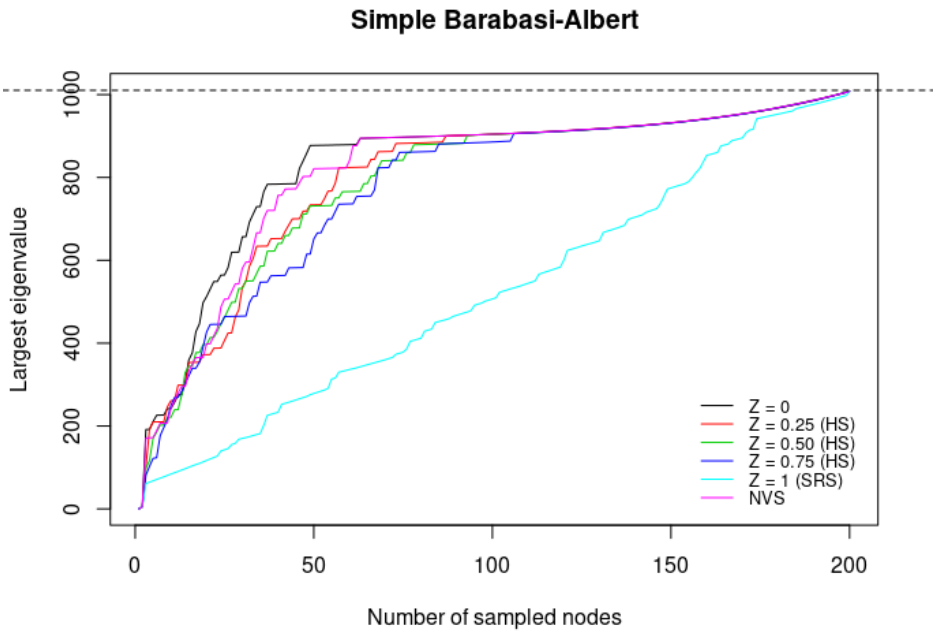


Figure 25: Comparison of the performance of sampling methods for simple Barabási–Albert graph.

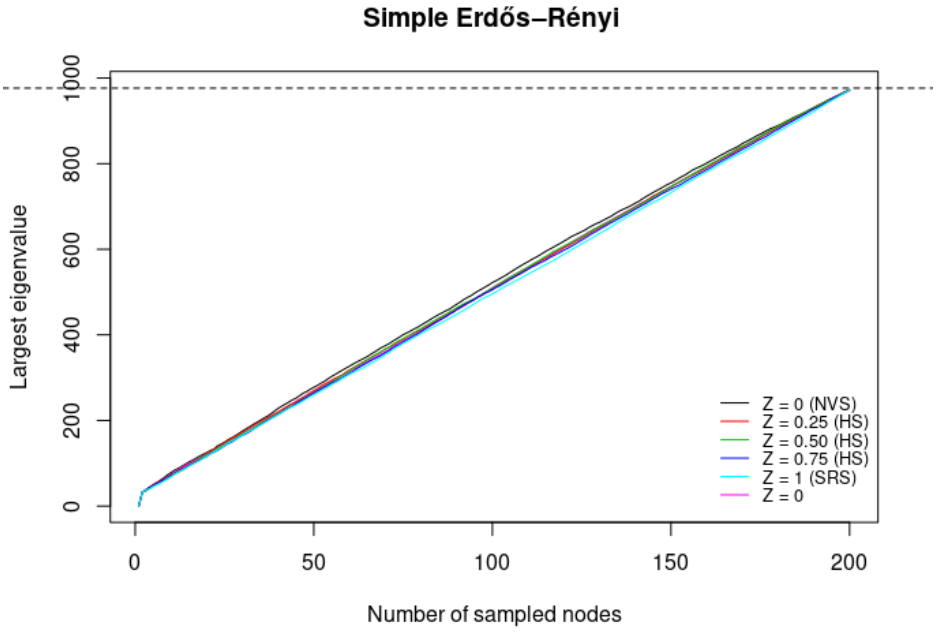


Figure 26: Comparison of the performance of sampling methods for simple Erdős–Rényi graph.

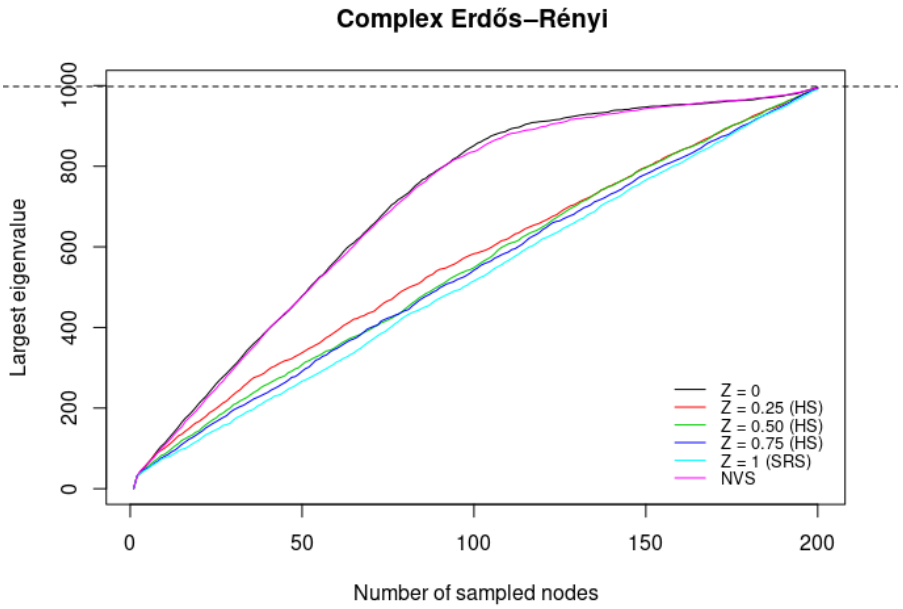


Figure 27: Comparison of the performance of sampling methods for strongly connected complex Erdős–Rényi graph.

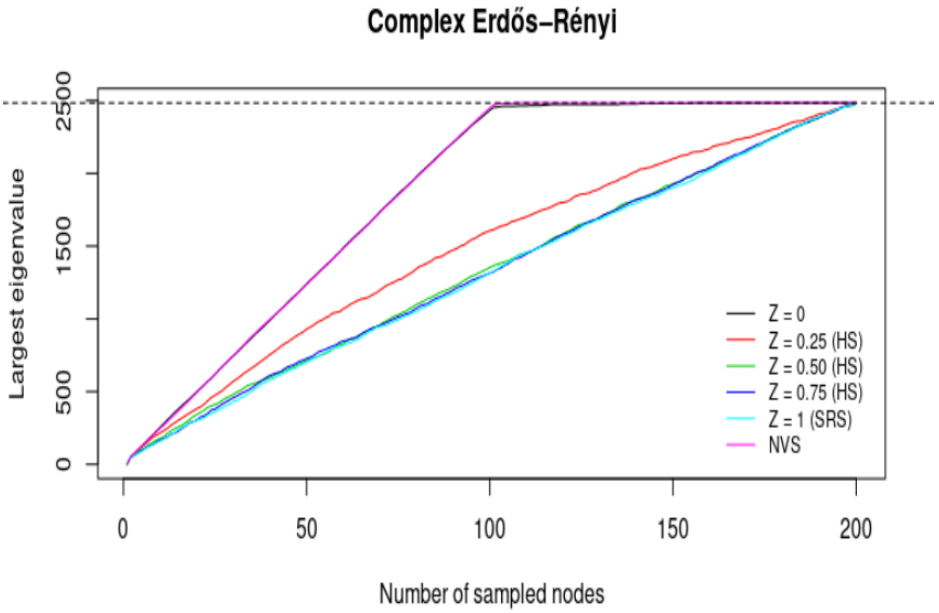


Figure 28: Comparison of the performance of sampling methods for weakly connected complex Erdős–Rényi graph.

3.7 Results

The R software package is used to develop codes to implement the proposed methods. The performance of the proposed methods is analyzed using simulation over 100 iterations with a generated network of 200 nodes using the Barabesi-Albert (BA) and the Erodes-Renie (ER) network models. The underlying networks are simple (one cluster) and complex (two clusters) networks. The performance of the proposed methods is measured in terms of the largest eigenvalue of the sampled partially-known network. The performance of the hybrid sampling (HS) sampling method based on the NVS-B algorithm is compared to that of the random sampling method as well as with that of the NVS-A method denoted as NVS in the figure. The value Z of the component of the simple random sampling in the hybrid sampling method is assumed to take 5 different levels: 0, 25%, 50%, 75%, and 100%. A value of $Z = 0$ means that the hybrid sampling is purely based on network sampling while a value of $Z = 1$ means that the hybrid sampling is based entirely on simple random sampling. For the other levels, a value of $Z = x\%$ means that an $x\%$ of the sampling probability weight for every individual in the hybrid sampling is based partially on simple random sampling.

The results show that the performance of the proposed method is a function in several factors including the underlying network structure and the partial weight of the network sampling component Z. Figure 23 shows that the performance of all the levels of the HS is significantly better than that of the SRS method for a graph composed of two strongly connected BA clusters, especially for low levels of Z, i.e., when the percentage of the simple random component is low. For example, by sampling only about 12.5% of the individuals, the improvement in terms of the Perron eigenvalue of the sampled subnetwork for the pure NVS method, is about 233% of that of the SRS method. As the percentage of the NVS component in the probability weight of the HS sampling increases, the level of improvement in terms of the Perron eigenvalue of the sampled subnetwork increases too. From the Figure, we can observe that by sampling only about 25% of the individuals, the Peron eigenvalue of the sampled subnetwork for the pure NVS method is about 80% of the value of the true Perron eigenvalue of the full network, while the SRS method achieves only 30% of the true Perron eigenvalue. This shows that by sampling a small fraction of the population, the HS methods achieves good approximation of the full matrix by identifying

well-connected nodes who can be for example vaccinated to prevent outbreaks. Figure 24 shows that the performance of all the levels of the HS method is also significantly better than that of the SRS method for a graph generated based on two unbalanced weakly connected BA clusters. For this case, we note that the performance of the HS method, when the sampling is partially SRS and partially NVS, is better than that of the pure NVS-A and NVS-B marked as (Z=0) and (NVS) respectively on the Figure. This is because in such graphs, the variance in node degrees is high, with few nodes having very high degrees, and since the NVS is highly likely to span the nodes of one cluster before moving to the other, hence the NVS method in this case might get stuck in one cluster and misses sampling well-connected nodes in the other cluster especially in the early days. This shows that the HS method has the capability to overcome the limitations of the NVS method snice the random component of the HS method can enable avoiding getting trapped in one cluster. Similar to the case of two connected BA clusters, Figure 25 shows that also for the case of graphs of one BA cluster, the performance of the HS method is considerably better than that of the SRS method. When the underlying network is based on the ER method, then the performance of the HS method is similar to that of the SRS method especially for the graphs that consist of single cluster as can be seen from Figure 26. This result is expected since the ER graphs are typically similar to random graphs with low variance in the node degrees and therefore the NVS method has no advantage over the SRS method. On the other hand, for two cluster ER graphs, as can be seen from Figure 27 (strongly connected) and Figure 28 (weakly connected), the pure NVS method performs better than other alternatives, since the pure NVS method is highly likely to span nodes that from one cluster before moving to the other. Other alternatives, which are partially or completely random-based are highly likely to span nodes that are alternatively in different clusters. For the Erdős–Rényi graph, we observe that according to the proposed method, the largest eigenvalue of the evolving sub-matrix grows linearly with the number of sampled individuals. Note that as more individuals are sampled, i.e, as the size of the sampled submatrix increases, the Perron eigenvalue of the submatrix approaches that of the full matrix.

3.8 Future Work

In the current chapter we presented a hybrid sampling method to sample exactly one node at every time step. A future research direction is to explore the impact of sampling more than one node at every time step as well as the impact of the degree of the sampled neighbors on the performance of the neighbor voting and the hybrid sampling algorithms. Networks also are typically evolving over time and therefore there is a need to for future research to explore how the presented sampling methods can be extended in order to take the dynamic aspects of networks into consideration. In addition, this study assumes that the population size is known and constant throughout the simulation process. However, in certain situations, the population size might be unknown and therefore there is a need for future research to take this factor into consideration. The population size on the other hand might vary during the sampling process and consequently a future research direction is to analyze the impact of such a variation on the sampling outcomes. The proposed sampling method assumes that only local information about immediate neighbors is available at every time step and therefore there is a need to explore the impact of neighbors who are connected through distances of more than one link on the sampling process. Finally, multiple links might exist between any two nodes and hence future research needs to relax the assumption of binary links.

CHAPTER 4

Estimating the Infection and Recovery Rates for SIS Epidemic Models Using Hybrid Sampling in Partially Known Networks

4.1 Introduction

A significant application of statistical modeling in the analysis of the spread of infectious diseases when the connections in the population are characterized using a network structure, is exploring how different network and epidemic parameters can be estimated when the network is only partially known. In this chapter, we investigate how the network sampling and the infection process are jointly modelled assuming a virus propagation model that allows for reinfection, in order to estimate important epidemic parameters where the sampling process is used to move individuals from a high risk to a low risk class. The health status of individuals before they are sampled is determined using back tracing methods so the epidemic parameters can be estimated.

Early research efforts for the modeling of infectious diseases considered population-level models in which, the size of the population, the infection rate, and the recovery rate were the main factors in modeling the spread of a disease [39]. Later, researchers realized that these models were insufficient for analyzing the spread of infectious diseases. Therefore, individual-level models (ILMs) emerged as a more realistic alternative since they take into consideration the heterogeneity in infectivity of individuals. In these models, individuals can be people, animals, plants, farms, regions or combinations of these categories where for instance, several researchers indicate that when farms are considered to be the individuals, then distance between two farms might affect the ability of the disease to spread. In addition, milk-tanker movement patterns between farms might be different from a farm to a farm, which, also can influence the risk of disease spread. For example, [64, 37] model the foot and mouth disease as an ILM where they consider the farm to be

the individual and other information like types of animals, number of animals, and spatial location as covariates.

