Estimating Spread of Contact-Based Contagions in a Population Through Sub-Sampling

Sepanta Zeighami
USC
zeighami@usc.edu

Cyrus Shahabi
USC
shahabi@usc.edu

John Krumm
Microsoft Research
jckrumm@microsoft.com

ABSTRACT

Various phenomena such as viruses, gossips, and physical objects (e.g., packages and marketing pamphlets) can be spread through physical contacts. The spread depends on how people move, i.e., their mobility patterns. In practice, mobility patterns of an entire population is never available, and we usually have access to location data of a subset of individuals. In this paper, we formalize and study the problem of estimating the spread of a phenomena in a population, given that we only have access to sub-samples of location visits of some individuals in the population. We show that simple solutions that estimate the spread in the sub-sample and scale it to the population, or more sophisticated solutions that rely on modeling location visits of individuals do not perform well in practice. Instead, we directly model the co-locations between the individuals. We introduce PollSpreader and PollSusceptible, two novel approaches that model the co-locations between individuals using a *contact network*, and infer the properties of the contact network using the sub-sample to estimate the spread of the phenomena in the entire population. We analytically show that our estimates provide an upper bound and a lower bound on the spread of the disease in expectation. Finally, using a large high-resolution real-world mobility dataset, we experimentally show that our estimates are accurate in practice, while other methods that do not correctly account for co-locations between individuals result in entirely wrong observations (e.g, premature prediction of herd-immunity).

PVLDB Reference Format:
Sepanta Zeighami, Cyrus Shahabi, and John Krumm. Estimating Spread of Contact-Based Contagions in a Population Through Sub-Sampling. PVLDB, 14(9): 1557 - 1569, 2021.
doi:10.14778/3461535.3461544

1 INTRODUCTION

Phenomena that Spread through Contact. Viruses spread across a population through contacts and so do news, gossips, ideas and habits. Packages are passed between individuals when they meet to reach a destination. *Physical contacts* are responsible for such phenomena passing from one individual to another. Their *spread*, defined as how many people in a given population the phenomenon reaches, impacts our day to day lives, with COVID-19 as an on-going exhibit. Since the spread of such phenomena happens through contacts, it primarily depends on people's *mobility patterns* (e.g., how people move in a city and meet others), as well as the characteristics of the phenomenon which determine how it can be passed on from one person to another, or the *diffusion model* (e.g., the probability of transmission from one person to another given that they are co-located). The mobility pattern determines when and where contacts happen, while the diffusion model determines how a phenomenon spreads when there are contacts.

Role of Mobility Patterns. Mobility patterns play a fundamental role in the spread of any phenomena in a population and its analysis. To shed more light on their role we consider two aspects of mobility. (1) *Visits* form the *location sequence* of an individual which contain the information about the locations an individual has been to. (2) *Co-locations* between multiple individuals, a by-product of their visits, which contain the information on the contacts between individuals. The spread of a phenomena in a population depends on the second aspect, i.e., co-locations, since it determines which individuals were in contact and able to pass an item between them. This simple observation informs much of our later discussions.

Benefits of Analyzing Spread. The spread of the phenomena discussed has significance for both policy making and personal decision making. Given access to the mobility pattern of all individuals in a population and a diffusion model, we can simulate how a phenomena spreads in the population. This can be used to identify hotspots, compute location and individual risk-scores [8, 22, 31] and study various interventions or what-if scenarios [6, 13, 14, 16, 21]. For instance, in the case of COVID-19, we can model the impact of wearing a mask as a change in the diffusion model (e.g., by lowering the probability of a transmission during a contact), and see how the spread differs from not wearing a mask, as done in [9]. Having access to the mobility patterns for different cities at different times, allows us to perform these studies at a high resolution, e.g., for specific neighbourhoods and time periods.

Mobility Patterns of a Population. The benefits discussed above can only be materialized if we have access to mobility patterns for an entire population. However, that is not feasible in practice. Location sequences can be collected through cell-phones, but it is difficult, if not impossible, to convince every individual to share their location with a single entity. Nevertheless, running simulations with millions of users and billions of locations is computationally demanding.

Meanwhile, location data of subsets of individuals, obtained through their cell-phones, has become available. For instance, Veraset [2], a data-as-a-service company, provides anonymized population movement data collected through GPS signals of cell-phones across the US. Such datasets provide high-resolution and detailed mobility patterns of parts of the population, but will likely never contain the entire population. It is important to be able to fully and correctly utilize such datasets to analyze the spread of a phenomenon over an entire population.

Proceedings of the VLDB Endowment, Vol. 14, No. 9 ISSN 2150-8097.
doi:10.14778/3461535.3461544

Problem of Up-Sampling. In this paper, we formalize and study the problem of estimating the spread of a phenomena in a population, given access to sub-samples of location visits of individuals in the population. Our goal is to estimate how many people in the whole population a phenomenon will reach, given a diffusion model and sample visits of the population. For instance, by observing only a portion of the population in a city, solving this problem allows us to estimate how much COVID-19 will spread under different intervention strategies (e.g., if people wear or do not wear masks) for the entire population.

Existing Solutions and Challenges. A simple solution to the problem is to consider the spread in the sub-sampled population, and then scale the estimation to the whole population. However, we observed that such an approach vastly underestimates the spread because it ignores the co-locations between the unobserved individuals and the sampled individuals. In fact, the main challenge in solving the problem is accurately modeling co-locations between unobserved individuals and sampled individuals. We observed in our experiments that when this is not done accurately, the estimation can provide wrong infection patterns, e.g., estimating that the spread is stopping when the spread is actually increasing. This is because the underestimation gets amplified over time. For instance, underestimating the number of people who are currently infected in a population leads to further underestimating how many people the disease can be transmitted to in the future.

An alternative approach is to model the visits of the unobserved individuals using the visits of the sampled individuals, and use that model to infer co-locations between unobserved and sampled individuals. For instance, an approach can be to generate synthetic location trajectories, e.g., using [12, 26], based on the observed location sequences to create a larger dataset that contains both real and synthetic location sequences. However, we observed that this indirect formation of co-locations from synthetic location data has two limitations. First, the co-locations cannot be formed accurately because our model needs to generate extremely accurate synthetic locations, within a few meters (to generate correct co-locations), and for long periods of time (to be able to track infected individuals correctly). In addition, creating synthetic locations corresponding to real-world populations explodes the data size, rendering simulations at scale impractical. A detailed discussion of these methods is provided in Section 5 and we experimentally evaluated a representative of such approaches in Section 4 which confirmed the above observations.

Effectively, due to lack of access to individual location data, agent based simulations that are used to assess various interventions for containing the spread of a disease [6, 13, 14, 16, 21], utilized across the world for the COVID-19 pandemic, use fixed contact matrices to generate co-locations between individuals. These contact matrices contain, at an aggregate level, the rate of contacts between different compartments in the population (e.g., the rate of contacts between people of age 10 with people of age 50 in a population), and are created through surveys and interviews in various parts of the world [25, 29, 30]. In the absence of individual location data, such contact matrices can be useful for studying the spread in a population at an aggregate level (e.g. for a country). However, since the contact matrices are static, i.e., do not change with time, and are created at an aggregate level, they do not take into account the spatiotemporal changes in mobility (e.g., change in mobility on a day-to-day basis, or for different neighbourhoods in a city). Thus, they cannot be used to provide accurate estimate of the spread at a particular point in time and space. We believe using real location data and an accurate method for estimating the spread can help empower the above-mentioned studies to better understand the impact of different interventions.

Our Approach. We rigorously study the problem of up-sampling. Our approach is to statistically estimate the probability of a sampled individual getting infected. Such an estimate needs to take into account the probability of a sampled individual getting infected by unobserved individuals. Rather than modeling location visits of an individuals, from which co-locations can be indirectly inferred, we directly model co-locations between individuals. This follows our observation that the extra information associated with location visits (e.g., their exact Geo-coordinates and their entire sequence) is unnecessary for modeling co-locations, while modeling such extra information makes the model less accurate.

Our methods use a time-varying contact network [32] to model the co-location between individuals. The general approach is to use some statistics of the contact network, that can be estimated from sub-sampled individuals, to estimate the spread. We discuss two different ways this can be done. Our first approach, called Polling the Spreader, or PollSpreader, does this from the spreaders view: by modeling how many individuals a person can transmit the phenomena to. We observe that estimating statistics for this kind of modeling is difficult over long periods. Thus, we discuss our second approach, called Polling the Susceptible, or PollSusceptible, where we estimate the spread from a different perspective: we model different ways the phenomena can be transmitted to a particular individual. Using this approach, we provide lower and upper bounds on the spread of the phenomena in the whole population from the sub-sample. We experimentally observed that our estimates follow closely the spread in the whole population. Furthermore, our results show that the pattern of spread can be completely misjudged (e.g., showing early herd-immunity by mistake) if the co-locations are not estimated accurately.

Contributions and Organization. In this paper, we

- Define the problem of estimating spread through sub-sampled visits (Section 2);
- Present two novel methods, PollSpreader and PollSusceptible, that solve the problem accurately (Section 3);
- Theoretically study the problem and provide lower and upper bound estimates of the number of infected people in the population over time using PollSusceptible; and
- Experimentally show that our estimations are accurate in practice. (Section 4)

Furthermore, Sections 5 and 6 discuss the related work and our conclusion, respectively.

2 PROBLEM DEFINITION

We study the problem of the spread, across a population, of a phenomenon that passes from one person to another through their co-location. We assume only a sub-sample of the population are observed, and our goal is to estimate the number of individuals in

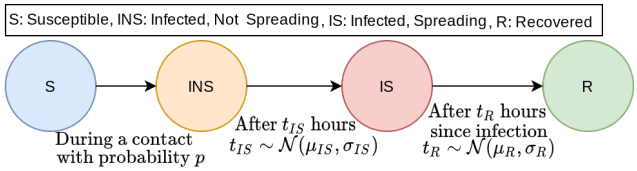

Figure 1: Diffusion Model

the true population who become subjected to the phenomenon. We next define the terminology used, summarized in Table 1.

