Abstract—The COVID-19 pandemic has emphasized the need for a robust understanding of epidemic models. Current models of epidemics are classified as either mechanistic or non-mechanistic: mechanistic models make explicit assumptions on the dynamics of disease, whereas non-mechanistic models make assumptions on the form of observed time series. Here, we introduce a simple mixture-based model which bridges the two approaches while retaining benefits of both. The model represents time series of cases as a mixture of Gaussian curves, providing a flexible function class to learn from data, and we show that it arises as the natural outcome of a stochastic process based on a networked SIR framework. This allows learned parameters to take on a more meaningful interpretation compared to similar non-mechanistic models, and we validate the interpretations using auxiliary mobility data collected during the COVID-19 pandemic. We provide a simple learning algorithm to identify model parameters and establish theoretical results which show the model can be efficiently learned from data. Empirically, we find the model to have low prediction error. Moreover, this allows us to systematically understand the impacts of interventions on COVID-19, which is critical in developing data-driven solutions for controlling epidemics.

Index Terms—Complex networks, epidemics, forecasting, mixture models.

I. INTRODUCTION

The COVID-19 pandemic has reinforced the need for a deep understanding of epidemic processes. The initial uncertainty which arose in the beginning of the pandemic led to new questions about infectious disease modeling and estimation, as well as an emphasis on robust control of the pandemic with attention to careful trade-offs between health outcomes and economic costs of interventions [1], [3], [8], [10], [27].

Epidemic processes are often understood through models, which attempt to simplify the complex process by which contagion travels between individuals in a population in order to provide insights and actionable policies [23], [29], [32]. Broadly, models in the literature are typically categorized into one of two types: Mechanistic models and non-mechanistic models [24]. Both types of models have benefits and drawbacks, and in this work we propose a single model which unifies the two approaches while retaining benefits of both.

Mechanistic models are well established and have been used by epidemiologists since the early 20th century to understand the dynamics of the spread of infectious disease [29]. This class of models makes assumptions on the underlying process by which disease spreads, through the use of differential equations or more complex agent-based approaches [2], [5], [8], [17], [22], [37]. Although these models are well understood, they have some drawbacks when applied to observed data [26]. Due to the indirect relationship between model parameters and forecasts, mechanistic models can have wide confidence intervals in forecasting and estimation of the pandemic [22]. Classical models also often make the assumption that individuals do not react to the state of the pandemic and hence a proper adjustment to such models requires the introduction of additional parameters, which are often time-varying, to accurately estimate the pandemic state, increasing the burden on parameter uncertainty [4], [9], [36].

Due to these limitations, which were particularly exacerbated during the initial stages of uncertainty during the COVID-19 pandemic, non-mechanistic models grew in popularity for epidemic forecasting [26], [35]. These models, also referred to as reduced form models, hypothesize a form for the observed time series. As opposed to assuming a prolonged period of exponential growth as is often the case for mechanistic models, reduced form models often make an implicit assumption that human behavior generates a sub-exponential trend in observed time series. This approach has been applied since the mid 19th century, and often results in better short-term forecasts with tighter confidence intervals [18], [38]. However, a major critique of such methods is that they are often not interpretable: the learned parameters do not have direct interpretations related to the spread of disease [26].

In this work, we focus on a specific reduced form model which bridges the two approaches to epidemic modeling while retaining benefits of both. Namely, we assume that the observed time series of cases has the form

\[ N(t) = \left( \sum_{k=1}^{r} e^{-\alpha_k t^2 + b_k t + c_k} \right) (1 + \varepsilon_t). \] (1)

Here, \( t \) denotes a discrete time index, \( r \) is a parameter which denotes the number of mixtures present in the time series, and \( \varepsilon_t \) represents noise which is bounded in absolute value almost

Manuscript received 17 October 2023; revised 7 February 2024; accepted 21 February 2024. Date of publication 11 March 2024; date of current version 25 March 2024. The work of Arnab Sarker and Ali Jadbabaie was supported by Vannevar Bush Fellowship from the Office of the Secretary of Defense. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hoi-To Wai. (Corresponding author: Arnab Sarker.)

Arnab Sarker is with the Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: arnabs@mit.edu).

Ali Jadbabaie is with the Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA 02139 USA, and also with the Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

Devavrat Shah is with the Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA 02139 USA, and also with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

This article has supplementary downloadable material available at https://doi.org/10.1109/TSIPN.2024.3375600, provided by the authors.

Digital Object Identifier 10.1109/TSIPN.2024.3375600
 surely by a parameter $0 < \delta < 1$. This noise need not be independent, which is a realistic assumption often during the COVID-19 pandemic as testing procedures often create a time dependence in observation noise. The model itself takes a reduced form approach to estimation, which allows us to perform a principled statistical analysis of a simple learning algorithm.

Moreover, we show that the model is the outcome of a simple Susceptible-Infected-Recovered (SIR) process on a network, which allows the learned parameters to become interpretable. The flexibility in the number of mixtures $r$ also allows the model to account for the possibility of additional peaks in the data, resulting in a benefit over unimodal models used in the literature [18], [35]. Because $r$ can vary, the model provides a non-parametric function class from which the trajectory of the epidemic can be learned.

The motivation for modeling cases as a mixture stems from the reality that the disease is spreading to a population which has diverse regional divisions and includes many jurisdictions [8]. Since each region has its own features and policies, we expect the observed case counts to take an additive form, and this was indeed the case within the United States during the initial stages of the pandemic (Fig. 1). The specific form of the Gaussian time series is chosen in part due to historical prevalence [18], [38], and the parameterization is rigorously justified in Section IV. A much wider variety of function classes can explain sub-exponential growth [11], and for example allowing the quadratic parameters of the form $a_k$ in (1) to be a time-varying function of the epidemic state is an important extension of this model [25]. Here, we restrict to the parsimonious class in (1) since the restrictive assumption better justifies applications to out-of-sample prediction.

**Summary of Contributions:** Overall, the contributions of this work may be summarized as follows:

- We provide an efficient algorithm for learning the model from data, and provide theoretical guarantees.
- We show that the proposed mixture model arises naturally as the outcome of a Susceptible-Infected-Recovered (SIR) process on a network model and thus provide an interpretation for the parameters of the model.
- We validate this interpretation of the parameters by associating learned parameters with empirical observations, and we find significant correlations which support the interpretation given by the generative model.
- We provide an alternative explanation of the model (1) in which the observations occur due to endogenous behavior, which furthers our understanding of the model.

## II. RELATED WORK

### A. Mechanistic Models for Epidemics

We begin with a discussion of mechanistic models in epidemiology, which will underlie some of the theoretical foundations of the generative network model which produces observations of the form in (1). The fundamental mechanistic models are originally based on a mean-field approach introduced in [29], and are known as compartmental models.

1) **Compartmental Models:** Compartmental models approximate the dynamics of infectious disease by classifying the population into different groups, and using these classifications to characterize the spread of the disease. Labels such as $S$, $E$, $I$, and $R$ are often used to categorize such compartments, which stand for susceptible, exposed, infected, and recovered, respectively. Mean-field models in epidemiology model the dynamics of the infectious disease through a series of differential equations which describe the transmission of the disease. Here, we discuss the Susceptible-Infected-Recovered (SIR) model of infectious disease due to its ability to capture many of the relevant features of the COVID-19 pandemic while retaining enough simplicity to enable tractable analysis. In the SIR model, the dynamics between the compartments are modeled as if every individual has an equal probability of coming into contact with another individual, i.e. the underlying contact network is assumed to be fully connected. This leads to the following set of differential equations to model the spread of disease.

\[
\frac{dS}{dt} = -\beta S(t)I(t)/N, \tag{2}
\]
\[
\frac{dI}{dt} = \beta S(t)I(t)/N - \gamma I(t), \tag{3}
\]
\[
\frac{dR}{dt} = \gamma I(t). \tag{4}
\]

Here, $t$ represents continuous time, and the process is assumed to begin at some time $t_0$. We assume that the number of births and deaths in the population is negligible, so that $S(t) + I(t) + R(t) = N$ for some positive constant $N$ for all $t$. Further, the initial conditions for such a system often assume that $S(t_0) \approx N$, $0 < I(t_0) \ll N$, and $R(t_0) = 0$. The value $\beta$ in this model represents a contact rate, in the sense that it is a measure of the number of contacts that an infected individual may have with the
susceptible population. Similarly, the parameter $\gamma$ represents a recovery rate, and represents on average how long an infected individual remains contagious.

Several important observations can be made about the SIR model considered above. For example, the ratio $\beta/\gamma$, often denoted as $R_0$, is referred to as the basic reproductive number of the system and provides a simple condition to determine if the number of infections of the population will increase. Namely, if $R_0 < 1$, then the number of infected individuals will decay to 0; otherwise, the number of infected individuals will increase until it reaches a maximum before decaying to 0 [23, Theorem 2.1]. It is also common to rewrite the SIR model with time-varying parameters by considering how the parameters $\beta$ and $\gamma$ may change over time due, for example, to social distancing or increased testing [24].

