Inference and Diffusion in Networks

by

Paolo Bertolotti

Submitted to the Institute for Data, Systems, and Society
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Social and Engineering Systems and Statistics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2022

Author .....
Institute for Data, Systems, and Society
December 10, 2021

Certified by .....
Ali Jadbabaie
JR East Professor of Engineering, Committee Chair

Certified by .....
Alberto Abadie
Professor of Economics, Committee Member

Certified by .....
Peter Kempthorne
Lecturer in Mathematics, Committee Member

Certified by .....
Devavrat Shah
Andrew (1956) and Erna Viterbi Professor, Committee Member

Accepted by .....
Fotini Christia
Chair, Social and Engineering Systems Program

Inference and Diffusion in Networks

by

Paolo Bertolotti

Submitted to the Institute for Data, Systems, and Society
on December 10, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Social and Engineering Systems and Statistics

Abstract

Networks provide a powerful and unified framework to study complex systems. By abstracting systems down to entities and their connections, network models provide insight into the structure and dynamics of critical systems across multiple domains. In this thesis, we study diffusion in social networks. Diffusion through networked systems corresponds to numerous consequential processes, and we focus on epidemic spread and information diffusion. We study these processes by applying and extending ideas from statistical inference. Inference, which focuses on estimation, testing, and uncertainty quantification, provides the mathematical tools to learn from data rigorously. This thesis utilizes both theory and data in order to address several real-world challenges.

In the first chapter, we study epidemic spread and consider the problem of identifying infected individuals in a population of size N. We introduce an approach that uses significantly fewer than N tests when infection prevalence is low. Our approach utilizes network structure to improve the performance of a classical approach called group testing. In the second chapter, we derive the performance of the most common form of group testing, Dorfman testing, under imperfect tests. We derive the full distribution of the number of tests needed, the number of false negatives, and the number of false positives, taking into account the conditions faced by medical practitioners. In the third chapter, we study information diffusion and introduce a statistical testing framework to identify cascades in network data. We define a test statistic that distinguishes between large, meaningful branches and the small branches formed during normal periods, and apply our statistic to identify information cascades in call detail record data. In the fourth chapter, we study the social network effects of drone strikes, focusing on information and physical diffusion around strikes. Utilizing a dataset of over 12 billion call detail records, we systematically analyze the impact of 74 U.S. drone strikes on communication and mobility in Yemen between 2010 and 2012.

Thesis Supervisor: Ali Jadbabaie
Title: JR East Professor of Engineering, Committee Chair

Thesis Supervisor: Alberto Abadie
Title: Professor of Economics, Committee Member

Thesis Supervisor: Peter Kempthorne
Title: Lecturer in Mathematics, Committee Member

Thesis Supervisor: Devavrat Shah
Title: Andrew (1956) and Erna Viterbi Professor, Committee Member

Acknowledgments

I would like to express my gratitude to the people who have made this thesis possible.

I would like to thank my advisor, Professor Ali Jadbabaie, for being a wonderful mentor over the past few years. Under his guidance, I have grown both as a researcher and as a person. Professor Jadbabaie taught me the importance of precision and rigor and, outside of research, has instilled in me a passion for knowledge and discovery. During my studies, he has provided an immense amount of time, support, and motivation.

It was an honor to have Professor Jadbabaie on my committee alongside Professors Alberto Abadie, Peter Kempthorne, and Devavrat Shah. Professor Abadie provided the foundation of my econometric knowledge. He has always made time to meet and discuss research questions, and has supported my work with insight and kindness. I greatly admire his character, research, and impact. Professor Kempthorne taught me statistics, and has shown me the power of using statistics to solve real-world problems. His research philosophy has greatly influenced mine. Professor Shah has been incredibly generous with his time. The creativity and impact of his work has guided my research, and I appreciate all the feedback he has given me.

I would also like to thank Professor Fotini Christia for being a wonderful collaborator and mentor. Professor Christia has taught me the importance of interdisciplinary viewpoints, the reach of political science, and the process for high-quality research. Many of my mentors have been heavily involved with the Institute for Data, Systems, and Society (IDSS), and I would like to thank IDSS for providing a great environment, and for encouraging the use of theory and data to tackle societal challenges. I would like to thank Prof. Munther Dahleh, Prof. Anette Hosoi, Prof. John Tsitsiklis, Eric Lai, and the NIH RADx team for their support and help.

I would like to acknowledge and thank the National Defense Science and Engineering Graduate Fellowship Program and Professor Jadbabaie's Office of Naval Research Vannevar Bush Faculty Fellowship, "A New Paradigm for Analysis of Complex, Networked, Social and Engineering Systems", for providing financial support for

my work.

There are many students and staff at MIT that have made my experience meaningful and enjoyable. Amir, Arnab, and Manon have been great friends and collaborators. I thank them for all the coffee chats and research discussions. I would also like to thank Beth, Kim, Laura, and Marygrace for all their support, assistance, and friendship throughout my time at MIT. I am beyond grateful to my friends. Adam, Joey, Michael, Ryan, Sam, and Victor have been supportive, kind, and uplifting. I would like to thank Jordan for his unfailing belief in me.

Finally, I would like to thank my family. My mother, father, and brother have given me so much throughout my life. Thank you for your love and support. I would also like to thank the entire Wyman family. This thesis is dedicated to my family and to Noelle, who makes life brighter.

Contents

List of Figures

List of Tables

Chapter 0

Introduction

0.1 Background

Networks provide a powerful and unified framework to study complex systems. By abstracting systems down to entities and their connections, network models provide insight into the structure and dynamics of critical systems across multiple domains. Notable examples of networks include societies and social networks [1, 2], the financial and banking sector [3, 4], and energy grids [5, 6]. As a field, network science utilizes a common set of mathematical tools to study the formation, structure, function, and dynamics of these systems [7–10].

In its simplest form, a network is a collection of nodes and edges (Figure 0-1). This simple representation can model a wide variety of systems and capture a surprising amount of information. For example, societies are incredibly intricate, but at their core, they are composed of individuals and the relationships between them. As a result, societies can be modeled as social networks like the one seen in Figure 0-2. The social network representation provides insight into the structure and dynamics of the society. For instance, the tightly connected nodes in Figure 0-2 correspond to closely knit communities. The weak connections between some communities indicate parts of the society may become disconnected if individuals leave. The network representation can also answer consequential questions, such as how quickly would information or a disease spread through the society?

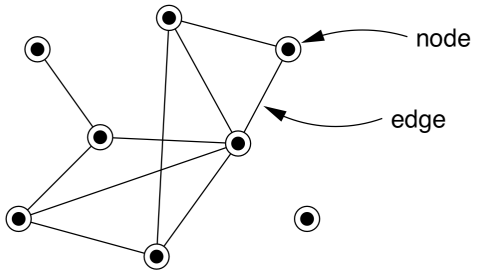


Figure 0-1: An example of a network with eight nodes and ten edges. Image from [7].

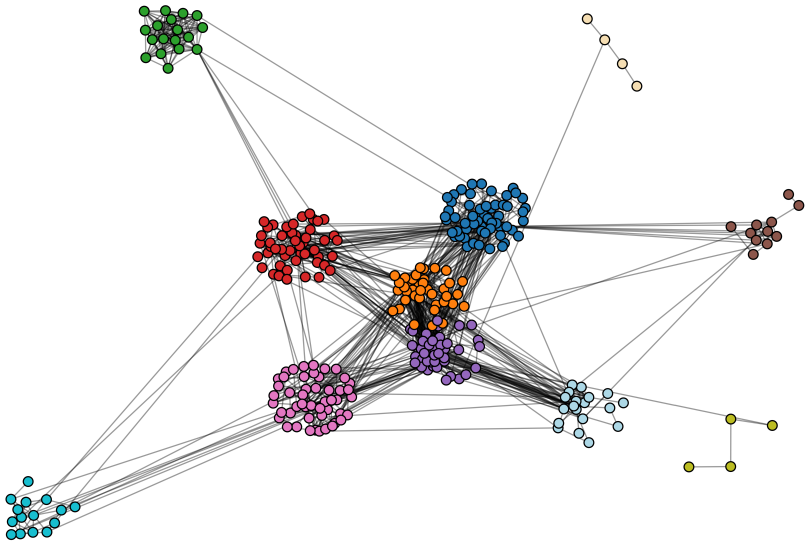


Figure 0-2: A social network of student interactions from the Technical University of Denmark. Nodes correspond to first-year students at the university and edges correspond to their physical interactions, recorded using Bluetooth-enabled smartphones. Node colors correspond to their community affiliation. Data provided by [11].

Network science, which studies complex systems using network models, is built on mathematical graph theory, a subfield of discrete mathematics. The study of graph theory began in 1736 with Leonhard Euler's solution to the Bridges of Königsberg problem [8]. The problem asked whether one could walk through the city of Königsberg by crossing each of its seven bridges once and only once. In network language, Euler modeled the bridges as edges and sections of the city as nodes. He then used the properties of the network to prove that such a walk was impossible.

Since the Bridges of Königsberg problem, graph theory has grown into a vibrant field. Theoreticians made notable advances in the 20th century with the introduction of random graphs, which integrated probability into network representations [12, 13]. Random graph models generate nodes and edges randomly according to predefined rules. Different probabilistic rules result in different network structures and properties. As a subfield of discrete math, graph theory has also benefited from advances in set theory [14], linear algebra [15], and control theory [16, 17].

The field of network science has exploded over the past two decades, driven by the large-scale generation of data and by advances in computing power. Extensive datasets have allowed researchers to study the structure and dynamics of real-world networked systems. Researchers have mapped the structure of the Internet and studied its resilience to connection issues [18]. Others have analyzed the billions of individuals on Facebook, their friendships, cliques, and communities [1]. Papers have analyzed Twitter and its role in spreading misinformation [2]. Research has studied the interconnection between banks and the potential for insolvency and illiquidity to spread through the financial system [4, 19]. In industry, network science ideas form the foundation of Google's search algorithm [20], Amazon's shipping optimization [21], Instagram's influencer advertising [22], and components of the Federal Reserve's stress test [23]. As the preceding examples demonstrate, a network science approach can address problems with major societal significance. By acknowledging the importance of intra-system connections, network models shed light on the structure and dynamics of critical settings.

Research within network science is divisible into three categories: 1) empirical

study of network structure, 2) development and study of network models, and 3) study of the dynamics of networks and processes on networks. First, empirical study of network structure focuses on specific properties of a given network. Popular properties include the importance of specific nodes (centrality) [3,24], the resilience of the network to the removal of edges and nodes (connectivity) [25], and how tightly knit sections of the network are (clustering and communities) [26, 27]. Second, network models define rules that generate networks. For example, we can model a network as having a fixed number of nodes and random edges that exist independently. Researchers study network models to understand their properties and how they compare to real-world systems [28, 29]. Third, a segment of research focuses on processes taking place on networks. Examples include network search and navigation [30], random walks [31], opinion dynamics [32], and diffusion processes [33].

Our work, which focuses on diffusion, falls within the third category. Diffusion through networked systems corresponds to numerous consequential processes such as information spread through societies, epidemics in populations, failure propagation in energy grids, and cascading defaults in banking networks. As a result, the emergence of diffusion has been studied and documented across several domains [2–5, 19, 34–38]. In this thesis, we study diffusion in social networks. We first study epidemic spread through societies and how network information can be used to contain infection. We then study information diffusion through populations and the structure and function of information cascades.

Epidemic spread is one of the original reasons, and is still a primary reason, for studying networks [7]. In an epidemic, a disease spreads from person to person through contact. Networks provide a natural way to model such spread. Individuals are represented by nodes and their physical connections are represented by edges. One or more nodes are infected at the start of the epidemic. These infected nodes can then infect their neighbors. Many network epidemic models have been introduced with different rules for infection status and infection passing [39, 40]. Research focuses on understanding how network structure affects the speed and extent of an infection [41,42]. In addition, researchers use network models to design approaches to slow and stop

epidemic spread [43, 44].

Information diffusion corresponds to the spread of information, news, rumors, or gossip through a social network. From a mathematical standpoint, information spread is often modeled using the same tools as epidemic spread. One or more individuals has or is "infected" with information at the start. The individuals then spread the information to their neighbors. Researchers have studied information diffusion in multiple consequential domains. Misinformation spread on social networks can affect elections [2]. News about violence in conflict regions can induce fleeing or more violence [45, 46]. Marketing campaigns leverage word-of-mouth sharing in social networks to spread products [47], vaccines [48], or financial programs [3].

We study diffusion in networks by applying and extending statistical inference to a network setting. Initial empirical work in network science often consists of exploratory data analysis and utilizes data science heuristics. We use statistical methods for disciplined analysis. Specifically, statistical inference, which focuses on estimation, testing, and uncertainty quantification, provides the mathematical tools to learn from data rigorously [49]. Inferential methods allow researchers to estimate key parameters and models and, crucially, add confidence levels to data-driven results. Coupled with identification, inference allows for the estimation of causal parameters, which quantify the effect and effectiveness of interventions and can directly inform policy decisions [50]. By applying and extending inference, our work contributes to the development and use of mathematical methods for data analysis in network science.

In our work, we utilize several aspects of inference to strengthen our results. We use accurate confidence intervals to quantify the uncertainty around estimates of effects and interventions. In addition, we include and discuss full distributions when deriving new estimators or quantities of interest. Distributions allow us to understand the variance of variables and provide confidence intervals. We also discuss and develop formal testing procedures. These procedures and their associated test statistics allow us to distinguish between meaningful results and background noise. Finally, we use inferential methods to add causal interpretation to our results. We leverage econometric methods to estimate the causal impact of certain events on other

events using only experimental data. Throughout our work, inference plays a key role in adding rigor to our results.

This thesis utilizes both theory and data in order to address several real-world challenges. On the theory side, we develop and apply ideas from network science, epidemic modeling, branching processes, econometrics, and statistics. On the data side, we use social network data generated by smartphones and traditional cellphones. We apply this theory and data to control epidemics, identify information diffusion, and understand conflict.

0.2 Summary of individual chapters

0.2.1 Network group testing

In the first chapter, we study epidemic spread and consider the problem of identifying infected individuals in a population of size N. We introduce an approach based on group testing that uses significantly fewer than N tests when infection prevalence is low. In its simplest form, group testing pools samples from individuals into groups for an initial stage of testing. If a pooled sample tests negative, all individuals that made up the pool are classified as negative for the disease. If a pooled sample tests positive, all individuals that made up the pool are retested individually to identify the infected members.

The most common form of group testing, Dorfman testing, groups individuals randomly in the first stage of testing [51]. However, as communicable diseases spread from individual to individual through underlying social networks, the position of each individual in the network affects their infection probability. As a result, we utilize network information to group individuals intelligently and improve the performance of group testing. Specifically, we group individuals by their community.

We first introduce an epidemic model based on branching processes and a network model based on a stochastic block model. We then analyze the performance of a network grouping approach and derive the number of tests needed to screen a

population. We prove the number of tests needed under network grouping is less than the number of tests needed under Dorfman testing. The extent of outperformance is driven by the strength of community structure in the network. When networks have strong community structure, network grouping achieves the lower bound for two-stage group testing procedures. When networks have no community structure, network grouping is equivalent to Dorfman testing. We also consider the performance of network grouping when we cannot perfectly identify community structure. We derive the number of tests needed under network grouping and imperfect community detection, and prove network grouping still outperforms Dorfman. Extending our analysis to general networks, we demonstrate network grouping outperforms Dorfman on any network that has positive modularity.[1]

We then analyze the performance of network grouping under imperfect tests. Imperfect tests, which result in false negatives and false positives, are a major consideration in practice. We derive the number of tests needed, the number of false negatives, and the number of false positives under network grouping. We prove that network grouping weakly dominates Dorfman testing across all metrics.

As a first empirical application, we consider the scenario of a university testing its student body for COVID-19. We use a social network dataset that captures physical interactions between first-year students at the Technical University of Denmark. After simulating epidemic processes on the network, we demonstrate network grouping requires significantly fewer tests than Dorfman testing to screen the population. As a second empirical example, we build a mobility network of the US and apply network grouping to screen the population of the country for COVID-19. We use mobility data provided by SafeGraph and epidemic data provided by the New York Times. We apply network grouping and demonstrate it significantly outperforms Dorfman and individual testing.

Our work demonstrates social network information can be used to improve group

[1]Modularity is a core metric in network science that measures the extent of community structure in a network [101, 102]. Given a community partition, modularity records the observed fraction of edges within communities minus the expected fraction of edges within communities. As a result, networks with community structure have positive modularities.

testing. We prove network grouping weakly dominates the most common form of group testing, Dorfman testing, across all metrics. The performance of network grouping depends on the strength of community structure in the network. Importantly, network grouping is simple for practitioners to implement. In practice, individuals should be grouped by family unit, social group, work group, or other community structure. Our work demonstrates this simple approach can significantly reduce the number of tests needed to keep populations healthy.

0.2.2 Performance of group testing under imperfect tests

In the second chapter, we continue our analysis of group testing and derive the performance of Dorfman testing under imperfect tests. As mentioned, Dorfman testing is the most common form of group testing in practice [52–55]. We derive the distribution of the number of tests needed, the number of false negatives, and the number of false positives.

The full distributions allow for the construction of confidence intervals and provide better guidance for medical practitioners. For example, consider a university testing its student population for COVID. Naive guidance would tell the university the number of tests it needs *on average* to screen the population. However, the university may need more tests in practice. Using confidence intervals, we can inform the university the number of tests it needs with high probability, which is much more relevant information. In our analysis, we recognize the flexibility available to medical practitioners and allow first and second stage false negative and false positive rates to differ. This modeling addition allows practitioners to use different tests in the first stage, when groups are being tested, and in the second stage, when individuals are being tested.

We model first-stage sensitivity as dependent on the number of samples in each group. Sensitivity is the fraction of infected individuals that correctly test positive, and is equal to one minus the false negative rate. Modeling first-stage sensitivity as a function of the number of samples accounts for viral-load dilution. As more samples are placed in each pool, any viral material present is diluted and false negative rates increase. In our work, we also derive and discuss optimal group sizes, approaches to

maximize the number of people tested, and the impact of infection prevalence.

To facilitate the use of group testing, we have built a dashboard that allows practitioners to analyze the performance of group testing under various parameters. The dashboard can be found at group-testing.herokuapp.com. The dashboard takes in various input parameters, including population size, infection prevalence, and first and second stage false negative and false positive rates, and returns the number of tests needed, the number of false negatives, and the number of false positives as a function of group size. The outputs can help practitioners design, understand, and implement various group testing programs.

Our work extends our understanding of the most common approach to group testing, Dorfman testing. We derive the performance of Dorfman testing under conditions faced by medical practitioners. Specifically, we derive the number of tests needed, the number of false positives, and the number of false negatives under group testing when tests are imperfect, tests have varying sensitivities and specificities, and samples are diluted. Our work provides a theoretical foundation for the group testing approaches used in practice; our derivations and discussion help practitioners design, understand, and implement group testing programs in order to efficiently identify infected individuals. By providing analytical results, we expand the understanding of group testing, its performance in medical clinics and testing centers, and its potential for large-scale surveillance testing of infectious diseases.

0.2.3 Tests for network cascades via branching processes

In the previous two chapters, we study epidemic spread. Epidemic spread in social networks behaves similarly to information diffusion. In information diffusion, information, news, or gossip spreads from individual to individual though a population, much like an infection. While information exchange in social networks is common, information cascades, in which a large number of individuals quickly contact each other, are rare. These cascade often correspond to consequential events such as the spreading of news following a violent event, the retweeting of viral fake news, or the spreading of gossip through a social clique. In this chapter, we focus on identifying

information cascades in social networks.

We introduce a statistical testing framework to identify information cascades in network data. In many empirical network science studies, diffusion processes are often described as cascades since they involves nodes contacting or "infecting" their neighboring nodes, who in turn infect their neighbors [56–59]. However, in many network settings, small scale diffusion regularly emerges during normal periods from normal behavior. Only a small number of large cascades occur, motivating the need to distinguish large, meaningful branch formation from the smaller, common branches formed during normal periods. Call detail records, which record calls between phone users, provide the motivating example for this chapter. Call detail records have been used to demonstrate the emergence of calling cascades after disruptive events [60–66]. However, as individuals make calls during normal periods as well, even when no event has occurred, call branches also form during normal periods.

We introduce a test statistic to distinguish between abnormally large branches, which we term cascades, and the common branches formed during normal periods. Our test statistic compares observed branch size to expected branch size under the null of normal periods. We define a semiparametric model of edge formation under the null. This model allows us to derive the expected size and variance of branches under the null using the machinery of branching processes [67–69]. The test statistic we introduce is semiparametric, consistent, and asymptotically distributed standard normal under the null. A formal statistic allows us to quantify the probability observed branches were formed during normal periods. Therefore, a rejection of the null indicates the observed branches are significantly large and correspond to cascades.

As an empirical application, we apply the test statistic to call detail records from Yemen. Our test statistic allows us to 1) add inference and significance results to observed branches, and 2) detect anomalous periods based on branch size. We find a significant calling cascade occurred after the Presidential Palace was bombed in 2011. The emergence of a cascade implies information regarding the bombing spread quickly and deeply through the underlying social network. In addition, we identify three periods with significantly large call branches originating in Sana'a, Yemen's capital,

during March 2011. The detected periods line up with key violent events during the 2011 Yemeni Revolution. Crucially, by adding inference to observed branch structures, our test statistic provides significance and confidence levels to our empirical findings.

0.2.4 The social network effects of drone strikes

In the fourth chapter, we study the social network effects of drone strikes, focusing on information and physical diffusion around strikes. Following the previous chapter, we analyze the emergence of information cascades around localized events. Here, we focus on the formation of calling cascades around drone strikes. Drone strikes have become a fixture of modern warfare, yet their effects and effectiveness remain unclear. Despite their prevalence, their covert nature and the often isolated nature of their targets have made strikes difficult to study quantitatively. We utilize a large dataset of call detail records to systematically study the reaction of civilians and communities to strikes.

We utilize a dataset of 12 billion call detail records (CDRs) to study the mobility and communication response to 74 U.S. drone strikes in Yemen between 2010 and 2012. As societies are intrinsically networked systems, we focus on the dynamics of the underlying social networks around these localized, violent events. Networks provide a powerful framework to study social interactions and structure as well as disruptions to the social fabric [2, 7, 25, 70, 71].

We find large calling cascades form after strikes, where branches of calls emanate out from the strike region. Over 95% of strikes are followed by cascades, with roughly one third exhibiting increased call volume through four levels of callers. Calling cascades allow information regarding strikes to diffuse quickly and deeply through the social network. Compared to non-strike periods, proximal individuals call their frequent and geographically close contacts more frequently. The shifts in calling patterns imply people call their friends, family, and neighbors after strikes. Notably, socially central individuals are called twice as often and proceed to spark large calling cascades. Our findings add evidence to the key role central individuals play in diffusing information.

We also study the mobility response to strikes. Physical mobility increases 27% on

strike days compared to the pre-strike mean and thousands of individuals flee their hometowns. While some return home quickly, a large number relocate permanently, highlighting a prolonged impact to communities. In addition, we find the social and physical network of the population explains where people choose to flee.

Our findings demonstrate drone strikes have a disruptive and widespread impact on civilian life. Furthermore, our results imply information, opinions, and emotions regarding strikes spread quickly through the population. The widespread impact is in contrast to the prevailing political and military position that strikes are surgical. As we discuss in the chapter, the disruptive impact of strikes has ethical, legal, and strategic implications.

Chapter 1

Network Group Testing

Abstract

We consider the problem of identifying infected individuals in a population of size N and introduce a group testing approach that uses significantly fewer than N tests when infection prevalence is low. The most common approach to group testing, Dorfman testing, groups individuals randomly. However, as communicable diseases spread from individual to individual through underlying social networks, our approach utilizes network structure to improve performance. We prove that network grouping, which groups individuals by community, weakly dominates Dorfman in terms of the expected number of tests used. When tests are imperfect, network grouping weakly dominates Dorfman in terms of the expected number of false positives and false negatives. Network grouping's outperformance is determined by the strength of community structure in the network. When networks have strong community structure, network grouping achieves the lower bound for two-stage testing procedures. Using social network data from multiple sources, we apply network grouping to screen populations for COVID-19. We demonstrate network grouping requires significantly fewer tests than Dorfman and individual testing. In contrast to many proposed group testing approaches, network grouping is simple for practitioners to implement. In practice, individuals can be grouped by family unit or social group.

1.1 Introduction

Group testing significantly improves testing capabilities for infectious diseases when resources are limited. The standard approach to identify infected individuals in a population of size N is to test all population members individually, which requires N tests. Group testing, in its simplest form, pools individual samples together into

groups of size n for an initial stage of testing. If a group tests negative, all individuals within the group are classified as negative for the disease. If a group tests positive, all individual samples from the group are individually retested to identify the infected members. To illustrate the power of group testing, consider the scenario where $N = 50$ and one individual is infected. If individuals are pooled into groups of size $n = 10$ for an initial stage of testing, one group will test positive and all 10 samples from the group will be retested. The group testing approach uses 15 tests compared to the 50 used under individual testing.

Group testing was introduced by [51] to screen for syphilis in the US military. Dorfman's insight was simple but powerful. As a result, group testing has been employed numerous times in the medical field for diseases including influenza, chlamydia, and malaria [52, 72, 73]. Within the US, group testing is commonly used in blood banks and infertility prevention programs, where large numbers of individuals are routinely tested [53, 74–76]. Group testing's efficient use of resources has made it a valuable technique in developing areas. Notably, group testing was used during the early stages of the HIV pandemic in Africa when polymerase chain reaction (PCR) test costs were high [77, 78]. By reducing testing costs and increasing access to diagnostic information, group testing plays an important role in increasing health equity.

Under Dorfman's approach, each individual's infection probability is treated as homogenous and individuals are placed into groups randomly, which is equivalent to ignoring any information regarding an individual's susceptibility to infection. However, as communicable diseases spread from individual to individual through underlying social networks, an individual's network location affects their infection probability. In this work, we utilize network information to pool individuals for group testing. Specifically, we group individuals by community as infections are more likely to spread between closely connected community members than between members of distinct communities.

In order to analyze the performance of a network grouping strategy, we first introduce a network model and epidemic model. We utilize a stochastic block model to generate the networks under consideration and the first stage of a branching process

epidemic model to generate infections. Our chosen models provide insight into the behavior of network grouping. With generative models in hand, we derive the number of tests used under network grouping. We prove the expected number of tests used under network grouping is upper bounded by Dorfman testing, which implies network grouping weakly dominates Dorfman. The outperformance of network grouping is driven by the strength of community structure in the network. In networks with strong community structure, network grouping achieves the lower bound for two-stage testing procedures. In networks with no structure, network grouping is equivalent to Dorfman testing.

Our analytical results demonstrate network grouping weakly dominates Dorfman testing across all metrics. We prove the number of tests used under network grouping is upper bounded by Dorfman even when imperfect community detection algorithms are employed. In addition, we analyze the performance of network grouping under imperfect tests, which can result in false positives and negatives. The expected number of false positives under network grouping is upper bounded by Dorfman testing and the expected number of false negatives under network grouping is equivalent to Dorfman. Extending our analysis to general networks, we demonstrate network grouping outperforms Dorfman on any network that has positive modularity.

As a first empirical application, we consider the scenario of a university testing its student body for COVID-19. We use a social network dataset that captures physical interactions between first-year students at the Technical University of Denmark. After simulating epidemic processes on the network, we demonstrate network grouping requires significantly fewer tests than Dorfman testing to screen the population. As a second empirical example, we build a mobility network of the US and apply network grouping to screen the population of the country for COVID-19. We use mobility data provided by SafeGraph and epidemic data provided by the New York Times. We apply network grouping and demonstrate it significantly outperforms Dorfman and individual testing.

Our work reinforces the benefit of group testing for communicable diseases, which is consequential for the current COVID-19 pandemic. Multiple labs have demonstrated

the efficacy of group testing for detecting the SARS-CoV-2 virus and several countries have implemented group testing to increase their testing capabilities [55, 79–81]. In the US, the authors' work on group testing has been used by the [82] to inform and implement group testing in schools and businesses [83]. As testing resources still remain constrained, we hope more institutions and governments will take advantage of the power of group testing [84, 85].

Our work contributes to recent literature on group testing with heterogenous probabilities and connects the fields of group testing and network science. Over the past few years, researchers have begun utilizing heterogenous infection probabilities to intelligently group individuals and further reduce the number of tests used under group testing [53, 76, 86]. These papers employ covariate information for each individual, such as demographics and clinical observations, to determine individuals with high probability of infection. In contrast, our work derives heterogenous probabilities from the underlying social network. Several previous papers have studied group testing on network structures using an information-theoretic approach [87, 88]. Our work takes an applied statistics approach, with the goal of providing practitioners with a group testing approach that is powerful, straightforward to understand, and simple to implement. Importantly, instead of viewing the underlying network as a constraint on the group testing problem, we use the information provided by the network to improve performance.

Since Dorfman's work in 1943, numerous group testing approaches with strong performance have been introduced [53, 76, 86–95]. However, the complexity of the proposed methods have limited their adoption in the medical field. As a result, Dorfman's original method of two-stage testing, in which individuals are grouped randomly for an initial stage of testing and samples from positive groups are then retested, remains the most common approach to group testing in practice [52–55]. Importantly, network grouping is simple for practitioners to implement. In practice, individuals should be grouped by family unit, social group, work group, or other community structure. Our work demonstrates this simple approach can significantly reduce the number of tests needed to keep populations healthy.

1.2 Setup

In this section, we describe Dorfman testing and the lower bound for two-stage group testing. Under two-stage testing, a population of size N is split into N/n groups of size n for an initial round of testing. Let G denote the number of positive groups after the initial stage. In the second stage of testing, all n samples from each positive group are retested individually. In total, $N/n + nG$ tests are used.

Under Dorfman testing, one individual is infected with probability one and the remaining $N - 1$ individuals are infected independently with probability v. Note, in Dorfman's paper, all N individuals are infected independently with probability v; we deviate slightly from his setup to ensure at least one individual is infected. The expected number of infected individuals is therefore $\mathrm{E}[I_D] = 1 + (N-1)v$. The expected number of tests used under Dorfman testing is

$$\mathrm{E}[T_D] = \frac{N}{n} + n\left[1 + \left(\frac{N}{n} - 1\right)v'\right] \tag{1.1}$$

where $v' = 1 - (1-v)^n$. The derivation of $\mathrm{E}[T_D]$ was provided by Dorfman and is provided in supplement 1.19.1 for completeness. When the infection prevalence v is low, Dorfman testing uses significantly fewer than N tests in expectation. As an example, consider the scenario where $N = 1000$ and $v = 0.05$ (5% infection prevalence). If we employ Dorfman testing and a group size of $n = 10$, only 507 tests are needed in expectation to test the entire population, a reduction of nearly 50% compared to the $N = 1000$ tests used under individual testing.

Given a population, a certain number of infected individuals, and a group size, the minimum number of tests under two-stage group testing is achieved by minimizing the number of positive groups G. G is minimized by perfect grouping, in which all infected individuals are pooled together into the minimum possible number of groups. The lower bound for two-stage testing procedures when $1 + (N-1)v$ individuals are infected is

$$T_{LB} = \frac{N}{n} + n \cdot \max\left(1, \frac{1 + (N-1)v}{n}\right) \tag{1.2}$$

The derivation is provided in 1.19.2. Revisiting our example, if $N = 1000$, $v = 0.05$, and $n = 10$, the minimum number of tests needed under two-stage group testing is 151. The lower bound is unattainable in most scenarios as we do not know which samples are infected a priori.

1.3 Model

In this work, we consider the population of N individuals to be embedded in a network, where each individual corresponds to a node and their physical interactions correspond to edges. In our framework, the network underlying the population is generated by a stochastic block model (SBM). Specifically, we consider an SBM with N nodes split into N/m communities of size m. Within each community of m nodes, edges exist between nodes independently with probability p. Edges exist between nodes in different communities independently with probability q, where $q \leq p$. As a result, nodes are more likely to be connected to other nodes in the same community than to nodes in other communities.

For our epidemic model, we consider the initial stage of a branching process model. Specifically, an epidemic starts with a single infected seed node, which is chosen at random from the population. The seed node infects each of its neighbors independently with probability α. The seed node has $m - 1$ possible neighbors within its community, each connected with probability p, and $N - m$ possible neighbors outside of its community, each connected with probability q. As a result, the expected number of infected individuals under this model, which will we use for network grouping, is $\mathrm{E}[I_{NG}] = 1 + (m-1)p\alpha + (N-m)q\alpha$.

The epidemic model describes the initial stage of an outbreak or, alternatively, a super-spreader event. We set α such that the expected number of infected individuals in the epidemic model is equal to the expected number of infected individuals in the Dorfman setting. Setting $\mathrm{E}[I_{NG}] = \mathrm{E}[I_D]$ and solving for α yields $\alpha = (N-1)v/[(m-1)p + (N-m)q]$. Figure 1-1a provides a visual example of an SBM and our epidemic model. For the remainder of this work, we assume the following.

Assumption 1. *Assume* $1 \leq n \leq N$, $1 < m < N$, $0 \leq q \leq p \leq 1$, *and* $v \in [0,1]$. *In addition, assume* $\alpha \in [0,1]$ *set such that* $\mathrm{E}[I_{NG}] = \mathrm{E}[I_D]$.

Assumption 1 states that pool and community sizes are less than the population size, and that the probability an edge exists between nodes in different communities is less than or equal to the probability an edge exists between nodes in the same community, which ensures the network has community structure as mentioned above. The condition on α ensures that the expected number of infected individuals in the epidemic model is equal to the expected number of infected individuals in the Dorfman setting, which allows us to compare the performance of network grouping to Dorfman testing.

1.4 Results

In this section, we introduce our main results on network grouping and its performance compared to Dorfman testing. Under network grouping, we group individuals by their community. In the simplest case, if communities have the same size as groups, $m = n$, each community is pooled into a unique group. If community size is divisible by group size, the m community members are pooled into m/n groups. If group size is divisible by community size, each group of size n consists of n/m communities. For example, if $m = 20$ and $n = 10$, each community is pooled into two groups and if $m = 5$ and $n = 10$, each group consists of two communities. When m not divisible by n and n not divisible by m, we keep communities intact as much as possible and remainder community members are pooled into the remaining groups.

The expected number of tests used under network grouping is

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n\left[1 + \left(\frac{m}{n} - 1\right)^{+} p' + \left(\frac{N}{n} - 1 - \left(\frac{m}{n} - 1\right)^{+}\right) q'\right] \tag{1.3}$$

where $p' = 1-(1-p\alpha)^n$, $q' = 1-(1-q\alpha)^n$, and $(x)^+ = \max(x, 0)$. The full distribution of the number of tests is provided in appendix 1.6.1 and derived in 1.19.3. The CDF and variance of the number of tests, which are useful for constructing confidence

intervals, are also provided in 1.6.1.

1.4.1 Performance of network grouping

With the number of tests used under network grouping, Dorfman testing, and the lower bound derived, we come to the main result of our work. The expected number of tests under network grouping is upper bounded by Dorfman testing and lower bounded by the two-stage testing lower bound.

Theorem 1. *Under the conditions of assumption 1,* $\mathrm{E}[T_{NG}]$ *is increasing in* q *and*

$$T_{LB} \leq \mathrm{E}[T_{NG}] \leq \mathrm{E}[T_D] \tag{1.4}$$

Proofs for this subsection are in 1.19.4 and 1.19.5. Theorem 1 states network grouping weakly dominates Dorfman testing in terms of the expected number of tests used. The performance of network grouping is determined by q, the probability an edge exists between nodes in different communities. When networks have strong community structure, network grouping significantly outperforms Dorfman testing. In fact, there are cases where network grouping achieves the lower bound, as seen in corollary 1.

Corollary 1. *Under the conditions of assumption 1, if* $q = 0$ *and* $n \geq m$*,*

$$\mathrm{E}[T_{NG}] = T_{LB} \tag{1.5}$$

Corollary 1 states there are cases where network grouping performs optimally. When $q = 0$, communities are disconnected from each other and all infected individuals will reside within the same community. When $n \geq m$, each group is large enough to capture each entire community and, as a result, all infected individuals will be grouped together. However, there are also scenarios where network grouping is equivalent to Dorfman testing, notably when $q = p$ as seen in corollary 2.

Corollary 2. *Under the conditions of assumption 1, if $q = p$,*

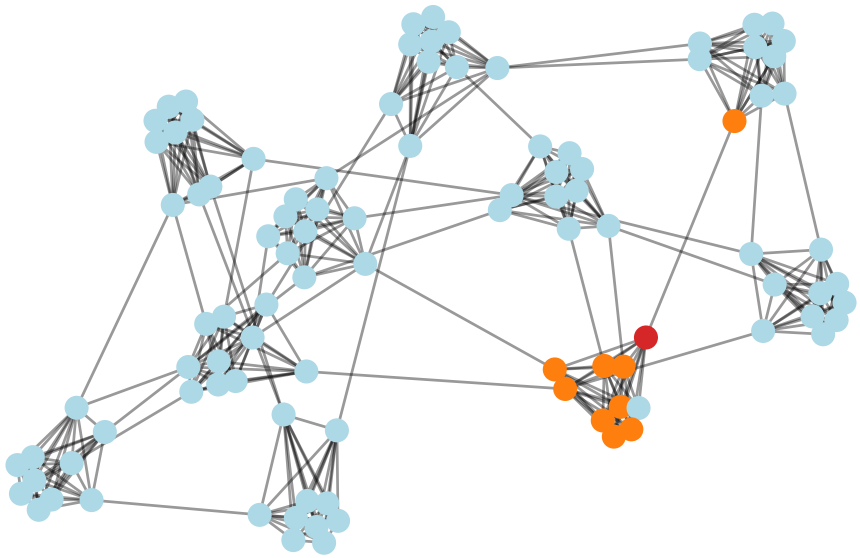
$$\mathrm{E}[T_{NG}] = \mathrm{E}[T_D] \tag{1.6}$$

Corollary 2 states network grouping is equivalent to Dorfman testing when the underlying network has no community structure. The reason is simple: since the network has no communities, all nodes have the same probability of being infected and the network provides no useful information for grouping. The corollary demonstrates Dorfman testing is a special case of network grouping that arises when the social network has no structure.

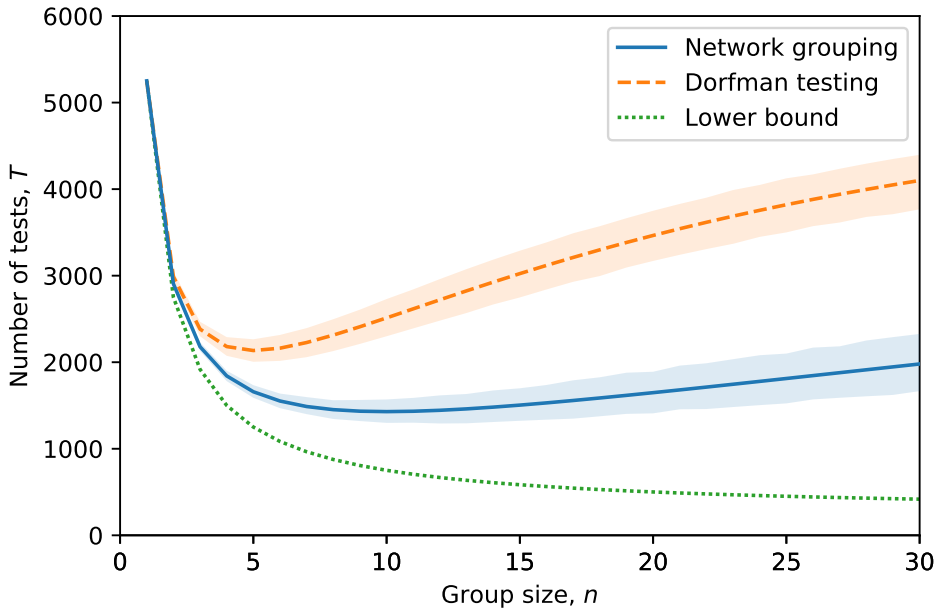
When $0 < q < p$, which is the setting one expects in practice, network grouping can significantly outperform Dorfman testing. Consider the scenario of a university testing its population for COVID-19 cases. Using MIT as our example, MIT has roughly $N = 5000$ undergraduates living in dorms of around $m = 400$ students. Assuming students within the same dorm are exposed to each other while students across dorms are rarely allowed to interact due to social distancing restrictions, we set $p = 0.8$ and $q = 0.02$. We set $v = 0.05$, which is in line with COVID infection prevalence in the US estimated by the COVID Tracking Project and [96]. Figure 1-1b displays the expected number of tests needed to test MIT's population under network grouping, Dorfman testing, and the two-stage lower bound as a function of group size. In this context, network grouping simply groups students by dorm. While Dorfman testing significantly improves upon individual testing, which uses $N = 5000$ tests, network grouping significantly improves upon Dorfman testing. When $n = 10$, network grouping uses only 1429 tests in expectation, a 43% reduction compared to Dorfman testing, which uses 2512 tests.

1.4.2 Imperfect community detection

Our previous results assume perfect community detection, where the precise community of each individual is known. In practice, the community structure of a real-world network may not be known a priori. In some cases, there may be natural communities



(a)



(b)

Figure 1-1: **(a)** Example of an SBM and our epidemic model with $N = 100$, $m = 10$, $p = 0.99$, $q = 0.01$, and $v = 0.05$. The infected seed node is colored red and its infected neighbors are colored orange. The community structure present in SBMs is clearly visible. **(b)** Comparison of network grouping, Dorfman testing, and the two-stage lower bound as a function of group size. The figure displays the expected number of tests used to test a population of size $N = 5000$ where $m = 400$, $p = 0.8$, $q = 0.02$, and $v = 0.05$. The shaded regions are 95% confidence intervals.

to consider, such as dorms, family units, or friend groups. In other cases, a community detection algorithm can be used to identify the underlying communities.

When precise community structure is not known, grouping individuals by community will rely on imperfect community detection methods. Let $\lambda \in [0, 1]$ denote the imperfection of the employed detection method. When $\lambda = 0$, community detection is perfect and we group individuals by their precise community, as in the previous subsection. When $\lambda = 1$, community detection is completely imperfect, meaning individuals are assigned to communities, and therefore to groups, uniformly at random. Any employed community detection method, such as heuristic grouping by family unit or systematic grouping by an algorithm like modularity maximization, will correspond to a λ between these two extremes.

In other words, λ captures the performance of the employed community detection method by measuring its error rate. When $\lambda = 0$, the community detection method has no error and perfectly detects the true communities of the network. When $\lambda = 1$, the community detection method has maximum error and is equivalent to randomly assigning nodes to communities. As above, any employed community detection method will perform between these two extremes and will correspond to a λ between 0 and 1.

The expected number of tests used under network grouping and imperfect community detection, $\mathrm{E}[T_{NG}^{\lambda}]$, is provided in 1.6.2 and derived in 1.19.6. The expected number of tests, $\mathrm{E}[T_{NG}^{\lambda}]$, is upper bounded by Dorfman testing and lower bounded by network group testing under perfect community detection.

Theorem 2. *Under the conditions of assumption 1 and imperfect community detection with* $\lambda \in [0, 1]$, $\mathrm{E}[T_{NG}^{\lambda}]$ *is increasing in* λ *and*

$$\mathrm{E}[T_{NG}] \leq \mathrm{E}[T_{NG}^{\lambda}] \leq \mathrm{E}[T_D] \tag{1.7}$$

If $\lambda = 0$, *then* $\mathrm{E}[T_{NG}^{\lambda}] = \mathrm{E}[T_{NG}]$ *and if* $\lambda = 1$, *then* $\mathrm{E}[T_{NG}^{\lambda}] = \mathrm{E}[T_D]$.

The proof of theorem 2 is in 1.19.7. Importantly, theorem 2 states network grouping under imperfect community detection still weakly dominates Dorfman testing in terms of the expected number of tests used. The expected number of tests used is increasing

in the imperfection of the community detection method, λ. When $\lambda = 0$, community detection is perfect and $\mathrm{E}[T_{NG}^{\lambda}]$ simplifies to our previous network grouping result. When $\lambda = 1$, community detection is completely imperfect, individuals are not grouped based on network information, and network grouping is equivalent to Dorfman testing. Interestingly, the theorem implies we recover Dorfman testing as a special case of network grouping when network information is not used to group individuals.

1.4.3 Imperfect tests

Imperfect tests are a major consideration in the medical field when testing for infectious diseases. False negatives miss infected individuals who may spread the infection further. False positives result in undue stress and unnecessary quarantines. In this subsection, we analyze the performance of network grouping under imperfect tests in terms of the number of tests used, false negatives, and false positives.

The performance of a diagnostic test is measured by two parameters: sensitivity and specificity. The sensitivity of a test is the fraction of infected individuals who correctly test positive. Therefore, sensitivity equals one minus the false negative rate. The specificity of a test is the fraction of non-infected individuals who correctly test negative. Therefore, specificity equals one minus the false positive rate. Relating the medical terms to statistical terminology, sensitivity is the power of the test and specificity is one minus the size of the test.

In our analysis, we allow the sensitivity and specificity of tests to differ between the first stage and second stage of testing in group testing procedures. This allows practitioners to use different tests for the first stage, when groups are tested, and the second stage, when individual samples are tested. We denote first-stage sensitivity as $s_{e_1,n}$, second-stage sensitivity as s_{e_2}, first-stage specificity as s_{p_1}, and second-stage specificity as s_{p_2}. We explicitly allow first-stage sensitivity, $s_{e_1,n}$, to depend on the group size n, since pooling samples dilutes the viral load of an infected sample and can therefore reduce test sensitivity. We leave $s_{e_1,n}$, s_{e_2}, s_{p_1}, and s_{p_2} as exogenous parameters for medical practitioners to input.

Under perfect tests, $s_{e_1,n}$, s_{e_2}, s_{p_1}, and s_{p_2} all equal 1, which results in no false

negatives or false positives. Under completely imperfect tests, $s_{e_1,n}$, s_{e_2}, s_{p_1}, and s_{p_2} all equal 0.5, which corresponds to tests marking samples positive or negative completely at random. Therefore, we include the following assumption in our analysis.

Assumption 2. *Assume* $s_{e_1,n}$, s_{e_2}, s_{p_1}, *and* s_{p_2} *are all* $\in [0.5, 1]$.

The expected number of tests, false positives, and false negatives under individual testing, Dorfman testing, the two-stage lower bound, and network grouping are provided in 1.6.3 and derived in 1.19.8–1.19.11. Under imperfect tests, the expected number of tests used under group testing changes since infected groups may incorrectly test negative and non-infected groups may incorrectly test positive. However, theorem 1 and corollaries 1 and 2 still hold under imperfect tests. Under imperfect tests, the expected number of tests used under network grouping is upper bounded by Dorfman testing and lower bounded by the two-stage lower bound. We formally state and prove this result in 1.19.12.

Under imperfect tests, the expected number of false positives under network grouping is upper bounded by the expected number under Dorfman testing, which in turn is upper bounded by the expected number under individual testing.

Theorem 3. *Under the conditions of assumptions 1 and 2 and imperfect tests,*

$$\mathrm{E}[FP_{NG}] \leq \mathrm{E}[FP_D] \leq \mathrm{E}[FP_I] \tag{1.8}$$

Both $\mathrm{E}[FP_{NG}]$ *and* $\mathrm{E}[FP_D]$ *are increasing in* n. *In addition,* $\mathrm{E}[FP_{NG}]$ *is increasing in* q *and when* $q = p$, $\mathrm{E}[FP_{NG}] = \mathrm{E}[FP_D]$.

The proof of theorem 3 is in 1.19.13. Theorem 3 states network grouping weakly dominates Dorfman testing and individual testing in terms of the expected number of false positives. Two-stage group testing approaches outperform individual testing since individuals are tested twice, making it less probable for a non-infected individual to test positive. Network grouping outperforms Dorfman testing since intelligent grouping results in fewer groups testing positive in the first stage. As a result, fewer individuals are tested in the second stage, which reduces the number of possible false

positives. The theorem also states false positives under both of the group testing approaches are increasing in n. As group size increases, the number of individuals who are tested in the second stage increases, increasing the opportunity for false positives. Lastly, the expected number of false positives under network grouping is increasing in q. When $q = p$, the network has no community structure and provides no useful information for network grouping. As a result, network grouping performs identically to Dorfman testing.

Under imperfect tests, the expected number of false negatives under network grouping is equal to the expected number under Dorfman testing and lower bounded by the expected number under individual testing.

Theorem 4. *Under the conditions of assumptions 1 and 2 and imperfect tests,*

$$\mathrm{E}[FN_I] \leq \mathrm{E}[FN_{NG}] = \mathrm{E}[FN_D] \tag{1.9}$$

The proof of theorem 4 is in 1.19.14. Theorem 4 states network grouping performs identically to Dorfman testing in terms of false negatives. Both of the two-stage group testing approaches underperform individual testing. The reasoning is simple: for an infected individual to test positive under a two-stage approach, they must correctly test positive twice, once in the first stage as part of a group and again in the second stage individually. As a result, infected individuals have a greater chance of being missed under two-stage group testing than under individual testing.

Recall our example of MIT screening its undergraduate population of $N = 5000$ for COVID-19. Under imperfect tests, the different testing approaches result in varying numbers of false positives and false negatives. For this example, we set first and second-stage specificity and second-stage sensitivity to 0.95. First-stage sensitivity for $n = 1$, $s_{e_1,n=1}$, is set to 0.95 and decreases linearly to $s_{e_1,n=30} = 0.90$ at $n = 30$. Figure 1-2a displays the expected number of false positives after testing the population using network grouping, Dorfman testing, and individual testing. Network grouping significantly outperforms the other two approaches. Using a group size of $n = 10$, network grouping produces only 42 false positives in expectation compared to 90

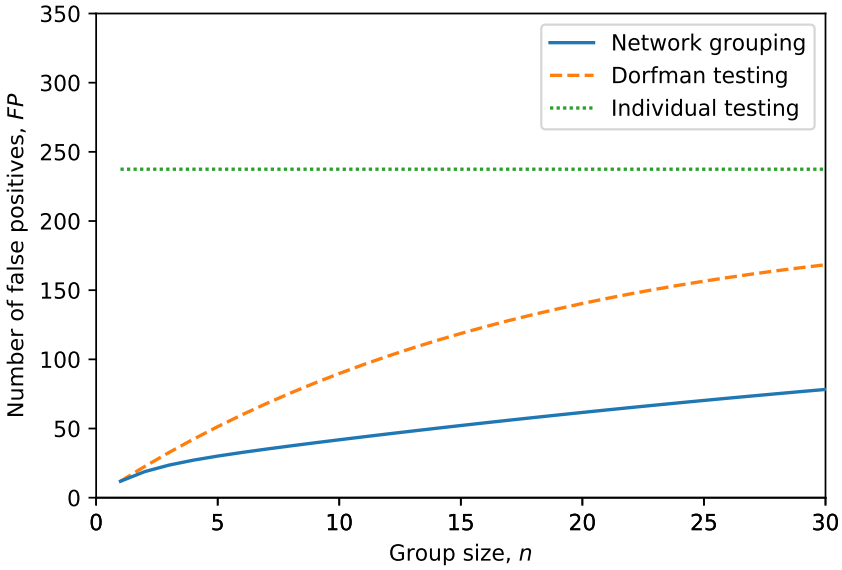
false positives under Dorfman testing and 237 under individual testing. In line with theorem 3, the number of false positives under the group testing approaches increases with group size.

Figure 1-2b displays the expected number of false negatives after testing MIT's population. As stated in theorem 4, the number of false negatives under network grouping equals the number under Dorfman testing. Both group testing approaches underperform individual testing. The number of false negatives increases with group size under network grouping and Dorfman testing as first-stage sensitivity $s_{e_1,n}$ decreases with n. Using tests with higher sensitivity or repeat testing of the population can reduce the number of false negatives under group testing. Since group testing often uses significantly fewer tests than individual testing, both approaches are economical in many scenarios.

