# Convex Risk Minimization to Infer Networks from Probabilistic Diffusion Data at Multiple Scales

Emre Sefer [*1] and Carl Kingsford [*2]

[*] *Lane Center of Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA*

[1] esefer@andrew.cmu.edu

[2] carlk@cs.cmu.edu

***Abstract*—SEIR (Susceptible-Exposed-Infected-Recovered) is a general and widely-used diffusion model that can model the diffusion in different contexts such as idea spreading and disease propagation. Here, we tackle the problem of inferring graph edges if we can only observe a SEIR diffusion process spreading over the nodes of a graph. This problem is of importance in the common case where node states can be estimated with less cost than the edges can be found. Some applications include inferring a contact network from disease spread data, inferring a reference network from idea spreading, or estimating influenza diffusion rates between U.S. states. We improve upon the existing approaches for this problem in three ways: (1) we assume we are provided only with the probabilistic information about the state of each node which may also be undersampled or incomplete; (2) we present a more general framework that better uses trace data to model edge non-existence under SEIR model; (3) we can infer the network at both micro and macro scales.**

**Experiments on both real and synthetic data show that our method is accurate under these challenging cases at multiple scales, and it performs consistently better than the existing methods. For instance, we can infer a high school human contact network at microscale by tracking influenza diffusion almost $10\%$ better than the existing methods as well as the estimated networks closely mimick the full range of properties of the true network. We also estimated the strength of the influenza diffusion between and inside the U.S. states from Google Flu Trends data at macroscale. Estimated rates are correlated with the human transportation rates between the states to a certain degree, and we gain interesting insight into the influenza diffusion in U.S. such as the importance of the less populous states in epidemics as well as the asymmetric influenza diffusion between U.S. states.**

## I. Introduction

Networks are being heavily used to model and analyze the properties of various social and biological systems. The phenomenon of study is often modeled as a dynamic process spreading over the network. Diffusion is special case of those processes in which a spread (e.g., an infection) starts from some part of the graph and spreads to other portions over time via the edges of the graph. Some examples are virus spreading [1], and idea spreading over Twitter [2]. A diffusion model defines a set of possible states that the nodes of the graph can be in as well as rules for probabilistically switching between those states. SEIR [3], for example, is a well-known example of a diffusion model that is often used to simulate the spread of infection. Other widely-used SI, SIS, SIR models [3] are special cases of the SEIR model.

In many situations, it is easier or less costly to observe the states of the nodes than it is to observe the edges of the network over which the diffusion process is spreading. For instance, we might easily observe opinion diffusion on social networks but it may not be possible to see the network due to privacy. Similarly, it is difficult to measure the human contact network for flu transmission [1] but it is easier to detect whether people are ill. In other cases, we are also interested in understanding the diffusion characteristics at macroscale since it is infeasible and unnecessary to learn it on micro level (person-to-person contact). For instance, we are interested in estimating the rates of influenza diffusion between the U.S. states but not on the person to person details of this transmission. In this paper, we study the problem of inferring the unknown network when all we are able to observe are traces of how the states of node change as the diffusion spreads over the graph. In terms of influenza diffusion, unknown network models the contacts between humans at microscale, whereas it represents influenza diffusion rates between U.S. states at macroscale. Recovery of the transmission network is important in designing better epidemic containment strategies and better vaccination strategies.

We present *CORMIN* (COnvex Risk Minimization to Infer Networks) that addresses the problem of inferring the graph from the diffusion data in less idealized and more applicable settings. First, we explore the case that diffusion data is not perfectly known. This uncertainty in the diffusion data is interpreted differently in different contexts. For instance, when tracking the spread of a disease, measured symptoms such as headache and fatigue only partially reveal a node's state since they are not perfect representatives of diffusion states (infected, etc.). Further, the infected person does not suddenly start showing all the symptoms but instead severity of the symptoms increase progressively over time. In this case, we cannot perfectly know the diffusion times but rather estimate our degree of belief (confidence) of being at certain states. When estimating the influenza diffusion rates between the U.S. states at macroscale, probabilistic modeling is mandatory since the diffusion data is an ensemble over many people, and probabilistic data in each U.S. state is interpreted as the percentage of people infected with influenza in that U.S. state. Second, obtaining diffusion data is often expensive, so we may not know status of nodes at each possible time step but rather observe them with frequency lower than that at which the diffusion model is operating. Lastly, we infer networks from SEIR model and its special cases at both micro and macro scales.

Our main innovation to tackle these challenges is to treat

diffusion data for each node and each possible state as probabilistic time series. This is in contrast to the existing diffusion-based inference methods [4], [5], [6], [7], [8] for which a node is in each state with either probability $0$ or $1$. We formulate the graph inference problem as L1 regularized risk (expected loss) minimization program from SEIR dynamics. When the diffusion data is perfect, L1 regularization can be removed and *CORMIN* can be run nonparametrically by adding constraints that force at least a single edge to exist between a newly infected node and the previously infected nodes that are not yet recovered. We applied *CORMIN* to infer synthetic networks, high school human contact network at microscale, and to estimate influenza diffusion rates between U.S. states at macroscale.