Ultimately, although mean-field compartmental models have been applied to various problems in epidemiology, they do not take into account potential heterogeneity in the physical contact networks of individuals who spread the infectious disease [7], [16], [20]. For example, because super-spreading events accelerate the spread of COVID-19, it is unlikely that a homogeneous contact rate $\beta$ is appropriate to model the dynamics of this infectious disease [19]. For this reason, many authors instead choose to use network-based models.

2) Network-Based Models of Epidemics: By modeling contact networks, one might explicitly take into account the interactions between different individuals [16], [20], [21], [28], [37]. In a network-based model, individuals are modeled using a set of nodes $V (|V| = n)$, and the interactions between the individuals are modeled using a set of edges $E \subseteq V \times V$. At each (discrete) time step $t \in \mathbb{N}$, each node in the network is assigned a label, $\xi_i(t) \in \{S, I, R\}$. The state of the entire network can then be summarized in the time-varying vector

$$\xi(t) = (\xi_1(t), \ldots, \xi_n(t)) \in \{S, I, R\}^n.$$  \hspace{1cm} \hspace{1cm} (5)

The model may be fully described by then stating the probability that an individual node transitions from one particular state to another. In many network-based models of epidemics, each infected node infects its neighbors independently with some probability $\beta$, and each infected node becomes recovered at each time step with probability $\gamma$. Hence, using these transition probabilities, one can study the dynamics of the 3rd state Markov chain represented by these underlying probabilistic rules [37]. In particular, as Ganesh et al. [20] find that if $\beta/\gamma$ is less than the inverse of the spectral radius of the graph, then the epidemic may die out quickly, whereas above this threshold the epidemic will last for a long period of time.

Several network-based mechanistic models of epidemics also incorporate human behavior in their modeling assumptions [4], [25], [27], [36], wherein the parameters which govern the spread of disease may be affected by the current state of the epidemic on the network. That is, as the epidemic size increases, these models assume that individuals take precautionary measures to mitigate the spread of disease.

While such results provide insight on the dynamics of infectious disease in a networked setting, this literature does not specifically attempt to forecast future progress of the infectious disease. In epidemic forecasting with network-based mechanistic models, practitioners must make assumptions about the underlying social network, for example through agent-based modeling [33]. By collecting information about how individuals interact with their environment, such models can be made to handle heterogeneity in ways that mean-field models can not. These approaches require large amounts of data and resources to generate accurate predictions [3]. Since millions of edges may be included in the mobility network, and observed mobility often changes over time due to community response to the state of the pandemic, the amount of data required to satisfy complex, time-varying network-based models can become prohibitive.

B. Non-Mechanic Modeling of Epidemics

In contrast to the mechanistic models of infectious disease considered above, several authors approach prediction of epidemics using a non-mechanistic approach [18], [31], [38]. In these approaches, the trajectory of key metrics such as case counts or fatalities is assumed to come from a specific class of functions, and optimization techniques are used to determine the parameters of the models and provide forecasts. While such approaches often lack interpretable parameters, they do have computational benefits, are data-driven, and are efficient in their use of data as they provide forecasts based only on the available time series.

The most commonly considered class of functions for predicting the number of infected individuals is the class of Gaussian bell curves, which have been considered as early as the mid 19th century with the work of William Farr [18]. The use of this function class, often referred to as Farr’s law, has been considered in predicting infections due to AIDS [6], smallpox [38], drug mortality [13], and COVID-19 [35], with mixed results. In particular, the function class makes the restrictive assumption that the spread of disease is unimodal in nature. Hence, a key feature of this work is to extend the original model of [18] in a flexible way that is more appropriate for epidemic forecasting.

Recently, additional function classes based on advances in machine learning have also been considered in the context of the spread of infectious disease [31]. With such models, the class of functions used for prediction is not necessarily unimodal, and practitioners are able to incorporate auxiliary data such as mobility data in order to make predictions with a non-mechanistic approach.

Our work presents a non-mechanistic approach to learning the model in (1). Since the approach is non-mechanistic, it has the benefit of being data-driven and having an efficient implementation. Further, we find that this function class may actually be thought of as arising from a mechanistic SIR process in which individuals react to the spread of disease. Specifically, we show that each individual component of the mixture in (1) can be viewed as arising from the population’s reaction to the epidemic as a function of time (Section IV-A), expanding on the work of [38] by providing a network-based stochastic model to justify the shape of the Gaussian curve. Alternatively, each
component of the mixture can also be viewed as a population’s reaction to the state of the epidemic (Section V), expanding on related work in [4], [25], [36] by specifying a way in which state-dependent epidemic control can result in a Gaussian curve. The number of peaks \( r \) in (1) then provides an indication of the number of communities in the network as opposed to a more granular assumption on network topology; in this sense, the approach strikes a balance between the unstructured mean-field SIR models and the overly explicit network-based SIR models, while retaining the benefits of non-mechanistic approaches.

III. LEARNING MIXTURES FROM DATA

Learning a mixture of Gaussians as a time series requires novel algorithmic insights as the setting of this problem is different than related work on learning probability distributions (c.f. [14], [34]). The input to the problem are a series of observations \( \{ N(t) \} \) which are drawn according to the model (1). The goal of the learning procedure is then to identify a number of mixtures \( r \) and the parameters \( \{ a_k, b_k, c_k \} \) for \( k = 1 \) through \( r \) which provide a best fit to the data. To operationalize a best fit to the data, we will use the \( \ell_2 \) loss, i.e. we wish to find parameters which minimize

\[
L_r(\{a_k, b_k, c_k\}_{k=1,\ldots,r}) = \sum_{t=1}^{T} \left( \sum_{k=1}^{r} e^{-a_k t^2 + b_k t + c_k} - N(t) \right)^2 . \tag{6}
\]

This is a difficult non-convex optimization problem, and it is worth noting that it is markedly different than learning a distribution from a mixture of Gaussians. Although a method such as expectation-maximization (EM) can be used to learn a mixture of Gaussians from a collection of samples from a distribution, learning from a time series provides a markedly different setting [34]. Rather than learning from a collection of samples, we must learn directly from the observed points of the time series, which would be comparable to being given a part of a density function and attempting to estimate the complete density. Thus far, the task of fitting parameters to the mixture of Gaussians form has been considered by Singhal et al. [41] who use general non-convex optimization techniques to fit model parameters [12]. Here, we provide an intuitive algorithm for learning model parameters and provide interpretable parameter estimation bounds for this proposed algorithm.

Our algorithm is based on analyzing a transformation of \( N(t) \) which allows us to find a reasonably separated set of mixtures by using a peak finding algorithm. That is, although the estimation procedure here is ad hoc in the sense that it is not based on a formal maximum likelihood procedure, it has a key conceptual benefit in being able to identify separate outbreaks in the time series data of the pandemic. That is, the algorithm is particularly well-suited to this application.

We will begin by providing an intuition for the algorithm, which is then described explicitly through Algorithms 1 and 2. We will then show a statistical guarantee showing that, under certain noise assumptions, the algorithm performs well when the number of peaks in the data is small and sufficient data is available.

A. Algorithmic Intuition: Single Peak With No Noise

To begin our algorithmic intuition, we first consider the simplest possible setting for learning a function class of the form in (1), which is the noiseless case with \( r = 1 \). That is, we first assume the observations have the form

\[
N(t) = e^{-at^2 + bt + c} .
\]

Optimizing the loss function in (6) in this simplified setting still remains non-convex in general, so the goal is to find a reasonable initialization point from which to compute gradient descent. While the final goal is to learn the parameters \( a, b, \) and \( c \), we note that the learning procedure is more intuitive if we rewrite the form of the Gaussian as

\[
N(t) = Me^{-a(t-C)^2} ,
\]

where \( M = e^{c+\frac{b^2}{2a}} \) and \( C = \frac{b}{2a} \). This re-writing allows for the parameters to take on specific meanings: \( M \) represents the maximum value attained by the Gaussian, \( C \) represents the time at which the maximum is reached, and \( a \) represents a curvature of the Gaussian.

Hence, in order to learn the parameters of a single Gaussian curve in the noiseless case, we can first estimate \( \hat{M} \) as the maximum of the observed counts, \( \hat{M} = \max_t N(t) \). \( \hat{C} \) would then be estimated as the value of \( t \) for which the peak is reached, \( \hat{C} = \arg \max_t N(t) \). To learn \( a \), we note that the function

\[
D(t) = \log \frac{N(t+1)}{N(t)} - \log \frac{N(t)}{N(t-1)} \tag{7}
\]

is precisely equal to \(-2a\) in this noiseless, single peak case. So, we estimate \( \hat{a} = -D(1)/2 \) in this noiseless case.

