CRISP: A Probabilistic Model for Individual-Level COVID-19 Infection Risk Estimation Based on Contact Data

Ralf Herbrich[*]
Zalando
Berlin, Germany
ralf@zalando.de

Rajeev Rastogi
Amazon
Bangalore, India
rastogi@amazon.com

Roland Vollgraf
Zalando
Berlin, Germany
roland.vollgraf@zalando.de

Abstract

We present CRISP (**C**OVID-19 **RI**sk **S**core **P**rediction), a probabilistic graphical model for COVID-19 infection spread through a population based on the SEIR model where we assume access to (1) mutual contacts between pairs of individuals across time across various channels (e.g., Bluetooth contact traces), as well as (2) test outcomes at given times for infection, exposure and immunity tests. Our micro-level model keeps track of the infection state for each individual at every point in time, ranging from susceptible, exposed, infectious to recovered. We develop a Monte Carlo EM algorithm to infer contact-channel specific infection transmission probabilities. Our algorithm uses Gibbs sampling to draw samples of the latent infection status of each individual over the entire time period of analysis, given the latent infection status of all contacts and test outcome data. Experimental results with simulated data demonstrate our CRISP model can be parametrized by the reproduction factor R_0 and exhibits population-level infectiousness and recovery time series similar to those of the classical SEIR model. However, due to the individual contact data, this model allows fine grained control and inference for a wide range of COVID-19 mitigation and suppression policy measures. Moreover, the algorithm is able to support efficient testing in a test-trace-isolate approach to contain COVID-19 infection spread. To the best of our knowledge, this is the first model with efficient inference for COVID-19 infection spread based on individual-level contact data; most epidemic models are macro-level models that reason over entire populations. The implementation of CRISP is available in Python and C++ at https://github.com/zalandoresearch/CRISP.

1 Introduction

The COVID-19 pandemic has spread rapidly around the world, with the number of infections and deaths steadily growing. Most governments around the world have been completely unprepared to deal with the COVID-19 outbreak, which UN Secretary-General Antonio Guterres has referred to as humanity's worst crisis since World War II. While governments around the world had plans in place in the event of a pandemic, the peculiarities of COVID-19 (e.g., delayed onset of symptoms, asymptomatic transmission) have challenged these preparations. Governments have reacted by implementing measures such as nationwide lock-downs, that require people to stay inside their homes, enforcing social distancing and therefore breaking the COVID-19 infection chain. However, a blunt mechanism such as a lock-down (over an extended period) can cause severe damage to the economy, and so, there is a need to find alternative measures to slow down or stop the spread without incremental effects in other areas of society. These alternatives have to be built in a solid foundation

[*] The ordering of authors is alphabetical. All authors contributed equally to the paper.

Preprint. Under review.

such as widespread testing and the isolation of infected (or potentially infected) individuals via contact-tracing.

Contact tracing technologies [20, 19] have shown promise in tracking the spread of the disease across the population. These mobile apps capture social contact information between users such as contact duration, distance, etc. using Bluetooth signals on devices. The fine-grained contact data of individuals collected by the apps can enable:

- *Individual risk score prediction.* The contact data, combined with information about COVID-19 positive test cases, can be used to predict the likelihood of infection for each individual. The individual risk scores can be leveraged by governments and organizations to prioritize testing as well as to identify individuals that need to enter isolation/quarantine.
- *Hotspot detection.* Tracing technologies can help authorities identify areas with a high density of contacts and/or individuals with high infection risk. This can allow policymakers to make more effective decisions, for example, by imposing highly restrictive measures such as lock-downs, shelter-at-home, or school closures only in COVID-19 hotspots while allowing activities to remain closer to normal in unaffected areas.
- *Insights about infection spread.* Contact tracing can provide insights into the relative importance of different modalities of disease transmission (e.g., through intermediate surfaces vs individual contact), risk of infection transmission based on contact characteristics such as duration and distance, most likely locations (e.g., schools, work, malls) for the spread of disease, and "super spreaders" who come in close proximity with a large number of individuals and so must be frequently tested for infection.

To achieve the above-mentioned benefits, we need to devise new models and inference algorithms for analyzing contact tracing data. This is because existing epidemics models [3, 15, 14, 6, 5] focus on estimating population-level statistics such as percentage of the population infected, number of days for the epidemic to peak, etc. as opposed to the infection state of each individual in the population. Other models [16, 17] that use ML-based inference techniques assume complete knowledge of the infection state of each individual at each time instant. However, in the COVID-19 scenario, (1) the infection status of individuals is not known until they are tested, and (2) infectious time of individuals are unknown since individuals may infect others while asymptomatic. Finally, governments are using contact tracing data [20, 19] to identify and test individuals who have come in direct contact with COVID-19 positive test cases. However, the fact that asymptomatic individuals may have infected a large number of people prior to displaying symptoms and being tested, delays the detection of these newly infected individuals by only using contact tracing apps.

Our main contributions can be summarized as follows:

- We propose CRISP (**C**OVID-19 **RI**sk **S**core **P**rediction), a probabilistic graphical model for COVID-19 infection spread through diverse contacts channels between individuals. Our model uses latent variables to represent the epidemiological states of individuals based on the SEIR model [12] at different points in time, and captures both the transitions between states as well as test outcomes.
- We develop a Monte Carlo EM algorithm to infer infection transmission probabilities across a range of contact channels. Our algorithm uses block-Gibbs sampling to draw samples of the latent infection status of each individual over the entire time period, given data about contacts and test results.
- We provide implementation details to accelerate both the block-Gibbs sampling and the forward sampling algorithm. A Python and C++ implementation of CRISP is available at https://github.com/zalandoresearch/CRISP.
- We conduct experiments with simulated data which demonstrate that our CRISP model can be parametrized by the reproduction factor R_0 and exhibits population-level infectiousness and recovery time series similar to those of the classical SEIR model. However, due to the individual contact data, this model allows fine grained control and inference for a wide range of COVID-19 mitigation and suppression policy measures. Furthermore, we show that a testing-and-quarantining policy based on infection risk scores computed by the CRISP algorithm is able to mitigate COVID-19 infection spread while quarantining fewer individuals compared to other policies based on contact-tracing and symptom-based testing.

To the best of our knowledge, this is the first comprehensive model for COVID-19 infection spread that (1) captures the infection states of individuals and transitions between them using the SEIR model, and (2) leverages contact tracing and test outcome data to infer model parameters such as contact-channel specific infection rates using scalable and computationally efficient inference algorithms.

2 Related Work

We classify related work into four broad categories: epidemic models, Machine Learning (ML) based inference of model parameters, influence maximization in social networks, and contact tracing apps.

2.1 Epidemic Models

In recent years, there has been research on modeling individual dynamics of epidemics [3, 15, 14]. However, this work typically resorts to mean-field theory to model virus spread over a network, and thus does not characterize the dynamic infectious state of each individual over time.

Ferguson et al. [6, 5] use a compartmental transmission model to simulate the spread of influenza across a population, and analyze interventions such as antiviral prophylaxis and social distancing to halt a pandemic. The authors use a stochastic model of individuals co-located in households that are randomly distributed across a geographical region, and infection risk from 3 sources – household, place and random contacts in the community. The infection transmission rates for the 3 sources and recovery rates are based on analysis of historical data. In contrast, we leverage real individual contact tracing data and outcomes of tests on individuals to infer the infection transmission rate for each contact and the likelihood of infection for each individual.

Lorch et al. [10] propose a spatiotemporal epidemic model that uses marked temporal processes to represent the epidemiological condition of each individual (based on a variation of the SEIR compartment models), individual mobility patterns, test outcomes, and testing and contact tracing strategies. The authors design an efficient sampling algorithm for the model using Monte Carlo roll-outs that is able to predict the spread of COVID-19 under different testing & tracing strategies, social distancing measures, and business restrictions, given contact histories of individuals. They use Bayesian optimization techniques to infer model parameters (e.g. infection transmission rate) that minimize the difference between the real positive COVID-19 cases and those in the Monte-Carlo simulations. In addition, they demonstrate the efficacy of their model using real COVID-19 data and mobility patterns of Tübingen, Germany. Our Monte Carlo EM inference algorithm for model parameters is computationally much more efficient than the Bayesian optimization techniques employed in [10].

2.2 Machine Learning-based Inference

In [16], the authors consider the problem of inferring latent social networks based on network diffusion or disease propagation data. Given the times when nodes become infected, but not who infected them, the authors identify the optimal network that best explains the observed data. The authors present a maximum likelihood approach based on convex optimization with a L_1-like penalty term (that encourages sparsity) to estimate the conditional probability of infection transmission between every node pair. A key difference from our work is that [16] assumes complete knowledge of infected nodes and infection times. In contrast, in the COVID-19 scenario, (1) the infection status of nodes is not known until they are tested, and (2) infection times of nodes are unknown since nodes may not show symptoms even though they are infected (and infecting others).