*CORMIN* is capable of inferring the graphs under many challenging cases, and we found it to perform consistently better than the existing methods in almost all cases due to its probabilistic formulation even though we run the competing methods with their best parameters. Performance of *CORMIN* is not significantly affected by the probabilistic data whereas the existing methods performance decreases even though we apply a non-naive rounding scheme to pre-process the input to make schemes designed for $0/1$ probabilities work with more general probabilities. For instance, *CORMIN* can achieve $F0.1$ score around $0.7$-$0.8$ over a human contact network if the traces are the only prior information available about the graph. It can also nicely model and infer the influenza diffusion between U.S. states at macroscale that cannot be done by the existing methods. At macroscale, we found the influenza transmission rates between U.S. states estimated by *CORMIN* on Google Flu Trends dataset to be correlated with the human transportation rates between those states. Estimated diffusion rates between U.S. states are asymmetric, and the diffusion rates between less populous states are high especially when they are close to each other.

In summary, probabilistic modeling of the observed data, and the ability to model both edge existence and non-existence is the main reason *CORMIN* outperforms the other methods on both real and synthetic data under various challenging cases. In contrast to the existing methods, we may also use *CORMIN* to estimate the diffusion rates at macroscale via its probabilistic formulation. *CORMIN* still performs reasonably well when the noise dynamics parameters that map exact transition times to the observed diffusion data are also unknown. In this case, it can simultaneously estimate the noise dynamics parameters and infer the graph which cannot be done by the existing methods.

### A. Related Work

Many existing methods [3], [9], [10] model the influenza transmission by differential equations; they make a homogenous network assumption by ignoring the effect of the network structure in diffusion. However, this assumption is not valid for many diffusion types at both micro and macro scales. For instance, influenza spreads over human contact network, and this network is mostly heterogeneous. Similarly, influenza spreads between U.S. states at macroscale but the transmission rates between the states are not the same. Recently, some methods have been suggested to infer social networks from diffusion data. Among them, both NetInf [5] and MultiTree [7] formulate inference as a maximum likelihood problem in terms of only the edge existence, and ConNIe [4], NetRate [6], KernelCascade [11] and InfoPath [8] predict the edges by estimating the diffusion probabilities. Another network inference method makes a prior assumption about the scale-freenees of the network [12].

These methods have a number of shortcomings that we attempt to address here. They assume perfect knowledge of diffusion events, and neglect the possibility of partially observable, under-sampled probabilistic diffusion data. Further, they cannot model the uncertainty inherent in the diffusion data. Another shortcoming is their inability to estimate the diffusion rates at macroscale. In this case, existing methods cannot treat multiple nodes as a single ensemble node which is mandatory especially for large-scale networks. Lastly, we define the inference problem for arbitrary loss functions without making any prior assumption about the graph structure, and show that it can be solved optimally for certain type of loss functions.

Similar problems have been previously considered when collective statistics instead of individual statistics are available [13], [14]. For instance, collective graphical models are shown to be useful for estimating the bird migration paths given collective bird location data over time instead of individual positions [13] where they formulate inference as an extension of maximum flow problem. They also develop efficient approximate inference methods under more general collective graphical models [14]. However, these methods are based on flow conservation where latent nodes change position without changing their states over time by interacting with other latent nodes. Then, these methods cannot be directly applied to our problem of estimating the connectivity structure and influenza transmission rates at macroscale under SEIR.

## II. Problem Formulation

Let $G = (V, E)$ be an unseen graph for which the edges $E$ are difficult to observe directly. Edges of $G$ may represent human contact events, interactions in PPI, relationships in social network, etc. We assume a uniform prior over the edges $E$ since we do not have additional information about the graph structure, or the node attributes. At each time step, each node of $G$ can be in one of several *states* $\mathcal{S}$. These states represent an abstraction of the node's status with respect to a diffusion process such as the spread of a virus. The model $\mathcal{M}$ governs how a node's state changes based on the states of its neighbors at previous time steps. Here, we focus on the general and widely-used SEIR model: the states $\mathcal{S}$ are Susceptible (*S*), Exposed but not contagious (*E*), Infected and contagious (*I*), and previously infected but now Recovered (*R*). The SI, SIS, SIR models are special cases of the SEIR model in which some states and transitions cannot occur. Those states are general enough abstractions to model various forms of diffusion in different contexts [15], [1]. The SEIR is Markovian, and it obeys the independent cascade [16] assumption which states that a single diffusion from one of node's neighbors is enough for node to become exposed.