To convert from \((\hat{a}, \hat{M}, \hat{C})\) to \((\hat{a}, \hat{b}, \hat{c})\), we note that in the form of the Gaussian, given the definitions of \( M \) and \( C \), we can solve for \( b \) and \( c \). Thus, we can estimate the parameters of the quadratic as

\[
\hat{a} = \hat{a} , \quad \hat{b} = 2\hat{C}\hat{a} , \quad \hat{c} = \log \hat{M} - \hat{C}^2\hat{a} .
\]

This initialization may not be perfect in the case where the true maximizing value \( C \) does not lie on an index \( t \) which is observed, as for example the maximizing value \( C \) could be between two integers. Hence, we apply gradient descent from this initialization point to find a local minimum which is a best fit for the Gaussian curve, which often recovers the true parameters of the Gaussian.

B. Algorithmic Intuition: Multiple Peaks With No Noise

The above handles the simplest possible case for learning, where \( r = 1 \). We next provide intuition to extend to the cases where \( r > 1 \). To begin, we consider the noiseless case where the observations come from a mixture of two Gaussian curves,

\[
N(t) = \sum_{k=1}^{2} e^{-a_k t^2 + b_k t + c_k} .
\]
The intuition for the learning procedure is as follows: We first identify the “mid-point” at which the two Gaussians have a similar number of cases, and then use this as a separating point to learn one Gaussian at a time.

To identify this mid-point, we make an observation from the computation of \( D(t) \) in this idealized case as shown in Fig. 2. Namely, we note that in regions where the two clusters have similar counts, \( D(t) \) increases to reach a local maximum.

In our algorithm, we exploit this local maximum in order to identify the midpoint between Gaussian components, creating disjoint intervals of time in which each interval corresponds to a single dominant Gaussian curve. Once these midpoints are defined, the problem is reduced to identifying the parameters of the dominant Gaussian components in each interval. As our theoretical results will indicate, the task of identifying such parameters is simple as long as the Gaussian curves are well-separated.

After determining the initialization parameters for each Gaussian, we again use gradient descent to optimize parameters due to the discrete nature of the observations. In the case of more than one Gaussian, we perform an alternating minimization, as the procedure works well to learn separated components of the mixture.

C. Algorithmic Intuition: Multiple Peaks With Noise

In order to account for the fact that there may be noise in observations, we average across multiple observations of \( D(t) \) to estimate \( a_k \), rather than use a single data point for estimation. In practice, this allows for a better estimation of the curvature of each Gaussian component in the mixture.

Moreover, the last question which remains is to determine the number of mixtures \( r \) used in the mixture. To determine a value of \( r \) which fits the data well but can be used in out-of-sample prediction, we use the BIC criterion [15, 39], and select the number of peaks as

\[
r^* \in \arg\min_r (3r) \ln T - 2 \ln L_r \left( \left\{ \hat{a}_k, \hat{b}_k, \hat{c}_k \right\}_{k=1,\ldots,r} \right).
\]  

(8)

Here, \( \{\hat{a}_k, \hat{b}_k, \hat{c}_k\}_{k=1,\ldots,r} \) represent the \( 3r \) learned parameters from the procedure outlined above. This allows us to justify that the model fits the data well without using too many parameters indicative of overfitting.

Our algorithms are formalized using pseudo-code in Algorithms 1 and 2. Specifically, we run Algorithm 1 followed by Algorithm 2 for several choices of \( r \), and then select a best choice of \( r \) using the BIC criterion in (8). In the next section, we provide theoretical results which indicate the effectiveness of the algorithm subject to bounded noise.

D. Parameter Estimation of the Reduced Form Model

We are able to provide a provable guarantee for Algorithm 1 for the case \( r = 2 \), and are able to extend to the case \( r > 2 \) so long as only two peaks are non-negligible at any given time. First, we recall that (1) can be written in the following form for \( r = 2 \):

\[
N(t) = \sum_{k=1}^{2} M_k e^{-a_k (t-C_k)^2} (1 + \varepsilon_t), \quad k = 1, 2.
\]  

(9)

Without loss of generality, we will let \( C_1 \leq C_2 \).

To present our theoretical result, we will make the assumption that \( a_1 = a_2 \), which will aid in clarity. This assumption can be relaxed, as is discussed in the Supplementary Material, and requires a more intricate bound on the size of the noise \( \varepsilon_t \) as well as an assumption that the proportion of cases observed from Community 2 is increasing on the observed interval, which are both reasonable in this context. The additional assumptions on underlying parameters are justified after the theorem statement, which is as follows.

**Theorem III.1 (Parameter Estimation Bounds):** Suppose \( a_1 = a_2 \), and suppose the parameters of the mixture model and

---

1This can be done using, e.g., a peak-finding algorithm from scipy [45]. The quantities \( x[1, \ldots, r-1] \) then represent the “midpoints” between individual Gaussian curves.
the bound on the noise $\delta$ satisfy the following properties for some $0 < \epsilon < \min\{M_1, M_2\}/5$, and $a, M > 0$.

1. $M_k \leq M$, $k = 1, 2$.
2. $a_k \geq a$, $k = 1, 2$.
3. $|C_1 - C_2| \geq 2\sqrt{\frac{2}{a_k} \log \frac{M}{\epsilon}}$.
4. $\delta \leq \delta'(a_1, C_1, C_2, M_1, M_2)$, defined as in (31) in the Appendix.
5. $C_1, C_2 \in [0, T]$.

Denote $\hat{M}_k$ as the estimate in line 6 and $\hat{C}_k$ as the estimate produced in line 7 of Algorithm 1. Then, for $k = 1, 2$

$$M_k [1 - \delta] e^{-a_k (|C_k| - C_k)^2} \leq \hat{M}_k \leq [M_k + \epsilon] [1 + \delta], \quad (10)$$

$$|\hat{C}_k - C_k| \leq \frac{1}{a_k} \log \left( \frac{M_k}{M_k e^{-a_k (|C_k| - C_k)^2} (1 + \delta)} \right) - \epsilon \leq \frac{1}{a_k} \left( \frac{1 - e^{-a_k (|C_k| - C_k)^2}}{e^{-a_k (|C_k| - C_k)^2} (1 + \delta) \frac{1}{1 + \delta} + \epsilon} \right). \quad (11)$$

**Proof:** Sketch. We provide a sketch of the three major steps of the proof, which are discussed in Section A of the Supplementary Material. The proof of this claim follows in three steps. First, we show that Assumptions 1–3 result in a condition where, if one component of the mixture model is dominant, then the other is at most $\epsilon$. Next, we show that because the noise is small due to Assumption 4, the value defined as

$$t_m = \operatorname{argmax}_{1 \leq t \leq T - 1} D(t),$$

which corresponds to the output of Line 2 in Algorithm 1, identifies an appropriate point which constitutes a midpoint between the two components of the mixture. That is, the estimate of $\hat{C}_1$ will be in a range where component 1 comprises the majority of cases, and $\hat{C}_2$ will be selected at a point in time where component 2 comprises the majority of cases. Finally, we combine these results to show that, because the contribution of the non-dominant component is at most $\epsilon$, and the estimates of $\hat{M}_k$ and $\hat{C}_k$ are associated with the correct component $k$, that the claim holds.

By providing parameter estimation bounds for Algorithm 1 in terms of epidemiological parameters $a_k$, $M_k$, and $C_k$, Theorem III.1 presents the first interpretable estimation bounds for learning the model in (1). In contrast, previous results on non-convex estimation in general, such as those discussed in [12] which apply to the methods used by Singhal et al. [41] to learn the model in (1) as a deterministic trend function, provide guarantees on the optimization objective rather than guarantees on the parameter estimates. Further, the results here differ from learning mixture models of probability distributions (c.f. [14], [40], [44]) as we learn directly from the time series as opposed from observed samples of a distribution.

**Interpretation of Assumptions:** The first condition of this theorem is straightforward, as it states that the number of daily cases in each mixture must be bounded. The second condition then states that each Gaussian component must have some curvature, i.e. each component can not be too flat. The third condition is that of temporal separation, and requires that the two components of the mixtures are spread apart in time. Such a condition is common to the study of identification of mixtures, as separation is key to acknowledging that the object of study should be described as distinct components forming a mixture (see, e.g. [14]). In the identification of distributions, separation is usually determined as a function of mean and variance of component distributions, whereas here the main requirement is temporal separation. The fourth condition simply requires that the observation noise is not particularly large. Moreover, the bound on $\delta$ is non-decreasing in $|C_1 - C_2|$, such that the temporal separation condition allows for the algorithm to become more robust. Finally, the fifth condition simply requires that for these bounds to hold, the peaks must be observed in the data.

**Example on US COVID-19 Case Data:** The conditions outlined above are reasonable to assume in practice: For $a = 0.0005$, which roughly corresponds to each Gaussian curve having 95% of total cases in the individual outbreak occur over roughly 100 days, and $M = 300,000$, which is approximately the highest number of daily COVID cases seen in the United States as of September 2021, we find that Assumption 3 of the Theorem requires that $|C_1 - C_2| \geq 105$ for $\epsilon = 13,000$. That is, under conservative estimates on the parameter bounds, and an $\epsilon$ which is an order of magnitude lower than the maximum number of cases $M$, the distance between peaks must be approximately 3.5 months, is reasonable as the first two local maxima in daily new cases in the United States were April 9th and July 24th, which have 106 days between them. For Assumption 4, estimates using the parameters above yields that $\delta$ may not exceed approximately 0.02%, which is restrictive, but becomes reasonable when case counts are large. Moreover, as we show in the following section, $\delta$ can be as large as 5% in synthetic experiments and the combination of Algorithms 1 and 2 can find parameters which fit the data well, so long as the midpoint between the two components is determined reasonably.

