NETWORK GROUP TESTING

Paolo Bertolotti and Ali Jadbabaie
Institute for Data, Systems, and Society
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{pbertolo,jadbabai}@mit.edu

ABSTRACT

We consider the problem of identifying infected individuals in a population of size N. We 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 information to improve performance. Network grouping, which groups individuals by community, weakly dominates Dorfman testing in terms of the expected number of tests used. 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. As an empirical example, we consider the scenario of a university testing its population for COVID-19. Using social network data from a Danish university, we demonstrate network grouping requires significantly fewer tests than Dorfman. 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, social group, or work group.

1 INTRODUCTION

Group testing improves testing capabilities for infectious diseases when resources are limited. In normal scenarios, infected individuals from a population of size N are identified by testing all population members individually, which uses N tests. In the simplest form of group 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 Dorfman (1943) 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 (Van et al., 2012; Currie et al., 2004; Taylor et al., 2010). Within the US, group testing is used in blood banks and infertility prevention programs where large numbers of individuals are routinely tested (FDA, 2012; Bilder et al., 2010). 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 (Emmanuel et al., 1988). 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 introduce a generative network model and epidemic model. We derive the number of tests used under network grouping and prove the expected number of tests is upper bounded by Dorfman testing, which implies network grouping weakly dominates Dorfman. The outperformance of network grouping is determined by the strength of community structure in the network. In networks with strong community structure, network grouping performs optimally and achieves the lower bound for two-stage testing procedures. In networks with no structure, network grouping is equivalent to Dorfman testing.

We end by considering the scenario of a university testing its population for COVID-19 cases. Using social network data from a Danish university, we demonstrate network grouping outperforms Dorfman 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 (Hogan et al., 2020; Yelin et al., 2020) and several countries have implemented group testing to increase their testing capabilities (FDA, 2020b; WSJ, 2020). As testing resources still remain constrained (WSJ, 2021), we hope more institutions and governments will take advantage of the power of group testing.

Since Dorfman's work in 1943, numerous group testing approaches with strong performance have been introduced (Litvak et al., 1994; Cheraghchi et al., 2012; Ghosh et al., 2020). However, the complexity of the proposed methods have limited their adoption in the medical field. As a result, Dorfman testing remains the most common approach to group testing in practice (McMahan et al., 2012; FDA, 2020a). Importantly, our proposed approach is simple for practitioners to implement. In practice, individuals can be grouped by family unit, social group, work group, or other community structure.

2 SETUP

In this section, we describe Dorfman two-stage testing and the lower bound for two-stage group testing procedures. 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, one individual is infected with probability one and the remaining $N - 1$ individuals are infected independently with probability v.[1] 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

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

where $v' = 1 - (1 - v)^n$. The derivation of $\mathrm{E}[T_D]$ was provided by Dorfman and can be found in appendix A.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%). 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 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

$$T_{LB} = \frac{N}{n} + n \cdot \max \left(1, \frac{1 + (N - 1)v}{n} \right) \tag{2}$$

---

[1] In Dorfman's original paper, all N individuals are infected independently with probability v. We deviate slightly from his original setup to ensure at least one individual is infected.

The derivation can be found in appendix A.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.

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

4 RESULTS

In this section, we introduce our main results regarding network grouping and its performance compared to Dorfman testing. Under network grouping, we group individuals by 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

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n \left[1 + \left(\frac{m}{n} - 1 \right)^+ p' + \left(\frac{N}{n} - 1 - \left(\frac{m}{n} - 1 \right)^+ \right) q' \right] \tag{3}$$

where $p' = 1 - (1 - p\alpha)^n$, $q' = 1 - (1 - q\alpha)^n$, and $(x)^+ = \max(x, 0)$. The derivation of the full distribution of the number of tests is provided in appendix A.3.

