Sampling manholes to home in on SARS-CoV-2 infections

Richard C. Larson[1¶], Oded Berman[2¶], Mehdi Nourinejad[3¶*]

[1] Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[2] Rotman School of Management, University of Toronto, Toronto, Canada

[3] Department of Civil Engineering, York University, Toronto, Canada

* Corresponding author

Email: mehdi.nourinejad@lassonde.yorku.ca

¶ All three authors contributed equally to this work.

ABSTRACT

Many individuals infected with the novel Coronavirus (SARS-CoV-2) suffer from intestinal infection as well as respiratory infection. These COVID-19-suffering individuals shed virus in their stool, resulting in municipal sewage systems carrying the virus and its genetic remnants. These viral traces can be detected in the sewage entering a wastewater treatment plant (WTP), often resulting in accurate estimates of the extent of infections over a community. In this paper, we develop algorithmic procedures that home in on locations and/or neighborhoods within the community that are most likely to have infections. Our novel data source is wastewater sampled and real-time tested from selected manholes. Our algorithms dynamically and adaptively point to a sequence of manholes to sample and test. The algorithms are often finished after 5 to 10 manhole samples, meaning that - in the field - the procedure can be carried out within one day. The goal is to provide timely information that will support faster more productive human testing for viral infection and thus reduce community spread of the disease. Leveraging the tree graph structure of the sewage system, we develop two algorithms, the first designed for a community that is certified at a given time to have zero infections and the second for a community known to have many infections. For the first, we assume that wastewater at the WTP has just revealed traces of SARS-CoV-2, indicating existence of a "Patient Zero" in the community. Our first algorithm identifies the city block in which the infected person resides. For the second, we home in on a most infected neighborhood of the community, where a neighborhood is usually several city blocks. We present computational results, some applied to the sewer system map of a New England town. The next step is to test our algorithmic procedures in the field, and to make appropriate adjustments.

1. Introduction

From its origins in Wuhan, China in December 2019, the novel Coronavirus (SARS-CoV-2) has spread ferociously around the world, creating a disease in humans - COVID-19 - that has infected millions and killed hundreds of thousands. The virus has many unusual properties that, as we write this, are still be researched and discovered. One distressing property is that 40 to perhaps 80 percent of infected and virus-shedding individuals are asymptomatic [1][2]. Another is a relatively long incubation time from becoming infected to showing physical symptoms, up to 14 days with a median time of about 5 days. [3]

1.1. SARS-CoV-2 in Human Wastewater

We are focusing on a third unusual property: SARS-CoV-2 not only attacks human lungs, but it can also reside and present symptoms in other parts of the body, including the human digestive track and particularly the intestines [4]. As a result, the SARS-CoV-2 virus and/or related genetic remnants (viral RNA - ribonucleic acid) often appear in the fecal matter of COVID-19 patients. These genetic materials, excreted in stool, positively affirm that the patient is infected with COVID-19. Not all COVID-19 patients present such in-stool markers, only about 50% [5] [6].

Since human waste is flushed down the toilet, any included genetic markers and/or chemicals are detectable in raw sewage. Much of the health condition of the human body can be determined by analyses of sewage, leveraging an emerging field called “Wastewater-Based Epidemiology” (WBE). Quoting a recent paper,

> "Wastewater-Based Epidemiology (WBE) is a new epidemiology tool that has potential to act as a complementary approach for current infectious disease surveillance systems and an early warning system for disease outbreaks. WBE postulates that through the analysis of population pooled wastewater, infectious disease and resistance spread, the emergence of new disease outbreak to the community level can be monitored comprehensively and in real-time." [7]

With COVID-19, while only about 50 per cent of infected individuals excrete remnants of the virus in their stool, any COVID-19-infected population of 20 or more will, almost certainly by the laws of probability, contain such remnants. A population of 1,000 will have virus remnants from about 500 contributors. Amazingly, very recent research [8] has asserted that detailed analysis of a community's sewage flowing into its sewage treatment plant (the ending terminal for the community's sewage system) can now theoretically detect the presence of virus remnants emanating from just one infected person in a contributing population of up to 2,000,000. Thus, the potential sensitivity of sewage testing is remarkably high.

Testing for COVID-19 is vital for adjusting public policies related to its control, such as "flattening the curve", "putting out local fires," and ultimately reducing greatly the incidence of the disease. But testing each human in a population is a time-consuming, expensive, arduous and often impossible task. Sewage testing is relatively inexpensive and can track accurately the ups and downs of COVID-19 prevalence in a community. Importantly, a Parisian study [9] suggests that a population newly experiencing a rise in COVID-19 infections will exhibit that rise in excreted virus-containing stool about one week before showing visible symptoms of infection. If the Parisian study is confirmed by others, timely sewage testing may provide a One-

Week (approximately) Early Warning System of population infection. Authorities acting on that information early could act to avert exponential rise in future infections. While testing sewage is not the same as testing individuals, it is fast, inexpensive and informative to public decision makers.

1.2. Wastewater Testing

As we write this, during the past few weeks there has been a virtual explosion of sewage testing "proof-of-concept" projects reported in the literature, many papers available in online pre-prints and still awaiting formal reviews. In addition to the aforementioned Parisian study [9], these include projects in Arizona [8], Montana [10], New Haven Conn. [11], Boston Massachusetts [12], Italy [13], the Netherlands [14] and many more. These reports demonstrate that the ups and downs of virus loads in the sewage track well the ups and downs of known cases in the population, and often with a one-week early warning. The theory is proving to be correct.

Testing sewage for COVID-19 is even a new business opportunity. MIT spinoff Biobot claims to be "the first company in the world to commercialize data from sewage." [15]. As we write this, Biobot is working with about 330 facilities in 40 states, returning analyses of sewage samples sent to them. Currently the service is *pro bono*. The University of Arizona and others are also performing tests for participating communities.

All of these activities are now evolving, with current test turn-around times disappointingly long, ranging from two weeks in one case to one week in another. Biobot claims to be headed for a 24-hour test in its Boston-area headquarters, which means perhaps a 3-day turnaround time total (two days for shipping back and forth and one for testing); that would get us closer to what is needed to have a true Early Warning System. There is another approach, building from tests successfully developed for other diseases in Africa, using only paper as the testing material to create what is being described as an instant "two-dollar test" [16]. The researcher is Dr. Zhugen Yang, a biomedical engineer at Cranfield University's Water Science Institute in the U.K., and he hopes to have a working system within two or three months.

In reviewing all of these current developments, we see that Wastewater-Based Epidemiology provides to us with an up to one week early warning system of an impending rise in cases of COVID-19 from a given community. But, even if these techniques – when refined and further tested – are proven correct, all we have is community-averaged predictions. There is no finer grain resolution relating to the likely locations within the community of the rising virus.

1.3. Framing of Our Contribution

The focus on this paper is to gain new information about locations within the community of new COVID-19 infections. To do this, we need a new data source. And that is derived from sampling sewage flows within the underground sewage pipeline network, prior to the final entrance of sewage into the sewage treatment plant. These samples are obtained from selected manholes within the system. Any sewage system can be modeled as a tree network,

with individual homes and businesses representing the input nodes with their corresponding input flows into the system. Any sewage will flow in one given path from origin towards the single terminal node of the tree graph, the local water treatment plant. Our research aims at finding a good limited sequence of manholes to sample and test in order to home in on the location of greatest new infection.