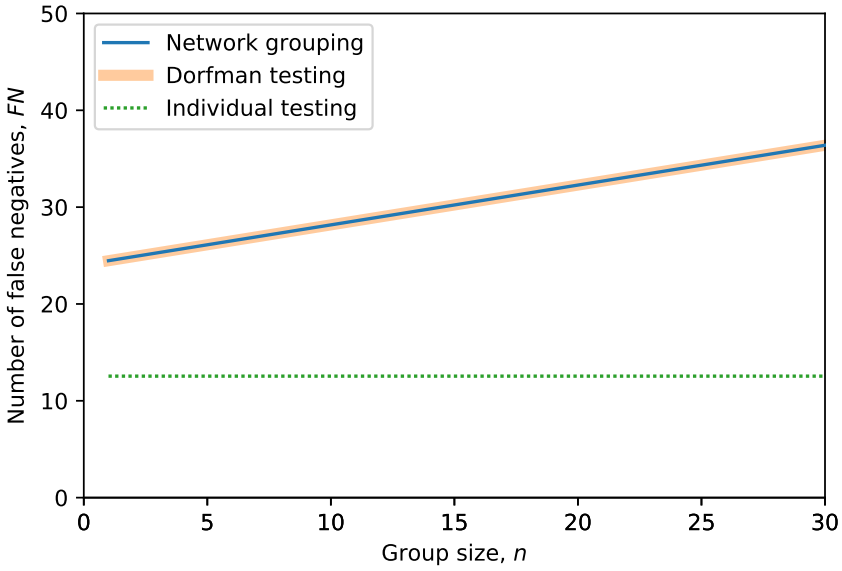
1.4.4 Application to a university social network

In this subsection, we apply network grouping to the scenario of a university testing its population for COVID-19 using social network data collected from a Danish university. In 2013, researchers distributed smartphones to first-year students at the Technical University of Denmark as part of the Copenhagen Networks Study [11]. The smartphones recorded physical interactions between students using Bluetooth. The dataset, which was published in 2019, is an ideal application for our methodology because it provides the social network of a university population where edges correspond to physical interactions. Using the dataset, we build the student social network, simulate an epidemic process, apply network grouping, and compare its performance to Dorfman testing.

We build the social network using one day's worth of Bluetooth interactions. After taking the largest connected component, the network has 310 nodes, which correspond to students, and 1503 edges, which correspond to their interactions. As we are dealing with a real-world network, we must identify the communities using a community detection algorithm. We use the Louvain method for community detection due to its popularity and strong performance [97]. The Louvain method is a heuristic approach



(a)



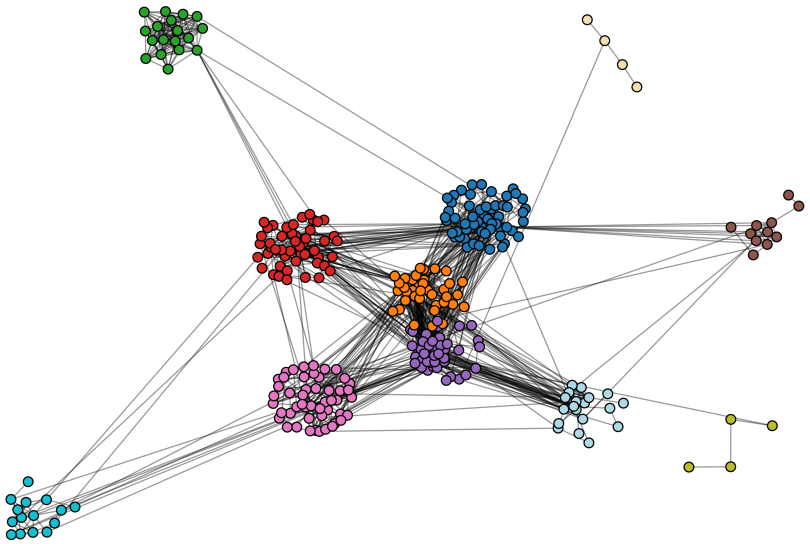
(b)

Figure 1-2: **(a)** Comparison of false positives under network grouping, Dorfman testing, and individual testing as a function of group size. The figure displays the expected number of false positives after testing a population of size $N = 5000$ where $m = 400$, $p = 0.8$, $q = 0.02$, $v = 0.05$, and $s_{p_1} = s_{p_2} = s_{e_2} = 0.95$. First-stage sensitivity for $n = 1$, $s_{e_1,n=1}$, equals 0.95 and decreases linearly to $s_{e_1,n=30} = 0.90$ at $n = 30$. **(b)** Comparison of false negatives under network grouping, Dorfman testing, and individual testing as a function of group size. The figure displays the expected number of false negatives after testing a population with the same parameters as in figure 1-2a.

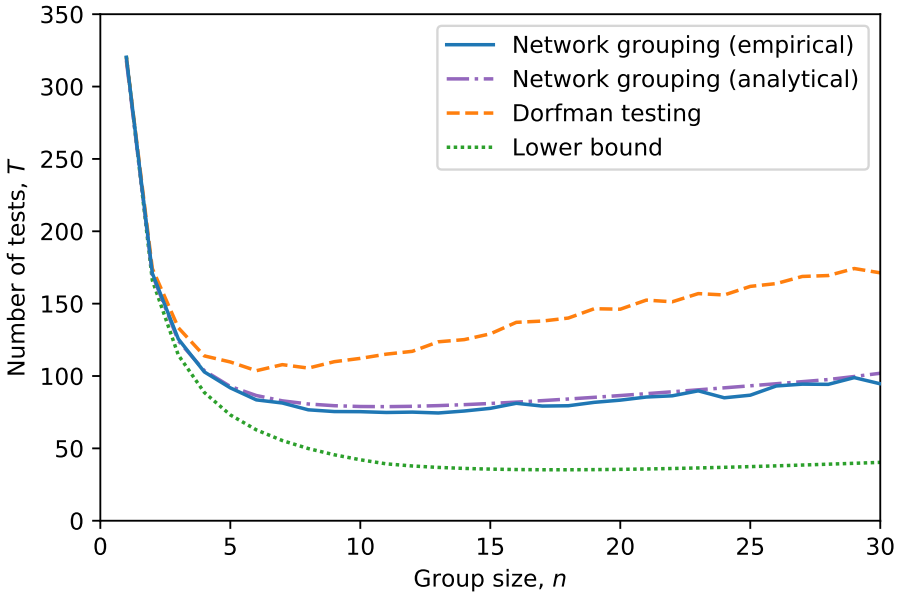
that aims to maximize the modularity of the resulting partition, which measures the density of edges within communities. Applying the Louvain method to our network results in 11 communities with an average size of 28 individuals.

The social network with nodes colored by their community is displayed in figure 1-3a. The network contains highly-connected communities that are weakly connected to other communities, such as the community in the upper left of the figure, as well as a few communities that are well connected to each other. We estimate p, the probability an edge exists between two nodes within the same community, as the number of edges that exist within communities divided by the total possible number of edges within communities. Similarly, we estimate q, the probability an edge exists between two nodes in different communities, as the number of edges that exist between communities divided by the total possible number of edges between communities. The resulting estimates are $\hat{p} = 0.18$ and $\hat{q} = 0.01$, indicating a sparse graph with strong community structure.

In order to analyze the performance of group testing, we simulate an epidemic process using the branching process model outlined in section 1.3. We set α, the probability of infection passing, to 0.95, which results in an estimated infection prevalence of 0.03. For each group size n, we run 1000 epidemic simulations, apply network grouping and Dorfman testing, and record the number of tests used under each approach. Network grouping groups individuals based on their detected communities while Dorfman testing groups individuals randomly. Figure 1-3b displays the average number of tests used as a function of group size. Network grouping strongly outperforms Dorfman testing. When $n = 10$, Dorfman testing uses 112 tests on average while network grouping uses 75 tests to screen the population of 310 students, a reduction of 33%. In addition, figure 1-3b demonstrates our analytical result for the number of tests used under network grouping, provided in equation 1.3, is a strong approximation for the number of tests used in a real network setting, even though the underlying network is not a stochastic block model.



(a)



(b)

Figure 1-3: **(a)** Social network of student interactions from the Technical University of Denmark. Nodes correspond to first-year students at the university and edges correspond to their physical interactions, recorded using Bluetooth-enabled smartphones. Nodes are colored by their community affiliation, which is determined using the Louvain algorithm. **(b)** Comparison of testing approaches applied to the Danish university social network. The figure displays the average number of tests used to screen the population of $N = 310$ as a function of group size.

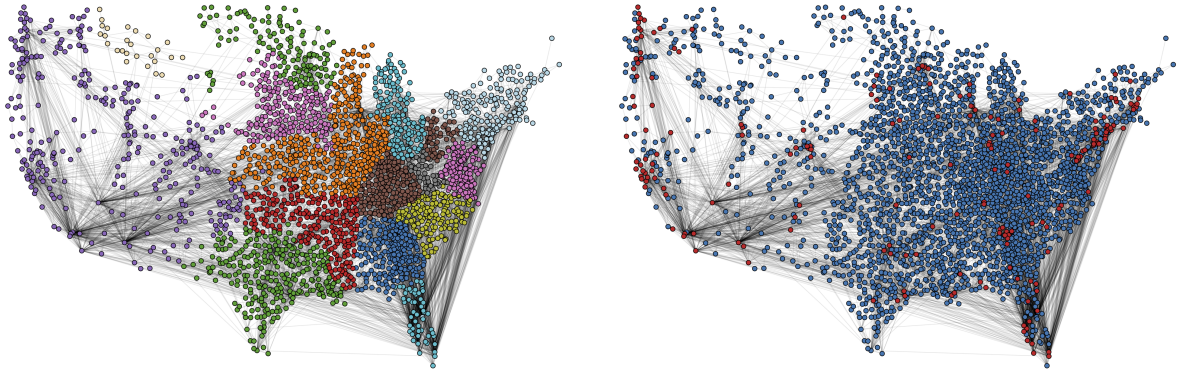
1.4.5 Application to a mobility network of the United States

In this subsection, we build a network of the United States (US) and apply network grouping to screen the population of the country for COVID-19. We use data provided by SafeGraph to build a mobility network where nodes correspond to counties and edges correspond to mobility between locations, measured using mobile devices. We then use COVID data from the New York Times to capture the spread of the pandemic through the country. By building a county-level network, we are able to use real epidemic spread data for our analysis. Finally, we apply two network grouping approaches to the data and compare their performance to Dorfman testing and individual testing.

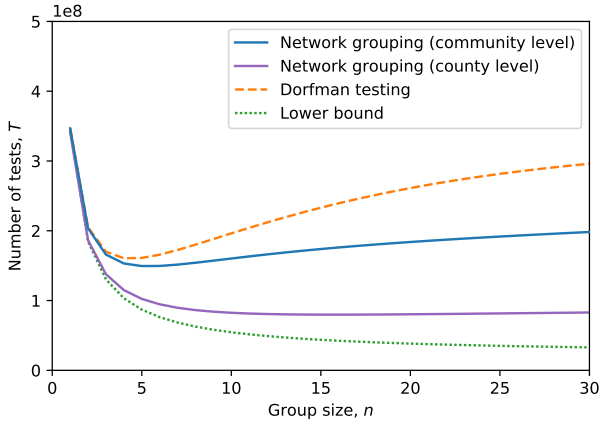
To begin, we build a mobility network using data provided by [98]. SafeGraph is a data provider that specializes in location and mobility data derived from mobile phone usage. Using their data, we build a network of the US where nodes correspond to counties and edges correspond to mobility between locations. Specifically, in our analysis, an edge exists between two nodes if more than 50 individuals traveled between the locations on March 2, 2020, the first Monday of the month. The resulting network has 2858 nodes and 14,473 edges. Similar to the previous subsection, we use the Louvain method to detect communities in the network. Applying the method to our network results in 18 communities with an average size of 159 nodes.

The mobility network with nodes colored by their community and positioned according to their geographical location is displayed in figure 1-4a. The outline of the US is clear from the network, even though no country borders are drawn. In addition, the Louvain method, which has no information about the geographical locations of the nodes, produces communities of nodes that are geographically clustered. Different communities in the network clearly correspond to different regions of the country. Similar to the previous subsection, we estimate p, the probability an edge exists between two nodes within the same community, and q, the probability an edge exists between two nodes in different communities. The resulting estimates are $\hat{p} = 0.04$ and $\hat{q} = 0.0006$, indicating a very sparse graph with strong community structure.

Instead of simulating an epidemic process on the network, we use real data from



(a) (b)



(c)

Figure 1-4: **(a)** Mobility network of the US where nodes correspond to counties and edges correspond to mobility between them, measured using mobile phones. Nodes are colored by their community affiliation, which is determined using the Louvain algorithm. **(b)** Mobility network of the US where nodes are colored red if the corresponding county had active COVID cases as of March 11, 2020. **(c)** Comparison of testing approaches applied to the mobility network. The figure displays the number of tests needed to screen the US population as a function of group size.

the COVID-19 pandemic. Using data provided by the [99], we record the number of active COVID cases in each county as of March 11, 2020, the day the [100] declared the COVID outbreak a pandemic. By March 11, 2020, 203 counties in the US had active cases, and infection prevalence in the population was around 6.8% [96]. Figure 1-4b displays the mobility network with infected counties colored red. By mid-March, the virus had spread to several areas of the country. Importantly, we see infections are localized in a few counties that are clustered together, which validates our modeling approach that models an infection as spreading from node to node through the network.

We apply two forms of network grouping to screen the entire US population of 330,000,000 individuals. We first group individuals by their community, as detected by the Louvain algorithm. Specifically, groups in the first stage of testing contain only individuals from the same community. For our second approach to group testing, we group individuals by their county, meaning groups in the first stage contain only individuals from the same county. The two approaches use different resolutions of network structure to form groups. We compare these two network grouping approaches to Dorfman testing, which groups individuals randomly.

Figure 1-4c displays the number of tests needed to screen the US population as a function of group size for the different approaches. Under individual testing, we need $N = 330,000,000$ tests to screen the population. Dorfman testing significantly improves upon individual testing, requiring 40% fewer tests when group size equals 10. Both network grouping approaches strongly outperform individual testing and Dorfman testing. When $n = 10$, network grouping by community uses 51% fewer tests than individual testing and 18% fewer tests than Dorfman. Network grouping by county uses 75% fewer tests than individual testing and 58% fewer tests than Dorfman. Network grouping by county performs strongly and approaches the lower bound for two-stage testing procedures. The approach performs well because infections are localized in a few counties, so grouping by county results in only a few positive groups.

This example uses real epidemic data to demonstrate the power of group testing, and network grouping specifically, for screening large populations. Reducing the number of tests needed to screen a population by 75% compared to the status quo

of individual testing results in significant cost and time savings. Fewer nasal swabs, reagents, PCR machines, medical professionals, and other resources are needed, making large-scale testing a viable option and a powerful tool to combat the pandemic. In addition, network grouping is intuitive and simple to implement. At the country level, the intuition is obvious: individuals should be grouped with nearby individuals. In this work, we have demonstrated this intuition holds for social networks for any size.

1.4.6 Extension to general networks

Our analysis uses a stochastic block model (SBM) to generate the networks under consideration. The SBM transparently models community structure and provides insight into the behavior of network grouping. However, as seen in our empirical example, SBMs are only an approximation for the structure of real-world networks.

Using SBMs, we have proven network grouping outperforms Dorfman testing under the assumption $q < p$, the probability edges exist between communities is less than the probability edges exist within communities. The $q < p$ condition guarantees the network has community structure. When $q = p$, the network has no community structure and network grouping is equivalent to Dorfman testing. Extending our analysis to general networks, we derive a condition on network structure that guarantees network grouping outperforms Dorfman testing.

The expected number of tests used under network grouping for general networks, $\mathrm{E}[T^*_{NG}]$, is provided in 1.6.4 and derived in 1.19.15. To ensure network grouping outperforms Dorfman, we provide a condition based on the modularity of the network. Modularity is a core metric in network science that measures the extent of community structure in a network [101, 102]. Given a community partition, modularity records the observed fraction of edges within communities minus the expected fraction of edges within communities. As a result, networks with no community structure have modularities near zero. Networks with community structure have positive modularities and networks with inverse community structure have negative modularity. Networks with inverse community structure have more edges between communities than within communities. The definition of modularity is provided below and is derived in 1.19.16.

Definition 1. *Given a community partition, the modularity Q of a network is*

$$Q = \frac{|int|}{|E|} - \frac{m-1}{N-1} \tag{1.10}$$

where $|int|$ is the number of edges within communities, $|E|$ is the number of edges in the network, m is the average community size, and N is the number of nodes.

We provide a lower bound on modularity that ensures network grouping outperforms Dorfman testing on a general network.

Theorem 5. *If*

$$Q \geq 1 - \frac{m-1}{N-1} - \frac{N(N-m)}{2|E|} \frac{\log\left(1 - \frac{|E|}{N(N-1)/2}\alpha\right)}{\log(1-\alpha)} \tag{1.11}$$

then $\mathrm{E}[T^*_{NG}] \leq \mathrm{E}[T_D]$. As $\alpha \to 0$, the condition in equation 1.11 goes to $Q \geq 0$.

The proof of theorem 5 is in 1.19.17. Theorem 5 states if a general network has strong community structure, then network grouping will use less tests than Dorfman in expectation. The condition simplifies nicely as α, the probability of infection passing, goes to zero. As $\alpha \to 0$, the condition simplifies to $Q \geq 0$. If the general network has community structure (measured by positive modularity), network grouping outperforms Dorfman testing.

Theorem 5 reinforces the main idea behind network group testing: when the underlying network has community structure, the structure can be used to intelligently group individuals, and network grouping will outperform the random grouping of Dorfman testing.

1.5 Conclusions

In this work, we have introduced the idea of using network information to improve group testing. Since communicable diseases spread from individual to individual through underlying social networks, grouping individuals by community can significantly reduce

the number of tests needed to screen a population. We demonstrate network grouping weakly dominates Dorfman testing, the most common group testing approach, in terms of the number of tests used, false positives, and false negatives. The outperformance of network grouping transparently depends on the strength of community structure in the network. We also establish a link between modularity, a core network science metric, and the outperformance of network group testing. Importantly, network grouping is simple to implement in practice, which is in contrast to many proposed group testing approaches. Practitioners can group individuals by family unit, friend group, office group, or other community structure. Our work aims to improve and increase diagnostic and surveillance testing, which are key methods for mitigating pandemics and advancing public health.

This work opens several fruitful areas for future research. Future work can analyze the performance of network grouping under different network structures, epidemic models, and community detection algorithms. In this paper, we implement network grouping by grouping individuals by community. Future research can utilize other network information and group individuals by clique, cluster, centrality, or some other network characteristic. In addition, covariate information, such as an individual's demographics and clinical results, can supplement and enhance network grouping. The network grouping approach can also be applied to one-stage group testing algorithms, which may produce fewer false negatives. Finally, network grouping can be applied to non-medical settings as in [103], such as communication networks, cybersecurity, and compressed sensing.

1.6 Appendix

1.6.1 Distribution of network grouping

The distribution of the number of tests used under network grouping is

$$T_{NG} \sim \frac{N}{n} + n\left[1 + Bin\left(\left(\frac{m}{n} - 1\right)^{+}, p'\right) + Bin\left(\frac{N}{n} - 1 - \left(\frac{m}{n} - 1\right)^{+}, q'\right)\right] \quad (1.12)$$

where $p' = 1 - (1 - p\alpha)^n$ and $q' = 1 - (1 - q\alpha)^n$. The equations in this subsection are derived in 1.19.3. The variance of T_{NG} is

$$\mathrm{Var}(T_{NG}) = n^2\left(\frac{m}{n} - 1\right)^{+} p'(1 - p') + n^2\left(\frac{N}{n} - 1 - \left(\frac{m}{n} - 1\right)^{+}\right) q'(1 - q') \quad (1.13)$$

where $p' = 1 - (1 - p\alpha)^n$ and $q' = 1 - (1 - q\alpha)^n$. The CDF of T_{NG}, which is useful for constructing confidence intervals and quantiles, is

$$\begin{aligned}\mathrm{P}(T_{NG} \le z) = \sum_{x=0}^{\lfloor k \rfloor}\sum_{y=0}^{x}\Bigg[&\binom{(m/n - 1)^{+}}{y}(p')^{y}(1 - p')^{(m/n-1)^{+}-y} \\ &\cdot \binom{N/n - 1 - (m/n - 1)^{+}}{x - y}(q')^{x-y}(1 - q')^{N/n-1-(m/n-1)^{+}-(x-y)}\Bigg]\end{aligned} \quad (1.14)$$

where $k = z/n - N/n^2 - 1$, $p' = 1 - (1 - p\alpha)^n$, and $q' = 1 - (1 - q\alpha)^n$.

1.6.2 Imperfect community detection

The expected number of tests used under network grouping and imperfect community detection is

$$\mathrm{E}[T_{NG}^{\lambda}] = \frac{N}{n} + n\left[1 + \left(\frac{m}{n} - 1\right)^{+} p'_{\lambda} + \left(\frac{N}{n} - 1 - \left(\frac{m}{n} - 1\right)^{+}\right) q'_{\lambda}\right] \quad (1.15)$$

$$p'_{\lambda} = 1 - \left[(1 - \lambda)(1 - p\alpha) + \lambda\left(\frac{m - 1}{N - 1}(1 - p\alpha) + \frac{N - m}{N - 1}(1 - q\alpha)\right)\right]^n$$

$$q'_{\lambda} = 1 - \left[(1 - \lambda)(1 - q\alpha) + \lambda\left(\frac{m - 1}{N - 1}(1 - p\alpha) + \frac{N - m}{N - 1}(1 - q\alpha)\right)\right]^n$$

where $(x)^+ = \max(x, 0)$. The derivation of equation 1.15 is provided in 1.19.6.

1.6.3 Imperfect tests

Individual testing, Dorfman testing, and the two-stage lower bound Individual testing uses N tests to screen a population of size N. The expected number of false negatives and false positives under individual testing, FN_I and FP_I, are given below.

$$\mathrm{E}[FN_I] = (1 + (N-1)v)(1 - s_{e_2}) \tag{1.16}$$

$$\mathrm{E}[FP_I] = (N-1)(1-v)(1 - s_{p_2}) \tag{1.17}$$

The derivations of equations 1.16 and 1.17 are in 1.19.8.

The expected number of tests used under Dorfman testing and imperfect tests is

$$\mathrm{E}[T_D] = \frac{N}{n} + n\left[s_{e_1,n} + \left(\frac{N}{n} - 1\right) v'_n\right] \tag{1.18}$$

where $v'_n = s_{e_1,n}(1 - (1-v)^n) + (1 - s_{p_1})(1-v)^n$. The derivation of Dorfman testing results under imperfect tests is in 1.19.9. The expected number of false negatives and false positives under Dorfman testing and imperfect tests are

$$\mathrm{E}[FN_D] = (1 + (N-1)v)(1 - s_{e_1,n}s_{e_2}) \tag{1.19}$$

$$\mathrm{E}[FP_D] = (n-1)(1-v)(1-s_{p_2})s_{e_1,n} + (N-n)(1-v)(1-s_{p_2})v'_{n-1} \tag{1.20}$$

where $v'_{n-1} = s_{e_1,n}(1 - (1-v)^{n-1}) + (1 - s_{p_1})(1-v)^{n-1}$.

The two-stage group testing lower bound under imperfect tests is

$$\begin{aligned}\mathrm{E}[T_{LB}] = \frac{N}{n} + n\Bigg[&s_{e_1,n} \cdot \max\left(1, \frac{1 + (N-1)v}{n}\right) \\ &+ (1 - s_{p_1})\left(\frac{N}{n} - \max\left(1, \frac{1 + (N-1)v}{n}\right)\right)\Bigg]\end{aligned} \tag{1.21}$$

The derivation of equation 1.21 is in 1.19.10.

Network grouping The expected number of tests used under network grouping and imperfect tests is

$$\begin{aligned}
\mathrm{E}[T_{NG}] &= \frac{N}{n} + n\left[s_{e_1,n} + \left(\frac{m}{n} - 1\right)^{+} p'_n + \left(\frac{N}{n} - 1 - \left(\frac{m}{n} - 1\right)^{+}\right) q'_n\right] \qquad (1.22)\\
p'_n &= s_{e_1,n}(1 - (1 - p\alpha)^n) + (1 - s_{p_1})(1 - p\alpha)^n\\
q'_n &= s_{e_1,n}(1 - (1 - q\alpha)^n) + (1 - s_{p_1})(1 - q\alpha)^n
\end{aligned}$$

where $(x)^+ = \max(x, 0)$. The derivation of network grouping results under imperfect tests is in 1.19.11. We note $\mathrm{E}[T_{NG}]$ under imperfect tests can be either higher or lower than $\mathrm{E}[T_{NG}]$ under perfect tests, depending on $s_{e_1,n}$ and s_{p_1}.

The expected number of false negatives under network grouping is

$$\mathrm{E}[FN_{NG}] = (1 + (m-1)p\alpha + (N-m)q\alpha)(1 - s_{e_1,n}s_{e_2}) \qquad (1.23)$$

The expected number of false positives under network grouping is

$$\begin{aligned}
\mathrm{E}[FP_{NG}] = {} & (\min(n,m) - 1)(1 - p\alpha)(1 - s_{p_2})s_{e_1,n} +\\
& (n - \min(n,m))(1 - q\alpha)(1 - s_{p_2})s_{e_1,n} +\\
& (m - n)^{+}(1 - p\alpha)(1 - s_{p_2})p'_{n-1} +\\
& (N - \max(n,m))(1 - q\alpha)(1 - s_{p_2})q'_{n-1} \qquad (1.24)\\
p'_{n-1} = {} & s_{e_1,n}(1 - (1 - p\alpha)^{n-1}) + (1 - s_{p_1})(1 - p\alpha)^{n-1}\\
q'_{n-1} = {} & s_{e_1,n}(1 - (1 - q\alpha)^{n-1}) + (1 - s_{p_1})(1 - q\alpha)^{n-1}
\end{aligned}$$

where $(x)^+ = \max(x, 0)$. Note, under our notation for p'_{n-1} and q'_{n-1}, the sensitivity $s_{e_1,n}$ depends on the original group size n.

The overall sensitivity of network grouping, $s_{e_{NG}}$, is one minus the overall false negative rate. Likewise, the overall specificity of network grouping, $s_{p_{NG}}$, is one minus the overall false positive rate. Therefore,

$$s_{e_{NG}} = 1 - \frac{\mathrm{E}[FN_{NG}]}{\mathrm{E}[I_{NG}]} \qquad (1.25)$$

$$s_{p_{NG}} = 1 - \frac{\mathrm{E}[FP_{NG}]}{N - \mathrm{E}[I_{NG}]} \tag{1.26}$$

where $\mathrm{E}[I_{NG}] = 1 + (m-1)p\alpha + (N-m)q\alpha$.

1.6.4 Extension to general networks

The expected number of tests used under network grouping for general networks is

$$\mathrm{E}[T^*_{NG}] = \frac{N}{n} + n\left[1 + \frac{1}{N}\sum_{i\in\mathcal{N}}\sum_{g\in\mathcal{G}\setminus g_i}(1-(1-\alpha)^{n_{g,i}})\right] \tag{1.27}$$

where $\mathcal{N}$ is the set of nodes, $\mathcal{G}$ is the set of groups, g_i is the group that contains node i, and $n_{g,i}$ is the number of nodes in group g connected to node i. The derivation of equation 1.27 is provided in 1.19.15. Note, given a network, the computation of $\mathrm{E}[T^*_{NG}]$ does not rely on any unknown information.

1.19 Supplementary materials

1.19.1 Derivation of Dorfman testing

Under Dorfman testing, a population of size N is split into N/n groups of size n for an initial round of testing. Let G denote the number of positive groups after the initial round. In the second round of testing, all n samples from each positive group are retested individually. In total, $N/n + nG$ tests are used. G is a random variable. Of the N/n groups, one is positive with probability one as there is at least one infected individual. The remaining $N/n - 1$ groups are positive independently with some probability v'. As a result, G is distributed $1 + Bin(N/n - 1, v')$.

The probability v' is derived as follows. Each of the remaining $N - n$ individuals (that are not in the first group) are infected with probability v and not infected with probability $1 - v$. The probability that all n individuals in a group are not infected is $(1 - v)^n$. The probability that at least one individual in the group is infected, and therefore the group tests positive, is $v' = 1 - (1 - v)^n$. Putting everything together, the number of tests used under Dorfman testing is distributed

$$T_D \sim \frac{N}{n} + n\left[1 + Bin\left(\frac{N}{n} - 1, v'\right)\right] \tag{1.28}$$

Taking the expectation of T_D provides $\mathrm{E}[T_D]$ as displayed in equation 1.1.

1.19.2 Derivation of the two-stage lower bound

Under two-stage group testing, a population of size N is split into N/n groups of size n for an initial round of testing. Let G denote the number of positive groups after the initial round. In the second round of testing, all n samples from each positive group are retested individually. In total, $N/n + nG$ tests are used. Given a population, a certain number of infected individuals, and a group size, the minimum number of tests is achieved by minimizing the number of positive groups G. G is minimized by perfect grouping, in which all infected individuals are pooled together into the minimum possible number of groups.

When $1 + (N-1)v$ individuals are infected, the minimum number of positive groups of size n is $[1 + (N-1)v]/n$. For example, if 20 individuals are infected and $n = 10$, the minimum number of positive groups is two. When the number of infected individuals is less than or equal to n, the minimum number of positive groups will be one. Note, there is always at least one infected individual in our framework. Putting everything together, the lower bound for the number of tests needed under two-stage group testing is

$$T_{LB} = \frac{N}{n} + n \cdot \max\left(1, \frac{1 + (N-1)v}{n}\right) \tag{1.29}$$

1.19.3 Derivation of network grouping

Under two-stage group testing, a population of size N is split into N/n groups of size n for an initial round of testing. Let G denote the number of positive groups after the initial round. In the second round of testing, all n samples from each positive group are retested individually. In total, $N/n + nG$ tests are used. G is a random variable.

The network contains N/m communities of size m. We consider cases where n divisible by m or m divisible by n. We consider cases where n not divisible by m and m not divisible by n in 1.19.18. First consider the case where $m \leq n$. Since we group individuals by community (as described at the beginning of section 1.4), the infected seed node and its $m-1$ community members will be contained in the same group. This group will be positive with probability one. The remaining $N/n - 1$ groups each contain n nodes that belong to different communities than the seed node. As a result, each node in the remaining $N/n - 1$ groups is not infected with probability $1 - q\alpha$, as they are only infected if they are both connected to the seed, with probability q, and infected by the seed, with probability α. The probability all n nodes within a group are not infected is $(1 - q\alpha)^n$. The probability that at least one individual in a group is infected, and therefore the group tests positive, is $q' = 1 - (1 - q\alpha)^n$. In summary, the remaining $N/n - 1$ groups are positive independently with probability q'. Putting everything together, the distribution of the number of tests used under

network grouping when $m \leq n$ is

$$T_{NG} \sim \frac{N}{n} + n\left[1 + Bin\left(\frac{N}{n} - 1, q'\right)\right] \tag{1.30}$$

where $q' = 1 - (1 - q\alpha)^n$.

Now consider the case where $m > n$. As we group individuals by community, there will be one group that contains the infected seed node and $n - 1$ of its community members. This group will be positive with probability one. The remaining $m - n$ nodes from the seed node's community will be pooled into $(m - n)/n = {}^{m}/_{n} - 1$ other groups. Each node in these groups will be infected with probability $1 - p\alpha$, as they are only infected if they are both connected to the seed, with probability p, and infected by the seed, with probability α. Following the same logic as the $m \leq n$ case, each of these ${}^{m}/_{n} - 1$ groups is positive independently with probability $p' = 1 - (1 - p\alpha)^n$. After accounting for the infected seed's group and the other ${}^{m}/_{n} - 1$ groups, ${}^{N}/_{n} - {}^{m}/_{n}$ groups still remain. Each of the n nodes in these groups are members of different communities than the seed node. Therefore, each of the ${}^{N}/_{n} - {}^{m}/_{n}$ groups is positive independently with probability $q' = 1 - (1 - q\alpha)^n$. Putting everything together, the distribution of the number of tests used under network grouping when $m > n$ is

$$T_{NG} \sim \frac{N}{n} + n\left[1 + Bin\left(\frac{m}{n} - 1, p'\right) + Bin\left(\frac{N}{n} - \frac{m}{n}, q'\right)\right] \tag{1.31}$$

where $p' = 1 - (1 - p\alpha)^n$ and $q' = 1 - (1 - q\alpha)^n$.

The two cases, $m \leq n$ and $m > n$, can be easily combined. Defining $(x)^+ = \max(x, 0)$, we have $({}^{m}/_{n} - 1)^+ = 0$ when $m \leq n$. Therefore, we can write the distribution of the number of tests used under network grouping in the general case as

$$T_{NG} \sim \frac{N}{n} + n\left[1 + Bin\left(\left(\frac{m}{n} - 1\right)^+, p'\right) + Bin\left(\frac{N}{n} - 1 - \left(\frac{m}{n} - 1\right)^+, q'\right)\right] \tag{1.32}$$

where $p' = 1 - (1 - p\alpha)^n$ and $q' = 1 - (1 - q\alpha)^n$. Taking the expectation of T_{NG} in equation 1.32 provides $\mathrm{E}[T_{NG}]$ as displayed in equation 1.3.

Variance and CDF of network grouping The distribution of the number of tests used under network grouping is the convolution (sum) of independent binomial distributions. The variance of the distribution is straightforward to derive as the variance of a sum of independent random variables is the sum of the variances. The variance of T_{NG} is therefore

$$\text{Var}(T_{NG}) = n^2 \left(\frac{m}{n} - 1\right)^+ p'(1-p') + n^2 \left(\frac{N}{n} - 1 - \left(\frac{m}{n} - 1\right)^+\right) q'(1-q') \quad (1.33)$$

where $p' = 1-(1-p\alpha)^n$ and $q' = 1-(1-q\alpha)^n$.

The CDF of T_{NG}, which is useful for constructing confidence intervals and quantiles, also follows from equation 1.32. The CDF is shown below, followed by its derivation.

$$\begin{aligned} \text{P}(T_{NG} \leq z) = \sum_{x=0}^{\lfloor k \rfloor} \sum_{y=0}^{x} \Bigg[& \binom{(m/n - 1)^+}{y} (p')^y (1-p')^{(m/n-1)^+ - y} \\ & \cdot \binom{N/n - 1 - (m/n - 1)^+}{x-y} (q')^{x-y} (1-q')^{N/n - 1 - \left(m/n - 1^+\right) - (x-y)} \Bigg] \end{aligned} \quad (1.34)$$

where $k = z/n - N/n^2 - 1$, $p' = 1-(1-p\alpha)^n$, and $q' = 1-(1-q\alpha)^n$.

To derive equation 1.34 from equation 1.32, we note

$$\text{P}\left(T_{NG} \leq z\right) = \text{P}\left(Bin(g_1, p') + Bin(g_2, q') \leq z/n - N/n^2 - 1\right) \quad (1.35)$$

where $g_1 = (m/n - 1)^+$ and $g_2 = N/n - 1 - (m/n - 1)^+$. We define $k = z/n - N/n^2 - 1$. Note, $Bin(g_1, p')$ and $Bin(g_2, q')$ are independent binomials. Their sum equals some value x if one equals $y \leq x$ and the other equals $x - y$. The probability their sum is less than or equal to some value k is the sum of probabilities that their sum equals all values from 0 to k. As a result,

$$\text{P}\left(Bin(g_1, p') + Bin(g_2, q') \leq k\right) = \sum_{x=0}^{\lfloor k \rfloor} \sum_{y=0}^{x} \text{P}\left(Bin(g_1, p') = y\right) \cdot \text{P}\left(Bin(g_2, q') = x - y\right) \quad (1.36)$$

Plugging in the standard binomial PMFs provides the CDF of T_{NG} as shown in equation 1.34.

1.19.4 Proof of theorem 1

Upper bound To prove $\mathrm{E}[T_{NG}] \leq \mathrm{E}[T_D]$, we prove $\mathrm{E}[T_{NG}]$ is increasing in q under assumption 1 and equals $\mathrm{E}[T_D]$ when q is set to its maximum value under assumption 1, $q = p$. This also proves corollary 2. To prove $\mathrm{E}[T_{NG}]$ is increasing in q, we consider the cases where $m > n$ and $m \leq n$ separately.

Case 1: We first consider the case where $m > n$. When $m > n$, $\mathrm{E}[T_{NG}]$ is given by

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n\left[1 + \left(\frac{m}{n} - 1\right)p' + \left(\frac{N}{n} - \frac{m}{n}\right)q'\right] \tag{1.37}$$

where $p' = 1 - (1 - p\alpha)^n$ and $q' = 1 - (1 - q\alpha)^n$. Under assumption 1, $\alpha = (N-1)v/[(m-1)p + (N-m)q]$. Taking the derivative of equation 1.37 with respect to q and simplifying yields

$$\frac{\partial \mathrm{E}[T_{NG}]}{\partial q} = \frac{nv(N-1)(N-m)\left[p(m-1)\left(1-q\alpha\right)^{n-1} - p(m-n)\left(1-p\alpha\right)^{n-1}\right]}{[(m-1)p + (N-m)q]^2} \tag{1.38}$$

Demonstrating equation 1.38 is nonnegative proves equation 1.37 is increasing in q. The denominator is nonnegative due to the square and n, v, $N-1$, and $N-m$ are nonnegative by assumption 1. Examining the bracket term in the numerator, we note $p(m-1) \geq p(m-n)$ as $n \geq 1$ and $(1-q\alpha)^{n-1} \geq (1-p\alpha)^{n-1}$ as $p \geq q$. Note, both $1 - q\alpha$ and $1 - p\alpha$ are probabilities between 0 and 1 as p and α are between 0 and 1. As a result, the bracket term is nonnegative and the entirety of equation 1.38 is nonnegative.

Case 2: We now consider the case where $m \leq n$. When $m \leq n$, $\mathrm{E}[T_{NG}]$ is given by

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n\left[1 + \left(\frac{N}{n} - 1\right)q'\right] \tag{1.39}$$

where $q' = 1 - (1 - q\alpha)^n$. Again, $\alpha = (N-1)v/[(m-1)p + (N-m)q]$. Taking the

derivative of equation 1.39 with respect to q and simplifying yields

$$\frac{\partial \, \mathrm{E}[T_{NG}]}{\partial q} = \frac{nvp(N-1)(N-n)(m-1)\left(1-q\alpha\right)^{n}}{\left[(m-1)p+(N-m)q\right]\left[(m-1)p+(N-m)q+q(1-N)v\right]} \tag{1.40}$$

All terms in the numerator and the first bracket term in the denominator are nonnegative by assumption 1. The second bracket term in the denominator is nonnegative if

$$(m-1)p+(N-m)q+q(1-N)v \geq 0 \tag{1.41}$$

Rearranging equation 1.41 yields

$$\frac{q(N-1)v}{(m-1)p+(N-m)q} \leq 1 \tag{1.42}$$

$$q\alpha \leq 1 \tag{1.43}$$

which is true by assumption 1. As a result, equation 1.40 is nonnegative and equation 1.39 is increasing in q.

Final step: We have shown $\mathrm{E}[T_{NG}]$ is increasing in q for $m > n$ and $m \leq n$. Setting q to its maximum value under assumption 1, $q = p$, we have $p' = q'$ as $1-(1-p\alpha)^{n} = 1-(1-q\alpha)^{n}$. In addition, α simplifies to v/p and $p\alpha = v$. Therefore, $p' = q' = v'$ where $v' = 1-(1-v)^{n}$. $\mathrm{E}[T_{NG}]$ in the general case simplifies to

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n\left[1+\left(\frac{m}{n}-1\right)^{+}p' + \left(\frac{N}{n}-1-\left(\frac{m}{n}-1\right)^{+}\right)q'\right] \tag{1.44}$$

$$= \frac{N}{n} + n\left[1+\left(\frac{N}{n}-1\right)v'\right] \tag{1.45}$$

and we have $\mathrm{E}[T_{NG}] = \mathrm{E}[T_D]$, completing the upper bound portion of the proof.

Lower bound To prove $\mathrm{E}[T_{NG}] \geq T_{LB}$, we prove $\mathrm{E}[T_{NG}] - T_{LB} \geq 0$ for the three cases of 1) group size larger than (or equal to) the expected number of infected individuals, $n \geq 1+(N-1)v$, 2) group size less than infected individuals and less than community size, $n < 1+(N-1)v$ and $n < m$, and 3) group size less than infected

individuals and greater than (or equal to) community size, $n < 1 + (N-1)v$ and $n \geq m$.

Case 1: When group size is larger than or equal to the expected number of infected individuals, $n \geq 1 + (N-1)v$, the lower bound in equation 1.2 simplifies to $N/n + n$. Therefore,

$$\mathrm{E}[T_{NG}] - T_{LB} = n\left(\frac{m}{n} - 1\right)^{+} p' + n\left(\frac{N}{n} - 1 - \left(\frac{m}{n} - 1\right)^{+}\right) q' \tag{1.46}$$

where $p' = 1 - (1 - p\alpha)^n$ and $q' = 1 - (1 - q\alpha)^n$. By assumption 1, $n \geq 1$ and both p' and q' are probabilities between 0 and 1, as $1 - p\alpha$ and $1 - q\alpha$ are between 0 and 1. In addition, the term $(N/n - 1 - (m/n - 1)^{+})$ is nonnegative as $N \geq n$ and $N > m$. As a result, the entirety of equation 1.46 is nonnegative.

Case 2: When group size is smaller than the expected number of infected individuals and community size, $n < 1 + (N-1)v$ and $n < m$, we can write the lower bound T_{LB} as

$$\begin{aligned} T_{LB} &= \frac{N}{n} + 1 + (N-1)v && (1.47)\\ &= \frac{N}{n} + 1 + (m-1)p\alpha + (N-m)q\alpha && (1.48)\\ &= \frac{N}{n} + n + (m-1)(1-(1-p\alpha)) + (N-m)(1-(1-q\alpha)) - (n-1) && (1.49) \end{aligned}$$

where the second equality makes use of the $E[I_D] = E[I_{NG}]$ equivalence specified in assumption 1. Equation 1.49 is a slight rearrangement of equation 1.48. As $n < m$, $\mathrm{E}[T_{NG}]$ becomes

$$\begin{aligned} \mathrm{E}[T_{NG}] &= \frac{N}{n} + n + (m-n)p' + (N-m)q' && (1.50)\\ &= \frac{N}{n} + n + (m-1)p' + (N-m)q' - (n-1)p' && (1.51) \end{aligned}$$

Subtracting equation 1.49 from equation 1.51 yields

$$\begin{aligned}&\mathrm{E}[T_{NG}] - T_{LB} = \\ &\quad (m-1)[p' - (1-(1-p\alpha))] + (N-m)[q' - (1-(1-q\alpha))] + (n-1)(1-p')\end{aligned} \tag{1.52}$$

By assumption 1, we have $m > 1$, $N > m$, and $n \geq 1$. In addition, $1 \geq p'$ as $p' = 1-(1-p\alpha)^n$ is a probability between 0 and 1. Lastly, $p' = 1-(1-p\alpha)^n \geq 1-(1-p\alpha)$ as $(1-p\alpha)^n \leq (1-p\alpha)$. Similarly, $q' = 1-(1-q\alpha)^n \geq 1-(1-q\alpha)$. As a result, $\mathrm{E}[T_{NG}] - T_{LB}$ is nonnegative.

Case 3: We consider the case where group size is smaller than the expected number of infected individuals but larger than (or equal to) community size, $n < 1+(N-1)v$ and $n \geq m$. Using equation 1.48 and the inequality $m \geq 1+(m-1)p\alpha$, we have the following inequality for the lower bound T_{LB}.

$$T_{LB} = \frac{N}{n} + 1 + (m-1)p\alpha + (N-m)q\alpha \tag{1.53}$$

$$\leq \frac{N}{n} + m + (N-m)q\alpha \tag{1.54}$$

$$= \frac{N}{n} + n + (N-m)(1-(1-q\alpha)) - (n-m) \tag{1.55}$$

Since $n \geq m$, $\mathrm{E}[T_{NG}]$ becomes

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n + (N-n)\,q' \tag{1.56}$$

$$= \frac{N}{n} + n + (N-m)\,q' - (n-m)q' \tag{1.57}$$

Subtracting equation 1.55 from equation 1.57 yields

$$\mathrm{E}[T_{NG}] - T_{LB} \geq (N-m)[q' - (1-(1-q\alpha))] + (n-m)(1-q') \tag{1.58}$$

By assumption, $N > m$ and $n \geq m$. In addition, $1 \geq q'$ as q' is a probability between 0 and 1. Lastly, $q' = 1-(1-q\alpha)^n \geq 1-(1-q\alpha)$ as $(1-q\alpha)^n \leq (1-q\alpha)$. As a

result, the difference $\mathrm{E}[T_{NG}] - T_{LB}$ is nonnegative.

We have proven $\mathrm{E}[T_{NG}] - T_{LB} \geq 0$ for the three cases under consideration, completing the lower bound portion of the proof and completing the proof of theorem 1. □

1.19.5 Proof of corollary 1 and 2

Corollary 1 follows directly from the definition of $\mathrm{E}[T_{NG}]$. When $q = 0$, we have $q' = 1 - (1 - q\alpha)^n = 0$. When $n \geq m$ and $q' = 0$, $\mathrm{E}[T_{NG}]$ simplifies to

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n \tag{1.59}$$

By assumption 1, $\alpha = (N-1)v/[(m-1)p + (N-m)q] \leq 1$. When $q = 0$, we have $(N-1)v/[(m-1)p] \leq 1$, which implies $v \leq (m-1)p/(N-1)$. Recall the number of infected individuals is $1 + (N-1)v$. We now have $1 + (N-1)v \leq 1 + (m-1)p$ and $1 + (m-1)p \leq m$ as $p \leq 1$. Since $m \leq n$, we have $1 + (N-1)v \leq n$. Therefore, the lower bound T_{LB} is

$$T_{LB} = \frac{N}{n} + n \tag{1.60}$$

and $\mathrm{E}[T_{NG}] = T_{LB}$, completing the proof. □

Note, the assumption $\alpha \leq 1$ in assumption 1 sets an upper bound for the infection prevalence v as α is a function of v. However, this is not restrictive as group testing is employed in cases when v is low.

Corollary 2 is proved during the proof of theorem 1. See the upper bound portion of the proof in 1.19.4.

1.19.6 Derivation of imperfect community detection

Under two-stage testing procedures, a population of size N is pooled into N/n groups of size n for the initial stage of testing. As described in 1.19.3, the N/n groups under network grouping can be split into three categories under perfect community detection:

1) one group that contains the infected seed individual, 2) $({}^{m}/_{n} - 1)^{+}$ groups that contain individuals from the same community as the infected seed, and 3) the remaining ${}^{N}/_{n} - 1 - ({}^{m}/_{n} - 1)^{+}$ groups that contain individuals from different communities than the infected seed.

Under imperfect community detection, individuals from communities other than the infected seed's may be incorrectly placed in the $({}^{m}/_{n} - 1)^{+}$ groups and individuals from the same community as the infected seed may be incorrectly placed in the ${}^{N}/_{n} - 1 - ({}^{m}/_{n} - 1)^{+}$ groups. As a result, the probabilities the $({}^{m}/_{n} - 1)^{+}$ groups and the ${}^{N}/_{n} - 1 - ({}^{m}/_{n} - 1)^{+}$ groups test positive change under imperfect community detection.

Under perfect community detection, the $({}^{m}/_{n} - 1)^{+}$ groups each contain n individuals that are each not infected with probability $1 - p\alpha$. Under completely imperfect community detection, individuals are placed into groups uniformly at random. There are $m - 1$ out of $N - 1$ individuals in the population from the same community as the infected seed and each is not infected with probability $1 - p\alpha$. In addition, there are $N - m$ out of $N - 1$ individuals from different communities than the infected seed and each is not infected with probability $1 - q\alpha$. As a result, when an individual is chosen uniformly at random from the population and placed into a group, the probability the individual is not infected is $({}^{m-1}/_{N-1})(1 - p\alpha) + ({}^{N-m}/_{N-1})(1 - q\alpha)$.

The parameter $\lambda \in [0, 1]$ records the imperfection of the community detection algorithm used. Perfect community detection corresponds to $\lambda = 0$ and completely imperfect community detection, under which individuals are grouped uniformly at random, corresponds to $\lambda = 1$. Therefore, the probability an individual in the $({}^{m}/_{n} - 1)^{+}$ groups is not infected can be written as

$$(1 - \lambda)(1 - p\alpha) + \lambda \left(\frac{m - 1}{N - 1}(1 - p\alpha) + \frac{N - m}{N - 1}(1 - q\alpha) \right) \tag{1.61}$$

In words, the probability an individual in a group is not infected is the convex combination of their probability under perfect community detection and their probability under completely imperfect community detection. When $\lambda = 0$, equation 1.61 simplifies

to $1-p\alpha$, the probability of non-infection under perfect community detection, and when $\lambda = 1$, equation 1.61 simplifies to $({}^{m-1}/_{N-1})(1-p\alpha) + ({}^{N-m}/_{N-1})(1-q\alpha)$, the probability of non-infection under completely imperfect community detection. As a result, for the $({}^{m}/_{n} - 1)^{+}$ groups, the probability that at least one individual in a group of size n is infected, and therefore the group tests positive, is

$$p'_{\lambda} = 1 - \left[(1-\lambda)(1-p\alpha) + \lambda\left(\frac{m-1}{N-1}(1-p\alpha) + \frac{N-m}{N-1}(1-q\alpha)\right)\right]^{n} \tag{1.62}$$

Under perfect community detection, the ${}^{N}/_{n} - 1 - ({}^{m}/_{n} - 1)^{+}$ groups each contain n individuals that are each not infected with probability $1-q\alpha$. The probability each group has at least one infected individual and therefore tests positive under imperfect community detection, which we denote q'_{λ}, is derived identically to p'_{λ}.

Putting everything together, under imperfect community detection, ${}^{N}/_{n}$ groups are tested in the first stage. In the second stage, nG samples are tested where G denotes the number of positive groups from the first stage. Of the ${}^{N}/_{n}$ groups, one will be positive with probability one as it contains the infected seed, $({}^{m}/_{n} - 1)^{+}$ groups will be positive independently with probability p'_{λ} as described above, and the remaining ${}^{N}/_{n} - 1 - ({}^{m}/_{n} - 1)^{+}$ groups will be positive independently with probability q'_{λ}. Therefore, the expected number of tests used under network grouping and imperfect community detection is as given in equation 1.15.

1.19.7 Proof of theorem 2

To prove $\mathrm{E}[T^{\lambda}_{NG}] \leq \mathrm{E}[T_D]$ under imperfect community detection, we prove $\mathrm{E}[T^{\lambda}_{NG}]$ is increasing in λ under assumption 1 and equals $\mathrm{E}[T_D]$ when λ is set to its maximum value of 1. To prove $\mathrm{E}[T^{\lambda}_{NG}]$ is increasing in λ, we consider the cases where $m \leq n$ and $m > n$ separately.

Case 1: We first consider the case where $m \leq n$. When $m \leq n$, $\mathrm{E}[T_{NG}^{\lambda}]$ under imperfect community detection is given by

$$\mathrm{E}[T_{NG}^{\lambda}] = \frac{N}{n} + n\left[1 + \left(\frac{N}{n} - 1\right) q'_{\lambda}\right] \tag{1.63}$$

$$q'_{\lambda} = 1 - \left[(1-\lambda)(1-q\alpha) + \lambda\left(\frac{m-1}{N-1}(1-p\alpha) + \frac{N-m}{N-1}(1-q\alpha)\right)\right]^{n} \tag{1.64}$$

Taking the derivative of equation 1.63 with respect to λ and simplifying yields

$$\frac{\partial\, \mathrm{E}[T_{NG}^{\lambda}]}{\partial\lambda} = \frac{n\alpha(N-n)(m-1)(p-q)}{N-1}\left(\frac{\lambda(1-p\alpha)(m-1) + (1-q\alpha)(N-1-\lambda(m-1))}{N-1}\right)^{n-1} \tag{1.65}$$

The terms n, α, $N-n$, $m-1$, $p-q$, $N-1$, λ, $1-p\alpha$, and $1-q\alpha$ are all nonnegative by assumption 1 and the assumption $\lambda \in [0,1]$. To show the remaining term $N - 1 - \lambda(m-1)$ is nonnegative, we note it is decreasing in λ. Therefore, setting $\lambda = 1$ lower bounds the term by $N - m$, which is positive. Therefore, the remaining term is nonnegative, equation 1.65 is nonnegative, and equation 1.63 is increasing in λ.