Mobility Pattern. We consider a population consisting of n individuals. For each individual, a visit provides the location of the individual at a particular point in time. Associated with each individual, u, is a *visit sequence* or *location sequence*, which is a sequence of the location of their visits over times. For two consecutive visits, c and c' of an individual at times t and t', we assume the individual is at location specified by c from time t to t'. Given the visit sequence of u, we define l_u to be a function returning the location of u at any point in time, i.e., for any t, $l_u(t) = (x_t^u, y_t^u)$ where x_t^u and y_t^u are the coordinates of the location of u at time t. We abuse the notation and also refer to l_u as u's visit sequence. **Diffusion Model**. Here we discuss how diffusion occurs over the population. For ease of discussion, and without loss of generality, we adopt a terminology commonly used when a disease spreads in a population and in particular the SIR model [20] which is commonly used in epidemiology. The model is shown in Figure 1. Each individual is in either the Susceptible (S), Infected (I) or Recovered (R) compartment. We refer to the compartment an individual belongs to as their *status*. Recovered individuals are assumed to be immune to the disease (or deceased), and thus will not contract the disease anymore. Infected individuals are either able to spread the disease (called Infected and Spreading, or IS) or they are not (called Infected and not Spreading, or INS). Susceptible individuals can contract the disease from Infected and Spreading individuals. After an S individual, u, contracts the disease at time t, u immediately become INS. Then, at time $t + t_{IS}$, where t_{IS} is a positive number sampled from some distribution Θ_{IS}, u becomes IS. Furthermore, at time $t + t_R$, where t_R is a positive number larger than t_{IS} and sampled from some distribution Θ_R, u becomes R.

Consider a Susceptible individual, u, and an Infected and Spreading individual, v, and their corresponding visit sequences l_u and l_v. At any time, t, with probability $p = f(l_u, l_v, t)$, u contracts the disease from v, where $f(l_u, l_v, t)$ is an application-dependent and user-defined function. The function $f(l_u, l_v, t)$ determines the probability of transmission of the disease from one person to another depending on the location of the individuals. Intuitively, there can be a non-zero probability of transmission when two individuals are close enough spatially for a long enough duration.

Individuals are initially infected based on some probability distribution. At time zero, a vector c is sampled from a distribution Θ_{init}, where $c \in \{0, 1\}^n$ and the i-th element of c determines whether the i-th individual in the population is initially infected or not. An initially infected user becomes INS, and transition to IS and R states following the same procedure discussed above.

Notation	Meaning
p_s	Prob. of an individual begin sampled
θ_{IS}	Distribution of time it takes to become IS
θ_R	Distribution of time it takes to become R
θ_{init}	Distribution of initial infections
μ_{IS}	Time it takes to become IS
μ_R	Time it takes to become R
d_{max}	Max. distance for transmission
t_{min}	Min. duration for transmission
p_{init}	Prob. of an initial infection
p_{inf}	Prob. of infection given a contact
S_u	R. v. denoting if u is sampled
X_u^t	R. v. denoting if u gets infected until time t
$\hat{p}_u^t$	Estimate of $P(X_u^t = 1)$ from sub-sample
$N(u, G_{s,t})$, N_u	Neighbourhood of u in $G_{s,t}$
N_u^A	$N(u, G_{s,t}) \setminus A$ ("\" is set difference operator)
EI_t	Estimate for the Up-Sampled Infected Count

Table 1: Table of Notations

Problem Statement. Our goal is to estimate the extent of the spread over the population from just observing the visit sequences of a *sub-sample* of the population. Specifically, consider a population $\mathcal{U}$ of n individuals and a set $\mathcal{S} \subseteq \mathcal{U}$, where $\mathcal{S}$ is sampled from $\mathcal{U}$ uniformly at random. Each individual is sampled independently and with probability p_s. We use the term *sub-sampled population* to refer to the set $\mathcal{S}$ and *whole population* to refer to $\mathcal{U}$. We refer to individuals in $\mathcal{U} \setminus \mathcal{S}$ as unobserved individuals.

Definition 2.1 (Up-Sampling Infected Count Problem). Give a sub-sampled population, $\mathcal{S}$, and the parameters of the diffusion model, for any given time t, the Up-Sampling Infected Count Problem is to return the expected number of individuals who have gotten infected in the whole population, $\mathcal{U}$, until time t.

EI_t denotes any estimation of the answer to the Up-Sampling Infected Count Problem. Solving the analogously problem for other compartments (e.g., Susceptible or Recovered) will be similar, and thus we focus only on estimating the number of Infected individuals.

3 ESTIMATING THE SPREAD

Solving Up-Sampling Infected Count Problem would have been trivial had we had access to the true status of the sub-sampled individuals. If we knew exactly k_t of the sub-sampled individuals are infected at time t (or knew their exact probability of being infected), then an unbiased estimate of the number of infected individuals in the population would have been $\frac{k_t}{p_s}$. However, except for $t = 0$, obtaining the correct value of k_t is difficult, since a sampled individual may get infected by individuals who were not sampled. In this section, we first discuss a concrete diffusion model and introduce the necessary terminology (Section 3.1) and illustrate the difficulty of obtaining such an estimate with two naive solutions (Sections 3.2 and 3.3). We then present our methodology (Section 3.4) and discuss its generalization to other diffusion models (Section 3.5).

3.1 Terminology and Diffusion Model

Diffusion Model. For ease of discussion and concreteness, we present our methodology on a diffusion model with the following parameter setting. The parameters Θ_{IS} (time to spreading) and Θ_{IR} (time to recovery) are set to a distribution that returns μ_{IS} and μ_R respectively with probability 1, i.e. they are deterministic. Furthermore, Θ_{init} (initial infections) is set such that every individual is initially infected independently with probability p_{init}. Finally, $f(l_u, l_v, t)$ is defined as follows. Intuitively, if an IS and an S individual are within d_{max} of each other for at least t_{min}, then the IS individual will infect the S individual with probability p_{inf}, for user-defined parameters p_{inf}, d_{max} and t_{min}. More formally, consider the time t_1 when the co-location between u and v starts. That is, $d(l_u(t_1), l_v(t_1))$ is at most the parameter d_{max} and that right before t_1, $d(l_u(t), l_v(t)) > d_{max}$. Let t_2 be the timestamp when the co-location between u and v ends. That is, t_2 is the first timestamp after t_1 such that $d(l_u(t_2), l_v(t_2)) > d_{max}$. If the duration of the co-location, i.e., the time difference from t_2 to t_1, is at least a parameter t_{min}, then $f(l_u, l_v, t_1) = p_{inf}$, for the parameter p_{inf} denoting the probability of infection. $f(l_u, l_v, t) = 0$ for all other timestamps. We use the term *contact* to refer to co-locations within a distance d_{max} that last for at least t_{min} units of time.

Terminology. For an individual u, the indicator random variables S_u is equal to 1 if u is sampled and X_u^t is equal to 1 if u gets infected at a time less than or equal to t. We refer to infections that were caused by transmission through k individuals starting from an initial infections as *k-hop infections*. That is, for a sequence of individual $< v_0, v_1, v_2, ..., v_k >$, we call an infection a k-hop infection if the individual v_k was infected by v_{k-1}, v_{k-1} was infected by v_{k-2} and so on, and that v_0 was an initial infection (patient zero).

3.2 First Attempt: Scaling

Our first simple solution is to use Monte Carlo simulation to estimate the spread of the disease in the sub-sampled population.

Methodology. Given a sub-sample, the randomness in the spread is due to the randomness both in the initial infections and the transmissions. To simulate the spread, a number of individuals are initially infected according to Θ_{init}. Then, the visit sequences of the individuals are used to determine contacts between IS individuals and S individuals, and for every such contact the S individuals gets infected with probability p_{inf}. At time t, define k_t to be the number of people that are infected in the simulation. We run the simulation r times, obtaining the estimate k_t^i for time t from the i-th run. We return $\frac{1}{p_s}\frac{\sum_i k_t^i}{r}$ as our estimate of expected number of infections at time t. Here $\frac{1}{r}$ is for taking the mean of the r simulations, and $\frac{1}{p_s}$ is for scaling up from the sub-sample to the full population.

Analysis. Such an estimate provides a biased estimate of the expected number of infections that occur after the initial infections. It specifically underestimates the number of infections for any time $t > 0$. To analyze the accuracy, for a specific run of the simulation, let $\hat{X}_u^t$ be a random variable denoting if an individual u gets infected by other sampled individuals until time t. Then, the estimate, EI_t is

$$EI_t = \frac{1}{p_s}\sum_u S_u \hat{X}_u^t \tag{1}$$

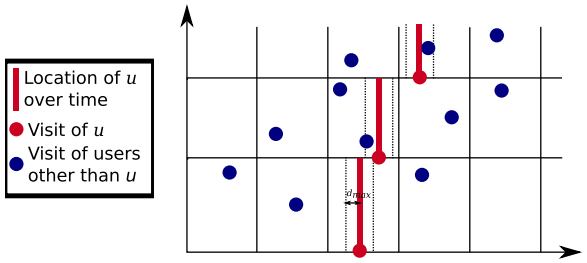

Figure 2: Modeling distribution of visits for an individual u. For ease of display, we do not show latitude.

First, note that at time $t = 0$, the estimate is unbiased. That is, taking the expected value of Eq. 1 for $t = 0$, we obtain $E[S_u\hat{X}_u^0] = E[S_u]E[\hat{X}_u^0] = p_s \times p_{init}$ and the expected number of infections in the whole population is $n \times p_{init}$. Thus, the estimate $\frac{1}{p_s}\sum_u S_u\hat{X}_u^0$ is an unbiased estimate of the number of infections at time $t = 0$.

However, consider $\hat{X}_u^t$ for $t > 0$. Observe that $P(\hat{X}_u^t = 1)$ can be less than $P(X_u^t = 1)$. This is because, with some probability, individuals with whom u has contacts may not be sampled, reducing the probability of u getting infected from the sampled individuals. In terms of the Monte Carlo simulation, this can be seen as the case of *false negatives*. That is, even assuming initial infections and the transmission are deterministic, u may get infected from someone in the whole population, but may not get infected from anyone in the sub-sample, because the transmission of the disease to u in the whole population is through unobserved individuals.