Warriyar et al. [17] introduce a novel R statistical software package EpiILM for simulating infectious disease spread, and carrying out Bayesian MCMC-based statistical inference for spatial and/or (contact) network-based models in the Deardon et al. [4] individual-level modelling framework. In individual-level models (ILMs), the epidemiological state of each individual (e.g., susceptible or infected) is assumed to be perfectly known at each time instant, which makes it relatively straightforward to estimate model parameters such as infection transmission probabilities (as a function of covariates) using maximum likelihood estimation or Bayesian inference using Metropolis-Hastings MCMC. However, in the COVID-19 scenario, epidemiological states of individuals are hidden until they are tested, and this complicates Bayesian inference in our probabilistic model setting.

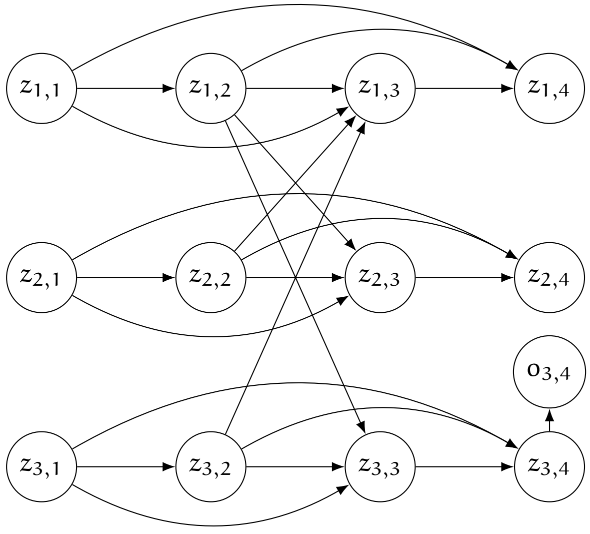

Figure 1: Graphical model of the CRISP contact infection spread model for 3 people over 4 time steps where individual $u = 1$ meets both individual $u = 2$ and $u = 3$ at time $t = 2$ and one test outcome of individual $u = 3$ at time $t = 4$. Note that this model has no cycles as we assume the infection status $z_{u,t}$ only depends on variables $z_{v,t'}$ *before* time step t, $t' < t$. However, due to the "memory" that the state $z_{u,t} = E$ and $z_{u,t} = I$ have, we require edges into the entire past of an infection trace.

2.3 Influence Maximization in Social Networks

The *Influence Maximization* problem aims to select k users in a social network that maximize influence spread, and was first modeled as an algorithmic problem by Kempe et al. [8]. [9] presents a comprehensive survey of different diffusion models that capture the information diffusion process and approximation algorithms to maximize influence. The papers assume that diffusion model parameters such as influence probabilities are given and focus on selecting k users to maximize influence spread. In contrast, our paper focuses on the problem of estimating model parameters related to infection transmission probabilities for each contact, given social contact information between users and COVID-19 test results for users.

[11] addresses the problem of finding the "backbone" of an influence network. It employs network sparsification to preserve only the links that play an important role in the propagation of information. [7] considers the problem of estimating influence probabilities between users in a social graph. Given a social graph and a log of actions by users, the Maximum Likelihood Estimator (MLE) of influence probability of node u on node v is simply the fraction of actions performed by u that are also performed by v. Unlike [7], in our setting, the infection status and times of nodes are latent, and need to be inferred by our algorithms.

2.4 Contact Tracing Apps

To combat the spread of COVID-19, governments have launched contact tracing apps[20, 19] that use Bluetooth signals on mobile phones to track contacts between users. Users who have come in direct contact with COVID-19 positive test cases are considered to be at high risk of infection, and subject to tests and quarantine actions. However, a key problem with this approach is that COVID-19 infected users are typically tested only after they show symptoms, and typically, infected users show symptoms 5-6 days post infection. These asymptomatic users may have infected a large number of users over multiple hops prior to displaying symptoms and being tested. This delays detection of infected users using contact tracing apps, and limits their effectiveness to proactively test and isolate infected users to contain the spread of COVID-19. In contrast, our probabilistic modeling algorithm CRISP predicts the likelihood of a user getting infected with COVID-19 through a chain of social contacts involving asymptomatic users, and identifies infected users early, even though they may be multiple hops from a user who has tested positive for COVID-19 and even before they begin to show symptoms. Our inference algorithm also learns infection transmission probabilities for each contact channel.

3 CRISP Infection Spread Model

Our CRISP model is an SEIR model at the level of every individual (see [12] for an introduction). Note that we consider discrete time steps t implicitly assumed to be at the level of single days. We assume that we are given the following two datasets for a given set $\mathcal{S}$ of individuals:

- $\mathcal{D}_{\text{contact}} = \{(u_i, v_i, t_i, \mathbf{x}_i)\}_{i=1}^{N} \subseteq \mathcal{S} \times \mathcal{S} \times \mathbb{N} \times \mathbb{N}^J$ of N quadruples of a pair of two individuals (u_i, v_i) who have met at time t_i with specific features $\mathbf{x}_i$. Here we assume that the feature vector $\mathbf{x}_i$ describes the overall contact between u_i and v_i via the number $x_{i,j}$ of mutual contacts over channel j (e.g., Bluetooth encounters, queuing together, sharing public transportation). We assume $\mathcal{D}_{\text{contact}}$ to be symmetric so that $(u, v, t, \mathbf{x}) \in \mathcal{D}_{\text{contact}} \leftrightarrow (v, u, t, \mathbf{x}) \in \mathcal{D}_{\text{contact}}$.
- $\mathcal{D}_{\text{test}} := \{(u_i, t_i, o_i)\}_{i=1}^{K} \subseteq \mathcal{S} \times \mathbb{N} \times \{0, 1\}$ of K triplets of individual u_i taking a test at time t_i with the test outcome o_i where $o_i = 0$ indicates a negative test outcome.

We model the T discrete time steps of infection status for each individual. Our model has $|\mathcal{S}| \times T$ many latent variables $\mathcal{Z} := \{z_{u,t}\}_{u \in \mathcal{S}, t=1,\ldots,T} \in \{S, E, I, R\}^{|\mathcal{S}| \times T}$ that represent the four stages of infection[2]:

- $z_{u,t} = S$: individual u has not been infected and is susceptible,
- $z_{u,t} = E$: individual u is infected but not contagious,
- $z_{u,t} = I$: individual u is infected and is contagious,
- $z_{u,t} = R$: individual u has recovered and is not contagious.

Let us use the notation $\mathcal{Z}_{u,t} := \{z_{u,1}, \ldots, z_{u,t}\}$ to denote the set of latent states $z_{u,t}$ of individual u up to and including time t and $\mathcal{Z}_t := \bigcup_u \mathcal{Z}_{u,t}$. In addition to the latent infection status of all individuals at each time step, we also model K variables $o_{u,t} \in \{0, 1\}$ for the test outcomes in $\mathcal{D}_{\text{test}}$, that is, $\mathcal{O} := \{o_{u,t} : (u, t, o_{u,t}) \in \mathcal{D}_{\text{test}}\}$. Then, our graphical model $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ between the variables $\mathcal{V} := \mathcal{Z} \bigcup \mathcal{O}$ has the following edges:

1. $\mathcal{E}_{\text{time}} = \bigcup_u \mathcal{E}_u$ and $\mathcal{E}_u := \{(z_{u,t}, z_{u,t'})_{t<t'}\}$. All edges between the latent infection states of a single individual u. The edges $\mathcal{E}_u$ will be used to describe the probability of $P(z_{u,t}|z_{u,t-1}, \ldots, z_{u,1})$ and describe the full time series of being susceptible, exposed, infectious and then recovered.
2. $\mathcal{E}_{\text{contact}} := \bigcup_{(u,v,t,\mathbf{x}) \in \mathcal{D}_{\text{contact}}} \{(z_{u,t}, z_{v,t+1})\}$. All edges between two individuals u and v who had a contact at time t.
3. $\mathcal{E}_{\text{test}} := \bigcup_{(u,t,o) \in \mathcal{D}_{\text{test}}} \{(z_{u,t}, o_{u,t})\}$. All edges between a test outcome at time t and the corresponding infection status $z_{u,t}$ of individual u. These edges will be used to describe the probabilities $P(o_{u,t}|z_{u,t})$ of a test outcome given the infection status of u at that same time.

The full edge set $\mathcal{E}$ is the union of these three edge types: $\mathcal{E} = \mathcal{E}_{\text{time}} \cup \mathcal{E}_{\text{contact}} \cup \mathcal{E}_{\text{test}}$. Figure 3 shows an example graphical model with these three edge components.

In order to define the joint probability distribution, note that all edges $\mathcal{E}_{\text{time}}$ and $\mathcal{E}_{\text{contact}}$ are pointing forward in time. Thus, all $\{z_{u,t+1}\}_{u \in \mathcal{S}}$ are conditionally independent of each other given all the past states $\mathcal{Z}_t$. Also, all edges $\mathcal{E}_{\text{test}}$ have the property that a test outcome of individual u at time t only depends on the infection status of u at t, $z_{u,t}$. The joint probability distribution is given by

$$
\begin{aligned}
P\left(\mathcal{Z}_T, \mathcal{O}\right) &= P\left(O|\mathcal{Z}_T\right) \cdot P\left(\mathcal{Z}_T\right), &(1)\\
P\left(\mathcal{O}|\mathcal{Z}_T\right) &= \prod_{u} \prod_{(t,o) \in \mathcal{T}_u} P(o|z_{u,t}), &(2)\\
P\left(\mathcal{Z}_T\right) &= \prod_{t} \prod_{u} P(z_{u,t+1}|\mathcal{Z}_t), &(3)
\end{aligned}
$$

where $\mathcal{T}_u = \{(t, o) : (u, t, o) \in \mathcal{D}_{\text{test}}\}$. Since we are using an SEIR model, the only non-zero probabilities $P(z_{u,t+1}|\mathcal{Z}_t)$ are the transitions $S \to S$, $S \to E$, $E \to E$, $E \to I$, $I \to I$, $I \to R$, $R \to R$.