4.1.1 The Virus Propagation Model

In 1927, Kermack and McKendrick introduced a mathematical model to describe the dynamics of the spread of an infectious disease. The model is known as the (SIR) model which stands for susceptible, infectious, and removed. In the SIR model, Kermack and McKendrick assumed that the rate of interaction between any two individuals in a population is the same, in other words they assumed that the population is homogeneous. Even though this model was very useful to describe the dynamics of an infectious diseases, however it is not realistic because in real life different individuals might have different interaction patterens. Therefore, researchers continued developing more realistic models that introduce heterogeneity in the modeling process.

The simplest mathematical model that describes the dynamics of the spread of an infectious disease is called the SI model, with only two states; susceptible and infected. The SI model is effective for an analysis of the dynamics of an infectious disease at the population level [84]. In the SI model, once an individual is infected then that individual remains infected forever, which is not true for most diseases. In real life, the immune system of an infected individual is going to fight the agent that caused the disease, and consequently after a certain period of time the infected individual is going to either recover or die. Dying or recovering can be considered as one state, since in either case, the infected individual is going to be removed from the infected state, and therefore a more realistic model; the susceptible-infectious-removed (SIR) framework was developed [84]. Later, researchers found that some recovered individuals become susceptible again and consequently the susceptible-infectious-susceptible (SIS) framework has emerged which is even more realistic than the SIR model for many common diseases. This study assumes that the virus propagation follows the SIS model. In this study we use the terms susceptible and healthy interchangeably and likewise, infected and sick are used interchangeably. In addition, we assume that infected individuals have the ability to infect others.

The links between individuals can be represented as a contact network, where the links

can affect the spread of an infectious disease. In real life, individuals have a set of links with friends, co-workers, ... etc, and in a large population, the chance of having a link between two individuals chosen at random is very small and may well be neglected. Therefore, it is not realistic to assume that any two individuals are equally likely to have a link between them. Several researchers have explored the use of network-based modeling to study the spread of infectious diseases. For example, Cauchemez et al. [22] studied the outbreak of the influenza (H1N1) in Hong Kong using social networks to model the spread of the disease. Groendyke et al. [52] also, modeled the outbreak of measles in Germany using a network-based modelling. In addition, Riley and Ferguson [97] used network-based models of infectious diseases to analyze the outbreak of smallpox in Great Britain. Also, Malik et al. [72] used ILM-based approach to study the Hong Kong pandemic of influenza (H1N1) during the 2008 to 2010 period.

4.2 Statistical Models

In this thesis we study the spread of infectious diseases using a network-based ILM. The simplest mathematical model that describes the dynamics of the spread of an infectious disease is the susceptible-infected (SI) model, where at any given time, each individual can be in any of the two states: susceptible (S) or infected (I). In the SI framework, as time $t \to \infty$, all the individuals who have a link to an infected individual are going to be infected. The number of infected individuals in the SI model is non-decreasing, since according to this model the I status is irreversible, where the infected individuals cannot recover and they remain infected forever, which is not true for most diseases. Therefore, it is realistic to model the spread of infectious diseases based on the susceptible-infectious-susceptible (SIS) framework. The SIS is an extension of the SI model since it accounts for the case when an individual could recover and get infected more than once (i.e. the SIS model allows for reinfection). Therefore, for the infection process we develop a network-based ILM which is modeled in terms of the SIS compartmental framework, such that at any given time, each individual can be in any of the two states: susceptible (S) or infected (I). Susceptible state includes individuals who are healthy, but they can catch the disease if they have contacts with an infected individual. Infected state includes individuals who are sick and if they have a contact with a susceptible individual, they probably transmit the disease to that individual. During the

epidemic, individuals move from the S state to the I state; $S \rightarrow I$, based on the transmission rate β, and once the individual recovers, based on the recovery rate γ, it moves to the S state again. We also assume that the recovery rate γ is independent of the individual's degree. The infection process is formulated as a Markov Chain process. Table 6 provides a list of our model parameters.

The focus of this thesis is on partially known networks where we use sampling to identify new individuals. In terms of the sampling process, we assume that any individual who becomes known remains known as time $t \rightarrow \infty$ which is similar to the the SI concept. The sampling process is formulated as a Markov Chain process where rather than the S and I compartments, we define two new compartments: unknown (N) and known (K). For the sampling process we developed a network-based ILM which is modeled in terms of the NK compartmental framework, such that at any given time, each individual can be in any of the two states: N or K. The unknown state represents individuals who have not been sampled yet, but they can be sampled if they have contacts with a sampled individual. The known state represents individuals who have been sampled. Note that according to the characteristics of the NK compartmental framework which are stated above, individuals move permanently from the N state to the K state; $N \rightarrow K$, during the sampling process.

The transmission rate β can depend on the behavior of a population, where for example in some countries when an individual is infected by a flu, then he or she wears a facial mask which can help protecting other individuals from being infected. In a network based modeling the infection process as well as the sampling process can be affected by the status of neighbors, since neighbors often share local information and might influence each other. Therefore in our model, we assume network locality, where the probability of a healthy individual i becoming infected is proportional to the status of its neighbors [75]. Likewise, based on the network locality assumption, the status of neighbors affects the probability of unknown individual i becoming known. In chapter 3, we developed a method to move individuals from the unknwon status to the known status using network sampling. Our proposed network sampling method is based on sampling without replacement where only unsampled nodes are considered for sampling in every iteration. A novel network sampling method is presented in [1], where a node is sampled proportional to number of its sampled neighbors. We call this method the partial neighbor voting sampling (NVS) method.

Table 6: The joint model parameters.

Parameter	Description
n	Number of individuals
t	The epidemic period (in days)
v	Vector of infected individuals
w	Vector of sampled individuals
e_i	A one-hot vector; a binary vector with zero everywhere except for the candidate individual in the i^{th} position
rep	Number of repetitions
edge	Number of edges to be added in each time step in Barabási–Albert model
power	The preferential attachment factor (1: linear) in Barabási–Albert model
prob	The probability that two nodes being connected in Erdős–Rényi model
common	Number of nodes that connect the two clusters
blobN	The cluster size
S	Susceptible state
I	Infectious state
N	Unknown class
K	Known class
Zeta (Z)	Probability of sampling by random
δ	Probability of sampling through a network
β	Infection transmission rate
γ	Recovery rate
A	Known symmetric contact adjacency matrix
$\hat{p}$	The prevalence of the disease
$\eta_{i,t}$	The number of infected neighbors of individual i at time t
$X_{i,t}$	The infection status of individual i at time t
$Y_{i,t}$	The sampled status of individual i at time t
n_I	The cumulative number of infected individuals
n_{tot}	The total number of both infected and susceptible individuals
f	The total number of infective individuals
n_u	The total number of unsampled individuals at iteration u

Since the partial neighbor voting sampling (NVS) algorithm samples individuals according to the votes of their neighbors, hence, one limitation of this algorithm is that it might get stuck in one cluster in case we have multi-cluster network especially if the network is disconnected or weakly connected. In chapter 3, and in order to avoid this limitation, we extended the NVS algorithm of [1] by developing a hybrid sampling (HS) method based on combined random sampling and network sampling. We implement this method by introducing a weight factor Z to represent the weight of the random sampling component. Note that $0 \leq Z \leq 1$ by definition, where $Z = 0$ denotes pure network sampling and $Z = 1$ denotes pure random sampling.

Definition 1. The infectious contact degree f_i for an individual i, is the total number of contacts between individual i and all other *infected* individuals. It is calculated as

$$f_i = v^t A e_i.$$

Lemma 1. *Let, β, v, A, e, and f be given. Then the probability of an $S \to S$ transition for individual i in a time unit t, $\forall i \in \{1, ..., n\}, t \in \{1, ..., l\}$, is*

$$P(X_{i,t} = S | X_{i,t-1} = S) = e^{-\beta v^T A e_i} = e^{-\beta f_i}.$$

Proof. Given a population with f infected individuals and an infection rate of β, we model the infection probability according to a Poisson distribution with a mean of βf [38, 65]. The probability of an individual avoiding the infection is equal to the probability that the individual remains healthy after contact with infected individuals, given that his or her status is healthy. Therefore, on the population level

$$P(X_t = S | X_{t-1} = S) = e^{-\beta f}.$$

However, for the network-based infection process and according to the network locality assumption,

$$P(X_{i,t} = S | X_{i,t-1} = S) = e^{-\beta f_i}. \tag{4.1}$$

According to Definition 1,

$$P(X_{i,t} = S | X_{i,t-1} = S) = e^{-\beta v^T A e_i}.$$

□

Note that, since our infection model has only two states S and I, hence the probability of an individual catching the infection $P(X_{i,t} = I | X_{i,t-1} = S)$ is complementary to $P(X_{i,t} = S | X_{i,t-1} = S)$. Therefore,

$$P(X_{i,t} = I | X_{i,t-1} = S) = 1 - e^{-\beta v^T A e_i}. \tag{4.2}$$

Lemma 2. *Let γ be given. Then the probability of an $I \to I$ transition for individual i in a time unit is*

$$P(X_{i,t} = I | X_{i,t-1} = I) = e^{-\gamma}.$$

Proof. It is assumed that the recovery probability follows a Poisson distribution with a mean of γ. Therefore, the probability of an infected individual remains infected (no recovery) in any iteration is expressed as

$$P(X_{i,t} = I | X_{i,t-1} = I) = e^{-\gamma}. \tag{4.3}$$

since the recovery of an individual is not affected by the health status of its neighbors. □

Table 7 displays the transition matrix for the SIS network-based models. Note that, since our infection model has only two states S and I. Hence the probability of an individual catching the infection $P(X_{i,t} = S | X_{i,t-1} = I)$ is complementary to $P(X_{i,t} = I | X_{i,t-1} = I)$. Therefore,

$$P(X_{i,t} = S | X_{i,t-1} = I) = 1 - e^{-\gamma}. \tag{4.4}$$

Table 7: Transition matrix of individual i for the Infection process.

	S	I
S	$e^{-\beta v^T A e_i}$	$1 - e^{-\beta v^T A e_i}$
I	$1 - e^{-\gamma}$	$e^{-\gamma}$

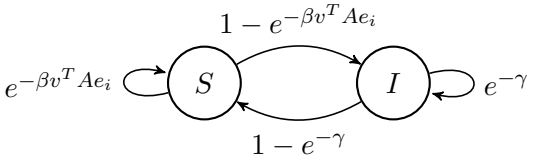


Figure 29: State diagram and a transition matrix of individual i for the SIS network-based models.

4.2.1 The Joint Sampling-Infection Process

In real life the infection process and the sampling process affect each other, and therefore a joint process is developed based on these two processes. Assuming conditional independence and assum-

ing that individuals will be either in state S or I and the class will be either known K or unknown N. Let us define the following events: A denotes that the next state is S, B denotes that the next class is N, C denotes that the current condition i.e, SN, SK, IN, or IK. Note that the event A and B are dependent, because the probability of a known individual moving from state S to state I is lower than the probability of an unknown individual having the same movement, since the known individuals are more likely to have precautions to avoid the infection. However, we assume that the two events A and B are conditionally independent given C (the current status of the individual) and the joint process is modeled as a Markov Chain process.

Theorem 2. *Let, $\beta, \delta, Z, A, n_u, v, w$, and e be given, then the probability of an $SN \rightarrow SK$ transition for individual i in a time unit t is*

$$\begin{aligned}
&P(X_{i,t} = S, Y_{i,t} = K | X_{i,t-1} = S, Y_{i,t-1} = N) \\
&= e^{(-\beta_N v^T A e_i)}[\frac{Z}{n_u} + (1 - Z)(\frac{1 - e^{-(\delta w^T A e_i)}}{\sum_{j=1:j\in SFS^c}^{n} 1 - e^{-(\delta w^T A e_j)}})].
\end{aligned}$$

Proof. Based on the conditional independence assumption, the probability that a health and unknown individual moves to the healthy and known class is

$$\begin{aligned}
&P(X_{i,t} = S, Y_{i,t} = K | X_{i,t-1} = S, Y_{i,t-1} = N) \\
&= P(X_{i,t} = S | X_{i,t-1} = S, Y_{i,t-1} = N) P(Y_{i,t} = K | X_{i,t-1} = S, Y_{i,t-1} = N).
\end{aligned}$$

Prior research indicates that there are significant differences in the behavior of different risk groups like known and uknown populations [51, 77, 56]. Therefore, we assume that the infection rate for a sampled individual (known individual) is β_k while the infection rate for an unsampled individual (unknown individual) is β_N and we assume that $\beta_K << \beta_N$. Therefore, based on Lemma 1

$$P(X_{i,t} = S | X_{i,t-1} = S, Y_{i,t-1} = N) = e^{-(\beta_N v^T A e_i)}$$

The sampling process in not affected by the infection process according to our assumptions. There-

fore, based on Theorem 1

$$
\begin{aligned}
&P(Y_{i,t} = K | X_{i,t-1} = S, Y_{i,t-1} = N) \\
&= [\frac{Z}{n_u} + (1 - Z)(\frac{1 - e^{-(\delta w^T A e_i)}}{\sum_{j=1:j\in SFS^c}^{n} 1 - e^{-(\delta w^T A e_)}})] \\
&= [\frac{Z}{n_u} + (1 - Z)(\frac{1 - e^{-(\delta m_i)}}{\sum_{j=1:j\in SFS^c}^{n} 1 - e^{-(\delta m_j)}})].
\end{aligned}
$$