More formally, a *trace* $d$ of the SEIR diffusion process measured at time steps $T_d$ provides us with a set of probabilities $\{s_v^d(t),\ e_v^d(t),\ i_v^d(t),\ r_v^d(t)\}$ for every node $v \in V$ and every time step $t \in T_d\}$, where $x_v^d(t)$ is the probability that

| Symbol | Definition |
|---|---|
| $d$ | A single trace |
| $D$ | Set of diffusion traces |
| $T_d$ | Set of time points observed in $D$ |
| $s_v^d(t_j)$, $e_v^d(t_j)$, $i_v^d(t_j)$, $r_v^d(t_j)$ | Probabilities of $v$ being in *S*, *E*, *I*, *R* states in trace $d$ at time $t_j$ |
| $b$ | A perfect trace: $b = \{b_v,\ v \in V\}$ |
| $b_v$ | Perfect trace for node $v$: $b_v = \{t^b_{e,v}, t^b_{i,v}, t^b_{r,v}\}$ |
| $t^b_{e,v}$, $t^b_{i,v}$, $t^b_{r,v}$ | Exact time $v$ passes into *E*, *I*, *R* in trace $b$ |

TABLE I: Notation for problem definition

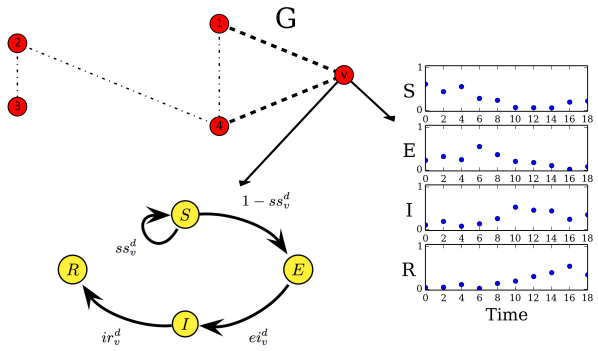



Fig. 1: Only S→E transition is being affected by $G$, while trace $d_v$ provides the set of state probabilities of node $v$

node $v$ is in state $x$ at time $t$ in trace $d$. For any node $v$ and time $t$, we assume $s_v^d(t) + e_v^d(t) + i_v^d(t) + r_v^d(t) = 1$ indicating that $v$ must be in one of the SEIR states. In fact, exact state transitions of node $v$ into *E*, *I*, *R* states in trace $d$ happen at $t^d_{e,v}$, $t^d_{i,v}$, $t^d_{r,v}$ respectively. We cannot observe these exact state transition times, but they are related to the observed trace $d$ via the noise dynamics function $\mathcal{N}$ which is explained in detail in Section IV-A. $\mathcal{N}$ provide the probability of observing a particular probabilistic state trace for a node instead of the true state trace. Thus, our computational problem is:

**Problem 1.** *Infer the set of edges $E$ given: the set of nodes $V$, a collection $D$ of traces of probabilistic node states of the form described above, estimates of the noise dynamics $\mathcal{N}$, and a model $\mathcal{M}$, such as SEIR, by which the diffusion process is assumed to have occurred.*

Notation for the problem and its input is summarized in Table I.

Our general framework for Problem 1 is this: we write down a set of probabilistic dynamic equations that model how the probability of each node being in each state changes under SEIR. This provides a theoretical trajectory through the space of state probabilities that depends on which edges exist in the graph and state transition times. We then formulate an optimization problem to find the choice of edges that makes the theoretical trajectories match the observed traces as best as possible under the expectation of the selected loss function over the exact state transition times.

## III. Diffusion Dynamics

We introduce $x_{uv}$ for every pair of nodes $u \neq v$ with the interpretation that $x_{uv} = 1$ if edge $(u, v)$ should exist. Assuming trace $d$ is known for sorted time steps $T_d = t_1, t_2, t_3, \ldots, t_w$, for each consecutive pair $t_{j-1}$, $t_j$ of this sample, SEIR can be thought as nonlinear discrete model and its dynamics can be written as in (1)–(4):

$$s_v^d(t_j) = s_v^d(t_{j-1})\, ss_v^d(t_j) \tag{1}$$

$$e_v^d(t_j) = e_v^d(t_{j-1})\ \left(1 - ei_v^d(t_j)\right) + s_v^d(t_{j-1})\ \left(1 - ss_v^d(t_j)\right) \tag{2}$$

$$i_v^d(t_j) = i_v^d(t_{j-1})\ \left(1 - ir_v^d(t_j)\right) + e_v^d(t_{j-1})\, ei_v^d(t_j) \tag{3}$$

$$r_v^d(t_j) = i_v^d(t_{j-1})\, ir_v^d(t_j) + r_v^d(t_{j-1}) \tag{4}$$

where $ss_v^d(t_j)$, $ei_v^d(t_j)$, and $ir_v^d(t_j)$ model the S→S, E→I, and I→R transition probabilities that will be explicitly defined ahead. The system of equations (1)–(4) give the probability of each node being in each state at time $t_j$. For instance, according to (2), node $v$ is exposed at time $t_j$ if it is exposed at $t_{j-1}$ and has not transitioned into *infected* state, or it was susceptible at $t_{j-1}$ and transitioned into *exposed* state. Among the all state transitions, only S→E is exogenous; it is affected by $x_{uv}$ and that dependence is captured in $ss_v^d(t_j)$ terms. Figure 1 illustrates this dependence. Only the states of nodes 1, 4 may affect S→E transition for node $v$ since there is an edge between $v$ and them, while trace $d_v$ provides the set of state probabilities of node $v$ for a restricted set of time points.