[2] Note that we are assuming that recovered individuals are immune until time step T.

$$P(z_{u,t+1}|\mathcal{Z}_t) = \begin{cases} f(u,t,\mathcal{Z}_t) & \text{if } z_{u,t} = S \wedge z_{u,t+1} = S \\ 1 - f(u,t,\mathcal{Z}_t) & \text{if } z_{u,t} = S \wedge z_{u,t+1} = E \\ 1 - g(u,t,\mathcal{Z}_{u,t}) & \text{if } z_{u,t} = E \wedge z_{u,t+1} = E \\ g(u,t,\mathcal{Z}_{u,t}) & \text{if } z_{u,t} = E \wedge z_{u,t+1} = I \\ 1 - h(u,t,\mathcal{Z}_{u,t}) & \text{if } z_{u,t} = I \wedge z_{u,t+1} = I \\ h(u,t,\mathcal{Z}_{u,t}) & \text{if } z_{u,t} = I \wedge z_{u,t+1} = R \\ 1 & \text{if } z_{u,t} = R \wedge z_{u,t+1} = R \\ 0 & \text{otherwise} \end{cases} . \tag{4}$$

Infection Model In order to define f, we assume that an infection occurs from exogenous influences with a fixed probability $p_0 \in [0,1]$ or with probability of $p_j \in [0,1]$ for every instance of a contact through the contact channel j if the contact was already in the state I. Thus, the probability that no infection occurred at time t equals

$$f(u,t,\mathcal{Z}_t) = (1-p_0) \cdot \prod_{(v,u,t,\mathbf{x}) \in \mathcal{D}_{\text{contact}} : z_{v,t}=I} \prod_{j=1}^{J} (1-p_j)^{x_j} . \tag{5}$$

Infection Status Model In order to define g and h, let us assume we have a point density function $q_E : \mathbb{N}^+ \mapsto [0,1]$ and $q_I : \mathbb{N}^+ \mapsto [0,1]$ for the probability $q_E(d_E)$ that the exposure ($z_{u,t} = E$) lasts for d_E time steps (and similarly for the duration of the infectiousness). Examples of functions q_E and q_I are the probability mass functions of the binomial, negative-binominal or geometric distributions. However, for the case of COVID-19, we will use discrete probabilities established from analysis of the population in [1] and [18]. Moreover, let

$$\pi(n;q) = \frac{q(n)}{1 - \sum_{i=1}^{n-1} q(i)} = \frac{P(d=n)}{P(d \geq n)} = P(d = n | d \geq n) , \tag{6}$$

be the conditional probability (according to q) that the duration is exactly n time steps given that the duration is at least n time steps. Then,

$$g(u,t,\mathcal{Z}_{u,t}) = \pi\left(t - \max_{t' \leq t}\{t' : z_{u,t'} = S\}; q_E\right) \tag{7}$$

$$h(u,t,\mathcal{Z}_{u,t}) = \pi\left(t - \max_{t' \leq t}\{t' : z_{u,t'} = E\}; q_I\right) \tag{8}$$

Note that the first argument to both g and h is the number of E and I states up to and including time t in the state sequence $\mathcal{Z}_{u,t}$.

Test Outcome Model Finally, we need to define the probability of a test outcome o given the infection status $z_{u,t}$ of individual u at time t. Since there are two types of mistakes of a test, we use

$$P(o|z_{u,t}) = \begin{cases} \alpha & \text{if } z_{u,t} = I \wedge o = 0 \\ 1-\alpha & \text{if } z_{u,t} = I \wedge o = 1 \\ 1-\beta & \text{if } z_{u,t} \in \{S,E,R\} \wedge o = 0 \\ \beta & \text{if } z_{u,t} \in \{S,E,R\} \wedge o = 1 \end{cases} . \tag{9}$$

We assume both $0 < \alpha \ll 1$ and $0 < \beta \ll 1$. It is easy to implement more sophisticated test accuracy models here, in particular to distinguish between different infection states. Also, we can easily model α and β which are dependent on how many days an individual has been in state I; this change would not affect the block-Gibbs sampling scheme in Section 4 in an adverse way.

Prior Model In order to complete the description of the full probabilistic model, we have to specify $P(\mathcal{Z}_1)$. For simplicity, we assume these probabilities to be a delta-peak at state S, that is, $P(z_{u,1} = S) = 1$ for all $u \in \mathcal{S}$.

4 Inference in the CRISP Model

For inference in the aforementioned model we are interested in computing the infection risk score of every individual u at every time step t given the test outcomes $\mathcal{O}$ available as well as the hyper-parameters $\theta := (p_0, p_1, \dots, p_J)$ which cannot be set by knowledge of the diseases: the J parameters p_j represent the probabilities of COVID-19 infection transmission through the contact channel j and p_0 captures the probability that an infection occurs at any time-step from exogenous influences.

In order to estimate θ, we will maximize the log-likelihood of the data $\mathcal{O}$, that is

$$\begin{aligned} \theta^* &= \text{argmax}_\theta \log\left(P(\mathcal{O}|\theta)\right) && (10)\\ &= \text{argmax}_\theta \log\left(\sum_{\mathcal{Z}_T} P(\mathcal{Z}_T|\mathcal{O},\theta)\cdot P(\mathcal{O}|\theta)\right) && (11)\\ &= \text{argmax}_\theta \log\left(\sum_{\mathcal{Z}_T} P(\mathcal{Z}_T,\mathcal{O}|\theta)\right), && (12)\end{aligned}$$

where the second decomposition explicitly contains the posterior $P\left(\mathcal{Z}_T|\mathcal{O}\right)$. However, this posterior is not analytically tractable and therefore we will approximate it by performing block-Gibbs sampling of an infection trace $\mathbf{z}_u := (z_{u,1}, \dots, z_{u,T})$ of individual u keeping all other infection traces $\{\mathbf{z}_{v:v\neq u}\}$ fixed. This requires a computationally efficient procedure to sample from the conditional probability distribution $P\left(\mathbf{z}_u|\{\mathbf{z}_{v:v\neq u}\},\mathcal{O},\theta\right)$ which we describe in the following subsection.

4.1 Infection Risk Score Inference

Since we assume that the total number of days of the model, T, is not large[3], we will enumerate all possible sequences of infection traces $\mathbf{z}_u$ and compute the un-normalized probability of $P\left(\mathbf{z}_u|\{\mathbf{z}_{v:v\neq u}\},\mathcal{O},\theta\right)$ for all terms that depend on elements of the trace $\mathbf{z}_u$ in order to re-normalize and draw from this distribution. Also, as our model is an SEIR model, we know that each sample $\mathbf{z}_u$ can be uniquely represented by a triple $\omega = (t_0, d_E, d_I) \in \mathbb{N}\times\mathbb{N}^+\times\mathbb{N}^+$ of time steps with t_0 being time steps individual u is in state S, d_E being time steps in state E, d_I being time steps in state I and the remaining $T - t_0 - d_E - d_I$ being time steps in state R.

There are three groups of factors that (might) involve $\mathbf{z}_u$ in the (un-normalized) conditional probability distribution $P\left(\mathbf{z}_u|\{\mathbf{z}_{v:v\neq u}\},\mathcal{O},\theta\right)$:

$$\underbrace{\prod_{t=1}^{T-1} P(z_{u,t+1}|\mathcal{Z}_t)}_{A(\mathbf{z}_u)} \cdot \prod_{v\neq u}\underbrace{\prod_{t=1}^{T-1} P(z_{v,t+1}|\mathcal{Z}_t)\cdot}_{B(\mathbf{z}_u)} \underbrace{\prod_{(t,o)\in\mathcal{T}_u} P(o|z_{u,t})}_{C(\mathbf{z}_u)} . \qquad (13)$$

The first set of factors, $A(\mathbf{z}_u)$, captures the temporal evolution of the infection state changes of $\mathbf{z}_u$ directly and can be reduced to three factors based on ω and all the contacts v that could have infected individual u. The second set of factors, $B(\mathbf{z}_u)$, captures the factors where the infectiousness of u might impact other individuals v. Finally, the third set of factors, $C(\mathbf{z}_u)$, captures the outcome of tests on individual u.

Factors $A(\mathbf{z}_u)$ In order to derive a compact representation of $A(\mathbf{z}_u)$, we assume that it can be written in terms of

$$A(\mathbf{z}_u) = l_0(t_0)\cdot l_E(d_E)\cdot l_I(d_I)\cdot l_{\text{infected}} \qquad (14)$$

[3] As of today, the COVID-19 pandemic is active for 90 days.