Observations. Our experiments show that this underestimation is significant, amplified because it cascades through multi-hop infections (e.g., second-hop infections are underestimated, which in turn makes the estimate of third-hop infections worse). Simply scaling up to the full population does not work, which we show in the experiments in Section 4. This observation suggests that we need to account for the visit sequence that were *lost* when sub-sampling.

3.3 Second Attempt: Location Modeling

3.3.1 Framework. Section 3.2 illustrates the need for accounting for the unobserved individuals. Our framework to do so is to estimate $P(X_u^t = 1)$. We denote by $\hat{p}_u^t$ an estimate of $P(X_u^t = 1)$. To solve the issues arising in Section 3.2, we need to account for the unobserved individuals when calculating $\hat{p}_u^t$. This differs from our first attempt in how we model the unobserved individuals to calculate $\hat{p}_u^t$. After we obtain $\hat{p}_u^t$ for all sub-sampled individuals, our final estimate is

$$EI_t = \frac{1}{p_s}\sum_u S_u \hat{p}_u^t \tag{2}$$

Comparing Eq. 2 with Eq. 1, the main difference in this approach is the flexibility it allows in estimating $P(X_u^t = 1)$.

3.3.2 Second Attempt: Density-Based Estimation. Our second attempt models the distribution of the visits of the unobserved individuals to calculate the probability of a sub-sampled individual being infected by unobserved individuals.

Methodology. Our approach follows four steps: (1) Modeling the visit probability distribution of the unobserved population, (2) calculating the probability of u getting infected by a single visit of an unobserved individual when u is in a particular cell, (3) estimating probability of u getting infected when it's in a particular cell, and (4) estimating $P(X_u^t = 1)$ by taking into account past visits of u.

Step (1), Modeling Visit Distribution: We discretize time and space using a grid. We use a uniform grid across space, but a per user grid across time. That is, for each user, we model the visit distribution with a different histogram, but all the histograms model the same distribution, i.e., the distribution of the check-ins of all the individuals. The grid across time uses event-based discretization for each individual. In particular, for an individual u, let $C =< t_0, t_1, ... >$ be the sequence of times we have observed a visit. According to the discussion in Section 2, u is in the same location during time t_i to t_{i+1} for all i. Thus, for user u, we discretize time based on C. The choice of griding for time simplifies our analysis as every visit now corresponds to a cell, but other griding strategies can be similarly applied. For each grid cell (t, i, j), let $n_{t,i,j}$ be the number of observed visits that fall into that cell, where we also count the individuals who are already in the cell $g = (t, i, j)$ when u enters the cell, to account for infections happened through individuals already in the cell. Figure 2 shows the grid used for modeling for an individual u, where $n_{t,i,j}$ is calculated by counting the number of blue circles. Note that for different individuals, we will have different horizontal partitioning in the figure. We define

$$p_{t,i,j} = \frac{n_{t,i,j}}{\sum_{i',j'} n_{t,i',j'}} \tag{3}$$

to model the probability of a visit falling into the cell (t, i, j). Furthermore, we assume visits inside a cell are distributed uniformly in space. In this step, we do not model the amount of time that an observed individual v (blue dot in Figure 2) is in the cell.

Step (2), Prob. of Infection from a Single Visit: Consider a visit c, corresponding to a grid cell $g = (t, i, j)$ of a sampled susceptible individual, u. Assume we know that a visit c' by some unobserved infected individual happens when u is in g. We need to calculate the probability of c' being within d_{max} of c and lasting for at least t_{min}. Let $p_{c'}$ be the probability that c' causes u to get infected and p_{long_enough} the probability of a co-location lasting at least t_{min}.

$$p_c = p_{t,i,j} \times \frac{\pi d_{max}^2}{\text{cell area}} \times p_{long\_enough} \tag{4}$$

where $\frac{\pi d_{max}^2}{\text{cell area}}$ is the probability of co-location given that the visits are in the same cell (ignoring the edge cases where co-locations happen from two different cells). In Figure 2, $p_{t,i,j} \times \frac{\pi d_{max}^2}{\text{cell area}}$ calculates the probability of a new blue circle (visit by others) falling within the dashed line of the red rectangle (within d_{max} of location of u) for a particular time. We model the duration an individual stays at a location with an exponential random variable, so p_{long_enough} is the probability of the random variable being more than t_{min}.

Step (3), Prob. of Infection in a Cell: If there are N_c visits by unobserved infected users for the duration that u is in the cell, probability of the user getting infected during that time period is

$$p_t = 1 - (1 - p_c)^{N_c} \tag{5}$$

To estimate N_c, we estimate (a) the number of unobserved IS individuals, N_{IS} and (b) the average number of visits per IS individual $N_{c \text{ per IS}}$. Then we estimate $N_c = N_{IS} \times N_{c \text{ per IS}}$. For (a), assume that the current time is t. An IS individual must have been infected some time between $t - \mu_{IS}$ and $t - \mu_R$. We use our estimate of the number of infected individuals for times $t - \mu_{IS}$ and $t - \mu_R$ to calculate the number of current IS individuals. For (b), we use $N_{c \text{ per IS}} = \frac{\sum_{i,j} n_{t,i,j}}{p_s}$, where $\sum_{i,j} n_{t,i,j}$ is the total number of observed visits during the time u is in cell g, and we scale it by $\frac{1}{p_s}$ to get to the whole population (here we have assumed that the average number of visits per IS individuals is the same as the average number of visits for all individuals).

Step (4), Prob. of Infection until current time: If u enters N_t cells until time t, the probability of u getting infected until time t is

$$\hat{p}_u^t = 1 - \prod_{i=0}^{N_t} (1 - p_i) \tag{6}$$

Observations. We experimentally observed that this approach still leads to underestimation (see Section 4). This happens for two reasons. First, on one hand, there are inevitable inaccuracies in modeling the visits spatially due to modeling assumptions. For instance, the use of a grid assumes locations that are spatially close have similar visit density. This may not be true in practice, because for instance a church may exist in a residential area. Although this may be possible to address by increasing the granularity of the grid cells, doing so will require a large set of sampled users to avoid over-fitting (as otherwise a visit will have zero probability of falling into most gird cells). On the other hand, to accurately model co-locations, a model that is accurate to within a few meters is needed. Second, the temporal correlations between the visits needs to be modeled. For instance assume that an infection happens in a particular grid-cell, g at time t. It means that an unobserved IS user, v, was in g at time t. This changes the probability distribution of the locations where v can be at time $t + 1$. Overall, both points imply that an approach that aims to up-sample the infection data by modeling the location sequences requires a very precise modeling of the location sequences both spatially and temporally. We present accuracy results from this model in Section 4.

An interesting observation is that our end-goal is to use the model of the location sequences to find possible co-locations between individuals. That is, the spatial information associated with a location sequence is only a means to an end, but is not necessary. Instead, we can directly model the possible co-locations.

3.4 Third attempt: Co-location Modeling

Motivated by the above observations, we directly model the contacts between the individuals. We present two approaches that provide bounds on the expected spread of the disease in the population. Our first approach, PollSpreader, models the problem from the spreaders' view, i.e., aims at calculating how many people will get infected given a number of spreaders (Section 3.4.2). Although this approach works well when modeling first-hop infections, up-sampling for multi-hops becomes difficult. Therefore, we present our second approach, PollSusceptible, which looks at the problem from the susceptible view, e.g., directly calculates the probability of a susceptible individual getting infected (Section 3.4.3). Both

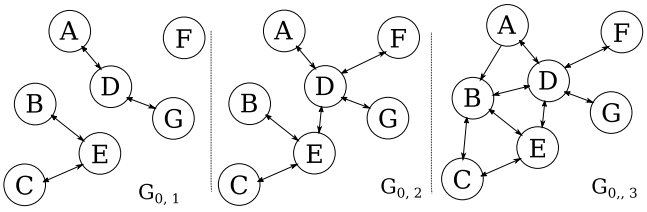

Figure 3: A contact network

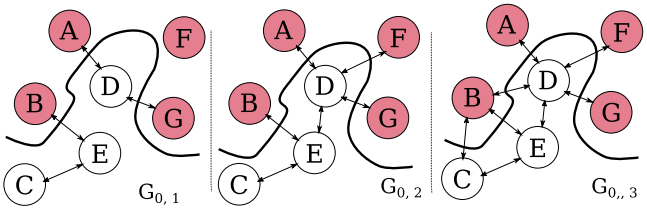

Figure 4: A cut on the network (Red: Initial infections, White: Rest of the individuals)

approaches use a time-dependent contact network which we first describe in Section 3.4.1. For the sake of space, we discuss the time and space complexity of our algorithms in our technical report [36].

3.4.1 Dynamic Contact Network. We use a dynamic weighted directed graph to model contacts between individuals. For the contacts from time s until t, we build a graph $G_{s,t} = (V, E_{s,t})$, where V contains a node for every individual and $E_{s,t}$ contain edges between users for whom there exists a co-location between during the time period [s, t]. Furthermore, the weight on the edge (u, v), denoted by $w_{u,v}$, is the probability that v gets infected from u given that u is infected at time s. Figure 3 shows an example of a contact network. We use $\hat{G}_{s,t} = (\hat{V}, \hat{E_{s,t}})$ to refer to a contact network built from a sub-sample. $\hat{G}_{s,t}$ has fewer nodes than $G_{s,t}$.

3.4.2 PollSpreader: Up-Sampling using Spreaders' View. We study how many first-hop infections can occur based on an initial number of infections. Our approach follows the framework of Section 3.3.1, and we aim at calculating $\hat{p}_u^t$. We first explain how the contact network can be used to do this if we have access to the entire population. Then, we discuss how the approach can be adjusted when we only observe a sub-sample of the population. Finally, we discuss the difficulty in extending the approach to multi-hop infections. We refer to this method as Polling the Spreader or PollSpreader.

Calculating $\hat{p}_u^t$ from the Whole Population . Assume we have access to the whole population. Our approach works as follows. (1) We calculate the total number of contacts between initial infections and the susceptible population. Then, (2) we use that to estimate the probability of a susceptible individual getting infected.