Case 2: We now consider the case where $m > n$. When $m > n$, $\mathrm{E}[T_{NG}^{\lambda}]$ under imperfect community detection is given by

$$\mathrm{E}[T_{NG}^{\lambda}] = \frac{N}{n} + n\left[1 + \left(\frac{m}{n} - 1\right) p'_{\lambda} + \left(\frac{N}{n} - \frac{m}{n}\right) q'_{\lambda}\right] \tag{1.66}$$

$$p'_{\lambda} = 1 - \left[(1-\lambda)(1-p\alpha) + \lambda\left(\frac{m-1}{N-1}(1-p\alpha) + \frac{N-m}{N-1}(1-q\alpha)\right)\right]^{n} \tag{1.67}$$

$$q'_{\lambda} = 1 - \left[(1-\lambda)(1-q\alpha) + \lambda\left(\frac{m-1}{N-1}(1-p\alpha) + \frac{N-m}{N-1}(1-q\alpha)\right)\right]^{n} \tag{1.68}$$

Taking the derivative of equation 1.66 with respect to λ and simplifying yields

$$\frac{\partial\, \mathrm{E}[T_{NG}^{\lambda}]}{\partial\lambda} = \frac{n\alpha(N-m)(p-q)}{N-1}\left[(m-1)\left(\frac{(1-q\alpha)(N-1) - \lambda\alpha(m-1)(p-q)}{N-1}\right)^{n-1} - (m-n)\left(\frac{(1-p\alpha)(N-1) + \lambda\alpha(N-m)(p-q)}{N-1}\right)^{n-1}\right] \tag{1.69}$$

All terms in the leading factor before the bracket are nonnegative by assumption 1. Within the bracket, we note $m-1 \geq m-n$. Therefore, we only have to show the first numerator, $(1-q\alpha)(N-1)-\lambda\alpha(m-1)(p-q)$, is larger than the second numerator, $(1-p\alpha)(N-1)+\lambda\alpha(N-m)(p-q)$, to prove the entire bracketed term is nonnegative. To do so, we first note the difference of the numerator terms is decreasing in λ, which can be seen by taking the derivative of the difference with respect to λ. Taking the derivative and simplifying yields $-\alpha(N-1)(p-q)$, which is nonpositive. Therefore, setting $\lambda = 1$ provides a lower bound on the difference of the numerator terms. When $\lambda = 1$, the difference of the numerator terms simplifies to 0, indicating the first numerator is greater than or equal to the second. As a result, equation 1.69 is nonnegative and equation 1.66 is increasing in λ.

Final step: We have shown $\mathrm{E}[T_{NG}^{\lambda}]$ under imperfect community detection is increasing in λ for $m \leq n$ and $m > n$. Setting λ to its maximum value of 1, plugging in $\alpha = (N-1)v/[(m-1)p+(N-m)q]$ which holds under assumption 1, and simplifying yields

$$p'_{\lambda=1} = q'_{\lambda=1} = 1 - \left(\frac{m-1}{N-1}(1-p\alpha) + \frac{N-m}{N-1}(1-q\alpha)\right)^n = 1-(1-v)^n \qquad (1.70)$$

Therefore, $p'_{\lambda=1} = q'_{\lambda=1} = v'$ and $\mathrm{E}[T_{NG}^{\lambda}]$ simplifies to

$$\mathrm{E}[T_{NG}^{\lambda}] = \frac{N}{n} + n\left[1 + \left(\frac{N}{n} - 1\right)v'\right] \qquad (1.71)$$

and we have $\mathrm{E}[T_{NG}^{\lambda}] = \mathrm{E}[T_D]$ under imperfect community detection when $\lambda = 1$.

Alternatively, when $\lambda = 0$, we have $p'_{\lambda} = 1-(1-p\alpha)^n = p'$ and $q'_{\lambda} = 1-(1-q\alpha)^n = q'$. Therefore, $\mathrm{E}[T_{NG}^{\lambda}] = \mathrm{E}[T_{NG}]$ when $\lambda = 0$. □

1.19.8 Derivation of individual testing under imperfect tests

Under our setup, the population of size N has $1+(N-1)v$ infected individuals in expectation. Under individual testing, each person in the population of is tested individually. We use s_{e_2} to denote the sensitivity of the tests, where sensitivity is 1

minus the false negative rate. Each infected individual will test falsely negative with probability $1 - s_{e_2}$ and, as a result, we have $(1 + (N-1)v)(1 - s_{e_2})$ false negatives in expectation.

The population of size N has $(N-1)(1-v)$ non-infected individuals in expectation. We use s_{p_2} to denote the specificity of the tests, where specificity is 1 minus the false positive rate. Each non-infected individual will test falsely positive with probability $1 - s_{p_2}$ and, as a result, we have $(N-1)(1-v)(1-s_{p_2})$ false positives in expectation.

Note, we use s_{e_2} and s_{p_2} for individual testing (as opposed to $s_{e_1,n}$ and s_{p_1}) because the second-stage tests screen individual samples whereas the first-stage tests screen group samples. Using s_{e_2} and s_{p_2} therefore provide a more natural benchmark when comparing individual testing to Dorfman testing and network group testing.

1.19.9 Derivation of Dorfman testing under imperfect tests

Number of tests The number of tests used by two-stage testing procedures is $N/n + nG$ where N is the population size, n is the group size, and G is the number of groups that test positive in the first stage of testing. G is a random variable. Under perfect tests, the expected number of tests used under Dorfman testing is provided in equation 1.1. Under imperfect tests, the expectation of G changes as truly positive groups may test negative incorrectly and truly negative groups may test positive incorrectly.

Under Dorfman testing, one group is positive with probability one and the remaining $N/n - 1$ are positive independently with probability $1 - (1-v)^n$. Under imperfect tests, the first group (which is truly positive) tests positive with probability $s_{e_1,n}$ and the remaining $N/n - 1$ groups test positive independently with some probability v'_n. The $N/n - 1$ groups test positive if they are truly positive and test positive correctly, which occurs with probability $s_{e_1,n}(1 - (1-v)^n)$, or if they are truly negative and test positive incorrectly, which occurs with probability $(1 - s_{p_1})(1-v)^n$. Therefore,

$v'_n = s_{e_1,n}(1-(1-v)^n) + (1-s_{p_1})(1-v)^n$. Putting everything together,

$$\mathrm{E}[G] = s_{e_1,n} + \left(\frac{N}{n} - 1\right) v'_n \tag{1.72}$$

and the expected number of tests used under Dorfman testing and imperfect tests is as shown in equation 1.18.

False negatives Under Dorfman testing, individual samples can be split into three categories: 1) the infected seed individual that is infected with probability one, 2) the $n-1$ individuals that are placed in the same group as the infected seed, and 3) the remaining $N-n$ individuals. To derive the expected number of false negatives, we derive the probability of a false negative for each category.

The infected seed tests falsely negative if its group tests falsely negative in the first stage, which occurs with probability $1-s_{e_1,n}$, or if its group tests positive in the first stage and then its sample tests negative in the second stage, which occurs with probability $s_{e_1,n}(1-s_{e_2})$. As a result, the infected seed tests falsely negative with probability $1-s_{e_1,n}+s_{e_1,n}(1-s_{e_2})$. Each of the $n-1$ individuals in the infected seed's group tests falsely negative if they are truly positive and their group tests falsely negative in the first stage or if they are truly positive, their group tests positive in the first stage, and then their sample tests negative in the second stage. As a result, each of these $n-1$ individuals test falsely negative with probability $v(1-s_{e_1,n})+vs_{e_1,n}(1-s_{e_2})$. The remaining $N-n$ individuals test falsely negative with the same probability, since they also test falsely negative if they are truly positive and their group tests falsely negative in the first stage or if they are truly positive, their group tests positive in the first stage, and then their sample tests negative in the second stage.

The expected number of false negatives under Dorfman testing and imperfect tests is therefore given by

$$\begin{aligned}\mathrm{E}[FN_D] &= 1 - s_{e_1,n} + s_{e_1,n}(1 - s_{e_2}) + \\ &\quad (n-1)[v(1 - s_{e_1,n}) + v s_{e_1,n}(1 - s_{e_2})] + \\ &\quad (N-n)[v(1 - s_{e_1,n}) + v s_{e_1,n}(1 - s_{e_2})] \\ &= (1 + (N-1)v)(1 - s_{e_1,n} s_{e_2}) \end{aligned} \tag{1.73}$$

False positives Under Dorfman testing, individual samples can be split into three categories: 1) the infected seed individual that is infected with probability one, 2) the $n-1$ individuals that are placed in the same group as the infected seed, and 3) the remaining $N-n$ individuals. To derive the expected number of false positives, we derive the probability of a false positive for each category.

The infected seed cannot be falsely positive. The $n-1$ individuals test falsely positive if they are truly negative, with probability $1-v$, their group tests positive in the first stage, with probability $s_{e_1,n}$, and they test falsely positive in the second stage, with probability $1 - s_{p_2}$. As a result, each of the $n-1$ individuals test falsely positive with probability $(1-v)s_{e_1,n}(1-s_{p_2})$. The $N-n$ individuals test falsely positive if they are truly negative, with probability $1-v$, their group tests positive in the first stage, with some probability v'_{n-1}, and they test falsely positive in the second stage, with probability $1 - s_{p_2}$. Since the individuals in question are truly negative, their group tests positive in the first stage if at least one of the remaining $n-1$ individuals in the group is truly positive and the group tests positive correctly, with probability $s_{e_1,n}(1-(1-v)^{n-1})$, or if the remaining $n-1$ individuals in the group are truly negative and the group tests positive incorrectly, with probability $(1-s_{p_1})(1-v)^{n-1}$. As a result, $v'_{n-1} = s_{e_1,n}(1-(1-v)^{n-1}) + (1-s_{p_1})(1-v)^{n-1}$ and the $N-n$ individuals test positive incorrectly with probability $(1-v)(1-s_{p_2})v'_{n-1}$.

The expected number of false positives under Dorfman testing and imperfect tests

is therefore given by

$$\mathrm{E}[FP_D] = (n-1)(1-v)(1-s_{p_2})s_{e_1,n} + (N-n)(1-v)(1-s_{p_2})v'_{n-1} \tag{1.74}$$

where $v'_{n-1} = s_{e_1,n}(1-(1-v)^{n-1}) + (1-s_{p_1})(1-v)^{n-1}$. Note, under our notation for v'_{n-1}, the sensitivity $s_{e_1,n}$ depends on the original group size n.

1.19.10 Derivation of the lower bound under imperfect tests

The number of tests used by two-stage testing procedures is $N/n + nG$ where N is the population size, n is the group size, and G is the number of groups that test positive in the first stage of testing. The lower bound is achieved by using perfect pooling, as described in 1.19.2. Under perfect pooling and perfect tests, the number of truly positive groups and the number of groups that test positive are equivalent, with both equal to $G = \max(1, (1+(N-1)v)/n)$. However, under imperfect tests, G changes as truly positive groups may test negative incorrectly and truly negative groups may test positive incorrectly. There are N/n groups in total, truly positive groups test positive with probability $s_{e_1,n}$, and truly negative groups test positive with probability $1 - s_{p_1}$. As a result, under perfect pooling and imperfect tests,

$$\mathrm{E}[G] = s_{e_1,n} \cdot \max\left(1, \frac{1+(N-1)v}{n}\right) + (1-s_{p_1})\left(\frac{N}{n} - \max\left(1, \frac{1+(N-1)v}{n}\right)\right) \tag{1.75}$$

and the two-stage testing lower bound is as shown in equation 1.21.

1.19.11 Derivation of network grouping under imperfect tests

Number of tests The number of tests used by two-stage testing procedures is $N/n + nG$ where N is the population size, n is the group size, and G is the number of groups that test positive in the first stage of testing. G is a random variable. Under perfect tests, the expected number of tests used under network grouping is provided in equation 1.3 and is derived in 1.19.3. Under imperfect tests, the expectation of

G changes as truly positive groups may test negative incorrectly and truly negative groups may test positive incorrectly.

Recall under network grouping, m is the number of individuals in each communities, p is the probability of an edge between two individuals in the same community, q is the probability of an edge between two individuals in different communities, and α is the probability an infected individual passes on the infection to their network neighbor. As described in 1.19.3, the N/n groups can be split into three categories: 1) the group that contains the infected seed individual, 2) the $(m/n - 1)^+$ groups that contain individuals from the same community as the infected seed, and 3) the $N/n - 1 - (m/n - 1)^+$ groups that contain individuals from different communties than the infected seed.

We derive the probability of testing positive for each of the three group categories. The group that contains the infected seed tests positive with probability $s_{e_1,n}$. The $(m/n - 1)^+$ groups test positive if they contain at least one infected individual, with probability $1 - (1 - p\alpha)^n$, and test positive correctly, with probability $s_{e_1,n}$, or if they contain no infected individuals, with probability $(1 - p\alpha)^n$, and test positive incorrectly, with probability $(1 - s_{p_1})$. The probability each of the $(m/n - 1)^+$ groups test positive is therefore $p'_n = s_{e_1,n}(1 - (1 - p\alpha)^n) + (1 - s_{p_1})(1 - p\alpha)^n$. Similarly, the $N/n - 1 - (m/n - 1)^+$ groups test positive if they contain at least one infected individual, with probability $1 - (1 - q\alpha)^n$, and test positive correctly, with probability $s_{e_1,n}$, or if they contain no infected individuals, with probability $(1 - q\alpha)^n$, and test positive incorrectly, with probability $(1 - s_{p_1})$. The probability each of the $N/n - 1 - (m/n - 1)^+$ groups test positive is therefore $q'_n = s_{e_1,n}(1 - (1 - q\alpha)^n) + (1 - s_{p_1})(1 - q\alpha)^n$.

Putting everything together, the expected number of groups that test positive in the first stage under network grouping and imperfect tests is

$$\mathrm{E}[G] = s_{e_1,n} + \left(\frac{m}{n} - 1\right)^+ p'_n + \left(\frac{N}{n} - 1 - \left(\frac{m}{n} - 1\right)^+\right) q'_n \tag{1.76}$$

where $p'_n = s_{e_1,n}(1 - (1 - p\alpha)^n) + (1 - s_{p_1})(1 - p\alpha)^n$, $q'_n = s_{e_1,n}(1 - (1 - q\alpha)^n) + (1 - s_{p_1})(1 - q\alpha)^n$, and $(x)^+ = \max(x, 0)$. The expected number of tests used under

network grouping and imperfect tests is therefore as shown in equation 1.22.

False negatives Under network grouping, individual samples can be split into three categories: 1) the infected seed individual, 2) the $m-1$ individuals from the same community as the infected seed, and 3) the $N-m$ individuals from different communities than the infected seed. To derive the expected number of false negatives, we derive the probability of a false negative for each category.

The infected seed tests negative if its group tests falsely negative in the first stage, which occurs with probability $1-s_{e_1,n}$, or if its group tests positive in the first stage and then its sample tests negative in the second stage, which occurs with probability $s_{e_1,n}(1-s_{e_2})$. As a result, the infected seed tests falsely negative with probability $1-s_{e_1,n}+s_{e_1,n}(1-s_{e_2})$. Each of the $m-n$ individuals from the same community as the infected seed tests falsely negative if they are truly positive, with probability $p\alpha$, and their group tests falsely negative in the first stage or if they are truly positive, their group tests positive in the first stage, and then their sample tests negative in the second stage. As a result, each of these $m-n$ individuals test falsely negative with probability $p\alpha(1-s_{e_1,n})+p\alpha s_{e_1,n}(1-s_{e_2})$. Each of the $N-m$ individuals from different communities than the infected seed tests falsely negative if they are truly positive, with probability $q\alpha$, and their group tests falsely negative in the first stage or if they are truly positive, their group tests positive in the first stage, and then their sample tests negative in the second stage. As a result, each of these $N-m$ individuals test falsely negative with probability $q\alpha(1-s_{e_1,n})+q\alpha s_{e_1,n}(1-s_{e_2})$.

The expected number of false negatives under network grouping and imperfect tests is therefore given by

$$\begin{aligned}
\mathrm{E}[FN_{NG}] &= 1-s_{e_1,n}+s_{e_1,n}(1-s_{e_2})+ \\
&\qquad (m-1)[p\alpha(1-s_{e_1,n})+p\alpha s_{e_1,n}(1-s_{e_2})]+ \\
&\qquad (N-m)[q\alpha(1-s_{e_1,n})+q\alpha s_{e_1,n}(1-s_{e_2})] \\
&= (1+(m-1)p\alpha+(N-m)q\alpha)(1-s_{e_1,n}s_{e_2}) \qquad (1.77)
\end{aligned}$$

False positives Under network grouping, individual samples can be split into five categories: 1) the infected seed individual, 2) the $(\min(n, m) - 1)$ individuals from the same community as the infected seed that are placed in the same group as the infected seed, 3) the $(n - \min(n, m))$ individuals from different communities than the infected seed that are placed in the same group as the infected seed, 4) the $(m-n)^+$ individuals from the same community as the infected seed that are placed in different groups than the infected seed, and 5) the $(N - \max(n, m))$ individuals from different communities than the infected seed that are placed in different groups than the infected seed. Note, when community size m is greater than group size n, the group sizes are 1, $n - 1$, 0, $m - n$, and $N - m$ respectively. When $m \leq n$, the group sizes are 1, $m - 1$, $n - m$, 0, and $N - n$ respectively. To derive the expected number of false positives, we derive the probability of a false positive for each category.

The infected seed cannot test falsely positive. The $(\min(n, m) - 1)$ individuals from the same community as the infected seed that are placed in the same group as the infected seed test falsely positive if they are truly negative, with probability $1 - p\alpha$, their group tests correctly positive in the first stage, with probability $s_{e_1,n}$, and they test incorrectly positive in the second stage, with probability $1 - s_{p_2}$. As a result, each tests falsely positive with probability $(1 - p\alpha)(1 - s_{p_2})s_{e_1,n}$. The $(n - \min(n, m))$ individuals from different communities than the infected seed that are placed in the same group as the infected seed test falsely positive if they are truly negative, with probability $1-q\alpha$, their group tests correctly positive in the first stage, with probability $s_{e_1,n}$, and they test incorrectly positive in the second stage, with probability $1 - s_{p_2}$. As a result, each tests falsely positive with probability $(1 - q\alpha)(1 - s_{p_2})s_{e_1,n}$.

The $(m - n)^+$ individuals from the same community as the infected seed that are placed in different groups than the infected seed test falsely positive if they are truly negative, with probability $1 - p\alpha$, their group tests positive in the first stage, with some probability p'_{n-1}, and they test incorrectly positive in the second stage, with probability $1 - s_{p_2}$. Since the individuals in question are truly negative, their group tests positive in the first stage if at least one of the remaining $n - 1$ individuals in the group is truly positive and the group tests positive correctly, with probability

$s_{e_1,n}(1-(1-p\alpha)^{n-1})$, or if the remaining $n-1$ individuals in the group are truly negative and the group tests positive incorrectly, with probability $(1-s_{p_1})(1-p\alpha)^{n-1}$. As a result, $p'_{n-1} = s_{e_1,n}(1-(1-p\alpha)^{n-1}) + (1-s_{p_1})(1-p\alpha)^{n-1}$ and the $(m-n)^+$ individuals test positive incorrectly with probability $(1-p\alpha)(1-s_{p_2})p'_{n-1}$.

Similarly, the $(N-\max(n,m))$ individuals from different communities than the infected seed that are placed in different groups than the infected seed test falsely positive if they are truly negative, with probability $1-q\alpha$, their group tests positive in the first stage, with some probability q'_{n-1}, and they test incorrectly positive in the second stage, with probability $1-s_{p_2}$. Following the derivation in the previous paragraph, $q'_{n-1} = s_{e_1,n}(1-(1-q\alpha)^{n-1}) + (1-s_{p_1})(1-q\alpha)^{n-1}$ and the $(N-\max(n,m))$ individuals test positive incorrectly with probability $(1-q\alpha)(1-s_{p_2})q'_{n-1}$.

The expected number of false positives under network grouping and imperfect tests is therefore given by

$$\begin{aligned}
\mathrm{E}[FP_{NG}] = &(\min(n,m)-1)(1-p\alpha)(1-s_{p_2})s_{e_1,n}+ \\
&(n-\min(n,m))(1-q\alpha)(1-s_{p_2})s_{e_1,n}+ \\
&(m-n)^+(1-p\alpha)(1-s_{p_2})p'_{n-1}+ \\
&(N-\max(n,m))(1-q\alpha)(1-s_{p_2})q'_{n-1} \qquad (1.78)\\
p'_{n-1} = &\, s_{e_1,n}(1-(1-p\alpha)^{n-1}) + (1-s_{p_1})(1-p\alpha)^{n-1} \\
q'_{n-1} = &\, s_{e_1,n}(1-(1-q\alpha)^{n-1}) + (1-s_{p_1})(1-q\alpha)^{n-1}
\end{aligned}$$

where $(x)^+ = \max(x,0)$. Note, under our notation for p'_{n-1} and q'_{n-1}, the sensitivity $s_{e_1,n}$ depends on the original group size n.

1.19.12 Test comparison under imperfect tests

Under imperfect tests, the expected number of tests used under network grouping is upper bounded by Dorfman testing and lower bounded by the two-stage lower bound.

Theorem 6. *Under the conditions of assumptions 1 and 2 and imperfect tests,* $\mathrm{E}[T_{NG}]$

is increasing in q and

$$\mathrm{E}[T_{LB}] \leq \mathrm{E}[T_{NG}] \leq \mathrm{E}[T_D] \tag{1.79}$$

If $q = 0$ and $n \geq m$, then $\mathrm{E}[T_{NG}] = \mathrm{E}[T_{LB}]$. If $q = p$, then $\mathrm{E}[T_{NG}] = \mathrm{E}[T_D]$.

Under imperfect tests, $\mathrm{E}[T_{LB}]$ is given in equation 1.21, $\mathrm{E}[T_{NG}]$ is given in equation 1.22, and $\mathrm{E}[T_D]$ is given in equation 1.18. The proof of theorem 6 is provided below. Theorem 6 states network grouping weakly dominates Dorfman testing in terms of the expected number of tests when using imperfect tests. Identically to the perfect test setting, the expected number of tests used is increasing in q, the probability edges exists between different communities. When $q = 0$ and $n \geq m$, communities are disconnected and group sizes are large enough to contain full communities. As a result, network grouping under imperfect tests performs optimally and achieves the two-stage lower bound under imperfect tests. Conversely, when $q = p$, the network has no community structure and network grouping under imperfect tests performs equivalently to Dorfman testing under imperfect tests.

Proof of theorem 6

Upper bound To prove $\mathrm{E}[T_{NG}] \leq \mathrm{E}[T_D]$ under imperfect tests, we prove $\mathrm{E}[T_{NG}]$ is increasing in q under assumptions 1 and 2 and equals $\mathrm{E}[T_D]$ when q is set to its maximum value under assumption 1, $q = p$. This also proves the statement "If $q = p$, then $\mathrm{E}[T_{NG}] = \mathrm{E}[T_D]$." To prove $\mathrm{E}[T_{NG}]$ is increasing in q, we consider the cases where $m > n$ and $m \leq n$ separately.

Case 1: We first consider the case where $m > n$. When $m > n$, $\mathrm{E}[T_{NG}]$ under imperfect tests is given by

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n\left[s_{e_1,n} + \left(\frac{m}{n} - 1\right)p'_n + \left(\frac{N}{n} - \frac{m}{n}\right)q'_n\right] \tag{1.80}$$

where $p'_n = s_{e_1,n}(1-(1-p\alpha)^n) + (1-s_{p_1})(1-p\alpha)^n$ and $q'_n = s_{e_1,n}(1-(1-q\alpha)^n) + (1-s_{p_1})(1-q\alpha)^n$. Under assumption 1, $\alpha = (N-1)v/[(m-1)p + (N-m)q]$. Taking

the derivative of equation 1.80 with respect to q and simplifying yields

$$\frac{\partial \, \mathrm{E}[T_{NG}]}{\partial q} = \frac{nvp(N-1)(N-m)(s_{e_1,n}+s_{p_1}-1)\left[(m-1)\,(1-q\alpha)^{n-1}-(m-n)\,(1-p\alpha)^{n-1}\right]}{[(m-1)p+(N-m)q]^2} \tag{1.81}$$

Demonstrating equation 1.81 is nonnegative proves equation 1.80 is increasing in q. The denominator is nonnegative due to the square and n, v, p, $N-1$, $N-m$, and $s_{e_1,n}+s_{p_1}-1$ are nonnegative by assumptions 1 and 2. Examining the bracket term in the numerator, we note $m-1 \geq m-n$ as $n \geq 1$ and $(1-q\alpha)^{n-1} \geq (1-p\alpha)^{n-1}$ as $p \geq q$. As a result, the bracket term is nonnegative and the entirety of equation 1.81 is nonnegative.

Case 2: We now consider the case where $m \leq n$. When $m \leq n$, $\mathrm{E}[T_{NG}]$ under imperfect tests is given by

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n\left[s_{e_1,n} + \left(\frac{N}{n}-1\right) q'_n\right] \tag{1.82}$$

where $q'_n = s_{e_1,n}(1-(1-q\alpha)^n)+(1-s_{p_1})(1-q\alpha)^n$. Again, $\alpha = (N-1)v/[(m-1)p+(N-m)q]$. Taking the derivative of equation 1.82 with respect to q and simplifying yields

$$\frac{\partial \, \mathrm{E}[T_{NG}]}{\partial q} = \frac{nvp(N-1)(N-n)(m-1)\,(1-q\alpha)^n\,(s_{e_1,n}+s_{p_1}-1)}{[(m-1)p+(N-m)q]\,[(m-1)p+(N-m)q+q(1-N)v]} \tag{1.83}$$

All terms in the numerator are nonnegative by assumptions 1 and 2. The denominator is equal to the denominator in equation 1.40 which we prove is nonnegative in 1.19.4. As a result, equation 1.83 is nonnegative and equation 1.82 is increasing in q.

Final step: We have shown $\mathrm{E}[T_{NG}]$ under imperfect tests is increasing in q for $m > n$ and $m \leq n$. Setting q to its maximum value under assumption 1, $q = p$, we have $p'_n = q'_n$. In addition, α simplifies to v/p and $p\alpha = v$. Therefore, $p'_n = q'_n = v'_n$ where $v'_n = s_{e_1,n}(1-(1-v)^n)+(1-s_{p_1})(1-v)^n$. $\mathrm{E}[T_{NG}]$ under imperfect tests in

the general case simplifies to

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n\left[s_{e_1,n} + \left(\frac{m}{n} - 1\right)^+ p'_n + \left(\frac{N}{n} - 1 - \left(\frac{m}{n} - 1\right)^+\right) q'_n\right] \tag{1.84}$$

$$= \frac{N}{n} + n\left[s_{e_1,n} + \left(\frac{N}{n} - 1\right) v'_n\right] \tag{1.85}$$

and we have $\mathrm{E}[T_{NG}] = \mathrm{E}[T_D]$ under imperfect tests, completing the upper bound portion of the proof.

Lower bound To prove $\mathrm{E}[T_{NG}] \geq T_{LB}$ under imperfect tests, we prove $\mathrm{E}[T_{NG}] - T_{LB} \geq 0$ for the two cases of 1) group size larger than (or equal to) the expected number of infected individuals, $n \geq 1 + (N-1)v$, and 2) group size less than the expected number of infected individuals, $n < 1 + (N-1)v$.

Case 1: When group size is larger than or equal to the expected number of infected individuals, $n \geq 1 + (N-1)v$, the lower bound in equation 1.21 simplifies to

$$T_{LB} = \frac{N}{n} + n\left[s_{e_1,n} + (1 - s_{p_1})\left(\frac{N}{n} - 1\right)\right] \tag{1.86}$$

Therefore,

$$\mathrm{E}[T_{NG}] - T_{LB} = n\left(\frac{m}{n} - 1\right)^+ p'_n + n\left(\frac{N}{n} - 1 - \left(\frac{m}{n} - 1\right)^+\right) q'_n - (N-n)(1 - s_{p_1}) \tag{1.87}$$

where $p'_n = s_{e_1,n}(1 - (1-p\alpha)^n) + (1 - s_{p_1})(1 - p\alpha)^n$, $q'_n = s_{e_1,n}(1 - (1-q\alpha)^n) + (1 - s_{p_1})(1 - q\alpha)^n$, and $(x)^+ = \max(x, 0)$. To prove equation 1.87 is nonnegative, we prove the first two terms, which are positive by assumptions 1 and 2, are larger than the final negative term. To do so, we first show $p'_n \geq q'_n$ and then show $q'_n \geq 1 - s_{p_1}$ before simplifying equation 1.87. First, we have $p'_n \geq q'_n$ as

$$p'_n - q'_n = s_{e_1,n}(1 - (1-p\alpha)^n) + (1 - s_{p_1})(1 - p\alpha)^n - s_{e_1,n}(1 - (1-q\alpha)^n) - (1 - s_{p_1})(1 - q\alpha)^n$$

$$= s_{e_1,n}[(1 - (1-p\alpha)^n) - (1 - (1-q\alpha)^n)] + (1 - s_{p_1})[(1-p\alpha)^n - (1-q\alpha)^n]$$

$$= ((1-q\alpha)^n - (1-p\alpha)^n)(s_{e_1,n} + s_{p_1} - 1) \tag{1.88}$$

which is nonnegative by assumptions 1 and 2. Now $q'_n \geq 1 - s_{p_1}$ as

$$\begin{aligned} q'_n - (1 - s_{p_1}) &= s_{e_1,n}(1 - (1-q\alpha)^n) + (1 - s_{p_1})(1-q\alpha)^n - (1 - s_{p_1}) \\ &= s_{e_1,n}(1 - (1-q\alpha)^n) - (1 - s_{p_1})(1 - (1-q\alpha)^n) \\ &= (1 - (1-q\alpha)^n)(s_{e_1,n} + s_{p_1} - 1) \end{aligned} \tag{1.89}$$

which is nonnegative again by assumptions 1 and 2. Therefore, we lower bound equation 1.87 by replacing p'_n with q'_n yielding

$$\mathrm{E}[T_{NG}] - T_{LB} \geq n\left(\frac{m}{n} - 1\right)^+ q'_n + n\left(\frac{N}{n} - 1 - \left(\frac{m}{n} - 1\right)^+\right) q'_n - (N-n)\,(1 - s_{p_1}) \tag{1.90}$$

$$= (N-n)\,q'_n - (N-n)\,(1 - s_{p_1}) \tag{1.91}$$

which is nonnegative as $q'_n \geq 1 - s_{p_1}$. As a result, equation 1.87 is nonnegative.

Case 2: When group size is smaller than the expected number of infected individuals, $n < 1 + (N-1)v$, the lower bound T_{LB} under imperfect tests becomes

$$T_{LB} = \frac{N}{n} + n\left[s_{e_1,n}\frac{1 + (N-1)v}{n} + (1 - s_{p_1})\left(\frac{N}{n} - \frac{1 + (N-1)v}{n}\right)\right] \tag{1.92}$$

Subtracting T_{LB} from $\mathrm{E}[T_{NG}]$ under imperfect tests and simplifying yields

$$\begin{aligned} \mathrm{E}[T_{NG}] - T_{LB} = (s_{e_1,n} + s_{p_1} - 1)\Big[&(N-1)(1-v) - (N-n)(1-q\alpha)^n \\ &+ n\left(\frac{m}{n} - 1\right)^+ ((1-q\alpha)^n - (1-p\alpha)^n)\Big] \end{aligned} \tag{1.93}$$

To prove equation 1.93 is nonnegative, we first show it is increasing in n. We then set n to its minimum value, providing a lower bound on equation 1.93, and we demonstrate this lower bound is nonnegative. We prove equation 1.93 is increasing in n by considering the $n \geq m$ and $n < m$ cases separately. When $n \geq m$, differentiating

with respect to n yields

$$\frac{\partial\left(\mathrm{E}[T_{NG}]-T_{LB}\right)}{\partial n}=(s_{e_1,n}+s_{p_1}-1)(1-q\alpha)^n(1+(n-N)\ln(1-q\alpha)) \tag{1.94}$$

We note $n-N$ and $\ln(1-q\alpha)$ are negative, ensuring $(n-N)\ln(1-q\alpha)$ is positive. Therefore, equation 1.94 is nonnegative by assumptions 1 and 2. When $n<m$, differentiating with respect to n yields

$$\begin{aligned}\frac{\partial\left(\mathrm{E}[T_{NG}]-T_{LB}\right)}{\partial n}=(s_{e_1,n}+s_{p_1}-1)[&(1-q\alpha)^n(m-N)\ln(1-q\alpha)\\&+(1-p\alpha)^n(1+(n-m)\ln(1-p\alpha))]\end{aligned} \tag{1.95}$$

We note $m-N$, $\ln(1-q\alpha)$, $n-m$, and $\ln(1-p\alpha)$ are all negative, ensuring $(m-N)\ln(1-q\alpha)$ and $(n-m)\ln(1-p\alpha)$ are positive. Therefore, equation 1.95 is nonnegative by assumptions 1 and 2. As a result, equation 1.93 is increasing in n. Setting n to its minimum value of 1 provides a lower bound on equation 1.93.

$$\begin{aligned}\mathrm{E}[T_{NG}]-T_{LB}&\geq(s_{e_1,n}+s_{p_1}-1)\Big[(N-1)(1-v)-(N-1)(1-q\alpha)+(m-1)\left((1-q\alpha)-(1-p\alpha)\right)\Big]\\&=(s_{e_1,n}+s_{p_1}-1)[1+(m-1)p\alpha+(N-m)q\alpha-(1+(N-1)v)]\\&=0\end{aligned} \tag{1.96}$$

where the second equality makes use of the $E[I_D]=E[I_{NG}]$ equivalence specified in assumption 1. As a result, equation 1.93 is nonnegative.

We have proven $\mathrm{E}[T_{NG}]-T_{LB}\geq 0$ under imperfect tests for the two cases under consideration, completing the lower bound portion of the proof.

Equivalence at the lower and upper bounds The statement "If $q=p$, then $\mathrm{E}[T_{NG}]=\mathrm{E}[T_D]$" is proved during the upper bound portion of the proof of theorem 6.

The statement "If $q=0$ and $n\geq m$, then $\mathrm{E}[T_{NG}]=\mathrm{E}[T_{LB}]$" follows directly from the definition of $\mathrm{E}[T_{NG}]$ under imperfect tests. When $q=0$, we have $q'_n=s_{e_1,n}(1-(1-q\alpha)^n)+(1-s_{p_1})(1-q\alpha)^n=1-s_{p_1}$. When $n\geq m$ and $q'_n=1-s_{p_1}$,

$\mathrm{E}[T_{NG}]$ simplifies to

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n\left[s_{e_1,n} + (1 - s_{p_1})\left(\frac{N}{n} - 1\right)\right] \tag{1.97}$$

Identical to the proof of corollary 1 in 1.19.5, $n \geq m$ and $q = 0$ implies $1 + (N-1)v \leq n$. Therefore, the lower bound $\mathrm{E}[T_{LB}]$ under imperfect tests becomes

$$\mathrm{E}[T_{LB}] = \frac{N}{n} + n\left[s_{e_1,n} + (1 - s_{p_1})\left(\frac{N}{n} - 1\right)\right] \tag{1.98}$$

and $\mathrm{E}[T_{NG}] = \mathrm{E}[T_{LB}]$, completing the proof. □

1.19.13 Proof of theorem 3

Under individual testing, $\mathrm{E}[FP_I] = (N-1)(1-v)(1-s_{p_2})$ and under Dorfman testing, $\mathrm{E}[FP_D] = (n-1)(1-v)(1-s_{p_2})s_{e_1,n} + (N-n)(1-v)(1-s_{p_2})v'_{n-1}$ where $v'_{n-1} = s_{e_1,n}(1-(1-v)^{n-1}) + (1-s_{p_1})(1-v)^{n-1}$. Both $s_{e_1,n}$ and v'_{n-1} are probabilities between 0 and 1 as v and s_{p_1} are also probabilities between 0 and 1. Therefore, $\mathrm{E}[FP_D] \leq (n-1)(1-v)(1-s_{p_2}) + (N-n)(1-v)(1-s_{p_2}) = \mathrm{E}[FP_I]$.

To prove $\mathrm{E}[FP_{NG}] \leq \mathrm{E}[FP_D]$, we prove $\mathrm{E}[FP_{NG}]$ is increasing in q under assumptions 1 and 2 and equals $\mathrm{E}[FP_D]$ when q is set to its maximum value under assumption 1, $q = p$. To prove $\mathrm{E}[FP_{NG}]$ is increasing in q, we consider the cases where $m \leq n$ and $m > n$ separately.

Case 1: We first consider the case where $m \leq n$. When $m \leq n$, $\mathrm{E}[FP_{NG}]$ is given by

$$\begin{aligned}\mathrm{E}[FP_{NG}] = {} & (m-1)(1-p\alpha)(1-s_{p_2})s_{e_1,n} + \\ & (n-m)(1-q\alpha)(1-s_{p_2})s_{e_1,n} + \\ & (N-n)(1-q\alpha)(1-s_{p_2})q'_{n-1}\end{aligned} \tag{1.99}$$

where $q'_{n-1} = s_{e_1,n}(1-(1-q\alpha)^{n-1}) + (1-s_{p_1})(1-q\alpha)^{n-1}$. Under assumption 1, $\alpha = (N-1)v/[(m-1)p + (N-m)q]$. Taking the derivative of equation 1.99 with

respect to q and simplifying yields

$$\frac{\partial \, \mathrm{E}[FP_{NG}]}{\partial q} = \frac{nvp(N-1)(N-n)(m-1)(1-q\alpha)^n(s_{e_1,n}+s_{p_1}-1)(1-s_{p_2})}{[(m-1)p+(N-m)q]\,[(m-1)p+(N-m)q+q(1-N)v]} \tag{1.100}$$

Demonstrating equation 1.100 is nonnegative proves equation 1.99 is increasing in q. The denominator is equivalent to the denominator in equation 1.40, which we have already proved to be nonnegative. In the numerator, n, v, p, $N-1$, $N-n$, and $m-1$ are nonnegative by assumption 1. The probability $(1-q\alpha)^n$ is nonnegative as q and α are both probabilities between 0 and 1. Finally, $s_{e_1,n}+s_{p_1}-1$ and $1-s_{p_2}$ are both nonnegative as the specificities and sensitives are probabilities between 0.5 and 1 by assumption 2. As a result, the bracket term is nonnegative and the entirety of equation 1.100 is nonnegative.

Case 2: We now consider the case where $m > n$. When $m > n$, $\mathrm{E}[FP_{NG}]$ is given by

$$\begin{aligned}\mathrm{E}[FP_{NG}] = {} & (n-1)(1-p\alpha)(1-s_{p_2})s_{e_1,n} + \\ & (m-n)(1-p\alpha)(1-s_{p_2})p'_{n-1} + \\ & (N-m)(1-q\alpha)(1-s_{p_2})q'_{n-1}\end{aligned} \tag{1.101}$$

where $p'_{n-1} = s_{e_1,n}(1-(1-p\alpha)^{n-1}) + (1-s_{p_1})(1-p\alpha)^{n-1}$ and
$q'_{n-1} = s_{e_1,n}(1-(1-q\alpha)^{n-1}) + (1-s_{p_1})(1-q\alpha)^{n-1}$. Again,
$\alpha = (N-1)v/[(m-1)p+(N-m)q]$. Taking the derivative of equation 1.101 with respect to q and simplifying yields

$$\begin{aligned}\frac{\partial \, \mathrm{E}[FP_{NG}]}{\partial q} = {} & \frac{pv(N-1)(N-m)\,(1-s_{p_2})}{((m-1)p+(N-m)q)^3} \cdot \\ & \Big[((m-1)p+(N-m)q)\big[(n-1)s_{e_1,n}+(m-n)p'_{n-1}-(m-1)q'_{n-1}\big] + \\ & (n-1)(s_{e_1,n}+s_{p_1}-1)\big[(m-1)(1-q\alpha)^{n-2}((m-1)p+(N-m)q-q(N-1)v) \\ & -(m-n)(1-p\alpha)^{n-2}((m-1)p+(N-m)q-p(N-1)v)\big]\Big]\end{aligned} \tag{1.102}$$

All terms in the leading term before the large bracket are nonnegative by assumptions 1 and 2. Within the large bracket, the second term is nonnegative as $n-1$ and $s_{e_1,n}+s_{p_1}-1$ are nonnegative by the same assumptions and $m-1 \geq m-n$, $1-q\alpha \geq 1-p\alpha$ as $q \leq p$, and $q(N-1)v \leq p(N-1)v$ as $q \leq p$. Within the large bracket, the leading term of the first term, $(m-1)p+(N-m)q$, is nonnegative by assumption 1.

We now prove the final piece, $f \coloneqq (n-1)s_{e_1,n} + (m-n)p'_{n-1} - (m-1)q'_{n-1}$, is nonnegative. To do so, we prove the term is increasing in $s_{e_1,n}$ and s_{p_1}, then set $s_{e_1,n}$ and s_{p_1} to their minimum values of 0.5, and finally demonstrate the term remains nonnegative. Note, p'_{n-1} and q'_{n-1} are functions of several terms, including $s_{e_1,n}$ and s_{p_1}. Taking the derivative of f with respect to s_{p_1} yields

$$\frac{\partial f}{\partial s_{p_1}} = (m-1)(1-q\alpha)^{n-1} - (m-n)(1-p\alpha)^{n-1} \tag{1.103}$$

Equation 1.103 is nonnegative as $m-1 \geq m-n$ and $1-q\alpha \geq 1-p\alpha$ as $q \leq p$. Therefore, f is increasing in s_{p_1}. Taking the derivative of f with respect to $s_{e_1,n}$ yields

$$\frac{\partial f}{\partial s_{e_1,n}} = (n-1) - (m-1)(1-(1-q\alpha)^{n-1}) + (m-n)(1-(1-p\alpha)^{n-1}) \tag{1.104}$$

We note $(1-(1-p\alpha)^{n-1})$ is a probability between 0 and 1 and therefore equation 1.104 is greater than or equal to

$$\begin{aligned}(n-1)(1-(1-p\alpha)^{n-1}) - (m-1)(1-(1-q\alpha)^{n-1}) + (m-n)(1-(1-p\alpha)^{n-1}) \\ = (m-1)(1-(1-p\alpha)^{n-1}) - (m-1)(1-(1-q\alpha)^{n-1})\end{aligned} \tag{1.105}$$

Equation 1.105 is nonnegative as $1-(1-p\alpha)^{n-1} \geq 1-(1-q\alpha)^{n-1}$ as $p \geq q$. Therefore, 1.104 is nonnegative and f is increasing in $s_{e_1,n}$. Setting $s_{e_1,n}$ and s_{p_1} to their minimum values of 0.5 provides a lower bound for f. When $s_{e_1,n}$ and s_{p_1} equal 0.5, p'_{n-1} and q'_{n-1} simplify to 0.5, and f equals 0. Therefore, f in the general case is nonnegative. As a result, equation 1.102 is nonnegative and equation 1.101 is increasing in q.

Final step: We have shown $\mathrm{E}[FP_{NG}]$ is increasing in q for $m \leq n$ and $m > n$.

Setting q to its maximum value under assumption 1, $q = p$, we have $p'_{n-1} = q'_{n-1}$. In addition, α simplifies to v/p and $p\alpha = v$. Therefore, $p'_{n-1} = q'_{n-1} = v'_{n-1}$. When $q = p$, $\mathrm{E}[FP_{NG}]$ in the general case simplifies to

$$\begin{aligned}
\mathrm{E}[FP_{NG}] &= (\min(n,m)-1)(1-v)(1-s_{p_2})s_{e_1,n} + \\
&\quad (n-\min(n,m))(1-v)(1-s_{p_2})s_{e_1,n} + \\
&\quad (m-n)^{+}(1-v)(1-s_{p_2})v'_{n-1} + \\
&\quad (N-\max(n,m))(1-v)(1-s_{p_2})v'_{n-1} \\
&= (n-1)(1-v)(1-s_{p_2})s_{e_1,n} + \\
&\quad (N-n)(1-v)(1-s_{p_2})v'_{n-1} \qquad (1.106)
\end{aligned}$$

and we have $\mathrm{E}[FP_{NG}] = \mathrm{E}[FP_D]$, completing the proof. □

Proof of false positives increasing with n

To prove $\mathrm{E}[FP_D]$ and $\mathrm{E}[FP_{NG}]$ are increasing with n, we show the derivative of each with respect to n is nonnegative. Taking the derivative of $\mathrm{E}[FP_D]$, shown in equation 1.20 ,with respect to n and simplifying yields

$$\frac{\partial\,\mathrm{E}[FP_D]}{\partial n} = (s_{e_1,n} + s_{p_1} - 1)(1-s_{p_2})(1-v)^n(1+(n-N)\ln(1-v)) \qquad (1.107)$$

By assumptions 1 and 2, $(s_{e_1,n} + s_{p_1} - 1)$, $(1-s_{p_2})$, and $(1-v)^n$ are nonnegative. Both $n-N$ and $\ln(1-v)$ are nonpositive as v is between 0 and 1. Therefore, $1+(n-N)\ln(1-v)$ is nonnegative, equation 1.107 is nonnegative, and $\mathrm{E}[FP_D]$ is increasing in n.

To prove $\mathrm{E}[FP_{NG}]$ is increasing in n, we consider the cases where $m \le n$ and $m > n$ separately.

Case 1: We first consider the case where $m \le n$. When $m \le n$, $\mathrm{E}[FP_{NG}]$ is given by

$$\mathrm{E}[FP_{NG}] = (m-1)(1-p\alpha)(1-s_{p_2})s_{e_1,n} +$$

$$(n-m)(1-q\alpha)(1-s_{p_2})s_{e_1,n} + (N-n)(1-q\alpha)(1-s_{p_2})q'_{n-1} \quad (1.108)$$

where $q'_{n-1} = s_{e_1,n}(1-(1-q\alpha)^{n-1}) + (1-s_{p_1})(1-q\alpha)^{n-1}$. Taking the derivative of equation 1.108 with respect to n and simplifying yields

$$\frac{\partial \mathrm{E}[FP_{NG}]}{\partial n} = (s_{e_1,n} + s_{p_1} - 1)(1-s_{p_2})(1-q\alpha)^n(1+(n-N)\ln(1-q\alpha)) \quad (1.109)$$

By assumptions 1 and 2, $(s_{e_1,n} + s_{p_1} - 1)$, $(1-s_{p_2})$, and $(1-q\alpha)^n$ are nonnegative. Both $n-N$ and $\ln(1-q\alpha)$ are nonpositive as $q\alpha$ is between 0 and 1. Therefore, $1+(n-N)\ln(1-q\alpha)$ is nonnegative, equation 1.109 is nonnegative, and equation 1.108 is increasing in n.

Case 2: We now consider the case where $m > n$. When $m > n$, $\mathrm{E}[FP_{NG}]$ is given by

$$\begin{aligned}\mathrm{E}[FP_{NG}] = {} & (n-1)(1-p\alpha)(1-s_{p_2})s_{e_1,n} + \\ & (m-n)(1-p\alpha)(1-s_{p_2})p'_{n-1} + \\ & (N-m)(1-q\alpha)(1-s_{p_2})q'_{n-1}\end{aligned} \quad (1.110)$$

where $p'_{n-1} = s_{e_1,n}(1-(1-p\alpha)^{n-1}) + (1-s_{p_1})(1-p\alpha)^{n-1}$ and $q'_{n-1} = s_{e_1,n}(1-(1-q\alpha)^{n-1}) + (1-s_{p_1})(1-q\alpha)^{n-1}$. Taking the derivative of equation 1.110 with respect to n and simplifying yields

$$\begin{aligned}\frac{\partial \mathrm{E}[FP_{NG}]}{\partial n} = (s_{e_1,n} + s_{p_1} - 1)(1-s_{p_2})\big[&(1-q\alpha)^n(m-N)\ln(1-q\alpha) \\ &+ (1-p\alpha)^n(1+(n-m)\ln(1-p\alpha))\big]\end{aligned} \quad (1.111)$$

By assumptions 1 and 2, $(s_{e_1,n} + s_{p_1} - 1)$, $(1-s_{p_2})$, $(1-q\alpha)^n$, and $(1-p\alpha)^n$ are nonnegative. The terms $(m-N)\ln(1-q\alpha)$ and $(n-m)\ln(1-p\alpha)$ are nonnegative as each term $(m-N)$, $\ln(1-q\alpha)$, $(n-m)$, $\ln(1-p\alpha)$ is nonpositive. Therefore, equation

1.111 is nonnegative, equation 1.110 is increasing in n, and $\mathrm{E}[FP_{NG}]$ is increasing in n in the general case.

1.19.14 Proof of theorem 4

Under individual testing, $\mathrm{E}[FN_I] = (1 + (N-1)v)(1 - s_{e_2})$ and under Dorfman testing, $\mathrm{E}[FN_D] = (1 + (N-1)v)(1 - s_{e_1,n}s_{e_2})$. As $s_{e_1,n}$ is a probability between 0.5 and 1, $(1 - s_{e_2}) \leq (1 - s_{e_1,n}s_{e_2})$ and $\mathrm{E}[FN_I] \leq \mathrm{E}[FN_D]$. Note, if $s_{e_1,n} = 1$, then $\mathrm{E}[FN_I] = \mathrm{E}[FN_D]$.

By assumption 1, $E[I_D] = E[I_{NG}]$ and $1 + (N-1)v = 1 + (m-1)p\alpha + (N-m)q\alpha$. Therefore, $\mathrm{E}[FN_D] = \mathrm{E}[FN_{NG}]$ where $\mathrm{E}[FN_D]$ is defined above and $\mathrm{E}[FN_{NG}]$ is defined in equation 1.23. Therefore, we have $\mathrm{E}[FN_I] \leq \mathrm{E}[FN_{NG}] = \mathrm{E}[FN_D]$. □

1.19.15 Derivation of network grouping under general networks

Under two-stage group testing, a population of size N is split into N/n groups of size n for an initial round of testing. Let G denote the number of positive groups after the initial round. In the second round of testing, all n samples from each positive group are retested individually. In total, $N/n + nG$ tests are used. G is a random variable.

Of the N/n groups, one contains the infected seed and tests positive with probability one. Denote the infected seed as node i. For the remaining $N/n - 1$ groups, each group tests negative if none of the n group members are infected by the infected seed. Of the n group members in group g, let $n_{g,i}$ denote the number of nodes in the group that are connected to node i. Therefore, the probability group g is not infected is $(1-\alpha)^{n_{g,i}}$ where α is the infection passing probability. The probability the group is infected and tests positive is $1 - (1-\alpha)^{n_{g,i}}$. The expected number of groups that test positive given infected seed i is $1 + \sum_{g \in \mathcal{G} \setminus g_i}(1 - (1-\alpha)^{n_{g,i}})$ where $\mathcal{G}$ is the set of groups and g_i is the group that contains node i. Lastly, the infected seed is chosen uniformly at random from the network so we take the average over all possible infected seeds to get the expected number of positive groups, $1 + \frac{1}{N}\sum_{i \in \mathcal{N}}\sum_{g \in \mathcal{G} \setminus g_i}(1 - (1-\alpha)^{n_{g,i}})$ where $\mathcal{N}$ is the set of nodes.

Putting everything together provides the expected number of tests under network grouping for general networks

$$\mathrm{E}[T^*_{NG}] = \frac{N}{n} + n\left[1 + \frac{1}{N}\sum_{i\in\mathcal{N}}\sum_{g\in\mathcal{G}\setminus g_i}(1-(1-\alpha)^{n_{g,i}})\right] \tag{1.112}$$

where $\mathcal{N}$ is the set of nodes, $\mathcal{G}$ is the set of groups, g_i is the group that contains node i, and $n_{g,i}$ is the number of nodes in group g connected to node i.