In SEIR, $v$ stays in state S at time $t$ if it does not become exposed by an infected neighbor until after $t$. Let $s_{uv}^d$ be the probability of diffusion from $u$ to $v$ in trace $d$. If $d$ diffuses from $u$ to $v$, diffusion from $u$ to $v$ happens at time $t^d_{i,u} + t$ where $t$ is distributed according to given pmf $p_{uv}^d$, and its cdf $f_{uv}^d$. Then, probability that node $u$ that was infected at time $t$ would expose a neighbor $v$ over interval $[t', t'']$ given it has not exposed $v$ until $t'$ ($\mathrm{p}_{uv}^d(t', t''|t)$) can be computed as in (5) by using Bayes rule when $t'' \geq t' \geq t$:

$$\begin{aligned}\mathrm{p}_{uv}^d(t', t''|t) &= \frac{P(u \text{ infected at } t \text{ exposed } v \text{ between } t' \text{ and } t'')}{P(u \text{ infected at } t \text{ has not exposed } v \text{ until } t')} \\ &= \frac{f_{uv}^d(t''-t)s_{uv}^d - f_{uv}^d(t'-t)s_{uv}^d}{1 - f_{uv}^d(t'-t)s_{uv}^d}\end{aligned} \tag{5}$$

where $f_{uv}^d(\Delta t)$ is the cdf of diffusion time from $u$ to $v$ in trace $d$, and the difference in the nominator is the probability of exposure from $u$ in the interval $[t', t'']$. Using (5), we can estimate $ss_v^d(t_j)$ in (6) in terms of the probability of $v$ not having been passed the infection from any node $u$:

$$ss_v^d(t_j) = \prod_{u \in V} \prod_{t < t_j} \left(1 - \mathrm{p}_{uv}^d(t_{j-1}, t_j|t)\right)^{x_{uv}\, \tilde{i}_u^d(t)\, \left(1 - \sum_{t' < t_j} \tilde{r}_u^d(t')\right)} \tag{6}$$

In other words, the probability that $v$ remains susceptible at time $t_j$ is estimated to be the product over all nodes $u$ for which $x_{uv} = 1$ of the probability that $u$ was infected at time $t < t_j$ without recovering until $t_j$ but did not spread to $v$ during the interval $[t_{j-1}, t_j]$. In (6), $\tilde{e}_v^d(t)$, $\tilde{i}_v^d(t)$ and $\tilde{r}_v^d(t)$ are boolean indicators that are 1 if $v$ enters *E*, *I*, *R* in trace $d$ at time

| Symbol | Definition |
|---|---|
| $s^d_{uv}$ | Probability of diffusion from $u$ to $v$ in trace $d$ |
| $p^d_{uv}$, $f^d_{uv}$ | Probability, cumulative distribution of diffusion time from $u$ to $v$ in trace $d$ |
| $p^{ei}_v$, $p^{ir}_v$ | Probability distribution of $E \to I$, $I \to R$ transition time for $v$ |
| $ei^d_v(t_j)$, $ir^d_v(t_j)$ | Probability of $E \to I$, $I \to R$ transition for node $v$ at time $t_j$ |
| $\mathrm{p}^d_{uv}(t', t''\|t)$ | Probability that $v$ changed to $E$ during $[t', t'']$ by $u$ infected at $t$ in trace $d$ given $u$ has not exposed $v$ until $t'$ |
| $ss^d_v(t_j)$ | Probability that $v$ does not leave state $S$ between $t_{j-1}$ and $t_j$ |
| $\tilde{e}^d_v(t)$, $\tilde{i}^d_v(t)$, $\tilde{r}^d_v(t)$ | Boolean indicator that is 1 if $v$ enters $E$, $I$, $R$ in trace $d$ at time $t$ |
| $g_s(a)$, $g_e(a)$, $g_i(a)$, $g_r(a)$ | Probability of observing $4 \times 1$ state vector $a$ instead of perfect $S$, . . ., $R$ states in any trace at any time. |
| $\alpha^s_m$, $\alpha^e_m$, $\alpha^i_m$, $\alpha^r_m$ | Dirichlet distribution parameter vector for mixture component $m$ and states $S$, . . ., $R$ |

TABLE II: Table of notation for diffusion model

$t$ respectively ($t^d_{e,v} = t$, $t^d_{i,v} = t$, $t^d_{r,v} = t$). The probabilities $ei^d_v(t_j)$, $ir^d_v(t_j)$ in (1)–(4) can be estimated by (7)–(8) in terms of E→I / I→R transition probabilities of $v$, and probability of $v$ being E, I at time $t$.

$$ei^d_v(t_j) = \sum_{t=t_1}^{t_j} p^{ei}_v(t_j - t)\, \tilde{e}^d_v(t) \tag{7}$$

$$ir^d_v(t_j) = \sum_{t=t_1}^{t_j} p^{ir}_v(t_j - t)\, \tilde{i}^d_v(t) \tag{8}$$

A summary of the notation for the diffusion model is in Table II.