We provide two algorithms. Algorithm 1 represents the case of a community known to have zero infections at time zero. Eventually, monitoring of the community's sewage at its sewage treatment plant indicates a new presence of COVID-19, likely from a "Patient Zero" within the community. Our algorithm dynamically selects a sequence of manholes to test for COVID-19, each selection dependent on the test results of the previous manholes. Eventually we converge to Patient Zero's neighborhood, often to within 100 feet or so of his/her address. With just a handful of manhole tests, we find the closest manhole to the address of Patient Zero. Algorithm 1 is applicable in a variety of settings. Being academics, our favorite is a large university campus that has been locked down and thus empty for months due to COVID-19. This university is now opening a new semester to students, faculty and staff, all of whom who must be tested to assure they are not 'positive' to COVID-19 upon their return. Later, after everyone returns, a Patient Zero becomes infected off campus, and he/she brings that infection onto the campus. Our algorithm helps to pinpoint the university location of that individual. If the infected person is a student, the likely locations are living quarters such as dormitories, fraternities or sororities. If faculty or staff, the likely location is a building having the office of

the individual. The algorithm also applies to various industrial parks, shopping centers, vacation communities, etc.

Moving now from one to many, Algorithm 2 takes the other extreme. There are assumed to be scores of infected individuals in the community, analogous to the situations of virtually all of the previously cited proof-of-concepts tests. Thus, the aggregate sewage-treatment-plant testing procedures previously described may be used to obtain community-averaged estimates on numbers of infections. Again, our concern is moving beyond the community average, to identify that neighborhood within the community that may be considered to be a "Hot Spot," that is, an area having infection rates significantly above the community average. Such information is again useful to those who seek to do high-productivity testing to identify, treat, and isolate infected individuals. This algorithm still uses data from manhole testing of sewage, sequentially selects the manholes to be tested, and stops once a presumed high-incidence neighborhood has been identified.

1.4. More on Manholes and their Testing

Each home in a community having a central sewer system has an underground pipe connecting the home's wastewater plumbing to the under-street local sewer system. In naming various connections, we use the descriptive phrasings of the Town of Surprise, Arizona [17]. The building's pipe to the property line is "the building sewer," which then connects directly to the "lateral sewer," which then connects to the community's sewer pipe network. That network, which we model as a tree graph, has at least three types of pipes, each is ascending diameter

for increased flows: branch sewer, main sewer and trunk sewer. The branch sewer, usually serving a limited number of buildings in a small geographic area, collects sewage from lateral sewers and conveys it to a main or trunk sewer. A main sewer collects sewage from two or more branch sewers acting as tributaries. The trunk sewer conveys sewage from many tributary main and branch sewers over large areas to the sewage treatment plant.

The sewage, once in the community-owned system, flows in a unique path "downward," leading to the sewage treatment plant. A typical town will have hundreds of miles of under-street sewage pipes, the pipes accessible by manholes that are typically 100 to 500 feet apart. The sewer system map of Surprise, Arizona [18] reveals that a typical sewer-equipped street has about 12 manholes per one-mile street segment, meaning about 440 feet between manholes. In contrast, the more densely populated Town of Lexington, Massachusetts has 4,924 manholes distributed of its 171-mile sewer network, corresponding to about 183 feet between manholes [19]. The U.S. EPA estimates the number of sewer manholes nationwide to be about 12 million, averaging approximately 300 feet apart [20].

One final word before we describe the two algorithms: We have verified that manhole sampling is feasible and reasonably priced. In Lexington, Massachusetts, to test four manholes, gathering up to one liter of fluid at each location, takes at most four hours of a two-person crew, from start to finish. Quoted cost: $350. If either of our algorithms should find a sequence of say eight or fewer manholes to test, such tests could easily be carried out in one day, assuming there is a fast, on-site test for COVID-19 that could be carried out by the crew. The

"two-dollar" paper test being developed by Dr. Zhugen Yang, if successful, would be perfect for this purpose. As one can see, our algorithmic work is currently a bit ahead of the curve in terms of all the ingredients to make location-specific identification feasible.

2. Methods

As discussed above, Algorithm 1 deals with a community with no known cases of COVID-19. Then, one day analysis of the inflow to the water treatment plant reveals the first presence of COVID-19, likely a single Patient Zero or household "Patients Zero." The goal is to act immediately to home in on the address of the infected person(s). Once the alarm is sounded, we go into action by opening a directed sequence of manholes, obtaining sewage fluid samples, testing them, and then extending our manhole testing, in a way that converges close to the address of Patient(s) Zero.

The home of Patient(s) Zero is likely to have two closest manholes, one "upstream" from the residence and the other "downstream." Any test of the upstream manhole, if it exists, would not reveal any COVID-19 from the residence of the infected person(s). But the closest downstream manhole would provide that evidence. We seek to find that closest downstream manhole. If successful, we can then reduce our search for the infected person(s) to residences of only a few houses (i.e., all those houses inputting to the same downstream manhole). In that way, we may be able to stop any spread from Patient Zero to the rest of the community. Our approach also extends in Algorithm 2 to cases where the virus has infected multiple residences

in different neighborhoods. In such a scenario, the objective is the find the Hot Spot neighborhood with the highest number of infected homes.

We use algorithms to dynamically select the sequence of manholes to test. Our task is made easier by the fact that the municipal sewage system pipe network is a directed tree network, comprising nodes and directed links, with inputs from homes and businesses and the output all converging to one single “terminal node" to the network, the wastewater treatment plant.

For ease of exposition in this paper, we are algorithmically more interested in showing how to exploit network tree topology than in searching straight-line multiple manholes along a given link. In examples, using our “stylized networks,” we use exact topology but place only one manhole on each link, that manhole being eligible for opening and sampling flow contents on that link. But our algorithmic methods are general and can handle any number of manholes on a link.

In Algorithm 1, the single downstream manhole closest to the infected residence is called the ***source manhole*** or ***source node***. It is the only source of SARS-CoV-2 remnants entering the sewage tree network. This is the manhole we seek to identify. In Algorithm 2, there will be multiple source nodes, each being a closest downstream manhole to an infected neighborhood.

We start by assigning Bayesian probabilities [21] to all possible source manholes, each probability reflecting our initial belief that the infected address is assigned to that manhole.

These probabilities reflect professional beliefs and need not be created by detailed data analysis, as the results are not that sensitive to their exact values. One could even use a simple heuristic, such as assigning each probability associated with a given manhole as proportional to the number of residents in that manhole's "catchment" zone.

Marlborough, Massachusetts in Middlesex County is a typical charming New England small city, covering 22.2 square miles with a population of 38,500. After describing the logic of our algorithms first with small "toy problems," we will utilize a portion of the sewer system network of Marlborough to illustrate a realistic application of the algorithms. Figure 1 depicts a portion of the sewer pipe network for the city of Marlborough, which has one water treatment plant and more than 3,000 manholes. In our stylized modeling of the Marlborough system, we consider a subset of the system having a reduced 844 manholes, each a possible source node. The wastewater treatment plant is denoted "WTP", the terminal node of the tree network. While a tree-network structure is not readily apparent by viewing the sewage system when superimposed on the town's map (Figure 1(a)), we can redraw its pipe network to reveal the tree structure as shown in Figure 1(b). We will operate on the tree network depiction of Figure 1(b).