Step (1), Calculate Number of Infectious Contacts: We estimate the number of *infection events*, informally defined as when an initially infected user, v, has contact with an individual v who was not initially infected. Formally, define the infection random variable $I_{u,v}^t$ as an indicator random variable which is equal to one if u is an initially infected person and u infects v until time t given that no other individual infects v. We call $I_{u,v}^t = 1$ an *infection event*. Our goal is to estimate how many *infection events* occur in total in the whole population by calculating $E[\sum_{u,v} I_{u,v}^t]$.

Calculating $E[\sum_{u,v} I_{u,v}^t]$ can be formulated as estimating the expected total weight of the edges crossing a random graph cut on the graph $G_{0,t}$. Figure 4 shows how this is done for a specific realization of initial infections, where Figure 4 is obtained by a random selection of initial infections from Figure 3, that is, the red nodes are randomly selected initial infections. Any such random selection defines a cut on the graph for every timestamp. The goal is to estimate the expected number of edges crossing such a cut. Specifically, consider the following random cut, where a node is in the set I if it is initially infected, and in the set S if it's not. Every cut corresponds to an initialization of the infections. Given a cut, the expected number of infection events until time t is equal to the total weight of the edges crossing the cut from I to S (for each different t, we consider the cut on a different graph, as $G_{0,t}$ changes with time). Thus, the expected number of infection events is equal to the expected total weight of the edges crossing the cut from I to S. For every edge (u, v), assign the random variable $Z_{u,v}$ to be equal to $w_{u,v}$ if $u \in I$ and $v \in S$, and zero otherwise. We have that

$$E[Z_{u,v}] = p_{init} \times (1 - p_{init}) \times w_{u,v} \tag{7}$$

and the expected total weight of the edges crossing the cut is

$$E[\sum_{u,v} I_{u,v}^t] = e_c^t = \sum_{(u,v) \in E_{0,t}} p_{init} \times (1 - p_{init}) \times w_{u,v} \tag{8}$$

Step (2), Estimating $\hat{p}_u^t$ from Infection Events: We assume the infection events occur uniformly at random across the susceptible population, or in other words, the edges from I fall uniformly at random over the nodes in S. Although more sophisticated modelling approaches may be possible, we observed that this assumption works well in practice. Thus, given $Z_{u,v}$ for all u and v, the probability of a node in S having at least one edge incident on them is $1 - (1 - \frac{1}{|S|})^{\sum Z_{u,v}}$. We estimate this quantity by $1 - (1 - \frac{1}{E[|S|]})^{e_c^t}$. Finally, an individual either gets infected initially (with probability p_{init}), or does not get infected initially but gets infected through contact with u. Thus, we estimate probability of getting infected as

$$\hat{p}_u^t = p_{init} + (1 - p_{init})(1 - (1 - \frac{1}{E[|S|]})^{e_c^t}) \tag{9}$$

Calculating $\hat{p}_u^t$ from the Sub-Sampled Population . An advantage of this modelling is how easily it can be adjusted when we only have access to a sub-sample of the population. In Eq. 9, we only need to be able to estimate $E[e_c]$. To do so, we slightly modify our graph cut formulation. Specifically, let $\hat{I}$ be the set of nodes that are sampled and initially infected, and let $\hat{S}$ be the set of nodes that are sampled and not initially infected. Furthermore, let $\hat{Z}_{u,v}$ be equal to $w_{u,v}$ if $u \in \hat{I}$ and $v \in \hat{S}$, and zero otherwise. We have that

$$E[\hat{Z}_{u,v}] = p_s^2 \times p_{init} \times (1 - p_{init}) \times p_{u,v} \tag{10}$$

Comparing Eq. 10 with Eq. 7, we observe that adjusting for sampling can now be done just by scaling. Thus, an unbiased estimate of the total expected number of edges crossing the cut is $\frac{\sum_{u,v} \hat{Z}_{u,v}}{p_s^2}$. So

$$E[\sum_{u,v} I_{u,v}^t] = \hat{e}_c^t = \frac{1}{p_s^2} \sum_{(u,v) \in \hat{E}_{0,t}} p_{init} \times (1 - p_{init}) \times w_{u,v} \tag{11}$$

Algorithm 1 Calculating Prob. of Transmission Through a Path, $calc\_prob\_transmision(s, t)$

Require: Sequence of $k+1$ individuals, s, $s =< v_0, v_1, ..., v_k >$ and infection time of v_k, t_I
Ensure: $P(T_s^t = 1 | v_k \text{ gets infected at } t_i)$
1: **if** $k == 0$ **then**
2: **return** 1
3: $t_{IS} \leftarrow t_I + \mu_{IS}$
4: $t_R \leftarrow t_I + \mu_R$
5: $c_{v_k, v_{k-1}} \leftarrow$ list of contacts between v_k and v_{k-1} from t_{IS} to t_R ordered by time
6: $p \leftarrow 0$
7: **for** c_i in $c_{v_k, v_{k-1}}$ **do**
8: $t_c \leftarrow$ time c_i occurs
9: $p_{inf\_by\_c} \leftarrow calc\_prob\_transmision(< v_0, ..., v_{k-1} >, t_c)$
10: $p \leftarrow p + p_{inf\_by\_c} \times p_{inf} \times (1 - p_{inf})^{i-1}$
11: **return** p

Algorithm 2 Calculating Prob. of Infection by a Path, $calc\_prob\_inf(s, t_I)$

Require: Sequence of $k+1$ individuals, s, $s =< v_0, v_1, ..., v_k >$ and current time t
Ensure: Lower and upper bounds estimates of $P(T_s^t = 1)$
1: $p_{not\_l} \leftarrow 1$
2: $p_{not\_u} \leftarrow 1$
3: **for** v' in $N(v_k, \hat{G}_{0,t}) \setminus \{v_0, ..., v_k\}$ **do**
4: $p_{v'\_l}, p_{v'\_u} \leftarrow calc\_prob\_inf(< v_0, ..., v_k, v' >)$
5: $p_{not\_l} \leftarrow p_{not\_l} \times (1 - p_{v'\_u})$
6: $p_{not\_u} \leftarrow p_{not\_u} \times (1 - p_{v'\_l})$
7: $p_{starts\_at\_v_k} \leftarrow calc\_prob\_transmision(s, 0)$
8: $\hat{p}_l \leftarrow p_{starts\_at\_v_k} + (1 - p_{starts\_at\_v_k}) \times (1 - (p_{not\_u})^{c_u})$
9: $\hat{p}_u \leftarrow p_{starts\_at\_v_k} + (1 - p_{starts\_at\_v_k}) \times (1 - (p_{not\_l})^{c_l})$
10: **return** $\hat{p}_l, \hat{p}_u$

Our final estimate simply replaces e_c^t in Eq. 9 with $\hat{e}_c^t$.

Challenge for Multi-Hop Infections. The main challenge associated with this method is generalizing it to more than one-hop neighbours. The contact network can be generalized to a *reachability network* [32], such that there is an edge from a node u to another node v if u can infect v through multi-hop infections. However, taking sampling into account becomes difficult, because the probability of an edge being sampled now depends on which of the intermediary nodes are sampled, and the number of k-hop paths from u to v in the population. Thus, we need to accounts for all possible multi-hop paths containing unobserved secondary infections.

3.4.3 PollSusceptible: Up-Sampling using Susceptibles' View. To account for unobserved secondary infections, here, we directly model the probability of a susceptible person getting infected. This is our second graph-based approach, which proved to be the most accurate in our experiments. By making some independence assumptions, this allows us to naturally model multi-hop infections. We refer to this method as Polling the Susceptible or PollSusceptible.

Calculating $\hat{p}_u^t$ from the Whole Population. We again start by assuming we have access to the entire population and use the same contact network discussed before. The general idea is to observe that an individual, u, who got infected was either initially infected, or got infected through some of its contacts that are modeled by the contact network. If u got infected through a contact, v, the same logic recursively applies. That is, v was either initially infected or got infected through a contact. To state this formally, we first introduce more terminology, then discuss the case of getting infected through neighbours and finally detail the recursion step.

Neighbourhood Terminology and Notation. Recall that the contact network represents, for each individual, u, if they are infected, whom they can infect and with what probability. Consider $v \in N(u, G_{s,t})$, where $N(u, G_{s,t})$ is the set of neighbours of u, that is the set of nodes, v, in $G_{s,t}$ such that there is an edge from u to v in $G_{s,t}$. $N(u, G_{s,t})$ represents the set of possible first-hop infections caused by u, given that u starts spreading the disease at time s and recovers at time t. We use the notation N_u to refer to $N(u, G_{s,t})$ when the contact network in question is clear from the context, and we use the notation N_u^A, for some set A, to denote $N(u, G_{s,t}) \setminus A$.

Calculating $\hat{p}_{v_0}^t$ through first-hop neighbours. Our goal is as before, calculating $\hat{p}_{v_0}^t$ for all users v_0 following the framework of Section 3.3.1. To do so, for individuals v_0 and v_1, let T_{v_0,v_1}^t be the indicator random variable equal to one if the disease is transmitted from v_1 to v_0 until time t. Now, the event that v_0 does not get infected until time t, $X_{v_0}^t = 0$, can be decomposed as the intersection of two events: one that v_0 does not get infected initially, $X_{v_0}^t = 0$, and that v_0 does not get infected by any of its contacts until time t, written as $T_{v_0,v_1}^t = 0$ for all $v_1 \in N_{v_0}$. The idea can be illustrated using Figure 3. For instance, for E to get infected until time 3, it either has to be initially infected, or not initially infected but get infected from B, C or D. Using this formulation, the probability that v_0 gets infected until time t is

$$\hat{p}_{v_0}^t = p_{init} + (1 - p_{init})(1 - \underbrace{P(T_{v_0,v_1}^t = 0, \forall v_1 \in N_{v_0})}_{p_{v_0}^N}) \quad (12)$$

where the two events of v_0 either gets initially infected (first term), or it does not get initially infected but gets infected through transmission (second term) are accounted for. Denote $p_{v_0}^N = P(T_{v_0,v_1}^t = 0, \forall v_1 \in N_{v_0})$ for ease of reference. Here, we make the modeling assumption that the random variables X_{v_0,v_1}^t are independent for all $v \in N_{v_0}$. Using the assumption,

$$p_{v_0}^N = \prod_{v_1 \in N_{v_0}} P(T_{v_0,v_1}^t = 0) \quad (13)$$

we acknowledge that this may not be necessary true for multi-hop infections, but we observed that the introduced error is negligible in practice per our experiments (see Section 4).