With the number of tests 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*

$$T_{LB} \leq \mathrm{E}[T_{NG}] \leq \mathrm{E}[T_D] \tag{4}$$

If $q = 0$ and $n \geq m$, then $\mathrm{E}[T_{NG}] = T_{LB}$.
If $q = p$, then $\mathrm{E}[T_{NG}] = \mathrm{E}[T_D]$.
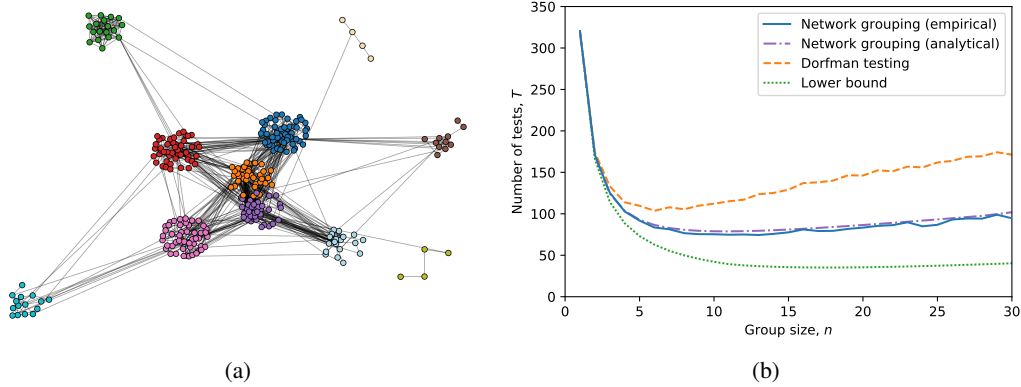
<center>(a)</center>
<center>(b)</center>

Figure 1: **(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. 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. Averages are computed after simulating 1000 epidemic processes per group size.

The proof of theorem 1 can be found in appendix A.4. Theorem 1 states network grouping weakly dominates Dorfman testing in terms of the expected number of tests used. The outperformance of network grouping is driven by q, the probability an edge exists between nodes in different communities. In settings where networks have extremely strong community structure, network grouping performs optimally and achieves the lower bound. Specifically, 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, notability when $q = p$. Interestingly, even though we assume a network model, epidemic model, and network grouping, we end up back where we started with Dorfman testing. The reasoning is simple: since the network has no structure, all nodes have the same probability of being infected and the network provides no useful information for grouping.

In normal cases when $0 < q < p$, network grouping significantly outperforms Dorfman testing. Consider the scenario of a university testing its population for COVID-19 cases. Using data from the Copenhagen Networks Study of Sapiezynski et al. (2019), we build the social network of first-year students at the Technical University of Denmark. The network contains 310 nodes, which correspond to students, and 1503 edges, which correspond to their physical interactions recorded using bluetooth-enabled smartphones. We apply the Louvain algorithm of Blondel et al. (2008) to detect communities, resulting in 11 communities with an average size of 28 individuals. The social network with nodes colored by their community is displayed in figure 1a. We estimate p to be 0.18 and q to be 0.01, indicating a sparse network with strong community structure.

We simulate 1000 epidemic processes on the network using the model outlined in section 3 with $\alpha = 0.95$, resulting in an estimated infection prevalence of 0.03 (3%). We apply network grouping, which groups individuals by community, and Dorfman testing, which groups individuals randomly. Figure 1b displays the average number of tests used under the different approaches 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 1b demonstrates our analytical result for the number of tests used under network grouping, provided in equation 3, is a strong approximation for the number of tests used in a real network setting.

In this work, we have introduced the idea of using social network information to improve group testing. When networks have strong community structure, network grouping outperforms Dorfman testing in terms of the number of tests used. It turns out network grouping also outperforms Dorfman

<center>4</center>

in terms of false positives and false negatives when tests are imperfect. However, we leave discussion of imperfect tests to future write-ups. Importantly, network grouping is simple for practitioners to implement; individuals can be grouped by family unit, friend group, or other community structure.