We now explain Algorithms 1 and 2 that find Patient Zero and the Hot Spot neighborhood respectively. We present two simple "toy" examples to illustrate the steps of each algorithm, each step representing the opening of a manhole and testing its contents. We then apply the

algorithms on the stylized network of Marlborough, Massachusetts, and report the required number of manhole-opening steps to find the source(s) of infection in the city.

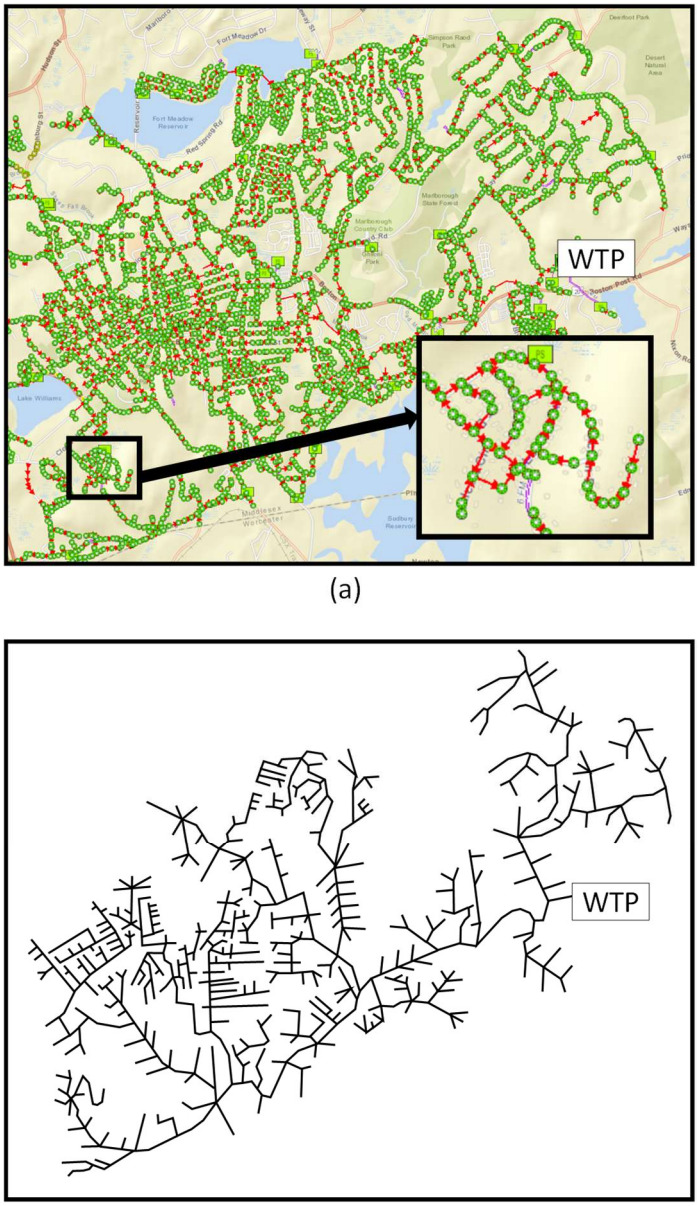


Figure 1- Top panel depicts a map of the Marlborough, Massachusetts wastewater removal system; the red arrows represent the direction of flow. The green circles are the manholes of the sewage pipe system. Bottom panel is a reduced stylized network depiction of Marlborough's sewage network. "WTP" in both panels represents the wastewater treatment plant.

2.1 Algorithm 1: Finding Patient Zero

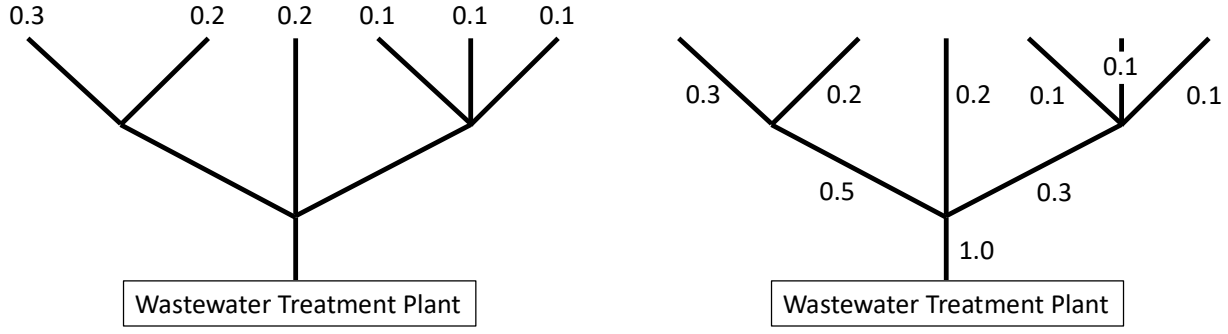
Iteration 1. We use a variant of "binary search" [22], an algorithm that finds the position of a target value within a sorted array. Binary search has the property that it "throws away" about half of the remaining solution space at each iteration. We seek to do the same, to throw away about half of the manholes at each iteration, none of them possibly being the desired source node. For each link we derive the "flow probability" as the probability that the infection is upstream of the link. The flow probability of each link is the sum of Bayesian probabilities of source nodes upstream of that link.

At Iteration 1 we seek a link in the tree that has about 50% of the Bayesian probabilities upstream from the link and the remaining Bayesian probabilities downstream from the link. Since it is unlikely that any subset of Bayesian probabilities will sum exactly to 0.5, we settle for the link whose probability is closest to 0.5. We identify a physical manhole within this link, open it, obtain a sewage sample, and then test it. If positive for SARS-CoV-2, then we know for sure that the residence we are seeking and its associated closest downstream manhole are upstream from this point; as a consequence, we throw away all downstream nodes of the network. Else we know the reverse and throw away all the upstream nodes.

Iteration 2. We are left with a tree network that is a subgraph of the original network, now with fewer nodes than the previous iteration. We readjust the Bayesian probabilities on the surviving subgraph so that they sum to 1.0. We repeat the logic of Iteration 1: We find a link on this subgraph that has about half of the Bayesian probabilities upstream from the link and the

remaining downstream. We identity a physical manhole on this "50-50" link, open it, obtain a sewage sample, and then test it. This is Test #2. If positive for SARS-CoV-2, then we know for sure that the residence we are seeking and its associated closest downstream manhole are upstream from this point; as a consequence, we throw away all downstream nodes of the network. Else we know the reverse and throw away all the upstream nodes. The above process continues until we have found Patient Zero. The algorithm always converges to Patient Zero.