Recall that m_i is the sampled contact degree for individual i as defined in chapter 3. Hence,

$$
\begin{aligned}
&P(X_{i,t} = S, Y_{i,t} = K | X_{i,t-1} = S, Y_{i,t-1} = N) \\
&= e^{(-\beta_N f_i)}[\frac{Z}{n_u} + (1 - Z)(\frac{1 - e^{-(\delta m_i)}}{\sum_{j=1:j\in SFS^c}^{n} 1 - e^{-(\delta m_j)}})].
\end{aligned}
$$

□

Theorem 3. *Let, $\gamma, \delta, Z, A, n_u, w,$ and e be given, then the probability of an $IN \to IK$ transition for individual i in a time unit t is*

$$
\begin{aligned}
&P(X_{i,t} = I, Y_{i,t} = K | X_{i,t-1} = I, Y_{i,t-1} = N) \\
&= e^{-(\gamma_K)}[\frac{Z}{n_u} + (1 - Z)(\frac{1 - e^{-(\delta w^T A e_i)}}{\sum_{j=1:j\in SFS^c}^{n} 1 - e^{-(\delta w^T A e_j)}})]
\end{aligned}
$$

Proof. Based on the conditional independence assumption, the probability that an infected and unknown individual moves to the infected and known class is

$$
\begin{aligned}
&P(X_{i,t} = I, Y_{i,t} = K | X_{i,t-1} = I, Y_{i,t-1} = N) \\
&= P(X_{i,t} = I | X_{i,t-1} = I, Y_{i,t-1} = N) P(Y_{i,t} = K | X_{i,t-1} = I, Y_{i,t-1} = N).
\end{aligned}
$$

Based on Lemma 2 and based on the assumption that the recovery rate for the unknown class γ_N

might be different from the recovery rate for the known class γ_k, therefore

$$P(X_{i,t} = I | X_{i,t-1} = I, Y_{i,t-1} = N) = e^{-(\gamma_N)}$$

The sampling process in not affected by the infection process. Therefore, based on Theorem 1

$$\begin{aligned} &P(Y_{i,t} = K | X_{i,t-1} = I, Y_{i,t-1} = N) \\ &= [\frac{Z}{n_u} + (1 - Z)(\frac{1 - e^{-(\delta w^T A e_i)}}{\sum_{j=1:j\in SFS^c}^{n} 1 - e^{-(\delta w^T A e_j)}})] \end{aligned}$$

Hence,

$$\begin{aligned} &P(X_{i,t} = I, Y_{i,t} = K | X_{i,t-1} = I, Y_{i,t-1} = N) \\ &= P(X_{i,t} = I | X_{i,t-1} = I, Y_{i,t-1} = N) P(Y_{i,t} = K | X_{i,t-1} = I, Y_{i,t-1} = N) \\ &= e^{-(\gamma_N)}[\frac{Z}{n_u} + (1 - Z)(\frac{1 - e^{-(\delta w^T A e_i)}}{\sum_{j=1:j\in SFS^c}^{n} 1 - e^{-(\delta w^T A e_j)}})] \\ &= e^{-(\gamma_N)}[\frac{Z}{n_u} + (1 - Z)(\frac{1 - e^{-(\delta m_i)}}{\sum_{j=1:j\in SFS^c}^{n} 1 - e^{-(\delta m_j)}})] \end{aligned}$$

□

Since the main assumption for the NK model is that the known individual cannot go back to the unknown state. Therefore, the probabilities $P(SK \to SN)$, $(SK \to IN)$, $P(IK \to SN)$, and $P(IK \to IN)$ are equal to 0. Note that the remaining transition probabilities are shown on Table 8.

Table 8: Transition matrix of individual i for the SIS and the NK processes.

	SN	SK	IN	IK
SN	$e^{(-\beta_N f_i)}(1 - [\frac{Z}{n_u} + (1 - Z)(\frac{1 - e^{-(\delta m_i)}}{\sum_{j=1:j\in SFS^c}^{n} 1 - e^{-(\delta m_j)}})])$	$e^{(-\beta_N f_i)}[\frac{Z}{n_u} + (1 - Z)(\frac{1 - e^{-(\delta m_i)}}{\sum_{j=1:j\in SFS^c}^{n} 1 - e^{-(\delta m_j)}})]$	$(1 - e^{(-\beta_N f_i)})(1 - [\frac{Z}{n_u} + (1 - Z)(\frac{1 - e^{-(\delta m_i)}}{\sum_{j=1:j\in SFS^c}^{n} 1 - e^{-(\delta m_j)}})])$	$(1 - e^{(-\beta_N f_i)})[\frac{Z}{n_u} + (1 - Z)(\frac{1 - e^{-(\delta m_i)}}{\sum_{j=1:j\in SFS^c}^{n} 1 - e^{-(\delta m_j)}})]$
SK	0	$e^{-(\beta_K f_i)}$	0	$(1 - e^{-(\beta_K f_i)})$
IN	$(1 - e^{-(\gamma_K)})(1 - [\frac{Z}{n_u} + (1 - Z)(\frac{1 - e^{-(\delta m_i)}}{\sum_{j=1:j\in SFS^c}^{n} 1 - e^{-(\delta m_j)}})])$	$(1 - e^{-(\gamma_K)})[\frac{Z}{n_u} + (1 - Z)(\frac{1 - e^{-(\delta m_i)}}{\sum_{j=1:j\in SFS^c}^{n} 1 - e^{-(\delta m_j)}})]$	$e^{-(\gamma_K)}(1 - [\frac{Z}{n_u} + (1 - Z)(\frac{1 - e^{-(\delta m_i)}}{\sum_{j=1:j\in SFS^c}^{n} 1 - e^{-(\delta m_j)}})])$	$e^{-(\gamma_N)}[\frac{Z}{n_u} + (1 - Z)(\frac{1 - e^{-(\delta m_i)}}{\sum_{j=1:j\in SFS^c}^{n} 1 - e^{-(\delta m_j)}})]$
IK	0	$1 - e^{-(\gamma_K)}$	0	$e^{-(\gamma_K)}$

4.3 Estimating the Infection and the Recovery Rates

One way to control the spread of an infectious disease is through a cycle of three stages: sampling, testing, and contact tracing, according to [62]. In case there is a need to test several individuals simultaneously then the multiplicity pool testing method described in chapter 2 can be used. Potential infected cases can be identified through random sampling or network sampling. Random sampling implies identifying individuals in a random manner repeatedly until the end of the sampling process. While for the network sampling method, the index case is asked to list his contact network through contact tracing and the contacts are immediately tested [62]. A main goal of contact tracing is to identify individuals who might have been infected as a result of having contact with an infected individual, and in case the identified individuals are infected then they are going to be either treated or isolated. Therefore, contact tracing is a powerful strategy for controlling an epidemic. On the other hand, taking the contact tracing into consideration when developing models of the spread of an epidemic increases the complexity of the modeling process [21]. When considering contact tracing it is important to distinguish between unreported cases and reported cases because only reported cases can lead to contact tracing. Contact tracing is efficient when the incubation period is long [68] which is the case for COVID-19 because it has a mean incubation period of 5.6 days [94]. Parameters in the case of partially-known individual-level networks can be estimated by adopting a process that consists of two main phases. In the first phase, individual level information is collected through sampling until the last day of the epidemic where the individual level information includes the daily infection status and the contact network of individuals. The second phase consists of estimating the probability of the infection of individuals in the days before they are sampled. These probabilities will be used to estimate the infection rate β and the recovery rate γ.

As indicated earlier, the proposed hybrid sampling is based on a blend of random sampling and network sampling. So, when for example an individual i is sampled in day t, then we trace the contacts of that individual. For illustration, let the adjacency matrix A contain information about the contacts of individuals. We define an $l \times n$ identification status matrix K which contains information about individuals who are sampled day by day. We also define an $l \times n$ infection status

matrix S that contains information about the infection status of individuals day by day. The S matrix contains the health status of individuals throughout the simulation where $S(i, j)$ represents the health status of individual j on day i. Specifically, $S(i, j) = 1$ if individual j is sick on day i and 0 otherwise. Likewise, the K matrix represents the identification status of individuals, where $K(i, j)$ represents whether individual j is known or not on day i such that, $K(i, j) = 1$ if individual j is known (have been identified on any previous day) on day i and 0 otherwise. At the end of the epidemic, all the individuals in matrix K must have been identified or otherwise they will be removed from the K matrix. Note that if any individual j is identified, then he or she will remain known throughout the remaining period while the health status of any individual can change from the sick to healthy status or vice versa according to a specific probability as explained in the next section. For simplicity, the scheme assumes that only one individual is sampled on any given day. The scheme starts by assuming that on day one, only one individual is known and its health status is sick, while all other individuals are assumed to be unknown, and their health status is healthy. On day two, the scheme passes over all individuals one by one to estimate, based on a specific probability, whether the health status of each individual will be changed or not and the second row of matrix S i.e. $S(2, 1)$ through $S(2, n)$ is updated accordingly. On day two, one new individual will be sampled according to a specific probability. For example, assume that the individual that has been sampled on day 2 is individual number 7, then $K(2, 7)$ will be set to 1. The process will be repeated for all the remaining days or until all individuals are sampled.

On the following days, the simulation passes over all individuals to update their health status day by day. The health status as well as the identification status of every individual are updated based on transition probabilities shown on Table 8, and then matrix S and matrix K will be updated accordingly. When an individual i is sampled on day t then the K matrix is updated by setting $K(i, t) = 1$ to denote that individual i becomes *known* on day t. The class of known individuals has an infection rate β_K and recovery rate γ_K while the unknown class has an infection rate of β_N and a recovery rate of γ_N. Also, every individual's probability of infection is estimated and the S matrix will be updated accordingly, such that if the individual i is confirmed to be infected on day t, then $S(i, t)$ will be set to 1, otherwise it will be set to 0. Furthermore, individual i will be asked to list its contact network; where during simulation the adjacency information of

this individual is retrieved from the already-generated adjacency matrix A. As explained in chapter 3, at any point based on the network sampling method, the probability of sampling any individual is a function in the number of its neighbors who have been sampled so far.

In the second stage of the simulation, we build a new matrix called Time-Individual Status matrix TIS with n rows and l columns, where n is the number of individuals and l is the number of time units. The TIS matrix represents the combination over time of the adjacency matrix A and the infection status matrix S based on the identification status matrix K. The rows of the TIS matrix represent the individuals and the columns represent the days. $TIS(i,t) = 1$ if individual i on day t is confirmed to be infected although the exact infection time might be unknown, and $TIS(i,t) = 0$ if individual i on day t is found to be susceptible. Note that the matrix TIS is not symmetric because an element $TIS(x,y)$ displays the status of individual x on day y while element $TIS(y,x)$ displays the status of individual y on day x, and the values of these two elements might not be equal when $x \neq y$. The K matrix is initialized as a zero matrix with only one uniformly-sampled individual marked as known. Note that any individual who is sampled in any specific day remains known throughout the process. Therefore, when individual j is sampled on day t, then all entries of column j starting with row t are set to 1. In other words, for individual j

$$K(i,j) = 1 \forall i \geq t.$$

To implement the parameter estimation phase efficiently, the K matrix is reordered, by placing the entries of the column corresponding to the first sampled individual on column 1, the entries of the column corresponding to the second sampled individual of column 2, and so on. This way, the lower left triangle contains the outcomes of the sampling process. As a result of this ordering, the lower left triangle of the K matrix will contain all $1s$, i.e.

$$K(i,j) = \begin{cases} 1 & \forall i \geq j \\ 0 & \text{otherwise} \end{cases}$$

According to this arrangement, $K(i,j) = 1$ indicates that individual j has been sampled on day i or earlier. The TIS matrix is updated according to the reordered K matrix where the upper right

triangle contains the status of individuals who were unknown in any specific day. In other words, assuming that n and l are equal, if $i > t$ then $TIS(t,i) = -1$, $\forall t = 1, ..., l, \forall i = 1, ..., n$.

Accordingly,

$$TIS(t,i) = \begin{cases} \text{-1} & \text{individual } i \text{ was unknown on day } t \\ 1 & \text{individual } i \text{ was known and sick on day } t \\ 0 & \text{individual } i \text{ was known and healthy on day } t \end{cases}$$

Our approach to estimate the value of infection rate β and the recovery rate γ starts by adding any newly discovered individual as well as its health status to the TIS matrix. Then all the contacts of this individual are added to the TIS. This process is repeated until the last (the current) day of the epidemic or until the last individual is identified. Then, we use back-tracing to identify the status of the individual as well as the status of its contacts in the previous days since the beginning of the epidemic. Unlike forward tracing which aims to identify individuals who might have been infected by an index case, back tracing has been used in to identify the source who infected a current case [41, 69, 95, 81, 80]. In this thesis, we adopt the concept of back tracing with the objective of identifying the health status of every individual in the early days before they are sampled, rather than identifying the source of infection for every individual. The reason behind our approach is that we want to identify if the individual was actually ever infected before the infection confirmation date if any. This is because the individual might have been infected through its contact with a particular asymptomatic neighbor and it might happen that this neighbor has been infected and recovered without knowing it. The R code implementations of the algorithms are available at https://github.com/ralsehib/Joint-Sampling-Infection-Processes.git. Algorithm 5 provides the psudocode code for building the TIS matrix.