REFERENCES

Christopher R. Bilder, Joshua M. Tebbs, and Peng Chen. Informative retesting. *Journal of the American Statistical Association*, 105(491):942–955, 2010. ISSN 01621459. URL http://www.jstor.org/stable/27920118.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. doi: 10.1088/1742-5468/2008/10/p10008. URL https://doi.org/10.1088/1742-5468/2008/10/p10008.

Mahdi Cheraghchi, Amin Karbasi, Soheil Mohajer, and Venkatesh Saligrama. Graph-constrained group testing. *IEEE Transactions on Information Theory*, 58(1):248–262, 2012. URL https://ieeexplore.ieee.org/document/6121984.

Marian J. Currie, Michelle McNiven, Tracey Yee, Ursula Schiemer, and Francis J. Bowden. Pooling of clinical specimens prior to testing for chlamydia trachomatis by PCR is accurate and cost saving. *Journal of Clinical Microbiology*, 42(10):4866–4867, 2004. ISSN 0095-1137. doi: 10.1128/JCM.42.10.4866-4867.2004. URL https://europepmc.org/articles/PMC522303.

Robert Dorfman. The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14(4):436–440, 1943. ISSN 00034851. URL http://www.jstor.org/stable/2235930.

JC Emmanuel, MT Bassett, HJ Smith, and JA Jacobs. Pooling of sera for human immunodeficiency virus (HIV) testing: an economical method for use in developing countries. *Journal of Clinical Pathology*, 41(5):582–585, 1988. ISSN 0021-9746. doi: 10.1136/jcp.41.5.582. URL https://europepmc.org/articles/PMC1141517.

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.

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, 2020a.

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, 2020b.

Sabyasachi Ghosh, Ajit Rajwade, Srikar Krishna, Nikhil Gopalkrishnan, Thomas E Schaus, Anirudh Chakravarthy, Sriram Varahan, Vidhya Appu, Raunak Ramakrishnan, Shashank Ch, et al. Tapestry: a single-round smart pooling technique for COVID-19 testing. *medRxiv*, May 2020. URL https://www.medrxiv.org/content/10.1101/2020.04.23.20077727v2.

Catherine A. Hogan, Malaya K. Sahoo, and Benjamin A. Pinsky. Sample pooling as a strategy to detect community transmission of SARS-CoV-2. *Journal of the American Medical Association*, April 2020. ISSN 0098-7484. doi: 10.1001/jama.2020.5445. URL https://doi.org/10.1001/jama.2020.5445.

Eugene Litvak, Xin M. Tu, and Marcello Pagano. Screening for the presence of a disease by pooling sera samples. *Journal of the American Statistical Association*, 89(426):424–434, 1994. ISSN 01621459. URL http://www.jstor.org/stable/2290842.

Christopher S. McMahan, Joshua M. Tebbs, and Christopher R. Bilder. Informative Dorfman screening. *Biometrics*, 68(1):287–296, 2012. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/41434069.

Piotr Sapiezynski, Arkadiusz Stopczynski, David Dreyer Lassen, and Sune Lehmann. Interaction data from the Copenhagen Networks Study. *Scientific Data*, 6(1):315, 2019. doi: 10.1038/s41597-019-0325-x. URL https://doi.org/10.1038/s41597-019-0325-x.

Steve M. Taylor, Jonathan J. Juliano, Paul A. Trottman, Jennifer B. Griffin, Sarah H. Landis, Paluku Kitsa, Antoinette K. Tshefu, and Steven R. Meshnick. High-throughput pooling and real-time PCR-based strategy for malaria detection. *Journal of Clinical Microbiology*, 48(2):512–519, 2010. ISSN 0095-1137. doi: 10.1128/JCM.01800-09. URL https://jcm.asm.org/content/48/2/512.