1.19.16 Derivation of modularity

Modularity is a core metric in network science that measures the amount of community structure in a network. The metric was introduced by Newman in [101]. Given a network and community partition, it is defined as the observed fraction of internal edges minus the expected fraction of internal edges. The expected fraction of internal edges depends on a null model for generating the expected network structure. As discussed in [102], multiple null models have been used. For our work, we use an Erdös-Rényi (ER) null model as it is the simplest and most transparent choice.

In an ER network, the number of nodes N is fixed. Edges are generated iid with probability $\tilde{p}$ between all possible pairs of nodes, where iid stands for independent and identically distributed. The expected number of internal edges is therefore the total possible number of internal edges times $\tilde{p}$. An undirected network with N nodes has $N(N-1)/2$ possible edges. If the network is split into N/m communities of size m, the network has $(m(m-1)/2)(N/m) = N(m-1)/2$ possible internal edges and $N(N-m)/2$ possible external edges.

Given a network, the probability of edge formation is simply the number of edges present $|E|$ divided by the total possible number of edges, which gives $\tilde{p} = \frac{|E|}{N(N-1)/2}$. Therefore, using an ER model, the expected number of internal edges given a network is $\frac{N(m-1)}{2}\tilde{p} = \frac{N(m-1)}{2}\frac{2|E|}{N(N-1)} = \frac{(m-1)|E|}{N-1}$. The expected fraction of internal edges is $\frac{(m-1)|E|}{N-1}\frac{1}{|E|} = \frac{m-1}{N-1}$.

Returning to modularity, modularity Q is the observed fraction of internal edges

minus the expected fraction of internal edges. The observed fraction of internal edges is simply $\frac{|int|}{|E|}$ where $|int|$ is the number of internal edges. Therefore, $Q = \frac{|int|}{|E|} - \frac{m-1}{N-1}$.

1.19.17 Proof of theorem 5

We prove our condition on modularity Q in equation 1.11 implies $\mathrm{E}[T^*_{NG}] \leq \mathrm{E}[T_D]$. If

$$Q \geq 1 - \frac{m-1}{N-1} - \frac{N(N-m)}{2|E|} \frac{\log\left(1 - \frac{|E|}{N(N-1)/2}\alpha\right)}{\log(1-\alpha)} \tag{1.113}$$

where $Q = \frac{|int|}{|E|} - \frac{m-1}{N-1}$ then

$$Q + \frac{m-1}{N-1} \geq 1 - \frac{N(N-m)}{2|E|} \frac{\log\left(1 - \frac{|E|}{N(N-1)/2}\alpha\right)}{\log(1-\alpha)} \tag{1.114}$$

$$\frac{|int|}{|E|} \geq 1 - \frac{N(N-m)}{2|E|} \frac{\log\left(1 - \frac{|E|}{N(N-1)/2}\alpha\right)}{\log(1-\alpha)} \tag{1.115}$$

$$\left(\frac{|E|}{|E|} - \frac{|int|}{|E|}\right) \frac{2|E|}{N(N-m)} \leq \frac{\log\left(1 - \frac{|E|}{N(N-1)/2}\alpha\right)}{\log(1-\alpha)} \tag{1.116}$$

$$\frac{|ext|}{N(N-m)/2} \leq \frac{\log\left(1 - \frac{|E|}{N(N-1)/2}\alpha\right)}{\log(1-\alpha)} \tag{1.117}$$

where we have used $|E| - |int| = |ext|$ in equation 1.117. Note, the left hand side of equation 1.117 is $\hat{q}$, the empirical probability an external edge exists in the network. Equation 1.117 enforces an upper bound on the probability of external edges, similar to the $q \leq p$ constraint in the SBM model, and provides another way of writing theorem 5.

Before proceeding, we show $v = \frac{|E|}{N(N-1)/2}\alpha$, establishing a connection between infection prevalence v and infection passing probability α. Recall α is set such that the expected number of infected individuals is equal under the models under consideration. The expected number of infected individuals under Dorfman testing is $1 + (N-1)v$. We now derive the expected number of infected individuals in a general network under the epidemic model specified in section 1.3. One node is chosen uniformly at random

from the network to serve as the infected seed. The seed infects each of its immediate neighbors with probability α. Therefore, given node i is the infected seed, the expected number of infected individuals is $1 + k_i\alpha$ where k_i is the number of neighbors of node i. To determine the expected number of infected individuals, we average over all possible infected seeds, yielding $1 + \frac{1}{N}\sum_{i\in\mathcal{N}} k_i\alpha$ where $\mathcal{N}$ is the set of nodes. We can simplify the equation because summing over all the degrees in a network yields two times the number of edges, $\sum_{i\in\mathcal{N}} k_i = 2|E|$. Therefore, the expected number of infected individuals under our epidemic model for a general network is $1 + \frac{2|E|}{N}\alpha$. Setting $1 + (N-1)v = 1 + \frac{2|E|}{N}\alpha$ and solving for v yields $v = \frac{|E|}{N(N-1)/2}\alpha$.

Returning to our condition on $\hat{q}$, we set group size n equal to community size m, yielding

$$\frac{|ext|}{N(N-n)/2} \leq \frac{\log(1-v)}{\log(1-\alpha)} \tag{1.118}$$

We now show $\sum_{i\in\mathcal{N}}\sum_{g\in\mathcal{G}\setminus g_i} n_{g,i} = 2|ext|$ where $\mathcal{G}$ is the set of groups, g_i is the group that contains node i, and $n_{g,i}$ is the number of nodes in group g connected to node i. Given a node i, summing over the number of nodes in group g connected to node i, $n_{g,i}$, for all groups g except the group containing i provides the external degree of node i. The external degree of node i is the number of edges incident to node i that connect to nodes in groups other than node i's group. Formally, $\sum_{g\in\mathcal{G}\setminus g_i} n_{g,i} = k_i^{ext}$ where k_i^{ext} is the external degree of node i. Summing over k_i^{ext} for all nodes i yields $2|ext|$ because we record all external edges in the network twice, $\sum_{i\in\mathcal{N}} k_i^{ext} = 2|ext|$. Therefore, our condition becomes

$$\frac{\sum_{i\in\mathcal{N}}\sum_{g\in\mathcal{G}\setminus g_i} n_{g,i}}{N(N-n)} \leq \frac{\log(1-v)}{\log(1-\alpha)} \tag{1.119}$$

$$\left(\frac{1}{N}\frac{1}{N/n - 1}\sum_{i\in\mathcal{N}}\sum_{g\in\mathcal{G}\setminus g_i} n_{g,i}\right)\log(1-\alpha) \geq n\log(1-v) \tag{1.120}$$

$$(1-\alpha)^{\left(\frac{1}{N}\frac{1}{N/n-1}\sum_{i\in\mathcal{N}}\sum_{g\in\mathcal{G}\setminus g_i} n_{g,i}\right)} \geq (1-v)^n \quad (1.121)$$

$$\frac{1}{N}\frac{1}{N/n-1}\sum_{i\in\mathcal{N}}\sum_{g\in\mathcal{G}\setminus g_i}(1-\alpha)^{n_{g,i}} \geq (1-\alpha)^{\left(\frac{1}{N}\frac{1}{N/n-1}\sum_{i\in\mathcal{N}}\sum_{g\in\mathcal{G}\setminus g_i} n_{g,i}\right)} \geq (1-v)^n \quad (1.122)$$

$$\frac{1}{N}\frac{1}{N/n-1}\sum_{i\in\mathcal{N}}\sum_{g\in\mathcal{G}\setminus g_i}(1-\alpha)^{n_{g,i}} \geq (1-v)^n \quad (1.123)$$

where equation 1.122 uses Jensen's inequality to remove the empirical expectations from the exponent. Finally, we have

$$\frac{1}{N}\frac{1}{N/n-1}\sum_{i\in\mathcal{N}}\sum_{g\in\mathcal{G}\setminus g_i}(1-\alpha)^{n_{g,i}} \geq (1-v)^n \quad (1.124)$$

$$\frac{1}{N}\frac{1}{N/n-1}\sum_{i\in\mathcal{N}}\sum_{g\in\mathcal{G}\setminus g_i}(1-(1-\alpha)^{n_{g,i}}) \leq 1-(1-v)^n \quad (1.125)$$

$$\frac{N}{n}+n\left[1+\frac{1}{N}\sum_{i\in\mathcal{N}}\sum_{g\in\mathcal{G}\setminus g_i}(1-(1-\alpha)^{n_{g,i}})\right] \leq \frac{N}{n}+n\left[1+\left(\frac{N}{n}-1\right)(1-(1-v)^n)\right] \quad (1.126)$$

$$\mathrm{E}[T^*_{NG}] \leq \mathrm{E}[T_D] \quad (1.127)$$

To end the proof, we show the right hand side of the condition in equation 1.11 goes to 0 as $\alpha \to 0$. To do so, we first show

$$\frac{\log\left(1-\frac{|E|}{N(N-1)/2}\alpha\right)}{\log(1-\alpha)} \to \frac{|E|}{N(N-1)/2} \quad (1.128)$$

as $\alpha \to 0$. We use the following bound on logarithmic functions

$$\frac{x-1}{x} < \log(x) < x-1 \quad (1.129)$$

for all $x > 0$ with $x \neq 1$. Let $\tilde{p} = \frac{|E|}{N(N-1)/2}$. Applying the bound to the top and bottom logarithms in equation 1.128 yields

$$\frac{-\tilde{p}\alpha}{1-\tilde{p}\alpha} < \log(1-\tilde{p}\alpha) < -\tilde{p}\alpha \quad (1.130)$$

$$\frac{-\alpha}{1-\alpha} < \log(1-\alpha) < -\alpha \tag{1.131}$$

Taking the reciprocal of inequality 1.131 and multiplying both inequalities by -1 provides

$$\tilde{p}\alpha < -\log(1-\tilde{p}\alpha) < \frac{\tilde{p}\alpha}{1-\tilde{p}\alpha} \tag{1.132}$$

$$\frac{1-\alpha}{\alpha} < \frac{1}{-\log(1-\alpha)} < \frac{1}{\alpha} \tag{1.133}$$

Both inequalities are comprised of positive terms. Combining the inequalities yields

$$\tilde{p}(1-\alpha) < \frac{\log(1-\tilde{p}\alpha)}{\log(1-\alpha)} < \frac{\tilde{p}}{1-\tilde{p}\alpha} \tag{1.134}$$

The upper and lower bounds both go to $\tilde{p}$ as $\alpha \to 0$, confirming the limit in equation 1.128.

Returning to the condition in equation 1.11, the limit of the equation is

$$Q \geq 1 - \frac{m-1}{N-1} - \frac{N(N-m)}{2|E|}\frac{|E|}{N(N-1)/2} \tag{1.135}$$

as $\alpha \to 0$. Simplifying yields

$$Q \geq 1 - \frac{m-1}{N-1} - \frac{N-m}{N-1} = 0 \tag{1.136}$$

□

1.19.18 Remainder correction

In 1.19.3, we derive the expected number of tests used under network grouping. Recall, N denotes the population size, n denotes the group size, and m denotes the community size. 1.19.3 considers the case where n divisible by m or m divisible by n. This assumption ensures communities are kept intact or are split evenly when placed in groups. For example, when $n \geq m$, each group contains several intact communities and, when $n < m$, each community is split into an integer number of groups. The

assumption guarantees each group contains either individuals from the infected seed community or individuals from different communities than the infected seed, but not both.

The divisibility assumption results in a clean equation for the number or tests used under network grouping, as shown in equation 1.3. The equation is transparent, easy to work with, and provides insight into the behavior of network grouping. As a result, we use equation 1.3 in the main text and as the foundation for the main results in our work.

However, when n not divisible by m and m not divisible by n, equation 1.3 is only an approximation, albeit a strong one. Therefore, in this subsection, we derive the expected number of tests used under network grouping when n not divisible by m and m not divisible by n. The expected number of tests used is shown below, followed by its derivation and a discussion.

When $n < m$:

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n\left[1 + \left\lfloor \frac{m}{n} - 1 \right\rfloor p' + \left(\left\lceil \frac{m}{n} - 1 \right\rceil - \left\lfloor \frac{m}{n} - 1 \right\rfloor\right) p'' + \left(\frac{N}{n} - 1 - \left\lceil \frac{m}{n} - 1 \right\rceil\right) q'\right] \tag{1.137}$$

$$p'' = 1 - \left[\frac{m\%n}{m}(1-p\alpha)^n + \left(1 - \frac{m\%n}{m}\right)(1-p\alpha)^{m\%n}(1-q\alpha)^{n-(m\%n)}\right]$$

where $p' = 1-(1-p\alpha)^n$, $q' = 1-(1-q\alpha)^n$, and % is the modulo operator.

When $n \geq m$:

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n\left[1 + \left\lceil \frac{m}{n\%m} - 1 \right\rceil q'' + \left(\frac{N}{n} - 1 - \left\lceil \frac{m}{n\%m} - 1 \right\rceil\right) q'\right] \tag{1.138}$$

$$q'' = 1 - \left[\left(1 - \frac{n\%m}{n}\right)(1-q\alpha)^n + \frac{n\%m}{n}(1-p\alpha)^{n\%m}(1-q\alpha)^{n-(n\%m)}\right]$$

where $q' = 1-(1-q\alpha)^n$, % is the modulo operator, and ${}^{m}/_{(n\%m)} = 0$ if $n\%m = 0$.

Derivation of the $n < m$ case To derive equation 1.137, we note two-stage testing procedures use ${}^{N}/_{n} + nG$ tests where G is the number of positive groups from the first stage of testing. To understand the expected value of G, we first focus on the

groups that contain individuals from the infected seed's community. The infected seed's community contains m individuals. As $m > n$, the community is split into $\lceil m/n \rceil$ groups. We keep communities intact as much as possible, meaning there will be $\lfloor m/n \rfloor$ intact groups that contain only individuals from the infected seed's community and there will be one split group that contains the remainder individuals from the infected seed's community as well as individuals from other communities. As an example, if $n = 5$ and $m = 12$, there will be two intact groups that contain only individuals from the infected seed's community and there will be one split group that contains 2 individuals from the infected seed's community and 3 individuals from different communities.

To determine the probability each group tests positive, we must first consider which group the infected seed is located in. If the infected seed is placed in the split group, that group tests positive with probability one, and the intact groups test positive with probability $(1-p\alpha)^n$. If the seed individual is placed in one of the intact groups, that group tests positive with probability one, the other intact groups tests positive with probability $(1-p\alpha)^n$, and the split group tests positive with probability $1-(1-p\alpha)^{m\%n}(1-q\alpha)^{n-(m\%n)}$, since the split group contains $m\%n$ individuals from the same community as the infected seed and $n-(m\%n)$ individuals from different communities. Here, % denotes the modulo operator. For example, $m\%n = 12\%5 = 2$ means the split community contains 2 individuals from the infected seed's community and 3 individuals from different communities.

The infected seed is placed into the split group with probability $m\%n/m$, as there are $m\%n$ remainder individuals out of the m community members in the infected seed's community. As a result, one group contains the infected seed and tests positive with probability one, $\lfloor m/n - 1 \rfloor$ groups only contain individuals from the infected seed's community and test positive with probability $p' = (1-p\alpha)^n$, and the remaining $\lceil m/n - 1 \rceil - \lfloor m/n - 1 \rfloor$ groups test positive with probability

$$p'' = \frac{m\%n}{m}(1-(1-p\alpha)^n) + \left(1-\frac{m\%n}{m}\right)(1-(1-p\alpha)^{m\%n}(1-q\alpha)^{n-(m\%n)})$$

$$= 1 - \left[\frac{m\%n}{m}(1-p\alpha)^n + \left(1 - \frac{m\%n}{m}\right)(1-p\alpha)^{m\%n}(1-q\alpha)^{n-(m\%n)} \right]$$

as they only contains individuals from the same community as the infected seed with probability ${}^{m\%n}/_{m}$ and they contains individuals from multiple communities with probability $1 - {}^{m\%n}/_{m}$.

The remaining ${}^{N}/_{n} - 1 - \lceil {}^{m}/_{n} - 1 \rceil$ groups contain only individuals from different communities than the infected seed and therefore test positive with probability $q' = 1 - (1-q\alpha)^n$. Putting everything together yields the expected number of tests as shown in equation 1.137.

Derivation of the $n \geq m$ case To derive equation 1.138, we note two-stage testing procedures use ${}^{N}/_{n} + nG$ tests where G is the number of positive groups from the first stage of testing. To understand the expected value of G, we first focus on the groups that contain individuals from the infected seed's community. The infected seed's community contains m individuals. As $n \geq m$ and n not necessarily divisible by m, the community will either be pooled in its entirety into one group or split into several groups. Similar to the $n < m$ case, we keep communities intact as much as possible when pooling. For example, if $N = 80$, $n = 10$ and $m = 8$, each of the ${}^{N}/_{n} = 8$ groups will contain one full community of 8 as well as 2 individuals from the remaining communities.

If the infected seed's community is split, multiple groups will contain individuals from the seed's community. Each group of size n contains $n\%m$ individuals from split communities and $n - (n\%m)$ individuals from intact communities, where % is the modulo (remainder) operator. In the example above, $n\%m = 2$ individuals came from split communities while $n - (n\%m) = 8$ individuals came from an intact community. Therefore, if the infected seed's community of size m is split, individuals from the community will be placed into $\lceil {}^{m}/_{(n\%m)} \rceil$ groups. One of the groups will contain the infected seed and will test positive with probability one. The remaining $\lceil {}^{m}/_{(n\%m)} - 1 \rceil$ groups contain $n\%m$ individuals from the seed's community, each not infected with probability $1 - p\alpha$, and $n - (n\%m)$ individuals from different communities, each not

infected with probability $1 - q\alpha$.

As a last step, we derive the probability the infected seed's community is split. In each of the N/n groups, $n\%m$ individuals come from split communities. Therefore, a total of $(N/n)(n\%m)/m$ communities are split, out of a total N/m communities. As a result, the probability the infected seed's community is split is

$$\frac{(N/n)(n\%m)/m}{N/m} = \frac{n\%m}{n} \tag{1.139}$$

We now summarize our results. One group contains the infected seed and tests positive with probability one. With probability $\frac{n\%m}{n}$, the infected seed's community is split into $\lceil m/(n\%m) - 1 \rceil$ groups, each of which tests positive with probability $(1 - (1 - p\alpha)^{n\%m}(1 - q\alpha)^{n-(n\%m)})$. With probability $1 - \frac{n\%m}{n}$, the infected seed's community remains intact and the $\lceil m/(n\%m) - 1 \rceil$ groups contain only individuals from different communities than the infected seed and test positive with probability $1 - (1 - q\alpha)^n$. Therefore, the probability the $\lceil m/(n\%m) - 1 \rceil$ groups test positive is

$$\begin{aligned} q'' &= \left(1 - \frac{n\%m}{n}\right)(1 - (1 - q\alpha)^n) + \frac{n\%m}{n}(1 - (1 - p\alpha)^{n\%m}(1 - q\alpha)^{n-(n\%m)}) \\ &= 1 - \left[\left(1 - \frac{n\%m}{n}\right)(1 - q\alpha)^n + \frac{n\%m}{n}(1 - p\alpha)^{n\%m}(1 - q\alpha)^{n-(n\%m)}\right] \end{aligned}$$

The remaining $N/n - 1 - \lceil m/(n\%m) - 1 \rceil$ groups contain only individuals from different communities than the infected seed and test positive with probability $q' = 1-(1-q\alpha)^n$. Therefore, the expected number of tests used under network grouping after correcting for split communities and remainder individuals when $n \geq m$ is as shown in equation 1.138.

Discussion As mentioned above, the expected number of tests shown in equation 1.3, which is the equation used in the main text, holds exactly when m is divisible by n or n is divisible by m. When m is not divisible by n and n is not divisible by m, equations 1.137 and 1.138 provide more accurate forms for the expected number of tests used under network grouping. Note, equations 1.137 and 1.138 depend on how remainder

individuals are handled. In the above derivations, we keep communities intact as much as possible when pooling. To visualize an example of the remainder correction, we simulate a network and epidemic, apply network grouping, and determine the number of tests used. Specifically, we simulate a stochastic block model where $N = 100$, $m = 10$, $p = 0.90$, and $q = 0.02$. We simulate an epidemic process on the network following the description in section 1.3 where $v = 0.05$. We pool individuals into groups following the network grouping procedure outlined in section 1.4. We run the simulation 5000 times and average to understand the expected number of tests used under network grouping. The results are shown in figure 1-5. The network grouping line, given by equation 1.3, lines up exactly with the simulated results when m is divisible by n or n is divisible by m. When m is not divisible by n and n is not divisible by m, the remainder correction line, given by equations 1.137 and 1.138, is closer to the simulated results. Note, for large networks, where N is large, the remainder correction has a negligible impact.

The expected number of tests after correcting for split communities and remainder individuals collapses to the expected number of tests used in the main text and shown in equation 1.3 when m is divisible by n or n is divisible by m. When $n < m$ and m is divisible by n, m/n is an integer. Therefore, equation 1.137 becomes

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n\left[1 + \left(\frac{m}{n} - 1\right)p' + \left(\frac{N}{n} - 1 - \left(\frac{m}{n} - 1\right)\right)q'\right] \tag{1.140}$$

which equals equation 1.3 in the $n < m$ case. When $n \geq m$ and n is divisible by m, $n\%m = 0$ and $m/(n\%m) = 0$ by definition. Therefore, equation 1.138 becomes

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n\left[1 + \left(\frac{N}{n} - 1\right)q'\right] \tag{1.141}$$

which equals equation 1.3 in the $n \geq m$ case.

In addition, when $p = q$, the remainder corrected expected number of tests collapses to the expected number of tests under Dorfman testing. When $p = q$, $p\alpha = q\alpha = v$ as shown in 1.19.4. Therefore, $p' = p'' = q' = q'' = v'$ where $v' = 1 - (1 - v)^n$. As a

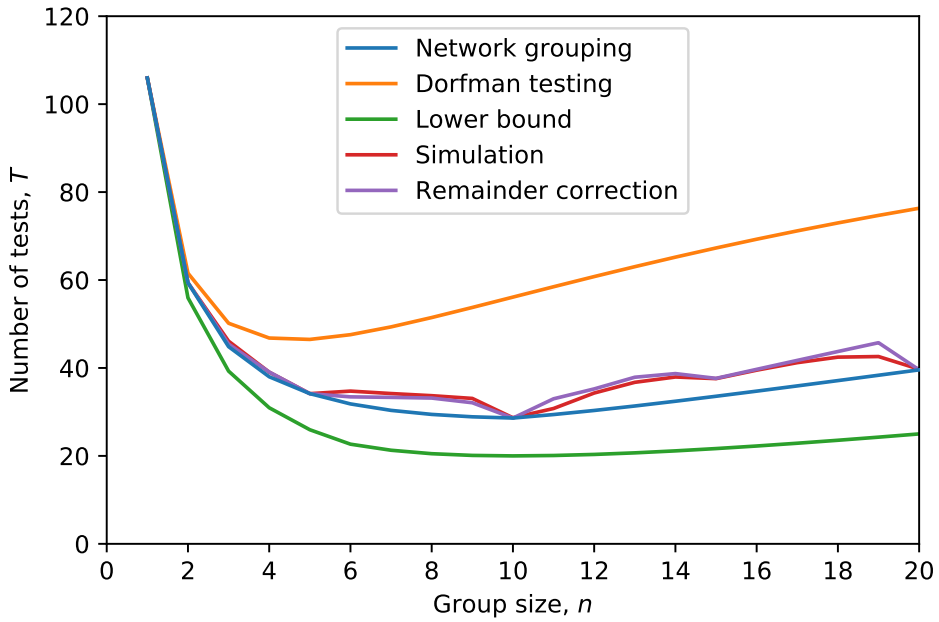


Figure 1-5: Number of tests used to screen a population of $N = 100$ when $v = 0.05$, $m = 10$, $p = 0.90$, and $q = 0.02$. The network grouping line demonstrates the expected number of tests used under network grouping as given in equation 1.3. The Dorfman testing and lower bound lines correspond to the number of tests used under Dorfman testing and the two-stage lower bound respectively, as given in equations 1.1 and 1.2. The simulation line corresponds to the average number of tests used under network grouping when the stochastic block model and epidemic model are simulated and network grouping is applied to the simulated data. The remainder correction line corresponds to the expected number of tests used under network grouping after accounting for split communities and remainder nodes, as given in equations 1.137 and 1.138.

result, both the $n < m$ and $n \geq m$ cases simplify to

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n\left[1 + \left(\frac{N}{n} - 1\right)v'\right] \tag{1.142}$$

which equals the expected number of tests used under Dorfman, as shown in equation 1.1.

Chapter 2

Performance of Group Testing under Imperfect Tests

Abstract

We consider the problem of identifying infected individuals in a population of size N. Group testing provides an approach that uses significantly fewer than N tests when infection prevalence is low. In this chapter, we derive the performance of the most common form of group testing, Dorfman testing, under imperfect tests. We derive the distribution of the number of tests needed, the number of false negatives, and the number of false positives. The full distributions allow for the construction of confidence intervals and provide better guidance for medical practitioners. Acknowledging the flexibility available to practitioners, we allow for different test sensitivity and specificity in the first and second stage of testing. We explicitly model first-stage sensitivity as dependent on the number of samples in each group, which accounts for viral-load dilution.

We have built a dashboard that allows practitioners to analyze the performance of group testing under various parameters. The dashboard can be found at `group-testing.herokuapp.com`. Documentation for the dashboard is provided in section 2.5.

2.1 Introduction

Group testing improves testing capabilities for infectious diseases when resources are limited. Given a population of size N, the standard approach to identify infected individuals is to test all population members individually, which requires N tests. In

the most common form of group testing, called Dorfman testing, individual samples are pooled together into groups of size n for an initial stage of testing. If a group tests negative, all individuals within the group are classified as negative for the disease. If a group tests positive, all individual samples from the group are retested individually to identify the infected members. To illustrate the power of group testing, consider the scenario where $N = 50$ and one individual is infected. If individuals are pooled into groups of size $n = 10$ for an initial stage of testing, one group will test positive and all 10 samples from the group will be retested. The group testing approach uses 15 tests compared to the 50 used under individual testing.

Group testing was introduced by the statistician Robert Dorfman in 1943 to screen for syphilis in the US military [51]. Dorfman's idea was simple but powerful. As a result, group testing has been employed numerous times in the medical field for diseases including influenza, chlamydia, and malaria [52, 72, 73]. Within the US, group testing is used in blood banks and infertility prevention programs where large numbers of individuals are routinely tested [53, 74, 76, 86]. Group testing's efficient use of resources has made it a valuable technique in developing areas. Notably, group testing was used during the early stages of the HIV pandemic in Africa when polymerase chain reaction (PCR) test costs were high [77]. By reducing testing costs and increasing access to diagnostic information, group testing plays an important role in increasing health equity.

Since Dorfman's work in 1943, numerous group testing approaches with strong performance have been introduced [53, 76, 86–90, 92–95]. However, Dorfman testing remains the most common approach to group testing in practice because it is straightforward for labs to implement [52–55]. For a formal comparison of Dorfman testing to individual testing and more advanced approaches, see the previous chapter on network group testing.

In this chapter, we derive the performance of Dorfman testing under perfect and imperfect tests, which result in false negatives and positives. We derive the distribution of the number of tests needed, the number of false negatives, and the number of false positives. The full distributions allow for the construction of confidence intervals

and provide better guidance for medical practitioners. Acknowledging the flexibility available to practitioners, we allow for different test sensitivity and specificity in the first and second stage of testing. We explicitly model first-stage sensitivity as dependent on the number of samples in each group, which accounts for viral-load dilution.

To facilitate the use of group testing, we have built a dashboard that allows practitioners to analyze the performance of group testing under various parameters. The dashboard can be found at `group-testing.herokuapp.com`. The dashboard takes in various input parameters, including population size, infection prevalence, and first and second stage sensitivity and specificity, and returns the number of tests needed, false negatives, and false positives as a function of group size. The outputs can help practitioners design, understand, and implement various group testing programs.

The remainder of this chapter is organized as follows. Section 2.2 describes the performance of group testing under perfect tests. Section 2.3 describes the performance of group testing under imperfect tests, including the number of tests, false negatives, and false positives. The section also includes a discussion of overall sensitivity and specificity, confidence intervals, optimal group size, maximizing the number of individuals tested, and sensitivity as a function of group size. Section 2.4 discusses the impact of infection prevalence on the performance of group testing. Section 2.5 provides documentation for the dashboard. Section 2.6 provides an example of group testing using our formulae and dashboard. Section 2.7 concludes. Derivations are provided in the appendix.

2.2 Perfect tests

To begin, we describe the performance of Dorfman testing under perfect tests, which result in no false negatives or positives. Dorfman testing is a two-stage group testing procedure. Under two-stage testing, a population of size N is split into N/n groups of size n for an initial stage of testing. Let G denote the number of positive groups after the initial stage. In the second stage of testing, all n samples from each positive

group are retested individually. In total, $N/n + nG$ tests are used.

Under Dorfman testing [51], the N individuals are infected independently with probability p. The expected number of infected individuals, I, is simply $\mathrm{E}[I] = Np$. The number of tests needed under Dorfman testing, T, is a random variable that is distributed

$$T \sim \frac{N}{n} + n \cdot \mathrm{Bin}\left(\frac{N}{n}, q\right) \tag{2.1}$$

where $q = 1 - (1-p)^n$. Taking the expectation of equation 2.1 provides the expected number of tests needed under Dorfman testing

$$\mathrm{E}[T] = \frac{N}{n} + Nq \tag{2.2}$$

where $q = 1 - (1-p)^n$. The derivation of T is provided in appendix 2.8.1. When the infection prevalence p is low, Dorfman testing uses significantly fewer than N tests in expectation. As an example, consider the scenario where $N = 1000$ and $p = 0.05$ (5%). If we employ Dorfman testing and a group size of $n = 10$, only 507 tests are needed in expectation to test the entire population, a reduction of nearly 50% compared to the $N = 1000$ tests needed under individual testing.

2.3 Imperfect tests

In this section, we analyze the performance of Dorfman testing under imperfect tests, which result in false negatives and positives. We derive the distribution of the number of tests needed, number of false negatives, and number of false positives.

The performance of a diagnostic test is measured by two parameters: sensitivity and specificity. The sensitivity of a test is the fraction of infected individuals who correctly test positive. Therefore, sensitivity equals one minus the false negative rate. The specificity of a test is the fraction of non-infected individuals who correctly test negative. Therefore, specificity equals one minus the false positive rate. Relating the medical terms to statistical terminology, sensitivity is the power of the test and

specificity is one minus the size of the test.

In our analysis, we allow test sensitivity and specificity to differ between the first stage and second stage of testing. This allows practitioners to use different tests for the first stage, when groups are tested, and the second stage, when individual samples are tested. We denote first-stage sensitivity as $s_{e_1,n}$, second-stage sensitivity as s_{e_2}, first-stage specificity as s_{p_1}, and second-stage specificity as s_{p_2}. We explicitly allow first-stage sensitivity, $s_{e_1,n}$, to depend on the group size n, since pooling samples dilutes the viral load of an infected sample and can therefore reduce test sensitivity. We leave $s_{e_1,n}$, s_{e_2}, s_{p_1}, and s_{p_2} as exogenous parameters for medical practitioners to input.

In section 2.6, we provide an example of a university testing its undergraduate population for COVID-19. The example applies the quantities and concepts discussed in this section.

2.3.1 Number of tests needed

The number of tests needed under Dorfman testing and imperfect tests differs from the number under perfect tests because infected groups may incorrectly test negative and non-infected groups may incorrectly test positive. The number of tests needed under Dorfman testing and imperfect tests, T', is distributed

$$T' \sim \frac{N}{n} + n \cdot \text{Bin}\left(\frac{N}{n}, q'\right) \tag{2.3}$$

where $q' = s_{e_1,n}(1-(1-p)^n) + (1-s_{p_1})(1-p)^n$. Taking the expectation of equation 2.3 provides the expected number of tests needed under Dorfman testing and imperfect tests

$$\text{E}[T'] = \frac{N}{n} + Nq' \tag{2.4}$$

where $q' = s_{e_1,n}(1-(1-p)^n) + (1-s_{p_1})(1-p)^n$. The derivation of T' is provided in appendix 2.8.2. The number of tests needed under Dorfman testing and imperfect

tests can be either less than or greater than the number of tests needed under perfect tests, depending on test sensitivity and specificity.

2.3.2 Number of false negatives

Under imperfect tests, some infected individuals incorrectly test negative. It is instructive to compare the number of false negatives under Dorfman testing and individual testing. Recall, Np individuals are infected in expectation. Under individual testing, each infected individual tests negative incorrectly with probability $1 - s_{e_2}$. We use second-stage sensitivity when discussing individual testing because individual samples are tested in the second stage. Therefore, there are $Np(1 - s_{e_2})$ false negatives in expectation under individual testing.

The number of false negatives under Dorfman testing and imperfect tests, FN, is distributed

$$FN \sim \text{Bin}\left(N, p(1 - s_{e_1,n}s_{e_2})\right) \tag{2.5}$$

Taking the expectation of equation 2.5 provides the expected number of false negatives under Dorfman testing and imperfect tests

$$\text{E}[FN] = Np(1 - s_{e_1,n}s_{e_2}) \tag{2.6}$$

The derivation of FN is provided in appendix 2.8.3.

Comparing the two approaches, Dorfman testing results in more false negatives than individual testing in expectation (since $1 - s_{e_1,n}s_{e_2} \geq 1 - s_{e_2}$). The reason is simple: Dorfman testing uses two stages of testing so infected individuals must test positive correctly twice, compared to just once under individual testing. In addition, first-stage sensitivity often decreases with group size n. Although individual testing results in fewer false negatives, the significant reduction in the number of tests needed under group testing provides several economical options to reduce false negatives. Practitioners can use more sensitive tests when conducting group testing than when

conducting individual testing, or can screen the population more frequently under group testing than individual testing (e.g., twice a week rather than once a week).

2.3.3 Number of false positives

Under imperfect tests, some non-infected individuals incorrectly test positive. Again, we can compare Dorfman testing to individual testing. Recall, $N(1-p)$ individuals are not infected in expectation. Under individual testing, each non-infected individual tests positive incorrectly with probability $1-s_{p_2}$, where we use second-stage sensitivity for our comparison since individual samples are tested in the second stage. Therefore, there are $N(1-p)(1-s_{p_2})$ false positives in expectation under individual testing.

The number of false positives under Dorfman testing and imperfect tests, FP, is distributed

$$FP \sim \text{Bin}\left(N, (1-p)(1-s_{p_2})q'_{n-1}\right) \tag{2.7}$$

where $q'_{n-1} = s_{e_1,n}(1-(1-p)^{n-1}) + (1-s_{p_1})(1-p)^{n-1}$. Taking the expectation of equation 2.7 provides the expected number of false positives under Dorfman testing and imperfect tests

$$\text{E}[FP] = N(1-p)(1-s_{p_2})q'_{n-1} \tag{2.8}$$

where $q'_{n-1} = s_{e_1,n}(1-(1-p)^{n-1}) + (1-s_{p_1})(1-p)^{n-1}$. The derivation of FP is provided in appendix 2.8.4.

Dorfman testing results in fewer false positives than individual testing (since $q'_{n-1} \leq 1$). Again, the reason is simple: for an individual to test positive incorrectly under Dorfman testing, they must test positive incorrectly during the second stage. Fewer individuals are tested during the second stage than during individual testing.

2.3.4 Overall sensitivity and specificity

As we have discussed, diagnostic tests are graded by their sensitivity and specificity. Testing procedures can also be graded by their overall sensitivity and specificity. Sensitivity measures the fraction of infected individuals that correctly test positive. Specificity measures the fraction of non-infected individuals that correctly test negative.

The overall sensitivity of Dorfman testing, s_{e_D}, is one minus the overall false negative rate. Likewise, the overall specificity of Dorfman testing, s_{p_D}, is one minus the overall false positive rate. Therefore,

$$s_{e_D} = 1 - \frac{\mathrm{E}[FN]}{\mathrm{E}[I]} \quad = s_{e_1,n} s_{e_2} \tag{2.9}$$

$$s_{p_D} = 1 - \frac{\mathrm{E}[FP]}{N - \mathrm{E}[I]} = 1 - (1 - s_{p_2}) q'_{n-1} \tag{2.10}$$

where $\mathrm{E}[I] = Np$ and $q'_{n-1} = s_{e_1,n}(1 - (1-p)^{n-1}) + (1 - s_{p_1})(1-p)^{n-1}$. Compared to individual testing, Dorfman testing has lower sensitivity (as $s_{e_1,n} s_{e_2} \leq s_{e_2}$) and higher specificity (as $1 - (1 - s_{p_2}) q'_{n-1} \geq s_{p_2}$ since $q'_{n-1} \leq 1$), mirroring our discussion of false negatives and positives.

The positive predictive value (PPV) of a test or testing approach represents the fraction of positive results that correspond to true positives. Likewise, the negative predictive value (NPV) represents the fraction of negative results that correspond to true negatives. PPV and NPV both depend on the prevalence of the disease as well as sensitivity and specificity. The PPV and NPV of Dorfman testing are

$$PPV_D = \frac{p \cdot s_{e_D}}{p \cdot s_{e_D} + (1-p)(1 - s_{p_D})} \tag{2.11}$$

$$NPV_D = \frac{(1-p) s_{p_D}}{p(1 - s_{e_D}) + (1-p) s_{p_D}} \tag{2.12}$$

2.3.5 Confidence intervals

Group testing work often focuses on the expected values of the quantities of interest. However, relying on expectations can result in serious issues in practice. Expectations define values that can be expected on average. When implementing group testing,

observed values will be either larger or smaller than expectation.

Consider the following scenario: a university is implementing group testing to screen its undergraduate population for COVID-19. The university has an undergraduate population of 5000 students and infection prevalence of 2%. Under Dorfman testing, the university needs 1415 tests in expectation to screen its population using groups of size 10 and assuming perfect tests. However, the university may actually need more tests to account for the randomness in group testing. In fact, there is a 49% probability the university will need more than 1415 tests. Better guidance would be "with 95% confidence, the university needs x tests or less to screen its population of 5000." To provide such guidance, confidence intervals are needed.

Because we have defined the full distributions of the quantities of interests, confidence intervals can easily be derived. The number of tests needed is an affine transformation of a binomial random variable. The number of false negatives and false positives are binomial random variables.

To provide guidance to the university in our motivating example, we derive the upper confidence bound for the number of tests needed. The CDF of T', the number of tests needed under imperfect tests, is

$$\begin{aligned} \mathrm{P}(T' \leq z) &= \mathrm{P}\left(\frac{N}{n} + n \cdot \mathrm{Bin}\left(\frac{N}{n}, q'\right) \leq z\right) && (2.13) \\ &= \mathrm{P}\left(\mathrm{Bin}\left(\frac{N}{n}, q'\right) \leq \frac{z}{n} - \frac{N}{n^2}\right) && (2.14) \\ &= F_{\mathrm{Bin}\left(\frac{N}{n}, q'\right)}\left(\frac{z}{n} - \frac{N}{n^2}\right) && (2.15) \end{aligned}$$

where $F_X(t)$ is the CDF of random variable X evaluated at t. To build a confidence interval for T', we simply evaluate the quantile function (also called the inverse CDF or percent point function) at our desired probability level. In our scenario, we are interested in the number of tests z such that T' is less than z with $(1-\alpha)\%$ probability, where α is our significance level. Therefore, we are looking for z such that

$$F_{\mathrm{Bin}\left(\frac{N}{n}, q'\right)}\left(\frac{z}{n} - \frac{N}{n^2}\right) = 1 - \alpha \tag{2.16}$$

Using the quantile function, we have

$$Q_{\text{Bin}\left(\frac{N}{n},q'\right)}(1-\alpha) = \frac{z}{n} - \frac{N}{n^2} \tag{2.17}$$

where $Q_X(t)$ is the quantile function of random variable X evaluated at t. Solving for z provides

$$z = \frac{N}{n} + n \cdot Q_{\text{Bin}\left(\frac{N}{n},q'\right)}(1-\alpha) \tag{2.18}$$

The binomial distribution does not have a closed-form quantile function. However, equation 2.18 can easily be evaluated computationally. Using Python, this is implemented as

```
from scipy.stats import binom
qprime = se1[n] * (1-(1-p)**n) + (1-sp1) * (1-p)**n
x = binom.ppf(1 - alpha, N/n, qprime)
z = N/n + n*x
```

Returning to our motivating example, we can use the above derivation to analyze the number of tests the university needs to screen its population. Recall, the university has an undergraduate population of size $N = 5000$, an infection prevalence of $p = 0.02$, perfect tests, and is using a group size of $n = 10$. Using equation 2.18 and our code snippet, the university needs 1560 tests or less to screen its population with 95% confidence.

2.3.6 Optimal group size

In this subsection, we discuss the optimal group size, which minimizes the number of tests needed to screen a population. In practice, group size is often determined by medical considerations, such as test sensitivity, or logistical factors. If practitioners are considering a certain group size or range of sizes, they can use our dashboard to analyze the performance of group testing for the group sizes under consideration. In addition, it is easy to visually determine the optimal group size using our dashboard

for various population and test parameters. For a full example using our dashboard, see section 2.6.

It is instructive to consider the optimal group size analytically. The optimal group size that minimizes the expected number of tests needed can be derived in closed form. Taking the derivative of $\mathrm{E}[T']$ in equation 2.4, setting the derivative equal to 0, and solving for n provides the optimal group size n^*

$$n^* = \frac{2}{\ln(1-p)} \cdot W\left[-\frac{1}{2}\left(\frac{\ln(1-p)}{1-s_{e_1,n}-s_{p_1}}\right)^{1/2}\right] \tag{2.19}$$

where $W[x]$ is the Lambert W function (product log). Note, we do not treat $s_{e_1,n}$ as a function of n for this derivation. Interestingly, the optimal group size does not depend on the population size N. As an example, if we are testing a population with infection prevalence $p = 0.02$ where first-stage sensitivity is a constant $s_{e_1,n} = 0.90$ and first-stage specificity is $s_{p_1} = 0.95$, the optimal group size is $n^* = 8.3$ using equation 2.19. For a population of size $N = 5000$, 1509 tests are needed in expectation to screen the population using groups of size 8, a 70% reduction in tests.

Equation 2.19 provides the optimal group size that minimizes the number of tests needed in expectation. The optimal group size that minimizes the number of tests needed with high probability cannot be derived analytically, because the binomial distribution does not have a closed-form quantile function. However, the optimal group size that minimizes the number of tests needed (either in expectation or with high probability) can easily be determined using our dashboard.

2.3.7 Maximizing the number of people tested

In this subsection, we discuss the maximum number of people that can be tested given a fixed number of tests. The discussion is useful for resource-constrained institutions who may not have enough tests to screen their population. The derivation highlights another benefit of knowing the full distribution of the number of tests needed.

The optimal group size that maximizes the number of screened individuals for a given number of tests in expectation is equal to n^* in equation 2.19, the optimal

group size that minimizes the number of tests needed for a given population size. This is because n^* does not depends on N, the population size. The derivation can be double checked by solving for N in equation 2.4, taking the derivative with respect to n, setting the derivative equal to 0, and solving for n. Solving for N in equation 2.4 yields the population size as a function of the expected number of tests and group size

$$N = \frac{n \cdot \mathrm{E}[T']}{1 + nq'} \tag{2.20}$$

where $q' = s_{e_1,n}(1 - (1-p)^n) + (1 - s_{p_1})(1-p)^n$. Plugging n^* from equation 2.19 into equation 2.20 provides the maximum number of individuals that can be screened given a fixed number of tests in expectation.

If an institution only has a fixed number of tests, they will likely run out of tests if they try to screen the maximum number of individuals provided by equation 2.20 and n^*. This is because equation 2.20 and n^* provide the maximum given a fixed number of tests *in expectation*. Instead, better guidance would be "an institution with t tests can screen up to N individuals using group testing with 95% confidence." To provide such guidance, we can employ the distribution of the number of tests.

We cannot derive the maximum population size that can be screened with high probability analytically because the binomial distribution does not have a closed-form quantile function. However, the maximum population size can easily be found by searching over a range of population sizes using equation 2.16. In subsection 2.3.5 on confidence intervals, equation 2.16 provides the CDF of the number of tests needed. Using equation 2.16, we set z equal to our given number of tests, n equal to the optimal group size n^* from equation 2.19, and use infection prevalence and test accuracy to determine q'. We then evaluate the left-hand side for a range of population N values. Each evaluation returns the probability that group testing uses fewer than z tests. For small values of N, the probability will be 1. For large values of N, the probability will fall below 1 and fall to zero. We can then choose an N such that the probability is equal to our desired confidence level.

For example, consider the scenario of a university testing its student body for

COVID-19. The university is resource constrained and only has 1000 tests available. Infection prevalence is $p = 0.02$, first-stage sensitivity is a constant $s_{e_1,n} = 0.90$, and first-stage specificity is $s_{p_1} = 0.95$. Using equation 2.19, we determine the optimal group size is $n^* = 8$. Using equation 2.16, we find the university can test $N = 1000$ students with 1000 tests with probability 1. However, the probability the university can test $N = 5000$ students with 1000 tests is 0. Iterating over values of N, we find the university can test $N = 2999$ students with probability 0.95.

2.3.8 Sensitivity as a function of group size

In this subsection, we discuss various ways to model first-stage sensitivity. We denote first-stage as $s_{e_1,n}$ and explicitly allow it to depend on the group size n, since pooling samples dilutes the viral load of an infected sample and can therefore reduce test sensitivity.

At the beginning of this section, we mention that we leave $s_{e_1,n}$ as an exogenous parameter for medical practitioners to input. Medical practitioners can evaluate the sensitivity of their tests in a controlled laboratory environment and determine the sensitivity of their tests for a range of group sizes. They can then explicitly input test sensitivity $s_{e_1,n}$ for each value of n and compute the performance of group testing using the formulae we have introduced here. However, evaluating sensitivity for all values of n is demanding. In many scenarios, practitioners can interpolate test sensitivity for a range of group sizes.

In our dashboard, we model test sensitivity as linearly decreasing with group size. We ask for test sensitivity when evaluating individual samples and test sensitivity when evaluating pools of size *input*, where *input* is entered by the user. We linearly interpolate between the two sensitivity values and linearly extrapolate to determine sensitivity for all group sizes. Sensitivity is floored at 0. An example is provided in section 2.6.

Linearly interpolation is simple to understand and implement, and captures decreasing sensitivity with group size. However, many other schemes can be used. For example, a logistic curve starting at a high value for individual samples and falling

to 0 for large group sizes provides another approach to model sensitivity. Another possibility is a step function, which models first-stage sensitivity as constant until some threshold group size, after which it steps down to a new value (or 0). Step functions may be appropriate in practice, since many practitioners evaluate the limit of detection of their tests. The limit of detection is the maximum dilution (minimum viral load) that can still be detected using a specific test.

By explicitly allowing first-stage sensitivity to depend on the group size n, we allow for a variety of methods to model $s_{e1,n}$. All of these approaches take viral-load dilution into account.

2.4 Impact of infection prevalence

In this section, we discuss the impact of infection prevalence on our quantities of interest. Group testing works best when infection prevalence is low (less than 20%). In many practical scenarios, infection prevalence is low enough to use group testing. For example, COVID infection prevalence in the US estimated by the COVID Tracking Project and Johns Hopkins University has been below 10% for the majority of the pandemic [96].

Figure 2-1 provides insight into the impact of infection prevalence on the performance of group testing. We see the number of tests needed under group testing varies for different values of p. Group testing works very well when p is low ($p = 0.001$, $p = 0.01$) and requires far fewer tests than individual testing. However, when p is large ($p = 0.1$), group testing does not provide as dramatic of an improvement. As we observe in the figure, the number of tests needed to screen a population using group testing increases with infection prevalence.

Group testing works well when p is low because many groups will not contain an infected individual and will test negative. These groups of individuals will be cleared with only one test. However, when p is large, groups are more likely to contain an infected individual and test positive. The individuals that make up positive groups then need to be retested in the second stage, diminishing the advantage of group

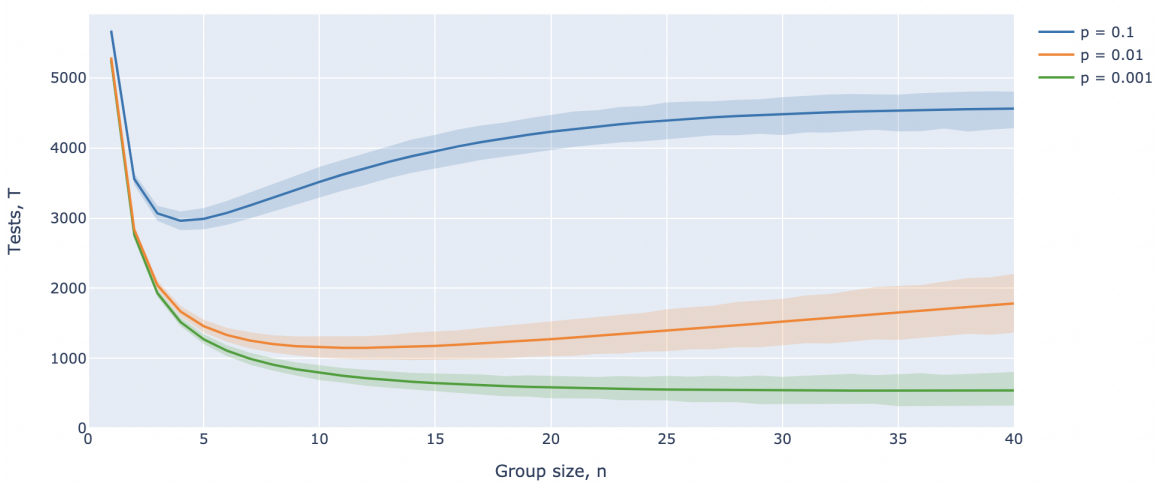


Figure 2-1: Number of tests needed for various values of infection prevalence. The figure displays the number of tests needed to screen a population of size 5000 as a function of group size. We set $s_{e_1,n} = 0.90$ for all n and $s_{p_1} = 0.95$. The three curves represent the number of tests needed for different values of infection prevalence p. Solid lines displays the expected number of tests and shaded regions provide 95% confidence intervals.

testing.

It is also instructive to analyze the impact of infection prevalence on optimal group size. Figure 2-2 displays the optimal group size as a function of p. We see optimal group size decreases with p. The reason is simple: when p is low, large groups can be used because the probability they contain an infected individual is low. As a result, large groups of individuals can be cleared with one test each. However, when p is large, large groups will contain infected individuals with high probability, and individuals will have to be retested. As a result, it is more efficient to use smaller groups when p is large.