Algorithm 5 Simulating the joint infection-sampling processes to build the Time-Individual Status matrix

Generate A^{nxn}
Pick a seed node at random, mark it as known and infected
Initialize K^{lxn} and S^{lxn}
for $rep = 1, ..., repetitions$ **do**
    (Simulate the infection and sampling processes)
    for $t = 2, ..., l$ **do**
        for $i = 1, ..., n$ **do**
            (Calculate the transition probabilities)
            if the status of i is SN **then**
                $P(S|S) = e^{-\beta_N v^T A e_i}$
                $S[t, i] = sample(0 : 1, prob = c(P(S|S), 1 - P(S|S)))$
                $P(N|N) = e^{-\delta w^T A e_i}$
                $K[t, i] = sample(0 : 1, prob = c(P(N|N), 1 - P(N|N)))$
            else if the status of i is IN **then**
                $P(I|I) = e^{-\gamma_N}$
                $S[t, i] = sample(0 : 1, prob = c(P(I|I), 1 - P(I|I)))$
                $P(N|N) = e^{-\delta w^T A e_i}$
                $K[t, i] = sample(0 : 1, prob = c(P(N|N), 1 - P(N|N)))$
            else if the status of i is SK **then**
                $P(S|S) = e^{-\beta_K v^T A e_i}$
                $S[t, i] = sample(0 : 1, prob = c(P(S|S), 1 - P(S|S)))$
            else if the status of i is IK **then**
                $P(I|I) = e^{-\gamma_K}$
                $S[t, i] = sample(x = 0 : 1, prob = c(P(I|I), 1 - P(I|I)))$
            end if
            Update S and K based on the transition probabilities
        end for
    end for
    $S = subset(S, select = order(colSums(K)))$
    $TIS \Longleftarrow S$
    for $t = 1, ..., (l - 1)$ **do**
        for $i = (t + 1), ..., n$ **do**
            $TIS[t, i] = -1$
        end for
    end for
end for

The goal of this study is to estimate the infection rate β and the recovery rate γ from partially- known information such that a likelihood function $L(\beta, \gamma)$ is maximized by conditioning on the number of infected neighbors η_i for an individual i, as described below. Based on our SIS network-based model, an individual moves from the susceptible state to the infectious state with probability $1 - e^{-\beta \eta_{t-1,i}}$, while an individual moves from the infectious state to the susceptible state

with probability $1 - e^{-\gamma}$. Note that γ is independent of i and t, and it can be estimated with only partially-known information.

Definition 5. Let, N_{IS} be the number of times that an individual whose status was infected on a specific day, becomes susceptible the next day, $\forall i, t$, and let N_{II} be the number of times that an individual whose status was infected on a specific day, remains infected the next day, $\forall i, t$.

Note that in case the status of an individual is infected in day t and remains infected in day $t+1$ then this case will be counted twice. Likewise, in case an individual is susceptible in day t and remains susceptible in day $t+1$ then this case will be counted twice.

Theorem 4. *The recovery rate $\hat{\gamma}$ that maximizes the likelihood of the TIS matrix is calculated as*

$$\hat{\gamma} = log(\frac{N_{I\cdot}}{N_{II}}). \tag{4.5}$$

Proof. The likelihood function is expressed as:

$$L(\beta, \gamma) = P[X_0 = x_0] \prod_{t=1}^{m} P[X_t = x_t | X_{t-1} = x_{t-1}]$$
$$= P[X_0 = x_0] \prod_{t=1}^{m} \left(\prod_{i=1}^{n} P[X_{t,i} = x_{t,i} | X_{t-1,i} = x_{t-1,i}] \right)$$

where,

$$P[X_{t,i} = x_{t,i} | X_{t-1,i} = x_{t-1,i}] = \begin{cases} e^{-\beta \eta_{t-1,i}} & X_{t-1,i} = S, X_{t,i} = S \\ 1 - e^{-\beta \eta_{t-1,i}} & X_{t-1,i} = S, X_{t,i} = I \\ 1 - e^{-\gamma} & X_{t-1,i} = I, X_{t,i} = S \\ e^{-\gamma} & X_{t-1,i} = I, X_{t,i} = I \end{cases}$$

according to Lemma 1 and Lemma 2 above.

Taking the log of $L(\beta, \gamma)$ we get

$$l(\beta, \gamma) = \sum_{t=1}^{m} \sum_{i=1}^{n} f(\beta, \gamma) \tag{4.6}$$

where,

$$f(\beta,\gamma) = \begin{cases} -\beta\eta_{t-1,i} & X_{t-1,i} = S, X_{t,i} = S \\ log(1 - e^{-\beta\eta_{t-1,i}}) & X_{t-1,i} = S, X_{t,i} = I \\ log(1 - e^{-\gamma}) & X_{t-1,i} = I, X_{t,i} = S \\ -\gamma & X_{t-1,i} = I, X_{t,i} = I \end{cases}$$

From Equation 4.6

$$l(\beta,\gamma) = \left(\sum_{\substack{X_{t-1,i}=S \\ X_{t,i}=S}} -\beta\eta_{t-1,i} + \sum_{\substack{X_{t-1,i}=S \\ X_{t,i}=I}} log(1 - e^{-\beta\eta_{t-1,i}}) + \sum_{\substack{X_{t-1,i}=I \\ X_{t,i}=S}} log(1 - e^{-\gamma}) + \sum_{\substack{X_{t-1,i}=I \\ X_{t,i}=I}} -\gamma \right) \tag{4.7}$$

Let

$$g(\beta) = \sum_{\substack{X_{t-1,i}=S \\ X_{t,i}=S}} -\beta\eta_{t-1,i} + \sum_{\substack{X_{t-1,i}=S \\ X_{t,i}=I}} log(1 - e^{-\beta\eta_{t-1,i}}) \tag{4.8}$$

and let

$$h(\gamma) = \sum_{\substack{X_{t-1,i}=I \\ X_{t,i}=S}} log(1 - e^{-\gamma}) + \sum_{\substack{X_{t-1,i}=I \\ X_{t,i}=I}} -\gamma \tag{4.9}$$

Then, Equation 4.7 can be written as

$$l(\beta,\gamma) = g(\beta) + h(\gamma)$$

Equation 4.8 can be simplified as,

$$g(\beta) = -\beta \sum_{\substack{X_{t-1,i}=S \\ X_{t,i}=S}} \eta_{t-1,i} + \sum_{\substack{X_{t-1,i}=S \\ X_{t,i}=I}} log(1 - e^{-\beta\eta_{t-1,i}})$$

Let, N_{IS} be the number of times that an individual whose status was infected on a specific day, becomes susceptible the next day, $\forall i, t$, and let N_{II} be the number of times that an individual whose status was infected on a specific day, remains infected the next day, $\forall i, t$.

Then , Equation 4.9 can be written as

$$h(\gamma) = log(1 - e^{-\gamma})N_{IS} - \gamma N_{II}$$

□

Then, $l(\beta, \gamma)$ is maximized when $g(\beta)$ is maximum and $h(\gamma)$ is maximum. Taking the derivative of the $h(\gamma)$

$$h'(\gamma) = N_{IS}\frac{1}{1-e^{-\gamma}}e^{-\gamma} - N_{II} = \frac{N_{IS}e^{-\gamma}}{1-e^{-\gamma}} - N_{II}$$

Factorization

$$= \frac{N_{IS}e^{-\gamma} - N_{II} + N_{II}e^{-\gamma}}{1-e^{-\gamma}}$$

Set the derivative to zero

$$(N_{IS} + N_{II})e^{-\gamma} - N_{II} = 0$$

$$e^{-\gamma} = \frac{N_{II}}{N_{IS} + N_{II}} = \frac{N_{II}}{N_{I\cdot}}$$

where $N_{I\cdot}$ is the number of daily active infections. Thus, $\hat{\gamma} = log(\frac{N_{I\cdot}}{N_{II}})$. The code for implementing Equation 4.5 is given in Algorithm 6.

Algorithm 6 Calculating the Recovery Rate $\hat{\gamma}$

```
Generate TIS matrix as in Algorithm 5
for rep =1, ...., repetitions do
    (Count the total number of Is in TIS)
    ni = length(which(TIS = 1))
    (Count the total number of Ss in TIS)
    ns = length(which(TIS = 0))
    (Count the total number of individuals with known status)
    total = ni + ns
    (Calculate the percent of infected individual)
    p̂ = ni/total
    counter = 0
    for t = 2, ..., l do
        for i = 1, ..., n do
            if (TIS[t − 1, i] = 1 ∧ TIS[t, i] = 1) then
                counter = counter +1
            end if
        end for
    end for
    γ̂ = log(ni/counter)
end for
```

Note that with only partially-known information, β cannot be estimated analytically. This is because the probability that an individual will move from the susceptible state to the infectious state is equal to

$$P(X_{i,t} = I | X_{i,t-1} = S) = 1 - e^{(-\beta \eta_{t-1,i})}$$

which depends on the infection rate and the number of infected neighbors $\eta_{t-1,i}$ of individual i on day $t-1$. However, note that an individual i might have a degree of 5 or more, which means that $\eta_{t-1,i} \geq 5$ for many individuals. According to the Abel–Ruffini theorem there is no general solution in radicals for polynomial equations of degree 5 or more [5]. Therefore, we need to estimate β numerically. To estimate the infection rate β in partially known networks, we developed two backward fill-up methods; namely the long back tracing method and the shortcut method. The long back tracing method is based on an expectation-maximization approach using Gibbs sampling and maximum likelihood estimation process. The shortcut method on the other hand, is based on maximum likelihood estimates where the fill-up phase simply replaces the unknown values with the estimated prevalence value.

4.3.1 The Long Back Tracing Method

The back-tracing method starts by moving backwards in terms of time estimating missing information. When we find an individual with unknown infection status at day, say $t-1$, we estimate its probability of being infected or susceptible at day $t-1$ conditioning on their status on day t. Suppose that an individual i was sick on day t i.e., $TIS(i,t) = 1$, the probability that this individual was in fact sick on day $t-1$, can be estimated as follows:

Based on the low of total probability

$$\begin{aligned} P[X_{i,t-1} = 1 | X_{i,t} = 1] &= \frac{P[X_{i,t} = 1 | X_{i,t-1} = 1] P[X_{i,t-1} = 1]}{P[X_{i,t} = 1 | X_{i,t-1} = 1] P[X_{i,t-1} = 1] + P[X_{i,t} = 1 | X_{i,t-1} = 0] P[X_{i,t-1} = 0]} \\ &\simeq \frac{(e^{-\hat{\gamma}})\hat{p}}{(e^{-\hat{\gamma}})\hat{p} + (1 - e^{(-\hat{\beta}\eta_{i,t-1})})(1 - \hat{p})} \end{aligned} \tag{4.10}$$

While if an individual i was susceptible on day t i.e., $TIS(i,t) = 0$, the probability that this individual was sick on day $t-1$ can be estimated as follows:

$$P[X_{i,t-1} = 1 | X_{i,t} = 0] = \frac{P[X_{i,t} = 0 | X_{i,t-1} = 1] P[X_{i,t-1} = 1]}{P[X_{i,t} = 0 | X_{i,t-1} = 1] P[X_{i,t-1} = 1] + P[X_{i,t} = 0 | X_{i,t-1} = 0] P[X_{i,t-1} = 0]}$$
$$\simeq \frac{(1 - e^{-\hat{\gamma}})\hat{p}}{(1 - e^{-\hat{\gamma}})\hat{p} + (e^{(-\hat{\beta}\eta_{i,t-1})})(1 - \hat{p})} \tag{4.11}$$

The prevalence of the disease ($\hat{p}$) is calculated as follows:

$$\hat{p} = \frac{n_I}{n_{tot}} \tag{4.12}$$

where n_I is the cumulative number of infected individuals identified so far, and n_{tot} is the total number of both infected and susceptible individuals.

This process is repeated for all the n individuals and the l time units to have an estimate of the health status of individuals even in the days before they have been sampled, i.e., the TIS matrix to be fully-known (all elements of TIS are either 0 or 1) so we can determine the maximum likelihood estimate of the infection rate β. We will illustrate our approach through an example that covers 11 days and 10 individuals. Lets assume that on average we will discover 1 individual per day. Moreover, we perform contact tracing for every individual we discover. Let the first case be discovered on day 1 and denote this case as individual No.1. Assume that the status of this individual is infected. Therefore, we set $TIS(1,1)$ to be equal to 1. For simplicity, assume that tracing the contacts of individual 1 shows that this individual has 5 contacts in its network. We continue discovering individuals and tracing their contact until day 11. Table 9 shows a sample adjacency matrix for this example.

Table 9: An example adjacency matrix.

	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	1	1	1	0	0	1
2	1	0	0	0	0	0	0	0	1	0
3	0	0	0	0	1	0	1	0	1	1
4	0	0	0	0	0	0	1	0	0	1
5	1	0	1	0	0	0	1	0	0	0
6	1	0	0	0	0	0	0	1	1	0
7	1	0	1	1	1	0	0	0	0	1
8	0	0	0	0	0	1	0	0	1	1
9	0	1	1	0	0	1	0	1	0	1
10	1	0	1	1	0	0	1	1	1	0

After initializing the TIS matrix with the partial information, we get Table 10(a), then we apply our back-tracing method in order to have a fully-known TIS matrix. Day by day we go back in reverse order and estimate the probability of being infected for each individual (the days before they were identified) until we reach the first day of the epidemic. Since the TIS matrix is now fully-known, hence, we can estimate the value of the infection rate β.

Example: Assume we have 10 individuals with adjacency matrix of Table 9. The TIS matrix is developed initially as in Table 10(a). From this table we see that on day 11 ($t = 11$) the infection status of all the individuals is known. Lets assume that on day 10 ($t = 10$) the infection status of all the individuals (except individual 10) is known. Now, we can use our model to estimate the probability of infection status of individual 10 on day 10 ($t = 10$). Our models estimate this probability conditioning on the infection status of the individual on day $t+1$. Based on our models, since $TIS(10, 11) = 0$, then Equation 4.11 will be applied.