Tam T. Van, Joseph Miller, David M. Warshauer, Erik Reisdorf, Daniel Jernigan, Rosemary Humes, and Peter A. Shult. Pooling nasopharyngeal/throat swab specimens to increase testing capacity for influenza viruses by PCR. *Journal of Clinical Microbiology*, 50(3):891–896, 2012. ISSN 0095-1137. doi: 10.1128/JCM.05631-11. URL https://europepmc.org/articles/PMC3295167.

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.

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.

Idan Yelin, Noga Aharony, Einat Shaer Tamar, Amir Argoetti, Esther Messer, Dina Berenbaum, Einat Shafran, Areen Kuzli, Nagham Gandali, Omer Shkedi, Tamar Hashimshony, Yael Mandel-Gutfreund, Michael Halberthal, Yuval Geffen, Moran Szwarcwort-Cohen, and Roy Kishony. Evaluation of COVID-19 RT-qPCR test in multi sample pools. *Clinical Infectious Diseases*, May 2020. ISSN 1058-4838. doi: 10.1093/cid/ciaa531. URL https://doi.org/10.1093/cid/ciaa531.

A   APPENDIX

A.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 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. 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

$$T_D \sim \frac{N}{n} + n \left[1 + Bin \left(\frac{N}{n} - 1, v' \right) \right] \tag{5}$$

Taking the expectation of T_D provides $\mathrm{E}[T_D]$ as displayed in equation 1.

A.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 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. 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 greater than or equal to one but 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

$$T_{LB} = \frac{N}{n} + n \cdot \max \left(1, \frac{1 + (N - 1)v}{n} \right) \tag{6}$$

A.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 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.

The network contains N/m communities of size m. We consider cases where n divisible by m or m divisible by n. First consider the case where $m \leq n$. Since we group individuals by community (as described at the beginning of section 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

$$T_{NG} \sim \frac{N}{n} + n \left[1 + Bin\left(\frac{N}{n} - 1, q'\right) \right] \tag{7}$$

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

$$T_{NG} \sim \frac{N}{n} + n \left[1 + Bin\left(\frac{m}{n} - 1, p'\right) + Bin\left(\frac{N}{n} - \frac{m}{n}, q'\right) \right] \tag{8}$$

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

$$T_{NG} \sim \frac{N}{n} + n \left[1 + Bin\left(\left(\frac{m}{n} - 1\right)^+, p'\right) + Bin\left(\frac{N}{n} - 1 - \left(\frac{m}{n} - 1\right)^+, q'\right) \right] \tag{9}$$

where $p' = 1 - (1 - p\alpha)^n$ and $q' = 1 - (1 - q\alpha)^n$. Equation 9 holds exactly when N divisible by n and either m divisible by n or n divisible by m. It is a strong approximation otherwise. Taking the expectation of T_{NG} in equation 9 provides $\mathrm{E}[T_{NG}]$ as displayed in equation 3.

A.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, $q = p$. 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

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n \left[1 + \left(\frac{m}{n} - 1\right)p' + \left(\frac{N}{n} - \frac{m}{n}\right)q' \right] \tag{10}$$

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 10 with respect to q and simplifying yields

$$\frac{\partial \mathrm{E}[T_{NG}]}{\partial q} = \frac{nv(N - 1)(N - m)\left[p(m - 1)(1 - q\alpha)^{n-1} - p(m - n)(1 - p\alpha)^{n-1}\right]}{[(m - 1)p + (N - m)q]^2} \tag{11}$$

Demonstrating equation 11 is nonnegative proves equation 10 is weakly 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 11 is nonnegative.