2.5 Dashboard documentation

We have built a dashboard that allows practitioners to analyze the performance of group testing under various parameters. The dashboard can be found at `group-testing.herokuapp.com`. The dashboard applies the most common form of group testing, Dorfman testing, which is a two-stage group testing procedure where individuals are

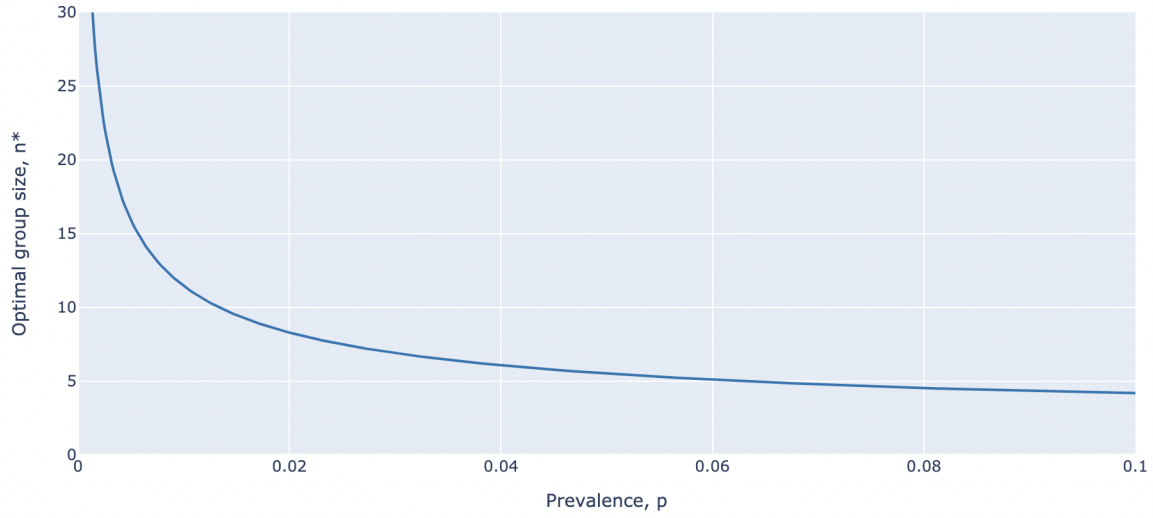


Figure 2-2: Optimal group size as a function of infection prevalence. The figure displays the optimal group size n^*, which minimizes the number of tests needed to screen a population, for various values of infection prevalence p. We set $s_{e_1,n} = 0.90$ for all n and $s_{p_1} = 0.95$.

grouped randomly.

The dashboard takes in various input parameters, including population size, infection prevalence, and first and second stage sensitivity and specificity, and returns the number of tests needed, number of false negatives, and number of false positives as a function of group size. Acknowledging the flexibility available to practitioners, we allow for different test sensitivity and specificity in the first and second stage of testing. We also provide confidence intervals in the outputs. The outputs, which use the quantities derived in this chapter, can help practitioners design, understand, and implement various group testing programs. In the following section, we provide an example using our dashboard.

We explicitly model first-stage sensitivity as dependent on the number of samples in each group, which accounts for viral-load dilution. Specifically, the dashboard models first-stage sensitivity as decreasing linearly with group size. Users input the sensitivity of the first-stage tests when testing individual samples. They then input the sensitivity of the first-stage tests when testing a pooled sample containing *input* individuals. The dashboard then linearly interpolates between the two points and linearly extrapolates for all other group sizes. First-stage sensitivity is floored at 0.

As a result, practitioners can evaluate test sensitivity as a function of group size in a laboratory setting, and input the correct sensitivities in the dashboard.

The dashboard inputs are

- **Population:** The size of the population (number of individuals) to be tested using group testing. This corresponds to the variable N in this chapter.
- **Prevalence (%):** The infection prevalence in percent. For example, set equal to 2 if 2% of the population is infected with the disease. Often, infection prevalence is an estimate based on preliminary testing results or nearby areas. This corresponds to the variable p in this chapter.
- **Max group size:** The maximum group size to be used during group testing. The output plots will plot results for group sizes ranging from 1 to max group size.
- **Specificity, 1st stage (%):** Specificity of the tests used in the first stage of testing. Specificity is the fraction of non-infected individuals that correctly test negative. It is equal to one minus the false positive rate. Pooled samples are tested in the first stage. Enter as a percentage. For example, enter 95 if the specificity of the test is 95%. This corresponds to the variable s_{p_1} in this chapter.
- **Specificity, 2nd stage (%):** Specificity of the tests used in the second stage of testing. Specificity is the fraction of non-infected individuals that correctly test negative. It is equal to one minus the false positive rate. Individual samples are tested in the second stage. Enter as a percentage. For example, enter 95 if the specificity of the test is 95%. This corresponds to the variable s_{p_2} in this chapter.
- **Sensitivity, 1st stage, n=1 (%):** Sensitivity of the tests used in the first stage of testing when group size equals 1. In other words, sensitivity of the tests used in the first stage when testing individual samples. Sensitivity is the

fraction of infected individuals that correctly test positive. It is equal to one minus the false negative rate. Pooled samples are tested in the first stage. Enter as a percentage. For example, enter 95 if the sensitivity of the test is 95%. This corresponds to the variable $s_{e_1,n=1}$ in this chapter.

- **Sens., 1st stage, input n:** The dashboard models first-stage sensitivity as decreasing linearly with group size (see explanation at the start of this section). The dashboard linearly interpolates between first-stage sensitivity when $n = 1$, which is entered in the previous field, and first-stage sensitivity when $n = input$, where *input* is entered here.

- **Sens., 1st stage, n=input (%):** Sensitivity of the tests used in the first stage of testing when group size equals *input*, the input value from the previous field. Sensitivity is the fraction of infected individuals that correctly test positive. It is equal to one minus the false negative rate. Pooled samples are tested in the first stage. This corresponds to the variable $s_{e_1,n=input}$ in this chapter. The dashboard linearly interpolates between first-stage sensitivity when $n = 1$, which is entered in a previous field, and first-stage sensitivity when $n = input$, where *input* is entered in the previous field and $s_{e_1,n=input}$ is entered here. For example, if first-stage sensitivity is equal to 0.90 when testing individual samples, but equals 0.80 when testing groups of size 30, we set $s_{e_1,n=1} = 0.90$, $input = 30$, and $s_{e_1,n=input} = 0.80$.

- **Sens., 2nd stage (%):** Sensitivity of the tests used in the second stage of testing. Sensitivity is the fraction of infected individuals that correctly test positive. It is equal to one minus the false negative rate. Individual samples are tested in the second stage. Enter as a percentage. For example, enter 95 if the sensitivity of the test is 95%. This corresponds to the variable s_{e_2} in this chapter.

- **Confidence interval (%):** Size of the confidence interval to display in the output plots. For example, if 95 is entered, 95% confidence intervals will be

shown in the output plots. As a specific example, if 95 is entered, the confidence interval in the "number of tests needed" plot shows the interval of the number of tests needed to screen the population with 95% confidence. See subsection 2.3.5 for a longer discussion of confidence intervals.

The dashboard outputs are

- **Plot of the number of tests needed as a function of group size:** The number of tests needed under Dorfman testing and imperfect tests as a function of group size. The y-axis records the number of tests needed to screen the population. The x-axis records the group size (number of samples pooled into each group in the first stage of testing). The solid line reports the expected number of tests needed, which corresponds to equation 2.4 in this chapter. Confidence intervals are displayed as shaded regions and are derived from equation 2.3 in this chapter. Hovering over the plot with your cursor will provide the exact values for given group sizes.

- **Plot of the number of false negatives as a function of group size:** The number of false negatives under Dorfman testing and imperfect tests as a function of group size. The y-axis records the number of false negatives. The x-axis records the group size (number of samples pooled into each group in the first stage of testing). The solid line reports the expected number of false negatives, which corresponds to equation 2.6 in this chapter. Confidence intervals are displayed as shaded regions and are derived from equation 2.5 in this chapter. Hovering over the plot with your cursor will provide the exact values for given group sizes.

- **Plot of the number of false positives as a function of group size:** The number of false positives under Dorfman testing and imperfect tests as a function of group size. The y-axis records the number of false positives. The x-axis records the group size (number of samples pooled into each group in the first stage of testing). The solid line reports the expected number of false positives,

which corresponds to equation 2.8 in this chapter. Confidence intervals are displayed as shaded regions and are derived from equation 2.7 in this chapter. Hovering over the plot with your cursor will provide the exact values for given group sizes.

2.6 Example

In this section, we provide an example of the performance of group testing under specific parameters. For our example, we consider the scenario of a university testing its undergraduate population of $N = 5000$ students for COVID-19. The university estimates infection prevalence to be $p = 0.05$, either from prior testing or by observing the prevalence in the surrounding community. We note $p = 0.05$ is in line with COVID infection prevalence in the US estimated by the COVID Tracking Project and Johns Hopkins University [96].

The university contracts a nearby laboratory to conduct PCR tests to screen the population. The lab uses a highly specific test in the first stage with $s_{p_1} = 0.98$. In the second stage, the lab uses a less specific test with $s_{p_2} = 0.95$. The first-stage test is also highly sensitive with $s_{e_1,n=1} = 0.90$. The lab knows the sensitivity of the test on individual samples from previous experience and from FDA guidance [104]. It analyzes the sensitivity of the test under sample dilution. After running lab experiments, it estimates the sensitivity of the test to be 0.75 when $n = 20$. Therefore, $s_{e_1,n=20} = 0.75$. The sensitivity of the second-stage test is $s_{e_2} = 0.88$. Note, these specificities and sensitivities are in line with true COVID test performance [105].

To analyze the performance of group testing under these parameters, we use our dashboard (`group-testing.herokuapp.com`). The dashboard visualizes the quantities derived in this chapter. We input our parameters, which are defined above, into the dashboard (Figure 2-3). The dashboard models first-stage sensitivity as decreasing linearly with group size. As a result, it linearly interpolates between $s_{e_1,n=1} = 0.90$ and $s_{e_1,n=20} = 0.75$ and linearly extrapolates to provide first-stage sensitivities for all values of n. First-stage sensitivity is floored at 0.

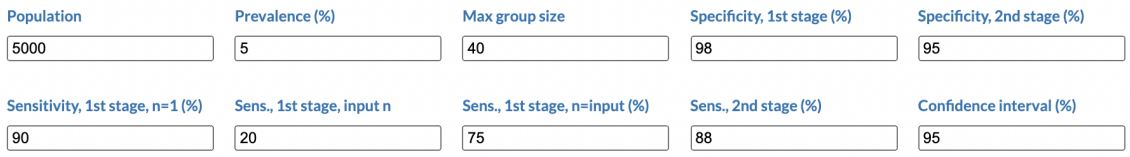
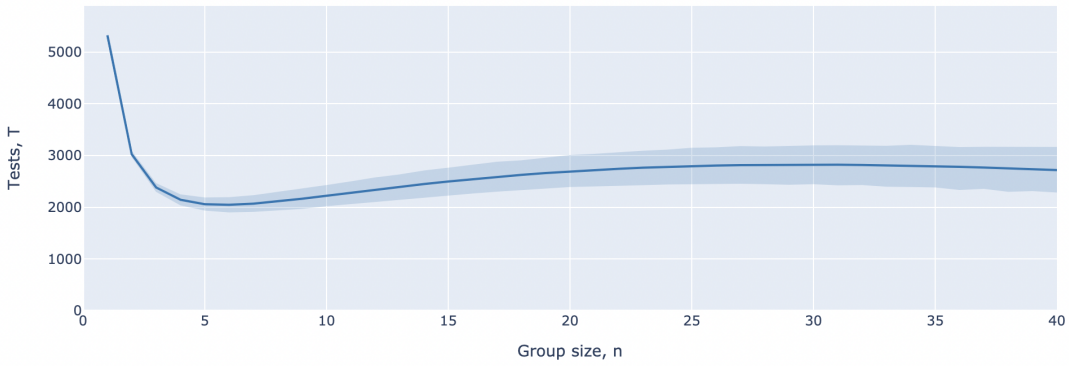


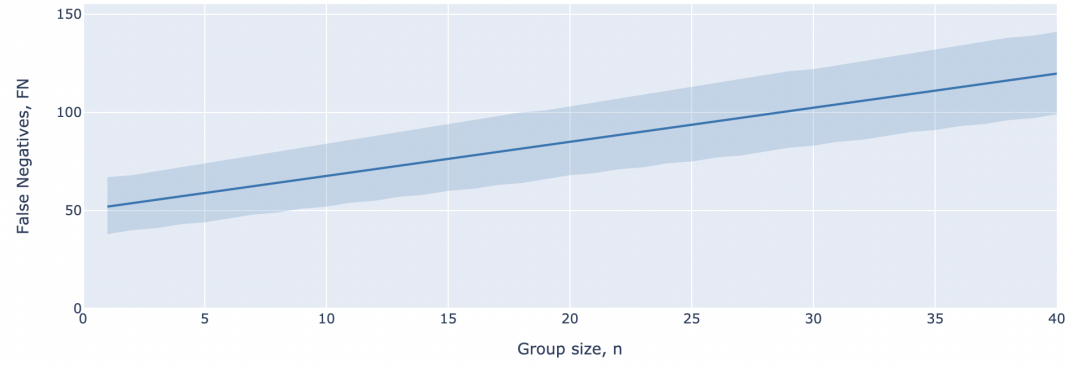
Figure 2-3: Dashboard parameter inputs for our university testing example. The dashboard analyzes the performance of group testing under specific parameters. In the dashboard input fields, we specify the population size, infection prevalence, and test specificity and sensitivity. Details regarding specificity and sensitivity are provided in the main text. We leave max group size as its default value, meaning the dashboard will provide the performance of group testing for groups of size 1 to 40. We leave confidence interval as its default value, meaning 95% confidence intervals will be displayed.

Figure 2-4a provides the number of tests needed under group testing in this scenario. Under individual testing, the university would need 5000 tests to screen its population. However, the university would only need 2047 tests in expectation to screen its entire population using group testing and groups of size 6. With 95% probability, the university would need between 1,901 and 2,195 tests using groups of size 6. Requiring less than 2200 tests to screen 5000 students is a reduction of 56% compared to individual testing.

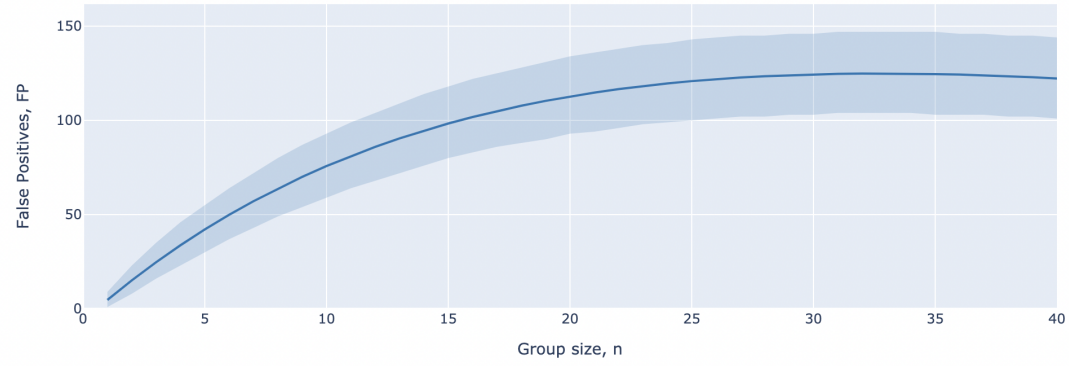
From the figure, we see groups of size 6 minimize the number of tests needed in expectation. As mentioned in subsection 2.3.6, it is easy to determine the optimal group size that minimizes the number of tests needed (either in expectation or with high probability) using our dashboard. The number of tests needed first decreases with group size n, highlighting the power of group testing, because many groups of individuals are classified as negative for the disease using only one test. After reach a minimum at $n = 6$, the number of tests starts increasing with group size. This is because large groups are more likely to include infected individuals and test positive, which means the individual samples must be retested in the second stage. Interestingly, we see the number of tests decreases again when n becomes very large. This is because first-stage sensitivity becomes very low for large n and large pools begin testing negative incorrectly, which means the individual samples are not retested in the second stage.



(a)



(b)



(c)

Figure 2-4: Performance of group testing for our university testing example. The subfigures display the **(a)** number of tests needed, **(b)** number of false negatives, and **(c)** number of false positives, all as a function of group size. The solid line provides the expected value and the shaded region is a 95% confidence interval.

Viewing Figure 2-4b, we see the number of false negatives increases steadily with group size. False negatives increase with group size because infected samples are diluted in the first stage. If the university uses groups of size 6, they will have 61 false negatives in expectation. With 95% confidence, they will have between 46 and 76 false negatives. Under individual testing, there would be $Np(1 - s_{e_2}) = 30$ false negatives in expectation. As discussed in section 2.3.2, Dorfman testing results in more false negatives than individual testing. As mentioned, the significant reduction in the number of tests needed under group testing provides several economical options to reduce false negatives. The university may choose to use more sensitive tests when conducting group testing than when conducting individual testing, or may choose to screen the population more frequently under group testing than individual testing (e.g., twice a week rather than once a week).

Figure 2-4c provides the number of false positives under our scenario. If the university uses groups of size 6, there will be 50 false positives in expectation. With 95% confidence, false positives will be between 37 and 64. Under individual testing, there would be $N(1 - p)(1 - s_{p_2}) = 238$ false positives in expectation. As discussed in section 2.3.3, Dorfman testing results in fewer false positives than individual testing. From the figure, we see the number of false positives first increases with group size. This is because the number of samples tested in the second stage increases with group size as large groups are more likely to contain an infected individual and require retesting in the second stage. However, the number of false positives then begins decreasing for large group sizes. This is because first-stage sensitivity becomes very low for large group sizes and many groups test negative incorrectly in the first stage. As a result, few samples are tested in the second stage.

The quantities we have provided in this chapter, along with the dashboard, provide guidance to the hypothetical university on the performance of a group testing approach to screen their undergraduate population. Specifically, we provide information on the number of tests needed, the number of false negatives, and the number of false positives for their unique circumstances and test parameters. The quantities derived in this chapter include several advances that reflect the realities of group testing

in practice. First, we allow first and second stage test sensitivity and specificity to differ. Second, we explicitly model first-stage sensitivity as dependent on group size, which accounts for viral-load dilution. Third, we provide the full distributions of the quantities of interest. As we have seen, full distributions allow for the construction of confidence intervals and provide better guidance to practitioners.

2.7 Conclusions

Our work extends our understanding of the most common approach to group testing, Dorfman testing. We derive the performance of Dorfman testing under conditions faced by medical practitioners. Specifically, we derive the number of tests needed, the number of false positives, and the number of false negatives under group testing when tests are imperfect, tests have varying sensitivities and specificities, and samples are diluted. We provide the full distributions for these quantities of interest, which allow for the construction of confidence intervals and provide better guidance for medical practitioners. In addition, we have built a dashboard that allows practitioners to analyze the performance of group testing under various parameters.

Our work provides a theoretical foundation for the group testing approaches used in practice. Our derivations and discussion help practitioners design, understand, and implement group testing programs in order to efficiently identify infected individuals. By providing analytical results, we expand the understanding of group testing, its performance in medical clinics and testing centers, and its potential for large-scale surveillance testing of infectious diseases.

2.8 Appendix

2.8.1 Derivation of the number of tests needed under perfect tests

Under Dorfman testing, a population of size N is split into N/n groups of size n for an initial stage of testing. Let G denote the number of positive groups after the initial stage. In the second stage of testing, all n samples from each positive group are retested individually. In total, $N/n + nG$ tests are used. G is a random variable. Because individuals are infected independently, the N/n groups are positive independently with some probability q. As a result, G is distributed $\text{Bin}(N/n, q)$.

The probability q is derived as follows. All N individuals are infected with probability p and not infected with probability $1 - p$. The probability that all n individuals in a group are not infected is $(1 - p)^n$. The probability that at least one individual in the group is infected, and therefore the group tests positive, is $q = 1 - (1 - p)^n$. Putting everything together, the number of tests needed under Dorfman testing is distributed

$$T \sim \frac{N}{n} + n \cdot \text{Bin}\left(\frac{N}{n}, q\right) \tag{2.21}$$

where $q = 1 - (1 - p)^n$.

2.8.2 Derivation of the number of tests needed under imperfect tests

The number of tests used by two-stage testing procedures is $N/n + nG$ where N is the population size, n is the group size, and G is the number of groups that test positive in the first stage of testing. G is a random variable. Under perfect tests, the number of tests needed under Dorfman testing is provided in the previous subsection. Under imperfect tests, the distribution of G changes because truly positive groups may test negative incorrectly and truly negative groups may test positive incorrectly.

Individuals are infected independently and test results are independent. Therefore, the $^N/_n$ groups test positive independently with some probability q'. As a result, G is distributed $\text{Bin}(^N/_n, q')$.

The $^N/_n$ groups are infected (contain at least one infected individual) independently with probability $1-(1-p)^n$, as derived in the previous subsection. The groups test positive if they are infected and test positive correctly, which occurs with probability $s_{e_1,n}(1-(1-p)^n)$, or if they are not infected and test positive incorrectly, which occurs with probability $(1-s_{p_1})(1-p)^n$. Therefore, $q' = s_{e_1,n}(1-(1-p)^n)+(1-s_{p_1})(1-p)^n$. Putting everything together,

$$T' \sim \frac{N}{n} + n \cdot \text{Bin}\left(\frac{N}{n}, q'\right) \tag{2.22}$$

where $q' = s_{e_1,n}(1-(1-p)^n) + (1-s_{p_1})(1-p)^n$.

2.8.3 Derivation of the number of false negatives

Each of the N individuals is infected and tests falsely negative independently with some probability j. Therefore, the number of false negatives, FN, is distributed $\text{Bin}(N, j)$. Each of the N individuals tests falsely negative if (1) they are truly positive and their group tests falsely negative in the first stage, or if (2) they are truly positive, their group tests positive correctly in the first stage, and then their sample tests negative incorrectly in the second stage. Scenario (1) occurs with probability $p(1-s_{e_1,n})$. Scenario (2) occurs with probability $ps_{e_1,n}(1-s_{e_2})$. Therefore, each individual is infected and tests falsely negative with probability $j = p(1-s_{e_1,n}) + ps_{e_1,n}(1-s_{e_2})$. This simplifies to $j = p(1-s_{e_1,n}s_{e_2})$.

The number of false negatives under Dorfman testing and imperfect tests is therefore distributed

$$FN \sim \text{Bin}\left(N, p(1-s_{e_1,n}s_{e_2})\right) \tag{2.23}$$

2.8.4 Derivation of the number of false positives

Each of the N individuals is not infected and tests falsely positive independently with some probability k. Therefore, the number of false positives, FP, is distributed $\text{Bin}(N, k)$. The N individuals test falsely positive if they are truly negative, with probability $1 - p$, their group tests positive in the first stage, with some probability q'_{n-1}, and they test falsely positive in the second stage, with probability $1 - s_{p_2}$. Since the individuals in question are truly negative, their group tests positive in the first stage if at least one of the remaining $n - 1$ individuals in the group is truly positive and the group tests positive correctly, with probability $s_{e_1,n}(1 - (1 - p)^{n-1})$, or if the remaining $n - 1$ individuals in the group are truly negative and the group tests positive incorrectly, with probability $(1 - s_{p_1})(1 - p)^{n-1}$. As a result, $q'_{n-1} = s_{e_1,n}(1 - (1 - p)^{n-1}) + (1 - s_{p_1})(1 - p)^{n-1}$. Therefore, the N individuals and not infected and test positive incorrectly with probability $k = (1 - p)(1 - s_{p_2})q'_{n-1}$.

The number of false positives under Dorfman testing and imperfect tests is therefore distributed

$$FN \sim \text{Bin}\left(N, (1 - p)(1 - s_{p_2})q'_{n-1}\right) \tag{2.24}$$

where $q'_{n-1} = s_{e_1,n}(1 - (1 - p)^{n-1}) + (1 - s_{p_1})(1 - p)^{n-1}$. Note, under our notation for q'_{n-1}, the sensitivity $s_{e_1,n}$ depends on the original group size n.

2.9 Transition from epidemic spread to information diffusion

In the first half of this thesis, we have studied epidemic spread in social networks. We have considered the scenario of an infectious disease spreading from individual to individual through a population. In our work, we have designed and analyzed testing approaches to identify infected individuals.

Testing for infectious diseases is a key component of controlling their spread. Once identified, infected individuals can receive treatment and can be quarantined from the rest of the population, improving their health outcomes and hindering the spread of the disease. Combined with vaccination and other non-pharmaceutical interventions like masking, testing has played a key role in combatting numerous epidemics, including the current COVID pandemic.

Normally, infected individuals in a population of size N are identified by testing each person individually, which uses N tests. However, group testing provides an approach to screen the entire population using significantly fewer than N tests when infection prevalence is low. In our work, we apply and extend group testing. We derive its performance under the conditions faced by medical practitioners and provide guidance for its implementation in practice.

We also improve group testing by utilizing social network information. We make the simple observation that communicable diseases spread from person to person through underlying social networks. As a result, the position of an individual in a social network affects their infection probability. We use social network structure to intelligently group individuals and further reduce the number of tests needed under group testing.

Group testing, including our method of network group testing, allow for the efficient screening of large populations. Compared to individual testing, group testing saves time, money, and other scarce resources like chemical reagents. As a result, it provides a powerful tool for combatting epidemic spread.

In the second half of this thesis, we transition to studying information diffusion in

social networks. Both epidemic spread and information diffusion are classical examples of network diffusion processes. In these processes, something spreads from node to node through a network. During an epidemic, an infectious disease spreads from person to person while during information diffusion, information, news, rumors, or gossip spreads through the social network.

From a mathematical standpoint, information spread is often modeled using the same tools as epidemic spread. One or more individuals has or is "infected" with information at the start. The individuals then spread the information to their neighbors.

In the next two chapters, we theoretically and empirically study information diffusion. We begin by introducing an approach to identify information cascades in network data. In observational network data, it can be difficult to distinguish between large, meaningful cascades and the small, common branches that form during normal periods. We introduce a test statistic that compares observed average branch size to expected branch size during normal periods, which allows us to quantify the probability that a cascade has occurred.

We apply our test statistic to call detail records from Yemen. Our approach allows us to 1) add inference and significance results to observed branches, and 2) detect anomalous periods based on branch size. We study the calling cascades that form after violent events during the Yemeni Revolution. Calling cascades are consequential, as they allow information to spread deeply and quickly through the population.

We then empirically study information diffusion around violent events, focusing on drone strikes. Using a dataset of over 12 billion CDRs, we study the social network effects of 74 drone strikes in Yemen between 2010 and 2012. We quantitatively and systematically analyze the impact of strikes on civilians and their communities. As societies are intrinsically networked systems, we use a social network approach in our analysis.

We study the communication response of civilians and look for information diffusion after strikes. As mentioned, information diffusion is extremely consequential in conflict areas. Diffusion facilitates the spread of information, opinions, and emotions regarding

strikes through the population. Strikes, which are unpopular and sometimes result in civilian casualties, have the potential to shift civilian sentiments and loyalties, which can affect the trajectories of modern conflicts. In addition, we study physical diffusion after strikes, analyzing the extent of displacement and fleeing.

This thesis reinforces the importance of network diffusion processes, specifically epidemic spread and information diffusion. As we have seen, understanding epidemic spread allows us to design better approaches to control the spread of a disease. In the coming chapters, we study the role diffusion plays in spreading information and news through social networks.

Chapter 3

Tests for Network Cascades via Branching Processes

Abstract

In the previous two chapters, we study epidemic spread. Epidemic spread in social networks behaves similarly to information diffusion. In information diffusion, information, news, or gossip spreads from individual to individual though a population, much like an infection. While information exchange in social networks is common, information cascades, in which a large number of individuals quickly contact each other, are rare. These cascade often correspond to consequential events such as the spreading of news following a violent event, the retweeting of viral fake news, or the spreading of gossip through a social clique. In this chapter, we focus on identifying information cascades in social networks.

We consider a network setting where branches form under the null of normal periods and larger branches form under the alternative. Our goal is to distinguish abnormally large branches, which we term cascades, from the common branches formed under the null. Call detail records provide the motivating example, as large call branches form after disruptive events, yet call branches also form during normal periods. We introduce a formal statistical testing framework to distinguish between branches formed under the null and alternative based on expected branch size. After defining the characteristics of edge formation under the null, we derive the expected size and variance of branches using the machinery of branching processes. We introduce a test statistic that compares observed average branch size to expected branch size under the null, which allows us to quantify the probability a cascade has occurred. Our test statistic is semiparametric, consistent, and asymptotically distributed standard normal under the null. Using call detail records from Yemen, we find a significant calling cascade occurred after the Presidential Palace was bombed in 2011. Lastly, we employ our statistic for event detection and successfully detect key violent events

during the 2011 Yemeni Revolution.

3.1 Introduction

Diffusion through networked systems corresponds to numerous consequential processes such as information spread through societies, epidemics in populations, failure propagation in energy grids, and systemic risk in banking networks. As a result, the emergence of diffusion has been studied and documented in several domains in empirical network science [2–5, 19, 34–38]. These processes are often described as cascades since they involve nodes contacting or "infecting" their neighboring nodes, who in turn infect their neighbors [56–59]. However, in many network settings, small scale diffusion regularly emerges during normal periods from normal behavior. Only a small number of large cascades occur, motivating the need to distinguish large, meaningful branch formation from the smaller, common branches formed during normal periods. As a motivating example, call detail records, which record calls between phone users, have been used to demonstrate the emergence of calling cascades after disruptive events [60–66]. However, as individuals make calls during normal periods as well, even when no event has occurred, call branches also form during normal periods. In this paper, our goal is to distinguish abnormally large branches, which we term cascades, from the common branches formed by normal activity. To this end, we introduce a formal statistical testing framework that distinguishes cascades in networked systems from the common branches formed during normal periods.

With a formal testing framework in place, we introduce a test statistic that compares observed branch size to expected branch size under the null of normal periods. We define a semiparametric model of independent and identically distributed edge formation under the null. This model allows us to derive the expected size and variance of branches under the null using the machinery of branching processes [67–69]. The test statistic we introduce is semiparametric, consistent, and asymptotically distributed standard normal under the null. A formal statistic allows us to quantify the probability observed branches were formed under the null of normal periods.

Therefore, a rejection of the null indicates the observed branches are significantly large and correspond to cascades. As an empirical application, we apply the test statistic to call detail records from Yemen. Our test statistic allows us to 1) add inference and significance results to observed branches and 2) detect anomalous periods based on branch size. We find a significant calling cascade occurred after the Presidential Palace was bombed in 2011. The emergence of a cascade implies information regarding the bombing spread quickly and deeply through the underlying social network. In addition, we identify three periods with significantly large call branches originating in Sana'a, Yemen's capital, during March 2011. The detected periods line up with key violent events during the 2011 Yemeni Revolution. Crucially, by adding inference to observed branch structures, our test statistic provides significance and confidence levels to our empirical findings.

The remainder of the chapter is organized as follows. Section 3.2 defines branches, introduces the model of edge formation under the null, and states the formal hypothesis testing framework. Section 3.3 reports our results, including the size and variance of branches under the null, the proposed test statistic, and empirical results using call detail records. Section 3.4 concludes.

3.2 Model

In this section, we define the model governing branch formation during normal periods. The normal period dynamics constitute our null model and allow us to derive the expected size and variance of branches under the null. We conclude the section by introducing the hypothesis testing framework, which characterizes expected branch size under the null and alternative.

3.2.1 Branch formation

The branches we consider correspond to classical branching processes and are defined as follows. A branch begins with a single origin node, constituting the zeroth generation of the branch. The origin nodes comprise the set G_0. Each origin node forms a

random number of edges to new nodes over a period of duration t. Let $X_i(t)$ be the non-negative integer-valued random variable defining the number of edges formed by node i during a period of duration t. We refer to the nodes that G_0 connect to as "contacted by G_0". The nodes contacted by G_0 form the first generation of their branches and comprise the set G_1. After being contacted, each node in G_1 then forms a random number of edges to new nodes, again governed by $X_i(t)$. These new nodes form the second generation of their branches and comprise the set G_2. The branches continue to grow in this manner until they reach a generation of nodes that fail to form any edges during the period of duration t. We define B_i as the size of the branch originating at node i, where size is defined as the total number of nodes in the branch. Figure 3-1 provides an example of branches and their respective random variable realizations.

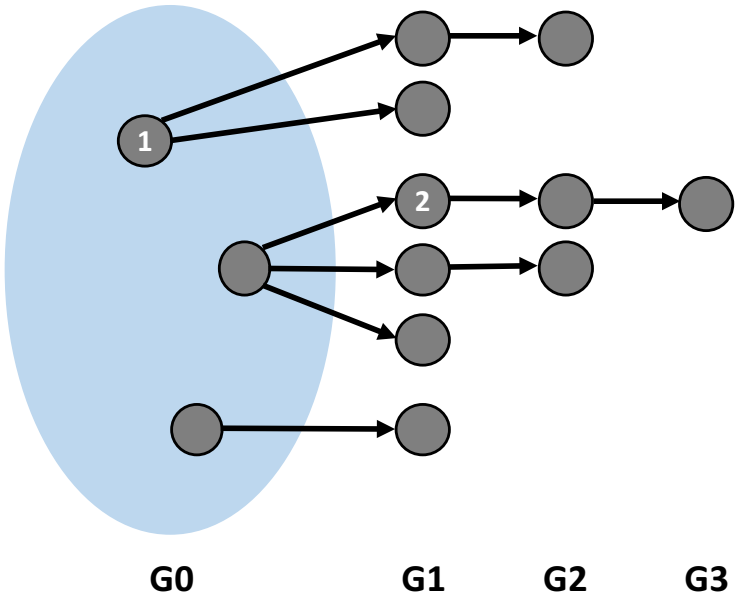


Figure 3-1: Example of branches. G_0 origin nodes contact first generation G_1 nodes who proceed to contact second generation G_2 nodes, and so on. B_1, the size of the branch originating at node 1, is 4. X_2, the number of nodes contacted by node 2, is 1.

Assumptions under the null (A1-4). Under the null of normal periods, we assume

1. $X_i(t)$ is independent and identically distributed (iid) for all i and for all non-overlapping periods of duration t.

2. The second moment of $X_i(t)$ is finite for all i and for all finite periods of duration t.

3. The expected value of $X_i(t)$ equals 0 for periods of duration $t = 0$ and is monotonic increasing in t.

4. Nodes form edges to new nodes only. Specifically, nodes in the nth generation of a branch do not form edges back to nodes in previous generations, to other nodes in the same generation, or to nodes in other branches.

The first assumption states, under the null, edge formation is independent. During normal periods, if a node is contacted, it does not prompt the node to contact more nodes. The second assumption guarantees the variance and expected value of $X_i(t)$ are finite. We define $\gamma := \mathrm{E}[X_i(t)]$ and $\sigma^2 := \mathrm{Var}(X_i(t))$, where we suppress dependence on the period duration t. The third assumption reasonably states that nodes cannot form edges if they have no time to do so and, as nodes have more time to form edges, they form more edges on average. The final assumption guarantees branch sizes are independent as it ensures branches originating at different nodes do not merge, which allows us to use the machinery of branching processes. Alternatively, we can drop the final assumption if we instead explicitly define $X_i(t)$ to be the number of edges formed by node i to new nodes only.

Our model is semiparametric, as we specify restrictions on the moments of $X_i(t)$ but do not impose a specific distribution. This provides significant generality and flexibility to our framework and to the test statistic we introduce in the following section. Several familiar parametric models fall within the framework introduced above.

Example 1 (Poisson process model). Consider the setup where, under the null, every node has its own independent Poisson arrival process, identically governed by rate parameter λ. Every time a node receives an arrival, it forms an edge to a new node. Therefore, the number of edges formed by node i during a period of duration t, $X_i(t)$, is distributed Poisson(λt). The expected value and variance of $X_i(t)$ both equal λt. This formulation satisfies Assumptions 1-4.

Translating the model introduced above to our call detail record (CDR) motivating example, nodes correspond to individuals and edges correspond to phone calls between them. When analyzing the call branches formed after an event, individuals proximal to the event location at the time of the event comprise G_0. The individuals that G_0 contact within a period of duration t after the event form G_1. The individuals that G_1 contact within a period of duration t after being contacted by G_0 form G_2, and so on.

3.2.2 Hypothesis testing framework

Our goal is to distinguish abnormally large branches, which we term cascades, from the common branches formed under the null. Under the null dynamics we have just introduced where edge formation is iid, branches have expected size $\mathrm{E}[B_i] := \mu_0$. In the following section, we solve for μ_0 explicitly.

Under the alternative in our formulation, branches have a larger expected size. Formally, our testing framework is

$$\mathrm{H}_0 : \mathrm{E}[B_i] = \mu_0$$
$$\mathrm{H}_1 : \mathrm{E}[B_i] > \mu_0$$

This setup allows us to formally test for the emergence of cascades. Specifically, a test statistic allows us to compare observed branches to their expected size under the null and quantify the probability the observed branches were formed under the null of normal behavior. A rejection of the null indicates the observed branches are significantly larger than those formed during normal period dynamics. Therefore, we label branches that lead to a rejection of the null as cascades.

We purposefully do not specify the dynamics of edge formation under the alternative in order to leave our framework as general as possible. Several possible dynamics would result in larger branches than those formed under iid edge formation. As an example, edge formation may be dependent, where a contacted node proceeds to contact more nodes. This corresponds to an intuitive notion of cascades where nodes respond to being contacted or infected by passing on information or contagion. Alternatively

or additionally, the rate of edge formation may be larger under the alternative. For example, individuals may make more phone calls after disruptive events. Both setups would result in larger branch sizes.

Translating the framework to CDRs, small call branches form during normal periods as individuals make calls under normal behavior. In abnormal periods, such as after disruptive events, larger call branches form as iid call behavior breaks down and individuals spread information to their contacts, forming calling cascades.

3.3 Results

In this section, we first derive the expected size and variance of branches under the null dynamics. We then introduce a test statistic that compares observed average branch size to expected size under the null, which allows us to determine whether a cascade has occurred. We conclude the section with an empirical application using call detail records from Yemen. The proofs in this section can be skipped without loss of continuity.

3.3.1 Size and variance of branches under the null

In the hypothesis testing framework introduced in the previous section, branches have expected size $\mathrm{E}[B_i] = \mu_0$ under the null dynamics of normal periods. Assumptions A1-4, which hold under the null, provide enough structure to derive μ_0 explicitly. We derive both the expected size and variance of branches under the null of iid edge formation using ideas from branching processes. Recall, $X_i(t)$ is the number of edges formed by node i during a period of duration t, $\mathrm{E}[X_i(t)] = \gamma$, and $\mathrm{Var}(X_i(t)) = \sigma^2$.

Lemma 1. *Under the null and Assumptions A1-4, there exists a duration length $t = T$ such that $E[X_i(T)] = \gamma < 1$. Fixing $t = T$, we have*

$$\mathrm{E}[B_i] = \mu_0 = \frac{1}{1-\gamma} \tag{3.1}$$

$$\mathrm{Var}(B_i) = \frac{\sigma^2}{(1-\gamma)^3} \tag{3.2}$$

Proof. We derive both the mean and variance of branch size under the null and Assumptions A1-4. By Assumption 3, $\mathrm{E}[X_i(t)] = 0$ when $t = 0$ and is monotonic increasing in t. Therefore, there exists a $t = T$ such that $\mathrm{E}[X_i(T)] = \gamma < 1$. We fix $t = T$ for the rest of the proof and suppress dependence on t moving forward.

Let Z_n be the number of nodes in the nth generation of a branch and B be the size of the branch, where size is defined as the total number of nodes in the branch. We focus on a single branch and suppress dependence on the origin node. We begin with a single origin node and therefore $Z_0 = 1$. By our definitions,

$$B = \sum_{n=0}^{\infty} Z_n \tag{3.3}$$

$$Z_n = \sum_{i=1}^{Z_{n-1}} X_i \tag{3.4}$$

as the size of a branch is equal to the number of nodes across all generations and the number of nodes in the nth generation is the total number of nodes contacted by the previous generation.

Under iid edge formation, our branches are classical branching processes. The expected size of branches is a common result in the branching process literature [67–69]. The expected number of nodes in the nth generation of a branch is given by

$$\mathrm{E}[Z_n] = \mathrm{E}[\,\mathrm{E}[Z_n|Z_{n-1}]\,] = \mathrm{E}[\gamma Z_{n-1}] = ... = \gamma^n \tag{3.5}$$

As $\gamma < 1$, the expected size of a branch is then

$$\mathrm{E}[B] = \mathrm{E}[\sum_{n=0}^{\infty} Z_n] = \sum_{n=0}^{\infty} \mathrm{E}[Z_n] = \sum_{n=0}^{\infty} \gamma^n = \frac{1}{1-\gamma} \tag{3.6}$$

Fubini's theorem formally justifies the exchange of an infinite sum and expectation in (3.6), as Z_n is positive, expectation is a Lebesgue integral with respect to a probability measure, and infinite summation is a Lebesgue integral with respect to a discrete measure. $\gamma < 1$ is required for the infinite summation to converge in (3.6).

As an aside, the branches we consider here are called subcritical since they have

finite size with probability 1 [68]. Note the expected size of branches is strictly finite by (3.6) as $\gamma < 1$. Therefore, the probability of an infinite size branch is 0.

We now solve for the variance of B.

$$\begin{aligned} \mathrm{Var}(B) &= \mathrm{E}[B^2] - (\mathrm{E}[B])^2 \\ &= \mathrm{E}[\Big(\sum_{n=0}^{\infty} Z_n\Big)^2] - \Big(\frac{1}{1-\gamma}\Big)^2 \end{aligned} \tag{3.7}$$

Expanding the squared summation of Z_n yields

$$\mathrm{E}[\Big(\sum_{n=0}^{\infty} Z_n\Big)^2] = \sum_{n=0}^{\infty} \mathrm{E}[Z_n^2] + 2\sum_{k=1}^{\infty}\sum_{j=0}^{\infty} \mathrm{E}[Z_j Z_{j+k}] \tag{3.8}$$

where we again use Fubini's theorem to swap expectations and infinite summations. The expansion of the infinite summation holds as Z_n is positive and $\mathrm{E}[\sum_{n=0}^{\infty} Z_n]$ is finite, by (3.6).

We solve for the first term in (3.8) by computing the variance of Z_n. The variance of the number of nodes in the nth generation of a branch is another common result in branching processes [68, 69]. We extend these results to solve for the variance of the entire branch size B.

The variance of a sum with a random number of random terms is given by

$$\mathrm{Var}(\sum_{i=1}^{Z_{n-1}} X_i) = \sigma^2 \mathrm{E}[Z_{n-1}] + \gamma^2 \mathrm{Var}(Z_{n-1}) \tag{3.9}$$

when X_i are iid with finite second moments and Z_{n-1} is a non-negative integer-valued random variable independent of all X_i. Therefore,

$$\mathrm{Var}(Z_n) = \sigma^2 \gamma^{n-1} + \gamma^2 \mathrm{Var}(Z_{n-1}) \tag{3.10}$$

Using the initial condition $\mathrm{Var}(Z_0) = 0$, the solution to this recursion is

$$\mathrm{Var}(Z_n) = \frac{\sigma^2}{\gamma(1-\gamma)}\gamma^n - \frac{\sigma^2}{\gamma(1-\gamma)}\gamma^{2n} \tag{3.11}$$

which gives the variance of the size of the nth generation. Therefore, the second moment of Z_n is

$$\begin{aligned}\mathrm{E}[Z_n^2] &= \mathrm{Var}(Z_n) + (\mathrm{E}[Z_n])^2 \\ &= \frac{\sigma^2}{\gamma(1-\gamma)}\gamma^n - \frac{\sigma^2}{\gamma(1-\gamma)}\gamma^{2n} + \gamma^{2n} \end{aligned} \tag{3.12}$$

All that is left is to solve for the last term in (3.8), $\mathrm{E}[Z_j Z_{j+k}]$. Using the law of iterated expectations several times,

$$\begin{aligned}\mathrm{E}[Z_j Z_{j+k}] &= \mathrm{E}[\ \mathrm{E}[Z_j Z_{j+k}|Z_j]\ ] \\ &= \mathrm{E}[\ Z_j\ \mathrm{E}[Z_{j+k}|Z_j]\ ] \\ &= \mathrm{E}[\ Z_j\ \mathrm{E}[\ \mathrm{E}[Z_{j+k}|Z_{j+1}]\ |Z_j]\ ] \\ = \ldots &= \mathrm{E}[\ Z_j\ \mathrm{E}[\ \mathrm{E}[\ \ldots\ \mathrm{E}[Z_{j+k}|Z_{j+k-1}]\ \ldots\ |Z_{j+1}]\ |Z_j]\ ] \\ &= \mathrm{E}[\ Z_j\ \mathrm{E}[\ \mathrm{E}[\ \ldots\ \gamma Z_{j+k-1}\ \ldots\ |Z_{j+1}]\ |Z_j]\ ] \\ &= \mathrm{E}[\ Z_j\ \mathrm{E}[\gamma^{k-1} Z_{j+1}|Z_j]\ ] \\ &= \gamma^k\ \mathrm{E}[Z_j^2] \end{aligned} \tag{3.13}$$

Note expressions of the form $\mathrm{E}[\ \mathrm{E}[Z_{j+k}|Z_{j+1}]\ |Z_j]$ collapse to $\mathrm{E}[Z_{j+k}|Z_j]$ since the σ-field generated by Z_j is a sub σ-field of the σ-field generated by Z_{j+1}. To see this simply, note the information available at step $j+1$ is the σ-field generated by all X_i random variables in generations 0 to $j+1$, which is a superset of the information available at step j, given by the σ-field generated by all X_i in generations 0 to j.

Plugging (3.12) and (3.13) into (3.8) and simplifying yields

$$\begin{aligned}\mathbb{E}[\Big(\sum_{n=0}^{\infty} Z_n\Big)^2] =& \sum_{n=0}^{\infty}\left(\frac{\sigma^2}{\gamma(1-\gamma)}\gamma^n - \frac{\sigma^2}{\gamma(1-\gamma)}\gamma^{2n} + \gamma^{2n}\right) \\ &+2\sum_{k=1}^{\infty}\sum_{j=0}^{\infty}\gamma^k\left(\frac{\sigma^2}{\gamma(1-\gamma)}\gamma^j - \frac{\sigma^2}{\gamma(1-\gamma)}\gamma^{2j} + \gamma^{2j}\right) \\ =& \frac{\sigma^2-\gamma+1}{(\gamma-1)^2(\gamma+1)} + \frac{2\left(\gamma^2-\gamma\sigma^2-\gamma\right)}{(\gamma-1)^3(\gamma+1)}\end{aligned}$$

$$= \frac{\gamma - \sigma^2 - 1}{(\gamma - 1)^3} \tag{3.14}$$

Plugging (3.14) into (3.7) and simplifying yields the variance of branch size.

$$\mathrm{Var}(B) = \frac{\sigma^2}{(1-\gamma)^3} \tag{3.15}$$

□

With expected branch size and variance under the null derived, we are ready to introduce the test statistic.

3.3.2 The test statistic

Our goal is to distinguish cascades, abnormally large branches, from the common branches formed under the null. To this end, we introduce a test statistic that compares observed average branch size to expected branch size under the null, accounting for sample variability.

Definition 2. *Define the test statistic W as*

$$W = \frac{\hat{\mu} - \mu_0}{se(\hat{\mu})} \tag{3.16}$$

where μ_0 is the expected size of branches under the null, $\hat{\mu}$ is an estimator of μ_0, and $se(\hat{\mu})$ is the standard error of the estimator $\hat{\mu}$. Define $\hat{\mu}$ as the average observed branch size, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} B_i$, where the sum is taken over all n nodes in G_0, the set of zeroth generation origin nodes.

W is a Wald test statistic, which measures the linear distance between an estimator and its null value in standard deviation units. Intuitively, if the observed branch sizes are much larger than their expected size under the null, accounting for sampling variability, we reject the null that the branches were formed during normal periods. Wald tests, along with likelihood ratio and Lagrange multiplier tests, make up the classical trinity of test statistics in econometrics [106]. Unlike its counterparts, however,

Wald tests do not require knowledge of the likelihood function underlying the data generating process. Using the results of Lemma 1, we can fully specify W and its distribution under the null.

Theorem 7. *Under the null and Assumptions 1-4, the test statistic W becomes*

$$W = \frac{\hat{\mu} - \mu_0}{se(\hat{\mu})} = \sqrt{n}\left(\frac{\frac{1}{n}\sum_{i=1}^{n} B_i - \frac{1}{1-\gamma}}{\frac{\sigma}{(1-\gamma)^{3/2}}}\right) \tag{3.17}$$

W is consistent and asymptotically distributed standard normal as $n \to \infty$.

Proof. $\hat{\mu}$ is defined in Definition 1 and μ_0 is given in Lemma 1. Under the null, the estimator $\hat{\mu}$ is a sample average of iid random variables. The standard error of $\hat{\mu}$ is then

$$\begin{aligned} se(\hat{\mu}) &= [\ \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} B_i)\ ]^{1/2} = [\ \frac{1}{n}\mathrm{Var}(B_i)\ ]^{1/2} \\ &= \frac{1}{\sqrt{n}}\ \frac{\sigma}{(1-\gamma)^{3/2}} \end{aligned} \tag{3.18}$$

where the variance of B_i is given in Lemma 1.

Under the null, W is a demeaned average of iid random variables with finite mean and variance, scaled by $\sqrt{n}$. Therefore, the classical Lindeberg-Levy central limit theorem applies and W converges in distribution to a standard normal random variable, $N(0,1)$, as the number of branches, n, goes to infinity.

W is a consistent test since it is based on a consistent estimator. By the law of large numbers, $\hat{\mu}$ is a consistent estimator of μ_0 under the null and $\hat{\mu} \to \mu_0$ in probability as $n \to \infty$. Under the alternative, where $\mathrm{E}[B_i] > \mu_0$, $\hat{\mu} \not\to \mu_0$ in probability as $n \to \infty$. Since the linear distance $\hat{\mu} - \mu_0$ is scaled up by $\sqrt{n}$ in W, W grows arbitrarily large under the alternative as $n \to \infty$ and the probability of correctly rejecting the null, given by $\mathrm{P}(|W| > c)$ for any finite critical value c, goes to 1. □

Determining the distribution of the test statistic under the null is necessary to control the size of the test, the probability of falsely rejecting the null hypothesis. The test statistic is consistent, which is a desirable, albeit common, property meaning the

power of the test, the probability of correctly rejecting the null hypothesis when the alternative is true, goes to 1 as n, the number of branches, goes to infinity. In addition, W is semiparametric, as we do not impose a specific distribution on X_i, the number of edges formed by a node, or B_i, the branch size. Instead, we only specify restrictions on the first two moments of X_i, which adds considerable generality to the statistic.

The downside of this generality is we are only able to derive the distribution of W asymptotically. However, asymptotic distributions provided by central limit theorems work well in practice [107]. By the Berry-Esseen theorem, W converges to a standard normal distribution at rate $n^{-1/2}$ [108]. Imposing a few additional assumptions, we provide rules of thumb regarding the number of samples needed to confidently use the asymptotic distribution of the test statistic. Specifically, we consider the parametric model introduced in Example 1, where every node has its own independent Poisson arrival process identically governed by rate parameter λ under the null. In this setting, $X_i(t)$ is distributed Poisson(λt) and $\mathrm{E}[X_i(t)] = \mathrm{Var}(X_i(t)) = \lambda t$. Under these null dynamics, we simulate branches and compute the test statistic value W as a function of n, the number of branches. For each value of n, we compute W 10,000 times in order to build the distribution of the test statistic. We repeat the process for values of λt ranging between 0.2 and 0.8. Figure 3-2 displays the Kolmogorov-Smirnov (KS) distance between the distribution of the test statistic W and the standard normal distribution as a function of n. Here, KS distance is defined as the max distance between the CDF of W, F_W, and the CDF of a standard normal random variable, Φ. Formally, the KS distance is

$$D_n = \sup_x |F_{W_n}(x) - \Phi(x)| \tag{3.19}$$

Figure 3-2 demonstrates the distribution of W converges quickly to its asymptotic distribution under the null, with good agreement with only $n = 50$ branches for all values of λt. As a rule of thumb, 50 branches or more provide enough samples to confidently use the standard normal asymptotic distribution.