Note however that as we go back day by day, the number of individuals with unknown status of infection increases. So, if we have more than one individual with unknown status of infection, we need to pick one of them by random and estimate the probability that this individual was infected on that day. As shown in Table 10(b), on day 5 for example, individual 9 and individual 10 both have unknown status of infection, but we randomly picked individual 9. According to the adjacency matrix of Table 9, individual 10 is a neighbor of individual 9. Since the infection status of individual 10 in day 5 is unknown, hence, we need to assume a temporary value for the probability of infection

of individual 10 in order to be able to apply Equation 4.11. So, we assume that any individual with unknown infection status, will temporary be given the prevalence of the disease $\hat{p}$ as their probability of infection on that day, if they have a link with the individual that we are estimating its infection status (individual 9 in this case). Table 10(c) illustrates this step. After estimating the health status of individual 9 on day 5, $TIS(9, 5)$ is set accordingly. Since, the health status of individual 9 on day 5 is now known, then we are able to estimate the health status of individual 10. In general, when the health status of n_u individual is unknown and $n_u > 1$, then we pick up $n_u - 1$ individuals randomly and assume that their health status equals the parameter value $\hat{p}$. Consequently, the health status of the remaining non-selected individuals are estimated and the relevant locations of the TIS matrix are updated. Next, one of the remaining $n_u - 1$ individuals is selected uniformly and its health status is estimated accordingly. We repeat these steps until we have a fully-known TIS matrix. The R code implementations of the algorithms are available at https://github.com/ralsehib/Long-Backward-Tracing-Method.git. The implementation of the Long Back-Tracing Method is provided in the psuedocode of Algorithm 7.

Table 10: An example of the steps of the Long Back-Tracing method. Note that the rows represent individuals and columns represent days.

(a) Estimating the probability of infection status for $i = 10$ on $t = 10$.

10	?	?	?	?	?	?	?	?	?	?	0
9	?	?	?	?	?	0	0	1	1	1	1
8	?	?	?	?	1	1	1	1	1	1	1
7	?	?	?	0	0	0	1	1	1	1	1
6	?	?	1	1	0	0	0	1	1	1	1
5	?	0	0	1	1	0	0	1	1	1	0
4	?	1	1	1	1	1	1	1	1	1	1
3	?	0	0	0	0	0	1	1	1	0	1
2	?	1	1	1	1	1	1	1	1	1	1
1	1	1	1	0	1	1	1	1	1	1	1
	1	2	3	4	5	6	7	8	9	10	11

(b) Selecting i at random on $t = 5$.

10	?	?	?	?	?	0	1	0	0	0	0
9	?	?	?	?	?	0	0	1	1	1	1
8	?	?	?	?	1	1	1	1	1	1	1
7	?	?	?	0	0	0	1	1	1	1	1
6	?	?	1	1	0	0	0	1	1	1	1
5	?	0	0	1	1	0	0	1	1	1	0
4	?	1	1	1	1	1	1	1	1	1	1
3	?	0	0	0	0	0	1	1	1	0	1
2	?	1	1	1	1	1	1	1	1	1	1
1	1	1	1	0	1	1	1	1	1	1	1
	1	2	3	4	5	6	7	8	9	10	11

(c) Estimating the probability of infection status for $i = 9, t = 5$.

10	?	?	?	?	$\hat{p}$	0	1	0	0	0	0
9	?	?	?	?	?	0	0	1	1	1	1
8	?	?	?	?	1	1	1	1	1	1	1
7	?	?	?	0	0	0	1	1	1	1	1
6	?	?	1	1	0	0	0	1	1	1	1
5	?	0	0	1	1	0	0	1	1	1	0
4	?	1	1	1	1	1	1	1	1	1	1
3	?	0	0	0	0	0	1	1	1	0	1
2	?	1	1	1	1	1	1	1	1	1	1
1	1	1	1	0	1	1	1	1	1	1	1
	1	2	3	4	5	6	7	8	9	10	11

Algorithm 7 Estimating the Infection Rate $\hat{\beta}$ using the Long Backward-Tracing Method

Generate TIS matrix as in Algorithm 5
for rep =1, ..., repetitions **do**
  Replace the entries of the unknown-status in the TIS by $\hat{p}$ from Algorithm 6
  for $t = (l-1), ..., 1$ **do**
    Identify the list of Unknown Infected Status (UIS)
    if length(UIS)>0 **then**
      for w=1, ..., length(UIS) **do**
        Pick a random sample v uniformly from the UIS list
        Remove the sampled individual from the UIS list
        if $TIS[t+1, v] = 1$ **then**
          $P(I|I) = \frac{e^{-\gamma}\hat{p}}{e^{-\gamma}\hat{p}+(1-e^{-\beta\eta_{i,t}})(1-\hat{p})}$
          $TIS[t, v] = sample(0:1, prob = c(P(I|I), 1-P(I|I)))$
        else if $TIS[t+1, v] = 0$ **then**
          $P(I|S) = \frac{(1-e^{-\gamma})\hat{p}}{(1-e^{-\gamma})\hat{p}+(e^{-\beta\eta_{i,t}})(1-\hat{p})}$
          $TIS[t, v] = sample(0:1, prob = c(P(I|S), 1-P(I|S)))$
        end if
      end for
    end if
    for $t = 2, ..., l$ **do**
      for $i = 1, ..., n$ **do**
        if $TIS[t-1, i] = 0 \wedge TIS[t, i] = 0$ **then**
          $P(S|S) = e^{-\beta\eta_{t,i}}$
          $sum\_of\_log = sum\_of\_log + log(P(S|S))$
        else if $TIS[t-1, i] = 0$ and $TIS[t, i] = 1$ **then**
          $P(I|S) = (1-e^{-\beta\eta_{t,i}})$
          $sum\_of\_log = sum\_of\_log + log(P(I|S))$
        end if
      end for
    end for
    Estimate the MLE(β)
  end for
end for

4.3.2 The Shortcut Back Tracing Method

The shortcut method implies that all the individuals with the unknown status are equally likely to be infected. Also, we assume that the likelihood of being infected is equal to the prevalence of the disease $\hat{p}$. The R code of the algorithms are available at https://github.com/ralsehib/Shortcut-Method.git. Algorithm 8 shows the pseudocode needed to implement the Shortcut Method. Table

11 displays an example TIS matrix using the Shortcut method. As mentioned earlier, the probability that an individual i was susceptible on day t is equal to $e^{-\beta\eta_{i,t-1}}$. We define $\eta_{i,t-1}$ as the number of infected neighbors for individual i on day $t-1$. However, in this case when on day $t-1$ an individual i has neighbors whose infectious status is unknown, then $\eta_{i,t-1}$ will include the number of infected neighbors plus the number of the neighbors with unknown infectious status multiplied by the prevalence of the disease $\hat{p}$. More precisely,

$$\eta_{i,t-1} = BK_i + BN_i * \hat{p} \tag{4.13}$$

where, BK_i = the number of infected neighbors.

BN_i = the number of neighbors with unknown infection status.

Hence, the probability that indivudal i was healthy on day $t-1$ given that he or she was sick on day t can be estimated as follows,

$$P[X_{i,t} = 1 | X_{i,t-1} = 0] = 1 - e^{-\beta\eta_{i,t-1}} \tag{4.14}$$

$$P[X_{i,t} = 1 | X_{i,t-1} = 1] = e^{-\beta\eta_{i,t-1}} \tag{4.15}$$

Table 11: An example TIS matrix for the Shortcut method.

10	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	1	1
9	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	1	1
8	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	0	0	0
7	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	1	0	0	1
6	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	0	0	1	1	1
5	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	1	1	1	0	0	0
4	$\hat{p}$	$\hat{p}$	$\hat{p}$	$\hat{p}$	0	0	0	1	1	1	0
3	$\hat{p}$	$\hat{p}$	$\hat{p}$	1	1	1	1	0	0	0	1
2	$\hat{p}$	0	0	0	0	1	1	1	1	0	0
1	1	1	1	1	1	0	0	0	0	1	1
	1	2	3	4	5	6	7	8	9	10	11

Algorithm 8 Estimating the Infection Rate $\hat{\beta}$ using the Shortcut Method

Generate TIS matrix as in Algorithm 5
for rep =1, ..., repetitions **do**
    Replace the entries of the unknown-status in the TIS by $\hat{p}$
    for $t = 2, ..., l$ **do**
        for $i = 1, ..., n$ **do**
            if $TIS[t-1, i] = 0 \wedge TIS[t, i] = 0$ **then**
                $P(S|S) = e^{-\beta \eta_{t,i}}$
                $sum\_of\_log = sum\_of\_log + log(P(S|S))$
            else if $TIS[t-1, i] = 0 \wedge TIS[t, i] = 1$ **then**
                $P(I|S) = (1 - e^{-\beta \eta_{t,i}})$
                $sum\_of\_log = sum\_of\_log + log(P(I|S))$
            end if
        end for
    end for
    Estimate the MLE(β)
end for

4.4 Simulation

The performance of the proposed methods are evaluated using simulation with a population of 200 individuals and 200 time units assuming an SIS network-based framework. The R software package used is Rstudio version 1.1.419. We assume the probability of two individuals being connected is $p = 0.6$, the rate of transmission is $\beta = 0.005$, and the rate of recovery is $\gamma = 0.06$. During the simulation, the individuals will either remain susceptible and unknown (SU), will move to susceptible and known (SK) state, will move to infected and unknown (IU) state, or will move to infected and known (IK) state. Throughout the simulation we assume that the recovery rate is independent of the network, i.e., recommendations of some control method cannot affect the recovery rate.

4.4.1 Results

The impact of the sampling strategy on the daily infections is shown in Figure 30 where the underlying network is generated according to Erdős–Rényi model of 200 nodes connecting to each other with a probability of 0.06 and 200 time steps. The SIS parameters are set as follows $\beta_N =$

0.001, $\gamma_N = 0.01$, $\beta_K = 0.0003$, and $\gamma_K = 0.06$. In this experiment we assume that we sample a maximum of one individual per day. From this Figure we can see that as the number of sampled individuals increases the number of daily infections increases until day 17 then starts to gradually decrease. Throughout the simulation, whenever an individual is sampled, then this individual is moved from the unknown class to the known class. Since we assume that the infection rate for the known class is less than that of the unknown class, then this sampled individual will be less likely to be infected. Furthermore, if the sampled individual is found to be infected, then he or she will be more likely to recover under the known class compared to the unknown class. Therefore, since in the early days, very few people are known, then the disease will spread rapidly until a certain date, beyond which as more individuals are sampled then the spread of the disease will be decreasing.

Note that from the results of chapter 3, we found that when the underlying network structure is based on the ER graph model, then the different sampling methods have comparable performance. On the other hand, when the underlying network is based on BA graphs that are composed of two weakly connected clusters, the proposed hybrid sampling method yields higher performance compared to other alternatives. Therefore, in this chapter, we will focus on the performance of the joint sampling and infection process where the sampling method is based on the HS algorithm and the underlying graph is based on the BA model. Figure 31 and Figure 32 show the results for the joint sampling-infection process for the HS method compared to that of the SRS method. The results indicate that the three levels of the HS method, i.e. when $0 < Z < 1$ lead to smaller total number of infections as well as smaller peak infection rates.

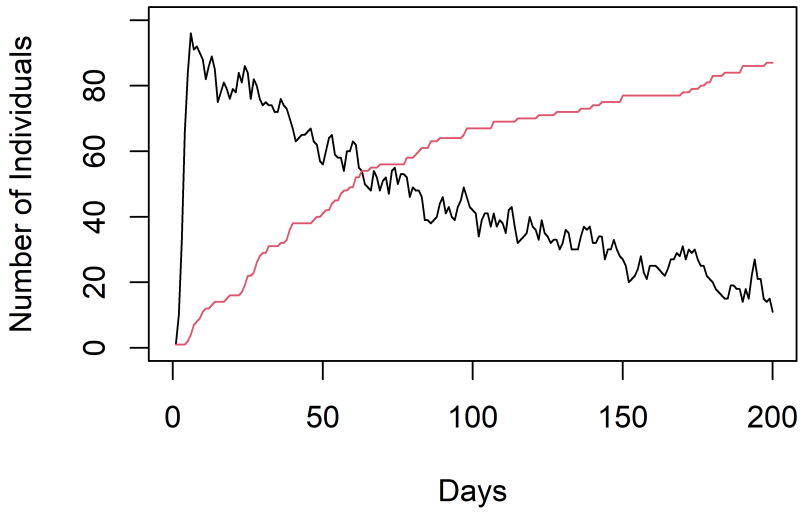


Figure 30: The number of infected and the number of sampled individuals per day. The black curve represents the number of infected individuals. The red curve represents the number of sampled individuals.

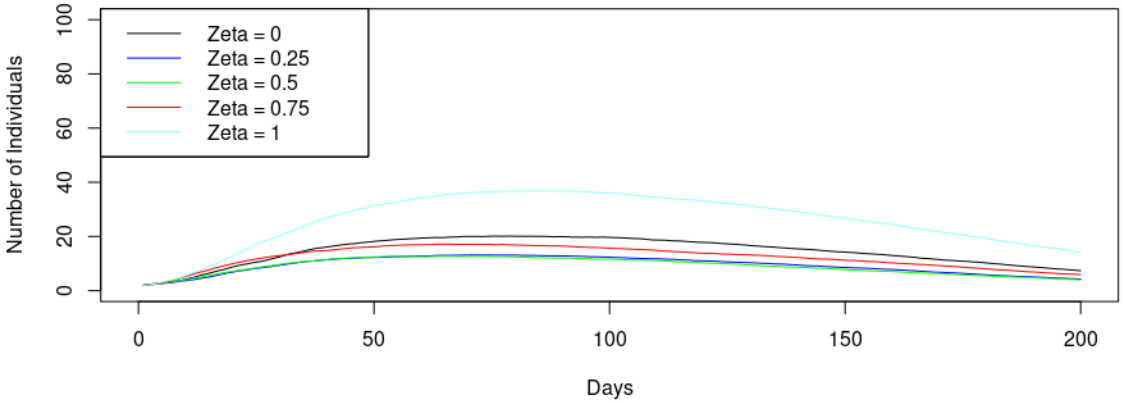


Figure 31: The number of infections for BA graph with two weakly connected clusters.