## IV. Convex Risk (Expected Loss) Minimization Based Formulation

Having defined the diffusion dynamics, our goal is now to formulate the inference Problem 1. We assume that diffusion data is given, and we have an estimate of noise dynamics $\mathcal{N}$, so $x_{uv}$ will be the only variables in diffusion dynamics (1)–(4). Let $b = \{b_v,\, v \in V\}$ be a noiseless trace, where $b_v = \{t^b_{e,v}, t^b_{i,v}, t^b_{r,v}\}$ and $t^b_{e,v}$, $t^b_{i,v}$, $t^b_{r,v}$ are the exact exposure, infection and recovery times of node $v$ in perfect trace $b$ respectively. Let $B$ be a set of noiseless traces, and $L_b : X \times b \to R$, $L_B : X \times B \to R$ be real-valued loss functions that estimate the loss (cost) of the set of edges $X$ given $b$ and $B$ respectively from the dynamic equations (1)–(4). In our case, set of true diffusion data $B$ is hidden, but we observe $D$ instead which defines the probabilities of being at states $S$, $E$, $I$, $R$ for each time step as discussed in Section II. Given $D$, the most probable set of edges $X \subseteq V \times V$ can be found by minimizing the risk (expected loss) over all realizations of $D$:

$$R(X, D) = \mathbb{E}_B[L_B] = \sum_B L_B(X, B)\, P(B|D) \tag{9}$$

where $P(B|D)$ models the noise dynamics $\mathcal{N}$; it is the probability that the set of observed traces $D$ are generated from the latent true diffusion data $B$. We assume that each trace $d$ is independent and noise affects each trace $d$ independently, so $P(B|D) = \prod_{d \in D} P(b|d)$. Then, overall risk can be expressed as:

$$R(X, D) = \sum_{d \in D} R(X, d) = \sum_{d \in D} \sum_{b \in \mathcal{Q}(d)} L_b(X, b)\, P(b|d) \tag{10}$$

where $\mathcal{Q}(d) = \{(t_e(v), t_i(v), t_r(v)) : t_e(v) \in T_d,\, t_e(v) < t_i(v) < t_r(v)\,, v \in V\}$ is the set of all latent valid trace realizations that might explain the observed $d$.

### A. Estimating $P(b|d)$

The noise affects each node independently, so $P(b|d) = \prod_{v \in V} P(b_v|d_v)$. From Bayes theorem, $P(b_v|d_v)$ can be expressed as:

$$P\left(b_v|d_v\right) = \frac{P\left(d_v|b_v\right)\, P\left(b_v\right)}{\underbrace{\sum_{b^*_v \in \mathcal{Q}(d)[v]} P\left(d_v|b^*_v\right)\, P(b^*_v)}_{P(d_v)}} \tag{11}$$

The probability $P(d_v|b_v)$ of observing $d_v$ given $b_v$ can be expressed as in (12) since observations at each time step are also independent:

$$\begin{aligned} P(d_v|b_v) = & \prod_{t<t^b_{e,v}} g_s(d_v[t]) \prod_{t^b_{e,v} \le t < t^b_{i,v}} g_e(d_v[t]) \\ & \prod_{t^b_{i,v} \le t < t^b_{r,v}} g_i(d_v[t]) \prod_{t^b_{r,v} \le t} g_r(d_v[t]) \end{aligned} \tag{12}$$

In (12), set of functions $g_x(d_v[t])$ for $x \in \{s, e, i, r\}$ give the probability of observing the $4 \times 1$ vector $d_v[t]$ at time $t$ instead of perfect $S$, $E$, $I$, $R$ traces respectively. Entries of $d_v[t]$ sum up to 1, so we model each $g_x(d_v[t])$ by a mixture of 4-dimensional Dirichlet distributions with $M$ components as in (13) which may approximate any functional shape arbitrarily well:

$$g_x(d_v[t]) = \sum_{m \in M} w^x_m\, g^m_x(d_v[t]) \tag{13}$$

Each mixture component $m$ for state $x$, trace $d$ and time $t$ is distributed according to the concentration parameters $\alpha^{x,d,t}_m$. For simplicity, we assume the same concentration parameters for every time $t$ and trace $d$ $\alpha^{x,d,t}_m = \alpha^x_m$. We also assume mixture weights $w^x_m$ to be same for every trace $d$. Each Dirichlet component in (13) is explicitly written in (14) where $d^y_v[t]$ is the value of state $y$ in $d_v[t]$, $\alpha^x_m[y]$ is the concentration parameter for state $y$, and $\mathbf{B}(\alpha^{\mathbf{x}}_{\mathbf{m}})$ is the normalizing constant:

$$g^m_x(d_v[t]) = \frac{1}{\mathbf{B}(\alpha^{\mathbf{x}}_{\mathbf{m}})} \prod_{y \in \{s,e,i,r\}} (d^y_v[t])^{\alpha^x_m[y]-1} \tag{14}$$

On the other hand, prior $P(b_v)$ in (11) can be explicitly written as in (15) in terms of state transition probabilities:

$$\begin{aligned} P(b_v) &= P(t^b_{e,v})\, P(t^b_{i,v}|t^b_{e,v})\, P(t^b_{r,v}|t^b_{i,v}) \\ P(b_v) &= P(t^b_{e,v})\, p^{ei}_v\, p^{ir}_v \end{aligned} \tag{15}$$

where $P(t^b_{i,v}|t^b_{e,v}) = p^{ei}_v$, $P(t^b_{r,v}|t^b_{i,v}) = p^{ir}_v$, and $P(t^b_{e,v}) = \frac{1}{|T_d|+1}$ is uniform since we do not have any prior information about the node transition times. Additional 1 in the denominator of $P(t^b_{e,v})$ models the case of $v$ not ever becoming exposed.