Potential alternatives to the test statistic W in Theorem 1 are classical difference-

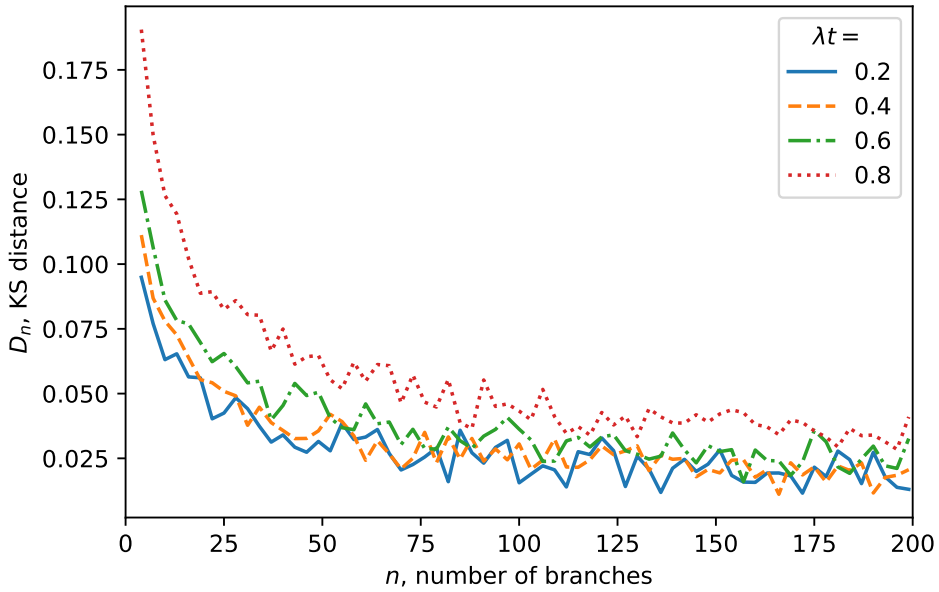


Figure 3-2: KS distance between the distribution of the test statistic W and the standard normal distribution as a function of n, the number of branches used to form W. KS distance is defined as the max distance between the two CDFs. The different curves correspond to different values of $\lambda t = \mathrm{E}[X_i(t)] = \mathrm{Var}(X_i(t))$, the expected number and variance of edges formed by each node. By Lemma 1, larger values of λt correspond to larger and more variable branch sizes.

in-means tests, such as Welch's t-tests or regressions on a binary indicator. However, as W uses analytic expressions for the expected size and variance of branches, it will be more accurate in smaller samples as long as the null assumptions hold.

3.3.3 Empirical application using call detail records

In this subsection, we apply our test statistic to call detail records from Yemen. Our call detail record (CDR) dataset includes over 2 billion calls from Yemen during 2011. The records were obtained from one of the major cellphone service providers in the country and contain the time of calls, the anonymized callers and call recipients, and the towers that handled each end of the calls. Coupled with the geographic coordinates of the towers, the records identify approximate locations for each individual at the time of their calls. CDRs naturally provide the social network of a population by highlighting communication between individuals. In addition, they offer high temporal and spatial resolution for studying reactions to localized events. In 2011, Yemen had a population of around 24 million [109]. Cellphone usage was around 50% of the population [110], while internet penetration was only 15% [111], making cellphones crucial to the spread of news and information. Our dataset covers roughly 6 million subscribers, accounting for a large fraction of the cellphone-using population.

Adding significance results to a given event.

We first test the call branches formed after a given event. A few minutes after 1pm on June 3rd, 2011, the Presidential Palace in Sana'a, Yemen's Capital, was bombed by opposition forces [112]. The president at the time, Ali Abdullah Saleh, was badly injured in the blast. Call volume in the vicinity of the palace spiked dramatically immediately after the bombing. Figure 3-3 displays outgoing call volume from individuals within a 5 km radius of the palace on the day of the bombing compared to call volume of the same area on the same day of the week for four weeks prior. The same day of the week from previous weeks forms a reasonable baseline as call volume exhibits strong weekly periodicity. Call volume reached a peak increase of

53% over the baseline days after the bombing occurred.

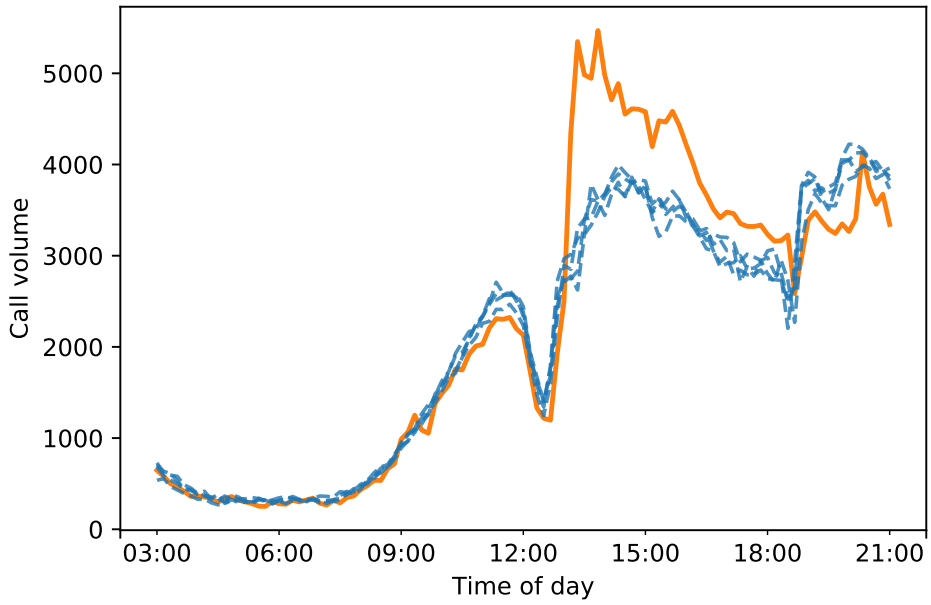


Figure 3-3: Outgoing call volume from a 5 km radius around the Presidential Palace in Sana'a. Call volume on June 3rd, 2011, the day the palace was bombed, in shown in solid orange. Call volume for the same day of the week for four weeks prior is shown in dashed blue.

We designate individuals within a 5 km radius of the palace who make calls between 1pm and 2pm on the day of the bombing as our G_0 origin nodes. We then build the subsequent call branches as follows, setting $t = 1$. For each G_0 individual, we record the calls they make during the one hour period between 1 and 2pm, labelling the individuals contacted by G_0 as G_1. We then record the calls made by G_1 within one hour after being contacted by G_0. We label the contacted individuals G_2. We continue in this manner, recording the calls made by G_2 within one hour after being contacted, and so on, until we reach a generation of callers that do not make any calls within one hour after being contacted. Multiple calls between two individuals are recorded as a single edge. Branches are built sequentially and each distinct call is only allowed to appear in a single call branch, ensuring branches do not merge and branch sizes are not coupled.

After the bombing, 17,302 call branches formed with an average size of 2.99 nodes.

Table 3.1 provides summary statistics for the observed branches. The majority of branches were small, with 62% containing only 2 nodes through 2 generations of callers (including G_0). The largest branch contained 94 nodes though 12 generations of callers. The branches reached a large number of individuals, with the 17,302 branches containing 51,674 nodes in total. In addition, the branches had substantial geographic reach. Figure 3-4 demonstrates the call branches spread across the country and reached the majority of populated areas in Yemen.

Table 3.1: Summary statistics for the palace bombing call branches

	Mean	*SD*	*Mode*	*Max*
Branch size	2.99	2.46	2	94
Generations	2.40	0.93	2	12
Breadth	1.43	0.83	1	15

17,302 call branches formed after the Presidential Palace bombing. Branch size is defined as the number of nodes in the branch, generations is defined as the number of G_i levels in the branch (including G_0), and breadth is defined as the max number of nodes in a single generation in the branch. SD stands for standard deviation.

Before we apply the test statistic W, we note a small subtlety when using CDRs. We only observe G_0 individuals who place a call and, as a result, branches have a minimum size of 2 nodes. Our branching process framework assumes G_0 nodes may form no edges, however. The solution is simple. Since edge formation in branching processes is iid, the expected size of a branch starting at a G_1 node, given we are at a G_1 node, is identical to the expected size of a branch starting at a G_0 node. For our test statistic, we therefore record branch sizes for branches starting at G_1 nodes and average over all G_1 nodes. We estimate γ and σ^2, the mean and variance of the number of edges formed by nodes, using the sample mean and sample variance of G_1 nodes from the same area on the same day of the week during the same one hour period for four weeks prior.

Average observed branch size is 1.48, averaged over 20,851 G_1 nodes. γ, the mean number of edges formed by nodes in one hour during normal periods, is 0.24 and σ^2, the variance of the number of edges formed, is 0.36. By Lemma 1, the expected

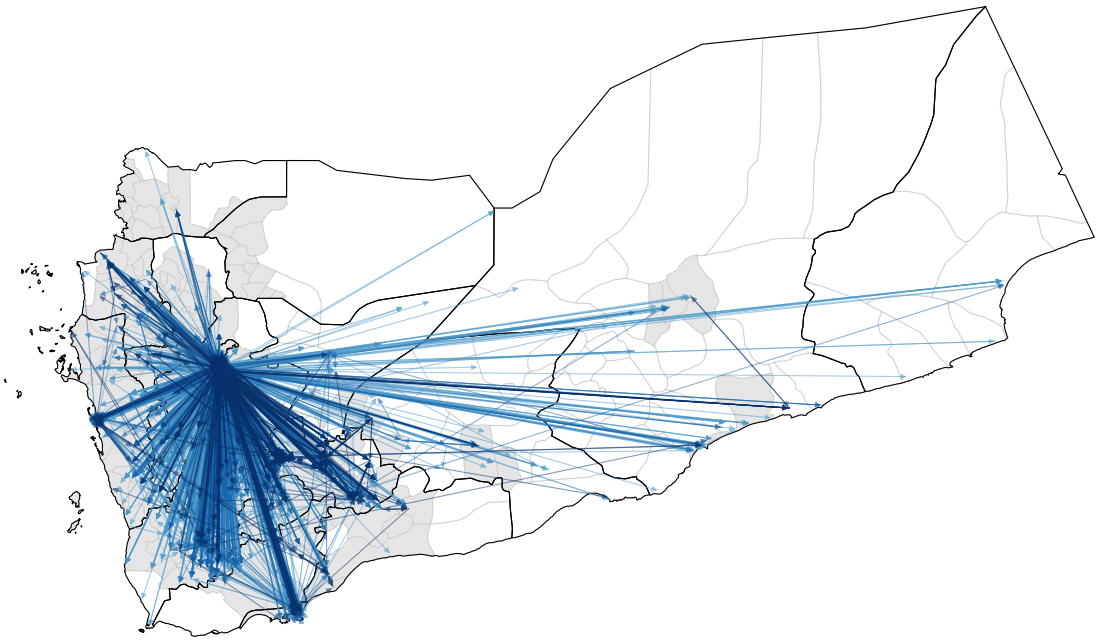
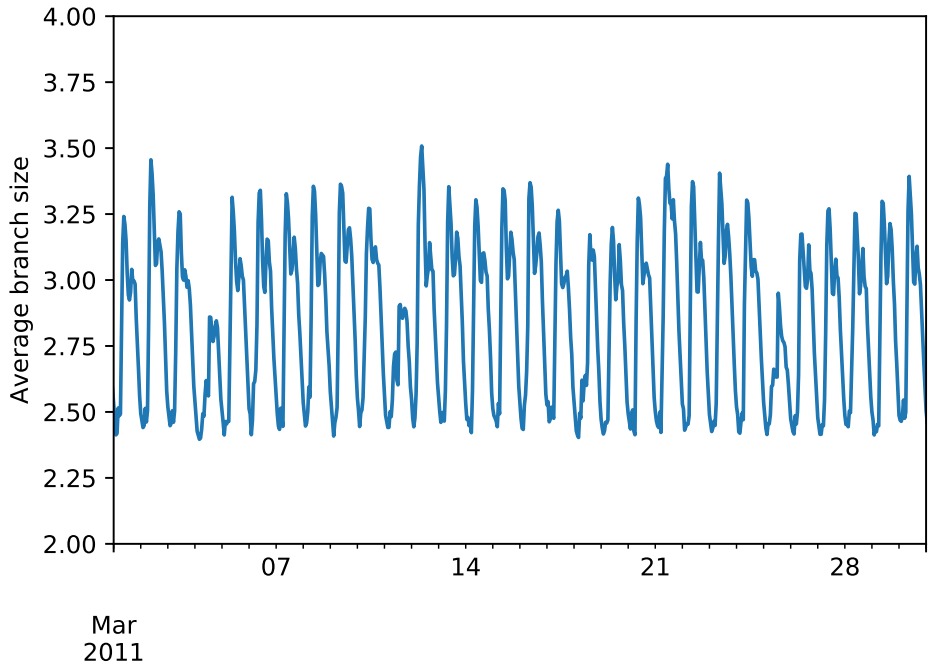
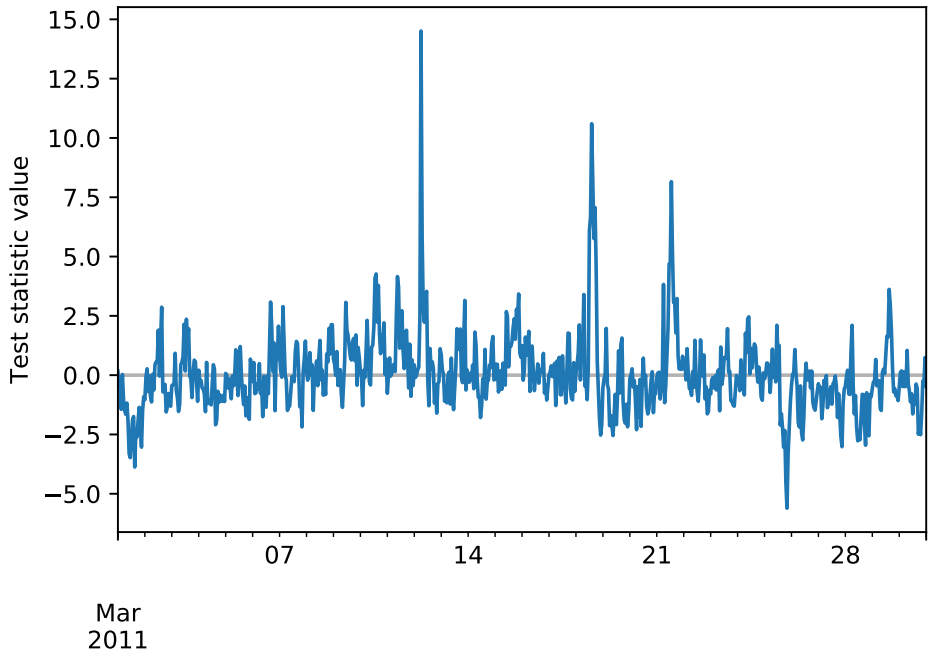


Figure 3-4: Branches formed after the Presidential Palace bombing superimposed on a map of Yemen. Darker edges are part of branches that contain more generations of callers. Governorates are outlined in black and districts are outlined in grey. Populated districts with a population density of greater than 30 people per square kilometer are shaded grey. Sparsely populated districts are left white.



(a)



(b)

Figure 3-5: **(A)** Average branch size for call branches originating in Sana'a during March 2011 at an hourly frequency. **(B)** Test statistic values for call branches originating in Sana'a during March 2011. The test statistic compares average branch size to expected branch size.

branch size under the null is 1.31. The test statistic we have introduced allows us to determine whether the observed average size of 1.48 is significantly larger than the expected size of 1.31. Plugging everything into the statistic from Theorem 1, W has a value of 27.8. The p-value of the test, the probability that a standard normal random variable is greater than W, is 5.1×10^{-170}, effectively zero. We reject the null at a 1% significance level, indicating the call branches formed after the Presidential Palace bombing were much larger than those formed during normal periods and, therefore, correspond to a calling cascade. The emergence of a calling cascade indicates contacted individuals proceeded to call their contacts after the bombing and implies information regarding the bombing spread quickly and deeply through the underlying social network. Crucially, our test statistic adds inference to the observed call branches, providing significance and confidence to our empirical findings.

Using the test statistic for event detection.

We now employ our testing framework to detect events based on large branch formation. We build the call branches originating in Sana'a every hour during March 2011, using a 10 km radius around the city center. Individuals who make a call within the radius form our G_0 nodes. We build branches as in the previous subsection for every hour period. Figure 3-5A plots the average call branch size originating in Sana'a during the month. The figure clearly shows daily periodicity in call branch size, with branch size peaking around 10am each day. Weekly periodicity is also evident. Fridays (March 4, 11, 18, and 25), the first day of the weekend in Yemen and a day of prayer and rest, display the smallest daytime call branch sizes. However, it is not immediately clear from the average branch sizes if any abnormal branches formed during the month.

We apply our test statistic to the time series of average branch sizes, using G_1 branches as in the previous subsection. We estimate expected branch size and variance for each hour period using the sample mean and sample variance of branch sizes formed by G_1 nodes from the same area during the same hour period on the same day of the week for four weeks prior. Figure 3-5B displays the test statistic values for March 2011. The periods with abnormally large call branches now stand out clearly. The

three largest spikes in test statistic values occur on the 12th, the 18th, and the 21st. We correct for multiple testing using the Benjamini-Hochberg procedure [113] and note the three spikes are all significant at a 1% significance (false discovery rate) level, indicating the observed branches are abnormally large. Checking media reports, we find the three dates correspond to key events during the 2011 Yemeni Revolution. On March 12th, police surrounded and fired on thousands of protesters who had gathered to call for President Saleh's resignation, killing four and injuring several hundred [114]. The attack was the most violent action against protestors up to that point. On March 18th, snipers fired on anti-government protesters, killing 45 and injuring 270 [115]. The gunmen were suspected pro-government forces and the attack led to unprecedented levels of anger among Yemenis. On March 21st, several military leaders defected [116]. Tanks and troops from both pro-government and anti-government forces were deployed in the city.

As we are able to match the three most significant spikes to noteworthy events, the test does not have any severe false positives (type I errors) in this sample. In order to check for false negatives (type II errors), we check Google News and other media sources for any notable activity in Sana'a during March 2011. The violence on March 12 and 18 were the only significant violent events reported during the month, adding confidence the test statistic does not have any severe false negatives in this sample. However, several large protests occurred during the month [117, 118], which the test statistic seemingly does not capture. There are several plausible explanations for the lack of significant diffusion during protests. Protests during this period were common, occurring near-daily since they began in mid-January [118, 119]. In addition, the protests were often planned in advance [118] and lasted for several hours and days [120]. Their routine, non-spontaneous, and non-instantaneous nature imply there may not have been any news or information worth spreading during the protests themselves, resulting in no significant branch formation and cascades. An interesting open question remains of which types of events cause cascades and how diffusion differs between different events.

In conclusion, our test statistic is able to successfully detect disruptive events by

highlighting periods with abnormally large call branches.

3.4 Conclusion

The test statistic we have introduced in this chapter allows us to determine whether a cascade has occurred in a network setting where branches also form during the null of normal periods. In addition, we have demonstrated the test works well in practice using call detail records from Yemen. Going forward, the test should be utilized in empirical network science research when cascade formation is being studied, as the statistic adds significance levels to observed branch structures.

The statistic can be applied to additional call detail record datasets to detect previously unreported disruptive events, such as state-sponsored attacks on civilians. It can also be applied to other network datasets such as social media networks to identify abnormally large cascades of information, news, and opinions. Applying the testing framework to Twitter data would highlight significant retweet chains. It could additionally identify substantial discussions and topics based on abnormally large branches of tweets. An application to shared posts on Facebook could be used to highlight viral posts, with a possible focus on identifying viral fake news.

Chapter 4

The Social Network Effects of Drone Strikes

Abstract

In the previous chapter, we observe the emergence of information cascades following localized events. These cascades are consequential as they quickly spread information through a population. In this chapter, we utilize call records to study whether calling cascades and physical diffusion emerge following drone strikes. Drone strikes have become a fixture of modern warfare, yet their effects on civilians, societies, and their underlying social networks remain opaque and fiercely debated.

Utilizing a new dataset of over 12 billion call detail records, we study the causal impact of 74 U.S. drone strikes on communication and mobility in Yemen between 2010 and 2012. Over 95% of strikes are followed by calling cascades, with roughly one third exhibiting increased call volume through four levels of callers. Compared to non-strike periods, proximal individuals call their frequent and geographically close contacts more frequently. Notably, socially central individuals are called twice as often and proceed to spark large calling cascades. Lastly, physical mobility increases 27% on strike days compared to the pre-strike mean and thousands of individuals flee their hometowns. These findings demonstrate drone strikes have a disruptive and widespread impact on civilian life. Furthermore, our results imply information, opinions, and emotions regarding strikes spread quickly through the population, which is in contrast to the prevailing political and military position that strikes are surgical.

4.1 Introduction

In November 2002, the United States (U.S.) conducted its first drone strike outside of an active battlefield, killing six al-Qaeda members in Yemen believed to be behind the bombing of an American destroyer [121]. Since then, the U.S. has carried out over 6000 confirmed strikes, mainly targeting suspected militants in Yemen, Pakistan, Afghanistan, and Somalia [121]. These strikes have become a mainstay of U.S. military strategy, allowing officials to target insurgents without deploying troops. Despite their prevalence, the covert nature and often isolated targets of drone strikes have made their effects difficult to study. As a result, their impact on civilians and effectiveness at countering terrorist organizations are subject to significant debate. While supporters argue they successfully disrupt terrorist networks by surgically removing key figures [122–126], critics claim they result in extrajudicial killing, civilian casualties, and increased militant sympathies [127–132]. Crucially, arguments on both sides suffer from a lack of available data [133–135].

This scarcity of information leaves open a fundamental question: How disruptive are drone strikes to civilians and their communities, and is the disruption limited to the immediate strike region? As societies are intrinsically networked systems, we focus on the dynamics of the underlying social networks around these localized, violent events. Networks provide a powerful framework to study social interactions and structure as well as disruptions to the social fabric [2, 7, 25, 70, 71]. Our goal is to identify and measure the social network effects of drone strikes, providing quantitative evidence of their impact on civilian life.

We utilize a new dataset of 12 billion call detail records (CDRs) to study the societal reaction to 74 U.S. drone strikes in Yemen between 2010 and 2012. During this period, al-Qaeda seized control of key territory across the country and the number of U.S. strikes increased in response, peaking in 2012 [136]. The CDRs we employ are uncommonly complete and contain the time of calls and texts, the anonymized callers and call recipients, and the towers that handled each end of the calls. Combined with the geographic coordinates of the towers, the records identify approximate locations

for each individual at the time of their calls. By highlighting communication between individuals, CDRs naturally provide the social network of a population and offer the high temporal and spatial resolution necessary to study reactions to localized events [137–142]. In addition, CDRs allow us to systematically study civilian reactions to strikes without relying on self-reported data such as surveys and interviews, which are the norm in the drone strike literature [132, 143, 144]. Data on the strikes was compiled by the Bureau of Investigative Journalism [121] and the New America think tank [136] from credible media reports and includes the date, approximate location, time, and casualty count for each strike.

At the start of our CDR data in 2010, Yemen had a population of around 23 million [109]. Cellphone usage rose steadily from 49% of the population in 2010 to 58% in 2012 [110], while internet penetration was only 12% [111], making cellphones crucial to the spread of news and information. Our dataset covers roughly 6 million subscribers, accounting for a large fraction of the cellphone-using population.

In our analysis, we assume drone strikes are exogenous shocks, which gives our results causal interpretation. To strengthen the argument for exogeneity, we confirm potentially confounding events, such as militant activity, do not take place at the same time and location as the drone strikes in our sample. Specifically, we confirm reported al-Qaeda attacks, public communications, and movements, as well as activity during the 2011 Battle of Zinjibar and the 2012 Abyan Offensive, do not overlap with the dates and locations in our strike dataset. The issue of confounding events highlights the advantage of using data with high temporal and spatial resolution, such as CDR data, to conduct event studies.

In this chapter, we employ our CDR data and modern inference methodology to study the impact of drone strikes on civilian communication and mobility. Quantifying the effects of strikes on civilian life provides a foundation for improved, data-driven policy and advances our understanding of drone strikes in modern conflict. In addition to their ethical and legal ramifications [121, 129, 132], the effects of strikes on civilians can alter the dynamics and outcomes of ongoing conflicts. Recent research has highlighted the key role civilians play in modern wars, demonstrating actions that

disrupt civilian life degrade support for the responsible party and affect subsequent levels of violence [145–148]. As a result, understanding the impact of strikes on civilians is of paramount importance for contemporary strategies for conflict prevention and resolution.

4.2 Results

4.2.1 Calling cascades

Previous studies employing CDRs have found call volume spikes dramatically in the vicinity of violent events [60–62, 149]. In addition, in the previous chapter, we have observed the emergence of information cascades around localized events. With these results as a starting point, we study whether calling calling cascades emerge around drone strikes, where branches of calls originating in the strike region spread through the underlying social network.

Specifically, we define proximal individuals as individuals within 15 miles (24.1 km) of the reported strike locations who make a call during the periods of elevated call volume following each strike. We label proximal individuals G0, as they are the zeroth generation of potential call branches originating in the strike region. Individuals contacted by G0 after the strike are labeled G1, individuals contacted by G1 are labelled G2, and so on. Figure 4-1 displays the call branches formed after a drone strike on January 20, 2010 and highlights the different generations of callers. All results presented are robust to strike region radii choices between 5 and 30 miles (8.0 and 48.3 km) (Supplementary materials (SM) section 4.5.3).

After the 74 drone strikes, average G0, G1, G2, and G3 call volumes all increase sharply (Fig. 4-2), indicating individuals contacted by proximal individuals proceed to call their contacts, who in turn call their contacts, and so on, forming calling cascades. G1 individuals, who are directly contacted by proximal individuals, respond most dramatically with a 168% call volume increase on average 40 minutes after the strikes occur compared to non-strike periods, with a standard error (SE) of 25.9. Non-strike

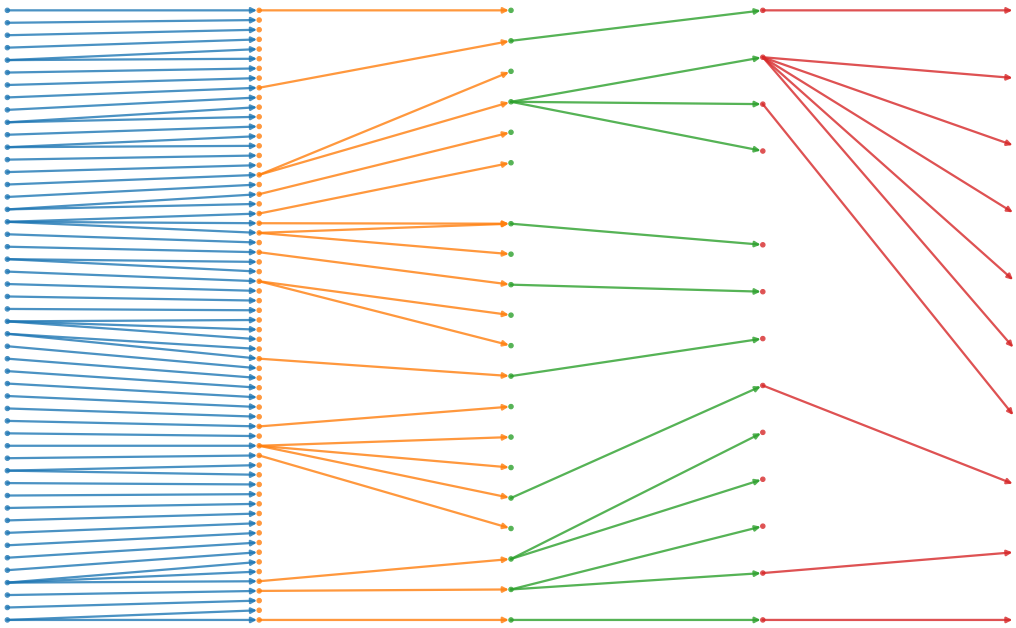


Figure 4-1: Call branches formed after a drone strike. Call branches formed after a drone strike on January 20, 2010, which targeted the home of an al-Qaeda leader in the Marib province. G0 individuals are proximal to the strike and contact G1 individuals, who proceed to contact G2 individuals, and so on. Calls from G4 individuals and higher generations are not shown. Only a subset of the call branches formed are displayed for clarity.

periods used as a baseline consist of the same day of the week as the strike for five weeks prior and five weeks in the future, exploiting the strong weekly periodicity in call volume. G0 call volume increases by 104% (SE 29.9) 30 minutes after strikes, G2 peaks at a 70% (SE 15.6) increase 60 minutes after, and G3 peaks at a 47% (SE 15.0) increase 70 minutes after. Notably, strikes that result in civilian casualties spark larger cascades. Regressing the increase in G1 call volume one hour after each strike on the number of civilians killed as well as several controls, we find strikes that kill 10 civilians correspond to a 115% (SE 38.6) greater increase in G1 call volume (Table 4.1A).

	(A)		**(B)**	
	Beta	T-Stat	Beta	T-Stat
Civilians Killed	11.46	(2.97)	4.08	(2.53)
Militants Killed	-0.66	(-0.51)	0.38	(1.67)
High-Ranking Militant	-14.86	(-0.45)	9.45	(1.62)
Morning	60.28	(1.61)	-2.41	(-0.52)
Evening	7.59	(0.28)	3.48	(0.81)
No. in Past Month	0.49	(0.03)	1.93	(0.73)
Population (10,000s)	3.90	(1.06)	1.50	(2.85)
Intercept	71.79	(2.00)	-9.29	(-1.28)
R^2	24.9%		55.3%	
NObs	74		74	

Table 4.1: **(A)** Strikes with civilian casualties are followed by larger cascades. We regress the increase in G1 call volume one hour after each strike on the reported number of civilians killed in the strike, the number of militants killed, a binary variable for whether a high-ranking militant was killed, binary variables for the time of the strike (morning is defined as 00:00-08:00 and evening is 16:00-23:59), the number of strikes that hit the same district in the past 30 days, and the population of the district. Interpreting the civilian coefficient, strikes that kill 10 civilians correspond to a 114.6% greater increase in G1 call volume. **(B)** Strikes with civilian casualties are followed by higher levels of fleeing. We regress the number of proximal individuals who live within the strike region, leave within 24 hours after strikes, and do not return within 30 days on the same covariates. Interpreting the civilian coefficient, a strike that kills 10 civilians corresponds to 40.8 more proximal individuals who flee their hometowns and do not return within 30 days. The covariates are provided by [121, 136, 150]. The regressions use heteroscedasticity robust standard errors.

To test the statistical significance of the call volume increases for each strike, we compare the number of calls made by each generation of callers on strike days to

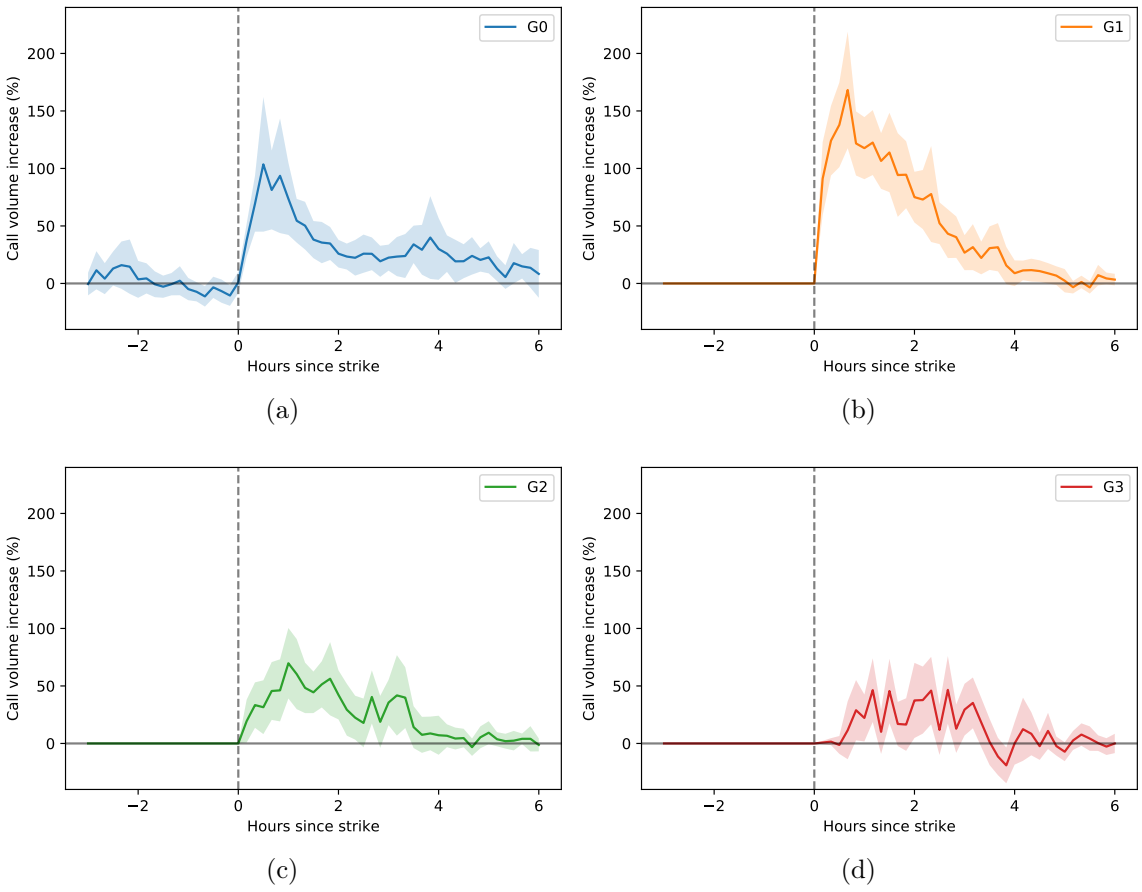


Figure 4-2: Emergence of calling cascades after drone strikes. Call volume by generation of caller, averaged across the 74 strikes, relative to call volume on non-strike days. The solid lines display average call volume for **(a)** G0 individuals, **(b)** G1 individuals, **(c)** G2 individuals, and **(d)** G3 individuals. The shaded regions provide 95% confidence intervals. Confidence intervals are computed by regressing the relative call volume series from all 74 strikes on a constant, providing sample averages as well as robust standard errors.

non-strike days with a regression framework. Specifically, for each strike and each generation of caller, we regress the number of calls made by each individual after being contacted on the strike day and during the same time period on the 10 baseline days on an indicator variable for the strike day. The coefficients of these regressions therefore report the average number of calls made by individuals on the strike day minus the average number of calls made by the same individuals on the baseline days. We test the significance of the regression coefficients at a 5% level via a one-sided t-test with heteroscedasticity robust standard errors. Out of the 74 strikes, the majority of strikes display significant cascades through several generations of callers. At a 5% level, 96% of strikes are significant through G1, 62% through G2, and 31% through G3. To address issues of multiple testing, we apply the Benjamini-Hochberg (BH) procedure to control the false discovery rate at 5% [113]. Under the BH step-up procedure, 96% of strikes are significant through G1, 59% through G2, and 24% through G3.

These cascades reveal the impact of drone strikes is not limited to the immediate strike region, but also propagates through the social network to individuals and communities several steps removed from the proximal individuals. Crucially, these cascades allow information, opinions, and emotions regarding the strikes to spread quickly through the population. The strong reaction to strikes and civilian casualties indicates civilians are acutely aware of the collateral damage caused by strikes and respond by informing their contacts.

4.2.2 Call patterns

The increased call volume around strikes raises the immediate question of whom proximal individuals are choosing to call. To answer this, we first determine each individual's baseline contact list by constructing the underlying social network using 30 days of calls before each strike.

The calls form an undirected graph where an edge exists between two individuals if they have communicated during the baseline period. These networks not only determine the list of people each proximal individual could potentially call after each strike but also allow us to rank those contacts along different metrics. For each

proximal individual, we rank their contacts by their frequency of communication with the proximal individual during the 30-day baseline period before the strike, by their home location proximity to the proximal individual's home location, and by their diffusion centrality, which measures the importance of each contact in the global social network. Home locations are defined as each individual's most common evening tower location during the 30-day baseline period.

Diffusion centrality was introduced by Banerjee et al. (2013) [3] to identify individuals ideally situated in a social network to spread information. Formally, diffusion centrality in vector form is defined as

$$\boldsymbol{D} = \left[\sum_{t=1}^{T}(p\boldsymbol{A})^t\right] \cdot \mathbf{1} \tag{4.1}$$

where $\boldsymbol{A}$ is the adjacency matrix of the network, T is the number of periods in which information can diffuse through the network, and p is the passing probability from node to node during each period. Intuitively, the diffusion centrality of node i measures the expected number of times all individuals in the network hear about a piece of information that node i is seeded with. Motivated by our empirical results, we set $T = 3$ and $p = 0.28$ as a large number of cascades are significant through three generations of callers and roughly 28% of G1-G3 individuals pass on a call after being contacted. Our results are robust to simpler measures of network importance such as degree centrality (SM section 4.5.5).

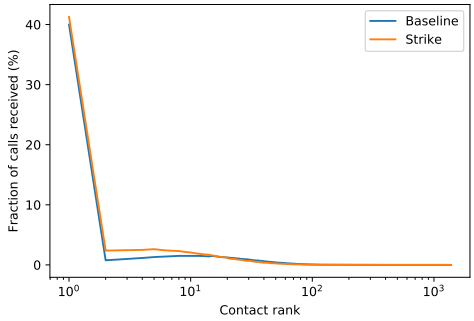
Both after strikes and during the baseline period, the majority of calls are made to important rank 1 contacts across the three different metrics (Fig. 4-3). However, after strikes, the distribution of calls made shifts and low rank contacts are called more frequently. 41% of the calls made after all 74 strikes are to rank 1 contacts by home location proximity, 28% are to rank 1 contacts by frequency of communication, and 5.5% are to rank 1 contacts by diffusion centrality, double the 2.7% of calls they receive during the baseline period. All three distributional shifts are statistically significant at a 5% level using a two-sample Kolmogorov-Smirnov test. In line with intuition, proximal individuals call their strong ties as well as their geographically close contacts

more frequently after strikes, most likely to inform or check in on family, friends, and neighbors. More surprisingly, a fraction of calls are made to central individuals who occupy ideal positions in the social network to diffuse information.

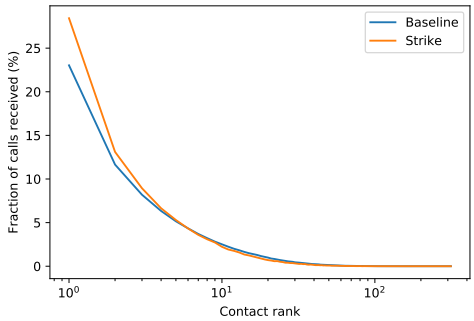
Calls to socially central individuals are consequential, as we find central G0 and G1 individuals spark larger cascades. To determine the relationship between cascade size and centrality, we first regress the number of individuals in G1, G2, and G3 in each call branch present after strikes (the branch size) on the diffusion centrality of the G0 individual (the origin node). To account for the possible endogeneity of diffusion centrality, we use an instrumental variable (IV) framework estimated via two-stage least squares (2SLS). We use the number of people that call a node on Eid al-Fitr, a major Muslim holiday marking the end of Ramadan, in 2011 and 2012 as instruments for the node's centrality.

Our instruments pass both the relevance and exclusion criteria for IV. Intuitively, the number of people who call a node during Eid is related to the node's social importance. Quantitatively, the number of people who call each node during 2011 Eid is 0.51 correlated with diffusion centrality and the number of people who call each node during 2012 Eid is 0.46 correlated. The two instruments are 0.54 correlated with each other. The instruments are excluded as, intuitively, the number of individuals who call a node during Eid is not related to the node's frequency of call origination after strikes, except through centrality. To strengthen this argument, we perform a Sargan over-identification test using the residuals of our 2SLS regression. With a p-value of 0.55, we fail to reject the null that our instruments are exogenous.

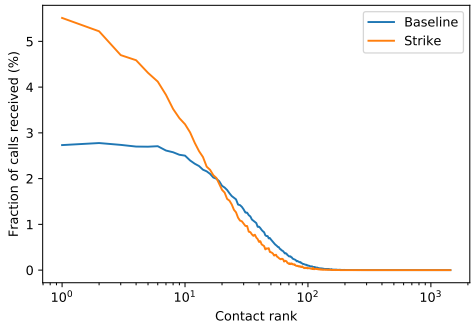
In addition to regressing branch size on the centrality of the G0 origin node, we also regress the number of individuals in G2 and G3 in each sub-branch on the diffusion centrality of the G1 individual that originates the sub-branch, again via 2SLS. We find high centrality G0 individuals originate call branches that reach 32 (SE 1.2) more people than low centrality G0 individuals (Table 4.2). Similarly, high centrality G1 individuals originate sub-branches that reach 25 (SE 2.5) more people. Both regressions indicate socially central individuals play a key role in diffusing information after drone strikes.



(a)



(b)



(c)

Figure 4-3: Shifts in calling patterns after strikes. **(a)**, **(b)**, and **(c)** display the fraction of calls received by contacts ranked by their home location proximity to the proximal individual, frequency of communication with the proximal individual during the baseline period, and diffusion centrality, respectively. Across all three metrics, important low rank contacts receive a larger fraction of calls after strikes than during the baseline period.

		Centrality	R^2	NObs
(A)	Beta	0.00320	7.1%	74,960
	T-Stat	(26.68)		
(B)	Beta	0.00249	8.3%	36,124
	T-Stat	(9.88)		

Table 4.2: Central individuals originate larger cascades. **(A)** regresses the number of individuals in G1, G2, and G3 for each call branch (the branch size) on the diffusion centrality of the G0 individual (the origin node) via two-stage least squares (2SLS), using the number of people that call each G0 individual during 2011 and 2012 Eid al-Fitr as instruments. **(B)** regresses the number of individuals in G2 and G3 in each sub-branch on the diffusion centrality of the G1 individual that originates the sub-branch, again via 2SLS. The regressions use strike fixed effects to account for strike-specific heterogeneity and heteroscedasticity robust standard errors. A highly central G0 individual with a centrality of 10,000 originates a call branch with 32 more individuals than a low centrality G0 individual. A highly central G1 individual with a centrality of 10,000 originates a sub-branch with 25 more individuals. For context, branch size ranges from 2 to 73 with a mean value of 3.0 and diffusion centrality ranges from 0.36 to 10,699.59 with a mean value of 238.78.

4.2.3 Mobility response

The location estimates provided by the CDRs allow us to analyze the physical response to drone strikes in addition to the communication response. For each proximal individual, we construct a time series of their locations and compute the distance between them, building an estimate of daily distance travelled. Plotting average daily distance travelled around strikes, we find mobility spikes substantially on strike days (Fig. 4-4). To quantify this increase, we regress the daily distance travelled by all proximal individuals for 14 days before strikes and during strike days on an indicator variable for the strike days. We find average mobility increases 7.6 km (SE 0.24) on strike days, an increase of 27% over the pre-strike mean of 28.5 km (Table 4.3). Individuals display a strong physical reaction to drone strikes, indicating strikes disrupt the course of daily life.

To test the mobility response for each strike separately, we regress the daily distance travelled by proximal individuals for 14 days before each strike and on the strike day on an indicator variable for the strike day. We then test the significance of the regression coefficients via a one-sided t-test with heteroscedasticity robust standard

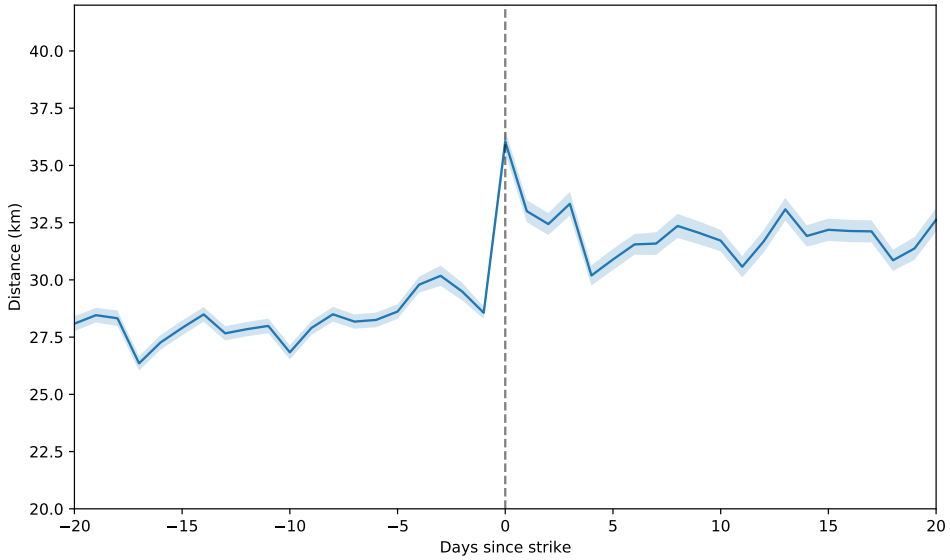


Figure 4-4: Spike in mobility around strikes. The solid line displays average daily distance travelled by proximal individuals around the 74 strikes. 95% confidence intervals are shown in light blue and are computed by regressing the mobility of individuals on a constant, providing sample averages as well as robust standard errors.

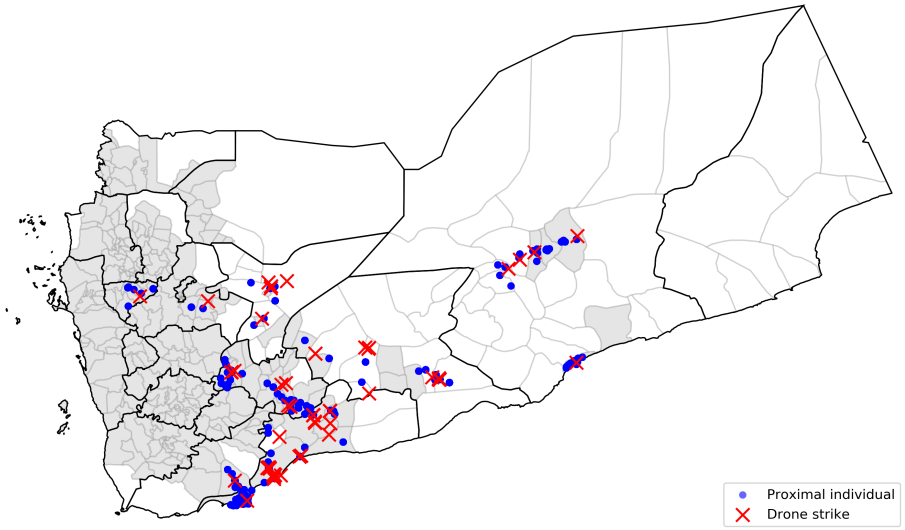
	Strike	R^2	NObs
Beta	7.64	0.1%	1,743,890
T-Stat	(32.37)		

Table 4.3: Mobility spikes on strike days. We regress the daily distance travelled by proximal individuals (in km) on a binary indicator for strike days. Specifically, for each individual, we include their daily distance travelled for 14 days preceding each strike, associated with a strike indicator variable of 0, and their daily distance travelled on the day of the strike, associated with a strike indicator variable of 1. Interpreting the coefficient, average daily distance travelled increases by 7.64 km on strike days compared to average distance travelled over the preceding two weeks, a 27% increase over the pre-strike mean of 28.5 km. The regression uses strike fixed effects to account for strike-specific heterogeneity and standard errors are clustered by individual.

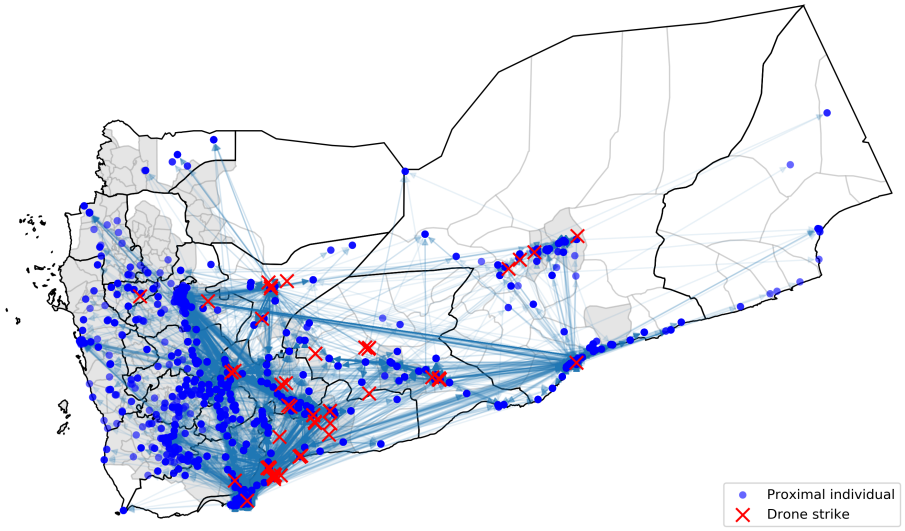
errors. Out of the 74 strikes, 58% display statistically significant increases in mobility on strike days at a 5% level. Applying Benjamini-Hochberg with a false discovery rate of 5% to correct for multiple testing, 54% of strikes display significant increases in mobility on strike days. Our results are robust to alternative measures of mobility, such as distance between proximal individuals' first and last call made each day (SM section 4.5.6).

Investigating the increase in mobility around strikes, we find many individuals flee the strike region. A large number of people leave the strike region quickly, dispersing around the country within 24 hours (Fig. 4-5). In the subset of the population covered by our CDR data, 4519 proximal individuals who live within the strike region leave within the first 24 hours after the 74 strikes occur and remain away for at least 24 hours, highlighting the disruptive impact of strikes. In percentage terms, 5.3% of proximal individuals who live within the strike region flee after each strike on average. Of those who flee, 51% return quickly and are home within five days. However, 1049 individuals do not return to their hometowns within a 30-day period, demonstrating a prolonged impact to communities. Notably, strikes that result in civilian casualties are followed by higher levels of fleeing. Regressing the number of individuals who do not return to their hometowns within 30 days on the number of civilians killed as well as several controls, we find strikes that kill 10 civilians correspond to 41 (SE 16.2) more individuals who flee and do not return compared to strikes with no civilian casualties (Table 4.1B). As the majority of strikes take place in medium to low population districts, the number of individuals displaced, even in our sample, is substantial and is an order of magnitude larger than the roughly 390 militants targeted in the 74 attacks.

Decomposing the movement of proximal individuals who flee, 54% end up near (within 5 miles / 8.0 km) a major city during their time away, 83% end up near the home location of one of their contacts, and 36% end up near the person they called immediately after the strike. Cities, contacts, and people called after strikes account for the locations of 91% of the proximal individuals who flee (SM section 4.5.6), demonstrating both physical and social networks explain the movement of individuals.



(a)



(b)

Figure 4-5: Dispersion of proximal individuals after drone strikes. **(a)** Locations of proximal individuals, in blue, at the time of the 74 strikes, which are marked with red X's. By definition, proximal individuals are within 15 miles (24.1 km) of the strikes at the time of their calls during the strike periods. Populated districts with a population density of greater than 30 people per square kilometer are shaded grey. Sparsely populated districts are left white. **(b)** Locations of the same proximal individuals 24 hours after the strikes occur, displaying rapid dispersion from the strike regions. The trajectories of individuals are shown as light blue edges.

Comparing the response to drone strikes to other events highlights the uniquely disruptive nature of strikes. We analyze the communication and mobility response to a factory explosion in the town of Ja'ar in March 2011, a bombing of the Presidential Palace in the capital city of Sana'a in June 2011, a bombing of the Presidential Palace in the coastal city of Al Mukalla in February 2012, a suicide bombing targeting soldiers in Sana'a during May 2012, and a suicide bombing targeting Sana'a's police academy in July 2012 (SM section 4.5.7). The communication response to these five events mirrors the response to drone strikes, with significant calling cascades forming after each event. In addition, the events led to shifts in calling patterns and spikes in mobility. However, the events did not induce substantial fleeing, and only 1.7% of proximal individuals left the areas within 24 hours. In contrast to drone strikes, no proximal individual fled and remained away for 30 days or more. Similarities in the communication response to strikes and these events can be attributed to their shared violent, instantaneous, and unexpected nature. Differences in fleeing, however, reveal the specifically damaging effect of strikes. Civilians likely view drone strikes differently than these focused comparison events, as strikes, especially those resulting in civilian causalities, may seem indiscriminate to those on the ground.

4.2.4 Event study approach to mobility

In this subsection, we revisit our mobility results using a panel-data event study, which provides a transparent quasi-experimental design. We regress outcome variables of interest on a set of lag and lead indicator variables. The resulting parameters estimate dynamic treatment effects that can be viewed graphically, and the identifying assumption of "no pre-event trend" can be checked graphically.