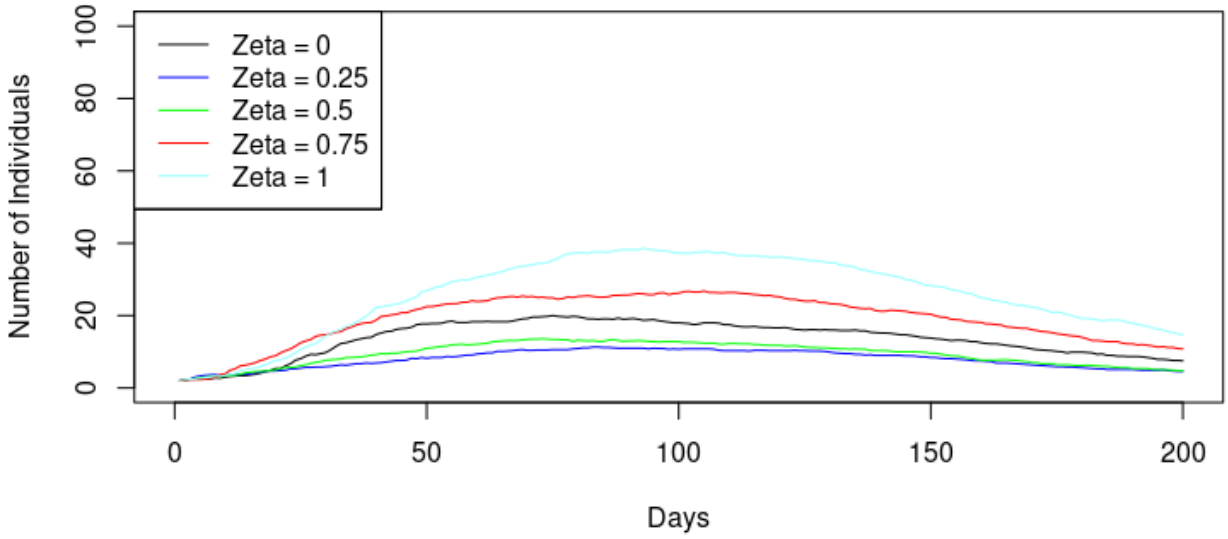
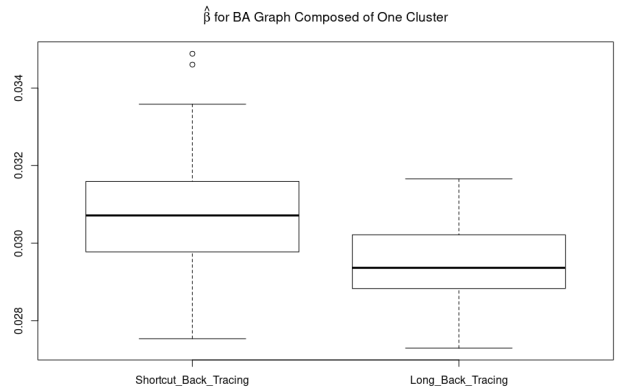
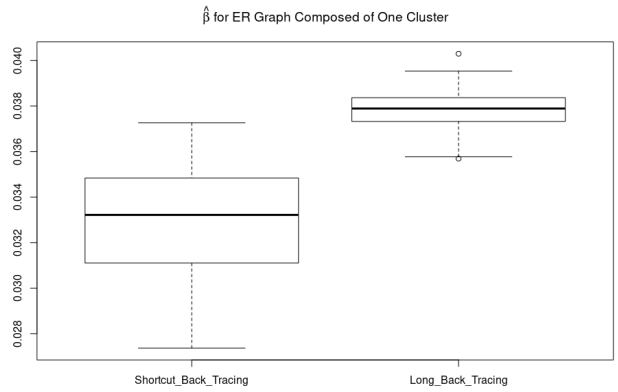


Figure 32: The number of infections for a single cluster BA graph.

The Shortcut method is expected to be faster than the Long Back-Tracing method. However, the estimated likelihood from the Shortcut method could be biased since the number of individuals with unknown infection status is proportional to the total number of sampled individuals. We applied Long Back-Tracing method and the Shortcut method on two different graphs based on the BA and the ER model. To make the two different graphs comparable, we fixed the number of nodes, and chose the number of edges to be almost the same. Figure 33 illustrates the values of $\hat{\beta}$ for a BA graph and an ER graph composed of one cluster for the Shortcut method and the Long Back-Tracing method. Both methods have overall underestimated the values of $\hat{\beta}$ for both graphs. The Long Back-Tracing method has estimated the value of $\hat{\beta}$ for the ER graph better than estimating the value of $\hat{\beta}$ for the BA graph. Also, there is a lower outlier in the box-plot for the ER graph when using the Long Back-Tracing method. Moreover, the Shortcut method has estimated the value of $\hat{\beta}$ for the ER graph better than estimating the value of $\hat{\beta}$ for the BA graph.



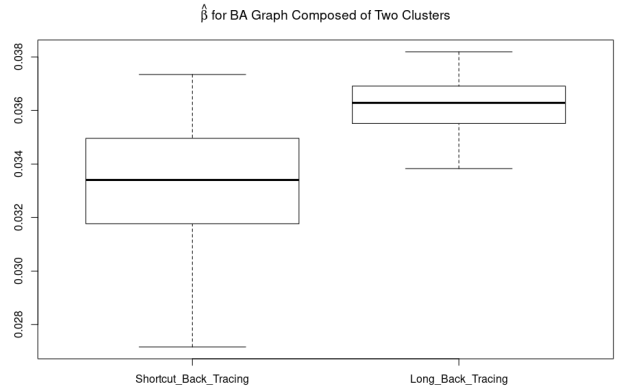
(a) BA graph



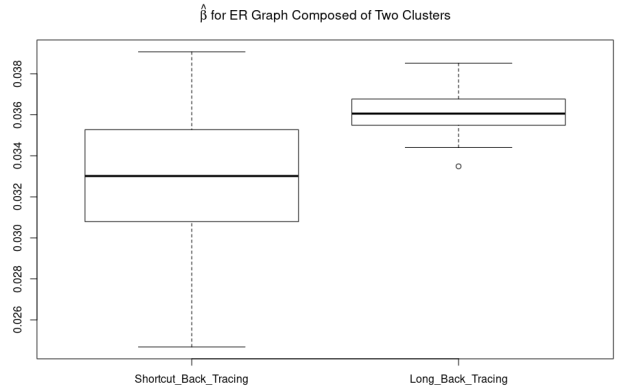
(b) ER graph

Figure 33: The box-plot of $\hat{\beta}$ for a graph composed of one cluster.

Figure 34 illustrates the values of $\hat{\beta}$ using the Long Back-Tracing method for a BA graph and an ER graph composed of two clusters each. We can see that our model has overall underestimated the value of $\hat{\beta}$ for both graphs. The Long Back-Tracing model estimated the value of $\hat{\beta}$ for the ER graph better than estimating the value of $\hat{\beta}$ for the BA graph. Also, there is a lower outlier in the box-plot for the ER graph using the Long Back-Tracing method. Figure 35 illustrates the values of $\hat{\gamma}$ for a BA graph and an ER graph composed of two clusters each. It is clear that the values of $\hat{\gamma}$ is close to each other for both graphs. However, the values of $\hat{\gamma}$ for the BA graph has more variability compared to the ER graph. Our model has overestimated the value of γ for both graphs.



(a) BA graph



(b) ER graph

Figure 34: The box-plot of $\hat{\beta}$ for a graph composed of two clusters.

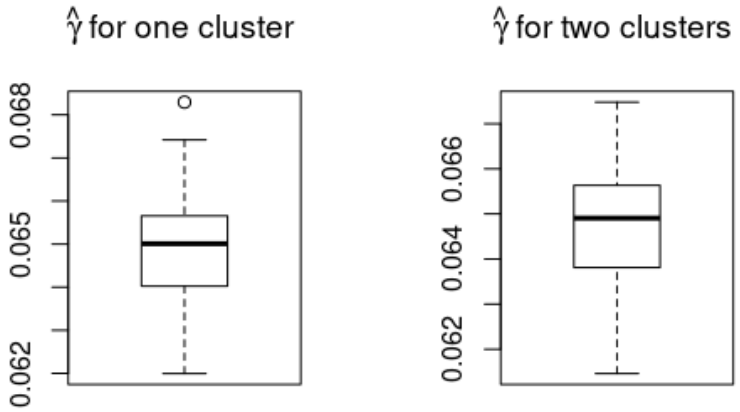


Figure 35: The box-plot of $\hat{\gamma}$ for a BA graph composed of one cluster and two clusters.

The results show that the Long Back-Tracing method has a better estimation of the transmission rate β for the ER graph compared to the BA graph. This is true whether the adjacency matrix is composed of one cluster or two clusters. Whereas the Shortcut method has a better estimation of the transmission rate β for the BA graph compared to the ER graph. This is true whether the adjacency matrix is composed of one cluster or two clusters.
In addition, the recovery rate is independent of the past missing information and therefore it can be calculated analytically, with only partially-known networks, by applying Equation (4.11). We can see that the values of $\hat{\gamma}$ are very close to each other in both methods; the Back-Tracing method and the Shortcut method, regardless of the underlying graph and the number of clusters. As for the speed of the two methods, we observe that the average computational time for the Long Back Tracing method is 81.8 seconds per iteration whereas the Shortcut method requires a significantly less average computational time of only 18.9 seconds per iteration resulting in an improvement of about 332%.

In conclusion, the Shortcut method is less costly in terms of the computation time compared to the Back-Tracing method, but the Back-Tracing method is better than the Shortcut method for estimating the transmission rate ($\hat{\beta}$) in the case of the ER graph. Whereas the Short-

cut method is better than the Back-Tracing method for estimating the transmission rate ($\hat{\beta}$) in the case of the BA graph.

4.5 Future Work

Typically, the infection rate as well as the recovery rate evolves over time as more nodes are sampled and moved from the unknown to the known state and therefore a potential future research direction is to explore the impact of variations in the value of the infection rates and the recovery rates as the simulation proceeds over time. Another future research direction is to study how pool testing can be jointly implemented with network sampling in order to reduce the total number of infections as well as the peak number of infection per time unit. In addition, there is a need to analyze the impact of the number of reinfections of an individual on the joint sampling-infection process. Moreover, in social networks, individuals might have interactions through more than one type of links where for example two individuals might be co-workers and at the same time they are family members. Therefore, future research should study the impact of non-binary networks where any two nodes might be connected through several links on the joint sampling-infection process. Finally, networks can be dynamic in nature where new individuals might join the network and current individuals are allowed to leave the network. Hence, a potential future direction is to explore how the proposed methods can be extended to include the case of dynamic networks.

CHAPTER 5

Conclusions

Reducing the negative impacts of epidemics require accurate detection and control methods. Effectiveness of the detection of infected individuals can be improved using pool testing. This thesis investigates the impact of pooling multiplicity on the accuracy of pool testing by developing models for higher levels of multiplicity pool testing, taking the probability of testing errors into consideration. Through simulation, the impact of several positivity classification protocols (thresholds) on pool testing accuracy: specificity and sensitivity, is evaluated using the ROC and the AUC. In addition, the impact of the batch size on the pool testing accuracy is also examined. The results indicate that under certain conditions multiplicity pool testing yields superior testing accuracy compared to individual testing without additional cost. The findings also demonstrate that pool testing gives higher gains in terms of pool testing sensitivity compared to individual testing in the case when the manufacturer reported sensitivity and the prevalence are low. The findings also reveal that the improvement in accuracy is a function in the multiplicity level, the classification threshold, and the batch size where the performance can be improved using a batch size that is inversely proportional to the prevalence level. The manufacturer's test sensitivity however has more significant impact on the accuracy of pool testing compared to that of manufacturer's test specificity.

Control of epedimics requires modeling the spread of the disease over social networks which are often only partially known and the identification of individuals in these types of networks is essential. Similar to the compartmental susceptible-infected (SI) virus propagation model in epidemiology, this thesis develops an unknown-known (NK) compartmental framework where individuals are sampled and moved permanently from the unknown state to the known state. An iterative hybrid sampling method composed of partial simple random sampling and partial network sampling is developed. Several levels of the partial components are implemented where the network

sampling method is based on the network locality of the sampled substructure in every iteration. The performance of the proposed method is evaluated in terms of the Perron eigenvalue of the sampled subnetwork using simulation. The performance evaluation shows that the hybrid sampling method has significantly superior performance compared to simple random sampling. The performance of the different levels of the partial combinations of the simple random sampling and the network sampling is also evaluated where we find that the different hybrid combinations give distinct outcomes under varying conditions.

The spread of infectious diseases leads to huge negative impacts on social and economic stability worldwide. One factor that contributes significantly to the fast spread of infectious diseases is the level of contact through social networks among individuals. Statistical modeling of the spread of the infectious diseases can enable researchers and decision makers to have a better understanding of the spread of infectious diseases and to develop more effective control measures to contain the outbreak. Individual-level models (ILMs) have emerged as a more realistic alternative to population-level models since they take into consideration the heterogeneity in the ability of individuals to infect or get infected by others. Therefore, researchers are increasingly adopting the use individual-level modelling (ILM) processes to analyze the spread of infectious diseases. Lately, a new line of research has emerged that takes advantage of developments in the network theory where relations among individuals can be accurately and efficiently represented using a network structure. Therefore, network-based ILMs began to attract the attention of researchers where the contact patterns among individuals as well as the transmission of a disease can be modeled using networks. Prior research in network-based ILMs, however, considered the contact network to be fully known, which is implausible and unrealistic for many reasons. Hence, there is a need to model and analyze the spread of infectious diseases in partially-known networks. This thesis develops statistical models to estimate the infection rate and the recovery rate in partially-known networks. The epidemic spread process is modelled as a Markov Chain process taking into consideration the virus propagation model, the network adjacency information, and the sampling process. The virus propagation model is assumed to follow the Susceptible-Infectious-Susceptible (SIS) model. A simulation model is developed to analyze the performance of the proposed process which starts by sampling individuals and applying the virus propagation model jointly to simulate the spread

of the disease. The recovery rate is then calculated and two back tracing methods are developed to estimate the health status of individuals in the days before the date they are sampled so the infection rate can be estimated for several network types. The simulation results show that there are tradeoffs between these two methods in terms of speed and accuracy.