The generative trace noise model expressed by (11) can also be seen as a variant of hidden semi-markov model (segment model) [17] where there is a hidden state for every time point in $T_d$ with 4 possible values *S*, *E*, *I*, *R*. In our case, each state also emits a duration to model the duration of being at a certain SEIR state, but each time step emits a distribution over 4 states instead of a single value as in basic hidden semi-markov model. Only a subset of state transitions are possible at each hidden state as they are restricted according to SEIR dynamics. Here, transition probabilities are defined by $p^{ei}_v$, $p^{ir}_v$ and $P(t^b_{e,v})$ whereas emission probabilities are from Dirichlet distribution mixture as in (13).

### B. Estimating $L_b(X, b)$

There are variety of loss functions for $L_b(X, b)$. Here, we are dealing with the probabilities so we use negative log-likelihood loss ($L_b(X, b) = -\log(\mathcal{L}(X|b))$) where the likelihood is defined as in (16)–(17), and the risk function turns into (18).

$$\mathcal{L}(X|b) = \prod_{v\in V} \left( \prod_{t<t^b_{e,v}} s^d_v(t) \prod_{t^b_{e,v}\le t<t^b_{i,v}} e^d_v(t) \prod_{t^b_{i,v}\le t<t^b_{r,v}} i^d_v(t) \right) \tag{16}$$

$$\mathcal{L}(X|b) = C \prod_{v\in V} \left( \left(1 - ss^d_v\left(t^b_{e,v}\right)\right) \prod_{t\in T_d, t<t^b_{e,v}} ss^d_v\left(t\right) \right) \tag{17}$$

$$R(X,D) = \sum_{d\in D} \sum_{b\in\mathcal{Q}(d)} P(b|d) \overbrace{\left( \sum_{v\in V} -\log\left(1 - ss^d_v\left(t^b_{e,v}\right)\right) + \sum_{v\in V}\sum_{u\in V} \sum_{t^b_{i,u}\le t<\min(t^b_{r,u},t^b_{e,v})} -\log\left(1-\mathrm{p}^d_{uv}\left(t-1,t|t^b_{i,u}\right)\right) x_{uv} \right)}^{-\log(\mathcal{L}(X|b))} \tag{18}$$

Likelihood (16) is the multiplication of the node state probabilities at each observed time point in perfect trace $b$ under SEIR. (17) is obtained from (16) by dynamic equations (1)–(4) where the constant $C$ is obtained from the state transitions that do not involve $X$. Risk for negative log-likelihood loss is written explicitly in (18) when combined with (6), and it is convex as proven in Theorem IV.1. Its proofs can be found in the appendix. $R(X, D)$ (18) is convex so it can be minimized optimally by the existing convex optimization methods [18].

**Theorem IV.1.** *Risk $R(X, D)$ with negative log-likelihood loss function in* (18) *is convex.*

### C. A More Efficient Relaxation

However, minimizing $R(X, D)$ (18) requires estimating the expectation of the loss function over set of all possible perfect transition time realizations defined by $\mathcal{Q}(d)$. This expectation estimation can be quite time-consuming since it may require an exponential number of summations in the worst case. To infer graphs efficiently, we can instead optimize the relaxed risk ($\hat{\mathcal{R}}(X, D)$) as in (19):

$$\hat{\mathcal{R}}(X,D) = \sum_{d\in D}\sum_{b\in\mathcal{Q}(d)} P(b|d) \left( \sum_{v\in V} \mathcal{T}^b_v + \sum_{v\in V}\sum_{u\in V} \sum_{t^b_{i,u}\le t<\min(t^b_{r,u},t^b_{e,v})} -\log\left(1-\mathrm{p}^d_{uv}(t-1,t|t^b_{i,u})\right) x_{uv} \right) \tag{19}$$

which is obtained by replacing each nonlinear term $\log(1 - ss^d_v(t_j))$ with its first-order Taylor approximation ($\mathcal{T}^b_v$) as estimated in (20):

$$\mathcal{T}^b_v = \sum_{u\in V} \log\left(\mathrm{p}^d_{uv}(t^b_{e,v}-1, t^b_{e,v}|t^b_{i,u})\right)\,(x_{uv}-1) \tag{20}$$