We follow a classical event study design [151]. We define y_{st} as our response variable where s indexes the strike and t indexes the time. We aim to estimate the effect of strikes, which occur at times e_s, on y_{st} over a window ranging from $\underline{j}$ periods

before the strike to $\overline{j}$ periods after. Our regression model is

$$y_{st} = \sum_{j=\underline{j}}^{\overline{j}} \beta_j x_{st}^j + \alpha_s + \varepsilon_{st} \tag{4.2}$$

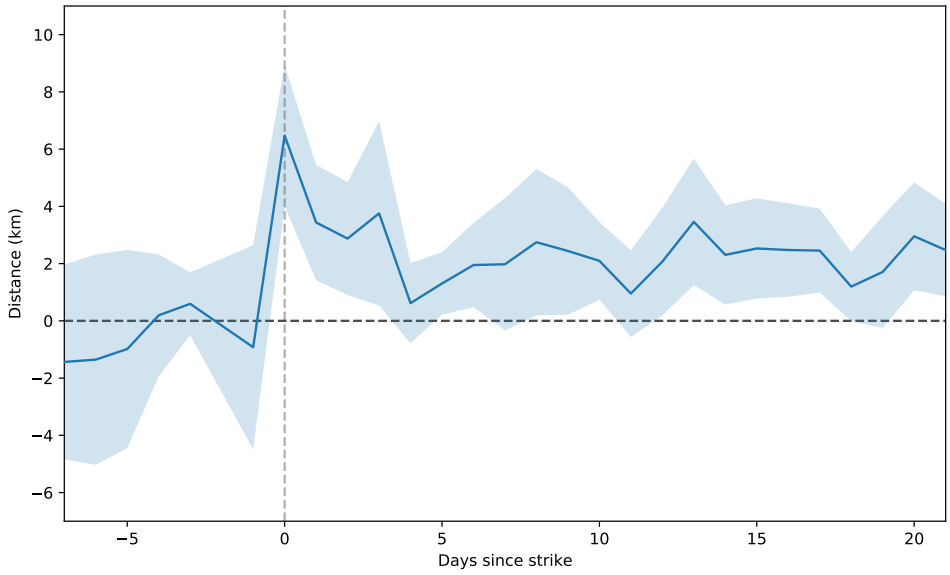
where j indexes the lags and leads, α_s captures strike-fixed effects, which account for unobserved heterogeneity across strikes, and ε_{st} is the error term. Our lag and lead indicator variable, x_{st}^j, is defined as

$$x_{st}^j = \begin{cases} \mathbb{1}[t \leq e_s + j] & \text{if } j = \underline{j} \\ \mathbb{1}[t = e_s + j] & \text{if } \underline{j} < j < \overline{j} \\ \mathbb{1}[t \geq e_s + j] & \text{if } j = \overline{j} \end{cases} \tag{4.3}$$

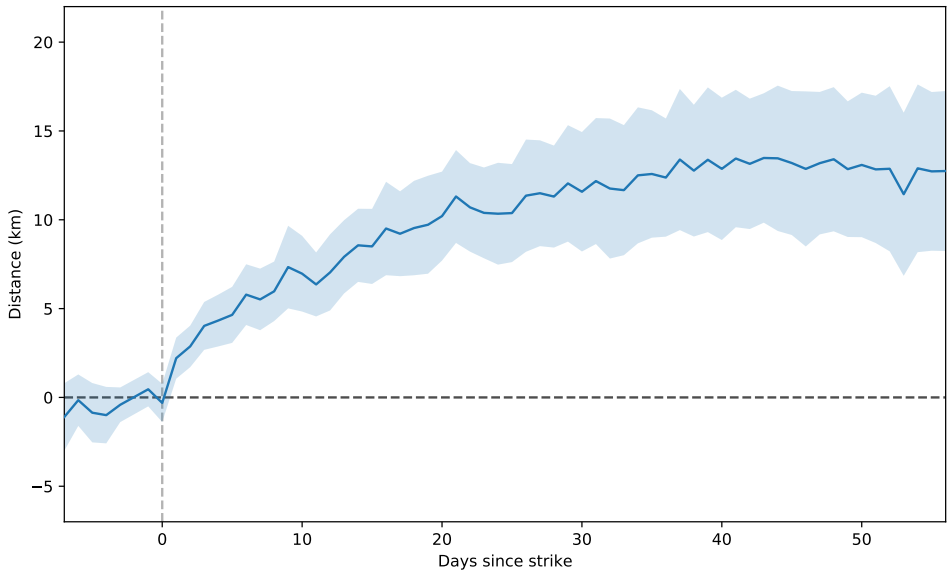
The variable x_{st}^j equals 1 if time t equals the strike-specific event time e_s plus the lead/lag index j, and equals 0 otherwise. The x_{st}^j variable is binned at the endpoints, meaning x_{st}^j equals 1 if the strike occurred $\underline{j}$ or more periods in the future or $\overline{j}$ or more periods in the past.

The coefficients β_j are our parameters or interest. β_j for $j \geq 0$ estimate the dynamic effects of strikes on the response variable. β_j for $j < 0$ capture any pre-strike trend in the response variable. To avoid multicollinearity, one of the x_{st}^j variables must be dropped. Following convention to drop a lag close to the event time, we drop x_{st}^j for $j = -2$. After dropping x_{st}^{-2}, the parameter β_j captures the effect j periods after the strike compared to the level two periods before the event.

As discussed in the previous subsection, the location estimates provided by the CDRs allow us to analyze the physical response to drone strikes. For each proximal individual, we construct a time series of their locations and compute the distance between them, building an estimate of daily distance travelled. Using the event study design in equation 4.2, we regress daily distance travelled for proximal individuals around all 74 strikes for 7 days before to 21 days after each strike. The resulting parameters estimate the dynamic effect of strikes on mobility. The parameter estimates with corresponding 95% confidence intervals are shown in Figure 4-6a.



(a)



(b)

Figure 4-6: Impact of drone strikes on proximal individual mobility. The figures display parameter estimates from an event study that regresses a dependent variable on lag and lead indicator variables. The full model is specified in equation 4.2. The dependent variables are **(a)** daily distance travelled by proximal individuals, and **(b)** distance of each proximal individual from the strike region. The shaded regions provide 95% confidence intervals. Following [152], standard errors are clustered by strike because each strike impacts a cluster of observations, rather than individual observations.

Viewing the results, we see mobility spikes significantly on strike days. Mobility increases 6.47 km (SE 1.26) on strike days compared to the period before strikes, an increase of 24% compared to the pre-strike mean. Mobility remains elevated for 21 days following strikes, with a mobility increase of 2.47 km (SE 0.83) remaining 21 days after strikes. Figure 4-6 also shows no significant trend in mobility before strikes, reinforcing our identifying assumption.

Investigating the increase in mobility around strikes, we find many individuals flee the strike region. We regress the distance of each proximal individual from the strike region around all strikes for 7 days before to 56 days (8 weeks) after each strike, again using equation 4.2. The results, shown in Figure 4-6b, demonstrate proximal individuals leave the strike region, with average distance from the region increasing steadily after strikes. Average distance reaches 5.52 km (SE 0.89) one week after strikes and 11.31 km (SE 1.33) 21 days after. Average distance continues to increase until it peaks at 13.48 km (SE 1.86) 43 days after strikes. The steady increase indicates individuals leave the region after strikes and travel far before stopping. Average distance does not return to pre-strike levels in our sample, indicating a subset of individuals do not return to the region.

4.3 Discussion

Our findings shed light on the effects of drone strikes, which have proven difficult to systematically study in the past. Utilizing new CDR data, we provide quantitative evidence of their impact on the communication and mobility patterns of civilians. Specifically, we find strikes induce calling cascades, shifts in calling patterns, and spikes in mobility. Notably, central individuals in the global social network are contacted and a significant number of individuals flee their hometowns. These results demonstrate strikes have a disruptive effect on civilians and their communities. Crucially, by highlighting the presence of both communication diffusion and physical diffusion after strikes, we find their impact is not only limited to the immediate strike region. The reverberations felt throughout the social network are in stark contrast to language

used by administration officials, who often describe strikes as a surgical military option [122, 123].

Our results provide a foundation for improved, data-driven policy and advance our understanding of drone strikes in modern conflict. Diffusion facilitates the spread of information, opinions, and emotions regarding strikes through the population. Strikes with civilian casualties are followed by larger cascades and greater amounts of fleeing, implying diffusion carries negative news and emotions. This spread has the potential to shift civilian loyalties and sentiments. In line with recent research on the strategic costs of civilian harm [145–148], civilians impacted by strikes will be less likely to aid counterinsurgents and more likely to harbor and support insurgents and their cause. As a result, fully characterizing and understanding the effects of strikes on civilians is crucial for improved policy. Updated strategies for conflict prevention and resolution are needed to resolve modern conflicts, including the ongoing wars in the Middle East, that have lasted decades and claimed hundreds of thousands of lives [153, 154].

Although we demonstrate a disruptive impact on civilian life and the presence of diffusion, we lack the content of the communications and are thus unable to analyze exactly how opinions and loyalties shift around these events. Future research should employ further data sources, such as media coverage, speeches, and sermons, to understand which information and emotions spread after strikes. An open question remains of whether the disruption induced by strikes increases or decreases militant recruitment. As the rate of U.S. drone strikes continues to steadily increase, their impact on civilians and communities cannot be ignored.

4.4 Methods

4.4.1 CDR and drone strike data

The call detail records were obtained from a major cellphone service provider in Yemen and contain subscriber communications between January 1, 2010 and October 31, 2012. Table 4.4 displays three calls from the dataset as an example. The set of 74 U.S. drone strikes was compiled by the Bureau of Investigative Journalism (BIJ) [121] and the New America think tank (NA) [136] from credible media outlets. In our subset of strikes, 6 occur in 2010, 15 in 2011, and 53 in 2012. As a summary, 18 of the 74 strikes took place in the morning (00:00-08:00), 30 took place during the day (08:00-16:00), and 26 took place in the evening (16:00-23:59). The average strike killed 1.1 civilians (4.2 std) and 9.0 militants (9.5 std), and occurred in a district with a population of 77,051.7 (47,723.6 std). 17 of the strikes killed a high-ranking militant (see SM section 4.5.3 for full definition).

4.4.2 Call branch construction

Proximal individuals are defined as individuals who place calls within 15 miles (24.1 km) of the strike location during the strike period. We label proximal individuals G0 as they form the zeroth generation of potential call branches. The individuals G0 contact during the strike period are labelled G1. We require G1 to be outside the strike region and to be distinct from the members of G0. This ensures any increase in G1 call activity is not due to them witnessing the event. The individuals G1 contact within an 80-minute window after being contacted by G0 are labelled G2. Again, we require G2 to be outside the strike region and to be distinct from the members of any previous generation. We continue in this manner, labelling the individuals G2 contact within an 80-minute window after being contacted by G1 as G3, and so on. The 80-minute response window is imposed as we are interested in calls made by call recipients directly after being contacted. Our results are robust to choices of response window length between 20 and 240 minutes (SM section 4.5.4).

4.5 Supplementary materials

4.5.1 Notes on causality

We assume drone strikes are exogenous shocks, which gives our results causal interpretation. Under exogeneity, the results of an event study analysis can be attributed to the event itself. We attribute the call volume increase in Figure 4-2, the shift in calling patterns in Figure 4-3, the spike in mobility in Figure 4-4 and Table 4.3, and the individuals who leave the strike region immediately (within 24 hours) to drone strikes.

The two main obstacles to exogeneity in event studies are simultaneity and joint response [155, 156]. Simultaneity occurs if the proposed dependent variable causes the proposed independent variable, in addition to or in lieu of the independent variable causing the dependent variable. We do not believe increases in call volume, shifts in calling patterns, or spikes in mobility cause drone strikes, as there is no evidence or documented reports that they do.

Joint response is a more realistic issue and arises if both the dependent variable and independent variable are both caused by some other event / variable. If both variables jointly respond to this missing influence, we may mistakenly attribute the change in dependent variable to our proposed independent variable. This is essentially omitted variable bias in the context of an event study. In our analysis, call volume increases, spikes in mobility, and drone strikes may all be caused by militant activity (such as militant attacks or militants entering a town). For joint response to be an issue, the omitted event must take place at the same time and in the same location as the drone strike. We confirm key militant activity, including reported al-Qaeda attacks, public communications, and movements, as well as activity during the 2011 Battle of Zinjibar and the 2012 Abyan Offensive, do not overlap with the dates and locations in our drone strike dataset. Therefore, we have confidence our event study results, which utilize daily and intra-day time windows as well as 15-mile (24.1 km) radii geographic windows, are not confounded by militant activity. The joint response issue highlights the advantages of using data with high temporal and spatial resolution,

such as CDR data, when conducting event studies.

4.5.2 Call detail records and coverage

Our dataset of call detail records was obtained from a major cellphone service provider in Yemen and contains subscriber communications between January 1, 2010 and October 31, 2012. The subset of the dataset we employ includes over 12 billion calls and texts for around 6 million unique subscribers. Each call record includes the anonymized caller and call recipient, the towers that handled the caller and call recipient's call, and the time and duration of the call. Combining these records with the latitudes and longitudes of the towers provides the approximate locations of both the caller and call recipient at the time of each call. Table 4.4 displays three calls from the dataset as an example.

Figure 4-7 displays Yemen's population in 2010 at the district level. Yemen's last census was in 2004 and the CSO (Central Statistical Organization) is not currently active. However, UN OCHA (United Nations Office for the Coordination of Humanitarian Affairs) has projected district-level populations for 2016, allowing us to interpolate 2010 populations [150]. Administratively, Yemen is divided into 22 governorates and 333 districts. The population is concentrated on the west side of the country as the east is mainly desert. The right three governorates, which account for over half of the country by landmass, account for only 8% of the population.

To understand our dataset's coverage of Yemen's population at a granular level, we compute the fraction of the population covered in each district. We first form home location estimates for all subscribers, which we define as their most common tower location. Using a Voronoi tessellation, which splits the country into regions approximately covered by each cell tower, we allocate individuals to the districts covered by their home towers. Specifically, we determine the fraction of each district's population covered by each tower according to the area of overlap from the tessellation. We then allocate the individuals in our dataset with specific home towers to districts in proportion to the district level populations covered by each tower. Note, this partitioning provides a more accurate estimate of our dataset's coverage than simply

attributing all individuals to the district where their home tower is located, as tower coverage does not obey district boundaries. If a tower sits near the border of a district, it likely covers individuals in the neighboring district as well. Our dataset's coverage of Yemen's population at the district level is balanced, covering 18.4% of the population in each district on average with a median value of 12.1%.

4.5.3 Drone strike data and strike period selection

We use a set of 74 U.S. drone strikes taking place in Yemen between January 2010 and October 2012. The strike dates and approximate locations are provided by the Bureau of Investigative Journalism (BIJ) [121] and the New America think tank (NA) [136], which compile strikes reported by various news outlets. The original dataset sourced from [121] and [136] contains 108 strikes in total. However, 34 strikes have no cell towers within a 15-mile (24.1 km) radius or fall on a holiday and are dropped from our analysis. To date, BIJ claims 329 U.S. drone strikes have taken place in Yemen with the earliest occurring in 2002 and the program beginning in earnest in 2009. In our subset of strikes, 6 occur in 2010, 15 in 2011, and 53 in 2012.

BIJ and NA provide approximate strike locations for all strikes and approximate times for a subset of strikes. To determine precise start and end times for each strike, we systematically look for periods of abnormally high call volume. For each strike region (defined as a 15-mile (24.1 km) radius around each approximate strike location), we compare the outgoing call volume on the day of the strike to the average call volume of the area on the same day of the week for five weeks prior and five weeks in the future (which we term baseline days), exploiting the strong weekly periodicity in call patterns. Specifically, we form the z-score series

$$V_z = \frac{V_s - \bar{V_b}}{\sigma_{V_b}} \tag{4.4}$$

where V_s is the outgoing call volume of the area on the strike day, $\bar{V_b}$ is the average outgoing call volume of the area over the 10 baseline days, and σ_{V_b} is the standard deviation of the call volume over the 10 baseline days. When determining baseline

days, holidays and other drone strike days are skipped due to their atypical call volume. For strikes that occur during Ramadan, we only include other Ramadan weeks when constructing the baseline, as more people rest during the day and make calls at night, inverting normal call volume. Call volume is binned into 10 minute intervals to form the V_s, $\bar{V_b}$, and σ_{V_b} series. We use the periods where V_z exceeds one to determine the start and end time for each strike. For a subset of strikes, BIJ and NA provide the approximate time of day when the strike occurred. For these 25 strikes, we verify our start times correspond to the reported times. The strike period for each strike is defined as the period between our computed start and end time.

Our results and conclusions are robust to the choice of strike radius. Figure 4-8 demonstrates similar results to those discussed in the main text using a 5-mile (8.0 km) strike radius. Fig. 4-9 displays call volume and mobility as a function of radius, demonstrating strikes have an identifiable impact and our conclusions hold for radii choices up to 30 miles (48.3 km).

We also use strike-specific characteristics provided by BIJ and NA in the regressions shown in Table 4.1. These characteristics include estimates of the number of civilians killed in each strike, the number of militants killed, and whether a high-ranking militant was killed. High-ranking militants are defined as provincial al-Qaeda in the Arabian Peninsula commanders or higher rank commanders. Non-high-ranking militants are defined as local commanders as well as unranked/unspecified militants. Any militant who was known to be involved in the planning of attacks on American targets was coded as high-rank.

As a summary, 18 of the 74 strikes took place in the morning (00:00-08:00), 30 took place during the day (08:00-16:00), and 26 took place in the evening (16:00-23:59). The average strike killed 1.1 civilians (4.2 std) and 9.0 militants (9.5 std), and occurred in a district with a population of 77,051.7 (47,723.6 std). 17 of the strikes killed a high-ranking militant.

4.5.4 Methodology for cascade analysis

Proximal individuals are defined as individuals who place calls within 15 miles (24.1 km) of the strike location during the strike period. We label proximal individuals G0 as they form the zeroth generation of potential call branches. The individuals G0 contact during the strike period are labelled G1. We require G1 to be outside the strike region and to be distinct from the members of G0. This ensures any increase in G1 call activity is not due to them witnessing the event. The individuals G1 contact within an 80-minute window after being contacted by G0 are labelled G2. Again, we require G2 to be outside the event region and to be distinct from the members of any previous generation. We continue in this manner, labelling the individuals G2 contact within an 80-minute window after being contacted by G1 as G3, and so on. Figure 4-1 in the main text provides an example. The 80-minute response window is imposed as we are interested in calls made by call recipients directly after being contacted. These calls are more likely to be related to the call they just received, and therefore to the strike, than calls they make much later. At the end of this section, we demonstrate our results are robust to the choice of response window length.

For each strike and for each level of caller, we build a series of relative outgoing call volume, which compares call volume during the strike period to call volume during the 10 baseline days. As in the previous section, baseline days are the same day of the week for five weeks prior and five weeks in the future. For proximal individuals (G0), we construct the following series. The relative change in call activity for proximal individuals, R_{G0}, is defined as the number of calls made within the strike region on the day of the strike, V_s, minus the average number calls made from the strike region over the baseline days, $\bar{V_b}$, and divided by the average number of baseline calls. Formally,

$$R_{G0} = \frac{V_s - \bar{V_b}}{\bar{V_b}} \tag{4.5}$$

We bin calls into 10 minute intervals to create the time series. To build relative call volume series for G1 and higher generations, we compare the number of calls individuals make during their 80-minute windows after being contacted on strike

days to the number of calls the same individuals make during the same 80-minute windows on the baseline days. For example, if a G1 individual places a call after being contacted by a G0 individual following a strike, we check to see if the same G1 individual normally places a call during that 80-minute window on the baseline days. For $i \geq 1$, the Gi series, R_{Gi}, are defined as the number of calls made by the Gi individuals within 80 minutes after being contacted on the strike day, $V_{Gi,s}$, minus the average number of calls made by the same individuals during the same time periods on the 10 baseline days, $\bar{V}_{Gi,b}$, and then divided by $\bar{V}_{Gi,b}$. Formally,

$$R_{Gi} = \frac{V_{Gi,s} - \bar{V}_{Gi,b}}{\bar{V}_{Gi,b}} \tag{4.6}$$

Call volume series allow us to study the difference between strike day call volume and normal call activity. Dividing this difference by normal call activity converts the change in call volume to a percent change, which can be directly compared across strikes that otherwise differ in the magnitude of normal call activity. Figure 4-2 of the main text plots the relative call volume series for G0, G1, G2, and G3 averaged across the 74 strikes. The 0 on the x-axis corresponds to the start of the strike period, as defined in the previous section. Note the G1, G2, and G3 series are 0 before the strike period begins by definition as we only record calls made by these individuals after they are contacted by G0 during the strike period. The confidence intervals are formed by regressing the relative call volume series from all 74 strikes on a constant, providing sample averages as well as robust standard errors for the estimates.

Construction of the G0 series differs from the G1 and higher series as G0 individuals make at least one call during the strike period by definition. Therefore, comparing the number of calls made by G0 individuals on strike days to the number of calls made by the same individuals on baseline days would result in a mechanical increase in call volume. For this reason, we compare the number of calls made from the strike region on strike days to the number of calls made from the strike region on baseline days. Note, as we use increased strike period call volume to define a subset of the strike periods (see previous section), part of the increased strike period call volume captured

in the G0 series is mechanical. However, as multiple studies have already documented increased call volume in the vicinity of disruptive events [60–62, 149], we use the result as our starting point to study whether G1 and higher levels exhibit increased call volume. As G1+ individuals are outside of the strike region by definition, increased strike region call volume does not bias G1, G2, and G3 call volume.

To determine whether the increase in call activity of individuals several steps removed from the strike region is statistically significant, we again compare the number of calls made during strike days to the number made during baseline days. Following the approach introduced above, for each level in the cascade we record the number of calls each call recipient makes within an 80-minute window after being contacted. These observations are directly compared to the number of calls made by the same individuals during the same 80-minute periods on the 10 baseline days. For each strike and each generation of caller, we regress the number of calls made by individuals after each strike and during baseline periods on an indicator variable for strikes. The coefficient on the strike indicator therefore records the average number of calls made by call recipients on the strike day minus the average number made during the baseline days. The t-statistic of the coefficient allows us to test whether individuals make more calls on the strike day after being contacted than they normally would, via a one-sided test at a 5% level using heteroscedasticity robust standard errors. As an example, G1 individuals make 0.75 calls each on average after a strike in Lawdar on January 30th, 2012, compared to 0.21 calls during the same 80-minute windows on the 10 baseline days. The beta from the regression is therefore 0.54 and the t-statistic is 9.91, indicating a statistically significant increase in call volume during the strike period. For each strike, we first test G1 callers, then G2 callers, and so on. Once we reach a generation of callers with an insignificant increase in call volume, we do not test further generations and move on to the next strike.

Out of the 74 strikes, 71 are significant through the G1 level of callers, 46 through G2, 23 through G3, 13 through G4, and 6 through G5, at a 5% level. We restrict our primary analysis to the G0-G3 levels in the main text as only a small subset of strikes exhibit deeper cascades. To address issues of multiple testing, we apply the

Benjamini-Hochberg (BH) procedure to control the false discovery rate at 5% [113]. Under the BH step-up procedure, 71 strikes are significant through G1, 44 through G2, 18 through G3, 8 through G4, and 3 through G5.

Above, we detail our use of an 80-minute response time. Our results are robust to response time choices between 20 and 240 minutes (Fig. 4-10). Again, we test statistical significance via regression and a 5% significance level. With a 20-minute response time, 71 strikes are significant through G1, 32 through G2, and 7 through G3. With a 240-minute response time, 67 are significant through G1, 51 through G2, and 42 through G3.

4.5.5 Methodology for call pattern analysis

In order to determine which of their contacts proximal individuals choose to call, we build a baseline social network using 30 days of calls preceding each strike. These calls form an undirected graph where an edge exists between individuals if they have communicated during the month. Building the underlying network allows us to determine each proximal individual's list of contacts as well as the centrality of individuals. Proximal individuals have a median value of 24 contacts. For each proximal individual, we rank their contacts by their frequency of communication with the proximal individual during the baseline month, by their home location proximity to the proximal individual, and by their global diffusion centrality in underlying network. Frequency of communication is simply defined as the number of calls between the proximal individual and contact during the 30-day baseline period. Home location proximity is defined as the distance between the home location of the contact and the home location of the proximal individual. Home locations are defined as individuals' most frequent evening (between 6pm and midnight) tower location during the baseline period. Diffusion centrality measures the influence of a node with regards to spreading information and is formally defined below. As a global centrality measure, it is calculated using the entire baseline social network. Respectively, these metrics allow us to determine whether call recipients are the contacts proximal individuals talk to most frequently, are the contacts that live close to the individual, or are important

nodes in the underlying social network.

Diffusion centrality was introduced by Banerjee et al. (2013) [3] to identify individuals ideally situated in a social network to spread information. Formally, diffusion centrality in vector form is defined as

$$\boldsymbol{D} = \left[\sum_{t=1}^{T}(p\boldsymbol{A})^t\right] \cdot \mathbf{1} \tag{4.7}$$

where $\boldsymbol{A}$ is the adjacency matrix of the network, T is the number of periods in which information can diffuse through the network, and p is the passing probability from node to node during each period. Intuitively, the diffusion centrality of node i measures the expected number of times all individuals in the network hear about a piece of information that node i is seeded with. Motivated by our empirical results, we set $T = 3$ and $p = 0.28$ as a large number of cascades are significant through three generations of callers and roughly 28% of G1-G3 individuals pass on a call after being contacted.

Figure 4-11 provides a stylized example of ranking a proximal individual's contacts by their diffusion centrality. When ranking contacts, we break ties by assigning the tied contacts the lowest rank in the group. For example, if two contacts have diffusion centralities of 23 while the third has a diffusion centrality of 10, the two tied contacts would both be assigned rank 1 and the remaining contact would be assigned rank 3.

Figure 4-3 in the main text demonstrates the shift in calling patterns around strikes. For each of the three ranking metrics, the figure displays the fraction of calls received by each contact rank both during the 30-day baseline periods and after strikes. For example, rank 1 contacts by frequency of communication with the proximal individuals receive 28.4% of the calls during the strike periods compared to 23.0% of calls during the 30-day baseline periods. The calls made after the 74 strikes are aggregated to compute the fraction of calls received. The number of calls received by contact rank during strike periods can be compared to the number of calls received by contact rank during baseline periods via a two-sample Kolmogorov-Smirnov test, which non-parametrically tests whether two empirical distributions are sampled from

the same population distribution. We employ the test in the main text to determine whether the shifts in calling patterns around strikes are statistically significant.

Note, the ranking metrics are correlated as a proximal individual's rank 1 contact by distance can also be their rank 1 contact by frequency as well as their rank 1 contact by centrality. Across all contacts, diffusion rank is 0.26 correlated to home proximity rank and 0.45 correlated to frequency rank. Home proximity rank is 0.26 correlated to frequency rank. Reported correlations are Spearman rank-order correlation coefficients. To disentangle this correlation and determine if all of the metrics are important to whether a contact is called or not after strikes, we regress calls received after strikes for each contact on their three ranks (Table 4.5). All three ranks are statistically significant, indicating frequency of communication, home location proximity, and centrality help explain which contacts are called after strikes. Using all individuals in the baseline networks, we compute diffusion centrality percentiles. The rank 1 by centrality contacts who are called after strikes are highly globally central with a median centrality percentile of 99%.

Our results are robust to the choice of centrality measure. Figure 4-12 displays the shift in calls received after strikes for contacts ranked by their degree centrality, which is a simple centrality measure defined as the number of individuals each contact is connected to in the baseline social network. The distributional shift closely mirrors the shift discussed in the main text, with rank 1 contacts by degree centrality receiving 5.6% of calls after strikes compared to 2.8% of calls during the baseline periods. As a note, diffusion centrality with the T parameter set to 1 is proportional to degree centrality.

After the 74 strikes, 84% of calls made by proximal individuals are to one of their contacts from the baseline period. Therefore, 16% of calls made by proximal individuals correspond to edges that do not exist in the baseline network. Many of the new recipients are near the strike location, with 57% within 15 miles (24.1 km). Although the specific proximal individual-call recipient edges do not exist in the baseline network, all of the call recipient nodes are present in the baseline. For each proximal individual and new call recipient pair, we determine how close they are in

the baseline network in terms of shortest path length. 66% of proximal individual-call recipient pairs are 2 edges away from each other, meaning only one individual separates them in the baseline network. Only 4% of the new call recipients are outside of the strike region and more than 3 edges away, indicating the new contacts are geographically or socially close.

To determine the relationship between cascade size and centrality, we first regress the number of individuals in G1, G2, and G3 in each call branch present after strikes (the branch size) on the diffusion centrality of the G0 individual (the origin node). To account for the possible endogeneity of diffusion centrality, we use an instrumental variable (IV) framework estimated via two-stage least squares (2SLS). We use the number of people that call a node on Eid al-Fitr, a major Muslim holiday marking the end of Ramadan, in 2011 and 2012 as instruments for the node's centrality. Our instruments pass both the relevance and exclusion criteria for IV. Intuitively, the number of people who call a node during Eid is related to the node's social importance. Quantitatively, the number of people who call each node during 2011 Eid is 0.51 correlated with diffusion centrality and the number of people who call each node during 2012 Eid is 0.46 correlated. The two instruments are 0.54 correlated with each other. The instruments are excluded as, intuitively, the number of individuals who call a node during Eid is not related to the node's frequency of call origination after strikes, except through centrality. To strengthen this argument, we perform a Sargan over-identification test using the residuals of our 2SLS regression. With a p-value of 0.55, we fail to reject the null that our instruments are exogenous.

For robustness, we also run the reduced form model, where we regress branch size directly on our Eid instrument. We regress the number of individuals in G1, G2, and G3 in each call branch present after strikes (the branch size) on the number of people who call the G0 individual during 2011 Eid (the origin node). Using strike fixed effects and clustering standard errors by strike, the coefficient on the Eid variable is positive and significant with a value of 0.0788 and a standard error of 0.016 (T-statistic of 4.955). The regression has 74,960 observations and an R-squared of 0.075. The average number of people that call a G0 node during Eid is 6 and the max value is

111. Interpreting the regression result, if 100 people call a G0 node during Eid, which signifies that the node is important, the node will form a call branch with 7.88 more individuals than a node that receives no calls during Eid. The result is in line with our IV regression that uses diffusion centrality in the main text. However, the result is less pronounced because the number of people that call on Eid is only a proxy for centrality. Diffusion centrality is a much more direct metric for the importance of an individual, especially when it comes to diffusing information.

4.5.6 Methodology for mobility analysis

Each time an individual makes or receives a call, the tower that handles their end of the call provides their approximate location. For each individual, we can therefore build a time series of their locations and subsequently estimate their levels of mobility as well as their key locations. We use the most frequent location per hour to form an hourly time series of approximate locations for each individual. Hourly locations remove any small-scale variation that may be caused by different towers picking up calls from a stationary individual. Computing the distance between subsequent locations and summing the distances for each day provides an estimate for the daily distance travelled by each individual.

Figure 4-4 in the main text displays the daily distance travelled by proximal individuals, where mobility is averaged across all individuals from the 74 strikes. 95% confidence intervals are shown in light blue and are computed by regressing the mobility of individuals on a constant, providing sample averages as well as robust standard errors for the estimates. As a complement to Fig. 4-4, Table 4.3 reports results from a regression of daily distance travelled by all proximal individuals on a binary indicator for strike days. Specifically, we include the daily distance travelled by each proximal individual for 14 days preceding the strike, associated with a strike indicator variable of 0, and the daily distance travelled by proximal individuals on the day of the strike, associated with a strike indicator variable of 1. Interpreting the coefficient of the regression, we find daily distance travelled increases by 7.64 km on strike days (with a standard error of 0.24) compared to average distance travelled over

the preceding two weeks, a 27% increase over the pre-strike mean of 28.5 km. The full specification of the regression is provided in the table caption.

To determine whether the increase in mobility is statistically significant for each strike, we regress the daily distance travelled by proximal individuals on an indicator variable for strike days. Again, for each strike, we include the daily distance travelled by each proximal individual for 14 days preceding the strike, associated with a strike indicator variable of 0, and the daily distance travelled by proximal individuals on the day of the strike, associated with a strike indicator variable of 1. The coefficients of the regressions therefore report the average daily distance travelled on the strike day minus the average daily distance travelled over the preceding 14 days. The t-statistics of the coefficients allow us to test whether individuals are more mobile on strike days than normal, via a one-sided test at a 5% level using heteroscedasticity robust standard errors. Out of the 74 strikes, 43 display statistically significant increases in mobility on strike days at a 5% level. To address issues of multiple testing, we apply the Benjamini-Hochberg (BH) procedure to control the false discovery rate at 5% [113]. Under the BH step-up procedure, 40 strikes have significant increases in mobility on strike days.

A concern when using CDRs to estimate mobility is potential bias caused by correlation between call frequency and mobility levels. However, we first note we form location estimates at an hourly level, so our results are not sensitive to increased call frequency within hour intervals. To counter any further potential bias, we rerun the daily distance travelled analysis using the subset of individuals who have more location estimates on non-strike days than strike days. The average daily distance travelled by this subset spikes 23% on strike days compared to the average distance travelled over the preceding two weeks, a similar increase to the 27% reported in the main text. An alternative measure to daily distance travelled is distance between the first and last call made each day. This first-last measure captures spikes in mobility if proximal individuals are present in the strike region in the morning of strike days and leave the strike region after strikes. When restricted to the subset of individuals who make at least two calls each day, this measure has the added benefit of robustness to

call frequency. Average daily first-last distance spikes 17% on strike days compared to the average distance over the preceding two weeks.

Figure 4-5A in the main text shows the locations of all 74 strikes and the locations of all proximal individuals at the time of their calls during the strike periods. By definition, all individuals are located within 15 miles (24.1 km) of the strike locations. Fig. 4-5B then shows the locations of the same proximal individuals 24 hours after their strike period calls, demonstrating a fraction leave the strike region within 24 hours and disperse around the country.

We study the proximal individuals who live within the strike region, leave the strike region within 24 hours of strikes, and remain away for at least 24 hours. Again, home locations are defined as each individual's most frequent evening tower location during the 30-day baseline period preceding each strike. For this subset of proximal individuals, Fig. 4-13 displays the distribution of the duration of time they remain away from the strike region. While 51% return quickly and are home within five days, 1046 individuals do not return to their hometowns within a 30-day period.

We find physical and social networks explain the movement of individuals after strikes. 54% end up within 5 miles (8.0 km) of a major city during the time they are away from home. The cities we consider are Aden, Al Hudaydah, Al Qatn, Ataq, Azzan, Bayda, Dhamar, Ibb, Ja'ar, Marib, Mudiyah, Mukalla, Rada'a, Sana'a, Ta'izz, and Zinjibar. The population of each city, which is used in the regression described in Table 4.6, is provided by [157]. Table 4.6 regresses the fraction of proximal individuals who flee to each city after each strike on the city's distance to the strike location and population. The regression demonstrates individuals who flee move towards nearby, densely populated cities. Table 4.7 regresses the maximum distance from the strike region of each proximal individual who flees on the distance between their most important contacts and the strike region and on the distance between their nearby, major city and the strike region. The regression demonstrates physical and social information can be used to predict the mobility of those who flee.

4.5.7 Comparison to other disruptive events

We compare the drone strike response patterns to the responses to other disruptive events. Specifically, we study the CDR response to a factory explosion and four bombings. On March 28, 2011, a munitions factory exploded in Ja'ar, a town in southern Yemen, killing 150 people [158]. On June 3, 2011, the Presidential Palace in Sana'a was bombed by opposition forces and the president at the time, Ali Abdullah Saleh, was badly injured [159]. On February 25, 2011, a car bomb was set off outside the Presidential Palace in Al Mukalla, killing 26 people [160]. On May 21, 2012, al-Qaeda carried out a suicide bombing against Yemeni soldiers practicing for a parade in Sana'a, killing over 90 [161]. On July 11, 2012, a suicide bomber killed at least 10 people outside of a police academy in Sana'a [162]. Similar to drone strikes, these events are violent, unexpected, and localized in time and space.

Following the methodology in the main text, we analyze the calling cascades that emerged after each event. All five events display significant cascades (Fig. 4-14), with statistically significant increases in call volume through the G2, G4, G5, G5, and G5 level of callers at a 5% level, respectively. Analyzing the shifts in calling patterns after the events as in the main text, we see geographically close and frequent contacts receive a greater fraction of calls after the events than during the baseline periods, for the majority of events (Fig. 4-15). Calls to central contacts increase substantially after the events. Analyzing mobility as in the main text, we see average mobility (measured by daily distance travelled) spikes on the majority of events days (Fig. 4-16). Compared to baseline mobility (average mobility over the preceding two weeks), average mobility increased 7.5 km (SE 0.40), 0.4 km (SE 0.06), 3.9 km (SE 0.37), 1.4 km (SE 0.06), and 3.0 km (SE 0.14) respectively on event days over pre-event means of 8.9 km, 4.5 km, 7.6 km, 6.4 km, and 7.9 km. These increases correspond to percent increases of 84%, 8%, 52%, 23%, and 38% over the pre-event means. After the events, 143, 966, 106, 865, and 1292 individuals who lived within the event region, left within 24 hours and remained away for at least 24 hours, corresponding to 1.7% of the proximal individuals on average. Notably, zero individuals remained away for

30 days or more across all five events.

4.5.8 Supplementary figures

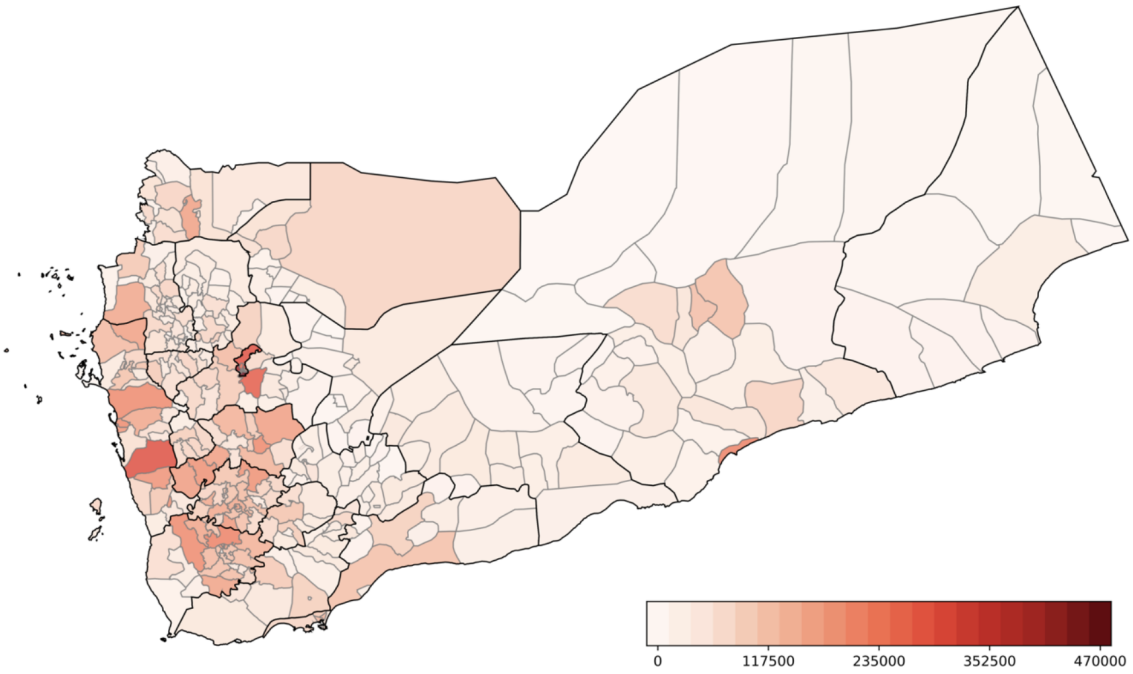
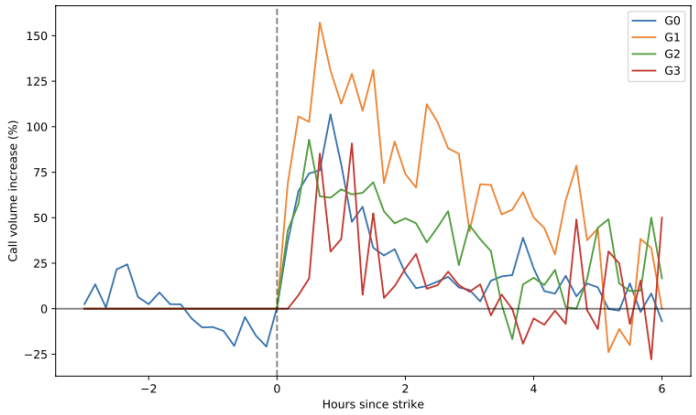
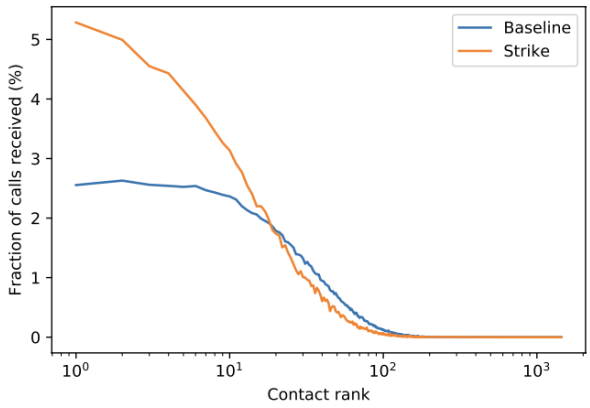


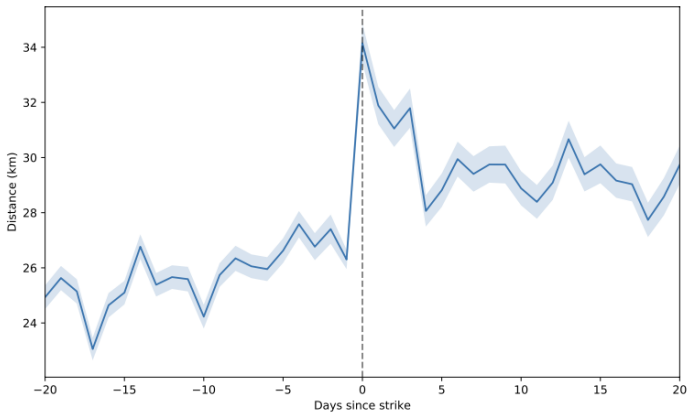
Figure 4-7: 2010 population of Yemen by district. Administratively, Yemen is split into 22 governorates and subdivided into 333 districts. The total population of the country in 2010 was around 23,607,000 [109]. District-level populations are provided by [150].



(A)



(B)



(C)

Figure 4-8: Main results using a 5-mile (8.0 km) strike radius for robustness. **(A)** displays call volume by generation of caller, averaged across strikes, relative to call volume on non-strike baseline days, **(B)** displays the fraction of calls received by contacts ranked by their diffusion centrality, and **(C)** displays average daily distance travelled by proximal individuals around strikes.

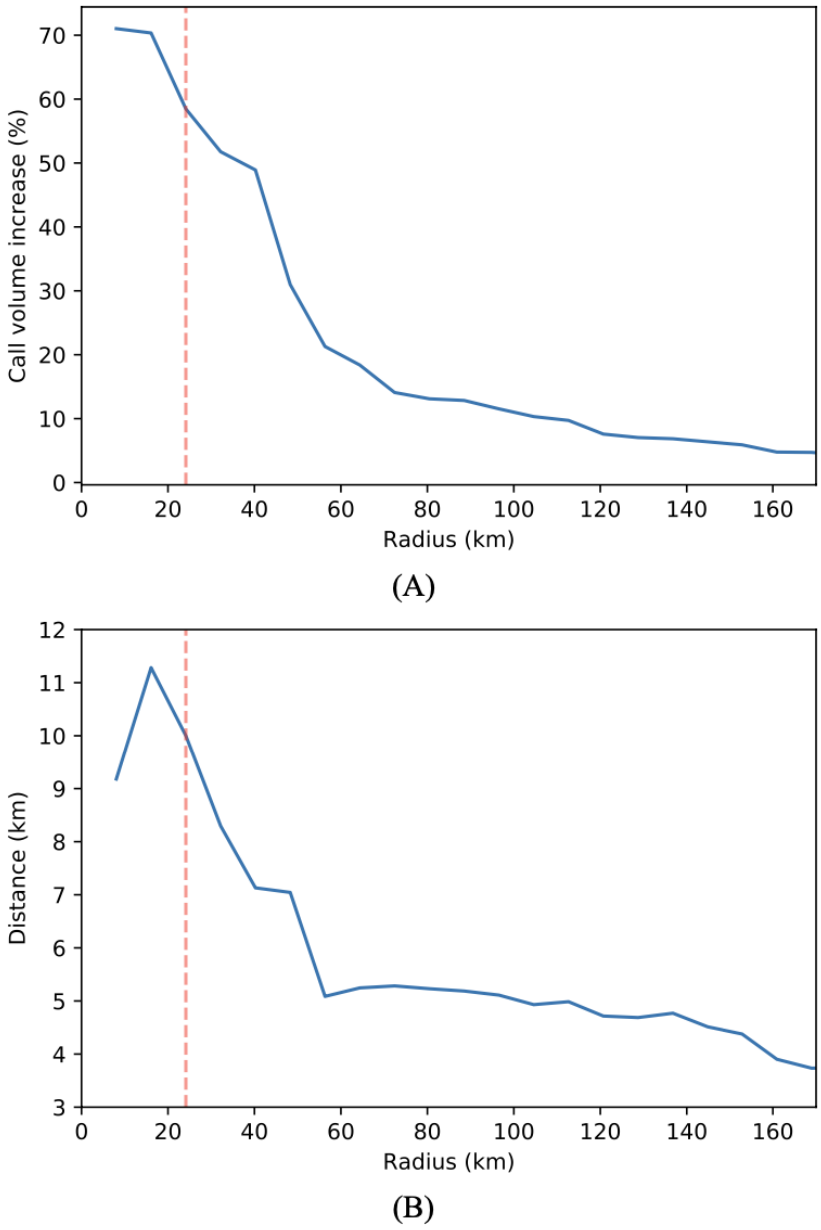


Figure 4-9: Call volume and mobility as a function of the strike region radius for robustness. **(A)** displays the average call volume increase of proximal individuals during the strike period relative to the call volume from the same area during the same period on non-strike baseline days. **(B)** displays the jump in average daily distance travelled by proximal individuals, defined as distance travelled on strike days minus average distance travelled over the preceding two weeks. Daily distance travelled is first averaged within strike and then averaged across strikes. Both (A) and (B) indicate strikes have an identifiable impact for radii choices between 5 and 30 miles (8.0 and 48.3 km). The dashed line is placed at 15 miles (24.1 km), the strike region radius used in our analysis.

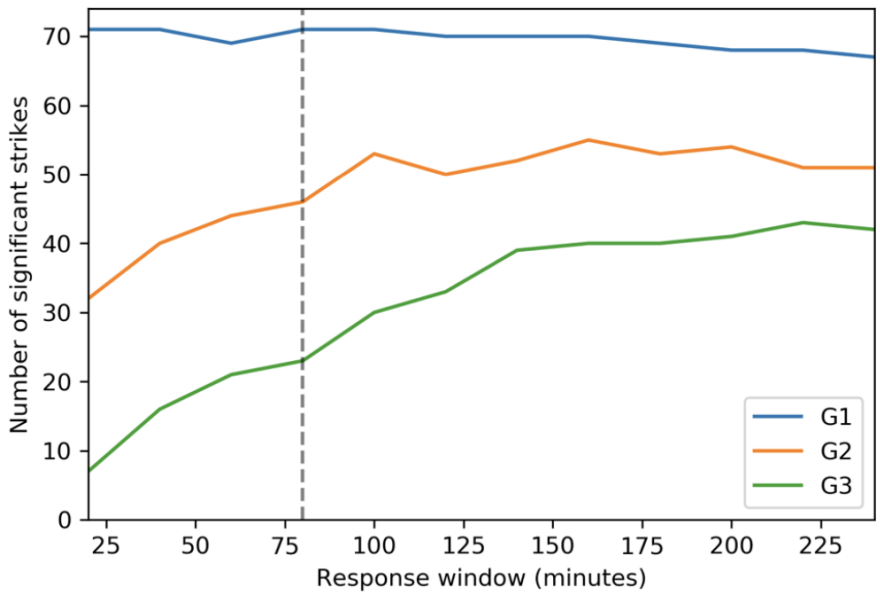


Figure 4-10: Number of strikes with significant cascade generations by response window length for robustness. In our analysis, the individuals G1 contact within an 80-minute window after being contacted by G0 are labelled G2. G3 individuals are defined similarly. Varying this response window length, we note our conclusions regarding cascades are robust, as many strikes have significant cascades through several generations of callers using window lengths between 20 and 240 minutes.

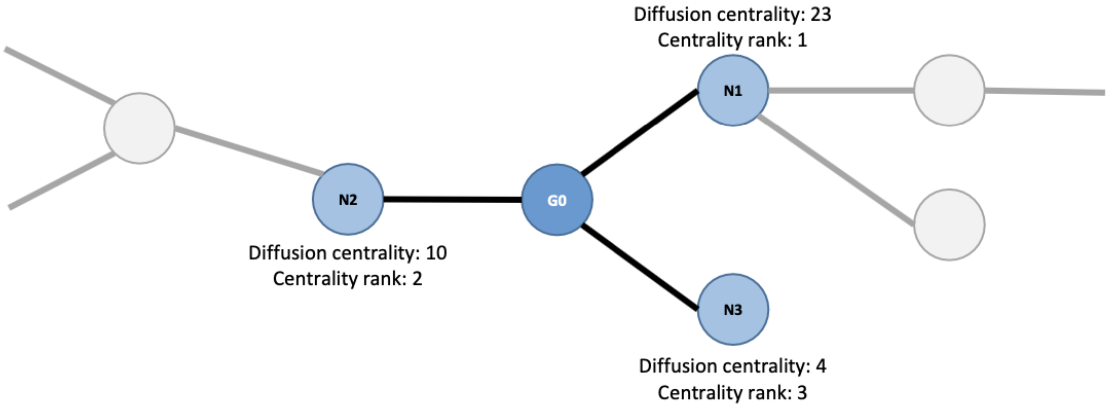


Figure 4-11: A stylized example of ranking the contacts (N1-3) of a proximal individual (G0) by their centrality scores. The proximal individual has three contacts with whom they have corresponded during the 30-day baseline preceding the strike. Their rank 1 contact by centrality has a diffusion centrality of 23, which is calculated using the entire baseline social network.

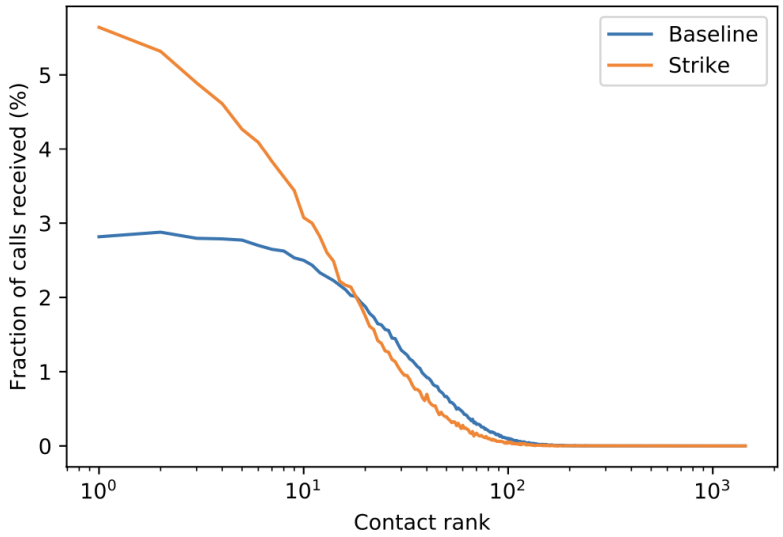


Figure 4-12: Fraction of calls received by contacts ranked by their degree centrality for robustness. Degree centrality of a contact is defined as the number of individuals they are connected to in the baseline social network. The increase in calls received by important low rank contacts after strikes mirrors the shift present for contacts ranked by diffusion centrality, as discussed in the main text. Rank 1 contacts by degree centrality receive 5.6% of calls after strikes compared to 2.8% of calls during the baseline periods.