BIBLIOGRAPHY

[1] Mohammed Al Mugahwi, Omar De la Cruz Cabrera, Silvia Noschese, and Lothar Reichel. Functions and Eigenvectors of Partially Known Matrices with Applications to Network Analysis. Applied Numerical Mathematics, 159:93–105, 2021.

[2] Khalid Al-Naamani, Issa Al-Jahdhami, Wafa Al-Tamtami, Kawther Al-Amri, Murtadha Al-Khabori, Siham Al Sinani, Elias A Said, Heba Omer, Hamad Al-Bahluli, Saada Al-Ryiami, et al. Prevalence and Persistence of SARS-CoV2 Antibodies among Healthcare Workers in Oman. Journal of Infection and Public Health, 14(11):1578–1584, 2021.

[3] Matthew Aldridge, Leonardo Baldassini, and Oliver Johnson. Group Testing Algorithms: Bounds and Simulations. IEEE Transactions on Information Theory, 60(6):3671–3687, 2014.

[4] Matthew Aldridge, Oliver Johnson, Jonathan Scarlett, et al. Group Testing: an Information Theory Perspective. Foundations and Trends in Communications and Information Theory, 15(3-4):196–392, 2019.

[5] Valerij Borisovič Alekseev. Abel's Theorem in Problems and Solutions: Based on the Lectures of Professor VI Arnold. Springer Science & Business Media, 2004.

[6] Thamer H Alenazi, Nasser F BinDhim, Meteb H Alenazi, Hani Tamim, Reem S Almagrabi, Sameera M Aljohani, Mada H Basyouni, Rasha A Almubark, Nora A Althumiri, and Saleh A Alqahtani. Prevalence and Predictors of Anxiety among Healthcare Workers in Saudi Arabia During the COVID-19 Pandemic. Journal of Infection and Public Health, 13(11):1645–1651, 2020.

[7] Areej AlFattani, Amani AlMeharish, Maliha Nasim, Khalid AlQahtani, and Sami AlMudraa. Ten Public Health Strategies to Control the COVID-19 Pandemic: The Saudi Experience. IJID Regions, 1:12–19, 2021.

[8] Abdullah A Algaissi, Naif Khalaf Alharbi, Mazen Hassanain, and Anwar M Hashem. Pre-

paredness and Response to COVID-19 in Saudi Arabia: Building on MERS Experience. Journal of Infection and Public Health, 13(6):834–838, 2020.

[9] Waleed Almutiry. Incorporating Contact Network Uncertainty in Individual Level Models of Infectious Disease within a Bayesian Framework. PhD thesis, 2018.

[10] Sultanah M Alshammari, Waleed K Almutiry, Harsha Gwalani, Saeed M Algarni, and Kawther Saeedi. Measuring the Impact of Suspending Umrah, a Global Mass Gathering in Saudi Arabia on the COVID-19 Pandemic. Computational and Mathematical Organization Theory, pages 1–26, 2021.

[11] Thamir M Alshammari, Ali F Altebainawi, and Khalidah A Alenzi. Importance of Early Precautionary Actions in Avoiding the Spread of COVID-19: Saudi Arabia as an Example. Saudi Pharmaceutical Journal, 28(7):898–902, 2020.

[12] Fahad Alsharif. Undocumented Migrants in Saudi Arabia: COVID-19 and Amnesty Reforms. International Migration, 60(1):188–204, 2022.

[13] Douglas G Altman and J Martin Bland. Diagnostic Tests. 1: Sensitivity and Specificity. BMJ: British Medical Journal, 308(6943):1552, 1994.

[14] Hrayer Aprahamian, Douglas R Bish, and Ebru K Bish. Optimal Risk-Based Group Testing. Management Science, 65(9):4365–4384, 2019.

[15] Hrayer Aprahamian, Ebru K Bish, and Douglas R Bish. Adaptive Risk-Based Pooling in Public Health Screening. IISE Transactions, 50(9):753–766, 2018.

[16] Esam I Azhar, Sherif A El-Kafrawy, Suha A Farraj, Ahmed M Hassan, Muneera S Al-Saeed, Anwar M Hashem, and Tariq A Madani. Evidence for Camel-to-Human Transmission of MERS Coronavirus. New England Journal of Medicine, 370(26):2499–2505, 2014.

[17] Allen C Bateman, Shanna Mueller, Kyley Guenther, and Peter Shult. Assessing the Dilution Effect of Specimen Pooling on the Sensitivity of SARS-CoV-2 PCR Tests. Journal of Medical Virology, 93(3):1568–1572, 2021.

[18] Niels G Becker. Analysis of Infectious Disease Data. Chapman and Hall, 1989.

[19] Mostefa Bensaada, Mohamed Amine Smaali, Oussama Bahi, Khalid Bouhedjar, Foudil Khelifa, Feriel Sellam, and Saad Mebrek. Improvement of SARS-CoV-2 Screening Using Pooled Sampling Testing in Limited RT-qPCR Resources. Journal of Virological Methods, 300:114421, 2022.

[20] Douglas R Bish, Ebru K Bish, Hussein El-Hajj, and Hrayer Aprahamian. A Robust Pooled Testing Approach to Expand COVID-19 Screening Capacity. PLoS One, 16(2):e0246285, 2021.

[21] Cameron Browne, Hayriye Gulbudak, and Glenn Webb. Modeling Contact Tracing in Outbreaks with Application to Ebola. Journal of Theoretical Biology, 384:33–49, 2015.

[22] Simon Cauchemez, Achuyt Bhattarai, Tiffany L Marchbanks, Ryan P Fagan, Stephen Ostroff, Neil M Ferguson, David Swerdlow, Pennsylvania H1N1 Working Group, et al. Role of Social Networks in Shaping Disease Transmission During a Community Outbreak of 2009 H1N1 Pandemic Influenza. Proceedings of the National Academy of Sciences, 108(7):2825–2830, 2011.

[23] Johns Hopkins Coronavirus Resource Center. Mortality Analyses. 2-15-2020.

[24] Anirudh Chakravarthy, Srikar Krishna, Sumana Ghosh, Ajay Tomar, Sriram Varahan, Ajit Rajwade, Sabyasachi Ghosh, Nimay Gupta, Rishi Agarwal, Himanshu Payal, et al. Large-Scale Testing for SARS-CoV-2 using Tapestry Pooling. medRxiv, 2020.

[25] Chun Lam Chan, Pak Hou Che, Sidharth Jaggi, and Venkatesh Saligrama. Non-Adaptive Probabilistic Group Testing with Noisy Measurements: Near-Optimal Bounds with Efficient Algorithms. In 2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pages 1832–1839. IEEE, 2011.

[26] Chun Lam Chan, Sidharth Jaggi, Venkatesh Saligrama, and Samar Agnihotri. Non-Adaptive Group Testing: Explicit Bounds and Novel Algorithms. IEEE Transactions on Information Theory, 60(5):3019–3035, 2014.

[27] Yi-Cheng Chen, Ping-En Lu, Cheng-Shang Chang, and Tzu-Hsuan Liu. A Time-Dependent

SIR Model for COVID-19 with Undetectable Infected Persons. IEEE Transactions on Network Science and Engineering, 7(4):3279–3294, 2020.

[28] Xiwei Cheng, Sidharth Jaggi, and Qiaoqiao Zhou. Generalized Group Testing. In International Conference on Artificial Intelligence and Statistics, pages 10777–10835. PMLR, 2022.

[29] Nicholas A Christakis and James H Fowler. Social Network Sensors for Early Detection of Contagious Outbreaks. PLoS One, 5(9):e12948, 2010.

[30] Ana Paula Christoff, Giuliano Netto Flores Cruz, Aline Fernanda Rodrigues Sereia, Dellyana Rodrigues Boberg, Daniela Carolina De Bastiani, Laís Eiko Yamanaka, Gislaine Fongaro, Patrícia Hermes Stoco, Maria Luiza Bazzo, Edmundo Carlos Grisard, et al. Swab Pooling: A New Method for Large-Scale RT-qPCR Screening of SARS-CoV-2 Avoiding Sample Dilution. PLoS One, 16(2):e0246544, 2021.

[31] Daniel KW Chu, Leo LM Poon, Mokhtar M Gomaa, Mahmoud M Shehata, Ranawaka APM Perera, Dina Abu Zeid, Amira S El Rifay, Lewis Y Siu, Yi Guan, Richard J Webby, et al. MERS Coronaviruses in Dromedary Camels, Egypt. Emerging Infectious Diseases, 20(6):1049, 2014.

[32] Dinh-Toi Chu, Suong-Mai Vu Ngoc, Hue Vu Thi, Yen-Vy Nguyen Thi, Thuy-Tien Ho, Van-Thuan Hoang, Vijai Singh, and Jaffar A Al-Tawfiq. COVID-19 in Southeast Asia: Current Status and Perspectives. Bioengineered, 13(2):3797–3809, 2022.

[33] Reuven Cohen, Shlomo Havlin, and Daniel Ben-Avraham. Efficient Immunization Strategies for Computer Networks and Populations. Physical Review Letters, 91(24):247901, 2003.

[34] Krista Conger. Testing Pooled Samples for COVID-19 Helps Stanford Researchers Track Early Viral Spread in Bay Area. Stanford Med., 2020.

[35] Peter Damaschke. Threshold Group Testing. In General Theory of Information Transfer and Combinatorics, pages 707–718. Springer, 2006.

[36] Timo de Wolff, Dirk Pflüger, Michael Rehme, Janin Heuer, and Martin-Immanuel Bittner. Evaluation of Pool-Based Testing Approaches to Enable Population-Wide Screening for COVID-19. PLoS One, 15(12):e0243692, 2020.

[37] Rob Deardon, Stephen P Brooks, Bryan T Grenfell, Matthew J Keeling, Michael J Tildesley, Nicholas J Savill, Darren J Shaw, and Mark EJ Woolhouse. Inference for Individual-Level Models of Infectious Diseases in Large Populations. Statistica Sinica, 20(1):239, 2010.

[38] Rob Deardon, Xuan Fang, and G Kwong. Statistical Modeling of Spatiotemporal Infectious Disease Transmission. Analyzing and Modeling Spatial and Temporal Dynamics of Infectious Diseases, page 211, 2014.

[39] Lorna E Deeth and Rob Deardon. Latent conditional individual-level models for infectious disease modeling. The International Journal of Biostatistics, 9(1):75–93, 2013.

[40] Robert Dorfman. The Detection of Defective Members of Large Populations. The Annals of Mathematical Statistics, 14(4):436–440, 1943.

[41] Akira Endo, Quentin J Leclerc, Gwenan M Knight, Graham F Medley, Katherine E Atkins, Sebastian Funk, Adam J Kucharski, et al. Implication of Backward Contact Tracing in the Presence of Overdispersed Transmission in COVID-19 Outbreaks. Wellcome Open Research, 5, 2020.

[42] Agustín Estévez, Pilar Catalán, Roberto Alonso, Mercedes Marín, Emilio Bouza, Patricia Muñoz, Luís Alcalá, COVID19 study group, et al. Sample Pooling is Efficient in PCR Testing of SARS-CoV-2: a Study in 7400 Healthcare Professionals. Diagnostic Microbiology and Infectious Disease, 100(1):115330, 2021.

[43] Benjamin Isac Fargion, Daniele Fargion, Pier Giorgio De Sanctis Lucentini, and Emanuele Habib. Purim: a Rapid Method with Reduced Cost for Massive Detection of COVID-19. arXiv preprint arXiv:2003.11975, 2020.

[44] Scott L Feld. Why Your Friends Have More Friends Than You Do. American Journal of Sociology, 96(6):1464–1477, 1991.

[45] HM Finucan. The Blood Testing Problem. Journal of the Royal Statistical Society: Series C (Applied Statistics), 13(1):43–50, 1964.

[46] Christopher M Florkowski. Sensitivity, Specificity, Receiver-Operating Characteristic (ROC) Curves and Likelihood Ratios: Communicating the Performance of Diagnostic Tests. The Clinical Biochemist Reviews, 29(Suppl 1):S83, 2008.

[47] Centers for Disease Control and Prevention. Lesson 1: Introduction to Epidemiology. 5-18-2012.

[48] Troy J Ganz, Rachel Donner, Kevin M Hines, Markus L Waithe-Alleyne, Deirdre L Slate, Gyorgy Abel, and Jared R Auclair. Two-Stage Hierarchical Group Testing Strategy to Increase SARS-CoV-2 Testing Capacity at an Institution of Higher Education: A Retrospective Analysis. The Journal of Molecular Diagnostics, 23(12):1691–1698, 2021.

[49] Sabyasachi Ghosh, Rishi Agarwal, Mohammad Ali Rehan, Shreya Pathak, Pratyush Agarwal, Yash Gupta, Sarthak Consul, Nimay Gupta, Ritesh Goenka, Ajit Rajwade, et al. A Compressed Sensing Approach to Pooled RT-PCR Testing for COVID-19 Detection. IEEE Open Journal of Signal Processing, 2:248–264, 2021.