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

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n \left[1 + \left(\frac{N}{n} - 1\right)q' \right] \tag{12}$$

where $q' = 1 - (1 - q\alpha)^n$. Again, $\alpha = (N-1)v/[(m-1)p + (N-m)q]$. Taking the derivative of equation 12 with respect to q and simplifying yields

$$\frac{\partial \mathrm{E}[T_{NG}]}{\partial q} = \frac{nvp(N-1)(N-n)(m-1)(1-q\alpha)^n}{[(m-1)p + (N-m)q][(m-1)p + (N-m)q + q(1-N)v]} \tag{13}$$

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

$$(m-1)p + (N-m)q + q(1-N)v \geq 0 \tag{14}$$

Rearranging equation 14 yields

$$\frac{q(N-1)v}{(m-1)p + (N-m)q} \leq 1 \tag{15}$$

$$q\alpha \leq 1 \tag{16}$$

which is true by assumption 1. As a result, equation 13 is nonnegative and equation 12 is weakly 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

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n\left[1 + \left(\frac{m}{n} - 1\right)^+ p' + \left(\frac{N}{n} - 1 - \left(\frac{m}{n} - 1\right)^+\right)q'\right] \tag{17}$$

$$= \frac{N}{n} + n\left[1 + \left(\frac{N}{n} - 1\right)v'\right] \tag{18}$$

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 2 simplifies to $N/n + n$. Therefore,

$$\mathrm{E}[T_{NG}] - T_{LB} = n\left(\frac{m}{n} - 1\right)^+ p' + n\left(\frac{N}{n} - 1 - \left(\frac{m}{n} - 1\right)^+\right)q' \tag{19}$$

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 19 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

$$T_{LB} = \frac{N}{n} + 1 + (N-1)v \tag{20}$$

$$= \frac{N}{n} + 1 + (m-1)p\alpha + (N-m)q\alpha \tag{21}$$

$$= \frac{N}{n} + n + (m-1)(1 - (1 - p\alpha)) + (N-m)(1 - (1 - q\alpha)) - (n-1) \tag{22}$$

where the second equality makes use of the $E[I_D] = E[I_{NG}]$ equivalence specified in assumption 1. Equation 22 is a slight rearrangement of equation 21. As $n < m$, $\mathrm{E}[T_{NG}]$ becomes

$$\mathrm{E}[T_{NG}] = \frac{N}{n} + n + (m-n)p' + (N-m)q' \tag{23}$$

$$= \frac{N}{n} + n + (m-1)p' + (N-m)q' - (n-1)p' \tag{24}$$

Subtracting equation 22 from equation 24 yields

$$
\begin{aligned}
\mathrm{E}[T_{NG}] - T_{LB} = \\
(m-1)[p' - (1-(1-p\alpha))] + (N-m)[q' - (1-(1-q\alpha))] + (n-1)(1-p') \quad (25)
\end{aligned}
$$

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 21 and the inequality $m \geq 1 + (m-1)p\alpha$, we have the following inequality for the lower bound T_{LB}.

$$
T_{LB} = \frac{N}{n} + 1 + (m-1)p\alpha + (N-m)q\alpha \quad (26)
$$

$$
\leq \frac{N}{n} + m + (N-m)q\alpha \quad (27)
$$

$$
= \frac{N}{n} + n + (N-m)(1-(1-q\alpha)) - (n-m) \quad (28)
$$

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

$$
\mathrm{E}[T_{NG}] = \frac{N}{n} + n + (N-n)\,q' \quad (29)
$$

$$
= \frac{N}{n} + n + (N-m)\,q' - (n-m)q' \quad (30)
$$

Subtracting equation 28 from equation 30 yields

$$
\mathrm{E}[T_{NG}] - T_{LB} \geq (N-m)[q' - (1-(1-q\alpha))] + (n-m)(1-q') \quad (31)
$$

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. $\qquad\square$

Final statements The statement "If $q = 0$ and $n \geq m$, then $\mathrm{E}[T_{NG}] = T_{LB}$" 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

$$
\mathrm{E}[T_{NG}] = \frac{N}{n} + n \quad (32)
$$

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

$$
T_{LB} = \frac{N}{n} + n \quad (33)
$$

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

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.

The statement "If $q = p$, then $\mathrm{E}[T_{NG}] = \mathrm{E}[T_D]$" is proved above during the upper bound portion of the proof.