Since the infectious status, $z_{v,t} = I$ of other individuals v that had contact with u only affect u in the susceptible state, we can derive $l_0(t_0)$ and l_{infected} from (4) by collecting the f terms (see (5)):

$$\prod_{t=1}^{t_0-1} f(u, t, \mathcal{Z}_t) \cdot (1 - f(u, t_0, \mathcal{Z}_{t_0})) \tag{15}$$

$$= \left(\prod_{t=1}^{t_0-1} p_{u,t}\right) \cdot (1 - p_0)^{t_0-1} \cdot (1 - (1 - p_0)p_{u,t_0}) \tag{16}$$

$$= \underbrace{\left(\prod_{t=1}^{t_0-1} p_{u,t}\right) \cdot \left(\frac{1 - (1 - p_0)p_{u,t_0}}{p_0}\right)}_{l_{\text{infected}}} \cdot \underbrace{(1 - p_0)^{t_0-1} p_0}_{l_0(t_0)}, \tag{17}$$

where $p_{u,t} := \prod_{(v,u,t,\mathbf{x}) \in \mathcal{D}_{\text{contact}}: z_{v,t}=I} \prod_j (1 - p_j)^{x_j}$. Note that $l_0(t_0)$ is the density function of the geometric distribution. Similarly, given (4) and (7) we can derive $l_E(d_E)$ as

$$l_E(d_E) = \prod_{d=1}^{d_E-1} (1 - g(u, t_0 + d, \mathcal{Z}_{u,t_0+d})) \cdot g(u, t_0 + d_E, \mathcal{Z}_{u,t_0+d_E}))$$

$$= \prod_{d=1}^{d_E-1} \left(1 - \frac{q_E(d)}{1 - \sum_{i=1}^{d-1} q_E(i)}\right) \cdot \frac{q_E(d_E)}{1 - \sum_{i=1}^{d_E-1} q_E(i)} \tag{18}$$

$$= \prod_{d=1}^{d_E-1} \left(\frac{1 - \sum_{i=1}^{d} q_E(i)}{1 - \sum_{i=1}^{d-1} q_E(i)}\right) \cdot \frac{q_E(d_E)}{1 - \sum_{i=1}^{d_E-1} q_E(i)} \tag{19}$$

$$= q_E(d_E). \tag{20}$$

A similar derivation shows that $l_I(d_I) = q_I(d_I)$ which proves that the computational complexity of computing $A(\mathbf{z}_u)$ has been reduced to one factor for each contact during the S states of u and three additional factors corresponding to the compact representation ω for $\mathbf{z}_u$. The number of factors do not directly scale up with T.

Factors $B(\mathbf{z}_u)$ In order to derive a compact representation of $B(\mathbf{z}_u)$, we note that only the cases of $z_{v,t} = S$ potentially contain the value of $z_{u,t}$ for $v \neq u$ (see the value range of the function g and h in (4)). In fact, looking at (5) it becomes evident that it requires $z_{u,t} = I$. Thus, $B(\mathbf{z}_u)$ is defined by

$$\prod_{t=1}^{T} \prod_{v \in \mathcal{C}_S(u,t)} f(v, t, \mathcal{Z}_t) \prod_{v \in \mathcal{C}_E(u,t)} (1 - f(v, t, \mathcal{Z}_t)), \tag{21}$$

where $\mathcal{C}_S(u, t)$ and $\mathcal{C}_E(u, t)$ are the individuals that u met at time t who were susceptible and have either stayed susceptible or got exposed, respectively:

$$\mathcal{C}_S(u, t) := \{v : (u, v, t, \mathbf{x}) \in \mathcal{D}_{\text{contact}} \wedge z_{v,t} = S \wedge z_{v,t+1} = S\},$$
$$\mathcal{C}_E(u, t) := \{v : (u, v, t, \mathbf{x}) \in \mathcal{D}_{\text{contact}} \wedge z_{v,t} = S \wedge z_{v,t+1} = E\}.$$

4.2 Efficient Block-Gibbs Sampling

In this subsection, we describe how we can speed up the block-Gibbs sampling step for drawing a sample infection trace $\mathbf{z}_u$ for individual u.

Constant terms A key observation is that the term $C(\mathbf{z}_u) = \prod_{(t,o) \in \mathcal{T}_u} P(o|z_{u,t})$ in (13)—corresponding to test outcomes for individual u—is a constant for each infection trace $\mathbf{z}_u$ irrespective of the values of other infection traces $\{\mathbf{z}_{v:v \neq u}\}$. Thus, $C(\mathbf{z}_u)$ can be pre-computed at the start of the block-Gibbs sampling algorithm for individual u and then reused every time we evaluate the likelihood of an infection trace $\mathbf{z}_u$. Similarly, the terms $l_0(t_0) = (1-p_0)^{t_0-1}p_0$, $l_E(d_E) = q_E(d_E)$ and $l_I(d_I) = q_I(d_I)$ in $A(\mathbf{z}_u)$ in (14) are constant for each infection trace $\mathbf{z}_u$ irrespective of the infection traces of other individuals. Hence, these terms can also be pre-computed at the start of the block-Gibbs sampling algorithm for individual u.

Contacts into u The remaining term l_{infected} in $A(\mathbf{z}_u)$ (see (14)) captures the contribution due to contacts into individual u from other (infectious) individuals v who are in state $z_{v,t} = \mathrm{I}$ prior to u herself getting infected at t_0. As a result, l_{infected} depends on the values of other infection traces $\{\mathbf{z}_{v:v\neq u}\}$ and needs to be recomputed during each block-Gibbs sampling step to draw sample $\mathbf{z}_u$. Let us define

$$l_{\text{infected}}(t) \quad := \quad p_{u,t}\,, \tag{22}$$

$$l'_{\text{infected}}(t) \quad := \quad \frac{1-(1-p_0)p_{u,t}}{p_0}\,, \tag{23}$$

where $p_{u,t} := \prod_{(v,u,t,\mathbf{x})\in\mathcal{D}_{\text{contact}}:z_{v,t}=\mathrm{I}} \prod_j (1-p_j)^{x_j}$ (see also (17)). Then, we have

$$l_{\text{infected}} \quad = \quad \prod_{t=1}^{t_0-1} l_{\text{infected}}(t)\cdot l'_{\text{infected}}(t_0) \tag{24}$$

Thus, at the start of each block-Gibbs sampling step, we pre-compute (22) and (23) for each time step t, and then use (24) to compute l_{infected} for a particular infection trace $\mathbf{z}_u$. Note that this involves only t_0 multiplications (or additions in the log-domain).

Contacts out from u We next turn our attention to computing $B(\mathbf{z}_u)$ for each infection trace $\mathbf{z}_u$ that captures the contribution due to contacts out from u. We introduce two states $\mathcal{Z}_t^{\mathrm{I}}$ and $\mathcal{Z}_t^{\neg\mathrm{I}}$ which are identical to $\mathcal{Z}_t$ except for the value of infection state $z_{u,t}$ which is I is $\mathcal{Z}_t^{\mathrm{I}}$ and one of $\{S, E, R\}$ in $\mathcal{Z}_t^{\neg\mathrm{I}}$. Now, let

$$B(\mathbf{z}_u, t, \mathcal{Z}_t) := \prod_{v\in\mathcal{C}_S(u,t)} f(v,t,\mathcal{Z}_t) \prod_{v\in\mathcal{C}_E(u,t)} (1-f(v,t,\mathcal{Z}_t)) \tag{25}$$

be the inner terms in (21). Note that for all contacts $(u,v,t,\mathbf{x}) \in \mathcal{D}_{\text{contact}}$, the terms $B(\mathbf{z}_u, t, \mathcal{Z}_t^{\mathrm{I}})$ and $B(\mathbf{z}_u, t, \mathcal{Z}^{\neg\mathrm{I}})$ differ only in the factor $\prod_j (1-p_j)^{x_j}$ that is in $f(v,t,\mathcal{Z}_t^{\mathrm{I}})$ but not in $f(v,t,\mathcal{Z}_t^{\neg\mathrm{I}})$ since u is infectious at this time t in $\mathcal{Z}_t^{\mathrm{I}}$ but not in $\mathcal{Z}_t^{\neg\mathrm{I}}$. Also, note that values of $\mathbf{z}_{u,t'}$ for $t' < t$ do not affect $B(\mathbf{z}_u, t, \mathcal{Z}_t)$. We can then obtain $B(\mathbf{z}_u)$ for each infection trace $\mathbf{z}_u$ value as

$$B(\mathbf{z}_u) := \text{Constant}\cdot \prod_{t=t_0+d_E}^{t_0+d_E+d_I} \frac{B(\mathbf{z}_u, t, \mathcal{Z}_t^{\mathrm{I}})}{B(\mathbf{z}_u, t, \mathcal{Z}_t^{\neg\mathrm{I}})}\,, \tag{26}$$

where Constant is the product of $B(\mathbf{z}_u, t, \mathcal{Z}_t^{\neg\mathrm{I}})$ over all time steps t and can be ignored due to normalization of the sampling distribution. Again, the ratio $B(\mathbf{z}_u, t, \mathcal{Z}_t^{\mathrm{I}})/B(\mathbf{z}_u, t, \mathcal{Z}_t^{\neg\mathrm{I}})$ can be pre-computed for all time steps t at the start of the block-Gibbs sampling step for $\mathbf{z}_u$, and then used to compute $B(\mathbf{z}_u)$ for each infection trace $\mathbf{z}_u$ as in (26).