Combining Eqs. 12 and 13, the problem of calculating $\hat{p}_{v_0}^t$ is now reduced to calculating $P(T_{v_0,v_1}^t = 0)$ for each v_1.

Recursion for Multiple Hops. To calculate $P(T_{v_0,v_1}^t = 0)$, a similar logic can be recursively applied. $T_{v_0,v_1}^t = 0$ can occur only if v_1 is initially infected, or if it is infected by another one of its neighbours. Continuing with our example, in Figure 3, for E to be infected by D until time 3, D must have either been infected initially or gotten

infected from A, F or G. Also note that for E to be infected by B until time 2, B must have gotten infected initially. Specifically, let $T^t_{v_0,v_1,v_2}$ be the random variable equal to one if v_0 gets infected until time t by v_1 who gets infected by v_2, $v_2 \in N^{\{v_0\}}_{v_1}$. We write

$$P(T^t_{v_0,v_1}=1) = p_{init}P(T^t_{v_0,v_1}=1|X^0_{v_1}=1)+ \\ (1-p_{init})(1-\prod_{v_2\in N^{\{v_0\}}_{v_1}} P(T^t_{v_0,v_1,v_2}=0)) \quad (14)$$

More generally, $P(T^t_s=1)$, when $s =< v_0, ..., v_k >$ can always be recursively calculated as

$$P(T^t_s=1) = p_{init}\underbrace{P(T^t_s=1|X^0_{v_k}=1)}_{\text{full-path transmission}}+ \\ (1-p_{init})(1-\prod_{v_{k+1}\in N^s_{v_k}}\underbrace{P(T^t_{s,v_{k+1}}=0)}_{\text{recursion}}) \quad (15)$$

where s, v_{k+1} denotes the concatenation of v_{k+1} to s. A simple base-case for the recursion is when all the neighbours of v_k are already on the path from v_0 to v_k (in practice when estimating the spread until time t, we are generally able to stop the recursion based on the number of hops. This is because we know the time it takes for a given disease to be transmitted during each hop, and thus the number of hops that can happen until time t can be estimated).

The term labeled *recursion* in Eq. 15 shows the recursive calculation. The term labeled *full-path transmission* is the probability of transmission happening given a complete path, e.g., the path starts with a patient zero, v_k, who was initially infected and continues until v_0. The probability of a full-path transmission can be calculated by considering the contacts between every pair of individuals, v_i and v_{i-1}, on the path as explained below. Before explaining the details of this calculation, note that using Eqs. 15 and 13, we can calculate Eq. 12, which provides our estimate of $p^t_{v_0}$.

Calculating Probability of Full-path Transmissions. Algorithm 1 shows how this is done. The idea is to divide the event $T^t_s=1$, for $s =< v_0, ..., v_k >$, into a set of mutually exclusive events and sum up the probabilities of the events to calculate the final probability. Let $c_{u,v}$ be the list of contacts between u and v ordered by time, let $W(u,v,c,t_i)$, for $c \in c_{u,v}$ denote the event that u infects v through contact c given that u was infected at time t_i. Observe that the event $E = (T^t_s=1|X^0_{v_k}=1)$ is the same as $E = \cup_{c_i\in c_{v_k,v_{k-1}}}(T^t_{s\setminus v_k}= 1 \wedge W(v_k,v_{k-1},c_i,0))$ and that each event is mutually exclusive (because an individual can get infected at most once), where $s \setminus v_k$ is the sequence of elements in s excluding v_k. Note that

$$P(T^t_s=1\wedge W(v_k,v_{k-1},c_i,0)) = \\ \underbrace{P(T^t_{s\setminus v_k}=1|W(v_k,v_{k-1},c,0))}_{\text{recursion}}\times\underbrace{P(W(v_k,v_{k-1},c_i,0))}_{(1-p_{inf})^{i-1}\times p_{inf}}$$

The first term is calculated recursively. To calculate the second term, recall that c_i is the i-th contact between v_k and v_{k-1} since v_{k-1} became IS. Thus, for v_{k-1} to become infected through c_i, it has to be true that none of the previous $i-1$ contacts caused infection, and that c_i did. Since each of the events are independent, the probability of this happening can be calculated as $(1-p_{inf})^{i-1}\times p_{inf}$.

Calculating $\hat{p}^t_u$ from the Sub-Sample Population. Now assume we only have access to a sub-set of the population. Calculating full-path transmission probabilities can be done in the same way as before (Alg. 1). However, not all the neighbours of a user v_0 are sampled. That is, in Eq. 12, we do not have access to $N(v_0,G_{0,t})$, but only its sub-sample $N(v_0,\hat{G}_{0,t})$. Thus, we need to adjust Eq. 13 to estimate $p^N_{v_0}$ from our sub-samples. We denote by $\hat{p}^N_{v_0} = P(T^t_{v_0,v_1}=0,\forall v_1\in N(v_0,\hat{G}_{0,t}))$, i.e., $\hat{p}^N_{v_0}$ this is our estimate of $p^N_{v_0}$ from the sub-sampled population. The challenge in making such an estimate is that we are estimating the product across a population using a sub-sample. Providing unbiased estimates of summations using sub-samples can be easily achieved by merely scaling the sample statistic, but such an approach does not work for estimating products. This is because there is no simple relationship between $E[\hat{p}^N_{v_0}]$ and $p^N_{v_0}$. Instead, to be able to compute reliable estimates, we provide lower and upper bounds on $p^N_{v_0}$ as discussed in the following theorem. The theorem makes use of the quantity p_{min}, which is the largest number such that $p_{min} \le P(T^t_{u,v}=0)$, and is a bound on the probability of transmission.

Theorem 3.1. *Let $c_l = \frac{1}{ps}$. We have that*

$$E[(\prod_{v_1\in N(v_0,G_{0,t})} P(T^t_{v_0,v_1}=0)^{S_{v_1}})^{c_l}] \le p^N_{v_0}$$

Furthermore, let c_u be the solution to $c_u p_s + \frac{c_u^2\log(p_{min})}{8} = 1$. If $p_s \ge \sqrt{\frac{-log(p_{min})}{2}}$, we have that

$$p^N_{v_0} \le E[(\prod_{v_1\in N(v_0,G_{0,t})} P(T^t_{v_0,v_1}=0)^{S_{v_1}})^{c_u}]$$

Proof Sketch. We take the log of $p^N_{v_0} = \prod_{v_1\in N(v_0,G_{0,t})} P(T^t_{v_0,v_1}=0)$ and use Jensen's inequality and Hoeffding's lemma to bound it. Full proof available in our technical report [36]. □

Theorem 3.1 gives us estimates that bound $p^N_{v_0}$ from above and below on expectation. However, due to sub-sampling we do not have access to $P(T^t_{v_0,v_1}=0)$ exactly. We recursively apply Theorem 3.1 to obtain lower and upper bounds on $P(T^t_{v_0,v_1}=0)$, and then use those in the statement of the theorem. Thus, when modeling multi-hop infections, Theorem 3.1 is recursively applied. Furthermore, in practice, we estimate p_{min} from our observations (we estimate the maximum number of co-locations, n_{max}, between two individuals from the sub-sample, then estimate p_{min} as $(1-p_{inf})^{n_{max}}$).

Final Algorithm. Alg. 2 depicts our final algorithm to find lower and upper bounds on $\hat{p}^t_u$, for an individual. This is obtained by calling *calc_prob_inf*$(< u >, t)$. The base case for Alg. 2 is when s contains all the nodes (the algorithm will not enter the for loop in that case), although a more efficient base case is to also consider the time t and check whether there is a non-zero probability that adding another hop (i.e., line 4) will be able to spread the disease to v_0 until time t. This is possible because it takes μ_{IS} for every infected individual to spread the disease, and thus, if there are k hops on the sequence $< v_0, ..., v_k >$, it takes at least $k \times \mu_{IS}$ for the disease to spread from v_k to v_0.

3.5 Generalizing the Diffusion Model

Our methods are applicable to other diffusion models as long as they satisfy the following requirements. First, initial infections need to be independent. Although we have so far assumed this probability to be the same and p_{init} for all individuals, the method can be easily modified to use probability $p_{init,u}$ for each individual u, i.e., the initial probabilities do not have to be the same for all individuals.

Second, consider the function $f(l_u, l_v, t)$ that defines probability of transmission for two location sequences. For our PollSpreader, $f(l_u, l_v, t)$ is only used to calculate the edge weights of the contact network. Thus, the only requirement is for it to be possible to calculate the probability of transmission from time s to time t from $f(l_u, l_v, t)$ if u is infected at time s. For PollSusceptible, observe that the formulation in Alg. 2 does not depend on the function f, but Alg. 1 is dependent on the specific transmission model. Here, f needs to be a function that can be used to evaluate $P(T^t_{v_0,v_1,..,v_k} = 1|v_k$ gets infected at $t_i)$. This is also the case for Θ_{IS} and Θ_{IR}.

Although our framework is general enough to be able to handle different diffusion models, the efficiency of the calculation can be an obstacle for complex distributions. For instance, Alg. 1 in line 5 takes advantage of the fact that only certain contacts between time t_{IS} and t_R can cause transmission, which is in turn because t_{IS} and t_R are deterministically calculated (in lines 3 and 4). However, modifying Θ_{IS} and Θ_{IR} will require checking all contacts between v_k and v_{k-1} when there is a non-zero probability that v_k is IS.

4 PERFORMANCE EVALUATION

4.1 Experimental Methodology

We ran our experiments on a machine running Ubuntu 18.04 LTS equipped with an Intel i9-9980XE CPU (3GHz) and 128GB RAM.