We present the steps of Algorithm 1 with a simple example of Figure 2. As mentioned above, each source node is assigned a Bayesian probability as shown in Figure 2(a). For each link of the tree we now assign a corresponding probability flow as shown in Figure 2(b). Our algorithm sequentially samples a set of links using Bayesian-induced probability flows as a measure of the virus intensity in each link. Note that the sum of probability flows into the wastewater treatment plant is one.

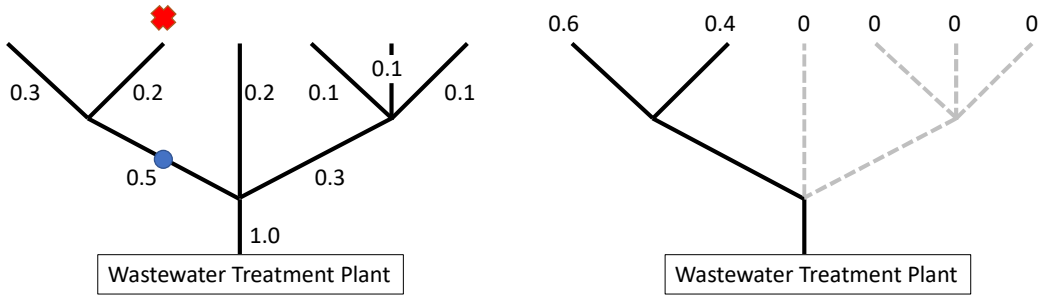


(a) Prior Bayesian Probabilities (b) Probability flows

Figure 2- Bayesian probabilities and probability flows of a simple network.

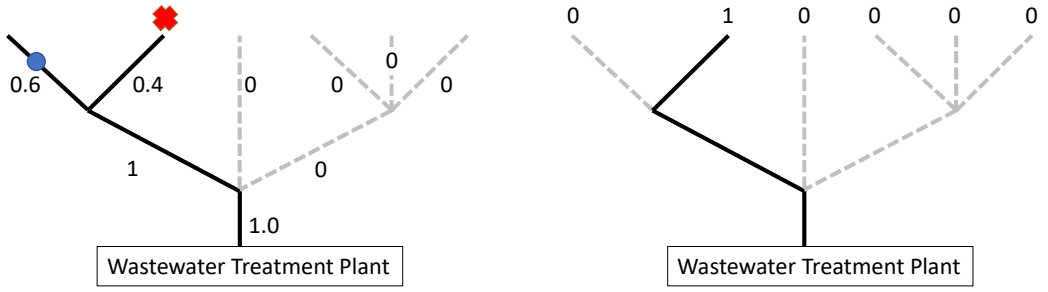
We demonstrate the iterations of Algorithm 1 in the example of Figure 3. The infected node (which is not known to the algorithm) is depicted with a red "X". In Iteration 1 we choose to

sample the link depicted with a blue circle in Figure 3a since it has probability flow of 0.5 and all other probability flows are less than 0.5. Because the sample tells us that the infection is upstream of the left link, we can discard the central and right links (Figure 3(b)). In Iteration 2 we normalize the probabilities (Figure 3(b)), recalculate the probability flows. Now we have a tie in the proximity to 0.5, and we choose to sample the link with the larger flow (i.e., 0.6 instead of 0.4), as shown in Figure 3(c). Because the sample tells us that the infection is downstream of the left link, we throw away the left-most link. We can now terminate the search since the infected link is identified (Figure 3(d)).



(a) Iteration 1: Choose a sample link

(b) Iteration1: Update source probabilities



(c) Iteration 2: Choose a sample link

(d) Algorithm Termination

Figure 3- Bayesian probabilities and associated probability flows of a simple network. The infected node is depicted with a red "X" and the sample link is depicted with a blue circle. The dashed lines depict eliminated links.

2.2 Algorithm 2: Finding the Hot Spot neighborhood

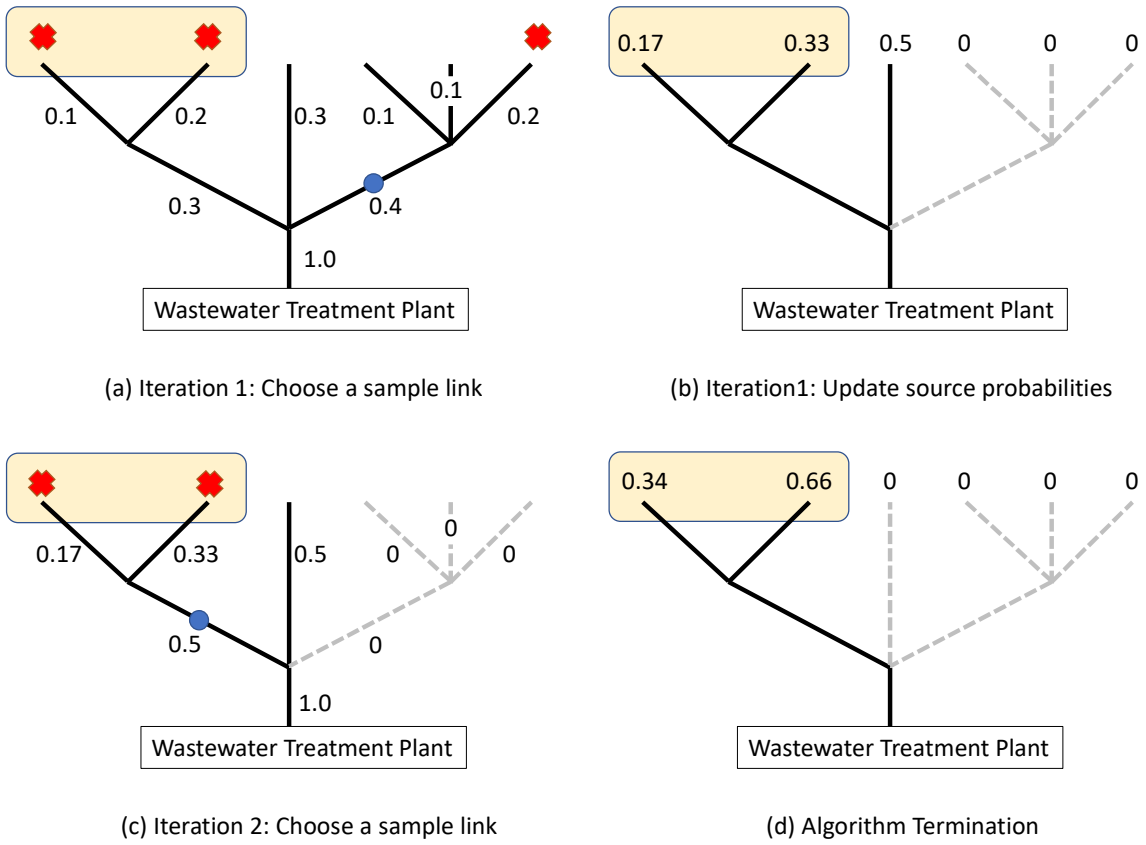
As mentioned above, Algorithm 2 is designed to find a neighborhood with the highest level of viral infection. We consider two or more nearby nodes as comprising a neighborhood, the exact definition is left to the operator of the algorithm. As a preliminary step in Algorithm 2, we first sample and record the virus load in the intake sewage flow at the stem link of the tree, the single link that connects the tree network to the wastewater treatment plant. This sample gives us the overall total virus load in the city. Our algorithm seeks to find the neighborhood with the largest contribution to total system viral load. The goal now is to sample the interior links of the sewage network to find the Hot Spot neighborhood. We now explain the steps of Algorithm 2.