Putting it all together Note that the quantities $l_{\text{infected}}(t)$, $l'_{\text{infected}}(t)$ and $B(\mathbf{z}_u, t, Z_t^{\mathrm{I}})/B(\mathbf{z}_u, t, \mathcal{Z}_t^{\neg\mathrm{I}})$ only depend on the infection status of individual u at time t because the infection traces $\mathbf{z}_v$ of all other individuals v are fixed when we are drawing a block-Gibbs sample for $\mathbf{z}_u$. Thus, the (un-normalized) conditional probability for each $\mathbf{z}_u$ value is obtained by taking the product of $A(\mathbf{z}_u)$, $B(\mathbf{z}_u)$ and $C(\mathbf{z}_u)$, which in turn are computed efficiently as described above from pre-computed values of $l_0(t_0)$, $l_E(d_E)$, $l_I(d_I)$ and $C(\mathbf{z}_u)$ at the start of the algorithm, and $l_{\text{infected}}(t)$, $l'_{\text{infected}}(t)$ and $B(\mathbf{z}_u, t, \mathcal{Z}_t^{\mathrm{I}})/B(\mathbf{z}_u, t, \mathcal{Z}_t^{\neg\mathrm{I}})$ for all time steps t at the start of the block-Gibbs sampling step.

Additional implementation optimizations We use two additional ideas to accelerate the implementation of the block-Gibbs sampling algorithm:

- We never materialize the infection trace $\mathbf{z}_u$ because it is uniquely described by the triple $\omega = (t_0, d_E, d_I)$; each value $z_{u,t}$ can be computed by no more than three comparisons of t with t_0, $t_0 + d_E$ and $t_0 + d_E + d_I$. Thus, the whole state of the latent variable model is represented by $3 \times |\mathcal{S}|$ integers.
- We carry out all computations of probabilities in the log-domain so all functions become sums and products instead of products and powers.

Algorithm 1: Block-Gibbs sampling algorithm for CRISP model

```
/* Initialization */
Initialize each $\mathbf{z}_u = S \cdot \mathbf{1}$
/* Precomputations independent of contact data */
forall $(t_0, d_E, d_I) \in \mathbb{N}^+ \times \mathbb{N}^+ \times \mathbb{N}^+$ such that $t_0 + d_E + d_I \leq T$ do
    Pre-compute $l_0(t_0)$, $l_E(d_E)$ and $l_I(d_I)$
    Construct the sequence $\mathbf{z}_u$ with $t_0$ states S, $d_E$ states E, $d_I$ states I and $T - t_0 - d_E - d_I$ states R
    Pre-compute $C(\mathbf{z}_u)$ according to (13)
repeat
    Pick a random index $u$
    /* Precomputations dependent on contact data */
    forall time steps $t$ do
        Pre-compute $l_{\text{infected}}(t)$ using (22) and $l'_{\text{infected}}(t)$ using (23)
        Pre-compute ratio $B(\mathbf{z}_u, t, \mathcal{Z}_t^I)/B(\mathbf{z}_u, t, \mathcal{Z}_t^{\neg I})$ using (25)
    forall $(t_0, d_E, d_I) \in \mathbb{N}^+ \times \mathbb{N}^+ \times \mathbb{N}^+$ such that $t_0 + d_E + d_I \leq T$ do
        /* Infection trace specific computations */
        Construct the sequence $\mathbf{z}_u$ with $t_0$ states S, $d_E$ states E, $d_I$ states I and $T - t_0 - d_E - d_I$ states R
        Compute $\log(A(\mathbf{z}_u)) = \log l_0(t_0) + \log l_E(d_E) + \log l_I(d_I) + \log(l_{\text{infected}})$ using (24)
        Compute $\log(B(\mathbf{z}_u))$ using (26)
        Set $l_{t_0,d_E,d_I} = \log(A(\mathbf{z}_u)) + \log(B(\mathbf{z}_u)) + \log(C(\mathbf{z}_u))$
    /* Block-Gibbs sampling step */
    Sample $(t_0^*, d_E^*, d_I^*)$ with probability $\propto \exp(l_{t_0^*,d_E^*,d_I^*} - \max_{t_0,d_E,d_I}(l_{t_0,d_E,d_I}))$
    Set $\mathbf{z}_u$ with (S,E,I,R) states corresponding to $(t_0^*, d_E^*, d_I^*)$
    return $\mathcal{Z}^i = \mathcal{Z}$
until convergence
```

Block Glibbs Sampling Algorithm Algorithm 1 is block-Gibbs sampling algorithm for sampling $\mathcal{Z}_T^i$ from our CRISP model. It cycles through (random) individuals u, sampling the vector of latent variables $\mathbf{z}_u$ from the conditional distribution $P(\mathbf{z}_u | \{\mathbf{z}_{v:v \neq u}\}, \mathcal{O}, \theta)$ until convergence. We can use the samples $\mathcal{Z}_T^1, \ldots, \mathcal{Z}_T^m$ drawn by this algorithm to compute the infection risk score for an individual u at time t by taking the fraction of samples $\mathcal{Z}_T^i$ in which the latent infection state $z_{u,t} \in \{E, I\}$.

4.3 Hyperparameter Inference

In order to estimate the hyper-parameters θ of the CRISP model, would like to find θ^* that maximizes the log-likelihood log (12). However, since this is intractable, we propose to use the Monte Carlo Expectation-Maximization (EM) algorithm [2]. We will use EM to refine θ in successive iterations. Let θ_{old} be the value of θ computed in the previous iteration. Then, in the E step of the current iteration, we will estimate the expected complete-data log-likelihood

$$\sum_{\mathcal{Z}_T} P(\mathcal{Z}_T | \mathcal{O}, \theta_{\text{old}}) \cdot \log\left(P(\mathcal{Z}_T, \mathcal{O} | \theta)\right) \,. \tag{27}$$

We will use the block-Gibbs sampling procedure described in Algorithm 1 to approximate the posterior distribution $P(\mathcal{Z}_T | \mathcal{O}, \theta_{\text{old}})$ over the latent infection status of individuals u. If the samples drawn from the posterior $P(\mathcal{Z}_T | \mathcal{O}, \theta_{\text{old}})$ are $\mathcal{Z}_T^1, \ldots, \mathcal{Z}_T^m$, then in the M step, we will compute θ that maximizes the expected complete-data log-likelihood

$$\theta_{\text{next}} = \operatorname{argmax}_\theta \sum_{i=1}^{m} \log\left(P\left(\mathcal{Z}_T^i, \mathcal{O} | \theta\right)\right) \tag{28}$$

$$= \operatorname{argmax}_\theta \sum_{i=1}^{m} \sum_{t=1}^{T-1} \sum_{u} \log\left(P\left(z_{u,t+1}^i, \mathcal{O} | \mathcal{Z}_t^i, \theta\right)\right) \,, \tag{29}$$

where $z^i_{u,t+1}$ is the infection state for individual u at time $t+1$ in sample $\mathcal{Z}^i_T$. If t^i_0 denotes the number of initial S states in the sample infection trace $\mathbf{z}^i_u$, we note that by virtue of (4) only the first t^i_0 terms depend on θ which reduces the above maximization term to only

$$\sum_{i=1}^{m}\sum_{u}\sum_{t=1}^{t^i_0-1}\log\left(f\left(u^i,t,\mathcal{Z}^i_t|\theta\right)\right)+\log\left(1-f\left(u^i,t^i_0,\mathcal{Z}^i_t|\theta\right)\right)\,.$$

We use stochastic gradient descent to compute the θ values that maximize the above expression. We also note that for numerical stability, we re-parameterize p_j via w_j as $p_j = \exp(w_j)/(1+\exp(w_j))$ which allows for an unconstrained optimization over $\mathbf{w}$.

4.4 Federated Block-Gibbs Sampling

We can extend the block-Gibbs sampling algorithm in CRISP to a federated learning setting [13] where local contact and test outcome data for an individual u are utilized to compute the block-Gibbs sample $\mathbf{z}_u$ on the individual's mobile device *without* ever needing to be shared with anyone else. This has two benefits: (1) We distribute the block-Gibbs sampling algorithm across hundreds of millions of mobile devices in the world and thereby utilize their distributed computational power, and (2) Contact and test outcome data for an individual are stored only on the individual's mobile device and not shared with other mobile devices—-this preserves a user's privacy. In the federated setting, the contact data is never centralized—instead for each individual u, her device executes the block-Gibbs sampling step to draw sample $\mathbf{z}_u$ only using the locally available contacts and test outcome data for u, as well as additional "minimal statistics" sent to u by the devices of its past contacts. In the following two paragraphs, we explain how to compute the factors $A(\mathbf{z}_u)$, $B(\mathbf{z}_u)$ and $C(\mathbf{z}_u)$ in (13) in a federated setting (see Algorithm 2 for the pseudo-code which runs on every mobile device).