Datasets. We consider two datasets. Our first dataset is the Veraset dataset. Veraset [2] is a data-as-a-service company that provides anonymized population movement data collected through location measurement signals of cell-phones across the US. We were provided access to this dataset for December 2019. The dataset consists of location signals of cell-phone devices, where each location signal is considered to be a visit, as defined in Section 2. Each record in the dataset consists of `anonymized_device_id`, `latitude`, `longitude`, `timestamp` and `horizontal_accuracy`. We assume each `anonymized_device_id` corresponds to a unique individual. We discard any location signal with `horizontal_accuracy` of worse than 25 meters. For a single day in December, there are 2,630,669,304 location signals across the US. Each location signal corresponds to an `anonymized_device_id` and there are 28,264,106 distinct `anonymized_device_ids` across the US in that day. Figure 5 shows the number of daily location signals recorded in the month of Dec. 2019 in the area of Manhattan, New York. Figure 6 shows the distribution of location signals across individuals in Manhattan in Dec. 2019. A point (x, y) in Figure 6 means that x percent of the individuals have at least y location signals in the month of Dec.

We use three subsets of this dataset containing 20,000 individuals in San Francisco, Manhattan and Cook (which contains Chicago) counties. Unless otherwise stated, we consider this 20,000 individuals as the entire population, based on which the ground-truth spread is calculated (our algorithms are given access to visits of only a subset of these individuals to perform their estimation). These areas were used since their different co-location patterns is expected to lead to differences in the disease spread with the same diffusion model. Specifically, for San Francisco, Manhattan and Cook counties, the average number of daily co-locations per individual (i.e. the count of the total number of co-locations in a day divided by the number of individuals) is 2.95, 1.72 and 0.23, respectively. We also use Gowalla [1], a publicly available dataset to allow for reproducability. Gowalla contains visits of users obtained from a social network over several months across the US. Since the data is very sparse, we select a 20 day period with the largest number of visits and 20,000 individuals. It contains 6, 760, 928 visits.

Parameter	Value
d_{max}	$\sim 11m$
t_{min}	15 min.
p_{inf}	0.01
p_{init}	0.01
μ_{IS}	5 days
μ_R	12 days
p_s	0.1

Table 2: Parameter Setting

Dataset	No. loc. signals
San Francisco	48,699,847
Manhattan	53,147812
Cook	130,236,901

Table 3: No. location signals for Dec.

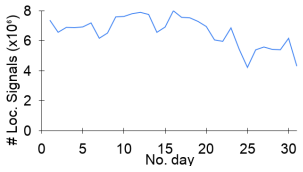

Figure 5: No. location signals per day for Manhattan

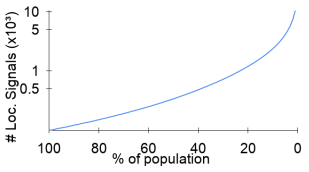

Figure 6: Distribution of location signals for Manhattan

Algorithms. We compare the performance of the algorithms discussed in Section 3.1. Scale is the algorithm discussed in Section 3.2, Density was discussed in Section 3.3.2, PollSpreader in Section 3.4.2 and PollSus_L and PollSus_U are the lower and upper bounds, respectively, from the method PollSusceptible discussed in Section 3.4.3. We use Density as a surrogate for approaches that generate synthetic trajectories, e.g. [12, 26], where they model the probability distribution of location sequences of the individuals. In contrast to [12, 26], Density allows for directly calculating the probability of co-locations, while [12, 26] rely on sampling to be able to find possible co-locations. For Density, we use a grid of 100×100 to discretize the space (results in width of each cell for Manhattan and San Francisco to be about 100m and for Cook county about 500m).

Metrics and Evaluation. For each dataset, we calculate the ground-truth using Monte-Carlo simulation. This is done by simulating the spread of the disease in the entire population for whom we have data, i.e., for Veraset dataset, the 20,000 individuals. We run the simulation 10 times and take the average. To evaluate each algorithm, we sample each individual independently and with probability p_s, for $p_s \in \{0.025, 0.05, 0.1, 0.2\}$ to create random sub-samples of the population. For each algorithm, we measure mean absolute error (MAE) in the estimation from the ground truth. This is done by, for each set of samples, measuring the difference between the ground

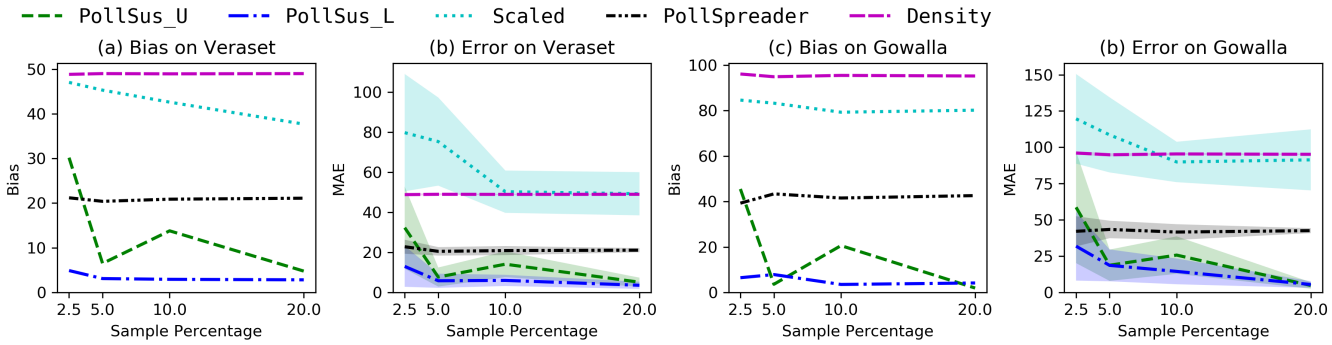

Figure 7: Error and Bias on San Francisco Veraset and Gowalla datasets

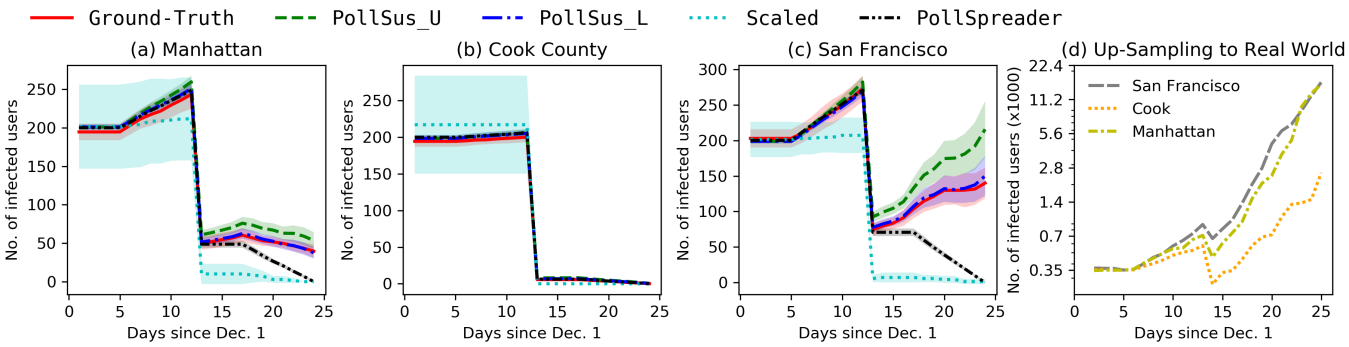

Figure 8: Infection Numbers in Different Counties

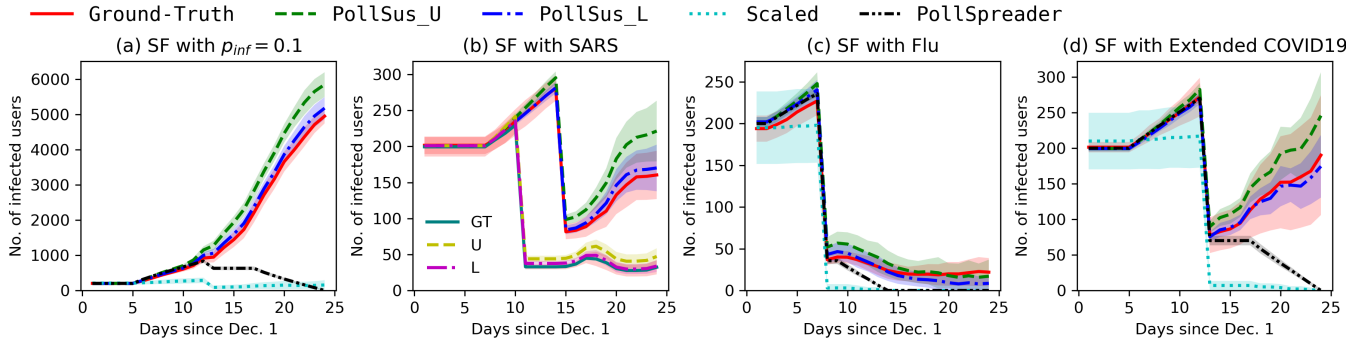

Figure 9: Other Diffusion Models

truth and the estimation. We also measure the *bias* in each estimation, calculated by taking the average of 10 runs for each algorithm and calculating its difference with the ground truth.

Parameter Setting. We use the diffusion model discussed in Section 3.1. Unless otherwise stated, the parameter setting is as shown in Table 2, referred to as COVID-19 diffusion model, designed to mimic the spread of COVID-19. We set μ_{IS} to 5 days similar to the mean incubation period reported in [23]. The work in [17] reports "Infectiousness was estimated to decline quickly within 7 days", so we set μ_R to $\mu_{IS} + 7$. We set t_{min} to 15 min as mentioned by [33].

Finally, we mention that our technical report [36] contains complimentary experimental results on the impact of sampling bias on the accuracy of our methods and a more comprehensive evaluation of our method on different diffusion models.

4.2 Results with COVID-19 Diffusion Model

4.2.1 Error and Bias. Figures 7 (a), (b) show the performance of the algorithms on the San Francisco dataset. First, we observe that PollSus_L provides very low error and bias, even when only 2.5% of the population is sampled, and thus can be used to estimate the spread accurately. PollSus_U performs well for higher sampling rates while Scale and Density perform poorly as they fail to adjust for infections that happen from unobserved individuals. Scale also has a higher variance compared with the other methods, which is a

result of the randomness of the Monte-Carlo simulation (the other algorithms are deterministic given a sub-sample).