Iteration 1. We use binary search to reduce the network until we are left with the Hot Spot neighborhood. We first record the total Coronavirus in-flow F entering the wastewater treatment plant and multiply all node probabilities and link probability flows by F. Given the Bayesian probabilities and total virus load F, the revised flow f on any given link is now the *expected* virus load on that link. The actual viral flows will differ, as to be revealed by sampling. We then sample the branch having expected virus flows upstream and downstream closest to a virus flow of $F/2$. We identify this link and obtain a sewage sample. This step is the same as Step 1 in Algorithm 1. If the recorded network viral load at the sampled link from upstream exceeds $F/2$, we through away the downstream nodes. Else we know the reverse and throw away all the upstream nodes.

Iteration 2. We now have a tree network that is a subgraph of the original network, usually with significantly fewer nodes. We re-normalize the Bayesian probabilities on the subgraph and repeat the logic of Iteration 1 to open and test a physical manhole on a new "50-50" link. We throw away upstream or downstream nodes of the tree in the same way as Iteration 1. The process continues until we are within a reasonable vicinity of the Hot Spot neighborhood. The stopping rule is human-based, subjective, depending on the size of the neighborhood and other considerations. By the viral-load-comparison mechanism of selection, we see that at each iteration the surviving subgraph has a higher average level of viral load per node than in any previous and now discarded subgraph. We can view the logic as a steepest ascent heuristic, always obtaining a subgraph with higher average viral load. Being a greedy algorithm, we are not guaranteed to find the hottest "Hot Spot," as there may exist a small neighborhood in a now-discarded subgraph having more viral intensity than that which we find with the algorithm.

We present the steps of Algorithm 2 with a simple example that has three infected nodes as shown in Figure 4(a). The total level of infection at the stem link is $F = 1$. The infected nodes (which are not known to the algorithm) are depicted with a red "X". In Iteration 1 we choose to sample the link depicted with a blue circle in Figure 4(a) since it has probability flow 0.4 (closest to 0.5) and all other flows are less than 0.4. We discard the right branch because the sample point detected only 1/3 of the total recorded viral load, *F* = 1 (Figure 4(b)). In Iteration 2 we re-normalize the probabilities (Figure 4(b)) and choose to sample the left branch which has a probability flow of 0.5. This sample tells us that 2/3 of the stem link infection is from the

upstream nodes. As there are only two nodes upstream of the sample point, we terminate the algorithm as we have found the Hot Spot neighborhood (Figure 4(d)).



(a) Iteration 1: Choose a sample link

(b) Iteration1: Update source probabilities

(c) Iteration 2: Choose a sample link

(d) Algorithm Termination

Figure 4- Bayesian probabilities and probability flows of a simple network. The infected nodes are depicted with a red "X" and the sample link is depicted with a blue circle. The dashed lines depict discarded links. The yellow box depicts the Hot Spot neighborhood.

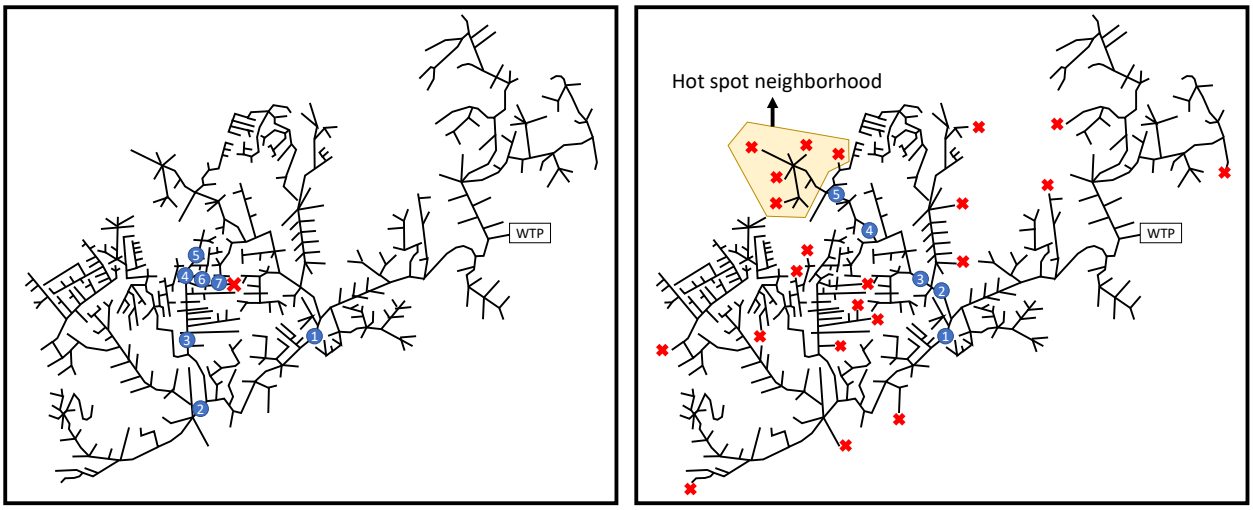
2.3 Returning to Marlborough, Massachusetts

We now implement the two algorithms on our stylized model of the Marlborough sewage network. To generate the nodal Bayesian probabilities, we independently draw random numbers from a unit uniform probability distribution for each potential source manhole and normalize the random draws so that the Bayesian probabilities sum to one. We first consider the Patient Zero scenario where only a single residence is infected, as shown by the red X near

the city center in Figure 5(a). We later vary the locations of the single infection and evaluate Algorithm 1's performance in finding many different infected nodes. In selecting the various different locations, we use Monte-Carlo sampling (without replacement) from the Bayesian probability distribution. In practice, municipal data sources can provide more accurate estimations of these Bayesian probabilities.

As shown in Figure 5(a), Algorithm 1 required only 7 samples to locate the infected source node. Sample 1 divided the network into two relatively equal size sub-graphs. Samples 2 and 3 helped us to determine the general whereabouts of the infected zone. Samples 4, 5, 6, 7 pinpointed the exact location of the infection.

Switching to Algorithm 2, we next consider the Hot Spot neighborhood scenario where several nodes are infected as shown in Figure 5(b). We have randomly generated the locations of infected nodes by taking random draws from the Bayesian probability distribution. As shown in Figure 5(b), the algorithm converged in five samples to the neighborhood having the largest number of infected nodes.



(a) (b)

Figure 5- Left and right panels depict the Patient Zero and the Hot Spot neighborhood scenarios, respectively. In the right panel, each red X is an infected manhole and the Hot Spot neighborhood is depicted in yellow.

We now discuss the performance of the two algorithms for Marlborough, assuming widely different locations of infection. Our previous analysis shown in Figure 5(a) depicted a single scenario of the problem. Subsequently, we randomly generated 100 scenarios, each having a new set of Bayesian probabilities. For each scenario, the infected node is randomly selected by taking a Monte Carlo random draw from the Bayesian probability distribution. For each scenario, we then use Algorithm 1 to find the infected node. The frequency distribution of the number of required samples is shown in Figure 6(a), ranging from a minimum of 7 to a maximum of 11 samples being required over 100 generated scenarios. This is a surprisingly small number of samples considering the size of the network with its 844 manholes.