Ultimately, our results show that parameter estimation bounds can be achieved which depend on the size of the overlap between the two Gaussian components of the mixture as well as the magnitude of the noise $\epsilon$. Even with this variability in parameter estimates from the initial results of Algorithm 1, we find empirically that after the alternating minimization process of Algorithm 2 is performed, that the empirical results of the algorithm are surprisingly close to observed data, both in and out of sample. These results are discussed in the following section.

**E. Synthetic Performance of the Mixture Approach**

We first show the performance of the algorithm on a synthetic dataset. We provide a synthetic example of a time series with the following parameters:

$$(a_1, M_1, C_1) = (0.0005, 250000, 50),$$

$$(a_2, M_2, C_2) = (0.0007, 300000, 170),$$

with 200 observations from $t = 0$ to 199, and noise $\epsilon_t$ which is uniform over the interval $[-0.05, 0.05]$. Fig. 3 highlights the value of the initialization from Algorithm 1 over an arbitrary
initialization selected at random. For the random baseline, we select \( a_1 \) and \( a_2 \) uniformly at random from \([0, 0.001]\), \( M_1 \) and \( M_2 \) uniformly at random from \([0, 300000]\), and \( C_1 \) and \( C_2 \) uniformly at random from \([0, 100]\) and \([100, 200]\), respectively. For both types of initialization, we then apply Algorithm 2, which performs gradient descent in an alternating fashion to minimize the cost function. These results hold over multiple random initializations. The \( R^2 \) score (coefficient of variation) from using the initialization from Algorithm 1 results in \( R^2 = 0.984 \).

In contrast, against the baseline of a random initialization run over 1,000 different initializations drawn as above, the median \( R^2 \) from this baseline model is \(-2.43 \times 10^5\), with none of the 1,000 initializations resulting in a fitted model which provides better performance than the initialization from Algorithm 1.

### F. Empirical Performance of the Mixture Approach

Fig. 4 shows the utility of our forecasting method. The median absolute percent error (MAPE) of our approach is 15.9%, for one week forecasts, compared to a median of 18.7% for the same metric across all models used by the COVID-19 forecasting hub. For two week forecasts, the MAPE of our approach is 20.6% compared to a median MAPE of 25.9% across other models in the forecasting hub. Hence, in the short term we find that our model has low median forecasting error. In general, using the BIC criterion to select \( r \), we find that our method is competitive in terms of forecasting COVID-19 cases, while not requiring any auxiliary data such as mobility or an excessive computational burden. In practice, we also compute error bounds for our predictions by estimating the variance of \( \epsilon_t \) in (1). Specifically, using the learned values of \( \tilde{a}_k, \tilde{b}_k, \tilde{c}_k \) in (1), we are able to estimate realizations of \( \epsilon_t \) to compute its variance. This procedure amounts to an additional assumption that \( \epsilon_t \) is stationary, and is useful in practice to quantify the uncertainty of the approach.

### G. Limitations

Our methodology relies on a non-mechanistic approach, and implicitly makes assumptions on human response to the pandemic, as discussed in the following section. When such assumptions are not satisfied, as is the case when there appears to be a new outbreak of cases emerging, our method does particularly poorly. As discussed in the following section, the quadratic term of this model makes an implicit assumption that individuals are continually reacting to the spread of the pandemic by removing contacts with one another. Allowing the quadratic parameter \( a_k \) to be time-varying is an important extension of this model, as illustrated by [25] in the context of a feedback model. However, given that each initial wave of the pandemic has roughly followed a Gaussian shape (Fig. 1), we feel that the model in (1) can be empirically useful for epidemic modeling.

The assumptions of this model can also be violated when there is little evidence of a new cluster within existing data and \( r \) can not be selected appropriately. Hence, in the initial stages of a new waves of the pandemic, our approach tends to deteriorate particularly in terms of long term performance. Indeed, these events skew the accuracy of the model for such periods, resulting in a mean average percent error of 34% (23rd of 34 available models) compared to the median percent error of 15.9% (14th of 34 available models). Such an issue provides clear avenues for future research, as predicting when a new wave of cases may occur is a critical problem in understanding the way in which epidemics spread.

### IV. GENERATIVE BASIS FOR MIXTURES

The results of the previous section highlight the benefits of using the simple, mixture-based model in (1) in COVID-19 modeling. The model has a clear basis in the epidemiological literature, as Farr’s law has been used to forecast the progress of epidemics since the 19th century, and still remains in use for forecasting drug mortality and the spread of infectious disease [13], [18], [38]. Moreover, the use of mixtures and the free parameter \( r \) selected for the model allows it to be non-parametric, such that it can adapt to the multiple waves of the pandemic which have been observed empirically in the case of COVID-19. While the model in (1) provides an intuitive mixture model, can be learned efficiently from data, and performs well empirically, it does not immediately provide interpretation in the same way as traditional mechanistic models, as it does not explicitly encode epidemiological parameters such as the infectiousness of disease into its parametric form.
To this end, we introduce a network model that captures population heterogeneity, incorporates traditional epidemiological dynamics, and provides a mechanistic justification for the model in (1). Our analysis consists of a tight statistical characterization of this stochastic model which provides the bridge between mechanistic models and the non-mechanistic model of (1). In Section IV-C, we then provide empirical evidence which lends validity to the generative model.

The model in (1) has two key features which we wish to explain through a generative model: the Gaussian shape of each component and the summation which allows for multiple components to be present in the data. First, we show how the Gaussian components in each term of the sum can arise when the individuals in the social network partake in a process we term degree pruning, which refers to a process by which individuals sever ties with one another over time. As we show, when degree pruning occurs at a constant rate and the population is sufficiently large, the case counts in a single community precisely follow a Gaussian shape. To show this result for a single component, we restrict to the case where the graph consists of a single community as modeled by an Erdős-Rényi graph and provide a detailed probabilistic analysis of the generative model in this case. Next, we show how the generative model can explain the $r$ separate components in the mixture model. Namely, by allowing the underlying graph structure to be drawn from a stochastic block model which explicitly encodes community structure, we are able to show that the observed time series can consist of multiple distinct components which each take the form of a Gaussian curve.

A. Justification of Gaussian Components

In this section, we provide a simple generative theoretical model which results in the Gaussian shape described in (1). We restrict to the case $r = 1$ for exposition, and consider the extension to $r > 1$ in the following sub-section.

Model Description for Single Community: The model begins with a set of nodes $V$, where each node $v \in V$ represents an individual within the community, and for notation we let $|V| = n$ represent the number of nodes. We consider the progression of infection among nodes in $V$ across discrete time steps $t = 0, 1, \ldots$. At each time $t$, an individual is either susceptible, infected, or recovered. We denote $S(t) \subseteq V$, $I(t) \subseteq V$, and $R(t) \subseteq V$ as the sets representing susceptible, infected, and recovered individuals, respectively. At each time $t$, we have that the sets $S(t)$, $I(t)$, and $R(t)$ are disjoint, and $S(t) \cup I(t) \cup R(t) = V$. That is, the model has an equivalent representation as the network-based models of Section II-A2. In our model, at $t = 0$, we assume the initial conditions

$$I(0) = \{v_0\}, \quad R(0) = \emptyset, \quad \text{and} \quad S(0) = V \setminus \{v_0\},$$

i.e. that there is a single initially infected node in the graph and that the remaining nodes are susceptible.

Next, we describe the mechanisms by which infection can spread between nodes. In this model, infection spreads based on edges which exist between pairs of nodes in $V$. For analytical purposes and clarity of presentation, we consider a model of deferred randomness, in the sense that at time $t$, the only edges which are revealed will be those associated with nodes in $I(t)$. For example, at time $t = 0$, the edges revealed will be those associated with $v_0$, as visualized in the first panel of Fig. 5. Edges are revealed between infected and susceptible nodes based on a probability which is time-varying, and we assume that the presence of each is determined independently of all other edges.

Specifically, for any node $v_i \in I(t)$ and any node $v_s \in S(t)$, the probability that an edge forms between $v_i$ and $v_s$ at time $t$ is equal to $\frac{\rho}{d_i}$, where we recall $n$ is the number of nodes, $0 < \rho < 1$ captures the aforementioned notion of degree pruning, and $d_i$ is a parameter which captures a notion of average degree measure of the graph, in the absence of any degree pruning. The term $\rho$ reflects the notion that at each time step $t$, each edge which would have been revealed is removed from the graph with probability $1 - \rho$. This introduction of the parameter $\rho$ is precisely what will allow for the quadratic term in the Gaussian form of (1) to appear in the analysis, as otherwise one would expect pure exponential growth in this model in initial stages when $n$ is large.