We have $P(b|d) = \prod_{v\in V} P(b_v|d_v)$ due to independence of noise for every node, so (19) becomes:

$$\hat{\mathcal{R}}(X,D) = \sum_{d\in D}\sum_{b\in\mathcal{Q}(d)} \sum_{u,v\in V\times V} P(b_u|d_u)\, P(b_v|d_v)\, \mathbf{M^b_{uv}}\, x_{uv} + C \tag{21}$$

where

$$\mathbf{M^b_{uv}} = \log(\mathrm{p}^d_{uv}(t^b_{e,v}-1, t^b_{e,v}|t^b_{i,u})) - \sum_{t^b_{i,u}\le t<\min(t^b_{r,u},t^b_{e,v})} \log(1-\mathrm{p}^d_{uv}(t-1,t|t^b_{e,u})) \tag{22}$$

Equation (21) is a linear function of $X$. In (21), each $x_{uv}$ depends only on the exact state transition times of $u$ and $v$ since the rest of the probabilities in $P(b|d)$ marginalize out when written as $P(b|d) = \prod_{v\in V} P(b_v|d_v)$.

We can express linear Eqn. (21) more explicitly in tensor form by (23) since expected loss for each edge $(u, v)$ depends only on the exact exposure time from $P_v(b|d)$ (sender), and exact infection and recovery times from $P_u(b|d)$ (receiver).

$$\hat{\mathcal{R}}(X,D) = \sum_{d\in D}\sum_{v\in V}\sum_{u\in V}\sum_{t^i_u\in T_d}\sum_{t^i_u<t^e_v}\sum_{t^e_v\le t^r_u} \Big( \mathbf{P^d_{v,e}}\left[t^e_v\right] \times \mathbf{P^d_{u,i,r}}\left[t^i_u, t^r_u\right] \mathbf{M^d_{uv}}\left[t^i_u, t^r_u, t^e_v\right]\, x_{uv} \Big) \tag{23}$$

In (23), $(|T_d|+1)\times 1$ vector $\mathbf{P^d_{v,e}}\left[t^e_v\right]$, and $(|T_d|+1)\times(|T_d|+1)$ matrix $\mathbf{P^d_{u,i,r}}\left[t^i_u, t^r_u\right]$ express these marginal probability distributions as defined in (24)–(25). In both equations, the $(|T_d|+1)$'th entries model the case of never transitioning into

the corresponding state:

$$\mathbf{P}^{\mathbf{d}}_{\mathbf{v},\mathbf{e}}\left[t^e_v\right] = \begin{cases} \sum_{t^e_v<t_1} \sum_{t_1<t_2} P_v(b=\{t^e_v,t_1,t_2\}|d) & \text{if } t \in T_d \\ 1-\sum_{t\in T_d} \mathbf{P}^{\mathbf{d}}_{\mathbf{v},\mathbf{e}}\left[t\right] & \text{else} \end{cases} \tag{24}$$

$$\mathbf{P}^{\mathbf{d}}_{\mathbf{u},\mathbf{i},\mathbf{r}}\left[t^i_u,t^r_u\right] = \begin{cases} \sum_{t_1<t^i_u} P_v(b=\{t_1,t^i_v,t^r_v\}|d) & \text{if } t^i_u < t^r_u \\ 0 & \text{else} \end{cases} \tag{25}$$

$(|T_d|+1)^3$ tensor $\mathbf{M}^{\mathbf{d}}_{\mathbf{uv}}\left[t^i_u,t^r_u,t^e_v\right]$ in (23) defines the coefficients for the edge from $u$ to $v$ to exist under the transition times $t^i_u$, $t^r_u$, $t^e_v \in (T_d+1)^3$ as explicitly defined below in (26):

$$\mathbf{M}^{\mathbf{d}}_{\mathbf{uv}}\left[t^i_u,t^r_u,t^e_v\right] = \begin{cases} \log\left(\mathrm{p}^d_{uv}\left(t^e_v-1,t^e_v|t^i_u\right)\right) & \text{if } t^i_u < t^e_v \le t^r_u \\ -\sum_{t<t^e_v} \log\left(1-\mathrm{p}^d_{uv}\left(t-1,t|t^e_u\right)\right) & \\ 0 & \text{else} \end{cases} \tag{26}$$

We can express (23) more compactly by (27) where each $x_{uv}$ coefficient is inner product of third-order tensor $\mathbf{M}^{\mathbf{d}}_{\mathbf{uv}}$ and the vector $\mathbf{P}^{\mathbf{d}}_{\mathbf{v},\mathbf{e}}$, and then it is sum of the entries of Hadamard product of the resulting matrix and matrix $\mathbf{P}^{\mathbf{d}}_{\mathbf{u},\mathbf{i},\mathbf{r}}$:

$$\hat{\mathcal{R}}(X,D) = \sum_{d\in D}\sum_{v\in V}\sum_{u\in V}\sum_{jk}\left(\mathbf{P}^{\mathbf{d}}_{\mathbf{u},\mathbf{i},\mathbf{r}} \odot \left(\mathbf{M}^{\mathbf{d}}_{\mathbf{uv}} \cdot \mathbf{P}^{\mathbf{d}}_{\mathbf{v},\mathbf{e}}\right)\right) x_{uv} \tag{27}$$