We use a similar approach to generate scenarios for the case of finding the Hot Spot neighborhood. Recall that each source manhole has an assigned Bayesian probability. For each

scenario, we use Monte Carlo Sampling of the Bayesian probabilities to determine whether or not any given node is infected, iterating over all nodes. We adopt a stopping rule that terminates the algorithm if the detected viral load divided by the number of upstream source nodes from two consecutive samples is less than a predefined threshold. (In practice, we believe that the stopping rule would be subjectively human created.) We generated 100 different scenarios. In Figure 6(b), we present the number of samples to find the Hot Spot neighborhood, showing a minimum of 4 and maximum of 7 samples. Algorithm 2 requires fewer samples because it does have to zero in on a single infected node, instead it finds the vicinity of the Hot Spot neighborhood.

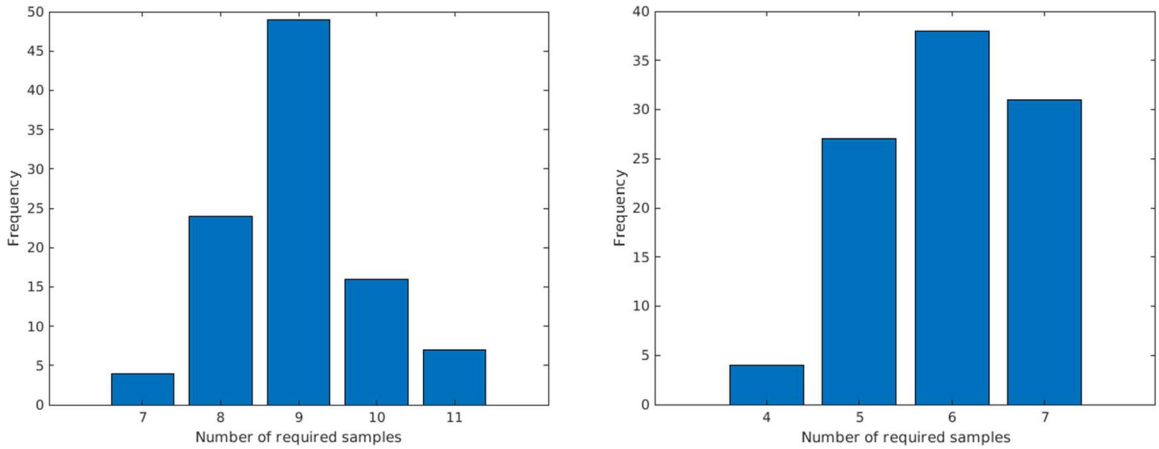


Figure 6. The number of sampled manholes to find (a) Patient Zero, and (b) the Hot Spot neighborhood in a stylized Marlborough, Massachusetts.

We now consider a scenario where we have prior information that two geographical areas of Marlborough have a higher chance of being infected. We consider two ***clusters*** as likely Hot Spots shown in Figure 7. The sum of Bayesian probabilities of the source nodes in Clusters 1 and 2 is 0.5 and 0.25, respectively. The sum of Bayesian probabilities over the rest of the city is 0.25. We generate the Bayesian probabilities in the same way as above but ensure that the sum of

probabilities in Cluster 1 is 0.5, Cluster 2 -- 0.25, and the remainder of the city -- 0.25. Similar to above, we randomly generate 100 scenarios, each having a new set of Bayesian probabilities, and use Monte Carlo sampling of the Bayesian probabilities to choose the infected nodes.

We present the number of required samples for the two algorithms in Figure 8. Algorithm 1 in Figure 8(a) has two peaks in the distribution of the required number of samples. The left peak is associated with Cluster 2 which has fewer source nodes and is less intricate to search than Cluster 1. It is easier to find the infected node in Cluster 2. The right peak is associated with Cluster 1 which typically required more samples to find Patient Zero. Algorithm 2 in Figure 8(b) also has two peaks for the two clusters.

We note that the maximum number of samples of Algorithm 1 with clusters is 13 (Figure 8(a)), in contrast to 11, its counterpart without clusters (Figure 6(a)). The cluster scenario may require more samples than the no-cluster scenario in cases when the infected node (or the infected neighborhood in the Hot Spot case) is outside of the clusters. In such cases, one might say that the Bayesian probabilities give misleading information on the whereabouts of the infection, thus, more samples are required.

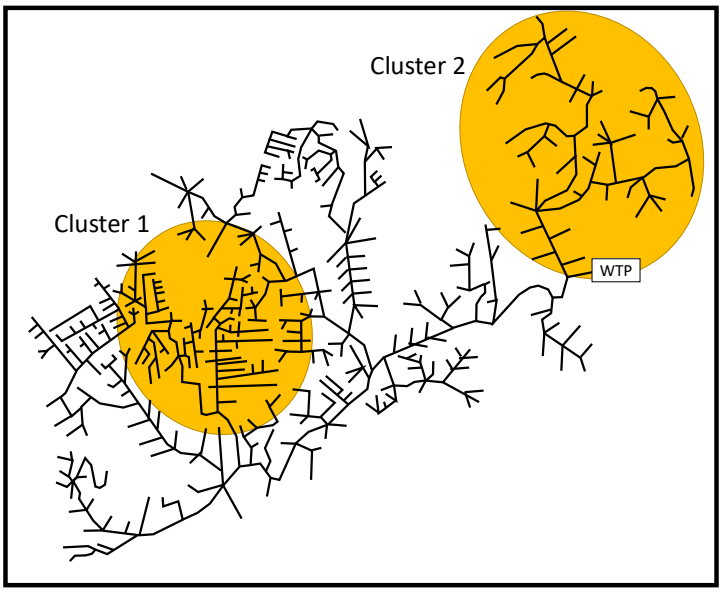


Figure 7. Marlborough, Massachusetts, with two infection clusters. Sum of source node Bayesian probabilities in Clusters and 1 and 2 is 0.5 and 0.25, respectively. For the remainder of the city, the sum of source node probabilities is 0.25.

We compare the two algorithms in the cluster and no-cluster scenarios in Table 1. The mean number of samples is smaller in the cluster compared to the no-cluster scenario. In Algorithms 1 and 2, the cluster scenario reduces the mean number of samples by 14% and 22%, respectively. However, the cluster scenario has a larger variance in the number of samples, which, as explained above, is predominantly due to cases where the infected nodes are outside of the clusters. According to Table 1, we have a larger right-tailed skew in the number of samples when the infection is clustered, showing that we can find the Hot Spots with fewer number of samples compared to the no-cluster scenario.