With the randomness of the graph structure presented, we now specify the epidemiological parameters associated with the infectiousness of the disease and the rate at which individuals who are infected become recovered. We assume there is a parameter $0 < \beta \leq 1$ such that, if there is an edge between an infected individual and a susceptible individual at time $t$, infection spreads with probability $\beta$. Hence, for any node $v_i \in I(t)$ and any node $v_s \in S(t)$, the probability that $v_i$ infects $v_s$ becomes exactly $\frac{\beta}{2} \rho d_i$.

In the model, we also assume that if an individual is infected at time $t$, then they become recovered at time $t + 1$. This assumption, which states that each infection lasts for a single time epoch, is no more restrictive than the assumption that nodes cure themselves with a constant probability at each time step, as is often considered in the literature (cf. [7], [16], [37]). In such literature, the assumption of a constant probability of cure at each epoch results in a distribution of infection time

$$\text{Fig. 5. Visualization of Degree Pruning with Deferred Randomness. At each time } t, \text{ the edges associated with infected nodes (red) are revealed, beginning with a single infected node at time } t = 0. \text{ After a single epoch, each infected node becomes recovered (black). As } t \text{ increases, individuals in the population are also continually severing ties, as indicated by the dashed red edges which represent those edges which would have been present had they not been pruned. For simplicity, this visualization assumes } \beta = 1, \text{ i.e. if there is an edge between an infected node and a susceptible node, the susceptible node will become infected at the next time step.}$$
which is memoryless, and is often not the case for infectious diseases. Here, by making each epoch last between one and two weeks, one can model the case in which the infectious period lasts for a constant amount of time. This still results in an approximation, but appears to perform well in practice as was shown in Section III.

With the model defined for a single community, we now move on to describe the results of the generative model, which provides the link between the Gaussian shape of (1) with the notion of degree pruning. Our first result exactly characterizes this connection, as we show that in expectation, the number of infected individuals in the model will follow the shape of the Gaussian curve.

In Theorem IV.1, we make use of $O$ notation with respect to $n$, in which we say, for two functions $f$ and $g$, $f(n) = O(g(n))$ if there exists an $m > 0$ and a value $n_0$ such that for all $n \geq n_0$, $f(n) \leq mg(n)$. That is, we provide results which hold for each $n$, and indicate that our results become apparent as the size of the population becomes large. Hence, our result relates to previous work which analyzes the SIR model in cases where the size of the susceptible population is large (cf. [32]). In standard SIR models, a key assumption is that decay of the susceptible population size results in diminishing rates of infection [29]. Here, we instead assume that it is human behavior which also “flattens the curve”, as is made apparent by this assumption that the population size becomes arbitrarily large compared to the number of infections.

**Theorem IV.1:** Define the quantity $N(t) := |I(t)|$, which represents the number of infected individuals at time $t$. Then,

\[
\mathbb{E}[N(t)] = e^{(\frac{1}{2} \log \rho) t^2 + (\log \frac{d}{\sqrt{\beta}}) t} - O \left( \frac{t^2 e^{\log^2 (d/\sqrt{\beta})} / \log (1/\rho)}{n} \right),
\]

where we recall that $\rho \in (0, 1)$ reflects the rate of degree pruning, $d$ reflects the average degree of the graph in the absence of any degree pruning, and $\beta$ is the probability that infection travels across an edge in the graph.

**Proof:** Sketch. The claim follows inductively by computing the expectation of $\mathbb{E}[N(t+1) \mid N(t)]$, with the base case that $\mathbb{E}[N(0)] = 1$. Formal algebraic details can be found in Section B.1 of the Supplementary Material.

Theorem IV.1 is critical to the connection between mechanistic and non-mechanistic models of epidemics. Farr’s law has been used since the mid-19th century to estimate epidemics, and here we see that there is a generative explanation which gives the parameters of the Gaussian form a mechanistic interpretation. Namely, the quadratic term in the exponent of the Gaussian represents the extent to which individuals are reacting to the progress of the virus, and the linear term in the exponent of the Gaussian represents a reproduction rate. A similar relationship has been noted in [38], and Theorem IV.1 provides a formalization of this prior argument in the context of a network model as well as a mechanistic interpretation of the parameters of the Gaussian curve. We later validate this interpretation using mobility data in Section IV-C. By introducing a simple mechanism by which individuals can respond to the progress of an epidemic, the Gaussian form becomes apparent, and the parameters of the Gaussian can be interpreted as well.

Our statistical analysis also details concentration results characterizing the number of cases for each time $t$.

**Lemma IV.2:** For the model above, again define $N(t) := |I(t)|$ as the number of cases in the model at time $t$. For any $\epsilon > 0$, and any $t$,

\[
N(t) \leq e^{\left(\frac{1}{2} \log \rho\right) t^2 + \left(\log \frac{d}{\sqrt{\beta}}(1 + \epsilon)\right) t},
\]

with probability at least $1 - \sum_{s=0}^{t-1} e^{-\min\{\epsilon^2, \epsilon\} d \beta r^s / 4}$.

**Proof:** Sketch. This claim follows inductively by showing that, conditional on data up to time $t$, $N(t + 1)$ follows a Poisson distribution. Hence the tail probabilities of large values of each $N(t + 1)$ can be bounded. Details of the proof can be found in Section B.2 of the Supplementary Material.

The Lemma suggests that it is likely that the number of cases differs from its expectation by what is effectively a change in contact rate, since only the linear term of the exponential is changed. This provides a formal high-probability bound on the expected number of cases, and still suggests a bound in the form of a Gaussian curve, further expanding the work of [38] which considers a deterministic formulation which leads to the Gaussian time series. For this particular claim, when $t$ tends towards infinity, the bound on the probability becomes vacuous as $\rho^t$ tends towards 0 and hence each term in the summation tends towards 1. However, for large enough $t$, the expectation of $N(t)$ tends towards 0 regardless due to the Gaussian shape, which is to say that a simple Markov inequality can be utilized instead for sufficiently large $t$. This regime, in which expected case counts are low, is not the focus of this work as we are primarily concerned with regimes in which number of cases are sufficiently large. In the following section, we show that the result of Lemma IV.2 is indeed non-trivial, by utilizing the result to show that distinct Gaussian components are likely to be observed when the model is generalized to handle multiple communities.

Ultimately, these statistical results show that (1), in the special case where $r = 1$, characterizes the outcome of this natural stochastic process which encodes human behavior into the traditional networked SIR framework.

**B. Justification of Mixture Structure**

In this section, we provide the full generative model which justifies the general (1) for the case $r > 1$. The primary difference between this model and the model of the previous section is the set of connections in the underlying graph. Here we assume that the underlying graph of connections is drawn from a stochastic block model, which allows for community structure to be encoded into the model.

With the introduction of this community structure, we see that each individual community will experience a Gaussian curve in expectation, stemming from the initial infected individual within each community. The primary result of this section is to show that if the number of connections between each community is sufficiently small, then it is likely that the Gaussian curves from
each community will be well-separated, resulting in observations according to (1) for \( r > 1 \). We first present the details of the model and modification from the previous section, and then present our theoretical results.

**Description of Connections for Multiple Communities:** In the model which accounts for community structure, we first decompose the vertex set \( V \) into two disjoint sets \( V_1 \) and \( V_2 \), such that \( V_1 \cup V_2 = V \), \( V_1 \cap V_2 = \emptyset \). To simplify notation, we will assume \( |V_1| = |V_2| = n \), such that the total number of nodes in the graph is now \( 2n \). We refer to each set \( V_1 \) and \( V_2 \) as communities, as they will represent groups of nodes with a high likelihood of knowing one another. We will assume an initial condition that \( I(0) = \{ v_0 \} \), where \( v_0 \in V_1 \) without loss of generality, \( R(0) = \emptyset \) and \( S(0) = V \setminus \{ v_0 \} \).

For this model, we again assume a model of deferred randomness where the edges of the graph are revealed according to the individuals who are infected. However, for an infected individual in \( V_1 \), the probability of an edge forming with a susceptible individual in \( V_2 \) will be different than the probability of forming a connection with an individual in \( V_2 \). Specifically, for given \( v_i \in I(t) \) such that \( v_i \in V_1 \), and a node \( v_s \in S(t) \), then the probability there is an edge between \( v_i \) and \( v_s \) at time \( t \) is equal to \( \frac{d_{in}}{\sqrt{n}} \rho^t \) if \( v_s \) is also a member of \( V_1 \), and \( \frac{d_{out}}{n} \rho^t \) if \( v_s \) is a member of \( V_2 \). Here, \( d_{in} \) and \( d_{out} \) are parameters which encode the average number of neighbors within and outside a node’s community, respectively, in the absence of any degree pruning. In the case where \( d_{out} \ll d_{in} \), we see that there is community structure in the underlying graph (see Fig. 6). For infection within \( V_2 \) itself, we will assume that degree pruning only begins in \( V_2 \) after a first node is infected in \( V_2 \). That is, if we let \( T \) represent the random variable denoting the time at which the first node in \( V_2 \) is infected, then for nodes \( v_i \in I(t) \cap V_2 \) and \( v_s \cap V_2 \), the probability \( v_i \) would infect \( v_s \) would be equal to \( \frac{d_{in}}{n} \rho^{-T} \). This feature of the model reflects that communities are not expected to react to the virus until it becomes an immediate threat to the individuals within the community.