Furthermore, Figures 7 (c), (d) show the results on the Gowalla dataset. For our experiments on Gowalla, we set d_{max} to about 110m and $p_{inf} = 0.1$, since otherwise we observed very few co-locations and infections. We observe similar trends as before with PollSus_L and PollSus_U performing the best. we remove Density from our experiments from here onwards due to its poor performance.

4.2.2 *Daily Infected Numbers across Counties*

Up-sampling with Known Ground-Truth. The goal of this experiment is to examine, for different spread patterns, how closely the estimations follow the ground truth. As before, we have assumed the 20,000 individuals to be the true population based on which the ground-truth is calculated. Figures 8 (a), (b) and (c) show the number of infected people per day for Manhattan, Cook and San Francisco counties respectively. The shaded area shows one standard deviation above and below the mean. The drop in the number of infections on day 13 in the figures is due to our choice of $\mu_R = 12$, since all 200 initial infections recover by then.

Overall, the spread of the disease is correlated with the average number of daily co-locations per individuals, where the spread of the disease ends after the initial infections in Cook county, but it lasts longer in Manhattan and in San Francisco. We observe the idea of *herd-immunity* in Figures 8 (b), where the virus stops spreading well before all the people in the population are infected. The pattern is different for Manhattan and San Francisco, with the number of cases in Manhattan remaining at a steady level, while increasing in San Francisco (this is interesting from an epidemiology perspective, because it shows that herd immunity is dependent on the co-location pattern for a population, given the same disease). We also see that PollSus_L and PollSus_U follow the ground-truth closely in all the datasets and throughout the studied period. PollSpreader follows ground-truth but mainly up-to day 10, as after that an adjustment for multi-hop infections is required which PollSpreader does not take into account. Scale accurately estimates the initial infections, as discussed in Section 3.2, but does not account for its false negatives, which gradually deteriorate its performance.

Here, we have assumed the 20,000 individuals to be the true population. However, the *real-world* population of each of the counties is larger (20,000 is about 0.39% of the real-world population of Cook county and about 2.27% of the real-world population of San Francisco). As a result the spread shown does not necessarily follow the real-world number of infections. Furthermore, since the real-world population of the counties differ, a relative comparison across the counties for the real-world number of infections is also not justified, as the up-sampling procedures would be different.

Up-Sampling to Real World. This experiment shows how our method can be applied in the real world. We consider our entire dataset, consisting of 20,000 users in each of the cities, as a sub-sample of each city's entire population. 20,000 is about 2.27%, 1.25% and 0.39% of the populations of San Francisco, Manhattan and Cook county, respectively, and these values are used as the sampling probability. For all the counties, we set p_{init} so that the expected number of initial infections is 350 people. Furthermore, we reduce $p_{inf} = 0.001$ and $d_{max} \sim 2m$ as otherwise number of infections would have been too large. For clarity, we only present the results of PollSus_L, the best algorithm in the previous experiments. Figure 8 (d) shows the estimate of the spread for the counties. The estimates show an increasing trend for all the counties, with number of infected in San Francisco and Manhattan being the highest.

Note that we do not have access to ground truth infection values to be able to evaluate the accuracy of the estimates (because our location dataset is from Dec. 2019, which is before the spread of COVID-19 in the considered counties). However, we can observe that the exponential growth trend matches the intuition and the real-world observations of the spread of COVID-19, with denser counties having a larger number of infections. Importantly, the patterns in Figure 8 (d) are different from Figure 8 (a), (b) and (c), which showed the spread of the disease fading for Manhattan and Cook counties. This emphasises the importance of correct estimation procedures, as the patterns observed in a sub-sample can be very different from the patterns in the true population.

4.3 Results with other Diffusion Models

4.3.1 Varying p_{inf}. We vary p_{inf} from 0.01 to 0.1 to see how it impacts the spread. p_{inf} determines the probability of the disease spreading from one person to another given a co-location and can be used to model different transmission scenarios, e.g., low p_{inf} values can be for the case when people are wearing a mask and larger value for when they aren't. Figure 9 (a) shows the impact of increasing p_{inf}. Comparing Figure 9 (a) and Figure 8 (c) shows how increasing the probability of transmission increases the spread of the disease. Overall, the general trends are the same as before, with PollSus_L and PollSus_U closely following Ground-Truth. Moreover, in Figure 9 (a), PollSpreader shows a herd immunity pattern, where the infections initially increase and then start decreasing after a certain time, while the true number of infections is increasing. This is also true for Scale, where the underestimation is amplified over time, i.e., underestimating the number of current infections leads to further underestimating the number of future infections.

4.3.2 *SARS and Flu Diffusion Models*

SARS. To simulate the spread of SARS, we increase μ_{IS} to 7 and μ_R to 14 to reflect how viral load peaks during the second week of infection [7, 27], as opposed to COVID-19, where the peak is earlier [5]. The result of this experiment is shown in Figure 8 (b) (lines labeled Ground-Truth, PollSus_U and PollSus_L).

Furthermore, we consider the observation in [7] that "Low rate of viral shedding in the first few days of illness meant that early isolation measures would probably be effective", to study a possible intervention scenario. Specifically, we consider the scenario that people are isolated early after experiencing symptoms, which can be modeled by reducing the value of μ_R. In the real world, this happens for SARS due to high rates of hospitalization [28]. We set μ_R to 10 days, considering that average duration of symptoms onset to hospital admission was 3.8 days [11], and that the incubation period can be up to 7 days [28]. Lines labelled GT, U and L in Figure 8 (b) show the results of the algorithms Ground-Truth, PollSus_U and PollSus_L, respectively, on this diffusion model. PollSus_L and PollSus_U follow Ground-Truth closely in both scenarios. Note that such an intervention policy may not be effective for COVID-19, because the transmission of COVID-19 often happens before individuals experience symptoms [17]. Furthermore, because of

higher rates of hospitalization for SARS, people tend to be in the hospital when they are the most infectious [7, 28]. These explain reaching herd immunity for SARS in Figure 4.3 (b), but not for COVID-19 in Figure 8 (c), which is consistent with the real world.

Flu. For the spread of Flu, we set μ_{IS} to 2 and μ_R to 7, based on [4], where viral shedding peaks at day 2 and lasts for at most 5 days. Furthermore, we set $p_{inf} = 0.005$, since the basic reproduction number of Flu is lower than that of COVID-19 [10] (in our diffusion model, given the mobility pattern, the basic reproduction number is directly correlated with p_{inf}). Figure 9 (c) show the results for the spread of Flu. Overall, the performance of the algorithms is the same as on COVID-19 and SARS. We observe that, without any interventions, Flu spreads more slowly than COVID-19 and SARS.

4.3.3 Extended COVID-19 Diffusion Model. We also experimented with extending the diffusion model to see how it impacts our estimates. Specifically, we incorporated the idea that the probability of infection during a contact increases with the duration of the contact, and decreases with the distance between individuals [3, 15, 18, 24]. So far we have assumed that an infection happens with probability p_{inf} if an individual is within distance at most d_{max} for the duration of at least t_{min} (see Section 3.1) of an IS individual. We extend this definition so that an infection happens with probability p_{inf} if an individual is within distance at most d^1_{max} for the duration of at least t^1_{min} *or* within distance at most d^2_{max} for the duration of at least t^2_{min} *or* within distance at most d^3_{max} for the duration of at least t^3_{min}. We set $d^1_{max} = 2m$, $d^2_{max} = 4m$ and $d^3_{max} = 8m$ and $t^1_{min} = 15min$, $t^2_{min} = 30min$ and $t^3_{min} = 60min$. Figure 9 (d) shows the result for this experiment. The trends are similar to before, but the spread of the disease is faster than that shown in Figure 8 (c), as the diffusion model accounts for more transmission scenarios.

5 RELATED WORK

Related work fall into three categories: (1) synthetic trajectory creation, (2) contact matrices and synthetic co-location creation and (3) agent-based simulation modeling the spread of a disease.

Synthetic Trajectory Generation. Various methods have been proposed to generate location trajectories of people [12, 19, 26, 34]. Recent work [12, 26] use GANs [35] to model the distribution of the trajectories, learned from the available samples. This method can be applied to our problem by sampling new trajectories from the learned model and simulating the spread in the new population (which contains synthetic users). The simulation of [12] follows this approach, but in a discretized space. A *contact factor* is introduced (set to a fixed value) to model the rate of co-locations between people within the same cell. This is similar to our baseline, Density (see Section 3.3.2), which discretizes the space (the contact factor in our case is $\frac{\pi d^2_{max}}{\text{cell area}}$), and such discretization is also done in [26].

Overall, there are multiple problems with these approaches. (1) Locations need to be modeled accurately to within a few meters and for long periods. This is a challenging task, and discretizing space, done to improve the quality of the location trajectory, makes the quality of the co-locations worse, as we observed in our experiments. (2) Mobility patterns change over time and in different cities, and the approaches need to learn different models for each of them. (3) They need large number of samples for training a neural network and they provide no theoretical guarantees on their estimation. (4) Synthetic trajectories increase the data size, which in turn requires larger computational power to estimate the spread. Our approach circumvents issue (1) by directly looking at co-locations, (2) can be directly applied to any time and location, (3) provides theoretical guarantees on its estimates and (4) performs estimation on the sub-sampled population which requires less computational power.

Synthetic Contact Generation. Contact matrices [25, 29, 30] are commonly used to simulate contact between individuals in a population. They provide aggregate level (e.g., for an entire country) contact information between different compartments in the population (e.g., rate of contacts between people of different ages) and they are estimated through surveys and diaries. However, they do not change with time and are not available for specific cities or areas, which limits their usefulness for studying the spread at a particular time and in a specific area. They furthermore do not take into account the differences in individual mobility patterns, and consider the population as multiple monoliths. Our work addresses the above issue by using location sequences of the individuals for specific periods of time and in a specific area. Finally, although more sophisticated synthetic graph generation methods, e.g., [37], exist, we are unaware of any that do this for *physical contacts*. That is, the generated graph should correspond to physical co-locations, and be accurate for multiple weeks. We also note that such an approach would still not solve issues (2) and (3) mentioned above.

Agent-Based Simulations. Agent-based simulations [6, 13, 14, 16, 21] exhibit a use-case of our methodology, where spread of a disease is studied under different diffusion models and for different mobility patterns. Such simulations currently rely on fixed contact matrices which limits their focus and accuracy and do not allow for study of the spread at a specific time and for a particular area. Our approach can be readily used to address such inaccuracies using real data.