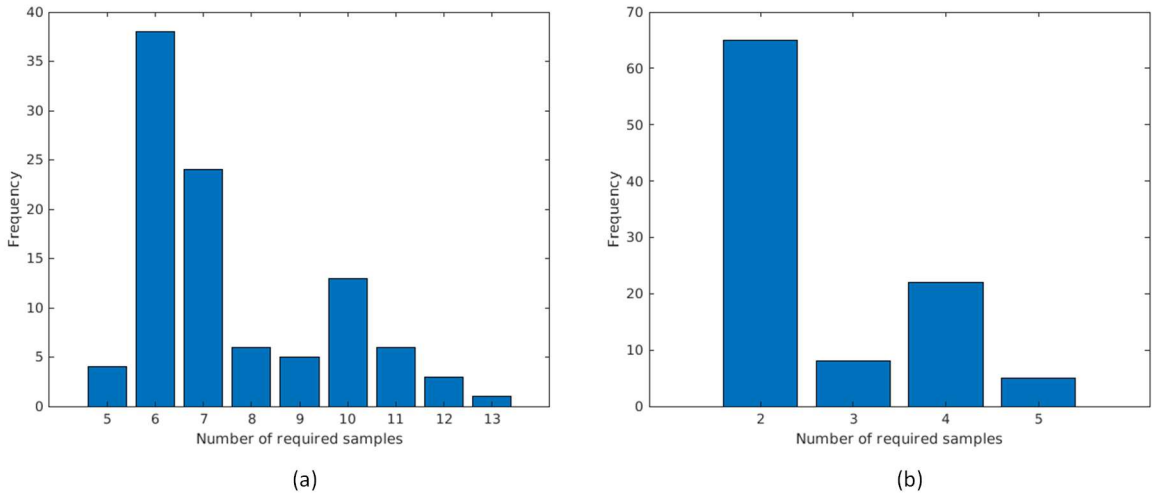


Figure 8. The number of sampled manholes to find (a) Patient Zero, and (b) the Hot Spot neighborhood in a stylized Marlborough, Massachusetts, when there are clustered infections.

Table 1. Mean, variance, and skewness of the number of samples required to find Patient Zero or the Hot Spot neighborhood with and without infection clusters.

	Mean	Variance	Skewness
Random Infections: Patient Zero	8.98	0.75	0.42
Clustered infections: Patient Zero	7.71	3.39	1.09
Random Infections: Hot Spot	5.96	0.75	-0.31
Clustered Infections: Hot Spot	4.66	1.48	1.42

Finally, moving away from clusters, the two algorithms work well in large as well as small networks. As an example, in a test of Algorithm 1 on randomly generated networks with approximately 3,000 links and 1,000 source nodes, we found typically that only 11 manhole samples were required to home in on the desired location. Applying to both simulated and actual sewer networks, we perform extensive computational and complexity analysis in [24].

3. Reflections

We have presented two algorithms that utilize viral markers of SARS-CoV-2 contained in human stool, and thus also present in human sewage, to help identify location(s) in a community of individuals currently infected. Each algorithm recommends a dynamically-dependent sequence of manholes to sample in order to home in on the locations(s) of the infected individuals. In the first case, we are seeking one or a small number of persons with COVID-19 living at the same address. In the second we are seeking a Hot Spot neighborhood within the community.

The programmed algorithms are very fast on any computer, and they suggest only a small fraction of a community's manholes need to be sampled and tested. For Algorithm 1, for instance, one could double the size of the community being analyzed and typically only add one extra manhole to sample and test! This important property derives from the binary search logic that attempts to cut away half of the eligible manholes at each iteration.

As mentioned earlier, our work is a bit ahead of the current science and technology of COVID-19 tracing within a community. Here are the major issues ahead of us, for our algorithmic approach to work well in practice.

3.1. Fast Testing.

For our methods to work, we need rapid testing on the spot at the manholes, each test administered to the newly obtained sewage sample. There are current efforts to create such tests, including one being labeled as the "fast two-dollar test."

3.2. Sewage system realities.

Municipal sewage systems are not pristine like hospital or university labs. They are notoriously under-maintained. Foreign substances other than human waste can occupy these systems. Thus, it remains to be seen how well these systems can accurately maintain and convey virus remnants during the winding tributary-and-main-line journey to the sewage treatment plant.

3.3. Lack of Time Averaging.

Most of the recent sewage content analysis by researchers cited in the opening of this paper uses time-averaged measurements of SARS-CoV-2 remnants. We do not have the luxury of time averaging with our fast testing and quick identification of the next manhole to test. This should not be a problem for Algorithm 2, since it focuses on a populated neighborhood of infected individuals and does not seek to estimate full viral load, only relative viral load in contrast to other neighborhoods. But it can be a problem with Algorithm 1 since we are dealing there with one or a small number of individuals. The viral signature they input to the sewage system at 9:00 AM is likely to be a burst of viral load, perhaps then followed by no input until later in the day or perhaps even the next day. So, there is a reasonable chance that this individual's viral signature in the sewage flow varies up and down during the day, possibly causing our sequential sampling system to miss the signature. The science of fluid dynamics suggests that the initial burst should dissipate and spread out over the journey through the system, thus smoothing the signature in the sewage flow. Such smoothing, if it occurs, could help our analysis. Also, many people having COVID-19 also have persistent diarrhea, and that

unfortunate condition would tend to even the signature flows over the course of a day. Finally, in private conversations with a senior hydraulic engineering research professor, we learn that no real-time sewage flow is required in our testing to reveal remnants of COVID-19. With no current flow, we sample and test the sludge in the manhole to verify (or not verify) upstream presence of the virus. Operationally, we see many question marks going forward. Only future research will help clarify these nontrivial issues.

3.4. Privacy.

The focus of each algorithm is to speed testing of residents in a community by identifying "most likely" addresses of infection. Algorithm 1 homes in on a few houses on one street, while Algorithm 2 focuses on identifying a likely neighborhood, say comprising several city or town blocks. Residents who are subsequently selected for testing by our focused new means may object on privacy worries. In the context COVID-19, privacy concerns have been expressed over use of cell phone tracking, in-house monitoring devices, facial recognition software and more [23]. It is too early to know if tracking of virus-laden sewage would spark similar concerns.

3.5. More Methodological Research Needed.

This is the first paper to offer a way to test manholes sequentially to home in on locations of individuals having COVID-19. Refinements to each algorithm can be made, and we have already begun their study.

We did not devote attention to evaluating the relative Bayesian probabilities. Their careful estimation for this procedure could be an entirely separate paper, and in practice, could result in much-improved performance. For instance, a neighborhood in which the majority of people have jobs that require leaving the house and working in an environment with substantial human interaction is likely to generate more COVID-19 cases than one in which most residents can work from home via the Internet. These differences can be expressed by markedly different values of the Bayesian probabilities assigned to neighborhoods.

Finally, once our algorithmic procedures are tried in practice, building a database of results along with practical problems, there undoubtedly will be a need to revise the algorithms and related procedures to adapt to the realities in the field.

3.6. Breaking News from Italy

As we submit this paper, we learn that analysis of wastewater in Italy has revealed the presence of SARS-CoV-2 in December, more than two months before known cases arose there [25]. Researchers at Italy's National Institute of Health (ISS) reported that recent re-analysis revealed remnants of SARS-CoV-2 in the wastewater of Milan and Turin in December 2019. These findings point to the invaluable attribute of wastewater as an early warning system for impending SARS-CoV-2 community infection. Given such early warnings, tracing methods such as our Algorithm 2 could help identify regions within cities having the highest likelihood of serous outbreak.