The primary result of this section shows that, when the stochastic block model is parameterized such that the number of edges between communities is sufficiently small, the Gaussian curves from each community will become temporally well-separated, resulting in a mixture for the observed time series of infection. The theorem is stated as follows.

**Theorem IV.3:** Define the following quantity, which represents the time at which the expected number of cases in community 1 is maximized:

\[
C_1(d_{in}, \beta, \rho) = -\log \left( \frac{d_{in} \beta}{\sqrt{\rho}} \right) / \log \rho .
\]

Suppose that, for some parameter \( \delta \in (0, 1/2) \), the following conditions hold:

1) The time at which expected number of cases in community 1 is maximized has the bound

\[
C_1(d_{in}, \beta, \rho) < \delta e^{T} + \log 20 / \log \frac{1}{\rho} . \tag{13}
\]

2) The expected number of connections between communities are not so large, specifically

\[
d_{out} \leq \frac{\log \left( \frac{1}{1-\delta} \right)}{2\beta \sum_{s=0}^{\infty} \left( \log \left( \frac{d_{in}}{n} \right) / \log \rho \right)^{-1} (2d_{in} \beta)^s \rho^{s(s+1)/2} . \tag{14}
\]

3) The number of individuals in each community is sufficiently large,

\[
n \geq 2\beta d_{out} . \tag{15}
\]

If the above conditions hold, then the random variable \( T \) which represents the time at which the first infection occurs in community 2 satisfies

\[
P \left( T > C_1(d_{in}, \beta, \rho) - \log 20 / \log \frac{1}{\rho} \right) \geq 1 - 2\delta . \tag{16}
\]

**Proof:** Sketch. The proof follows in two steps. First, we show that condition (13) results in an upper bound on the number of cases in the first community, which follows from Lemma IV.2. Next, we show that condition (14) then bounds the likelihood of infection in the second community, resulting in a lower bound on the likelihood that \( T \) is large. Formal proof details can be found in Section B.3 of the Supplementary Material.

Notably, the requirements of the theorem which result in temporal well-separation do not depend as much on the difference between \( d_{out} \) and \( d_{in} \), as much as it depends on the magnitude of each. In contrast, the identification of a stochastic block model depends on the difference \( d_{in} - d_{out} \) (c.f. [30]). Further, the result presented here provides a probabilistic guarantee on the amount of time we would expect between epidemics emerging in different communities, in contrast to results on deterministic models of epidemics which have community structure [43] or results based on simulations of epidemics on networks with community structure [42].

Theorem IV.3 indicates that, under certain model assumptions, with high probability the components of the mixture will be temporally well-separated, which is similar to the condition of identification required in Theorem III.1. From this theorem, we see that we can expect the observed components to be temporally well-separated if there are few connections between communities and the outbreak in community 1 is sufficiently
small. This observation highlights the different ways in which policy makers can prevent an outbreak from spreading between communities, as they can either focus on mitigating spread within their own community or ask members of a community with many infections to reduce their ties to other communities.

Example: Suppose that in the context of Theorem IV.3, we set \( \delta = 0.05 \), and choose the following parameters: \( \rho = 0.9 \), which represents that individuals remove approximately 10% of their contacts each epoch, \( d_{in} = 55.2 \), which represents expected contacts within the community in each epoch, \( \beta = 0.5 \), which represents the probability of infection given that a contact occurs, and \( d_{out} = 2 \times 10^{-5} \), which represents the expected number of contacts outside of the community. Then, \( C_1(d_{in}, \beta, \rho) \approx 32 \), and Theorem IV.3 shows that for any \( n \), with probability at least 90%, the time at which the first infection occurs in community 2 will be after at least 3.5 epochs, which occurs after about 10% of the time that it takes for cases in community 1 to reach their peak. In such a case, it would be likely to observe distinct mixtures.

The connection between the temporal well-separation in the generative model and the temporal separation condition required for learning can in fact be made precise, as shown in the following Corollary.

**Corollary IV.4:** Recall the definition of \( C_1(d_{in}, \beta, \rho) \), as

\[
C_1(d_{in}, \beta, \rho) = -\log \left( \frac{d_{in} \beta}{\sqrt{\rho}} \right) / \log \rho,
\]

and define the following random variable

\[
C_2^{d_{in}, \beta, \rho} = T - \log \left( \frac{d_{in} \beta}{\sqrt{\rho}} \right) / \log \rho = T + C_1(d_{in}, \beta, \rho),
\]

where \( T \) represents the time at which the first infection in community 2 occurs. \( C_2^{d_{in}, \beta, \rho} \) then represents the time at which the expected time series of infection in community 2 would be maximized. Then, under the conditions of Theorem IV.3, with probability at least \( 1 - 2\delta \), the following statement holds:

\[
|C_1(d_{in}, \beta, \rho) - C_2^{d_{in}, \beta, \rho}| \geq 2 \sqrt{\frac{1}{a} \log \frac{M}{\epsilon}},
\]

for parameters

\[
a = \frac{1}{2} \log \frac{1}{\rho}, \quad M = e^\frac{1}{2} \left( \log \left( \frac{d_{in} \beta}{\sqrt{\rho}} \right)^2 / \log \frac{2}{e} \right),
\]

\[
\epsilon \geq \frac{3}{8} - \frac{\log 20}{4} C_1(d_{in}, \beta, \rho) - \left( \frac{\log 20}{8} \right). \]

**Proof:** The Corollary follows directly from Theorem IV.3, as we see that \( |C_1(d_{in}, \beta, \rho) - C_2^{d_{in}, \beta, \rho}| = T \) since \( T \geq 0 \). Hence, with probability at least \( 1 - 2\delta \),

\[
|C_1(d_{in}, \beta, \rho) - C_2^{d_{in}, \beta, \rho}| \geq C_1(d_{in}, \beta, \rho) - \log 20 / \log \frac{1}{\rho}.
\]

Solving for \( \epsilon \) in the inequality, we find

\[
C_1(d_{in}, \beta, \rho) - \log 20 / \log \frac{1}{\rho} \geq 2 \sqrt{\frac{1}{a} \log \frac{M}{\epsilon}},
\]

with \( a \) and \( M \) defined above.

Corollary IV.4 highlights a temporal separation between the peaks of the expected time series, and makes explicit the relationship between the generative model presented here and the learning algorithm of Section III. Notably, the \( \epsilon \) obtained in the Corollary is nearly an order of magnitude lower than the value \( M \), which is desirable for the learning algorithm. We note that in Corollary IV.4, \( a \) is defined to be the quadratic coefficient in the expected time series and that the quantity \( M \) represents a maximum number of cases in the expected time series of infection for each community, specifically representing the maximum number of infections in \( V_1 \) due only to other infections which originated in \( V_2 \). We choose these particular expressions for clarity of presentation, as it could be the case that the maximum number of infections in \( V_1 \) can become larger due to the stochastic nature of the process within \( V_1 \) or due to additional cases in \( V_1 \) which can be traced to infections from individuals in \( V_2 \), after individuals in \( V_2 \) receive infection. Rather, the purpose of Corollary IV.4 is to highlight the connection between temporal well-separation in this model and the previous condition required for learning.

Taken together, these results provide a statistical characterization of a reasonable stochastic model, and motivate the use of the function class in (1). Of course, this is not the only possible model which results in observations which resemble a mixture of Gaussian curves. Rather, the above model provides but one formalization by which case counts of the form in (1) can arise, and we prefer this model for its tractability in analysis. In particular, the model in (1) may also arise from an SIR-model with a particular time-varying reproductive rate as we will show in Section V, or in other network-based models with a particular degree distribution which emulates the degree pruning parameter above. Even situations in which the spread has spatial heterogeneity can be captured, so long as the simultaneous outbreaks have similar features resulting in global observations of the Gaussian curve, in which case the temporal separation of Gaussian curves may reflect different waves of the pandemic [17].

C. Empirical Validation of the Networked SIR Interpretation

While our model is not the only one which explains the function class in (1), we provide empirical evidence that the data observed during the COVID-19 pandemic is consistent with the simple interpretation provided above. Namely, we are able to show that “degree pruning” correlates with observed mobility data. To do so, we compare degree pruning parameters learned from the data to mobility data taken from Google and SafeGraph. While these results do not provide causal evidence for the validity of our model, they are consistent with the interpretation of the generative model provided in Section IV and hence help to validate the interpretation of degree pruning. Indeed, we expect degree pruning strategies to be a multi-factorial concept, much like the classical reproduction number discussed in mechanistic models [23]. For example, degree pruning may be a function of unobserved depletion of the susceptible population through, e.g. vaccination or asymptomatic infections, for later parts of the pandemic.
We use two sets of mobility data in our analysis. The first, from Google, provides a time series showing percent reduction in six different types of mobility compared to a baseline measured in February 2020, before the United States government implemented any lockdown policies. The six types of mobility are RETAIL AND RECREATION, GROCERIES AND PHARMACY, PARKS, TRANSIT STATIONS, WORKPLACES, and RESIDENTIAL. The data is provided at county, state, and national granularity.