Factors $A(\mathbf{z}_u)$ **and** $C(\mathbf{z}_u)$ A key observation is that the terms $l_0(t_0)$, $l_E(d_E)$, $l_I(d_I)$ in $A(\mathbf{z}_u)$ as well as the factor $C(\mathbf{z}_u)$ can all be computed locally on the device with the contact and test outcome information available on the device. In order to compute the remaining term l_{infected} in $A(\mathbf{z}_u)$, we only require information on the individuals v who had a contact with u at each time step t and the infection status $z_{v,t}$ of v at the time of the contact. Individual u's mobile device already has the contact information for u; thus all that is required to compute l_{infected} are the current infection traces $\mathbf{z}_v$ for all individuals v who have had contacts with u. Since each infection trace is uniquely characterized by a (t_0, d_I, d_E) triple, we require the mobile devices of all individuals v who have had a contact with u to send u's device the (t_0, d_I, d_E) triple corresponding to $\mathbf{z}_v$.

Factor $B(\mathbf{z}_u)$ In order to compute $B(\mathbf{z}_u)$ as defined in (21), we require the term $f(v,t,\mathcal{Z}_t)$ for each individual v who has had a contact with u at time t and whose infection state $z_{v,t} = S$. Let $f_{-u}(v,t,\mathcal{Z}_t)$ be defined as in (5) over all contacts of v at time t except for individual u. Then, the device for each individual v who has had a contact with u at time t and whose infection state $z_{v,t} = S$ sends to u's device the quantity $f_{-u}(v,t,\mathcal{Z}_t)$ computed based on v's view of the infection traces of its contacts. These terms are used by u's device to compute $B(\mathbf{z}_u)$ as defined in (21).

5 Simulation-Based Experimental Results

In this section, we present two types of experimental evaluations:

1. *Population Level COVID-19 Infection Spread.* In the first set of experiments, we will demonstrate that CRISP is capable of modelling infection spread across an entire population. We will relate our individual-level parameters θ to more classical measures of infection spread such as reproduction factor R_0 and demonstrate that the structure of the contact patterns allow more fine grained control of the infection spread which can be used for alternative containment measures of the COVID-19 pandemic.
2. *Test and Quarantine Efficacy of CRISP Model.* In the second set of experiments, we will assess the test and quarantine efficacy of the CRISP model by comparing the population health after 5 months under three testing and quarantining policies: (1) symptom-based, (2) contact-tracing-based, and (3) CRISP model-based.

Algorithm 2: Federated block-Gibbs Sampling algorithm for CRISP model

```
/* Initialization */
Initialize $\mathbf{z}_u = S \cdot \mathbf{1}$
Initialize $\mathcal{N}_v = (\infty, 0, 0, \emptyset)$ for all $v$ // Stores minimal statistics from contacts

/* Precomputations independent of contact data */
repeat
    /* Update minimal statistic received in the incoming queue */
    forall $\{(t_0^v, d_E^v, d_I^v, \{f_{-u}(v, t, \mathcal{Z}_t)\})\}_v$ in the incoming message queue do
        $\mathcal{N}_v \leftarrow (t_0^v, d_E^v, d_I^v, \{f_{-u}(v, t, \mathcal{Z}_t)\})$
    /* Precomputations of test outcomes */
    forall $(t_0, d_E, d_I) \in \mathbb{N}^+ \times \mathbb{N}^+ \times \mathbb{N}^+$ such that $t_0 + d_E + d_I \leq T$ do
        Pre-compute $l_0(t_0)$, $l_E(d_E)$ and $l_I(d_I)$
        Pre-compute $C(\mathbf{z}_u)$ for this sequence according to (13)
    /* Precomputations dependent on contact data */
    forall time steps $t$ do
        Pre-compute $l_{\text{infected}}(t)$ using (22) and $l'_{\text{infected}}(t)$ using (23) and $(t_0^v, d_E^v, d_I^v)$ in $\mathcal{N}_v$ for all
          past contacts $v$
        Pre-compute ratio $B(\mathbf{z}_u, t, \mathcal{Z}_t^I)/B(\mathbf{z}_u, t, \mathcal{Z}_t^{-I})$ using (25) and $\{f_{-u}(v, t, \mathcal{Z}_t)\}$ in $\mathcal{N}_v$ for all
          past contacts $v$
    forall $(t_0, d_E, d_I) \in \mathbb{N}^+ \times \mathbb{N}^+ \times \mathbb{N}^+$ such that $t_0 + d_E + d_I \leq T$ do
        /* Infection trace specific computations */
        Construct the sequence $\mathbf{z}_u$ with $t_0$ states $S$, $d_E$ states $E$, $d_I$ states I and $T - t_0 - d_E - d_I$
          states $R$
        Compute $\log(A(\mathbf{z}_u)) = \log l_0(t_0) + \log l_E(d_E) + \log l_I(d_I) + \log(l_{\text{infected}})$ using (24)
        Compute $\log(B(\mathbf{z}_u))$ using (26)
        Set $l_{t_0, d_E, d_I} = \log(A(\mathbf{z}_u)) + \log(B(\mathbf{z}_u)) + \log(C(\mathbf{z}_u))$
    /* Block-Gibbs sampling step */
    Sample $(t_0^*, d_E^*, d_I^*)$ with probability $\propto \exp(l_{t_0^*, d_E^*, d_I^*} - \max_{t_0, d_E, d_I}(l_{t_0, d_E, d_I}))$
    Set $\mathbf{z}_u$ with (S,E,I,R) states corresponding to $(t_0^*, d_E^*, d_I^*)$
    /* Send minimal statistic to all contacts */
    forall $v$ in past contact list do
        $\mathcal{F} = \{f_{-v}(u, t, \mathcal{Z}_t) : (u, v, t, x) \in \mathcal{D}_{\text{contact}} \wedge t \leq t_0^*\}$
        Send message $(t_0^*, d_E^*, d_I^*, \mathcal{F})$ to $v$
until forever
```

In all these experiments, we use the parameters $\alpha = 0.001$ and $\beta = 0.01$ in (9) and match the distribution q_E and q_I of exposure and infectiousness duration to the empirical distributions provided in the medical literature [1, 18]. This is both used in the generation of the simulated test outcome data as well as for the CRISP inference algorithms as these parameters are publicly known. We will use the notation $\overline{q_I}$ for the expectation of the empirical distributions q_I.

In order to simulate realistic epidemiological spread, we need to translate a reproduction factor R_0 at $t = 0$ into contact data. By definition, R_0 is the average number of individuals that an infected person will infect over the entire period of being infectious. Thus, for a reproduction factor R_0 and a contact channel j with transmission probability $p_j \in [0, 1]$, we need to generate $C(R_0, p_j) := R_0/(\overline{q_I} \cdot p_j)$ many connections on average for all individuals in each time step. Conversely, for any process that generates η_j connections to unique and distinct individuals over channel j in each time step, the effective R_0 over contact channel j with 100% transmission probability equals $\overline{q_I} \sum_j \eta_j$. The actual number of contacts is drawn form a binomial distribution with $n = S - 1$ and a rate $p = \frac{C(R_0, p_j)}{2(S-1)}$. Note that the rate is one half of the target contact rate because all contacts are symmetrically mirrored.

5.1 Population Level COVID-19 Infection Spread

In order to assess if the CRISP model is able to provide realistic population-level statistics for COVID-19 infection spread, we simulate a population of $|\mathcal{S}| = 10,000$ individuals over a period of 274 days (9 months). We single out an individual u for whom we set $p_0 = 1$ so that she will get infected with probability 100% at $t = 1$ ("patient 0"); for all other people we assume a $p_0 = 10^{-6}$ to model a miniscule chance of infection spread from exogenous sources. We assume a single contact channel with a 1% chance of transmission, $p_1 = 0.01$. We simulate five scenarios:

- *No Mitigation.* Since R_0 of COVID-19 is estimated to be 2.5, at any time t we generate $C(2.5, p_1)$ random connections for every individual at every time step.
- *Social Distancing After 60 Days.* Intuitively, the "locality" of the contact patterns should play a role in the infection spread of COVID-19: if an individual is in contact with a broad range of other individuals, the spread should be faster than if unique number of people in contact over time is small. In order to demonstrate that this concept has indeed an effect on the infection spread, we performed an additional simulation where we kept the unique *number* of people that every individual meets in every time step at $C(2.5, p_1)$ but introduced the concept of "social bubbles" where all individuals form groups of 20 who have a large number of interactions with each other (i.e., equivalent to $C(2, p_1)$ but only rare interactions with people from other bubbles equivalent to $C(0.5, p_1)$ (see Figure 3 for a picture of the contact matrix with random connections and with "social bubbles").
- *Mitigation After 60 Days.* For $t \leq 60$, we generate $C(2.5, p_1)$ random connections for every individual at every time step. Afterwards, we assume that mitigation measures are taken which reduce the reproduction rate to 1.0. Thus, we generate only $C(1.0, p_1)$ random connections for every individual at every time step $t > 60$.
- *Suppression After 60 Days.* For $t \leq 60$, we generate $C(2.5, p_1)$ connections for every individual at every time step. Afterwards, we assume that lock-down measures are taken to suppress the pandemic which reduce the reproduction rate to 0.5.
- *Suppression After 60 Days and Release of Lock-down after 120 Days.* This scenario is similar to the previous scenario but we assume that due to very low infection numbers, the lock-down is released after 60 days. Thus, we generate $C(2.5, p_1)$ random connections for every individual for $t > 120$.