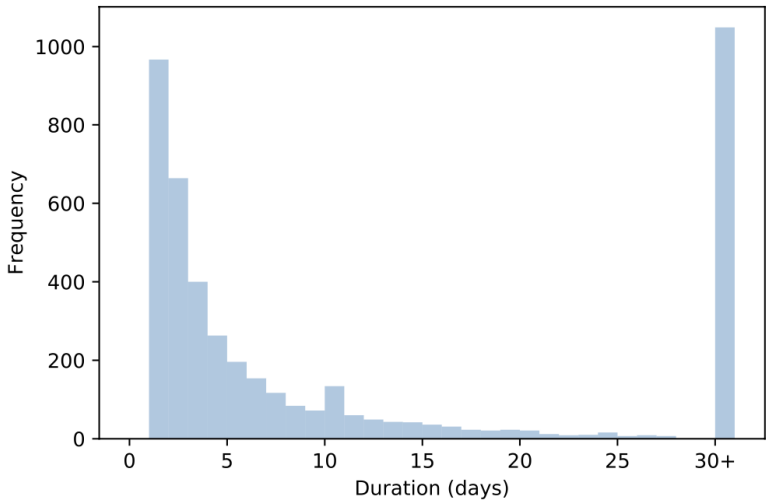


Figure 4-13: Distribution of the duration of time proximal individuals who leave after strikes remain away from the strike region. The subset of proximal individuals is restricted to those who live within the strike region, leave within 24 hours after strikes, and remain away for at least 24 hours.

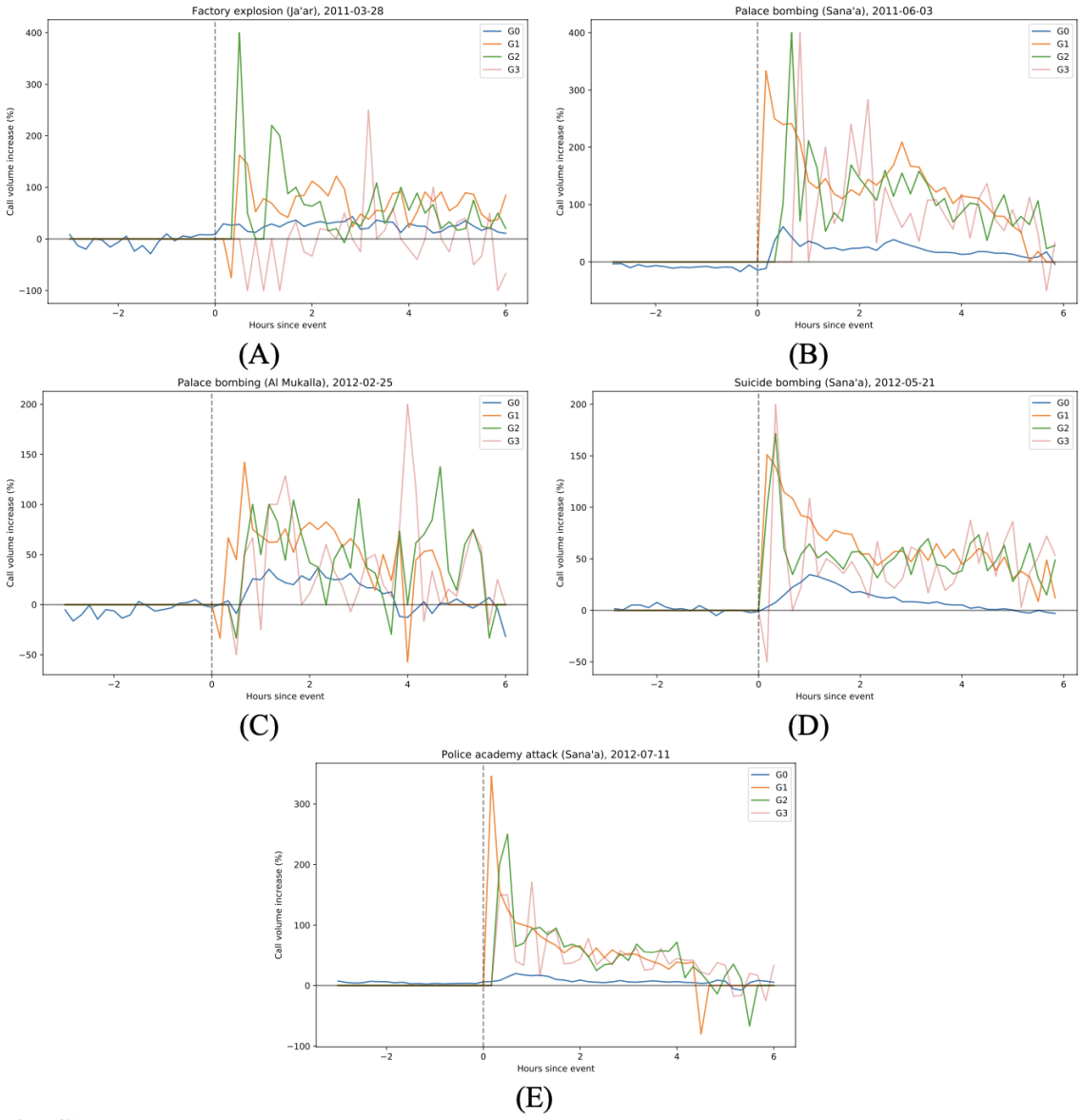


Figure 4-14: Call volume by generation of caller highlighting the emergence of calling cascades after the five comparison events: **(A)** factory explosion in Ja'ar, **(B)** Presidential Palace bombing in Sana'a, **(C)** Presidential Palace bombing in Al Mukalla, **(D)** suicide bombing in Sana'a, and **(E)** police academy bombing in Sana'a. Call volume on the day of the events is compared to call volume from the same area on baseline days to provide an increase in call volume, as in the main text. G0 individuals are proximal to the event and contact G1 individuals who contact G2 individuals, and so on.

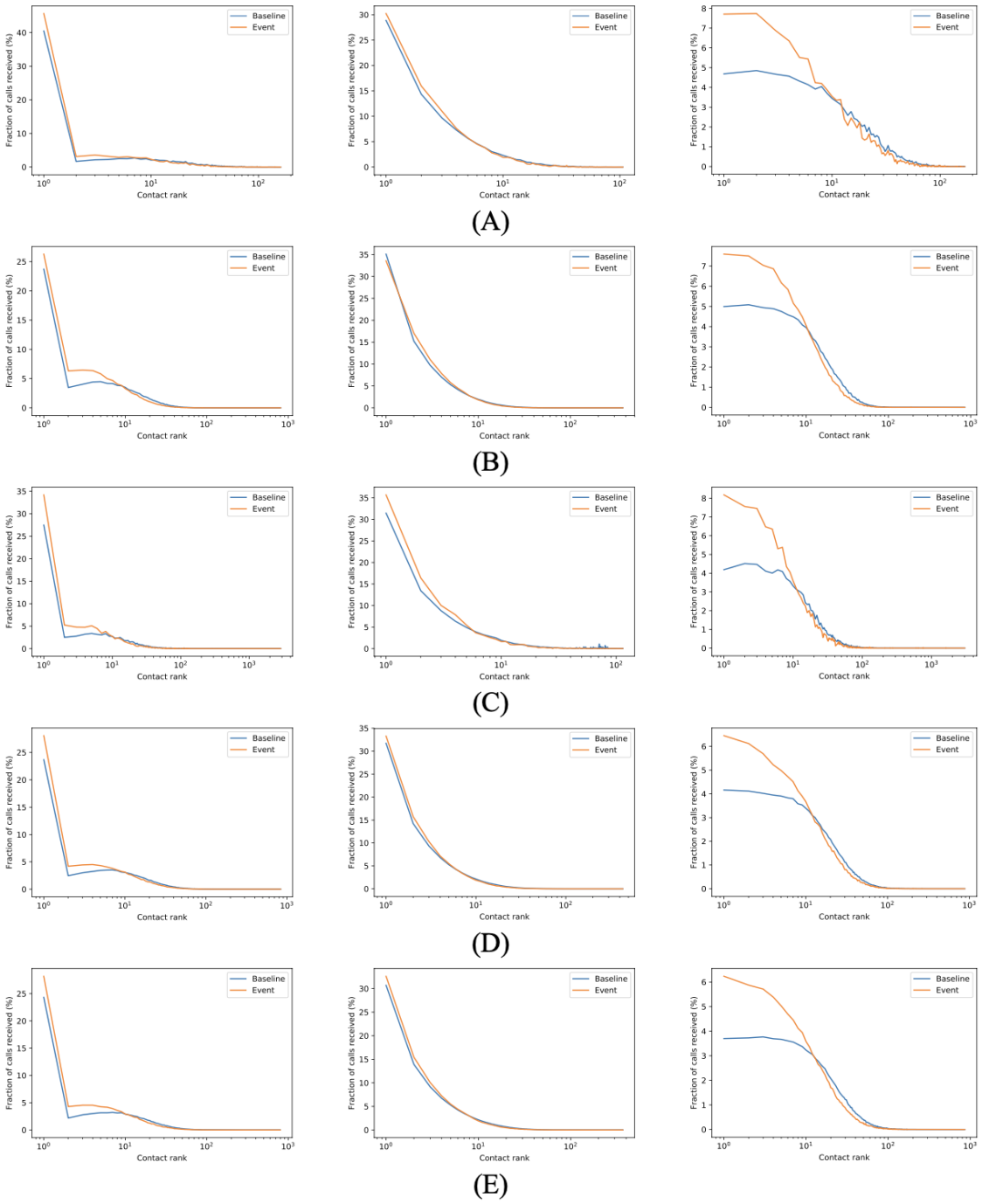


Figure 4-15: Shifts in calling patterns after the five comparison events: **(A)** factory explosion in Ja'ar, **(B)** Presidential Palace bombing in Sana'a, **(C)** Presidential Palace bombing in Al Mukalla, **(D)** suicide bombing in Sana'a, and **(E)** police academy bombing in Sana'a. The left, center, and right columns display the fraction of calls received by contacts ranked by their home location proximity to the proximal individual, frequency of communication with the proximal individual during the baseline period, and diffusion centrality, respectively. Across all three metrics, important low rank contacts receive a larger fraction of calls after the events than during the baseline periods.

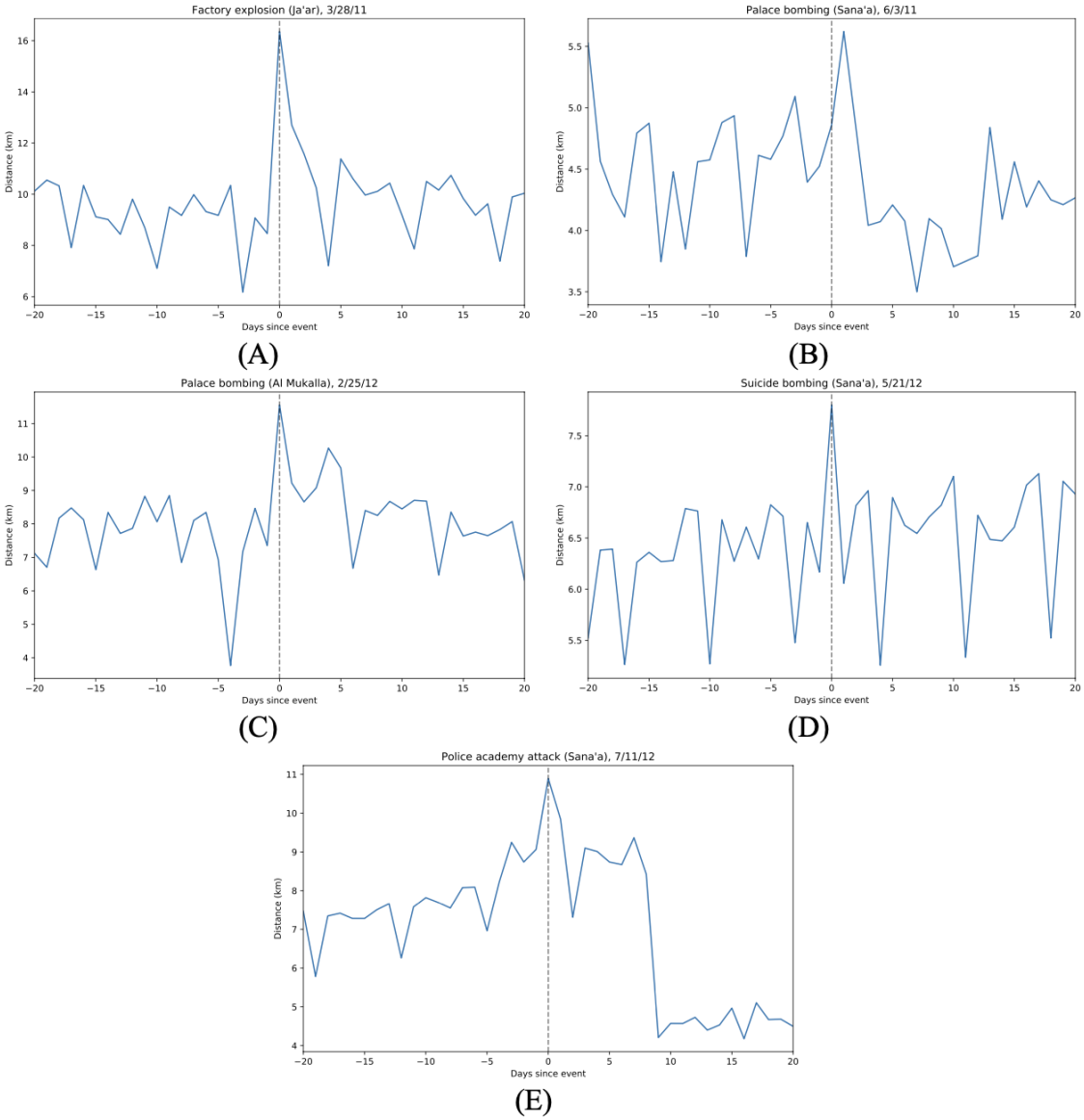


Figure 4-16: Spikes in mobility around the five comparison events: **(A)** factory explosion in Ja'ar, **(B)** Presidential Palace bombing in Sana'a, **(C)** Presidential Palace bombing in Al Mukalla, **(D)** suicide bombing in Sana'a, and **(E)** police academy bombing in Sana'a. The figures display average daily distance travelled by proximal individuals around the events. Daily distance travelled is computed for each proximal individual by constructing a time series of their locations and computing the distance between them.

4.5.9 Supplementary tables

Time	Anonymized ID		Tower ID		Duration
Start of Call	Caller	Recipient	Caller	Recipient	(seconds)
6/13/12 7:57:12	34331363	30002125	3995	623	1139
6/13/12 7:57:13	20918574	38389599	1373	1290	10
6/13/12 7:57:13	39496809	146551	3338	1401	32

Table 4.4: Three calls from the call detail record dataset, provided as an example.

	Rank Frequency	Rank Diffusion	Rank Distance	Intercept	R^2 / NObs
Beta	-0.001155	-0.000026	-0.000043	0.062	0.9%
T-Stat	(-194.2)	(-35.9)	(-56.3)	(244.3)	3,698,224

Table 4.5: Disentangling correlation between contact ranks. We regress the number of calls received by contacts during the strike period minus their expected number of calls on their frequency, proximity, and centrality ranks. Expected number of calls is defined as the rate of calls received during the 30-day baseline period (number of calls received divided by the period length in hours) times the length of the strike period in hours. For example, if a contact talks to a proximal individual once every two hours during the baseline period, we expect them to speak once during a two-hour strike period. All three ranks are statistically significant, indicating frequency of communication, centrality, and home location proximity help explain which contacts are contacted after strikes. The regression uses heteroscedasticity robust standard errors.

	Distance (km)	Pop. (10,000s)	Intercept	R^2 / NObs
Beta	-0.00040	0.00085	0.13556	22.5%
T-Stat	(-10.15)	(7.14)	(10.88)	908

Table 4.6: Individuals who flee move towards nearby, densely populated cities. 54% of proximal individuals who live within the strike region, leave within 24 hours after strikes, and remain away for at least 24 hours end up within 5 miles (8.0 km) of major cities. We regress the fraction of proximal individuals who flee to each city after each strike on the city's distance to the strike location and population. Interpreting the coefficients, a city of 100,000 people 30 km away from a strike would receive 13.2% of the proximal individuals who flee. The regression uses heteroscedasticity robust standard errors.

	To Rank 1 by Freq.	To Rank 1 by Centrality	To Call Recipient	To Predicted City	Intercept	R^2 / NObs
Beta	0.255	0.213	0.068	0.157	19.837	24.6%
T-Stat	(8.67)	(8.15)	(2.52)	(5.65)	(3.31)	4265

Table 4.7: Physical and social networks can be used to predict the mobility of those who flee. We regress the maximum distance from the strike region of each proximal individual who flees (those that live within the strike region, leave within 24 hours of the strike, and remain away for at least 24 hours) on the distance between their rank 1 contact by frequency of communication and the strike region, the distance between their rank 1 contact by diffusion centrality and the strike region, the distance between the individual they call directly after the strike and the strike region, and the distance between their predicted preferred city and the strike region. The predicted preferred city is determined for each strike using the regression in Table 4.6. Note these four locations account for one quarter of the variation in the distance proximal individuals travel. All distances are in kilometers and the regression uses heteroscedasticity robust standard errors.

Chapter 5

Conclusion

This thesis studies inference and diffusion in networks. We focus on epidemic spread and information diffusion in social networks, and analyze these processes by applying and extending ideas from statistical inference. We utilize estimation, testing, and uncertainty quantification to rigorously analyze data. This thesis utilizes both theory and data in order to address several real-world challenges.

In the first chapter, we study epidemic spread and introduce an approach to efficiently identify infected individuals. Our approach utilizes network structure to improve group testing. We demonstrate that grouping individuals by their community for an initial stage of testing outperforms the most common form of group testing, Dorfman testing, in terms of the number of tests needed, the number of false positives, and the number of false negatives. The extent of outperformance is determined by the strength of community structure in the network. Importantly, network grouping is simple for practitioners to implement. In practice, individuals can be grouped by family, social group, or some other community structure.

Our work on network grouping opens several fruitful areas for future research. Future work can analyze the performance of network grouping under different network structures, epidemic models, and community detection algorithms. In our work, we implement network grouping by grouping individuals by community. However, future work can utilize other network information and group individuals by clique, cluster, centrality, or some other network characteristic. In addition, covariate information,

such as an individual's demographics and clinical results, can supplement and enhance network grouping. The network grouping approach can also be applied to one-stage group testing algorithms, which may produce fewer false negatives. Finally, network grouping can be applied to non-medical settings, such as communication networks, cybersecurity, and compressed sensing.

In the second chapter, we continue our analysis of group testing. We analyze the performance of Dorfman testing, the most common approach to group testing in practice. We derive the distribution of the number of tests needed, the number of false positives, and the number of false negatives under conditions faced by medical practitioners. The full distributions provide confidence intervals and better guidance for practitioners. Recognizing real-world conditions, we allow for different first and second stage false positive and false negative rates. We also model first-stage false negative rates as dependent on the number of samples in each group, which accounts for viral-load dilution. We have built a dashboard that implements our results and allows practitioners to analyze the performance of group testing under various parameters.

Moving forward, theoreticians should work with medical practitioners to design and implement group testing. Both sides would benefit from close collaboration. Researchers can derive and explain the performance of the approaches currently used in practice and can also derive new, improved approaches. Practitioners can explain the flexibility and obstacles they face in practice, resulting in more realistic theoretical testing approaches. A close collaboration would result in better and more prevalent group testing, which would allow for efficient testing and screening of large populations. On the theory side, research can explore modeling sensitivity as a function of group size more deeply. Different functional forms should be considered and evaluated. In addition, optimal group size should be studied further as modeling sensitivity as a function of group size affects the optimum. Importantly, research should evaluate how to design approaches that balance different outcomes. Minimizing the number of tests needed may not be the right approach if the chosen group size results in a large number of false positives and negatives. Ideally, all of the performance metrics must be considered and balanced.

In the third chapter, we study information diffusion where, similar to the epidemic spread of previous chapters, something spreads from individual to individual through a social network. Instead of an infection spreading through a population, we now consider a piece of news, information, or gossip infecting individuals. We study information exchange between individuals and introduce a statistical testing framework to identify cascades in network data. We consider a network setting where branches form both during normal and abnormal periods. We introduce a test statistic that distinguishes between the large branches formed during abnormal periods, which we term cascades, and the small branches formed during normal periods. Call detail records provide the motivating example, because we would like to identify large call branches that form after disruptive events. Our test statistic compares observed average branch size to expected branch size under the null. We introduce a null model and derive the expected size and variance of branches under the null using ideas from branching processes. Our test statistic is semiparametric, consistent, and asymptotically normal. We apply our statistic to call detail records from Yemen to quantify the significance of a calling cascade formed after the Presidential Palace was bombed. We also use our statistic to identify several violent events during the Yemeni Revolution.

Going forward, the test statistic should be utilized in empirical network science research when cascade formation is being studied, as the statistic adds significance levels to observed branch structures. The statistic can be applied to additional call detail record datasets to detect previously unreported disruptive events, such as state-sponsored attacks on civilians. It can also be applied to other network datasets such as social media networks to identify abnormally large cascades of information, news, and opinions. Applying the testing framework to Twitter data would highlight significant retweet chains. It could additionally identify substantial discussions and topics based on abnormally large branches of tweets. An application to shared posts on Facebook could be used to highlight viral posts, with a possible focus on identifying viral fake news.

In the fourth chapter, we study the social network effects of drone strikes, focusing on information and physical diffusion around strikes. Following the previous chapter,

we analyze the emergence of information diffusion following localized events. We focus on the formation of calling cascades around drone strikes. Drone strikes have become a fixture of modern warfare, yet their effects and effectiveness remain opaque and fiercely debated. Utilizing a new dataset of over 12 billion call detail records, we study the causal impact of 74 U.S. drone strikes on communication and mobility in Yemen between 2010 and 2012. Over 95% of strikes are followed by calling cascades, with roughly one third exhibiting increased call volume through four levels of callers. Compared to non-strike periods, proximal individuals call their frequent and geographically close contacts more frequently. Notably, socially central individuals are called twice as often and proceed to spark large calling cascades. Lastly, physical mobility increases 27% on strike days compared to the pre-strike mean and thousands of individuals flee their hometowns. These findings demonstrate drone strikes have a disruptive and widespread impact on civilian life. Furthermore, our results imply information, opinions, and emotions regarding strikes spread quickly through the population, which is in contrast to the prevailing political and military position that strikes are surgical.

Although we demonstrate a disruptive impact on civilian life and the presence of diffusion, we lack the content of the communications and are thus unable to analyze exactly how opinions and loyalties shift around these events. Future research should employ further data sources, such as media coverage, speeches, and sermons, to understand which information and emotions spread after strikes. An open question remains of whether the disruption induced by strikes increases or decreases militant recruitment. Moving forward, researchers should coordinate with policy makers to design improved, data-driven policy. Updated strategies for conflict prevention and resolution are needed to resolve modern conflicts, including the ongoing wars in the Middle East, that have lasted decades and claimed hundreds of thousands of lives.

Bibliography

[1] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The anatomy of the facebook social graph," *arXiv preprint arXiv:1111.4503*, 2011.

[2] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[3] A. Banerjee, A. G. Chandrasekhar, E. Duflo, and M. O. Jackson, "The diffusion of microfinance," *Science*, vol. 341, no. 6144, 2013.

[4] M. Elliott, B. Golub, and M. O. Jackson, "Financial networks and contagion," *American Economic Review*, vol. 104, no. 10, pp. 3115–3153, 2014.

[5] I. Dobson, B. A. Carreras, V. E. Lynch, and D. E. Newman, "Complex systems analysis of series of blackouts: cascading failure, critical points, and self-organization," *Chaos*, vol. 17, no. 2, 2007.

[6] M. Rosas-Casals, S. Valverde, and R. V. Solé, "Topological vulnerability of the european power grid under errors and attacks," *International Journal of Bifurcation and Chaos*, vol. 17, no. 07, pp. 2465–2475, 2007.

[7] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[8] M. Newman, *Networks*. Oxford University Press, 2nd ed., 2018.

[9] D. Easley and J. Kleinberg, *Networks, crowds, and markets: reasoning about a highly connected world*. Cambridge University Press, 2010.

[10] M. O. Jackson, *Social and economic networks*. Princeton University Press, 2008.

[11] P. Sapiezynski, A. Stopczynski, D. D. Lassen, and S. Lehmann, "Interaction data from the Copenhagen Networks Study," *Scientific Data*, vol. 6, no. 1, p. 315, 2019.

[12] P. Erdos, A. Rényi, *et al.*, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, no. 1, pp. 17–60, 1960.

[13] E. N. Gilbert, "Random graphs," *The Annals of Mathematical Statistics*, vol. 30, no. 4, pp. 1141–1144, 1959.

[14] R. Durrett, *Random Graph Dynamics*. Cambridge University Press, 2007.

[15] S. U. Pillai, T. Suel, and S. Cha, "The perron-frobenius theorem: some of its applications," *IEEE Signal Processing Magazine*, vol. 22, no. 2, pp. 62–75, 2005.

[16] A. Jadbabaie, J. Lin, and A. S. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Transactions on automatic control*, vol. 48, no. 6, pp. 988–1001, 2003.

[17] H. G. Tanner, A. Jadbabaie, and G. J. Pappas, "Flocking in fixed and switching networks," *IEEE Transactions on Automatic control*, vol. 52, no. 5, pp. 863–868, 2007.

[18] R. Pastor-Satorras and A. Vespignani, *Evolution and structure of the Internet: A statistical physics approach*. Cambridge University Press, 2004.

[19] F. A. Longstaff, "The subprime credit crisis and contagion in financial markets," *Journal of Financial Economics*, vol. 97, no. 3, pp. 436–450, 2010.

[20] K. Bryan and T. Leise, "The $25,000,000,000 eigenvector: The linear algebra behind Google," *SIAM review*, vol. 48, no. 3, pp. 569–581, 2006.

[21] A. Sinha, "Network optimization algorithms at Amazon." `https://linkedin.com/pulse/network-optimization-algorithms-amazon-13-amitabh-sinha/`, 2019.

[22] M. D. Veirman, V. Cauberghe, and L. Hudders, "Marketing through instagram influencers: the impact of number of followers and product divergence on brand attitude," *International Journal of Advertising*, vol. 36, no. 5, pp. 798–828, 2017.

[23] Board of Governors, Federal Reserve System, "Dodd-Frank act stress test 2019: Supervisory stress test results." `https://federalreserve.gov/publications/files/2019-dfast-results-20190621.pdf`, 2019.

[24] M. E. Newman, "A measure of betweenness centrality based on random walks," *Social networks*, vol. 27, no. 1, pp. 39–54, 2005.

[25] P. S. Park, J. E. Blumenstock, and M. W. Macy, "The strength of long-range ties in population-scale social networks," *Science*, vol. 362, pp. 1410–1413, 2018.

[26] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, p. 026113, Feb 2004.

[27] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "Group formation in large social networks: membership, growth, and evolution," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 44–54, 2006.

[28] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *nature*, vol. 393, no. 6684, p. 440, 1998.

[29] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.

[30] J. M. Kleinberg, "Navigation in a small world," *Nature*, vol. 406, no. 6798, pp. 845–845, 2000.

[31] J.-C. Delvenne, S. N. Yaliraki, and M. Barahona, "Stability of graph communities across time scales," *Proceedings of the national academy of sciences*, vol. 107, no. 29, pp. 12755–12760, 2010.

[32] A. Tahbaz-Salehi and A. Jadbabaie, "A necessary and sufficient condition for consensus over random networks," *IEEE Transactions on Automatic Control*, vol. 53, no. 3, pp. 791–795, 2008.

[33] R. Cowan and N. Jonard, "Network structure and the diffusion of knowledge," *Journal of economic Dynamics and Control*, vol. 28, no. 8, pp. 1557–1575, 2004.

[34] S. Goel, D. J. Watts, and D. G. Goldstein, "The structure of online diffusion networks," in *Proc. 13th ACM Conf. Electronic Commerce (EC 12)*, pp. 623–638, 2012.

[35] G. Miritello, E. Moro, and R. Lara, "Dynamical strength of social ties in information spreading," *Physical Review E*, vol. 83, no. 4, 2011.

[36] K. Starbird and L. Palen, "(how) will the revolution be retweeted?: information diffusion and the 2011 egyptian uprising," in *Proc. ACM Conf. Computer Supported Cooperative Work (CSCW 12)*, pp. 7–16, 2012.

[37] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Physical Review Letters*, vol. 86, no. 14, pp. 3200–3203, 2001.

[38] C. Dye and N. Gay, "Modeling the sars epidemic," *Science*, vol. 300, no. 5627, pp. 1884–1885, 2003.

[39] M. J. Keeling and K. T. Eames, "Networks and epidemic models," *Journal of the royal society interface*, vol. 2, no. 4, pp. 295–307, 2005.

[40] P. Van Mieghem, "The n-intertwined sis epidemic network model," *Computing*, vol. 93, no. 2-4, pp. 147–169, 2011.

[41] M. Keeling, "The implications of network structure for epidemic dynamics," *Theoretical population biology*, vol. 67, no. 1, pp. 1–8, 2005.

[42] H. Wang, Q. Li, G. D'Agostino, S. Havlin, H. E. Stanley, and P. Van Mieghem, "Effect of the interconnected network structure on the epidemic threshold," *Physical Review E*, vol. 88, no. 2, p. 022801, 2013.

[43] K. Drakopoulos, A. Ozdaglar, and J. N. Tsitsiklis, "When is a network epidemic hard to eliminate?," *Mathematics of Operations Research*, vol. 42, no. 1, pp. 1–14, 2017.

[44] L. Chen and J. Sun, "Optimal vaccination and treatment of an epidemic network model," *Physics Letters A*, vol. 378, no. 41, pp. 3028–3036, 2014.

[45] B. Laughlin, "Information cascades and refugee crises: Evidence from kosovo," *Unpublished manuscript*, 2018.

[46] N. B. Weidmann, "Communication networks and the transnational spread of ethnic conflict," *Journal of Peace Research*, vol. 52, no. 3, pp. 285–296, 2015.

[47] S. Aral and D. Walker, "Creating social contagion through viral product design: A randomized trial of peer influence in networks," *Management science*, vol. 57, no. 9, pp. 1623–1639, 2011.

[48] A. Banerjee, A. G. Chandrasekhar, E. Duflo, and M. O. Jackson, "Using gossips to spread information: Theory and evidence from two randomized controlled trials," *The Review of Economic Studies*, vol. 86, no. 6, pp. 2453–2490, 2019.

[49] P. J. Bickel and K. A. Doksum, *Mathematical Statistics*, vol. 1. Chapman and Hall, 2015.

[50] G. W. Imbens and D. B. Rubin, *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[51] R. Dorfman, "The detection of defective members of large populations," *The Annals of Mathematical Statistics*, vol. 14, no. 4, pp. 436–440, 1943.

[52] T. T. Van, J. Miller, D. M. Warshauer, E. Reisdorf, D. Jernigan, R. Humes, and P. A. Shult, "Pooling nasopharyngeal/throat swab specimens to increase testing capacity for influenza viruses by PCR," *Journal of Clinical Microbiology*, vol. 50, no. 3, pp. 891–896, 2012.

[53] C. S. McMahan, J. M. Tebbs, and C. R. Bilder, "Informative Dorfman screening," *Biometrics*, vol. 68, no. 1, pp. 287–296, 2012.

[54] FDA, "In vitro diagnostics EUAs - molecular diagnostic template for laboratories." `https://www.fda.gov/medical-devices/coronavirus-disease-2019-covid-19-emergency-use-authorizations-medical-devices/vitro-diagnostics-euas`, July 28, 2020.

[55] C. A. Hogan, M. K. Sahoo, and B. A. Pinsky, "Sample pooling as a strategy to detect community transmission of SARS-CoV-2," *Journal of the American Medical Association*, April 2020.

[56] J. H. Fowler and N. A. Christakis, "Cooperative behavior cascades in human social networks," *Proc. National Academy of Sciences*, vol. 107, no. 12, pp. 5334–5338, 2010.

[57] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Transactions on the Web*, vol. 1, no. 1, 2007.

[58] S. Sreenivasan, K. S. Chan, A. Swami, G. Korniss, and B. K. Szymanski, "Information cascades in feed-based networks of users with limited attention," *IEEE Trans. Network Science and Engineering*, vol. 4, no. 2, pp. 120–128, 2016.

[59] D. J. Watts, "A simple model of global cascades on random networks," *Proc. National Academy of Sciences*, vol. 99, no. 9, pp. 5766–5771, 2002.

[60] J. P. Bagrow, D. Wang, and A. L. Barabasi, "Collective response of human populations to large-scale emergencies," *PLoS One*, vol. 6, no. 3, 2011.

[61] L. Gao, C. Song, Z. Gao, A. L. Barabasi, J. P. Bagrow, and D. Wang, "Quantifying information flow during emergencies," *Scientific Reports*, vol. 4, no. 1, 2014.

[62] J. Candia, M. C. Gonzalez, P. T. Wang, T. Schoenharl, G. Madey, and A. L. Barabasi, "Uncovering individual and collective human dynamics from mobile phone records," *Journal of Physics A: Mathematical and Theoretical*, vol. 41, no. 22, 2008.

[63] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, "A tale of one city: using cellular network data for urban planning," *IEEE Pervasive Computing*, vol. 10, no. 4, pp. 18–26, 2011.

[64] B. Moumni, V. Frias-Martinez, and E. Frias-Martinez, "Characterizing social response to urban earthquakes using cell-phone network data: the 2012 oaxaca earthquake," in *Proc. 2013 ACM Conf. Pervasive and Ubiquitous Computing (UbiComp 13)*, pp. 1199–1208, 2013.

[65] D. Pastor-Escuredo, A. Morales-Guzman, Y. Torres-Fernandez, J. M. Bauer, A. Wadhwa, C. Castro-Correa, and N. Oliver, "Flooding through the lens of mobile phone activity," *IEEE Global Humanitarian Technology Conf. (GHTC 2014)*, vol. 2014, pp. 279–286, 2014.

[66] A. Dobra, N. E. Williams, and N. Eagle, "Spatiotemporal detection of unusual human population behavior using mobile phone data," *PloS One*, vol. 10, no. 3, 2015.

[67] R. Durrett, *Random Graph Dynamics, Chapter: Branching processes*. Cambridge University Press, 2007.

[68] G. Grimmett and D. Stirzaker, *Probability and Random Processes, Chapter: Branching processes*. Oxford University Press, 2001.

[69] S. Ross, *Stochastic Processes, Chapter: Markov chains*. Wiley, 1996.

[70] D. Watts, "A twenty-first century science," *Nature*, vol. 445, no. 489, 2007.

[71] R. Bond, C. Fariss, J. Jones, A. Kramer, C. Marlow, J. Settle, and J. Fowler, "A 61-million-person experiment in social influence and political mobilization," *Nature*, vol. 489, pp. 295–298, 2012.

[72] M. J. Currie, M. McNiven, T. Yee, U. Schiemer, and F. J. Bowden, "Pooling of clinical specimens prior to testing for chlamydia trachomatis by PCR is accurate and cost saving," *Journal of Clinical Microbiology*, vol. 42, no. 10, pp. 4866–4867, 2004.

[73] S. M. Taylor, J. J. Juliano, P. A. Trottman, J. B. Griffin, S. H. Landis, P. Kitsa, A. K. Tshefu, and S. R. Meshnick, "High-throughput pooling and real-time PCR-based strategy for malaria detection," *Journal of Clinical Microbiology*, vol. 48, no. 2, pp. 512–519, 2010.

[74] FDA, "Guidance for industry - use of nucleic acid tests on pooled and individual samples from donors of whole blood and blood components, including source plasma, to reduce the risk of transmission of hepatitis B virus." `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-nucleic-acid-tests-pooled-and-individual-samples-donors-whole-blood-and-blood-components`, October 2012.

[75] A. Paxton, "HIV, HCV windows nearly shut with minipool NAT," *CAP Today*, October 2004.

[76] C. R. Bilder, J. M. Tebbs, and P. Chen, "Informative retesting," *Journal of the American Statistical Association*, vol. 105, no. 491, pp. 942–955, 2010.

[77] J. Emmanuel, M. Bassett, H. Smith, and J. Jacobs, "Pooling of sera for human immunodeficiency virus (HIV) testing: an economical method for use in developing countries," *Journal of Clinical Pathology*, vol. 41, no. 5, pp. 582–585, 1988.

[78] S. Vansteelandt, E. Goetghebeur, and T. Verstraeten, "Regression models for disease prevalence with diagnostic tests on pools of serum samples," *Biometrics*, vol. 56, no. 4, pp. 1126–1133, 2000.

[79] I. Yelin, N. Aharony, E. S. Tamar, A. Argoetti, E. Messer, D. Berenbaum, E. Shafran, A. Kuzli, N. Gandali, O. Shkedi, T. Hashimshony, Y. Mandel-Gutfreund, M. Halberthal, Y. Geffen, M. Szwarcwort-Cohen, and R. Kishony, "Evaluation of COVID-19 RT-qPCR test in multi sample pools," *Clinical Infectious Diseases*, May 2020.

[80] FDA, "Pooled sample testing and screening testing for COVID-19." `https://www.fda.gov/medical-devices/coronavirus-covid-19-and-medical-devices/pooled-sample-testing-and-screening-testing-covid-19`, August 24, 2020.

[81] WSJ, "Wuhan tests nine million people for coronavirus in 10 days." `https://www.wsj.com/articles/wuhan-tests-nine-million-people-for-coronavirus-in-10-days-11590408910`, May 25, 2020.

[82] National Institutes of Health, "When To Test, FAQ on pooled testing." https://whentotest.org/frequently-asked-questions#how-can-i-tell-if-pooled-testing-is-a-good-fit-for-my-organization, 2021.

[83] P. Bertolotti, "Group testing dashboard." https://group-testing.herokuapp.com/ via https://whentotest.org/frequently-asked-questions#how-can-i-tell-if-pooled-testing-is-a-good-fit-for-my-organization, 2021.

[84] The Wall Street Journal, "COVID-19 testing is hampered by shortages of critical ingredient." https://www.wsj.com/articles/covid-19-testing-is-hampered-by-shortages-of-critical-ingredient-11600772400, September 22, 2020.

[85] WSJ, "COVID-19 tests are still hard to get in many communities." https://www.wsj.com/articles/covid-19-testing-challenges-remain-for-many-urban-rural-communities-11611320400, January 22, 2021.

[86] M. S. Black, C. R. Bilder, and J. M. Tebbs, "Group testing in heterogeneous populations by using halving algorithms," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 61, no. 2, pp. 277–290, 2012.

[87] N. J. A. Harvey, M. Patrascu, Y. Wen, S. Yekhanin, and V. W. S. Chan, "Non-adaptive fault diagnosis for all-optical networks via combinatorial group testing on graphs," in *IEEE INFOCOM 2007 - 26th IEEE International Conference on Computer Communications*, pp. 697–705, May 2007.

[88] M. Cheraghchi, A. Karbasi, S. Mohajer, and V. Saligrama, "Graph-constrained group testing," *IEEE Transactions on Information Theory*, vol. 58, no. 1, pp. 248–262, 2012.

[89] A. Sterrett, "On the detection of defective members of large populations," *The Annals of Mathematical Statistics*, vol. 28, no. 4, pp. 1033–1036, 1957.

[90] E. Litvak, X. M. Tu, and M. Pagano, "Screening for the presence of a disease by pooling sera samples," *Journal of the American Statistical Association*, vol. 89, no. 426, pp. 424–434, 1994.

[91] M. Mézard and C. Toninelli, "Group testing with random pools: Optimal two-stage algorithms," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1736–1745, 2011.

[92] S. Ghosh, A. Rajwade, S. Krishna, N. Gopalkrishnan, T. E. Schaus, A. Chakravarthy, S. Varahan, V. Appu, R. Ramakrishnan, S. Ch, *et al.*, "Tapestry: a single-round smart pooling technique for COVID-19 testing," *medRxiv*, May 2020.

[93] J. Zhu, K. Rivera, and D. Baron, "Noisy pooled PCR for virus testing," *arXiv*, vol. 2004.02689, April 2020.

[94] A. Cohen, N. Shlezinger, A. Solomon, Y. C. Eldar, and M. Médard, "Multi-level group testing with application to one-shot pooled COVID-19 tests," *arXiv*, vol. 2010.06072, October 2020.

[95] J. Yi, R. Mudumbai, and W. Xu, "Low-cost and high-throughput testing of COVID-19 viruses and antibodies via compressed sensing: system concepts and computational experiments," *arXiv*, vol. 2004.05759, April 2020.

[96] Johns Hopkins University, "Daily testing trends in the US." `https://coronavirus.jhu.edu/testing/individual-states`, Accessed September 21, 2021.

[97] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[98] SafeGraph, "SafeGraph Data for Academics." `https://www.safegraph.com/academics`, Retrieved August 18, 2020.

[99] New York Times, "NYTimes COVID-19 Data." `https://github.com/nytimes/covid-19-data`, Retrieved May 18, 2020.

[100] World Health Organization, "WHO Director-General's opening remarks at the media briefing on COVID-19." `https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020`, March 11, 2020.

[101] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[102] M. E. J. Newman, "Spectral methods for community detection and graph partitioning," *Physical Review E*, vol. 88, p. 042822, Oct 2013.

[103] M. Aldridge, O. Johnson, and J. Scarlett, "Group testing: an information theory perspective," *arXiv*, vol. 1902.06002, 2019.

[104] FDA, "SARS-CoV-2 reference panel comparative data." `https://www.fda.gov/medical-devices/coronavirus-covid-19-and-medical-devices/sars-cov-2-reference-panel-comparative-data`, December 7, 2020.

[105] J. Watson, P. F. Whiting, and J. E. Brush, "Interpreting a COVID-19 test result," *The BMJ (British Medical Journal)*, vol. 369, 2020.

[106] R. Davidson and J. MacKinnon, *Estimation and Inference in Econometrics, Chapter: The classical hypothesis tests*. Oxford University Press, 1993.

[107] R. V. Hogg, E. A. Tanis, and D. L. Zimmerman, *Probability and Statistical Inference, Chapter: The central limit theorem.* Pearson, 2014.

[108] P. J. Bickel and K. A. Doksum, *Mathematical Statistics, Chapter: Further limit theorems and inequalities*, vol. 1. Chapman and Hall, 2015.

[109] United Nations Population Division, "Total population – both sexes." https://population.un.org/wpp/Download/Standard/Population/, Retrieved Jan. 2019 2019.

[110] Statista, "Mobile subscription penetration Yemen." https://statista.com/statistics/510648/mobile-cellular-subscriptions-per-100-inhabitants-in-yemen/, Retrieved Jan. 2019.

[111] World Bank, "World development indicators." https://datacatalog.worldbank.org/dataset/world-development-indicators, Retrieved Jan. 2019.

[112] BBC, "Yemen: President Saleh 'was injured by palace bomb'." https://bbc.com/news/world-middle-east-13892502, June 2011.

[113] B. Efron and T. Hastie, *Computer Age Statistical Inference, Chapter: Large-scale hypothesis testing and FDRs.* Cambridge University Press, 2016.

[114] The Guardian, "Yemen police kill protesters in crackdown on dissent." https://theguardian.com/world/2011/mar/12/yemen-police-kill-protesters-crackdown, March 2011.

[115] BBC, "Yemen unrest, dozens killed as gunmen target rally." https://bbc.co.uk/news/mobile/world-middle-east-12783585, March 2011.

[116] Star Advertiser, "Rival tanks deploy in streets of Yemen's capital." https://staradvertiser.com/2011/03/21/breaking-news/rival-tanks-deploy-in-streets-of-yemens-capital/, March 2011.

[117] Al Jazeera, "Yemen leader blames protests on US." https://aljazeera.com/news/middleeast/2011/03/20113191141211328.html, March 2011.

[118] Al Jazeera, "Anti-Saleh protests sweep Yemen." https://aljazeera.com/news/middleeast/2011/03/20113214474211863.html, March 2011.

[119] Al Jazeera, "New protests erupt in Yemen." https://aljazeera.com/news/middleeast/2011/01/2011129112626339573.html, January 2011.

[120] "Yemen's Saleh 'makes new offer to protesters'." https://aljazeera.com/news/middleeast/2011/03/201133014584368624.html, March 2011.

[121] Bureau of Investigative Journalism, "Drone warfare." https://thebureauinvestigates.com/projects/drone-war, Retrieved Jan. 2019.

[122] The White House, Office of the Press Secretary, "Press briefing by Press Secretary Jay Carney." https://obamawhitehouse.archives.gov/the-press-office/2012/01/31/press-briefing-press-secretary-jay-carney-13112, January 2012.

[123] The White House, Office of the Press Secretary, "Remarks of John O. Brennan on the ethics and efficacy of the President's counterterrorism strategy." https://lawfareblog.com/text-john-brennans-speech-drone-strikes-today-wilson-center, April 2012.

[124] U. S. Department of Defense, "Remarks by general townsend in a media availability in baghdad, iraq." https://defense.gov/Newsroom/Transcripts/Transcript/Article/1244058/remarks-by-general-townsend-in-a-media-availability-in-baghdad-iraq/, July 2017.

[125] P. Bergen and K. Tiedemann, "Washington's phantom war: the effects of the u.s. drone program in pakistan," *Foreign Affairs*, 2011.

[126] D. Byman, "Why drones work: the case for washington's weapon of choice," *Foreign Affairs*, 2013.

[127] K. Cronin, "Why drones fail: when tactics drive strategy," *Foreign Affairs*, 2013.

[128] J. Foust, "Unaccountable killing machines: the true cost of u.s. drones," *The Atlantic*, 2011.

[129] L. Hudson, C. S. Owens, and M. Flannes, "Drone warfare: blowback from the new American way of war," *Middle East Policy*, vol. 18, no. 3, pp. 122–132, 2011.

[130] M. J. Boyle, "Do counterterrorism and counterinsurgency go together?," *International Affairs*, vol. 86, no. 2, pp. 333–353, 2010.

[131] D. Kilcullen and A. M. Exummay, "Death from above, outrage down below," *The New York Times*, 2009.

[132] Human Rights Watch, "Between a drone and al-Qaeda: the civilian cost of U.S. targeted killings in Yemen." https://hrw.org/sites/default/files/reports/yemen1013_ForUpload_1.pdf, 2013.

[133] P. B. Johnston and A. K. Sarbah, "The impact of u.s. drone strikes on terrorism in pakistan," *International Studies Quarterly*, vol. 60, no. 2, pp. 203–219, 2016.

[134] L. Lewis, "Drone strikes in Pakistan: reasons to assess civilian casualties," *Defense Technical Information Center*, 2014.

[135] M. Smith and J. I. Walsh, "Do drone strikes degrade al-qaeda?," *Terrorism and Political Violence*, vol. 25, no. 2, pp. 311–327, 2013.

[136] New America, "Drone strikes: Yemen." https://newamerica.org/in-depth/americas-counterterrorism-wars/us-targeted-killing-program-yemen/, Retrieved Jan. 2019 2019.

[137] F. Christia, L. Yao, S. Wittels, and J. Leskovec, "Yemen calling: seven things cell data reveal about life in the republic," *Foreign Affairs*, 2015.

[138] X. Lu, L. Bengtsson, and P. Holme, "Predictability of population displacement after the 2010 haiti earthquake," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 109, no. 29, pp. 11576–11581, 2012.

[139] J. E. Blumenstock, N. Eagle, and M. Fafchamps, "Airtime transfers and mobile communications: Evidence in the aftermath of natural disasters," *Journal of Development Economics*, vol. 120, pp. 157–181, 2016.

[140] G. Palla, A. L. Barabasi, and T. Vicsek, "Quantifying social group evolution," *Nature*, vol. 446, pp. 664–667, 2007.

[141] M. C. Gonzalez, C. A. Hidalgo, and A. L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779–782, 2008.

[142] J. L. Toole, C. Herrera-Yaque, C. M. Schneider, and M. C. Gonzalez, "Coupling human mobility and social ties," *Journal of the Royal Society Interface*, vol. 12, no. 105, 2015.

[143] A. Shah, "Do u.s. drone strikes cause blowback?: evidence from pakistan and beyond," *International Security*, vol. 42, no. 4, pp. 47–84, 2018.

[144] C. C. Fair, K. Kaltenthaler, and W. J. Miller, "Pakistani opposition to American drone strikes," *Political Science Quarterly*, vol. 131, pp. 387–419, 2016.

[145] E. Berman, J. Felter, and J. Shapiro, *Small wars, big data*. Princeton University Press, 2018.

[146] L. N. Condra and J. Shapiro, "Who takes the blame? the strategic effects of collateral damage," *American Journal of Political Science*, vol. 56, no. 1, pp. 167–187, 2012.

[147] C. Kolenda, R. Reid, C. Rogers, and M. Retzius, "The strategic costs of civilian harm: applying lessons from Afghanistan to present and future conflicts," *Open Society Foundations*, 2016.

[148] S. Schutte, "Violence and civilian loyalties: evidence from afghanistan," *Journal of Conflict Resolution*, vol. 61, no. 8, 2017.

[149] F. Christia, S. I. Zoumpoulis, M. Freedman, L. Yao, and A. Jadbabaie, "The effect of drone strikes on civilian communication: evidence from yemen," *Political Science Research and Methods*, pp. 1–9, 2021.

[150] UN OCHA Yemen, Humanitarian Data Exchange, "Cso 2017 population projections." https://data.humdata.org/dataset/yemen-cso-2017-population-projections-by-governorate-district-sex-age-disaggregated, Retrieved Jan. 2019.

[151] K. Schmidheiny and S. Siegloch, "On event study designs and distributed-lag models," *IZA Discussion Paper*, 2019.

[152] A. Abadie, S. Athey, G. Imbens, and J. Wooldridge, "When should you adjust standard errors for clustering?," *NBER Working Paper*, 2017.

[153] T. Pettersson, S. Hogbladh, and M. Oberg, "Organized violence, 1989-2018 and peace agreements," *Journal of Peace Research*, vol. 56, 2019.

[154] Uppsala University, Department of Peace and Conflict Research, "Uppsala conflict data program." https://ucdp.uu.se, Retrieved June 2019.

[155] W. H. Greene, *Econometric analysis*. Pearson Education, 2003.

[156] J. Y. Campbell, A. W. Lo, and A. C. MacKinlay, *The econometrics of financial markets*. princeton University press, 2012.

[157] Central Statistical Organization, Republic of Yemen, "Yemen: governorates, major cities, and villages." https://citypopulation.de/Yemen.html, Retrieved Jan. 2019.

[158] BBC, "Yemeni arms factory blast toll rises amid protests." https://bbc.com/news/world-middle-east-12902310, Retrieved June 2019.

[159] BBC, "Yemen: President saleh 'was injured by palace bomb'." https://bbc.com/news/world-middle-east-13892502, Retrieved June 2019.

[160] BBC, "Yemen bomb attack 'kills at least 26 people' in Mukalla." https://bbc.com/news/world-17164558, Retrieved June 2019.

[161] BBC, "Al-Qaeda attack on Yemen army parade causes carnage." https://bbc.com/news/world-middle-east-18142695, Retrieved June 2019.

[162] Reuters, "Al-Qaeda suicide bomber attacks Yemen police academy." https://reuters.com/article/us-yemen-explosion/al-qaeda-suicide-bomber-attacks-yemen-police-academy-idUSBRE86A0H120120711, Retrieved June 2019.