The second dataset, provided by SafeGraph, tracks the census block location of particular devices, and results are aggregated daily. We process the data to measure the amount of time, on average, that a device spends outside of its census block. Since census blocks are typically small compared to the size of counties, the data provides another metric by which we may measure the extent to which individuals limited their mobility in response to the pandemic.

For validation of the interpretation of parameters, we use an estimate of degree pruning parameters which are the output of Algorithm 1. We compute these values in the period corresponding from March 1st, 2020, to May 20th, 2020, when the national mask mandate in the United States was implemented, as in this period of time we would expect mobility to be the best proxy for degree pruning. Among each state in the US, at a national level there are clear expected relationships between mobility and degree pruning parameters. Change in mobility in terms of time spent in RETAIL AND RECREATION, TRANSIT STATIONS, and WORKPLACES, as well as mobility as measured by the average amount of time a device spends outside of its census block, appear to correlate negatively with degree pruning parameters, which is to say that as mobility decreases in these measures, degree pruning rates increase, as expected. This is validated with correlation coefficients ranging from $-0.248$ to $-0.430$, each of which is statistically significant. Moreover, mobility as measured by change in time spent in RESIDENTIAL places is positively correlated, implying that as individuals increase their time spent at home, degree pruning rates increase, as expected.

### V. Closed-Loop Interpretation of Gaussian Curves

While the model above provides clear insight into the structure of the mixture model in (1), a drawback of the approach is that it is open-loop. That is, the degree pruning parameter $\rho$ is assumed to be constant to generate the Gaussian form, and the model does not explicitly allow for individuals to react to the state of the disease. In this section, we show that the Gaussian shape can in fact arise as the outcome of a closed-loop system, in which individuals explicitly react to the spread of disease. Specifically, by using an approximation of an SIR model, the Gaussian curve can be justified by assuming individuals react in response to the observed number of infections in an epidemic.

The networked SIR model in Section IV assumes that the number of individuals who are susceptible remains large compared to the infected population, i.e. that $n$ is large, and also that degree pruning occurs such that contact rates and removal rates are time-varying. With these assumptions, we can approximate the continuous time Susceptible-Infected-Recovered model discussed in Section II-A as

$$\frac{dI}{dt} = (\beta(t) - \gamma(t)) I(t) . \quad (17)$$

Denoting $\alpha(t) := \beta(t) - \gamma(t)$, we then see that in order to ensure that $I(t)$ follows the shape of a bell curve, $\beta(t) - \gamma(t)$ must follow a specific form as formalized in the following Theorem.

**Theorem V.1 (Gaussian Curve as Closed Loop Control):** In the setting of (17), suppose a population can guarantee the following form for $\alpha(t) := \beta(t) - \gamma(t)$

$$\alpha(t) = \begin{cases} \sqrt{b^2 - 4a(c - \log I(t))} & \text{if } \max_{t' \leq t} I(t') < e^{c+\frac{b^2}{4a}} \\ -\sqrt{b^2 - 4a(c - \log I(t))} & \text{otherwise} \end{cases}$$

Then, if $I(0) = e^c$,

$$I(t) = e^{-at^2+bt+c} ,$$

**Proof:** The proof follows from first plugging in the form of $I(t)$ into the equation of $\alpha(t)$. Casework then yields

$$\alpha(t) = -2at + b ,$$

in both cases, which directly implies the claim by substitution into (17). Hence, since the solution to the differential equation is unique given the initial condition, it must be the case that $I(t) = e^{-at^2+bt+c}$.

Theorem V.1 shows that the evolution of the number of infections in a particular community can be written in terms of the number of infections itself, implying that community reaction to the progress of an epidemic may be stated in terms of the infection prevalence. In particular, by equating $M = e^{c+\frac{b^2}{4a}}$ as in Section III-A, we see that the population must have a certain control in place until the cases reach their maximum $M$, and then have the opposite reaction to the state afterwards. While the proposed control provides a roundabout way in which the Gaussian may arise, parameterizing the form of the control input ultimately provides insight on policy. In this setting, a policy maker may try to adjust the value of $\alpha(t)$ to ensure that the maximum number of cases does not exceed some value $I_{\max}$ by ensuring that the rate $\alpha(t)$ evolves according to a rule for which the constants $a, b$, and $c$ satisfy $c + \frac{b^2}{4a} \leq \log I_{\max}$. This result is similar to those in other work on state-dependent responses to epidemics, but makes a specific parametric assumption on the
way in which individuals react to the state of the epidemic [4], [27], [36]. This indicates another possible way in which the Gaussian form of (1) can be interpreted through the lens of policy decisions. Namely, if observed case counts follow the Gaussian trend, then a possible explanation for this is that policy decisions are being formed as a function of the state of the epidemic, as opposed to through a constant pruning of edges.

VI. CONCLUSION

We provide a simple, non-mechanistic model for forecasting epidemics which, upon further inspection, bridges the two major approaches to epidemic forecasting while retaining benefits of both. By assuming that observed case counts follow a functional form represented by a sum of Gaussian curves, we benefit from the reduced form structure of the model because we are able to perform statistical inference to measure the parameters of each Gaussian. Moreover, we show that a generative model can yield observed case counts of the form of (1), which provides a benefit often associated with mechanistic models in that the learned parameters have an interpretation.

A key observation from the mechanistic perspective of the model is that, in order to attain the quadratic term of the Gaussian time series, ties must be continually severed within communities at a constant rate, through what we refer to as “degree pruning.” This assumption that degree pruning remains constant is inherently restrictive and need not hold throughout all time for a population, particularly as the number of cases in an epidemic begins to decrease. As such, it is important to consider models in which a population’s reaction to the state of the pandemic is state-dependent as discussed in [4], [25], [36]. In particular, our related work [25] shows that a time-varying, state-dependent degree pruning parameter can provide another way in which multiple peaks can occur for an epidemic, and this approach is particularly useful for later stages of the pandemic in which the geospatial heterogeneity of cases becomes difficult to infer from a single time series. That being said, the assumption that degree pruning is constant can be particularly valuable for short-term forecasting – if we expect degree pruning will not change much within a short time-period, then we may expect a Gaussian curve to provide a reasonable function class for forecasting.

Ultimately, degree pruning is a multi-factorial concept which encodes both the number of contacts severed at each time and reduction of the probability of disease transmission. Hence, policy strategies include but are not limited to the use of social distancing, masks, testing, and vaccinations in order to remove links between individuals in the community faced with an epidemic. Social distancing has been one of the primary ways in which communities have implemented degree pruning throughout the COVID-19 pandemic. By initiating stay-at-home orders and posting signs in public areas that encourage individuals to stay apart from one another, there are fewer links between individuals in the communities by which the virus may spread. Masks have had a similar effect on the extent of degree pruning. As the pandemic has progressed, mask mandates have become mandatory in many areas, and have reduced the total number of COVID-19 cases [10]. Testing provides another means of degree pruning, by limiting the number of contacts an individual has once they are infectious. By allowing individuals to know that they are infected with the virus, they may then self-isolate and limit the number of contacts they will have with susceptible individuals. In particular, by increasing testing rates over time, the rate of degree pruning can be changed. Finally, vaccinations provide another simple mechanism by which degree pruning may be achieved. For example, vaccinating a constant fraction of the susceptible population at each time step has the exact effect of degree pruning, as edges in the graph will be deleted proportionally to the rate of vaccination. Ultimately, a combination of the above strategies, with increased utilization over time, may allow policy makers to better understand and combat the spread of infectious disease, when considering the evolution of an epidemic through the lens of degree pruning.

The overarching goal of this work is to provide progress towards robust, data-driven control of epidemics for general outbreaks of infectious disease. This will require refined statistical algorithms for estimating the state of an epidemic subject to noisy observations, as well as an understanding of the mechanisms that policy makers can utilize to inhibit spread of the disease. Hence, further research can focus on determining optimal ways to estimate epidemic state from noisy data due to delays and testing variance, as well as understanding the impact that policy levers such as lockdowns and masks have on the spread of infectious disease. Moreover, in this work we have restricted to epidemics which are appropriately modeled with Susceptible, Infected, and Recovered compartments. Future work may build on the principles discussed here to understand if other non-mechanistic models can be viewed as special cases of mechanistic SI, SIS, or SEIR models (c.f. [23]). While our results are motivated by the spread of infectious disease, we note that the general form of the model lends itself to applications in epidemics more broadly. Future lines of work may also consider other epidemic processes such as correlated failures in financial networks [2] or the spread of information on social media [46]. By unifying the disparate approaches to epidemic forecasting, we hope to take a step towards reaping the benefits of both approaches in the design of public policy.