In Figure 2, we show the plot of $\sum_u P(z_{u,t} = z)$ over $t = 1, \ldots, 274$ days for $z \in \{E, I, R\}$ (orange = E, red = I, blue = R) from 100 forward samples of the CRISP model for these scenarios. As one can see, with no mitigation there is a high peak around day $t^* = 180$ and eventually herd-immunity is achieved at 85% of infected population. Even though the number of unique contacts in each time step is the same, "social bubbles" flatten the curve, thus slowing down the infection but growth rates of infected people are still super linear until large parts of the population had been in contact with the disease. Note that a similar mitigation policy is currently used in Belgium. In case of mitigation to $R_0 = 1.0$, growth rates are pushed to sub linear but the pandemic is still continuously going on after 9 months. Not surprisingly, suppression is most effective at bringing the infections back to nearly 0% after 120 days. However, if the lock-down is lifted after 120 days, a second wave of infections will cause an exponential increase in infectiousness after only two weeks (dashed lines). Note that all these effects were computable by simply forward sampling our individual-level CRISP model.

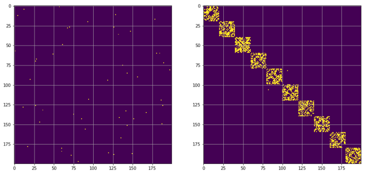

Figure 3: Snapshot of the contact matrix of the first 200 individuals for random connections and with "social bubbles".

5.2 Test and Quarantine Efficacy of CRISP Model

In order to assess the test and quarantine efficacy of the CRISP model, we consider a population of $|\mathcal{S}| = 1,000$ individuals for 150 days (5 months) with a uniformly random contact pattern of $C(2.5, 0.025) = 5.03$ contacts on average per individual and day. We simulate the actual infection spread by applying the following sequence in each time step (i.e., day): At the beginning of each time

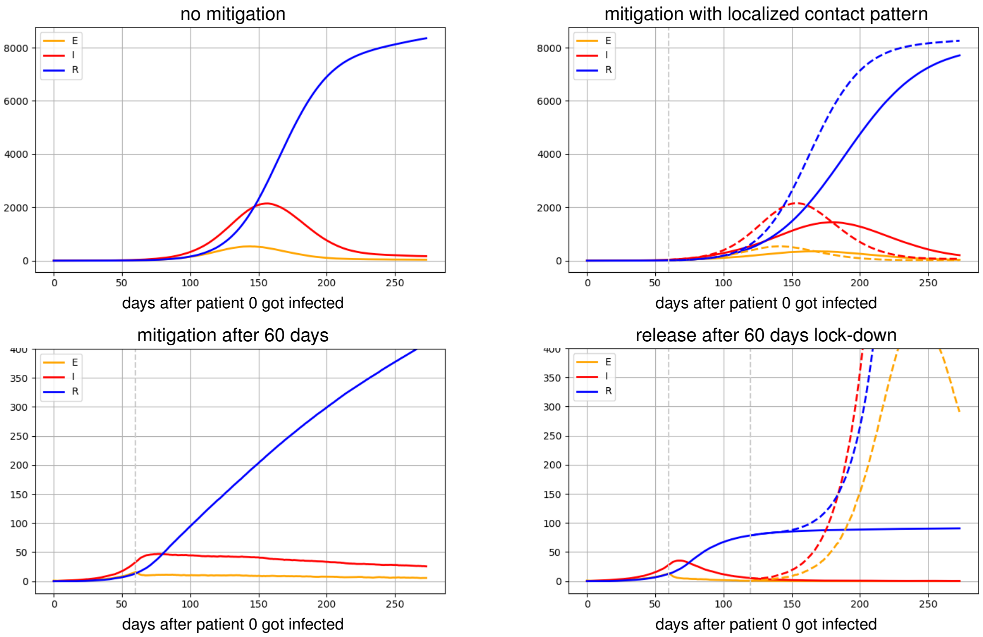

Figure 2: Population level COVID-19 infection spread for three different scenarios: *(top-left)* No mitigation ($R_0 = 2.5$). *(top-right)* No mitigation until day 60 and then using "social bubbles". Note that R_0 remains at 2.5 the entire time. *(bottom-left)* Mitigation after 60 days by reducing R_0 to 1.0. *(bottom-right)* Lock-down at day 60 and reduction of R_0 to 0.5 (solid lines). In dashed lines we show the effect of a subsequent re-opening of a subsequent contact rate increase to R_0 of 2.5 starting at day 120.

step, we query the testing-and-quarantining policy for a list of individuals which need to be tested and need to be in quarantine during this step (this will only be done after $t^* = 30$ to simulate an undetected initial outbreak). Each policy is constrained to select no more than 10 test candidates per day (1% of the total population). Given the quarantined individuals on that day, we remove contacts from and to the quarantined individuals for that day and then use the CRISP forward model (4) and CRISP test outcome model (9) to draw one sample of the next simulated infection state of every individual as well as the actual test outcomes of the requested test candidates. If the infection state of an individual changes from E to I in this sampling step, we assume that with 50% probability, the individual generates symptoms. Finally, at the end of the time step, the testing-and-quarantining strategy is revealed the test outcomes as well as the list of symptomatic individuals (again, provided $t \geq t^*$). We single out an individual u for whom we set $p_0 = 1$ so that she will get infected with probability 100% at $t = 1$ ("patient 0"); for all other people we assume a $p_0 = 10^{-4}$ to model a small chance of infection spread from exogenous sources.

1. *Symptom-Based Policy*. For every time step $t \geq t^*$, we will request testing for up to 10 symptomatic individuals from the previous time step. For all individuals with a positive test outcome on the previous day, we will institute a quarantine for ρ time steps where ρ ranges from 2 to 21 days in our evaluation.
2. *Contact-Tracing Policy*. For every time step $t \geq t^*$, we will request testing for up to 10 symptomatic individuals from the previous time step. If there are less than 10 symptomatic individuals, then we will request the remaining tests for individuals in quarantine sorted in descending order of the number of contacts they have had in the past 7 days with people who have tested positive. For every individual with a positive test outcome, we will not only quarantine her but also all the contacts she had in the past 7 days for ρ time steps where ρ ranges from 2 to 21 days in our evaluation; for every individual with a negative test outcome, we will remove her from quarantine.
3. *CRISP Model-Based Policy*. For every time step $t \geq t^*$, we will use block-Gibbs sampling of 100 infection traces $\mathbf{z}_u$ to estimate $P(z_{u,t})$ for every individual u at the current time step t based on the contacts and test outcomes prior to time step t. We will request testing

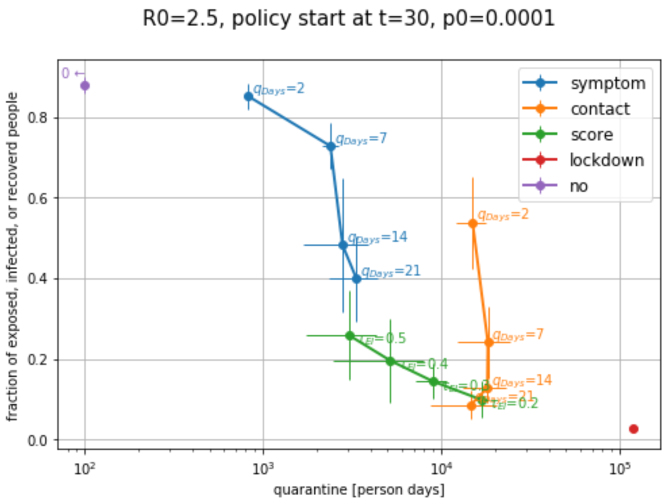

Figure 4: Effect of different mitigation policies on the infection percentage and quarantine days after $T = 150$ days (5 months). The y-axis shows the percentage of population that got infected with COVID-19 during the 150 days. The x-axis shows the total number of days that individuals were quarantined. The error-bars are computed as the standard deviation over 20 random initializations of the forward model simulating the $T = 150$ days while not affecting the randomization of the contact matrices.

for up to 10 symptomatic individuals from the previous time step. If there are less than 10 symptomatic individuals, then we will request the remaining tests for individuals (who have not tested positive before) in descending order of $\hat{P}(z_{u,t} = I)$. We will quarantine any individual who is not yet quarantined but whose estimated probability $\hat{P}(z_{u,t} \in \{E, I\})$ exceeds a given policy threshold τ_{EI}; we will release an individual from quarantine once their estimated probability $\hat{P}(z_{u,t} \in \{S, R\})$ exceeds a given policy threshold τ_{SR}. Note that we increase p_0 in the block-Gibbs sampling by a factor of 10 to account for "patient 0".

In order to gauge the efficacy of each policy, we measure two quantities at the end of the simulation ($t = 150$): (1) Percentage of population that got infected during the 150 days, and (2) total number of days that individuals were quarantined (e.g., if a policy locks down for the entire 150 days, this would result in 150,000 quarantine days). Varying the policy parameters ρ, τ_{EI} and τ_{SR} results in curves on the two dimensions of infection percentage and quarantine days. The closer a curve is to the origin, the more effective is the policy in terms of "health" (infection) and "economic" (quarantining) cost.