[50] Katherine R Goetzinger and Anthony O Odibo. Statistical Analysis and Interpretation of Prenatal Diagnostic Imaging Studies, Part 1: Evaluating the Efficiency of Screening and Diagnostic Tests. Journal of Ultrasound in Medicine, 30(8):1121–1127, 2011.

[51] M Gogia, C Lawlor, N Shengelia, K Stvilia, and HF Raymond. Hidden Populations: Discovering the Differences between the Known and the Unknown Drug using Populations in the Republic of Georgia. Harm Reduction Journal, 16(1):15, 2019.

[52] Chris Groendyke, David Welch, and David R Hunter. A Network-Based Analysis of the 1861 Hagelloch Measles Data. Biometrics, 68(3):755–765, 2012.

[53] Gregory Haber, Yaakov Malinovsky, and Paul S Albert. Is Group Testing Ready for Prime-Time in Disease Identification? Statistics in Medicine, 40(17):3865–3880, 2021.

[54] Rudolf Hanel and Stefan Thurner. Boosting Test-Efficiency by Pooled Testing for SARS-CoV-2—Formula for Optimal Pool Size. PLoS One, 15(11):e0240652, 2020.

[55] Harapan Harapan, Naoya Itoh, Amanda Yufika, Wira Winardi, Synat Keam, Haypheng Te, Dewi Megawati, Zinatul Hayati, Abram L Wagner, and Mudatsir Mudatsir. Coronavirus Disease 2019 (COVID-19): A Literature Review. Journal of Infection and Public Health, 13(5):667–673, 2020.

[56] Julian Heidecke, Jan Fuhrmann, and Maria Vittoria Barbarossa. A Mechanistic Model to Assess the Effectiveness of Test-Trace-Isolate-and-Quarantine Under Limited Capacities. arXiv preprint arXiv:2207.09551, 2022.

[57] Brianna D Hitt. Group Testing Identification: Objective Functions, Implementation, and Multiplex Assays. PhD thesis, The University of Nebraska-Lincoln, 2020.

[58] Brianna D Hitt, Christopher R Bilder, Joshua M Tebbs, and Christopher S McMahan. The Objective Function Controversy for Group Testing: Much Ado about Nothing? Statistics in Medicine, 38(24):4912–4923, 2019.

[59] Petter Holme and Naoki Masuda. The Basic Reproduction Number as a Predictor for Epidemic Outbreaks in Temporal Networks. PLoS One, 10(3), 2015.

[60] Pili Hu and Wing Cheong Lau. A Survey and Taxonomy of Graph Sampling. arXiv preprint arXiv:1308.5865, 2013.

[61] Michael G Hudgens and Hae-Young Kim. Optimal Configuration of a Square Array Group Testing Algorithm. Communications in Statistics—Theory and Methods, 40(3):436–448, 2011.

[62] James M Hyman, Jia Li, and E Ann Stanley. Modeling the Impact of Random Screening and Contact Tracing in Reducing the Spread of HIV. Mathematical Biosciences, 181(1):17–54, 2003.

[63] William Kautz and Roy Singleton. Nonrandom Binary Superimposed Codes. IEEE Transactions on Information Theory, 10(4):363–377, 1964.

[64] Matt J Keeling and Ken TD Eames. Networks and Epidemic Models. Journal of the Royal Society Interface, 2(4):295–307, 2005.

[65] Matt J. Keeling and Pejman Rohani. Modeling Infectious Diseases in Humans and Animals. Princeton University Press, 2008.

[66] Hae-Young Kim and Michael G Hudgens. Three-Dimensional Array-Based Group Testing Algorithms. Biometrics, 65(3):903–910, 2009.

[67] Hae-Young Kim, Michael G Hudgens, Jonathan M Dreyfuss, Daniel J Westreich, and Christopher D Pilcher. Comparison of Group Testing Algorithms for Case Identification in the Presence of Test Error. Biometrics, 63(4):1152–1163, 2007.

[68] Istvan Z Kiss, Darren M Green, and Rowland R Kao. Disease Contact Tracing in Random and Clustered Networks. Proceedings of the Royal Society B: Biological Sciences, 272(1570):1407–1414, 2005.

[69] Sadamori Kojaku, Laurent Hébert-Dufresne, Enys Mones, Sune Lehmann, and Yong-Yeol Ahn. The Effectiveness of Backward Contact Tracing in Networks. Nature Physics, 17(5):652–658, 2021.

[70] Vineet Kumar, David Krackhardt, and Scott Feld. Interventions with Inversity in Unknown Networks can Help Regulate Contagion. arXiv preprint arXiv:2105.08758, 2021.

[71] Linyuan Lü, Duanbing Chen, Xiao-Long Ren, Qian-Ming Zhang, Yi-Cheng Zhang, and Tao Zhou. Vital Nodes Identification in Complex Networks. Physics Reports, 650:1–63, 2016.

[72] Rajat Malik, Rob Deardon, Grace PS Kwong, and Benjamin J Cowling. Individual-Level Modeling of the Spread of Influenza within Households. Journal of Applied Statistics, 41(7):1578–1592, 2014.

[73] Smriti Mallapaty et al. The Mathematical Strategy that could Transform Coronavirus Testing. Nature, 583(7817):504–505, 2020.

[74] L Daniel Maxim, Ron Niebo, and Mark J Utell. Screening Tests: a Review with Examples. Inhalation Toxicology, 26(13):811–828, 2014.

[75] Ryan Seamus McGee, Julian R Homburger, Hannah E Williams, Carl T Bergstrom, and

Alicia Y Zhou. Model-Driven Mitigation Measures for Reopening Schools During the COVID-19 Pandemic. Proceedings of the National Academy of Sciences, 118(39):e2108909118, 2021.

[76] Christopher S McMahan, Joshua M Tebbs, and Christopher R Bilder. Two-Dimensional Informative Array Testing. Biometrics, 68(3):793–804, 2012.

[77] Zelalem Mengesha, Esther Alloun, Danielle Weber, Mitchell Smith, and Patrick Harris. "Lived the Pandemic Twice": A Scoping Review of the Unequal Impact of the COVID-19 Pandemic on Asylum Seekers and Undocumented Migrants. International Journal of Environmental Research and Public Health, 19(11):6624, 2022.

[78] C.D. Meyer. Matrix Analysis and Applied Linear Algebra. Society for Industrial and Applied Mathematics (SIAM), 2000.

[79] Lauren Ancel Meyers, Babak Pourbohloul, Mark EJ Newman, Danuta M Skowronski, and Robert C Brunham. Network Theory and SARS: Predicting Outbreak Diversity. Journal of Theoretical Biology, 232(1):71–81, 2005.

[80] Johannes Müller and Mirjam Kretzschmar. Contact Tracing–Old Models and New Challenges. Infectious Disease Modelling, 6:222–231, 2021.

[81] Johannes Müller, Mirjam Kretzschmar, and Klaus Dietz. Contact Tracing in Stochastic and Deterministic Epidemic Models. Mathematical Biosciences, 164(1):39–64, 2000.

[82] Leon Mutesa, Pacifique Ndishimye, Yvan Butera, Jacob Souopgui, Annette Uwineza, Robert Rutayisire, Ella Larissa Ndoricimpaye, Emile Musoni, Nadine Rujeni, Thierry Nyatanyi, et al. A Pooled Testing Strategy for Identifying SARS-CoV-2 at Low Prevalence. Nature, 589(7841):276–280, 2021.

[83] Hamidah Nasution, Herlina Jusuf, Evi Ramadhani, and Ismail Husein. Model of Spread of Infectious Disease. Systematic Reviews in Pharmacy, 11(2), 2020.

[84] M. E. J. Newman. Networks: an Introduction. Oxford University Press, Oxford; New York, 2010.

[85] Hiroshi Nishiura, Tetsuro Kobayashi, Takeshi Miyama, Ayako Suzuki, Sung-mok Jung, Katsuma Hayashi, Ryo Kinoshita, Yichi Yang, Baoyin Yuan, Andrei R Akhmetzhanov, et al. Estimation of the Asymptomatic Ratio of Novel Coronavirus Infections (COVID-19). International Journal of Infectious Diseases, 94:154, 2020.

[86] Yitzchak Novick and Amotz Bar-Noy. Cost-Based Analyses of Random Neighbor and Derived Sampling Methods. Applied Network Science, 7(1):34, 2022.

[87] Nancy A Obuchowski. ROC Analysis. American Journal of Roentgenology, 184(2):364–372, 2005.

[88] Eamon B O'dea, Andrew W Park, and John M Drake. Estimating the Distance to an Epidemic Threshold. Journal of The Royal Society Interface, 15(143):20180034, 2018.

[89] World Health Organization. Middle East Respiratory Syndrome Coronavirus (MERS-CoV). https://www.who.int/en/news-room/fact-sheets/detail/middle-east-respiratory-syndrome-coronavirus-(mers-cov), 2022.

[90] World Health Organization. Middle East Respiratory Syndrome Coronavirus (MERS-CoV) - Saudi Arabia. https://www.who.int/emergencies/disease-outbreak-news/item/2022-DON422, 2022.

[91] Hendrik Bernd Petersen, Shankar Agarwal, Peter Jung, and Bubacarr Bah. Improving the Reliability of Pooled Testing with Combinatorial Decoding and Compressed Sensing. In 2021 55th Annual Conference on Information Sciences and Systems (CISS), pages 1–5. IEEE, 2021.

[92] RM Phatarfod and Aidan Sudbury. The Use of a Square Array Scheme in Blood Testing. Statistics in Medicine, 13(22):2337–2343, 1994.

[93] Jonathan Pugh, Dominic Wilkinson, and Julian Savulescu. Sense and Sensitivity: Can an Inaccurate Test be Better than no Test at All? Journal of Medical Ethics, 48(5):329–333, 2022.

[94] JA Quesada, A López-Pineda, VF Gil-Guillén, JM Arriero-Marín, F Gutiérrez, and C Carratala-Munuera. Incubation Period of COVID-19: A Systematic Review and Meta-Analysis. Revista Clínica Española (English Edition), 221(2):109–117, 2021.

[95] Joren Raymenants, Caspar Geenen, Jonathan Thibaut, Klaas Nelissen, Sarah Gorissen, and Emmanuel Andre. Empirical Evidence on the Efficiency of Backward Contact Tracing in COVID-19. Nature Communications, 13(1):4750, 2022.

[96] Irving S Reed and Gustave Solomon. Polynomial Codes over Certain Finite Fields. Journal of the Society for Industrial and Applied Mathematics, 8(2):300–304, 1960.

[97] Steven Riley and Neil M Ferguson. Smallpox Transmission and Control: Spatial Dynamics in Great Britain. Proceedings of the National Academy of Sciences, 103(33):12637–12642, 2006.

[98] Jonathan Scarlett and Oliver Johnson. Noisy Non-Adaptive Group Testing: A (Near-) Definite Defectives Approach. IEEE Transactions on Information Theory, 66(6):3775–3797, 2020.

[99] Christoph Schumacher and Matthias Täufer. The Statistics of Noisy One-Stage Group Testing in Outbreaks. arXiv preprint arXiv:2012.02101, 2020.

[100] Jin-Taek Seong. Theoretical Bounds on Performance in Threshold Group Testing Schemes. Mathematics, 8(4):637, 2020.

[101] Matthias Täufer. Rapid, Large-Scale, and Effective Detection of COVID-19 Via Non-Adaptive Testing. Journal of Theoretical Biology, 506:110450, 2020.

[102] Nicolas Thierry-Mieg. A New Pooling Strategy for High-Throughput Screening: the Shifted Transversal Design. BMC Bioinformatics, 7(1):1–13, 2006.

[103] Claudio M Verdun, Tim Fuchs, Pavol Harar, Dennis Elbrächter, David S Fischer, Julius Berner, Philipp Grohs, Fabian J Theis, and Felix Krahmer. Group Testing for SARS-CoV-2 Allows for up to 10-Fold Efficiency Increase Across Realistic Scenarios and Testing Strategies. Frontiers in Public Health, page 1205, 2021.

[104] Rui Wang, Yongkun Li, Shuai Lin, WeiJie Wu, Hong Xie, Yinlong Xu, and John CS Lui. Common Neighbors Matter: Fast Random Walk Sampling with Common Neighbor Awareness. IEEE Transactions on Knowledge and Data Engineering, 35(5):4570–4584, 2023.

[105] Yang Wang, Deepayan Chakrabarti, Chenxi Wang, and Christos Faloutsos. Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint. In 22nd International Symposium on Reliable Distributed Systems, 2003. Proceedings., pages 25–34. IEEE, 2003.

[106] Jessica Watson, Penny F Whiting, and John E Brush. Interpreting a COVID-19 Test Result. BMJ, 369, 2020.

[107] Xin Xu, Chul-Ho Lee, et al. A General Framework of Hybrid Graph Sampling for Complex Network Analysis. In IEEE INFOCOM 2014-IEEE Conference on Computer Communications, pages 2795–2803. IEEE, 2014.

[108] Idan Yelin, Noga Aharony, Einat Shaer Tamar, Amir Argoetti, Esther Messer, Dina Berenbaum, Einat Shafran, Areen Kuzli, Nagham Gandali, Omer Shkedi, et al. Evaluation of COVID-19 RT-qPCR Test in Multi Sample Pools. Clinical Infectious Diseases, 71(16):2073–2078, 2020.

[109] Zhuojie Zhou, Nan Zhang, and Gautam Das. Leveraging History for Faster Sampling of Online Social Networks. arXiv preprint arXiv:1505.00079, 2015.

[110] Lirong Zou, Feng Ruan, Mingxing Huang, Lijun Liang, Huitao Huang, Zhongsi Hong, Jianxiang Yu, Min Kang, Yingchao Song, Jinyu Xia, et al. SARS-CoV-2 Viral load in Upper Respiratory Specimens of Infected Patients. New England Journal of Medicine, 382(12):1177–1179, 2020.