6 CONCLUSION

We studied the problem of estimating the spread of a virus, or other phenomena that can be transmitted through physical contact, in a population by having access only to a subsample of the population. We observed that modeling co-locations of the individuals, and not their locations, allows for accurate estimations. To that end, we provided two methods, PollSpreader and PollSusceptible, that estimate properties of a contact network to calculate the spread in the original population. We theoretically showed that our estimates provide lower and upper bounds on the spread in the original population, and experimentally showed that they are close to the ground-truth in practice. Future work includes using our estimates to study different intervention strategies and studying the problem by relaxing independence and uniformity assumptions on sampling.

ACKNOWLEDGMENTS

This research has been funded in part by NSF grants IIS-1910950 and CNS-2027794, the USC Integrated Media Systems Center (IMSC), and unrestricted cash gifts from Microsoft. We would also like to acknowledge Veraset for providing us with high fidelity location signals. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors.

REFERENCES

[1] 2020. Gowalla Dataset. https://snap.stanford.edu/data/loc-gowalla.html. Accessed: 2020-11-14.

[2] 2020. Veraset Website. https://www.veraset.com/about-veraset. Accessed: 2020-10-25.

[3] Erin Bromage. 2020. The Risks-Know Them-Avoid Them. *Erin Bromage: COVID-19 Musings* (2020).

[4] Fabrice Carrat, Elisabeta Vergu, Neil M Ferguson, Magali Lemaitre, Simon Cauchemez, Steve Leach, and Alain-Jacques Valleron. 2008. Time lines of infection and disease in human influenza: a review of volunteer challenge studies. *American journal of epidemiology* 167, 7 (2008), 775–785.

[5] Muge Cevik, Matthew Tate, Ollie Lloyd, Alberto Enrico Maraolo, Jenna Schafers, and Antonia Ho. 2020. SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis. *The Lancet Microbe* (2020).

[6] Sheryl L Chang, Nathan Harding, Cameron Zachreson, Oliver M Cliff, and Mikhail Prokopenko. 2020. Modelling transmission and control of the COVID-19 pandemic in Australia. *Nature communications* 11, 1 (2020), 1–13.

[7] Peter KC Cheng, Derek A Wong, Louis KL Tong, Sin-Ming Ip, Angus CT Lo, Chi-Shan Lau, Eugene YH Yeung, and Wilina WL Lim. 2004. Viral shedding patterns of coronavirus in patients with probable severe acute respiratory syndrome. *The Lancet* 363, 9422 (2004), 1699–1700.

[8] Wen-Hao Chiang, Xueying Liu, and George Mohler. 2020. Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. *medRxiv* (2020). https://doi.org/10.1101/2020.06.06.20124149

[9] Steffen E Eikenberry, Marina Mancuso, Enahoro Iboi, Tin Phan, Keenan Eikenberry, Yang Kuang, Eric Kostelich, and Abba B Gumel. 2020. To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. *Infectious Disease Modelling* (2020).

[10] Joseph Eisenberg. 2020. R0: How Scientists Quantify the Intensity of an Outbreak Like Coronavirus and Its Pandemic Potential. https://sph.umich.edu/pursuit/2020posts/how-scientists-quantify-outbreaks.html. Accessed: 2021-02-17.

[11] Dan Feng, Na Jia, Li-Qun Fang, Jan Hendrik Richardus, Xiao-Na Han, Wu-Chun Cao, and Sake J De Vlas. 2009. Duration of symptom onset to hospital admission and admission to discharge or death in SARS in mainland China: a descriptive study. *Tropical Medicine & International Health* 14 (2009), 28–35.

[12] Jie Feng, Zeyu Yang, Fengli Xu, Haisu Yu, Mudan Wang, and Yong Li. 2020. Learning to Simulate Human Mobility. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3426–3433.

[13] Neil Ferguson, Daniel Laydon, Gemma Nedjati Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, ZULMA Cucunuba Perez, Gina Cuomo-Dannenburg, et al. 2020. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. (2020).

[14] Neil M Ferguson, Derek AT Cummings, Christophe Fraser, James C Cajka, Philip C Cooley, and Donald S Burke. 2006. Strategies for mitigating an influenza pandemic. *Nature* 442, 7101 (2006), 448–452.

[15] Caroline X Gao, Yuguo Li, Jianjian Wei, Sue Cotton, Matthew Hamilton, Lei Wang, and Benjamin J Cowling. 2021. Multi-route respiratory infection: when a transmission route may dominate. *Science of the Total Environment* 752 (2021), 141856.

[16] M Elizabeth Halloran, Neil M Ferguson, Stephen Eubank, Ira M Longini, Derek AT Cummings, Bryan Lewis, Shufu Xu, Christophe Fraser, Anil Vullikanti, Timothy C Germann, et al. 2008. Modeling targeted layered containment of an influenza pandemic in the United States. *Proceedings of the National Academy of Sciences* 105, 12 (2008), 4639–4644.

[17] Xi He, Eric HY Lau, Peng Wu, Xilong Deng, Jian Wang, Xinxin Hao, Yiu Chung Lau, Jessica Y Wong, Yujuan Guan, Xinghua Tan, et al. 2020. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature medicine* 26, 5 (2020), 672–675.

[18] Christopher I Jarvis, Kevin Van Zandvoort, Amy Gimma, Kiesha Prem, Petra Klepac, G James Rubin, and W John Edmunds. 2020. Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK. *BMC medicine* 18 (2020), 1–10.

[19] Shan Jiang, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C González. 2016. The TimeGeo modeling framework for urban mobility without travel surveys. *Proceedings of the National Academy of Sciences* 113, 37 (2016), E5370–E5378.

[20] William Ogilvy Kermack and Anderson G McKendrick. 1927. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115, 772 (1927), 700–721.

[21] Cliff C Kerr, Robyn M Stuart, Dina Mistry, Romesh G Abeysuriya, Gregory Hart, Katherine Rosenfeld, Prashanth Selvaraj, Rafael C Nunez, Brittany Hagedorn, Lauren George, et al. 2020. Covasim: an agent-based model of COVID-19 dynamics and interventions. *medRxiv* (2020).

[22] Mehrdad Kiamari, Gowri Ramachandran, Quynh Nguyen, Eva Pereira, Jeanne Holm, and Bhaskar Krishnamachari. 2020. COVID-19 Risk Estimation using a Time-varying SIR-model. *1st ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19* (2020).

[23] Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy SM Leung, Eric HY Lau, Jessica Y Wong, et al. 2020. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine* (2020).

[24] Atsushi Mizukoshi, Chikako Nakama, Jiro Okumura, and Kenichi Azuma. 2021. Assessing the risk of COVID-19 from multiple pathways of exposure to SARS-CoV-2: Modeling in health-care settings and effectiveness of nonpharmaceutical interventions. *Environment international* 147 (2021), 106338.

[25] Joël Mossong, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo Scalia Tomba, Jacco Wallinga, et al. 2008. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med* 5, 3 (2008), e74.

[26] Kun Ouyang, Reza Shokri, David S Rosenblum, and Wenzhuo Yang. 2018. A Non-Parametric Generative Model for Human Trajectories.. In *IJCAI*. 3812–3817.

[27] Joseph Sriyal Malik Peiris, Chung-Ming Chu, Vincent Chi-Chung Cheng, KS Chan, IFN Hung, Leo LM Poon, Kin-Ip Law, BSF Tang, TYW Hon, CS Chan, et al. 2003. Clinical progression and viral load in a community outbreak of coronavirus-associated SARS pneumonia: a prospective study. *The Lancet* 361, 9371 (2003), 1767–1772.

[28] Eskild Petersen, Marion Koopmans, Unyeong Go, Davidson H Hamer, Nicola Petrosillo, Francesco Castelli, Merete Storgaard, Sulien Al Khalili, and Lone Simonsen. 2020. Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. *The Lancet infectious diseases* (2020).

[29] Kiesha Prem, Alex R Cook, and Mark Jit. 2017. Projecting social contact matrices in 152 countries using contact surveys and demographic data. *PLoS computational biology* 13, 9 (2017), e1005697.

[30] Kiesha Prem, Kevin van Zandvoort, Petra Klepac, Rosalind M Eggo, Nicholas G Davies, Alex R Cook, Mark Jit, et al. 2020. Projecting contact matrices in 177 geographical regions: an update and comparison with empirical data for the COVID-19 era. *medRxiv* (2020).

[31] Sirisha Rambhatla, Sepanta Zeighami, Kameron Shahabi, Cyrus Shahabi, and Yan Liu. 2020. Towards Accurate Spatiotemporal COVID-19 Risk Scores using High Resolution Real-World Mobility Data. *arXiv preprint arXiv:2012.07283* (2020).

[32] Houtan Shirani-Mehr, Farnoush Banaei Kashani, and Cyrus Shahabi. 2012. Efficient Reachability Query Evaluation in Large Spatiotemporal Contact Datasets. *Proc. VLDB Endow.* 5, 9 (2012), 848–859. https://doi.org/10.14778/2311906.2311912

[33] W Joost Wiersinga, Andrew Rhodes, Allen C Cheng, Sharon J Peacock, and Hallie C Prescott. 2020. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): a review. *Jama* 324, 8 (2020), 782–793.

[34] Mogeng Yin, Madeleine Sheehan, Sidney Feygin, Jean-François Paiement, and Alexei Pozdnoukhov. 2017. A generative model of urban activities from cellular data. *IEEE Transactions on Intelligent Transportation Systems* 19, 6 (2017), 1682–1696.

[35] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Thirty-first AAAI conference on artificial intelligence*.

[36] Sepanta Zeighami, Cyrus Shahabi, and John Krumm. 2020. Estimating Spread of Contact-Based Contagions in a Population Through Sub-Sampling. *https://infolab.usc.edu/DocsDemos/SpreadEstimation.pdf* (2020).

[37] Dawei Zhou, Lecheng Zheng, Jiawei Han, and Jingrui He. 2020. A Data-Driven Graph Generative Model for Temporal Interaction Networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 401–411.