In Figure 4, we plot curves for the three policies with $\rho \in \{2, 7, 14, 21\}$, $\tau_{EI} \in \{0.2, 0.3, 0.4, 0.5\}$, $\tau_{SR} = 0.9$. For comparison, we also show the two extreme points corresponding to "no mitigation" (i.e., zero quarantine days but the largest infection percentage of 90%) and "full lock-down" (i.e., largest quarantine days of 120,000 and near-zero infection percentage). All three curves exhibit a negative slope where a higher percentage of quarantine days corresponds to a more effective mitigation of infection spread. Of the three policies, our CRISP-based policy achieves the best performance in terms of the smallest number of quarantine days for a given infection percentage (i.e., Pareto frontier). This is because our CRISP model is able to accurately identify infectious users (even though they may be asymptomatic) and test/quarantine them proactively– this helps to prevent infection from spreading across the population while at the same time quarantining fewer individuals with a high likelihood of getting infected. In contrast, the symptom-based policy only tests individuals with symptoms and then quarantines the individuals who have tested positive. As a result, since 50% of the infected individuals are asymptomatic, they never get tested and quarantined, thus resulting in a spread of infection to 60% of the population. Similarly, the contact-tracing policy, by isolating all contacts of positive tested individuals (many of whom may have low likelihoods of getting infected), is able to achieve the absolute smallest infection percentage but at the cost of massive quarantining (30% of the population). Figure 5 shows a visualization of these effects in one of the simulation runs for $\rho = 14$, $\tau_{EI} = 0.3$, and $\tau_{SR} = 0.9$.

6 Conclusions

In this paper, we proposed a probabilistic graphical model for COVID-19 infection spread through individual contacts that captures the epidemiological state of each individual based on the SEIR model. We developed a computationally efficient block-Gibbs sampling-based algorithm to infer

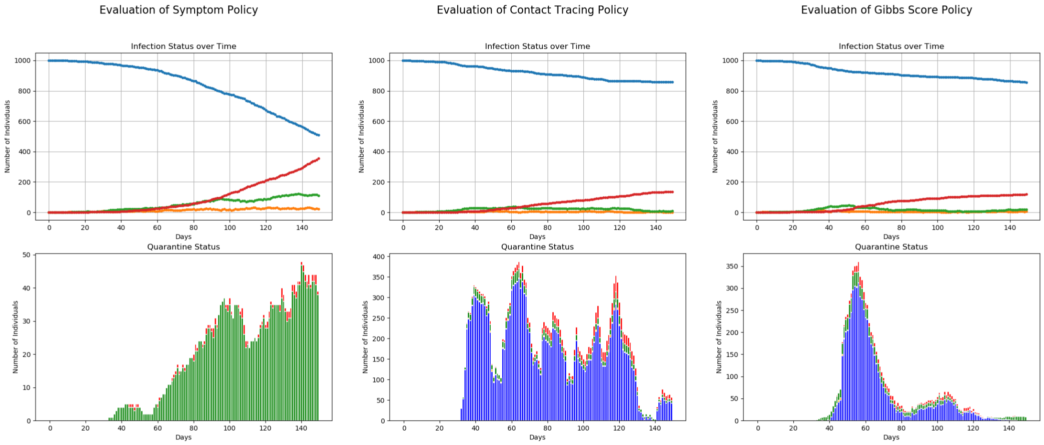

Figure 5: Infection trace and quarantining statistics for symptom-based (left), contact-tracing (middle), and CRISP ($\tau_{\mathrm{EI}} = 0.3, \tau_{\mathrm{SR}} = 0.9$) model-based (right) testing-and-quarantining policy over the duration of 150 simulated days (blue = S, orange = E, green = I, red = R). In the bottom plots, we show a stacked bar chart of quarantined individuals per day grouped by actual infection status. While the number of quarantined individuals for the symptom-based policy is small, the infection spread is not contained and the quarantining keeps growing exponentially. In contrast, the contact-tracing policy effectively suppresses infection spread while regularly quarantining more than 25% of the population. The CRISP model-based policy is initially picking a large number of individuals for quarantining but is then able to keep it at a low-level, in particular of susceptible individuals.

the COVID-19 infection risk score of all individuals at any time, given test outcome and mutual contact information between individuals. An efficient C++-based Python implementation of our inference algorithm is available at https://github.com/zalandoresearch/CRISP. Through experiments with simulated data, we showed that the CRISP model is able to model macro-level characteristics of the COVID-19 infection at county level ($\approx 10,000$ individuals) and effectively mitigate COVID-19 spread by pro-actively quarantining and testing individuals with high risk of infections.

As part of future work, we would like to further accelerate our inference procedure using other approximation techniques such as Variational Bayes and Expectation Propagation [2]. Our inference algorithm can also be speeded up by exploiting the parallelism inherent in our block-Gibbs Sampling algorithm. For example, it is possible to concurrently sample infection traces of two individuals with no contacts in common. It is also known that the hyper-parameters of the SEIR model vary with demographic attributes such as age, socio-economic status, or location (see, for example [10] who present a location-varying infection spread model). We would like to extend our model with group-level hyper-parameters to account for this variation. We would also like to explore the causal impact of mitigation or suppression policy measures (e.g., school closures, shop closures, small group gatherings) on COVID-19 infection spread when using contact-level data. Finally, we would like to consider more sophisticated models of COVID-19 transmission through different modalities, and contacts with varying duration and distance characteristics.

Acknowledgments We would like to thank Sebastian Munoz, Christopher Gandrud, Jasvinder Kandola and Peter Herbrich for all their valuable input and feedback on earlier drafts of this paper. We are also indebted to Christoph Thöns for his support with Python and C++ programming.

References

[1] Jantien A Backer, Don Klinkenberg, and Jacco Wallinga. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20-28 January 2020. *Euro Surveillance*, 25(5), 2020.

[2] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security*, 10(4), 2008.

[4] R. Deardon, S. P. Brooks, B. T. Grenfell, M. J. Keeling, M. J. Tildesley, N. J. Savill, D. J. Shaw, and M. E. Woolhouse. Inference for individual-level models of infectious diseases in large populations. *Statistica Sinica*, 20(1), 2010.

[5] N. M. Ferguson, D. A. Cummings, C. Fraser, J. C. Cajka, P. C. Colley, and D. S. Burke. Strategies for mitigating an influence pandemic. *Nature*, 442(7101):448–452, 2006.

[6] N. M. Ferguson, D.A. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn, and D. S. Burke. Strategies for containing an emerging influenza pandemic in southeast asia. *Nature*, 437(7056):209–214, 2005.

[7] Amit Goyal, Francesco Bonchi, and Laks Lakshmanan. Learning influence probabilities in social networks. In *Web Search and Data Mining (WSDM)*, 2010.

[8] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *International Colloquium on Automata, Languages and Programming*, 2003.

[9] Yuchen Li, Ju Fan, Yanhao Wang, and Kian-Lee Tan. Influence maximization on social graphs. *IEEE Transactions on Knwledge and Data Engineering*, 30(10):1852–1872, 2018.

[10] Lars Lorch, William Trouleau, Stratis Tsirtsis, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. A spatiotemporal epidemic model to quantify the effects of contact tracing, testing, and containment. *arXiv:2004.07641v2*, 2020.

[11] Michael Mathioudakis, Francesco Bonchi, Carlos Castillo, Aristides Gionis, and Antti Ukkonen. Sparsification of influence networks. In *International Conference on Knowledge Discovery and Data Mining*, 2011.

[12] Robert M. May. *Infectious diseases of humans: dynamics and control*. Oxford University Press, 1991.

[13] H.B. McMahan, E. Moore, D. Ramage, S. Hampson, and B.A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282, 2017.

[14] P. Van Mieghem and J. Omic. In-homogeneous virus spread in networks. *arXiv:1306.2588v2*, 2014.

[15] P. Van Mieghem, J. Omic, and R. Kooij. Virus spread in networks. *IEEE/ACM Transactions on Networks*, 2009.

[16] Seth Myers and Jure Leskovec. On the convexity of latent social network inference. In *Neural Information Processing Systems*, 2010.

[17] Vineetha Warriyar K. V., Waleed Almutiry, and Rob Deardon. Individual-level modeling of infectious disease data: Epiilm. *arXiv:2003.04963v1*, 2020.

[18] Roman Woelfel, Victor Max Corman, Wolfgang Guggemos, Michael Seilmaier, Sabine Zange, Marcel A Mueller, Daniela Niemeyer, Patrick Vollmar, Camilla Rothe, Michael Hoelscher, Tobias Bleicker, Sebastian Bruenink, Julia Schneider, Rosina Ehmann, Katrin Zwirglmaier, Christian Drosten, and Clemens Wendtner. Clinical presentation and virological assessment of hospitalized cases of coronavirus disease 2019 in a travel-associated transmission cluster. *medrxiv.org:10.1101/2020.03.05.20030502v1*, 2020.

[19] WWW. Aarogya setu. `https://www.mygov.in/aarogya-setu-app/`, 2020. Accessed: 2020-05-10.

[20] WWW. Trace together. `https://www.tracetogether.gov.sg`, 2020. Accessed: 2020-05-10.