REFERENCES

[1] D. Acemoglu, V. Chernoizhukov, I. Werning, and M. D. Whinston, “Optimal targeted lockdowns in a multigroup SIR model,” Amer. Econ. Rev. Insights, vol. 3, no. 4, pp. 487–502, 2021.
[2] D. Acemoglu, A. Ozdaglar, and A. Tahbaz-Salehi, “Systemic risk and stability in financial networks,” Amer. Econ. Rev., vol. 105, no. 2, pp. 564–608, 2015.
[3] A. Aleta et al., “Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19,” Nature Hum. Behav., vol. 4, no. 9, pp. 964–971, 2020.
[4] A. Bhimaraju, A. Chatterjee, and L. R. Varshney, “Expected extinction times of epidemics with state-dependent infectiousness,” IEEE Trans. Netw. Sci. Eng., vol. 9, no. 3, pp. 1104–1116, May/Jun. 2022.
[5] F. Brauer, C. Castillo-Chavez, and Z. Feng, Mathematical Models in Epidemiology, Berlin, Germany: Springer, 2019.
[6] D. J. Bregman and A. D. Langmuir, “Fan’s law applied to aids projections,” J. Amer. Med. Assoc., vol. 263, no. 11, pp. 1522–1525, 1990.
[7] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos, “Epidemic thresholds in real networks,” ACM Trans. Inf. Syst. Secur., vol. 10, no. 4, pp. 1–26, 2008.
[8] A. G. Chandrasekhar, P. Goldsmith-Pinkham, M. O. Jackson, and S. Thau, “Interacting regional policies in containing a disease,” Proc. Nat. Acad. Sci., vol. 118, no. 19, 2021, Art. no. e2021520118.

[9] X. Chan, J. Li, C. Xiao, and P. Yang, “Numerical solution and parameter estimation for uncertain sir model with application to COVID-19,” Fuzzy Opt. Decis. Mak., vol. 20, no. 2, pp. 189–208, 2021.

[10] V. Chernozhukov, H. Kasahara, and P. Schrimpf, “Causal impact of masks, policies, behavior on early COVID-19 pandemic in the us,” J. Econometrics, vol. 220, no. 1, pp. 23–62, 2021.

[11] R. Dandekar, C. Rackauckas, and G. Barbashidis, “A machine learning-aided global diagnostic and comparative tool to assess effect of quarantine control in COVID-19 spread,” Patterns, vol. 1, no. 9, 2020.

[12] M. Danilova et al., “Recent theoretical advances in non-convex optimization,” in High-Dimensional Optimization and Probability: With a View Towards Data Science. Berlin, Germany: Springer, 2022, pp. 79–163.

[13] S. Durakly, J. E. Brady, C. J. DiMaggio, and G. Li, “Applying Farr’s law to project the drug overdose mortality epidemic in the United States,” Int. J. Epidemic., vol. 1, no. 1, 2014, Art. no. 31.

[14] C. Daskalakis, C. Tzamos, and M. Zampetakis, “Ten steps of EM suffice for mixtures of two Gaussians,” in Proc. Int. Conf. Learn. Theor., 2017, pp. 704–710.

[15] J. Ding, V. Tarokh, and Y. Yang, “Model selection techniques: An overview,” IEEE Signal Process. Mag., vol. 35, no. 6, pp. 16–34, Nov. 2018.

[16] D. Easley et al., Networks, Crowds, and Markets, vol. 8, Cambridge, U.K.: Cambridge Univ. Press, 2010.

[17] J. M. Epstein, J. Parker, D. Cummings, and R. A. Hammond, “Coupled contagion dynamics of fear and disease: Mathematical and computational explorations,” PLoS One, vol. 3, no. 12, 2008, Art. no. e3955.

[18] W. Farr, “Progress of epidemics,” Second Report of the Registrar General of England and Wales, pp. 16–20, 1840.

[19] T. R. Frieden and C. T. Lee, “Identifying and interrupting superspreading events–implications for control of severe acute respiratory syndrome coronavirus 2,” Emerg. Inf. Dis., vol. 26, no. 6, pp. 1059–1066, 2020.

[20] A. Ganesh, L. Massoulié, and D. Towsley, “The effect of network topology on the spread of epidemics,” in Proc. IEEE 24th Annu. Joint Conf. Comput. Commun. Soc., 2005, pp. 1455–1466.

[21] M. Girvan, D. S. Callaway, M. E. J. Newman, and S. H. Strogatz, “Simple model of epidemics with pathogen mutation,” Phys. Rev. E, vol. 65, no. 3, 2002, Art. no. 031915.

[22] P. João et al., “Forecasting COVID-19 cases based on a parameter-varying stochastic SIR model,” Annu. Rev. Control, vol. 51, pp. 460–476, 2021.

[23] H. W. Hethcote, “The mathematics of infectious diseases,” SIAM Rev., vol. 42, no. 4, pp. 599–653, 2000.

[24] I. Holmdahl and C. Buckee, “Wrong but useful–what COVID-19 epidemiologic models can and cannot tell us,” New England J. Med., vol. 383, pp. 303–305, 2020.

[25] A. Jadbabaie, A. Sarker, and D. Shah, “Implicit feedback policies for COVID-19: Why “zero-COVID” policies remain elusive,” Sci. Rep., vol. 13, no. 1, 2023, Art. no. 3173.

[26] N. P. Jewell, J. A. Lewnard, and B. L. Jewell, “Caution warranted: Using the institute for health metrics and evaluation model for predicting the course of the COVID-19 pandemic,” Ann. Intern. Med., vol. 173, 2020, pp. 226–227.

[27] M. Kantner and T. Koprucki, “Beyond just “flattening the curve”: Optimal control of epidemics with purely non-pharmacological interventions,” J. Math. Ind., vol. 10, no. 1, pp. 1–23, 2020.

[28] M. J. Keeling and K. T. D. Eames, “Networks and epidemic models,” J. Roy. Soc. Interface, vol. 2, no. 4, pp. 295–307, 2005.

[29] W. O. Kermack and A. G. McKendrick, “A contribution to the mathematical theory of epidemics,” Proc. Roy. Soc. London Ser. A, vol. 115, no. 772, pp. 700–721, 1927.

[30] F. Krzakala et al., “Spectral redemption in clustering sparse networks,” Proc. Nat. Acad. Sci., vol. 110, no. 52, pp. 20935–20940, 2013.

[31] M. Le, M. Ibrahim, L. Sagun, T. Lacroix, and M. Nickel, “Neural relational autoregression for high-resolution COVID-19 forecasting,” Facebook AI Res., 2020.

[32] J. Ma, “Estimating epidemic exponential growth rate and basic reproduction number,” Inf. Dis. Mod., vol. 5, pp. 129–141, 2020.

[33] C. M. Macal and M. J. North, “Agent-based modeling and simulation,” in Proc. IEEE Winter Simul. Conf., 2009, pp. 86–98.

[34] T. K. Moon, “The expectation-maximization algorithm,” IEEE Signal Process. Mag., vol. 13, no. 6, pp. 47–60, Nov. 1996.

[35] IHME COVID-19 Health Service Utilization Forecasting Team and C. J. Murray, “Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by us state in the next 4 months,” medRxiv, 2020.

[36] M. Ogura and V. M. Preciado, “Epidemic processes over adaptive state-dependent networks,” Phys. Rev. E, vol. 93, no. 6, 2016, Art. no. 062316.

[37] N. A. Ruhi and B. Hassibi, “SIRS epidemics on complex networks: Concurrence of exact Markov chain and approximated models,” in Proc. IEEE 54th Conf. Decis. Control, 2015, pp. 2919–2926.

[38] M. Santillana et al., “Relatedness of the incidence decay with exponential adjustment (IDEA) model, ‘farr’s law’ and SIR compartmental difference equation models,” Infect. Dis. Modelling, vol. 3, pp. 1–12, 2018.

[39] G. Schwarz, “Estimating the dimension of a model,” Ann. Stat., vol. 6, pp. 461–464, 1978.

[40] N. Segol and B. Nadler, “Improved convergence guarantees for learning Gaussian mixture models by EM and gradient EM,” Electron. J. Stat., vol. 15, no. 2, pp. 4510–4544, 2021.

[41] C. Stegehuis, R. Van Der Hofstad, and J. S. H. Van Leeuwaarden, “Epidemic spreading on complex networks with community structures,” Sci. Rep., vol. 6, no. 1, 2016, Art. no. 29748.

[42] M. Van den Driessche, “Spatial structure: Patch models,” in Epidemic spreading on complex networks with community structures,” Sci. Rep., vol. 3, pp. 1–12, 2018.

[43] X. Xue, “Law of large numbers for the sir model with random vertex degree,” in Proc. IEEE 24th Annu. Joint Conf. Comput. Commun. Soc., 2005, pp. 1455–1466.

[44] P. Virtanen et al., “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python,” Nature Methods, vol. 17, pp. 261–272, 2020.

[45] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” Science, vol. 359, no. 6380, pp. 1146–1151, 2018.

[46] X. Xue, “Law of large numbers for the sir model with random vertex weights on Erdős–Rényi graph,” Physica A: Stat. Mechanics its Appl., vol. 486, pp. 434–445, 2017.