ACKNOWLEDGEMENTS

The authors would like to thank Evan C. Larson for taking the lead to explore and find publicly available wastewater treatment networks. His support accelerated the preparation of this paper and improved the experiments section.

We also thank Professor Ed Kaplan of Yale University for his careful reading of an earlier version of the manuscript and for sharing some of his team's research experiences in New Haven, Connecticut.

References

1. Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, *et. al.* spread of SARS-CoV-2 in the Icelandic population. *New England Journal of Medicine* April 14, 2020. DOI: 10.1056/NEJMoa2006100 https://www.nejm.org/doi/full/10.1056/NEJMoa2006100

2. Ing AJ, Cocks C, Green JP. COVID-19: in the footsteps of Ernest Shackleton. *Thorax BMJ*. May 2020. http://dx.doi.org/10.1136/thoraxjnl-2020-215091

3. Laurer SA, Grantz KH, Bi Q, Jones F, Zheng Q, *et. al.* the Incubation Period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of Internal Medicine*. 5 May 2020. https://doi.org/10.7326/M20-0504.

4. Lamers MM, Beumer J, van der Vaart J, Knoops K,' Puschhof J, *et. al.* SARS-CoV-2 productively infects human gut enterocytes. *Science* 01 May 2020: eabc1669 mDOI: 10.1126/science.abc1669 https://science.sciencemag.org/content/early/2020/04/30/science.abc1669

5. Wadman, M, Couzin-Frankel J, Kaiser J, Matacic C. how does coronavirus kill? clinicians trace a ferocious rampage through the body, from brain to toes. *Science*. April 17, 2020. https://www.sciencemag.org/news/2020/04/how-does-coronavirus-kill-clinicians-trace-ferocious-rampage-through-body-brain-toes

6. Xiao F, Tang M, Zheng X, Liu Y, Li X, *et. al*. evidence for gastrointestinal infection of SARS-CoV-2. March 3, 2020. *Gastroenterology*; 158:1831–1833. https://www.gastrojournal.org/article/S0016-5085(20)30282-1/fulltext

7. Sims N, Kasprzyk-Hordern B. future perspectives of wastewater-based epidemiology: monitoring infectious disease spread and resistance to the community level. Environment International. Published online, April 4, 2020. V. 139. June 2020. https://www.sciencedirect.com/science/article/pii/S0160412020304542?via%3Dihub

8. Hart, OE, Halden RU. computational analysis of SARS-Cov-2/COVID-19 surveillance by wastewater-based epidemiology locally and globally: feasibility, economy, opportunities and challenges. *Science of the Total Environment*. 5/23/2020. https://www.sciencedirect.com/science/article/pii/S0048969720323925?via%3Dihub

9. Wurtzer S, Marechal V, Mouchel J-M, Maday Y, Teyssou R, *et. al.* evaluation of lockdown impact on SARS-CoV-2 dynamics through viral genome quantification in Paris wastewaters. *medRxiv* Non-peer-reviewed paper. 05/06/20 https://doi.org/10.1101/2020.04.12.20062679

10. Nemudryi A, Nemudraia A, Surya K, Wiegand T, Buyukyoruk M, *et. al*. temporal detection and phylogenetic assessment of SARS-CoV-2 in municipal wastewater. *medRxiv* Non-peer-reviewed paper. April 20, 2020. https://www.medrxiv.org/content/10.1101/2020.04.15.20066746v1

11. Peccia J*,* Zulli A, Brackney DE, Grubaugh ND, Kapan EH, *et. al*. SARS-CoV-2 RNA concentrations in primary municipal sewage sludge as a leading indicator of COVID-19 outbreak dynamics. *medRxiv* Non-peer-reviewed paper. May 22, 2020. https://www.medrxiv.org/content/10.1101/2020.05.19.20105999v1

12. Wu F, Amy Xiao, Zhang J, Gu X, Lee WL, *et. al.* SARS-CoV-2 titers in wastewater are higher than expected from clinically confirmed cases. *medRxiv* Non-peer-reviewed paper. April 27, 2020. https://www.medrxiv.org/content/10.1101/2020.04.05.20051540v1

13. Rimoldi SG, Stefani F, Gigantiello A, Polesello S, Comandatore F, *et. al*. presence and vitality of SARS-CoV-2 virus in wastewaters and rivers. *medRxiv.* Non-peer-reviewed preprint. May 5, 2020. https://www.medrxiv.org/content/10.1101/2020.05.01.20086009v1

14. Medema G, Heijnen L, Elsinga G, Italiaander R, Brouwer A. presence of SARS-Coronavirus-2 in sewage. *medRXiv***,** Non-peer-reviewed preprint. March 30, 2020. https://doi.org/10.1101/2020.03.29.20045880

15. Gillespie C. is coronavirus sewage testing the key to slowing the spread of COVID-19? here's what the experts say. *Health*. May 8, 2020. https://www.health.com/condition/infectious-diseases/coronavirus/coronavirus-sewage-testing

16. Lesté-Lasserre C. coronavirus found in Paris sewage points to early warning system. *Science*. April 21, 2020. https://www.sciencemag.org/news/2020/04/coronavirus-found-paris-sewage-points-early-warning-system doi:10.1126/science.abc3799

17. Public Works Department, Engineering Services, City of Surprise, Arizona. CHAPTER 7 – SEWER SYSTEM DESIGN STANDARDS. pp 7-1, 7-2. Undated. https://www.surpriseaz.gov/DocumentCenter/View/3162/Sewer-System-Design-Standards?bidId=

18. Public Works Department, Engineering Services, City of Surprise, Arizona. City of Surprise water infrastructure master plan. 2004. https://www.surpriseaz.gov/DocumentCenter/View/1133/2004-Infrastructure-Master-Plan

19. Lexington Water and Sewer. https://www.lexingtonma.gov/water-and-sewer 2020.

20. Johnson S. so many manholes, so little time. *Concrete Construction.* April 9, 2012. https://www.concreteconstruction.net/projects/infrastructure/so-many-manholes-so-little-time_o

21. Zyphur MJ, Oswald FL. Bayesian probability and statistics in management research: a new horizon. *Journal of Management*, 39 (1), pp: 5-13. October 17, 2012. https://journals.sagepub.com/doi/10.1177/0149206312463183

22. Lin A. binary search algorithm. *WikiJournal of Science*, 2 (1). July 2, 2019, doi:10.15347/WJS/2019.005, Wikidata Q81434400

23. Lin, L, Martin TW. how is coronavirus eroding privacy. *Wall Street Journal.* April 15, 2020. https://www.wsj.com/articles/coronavirus-paves-way-for-new-age-of-digital-surveillance-11586963028 }

24. Nourinejad M, Berman, O, Larson RC. "Computational Results over Small and Large Networks on the New Manhole Sampling Algorithm to Home in on COVOD-19-Infected Addresses." Forthcoming.

25. Ide E. Virus already in Italy by December, sewers show. *Barron's*. June 19, 2020. https://www.barrons.com/news/virus-already-in-italy-by-december-waste-water-study-01592554804?tesla=y