$\hat{\mathcal{R}}(X,D)$ (27) is linear, convex, and it can be optimized quite fast since we can estimate all $x_{uv}$ coefficients by $O(|D||V|^2\max(|T_d|)^3)$ operations instead of $O(|D||V|^2\max(|T_d|)^V)$. $X$ can be found by minimizing $\hat{\mathcal{R}}(X,D)$ optimally by Program (28)–(30):

$$\underset{X}{\operatorname{argmin}} \quad \hat{\mathcal{R}}(X,D) + \lambda \sum_{(u,v)\in V\times V} x_{uv} \tag{28}$$

$$\text{s.t.} \sum_{u\in V,\, t^i_u<t^e_v\le t^r_u} x_{uv} \ge 1,\ \forall d\in D, v\in V \tag{29}$$

$$0 \le x_{uv} \le 1, \quad \forall (u,v)\in V\times V \tag{30}$$

Covering constraints (29) make sure that at least single edge exists between the newly infected node $v$ and the previously infected nodes that are not yet recovered for every trace $d$. When the diffusion data is not perfect, (29) are removed since we do not know $t^d_{i,u}$, $t^d_{e,v}$, $t^d_{r,u}$ from given diffusion data. We obtain the binary solution by randomly rounding $x_{uv}$.

## V. Possible Improvements

### A. Estimating Noise Dynamics Simultaneously With Graph Inference

We may not always know the noise dynamics parameters in Problem 1. In this case, we minimize the expected loss to simultaneously estimate the most possible $X$ and noise parameters under the generative noise model described in Section IV-A. However, their joint optimization is not convex anymore even for negative log-likelihood loss function.

To efficiently estimate both, we propose a two-step procedure similar to Monte Carlo Expectation Maximization [19]. In the first step, we estimate the optimal set of edges $X$ given $D$ and the estimated noise dynamics parameters, and we estimate the optimal noise dynamics parameters given $X$ and $D$ in the second step. First step is same as solving Problem 1, and both steps alternate until convergence to a local optimum. In the second step, we try to find the best mixture weights $w^x_m$ assuming dirichlet distribution concentration parameters $\alpha^x_m$ are fixed at uniformly sampled locations on a four-dimensional grid. Optimizing for the best $w^x_m$ over all latent valid trace realizations $B$ quickly becomes intractable for large number of traces, so we sample set of latent traces by turning the log-likelihoods estimated in the first step into probabilities via exponentiation. Let $\overline{B}$ be the set of sampled latent traces, and $W = \{w_{mx} \,|\, m \in M,\, x \in s,e,i,r\}$ be the set of weight variables where $w_{mx}$ is weight of mixture component $m$ for state $x$. Given $\overline{B}$, we minimize the negative logarithm of multiplications of the probabilities for $\overline{B}$ as in:

$$\mathcal{L}^p(W|X,D) = \sum_{b\in\overline{B}}\sum_{v\in V}\sum_{t\in T^b} -\log\left(\sum_{m\in M} w_{my} e^{bvt}_m\right) \tag{31}$$

where $y$ is the state of node $v$ at time $t$ in trace $b$, and $e^{bvt}_m$ are the coefficients estimated over fixed $\alpha^x_m$'s by Equation (14). Then, we solve the following Program (32)–(34) to estimate $W$:

$$\underset{W}{\operatorname{argmin}} \quad \mathcal{L}^p(W|X,D) \tag{32}$$

$$\text{s.t.} \sum_{m\in M} w_{mx} = 1, \quad \forall x \in s,e,i,r \tag{33}$$

$$w_{mx} \ge 0, \quad \forall m\in M,\, \forall x\in s,e,i,r \tag{34}$$

This optimization program is not under-constrained since we assume same mixture weights for each node which makes a total of $4M$ variables. Objective (32) is convex as in Theorem V.1 which proof follows from the fact that convexity is preserved under addition and negative logarithm of weighted multivariate linear function is also convex due to its positive semidefinite hessian matrix.

**Theorem V.1.** *Objective* (32) *is convex.*

Program (32)–(34) can be solved optimally by exponentiated gradient descent algorithm [20] since equality constraints (33) are non-overlapping. In this case, exponentiated gradient updates involve:

$$w^{t+1}_{mx} = \frac{w^t_{mx}\exp^{\left(-\eta\nabla_{mx}(w^t_{mx})\right)}}{Z^t_x} \tag{35}$$

where $Z^t_x = \sum_{m\in M} w^t_{mx}\exp^{\left(-\eta\nabla_{mx}(w^t_{mx})\right)}$ is the state-dependent normalization constant, parameter $\eta > 0$ is the learning rate, and $\nabla_{mx}(w^t_{mx})$ is the gradient of objective (32) with respect to $w_{mx}$. Weights estimated by (35) already satisfy the constraints (33), and this method iterates until convergence.

### B. Improvements For Special Cases of SEIR

Most of the expressions in the previous sections become slightly easier for SI and SIR models due to fewer states, and disappearance and modifications of the certain transitions. For instance, (15) turns into the uniform distribution for SI model since $p_v^{ei}$, $p_v^{ir}$ transitions disappear, and we do not have any prior information about the infection times. We estimate the coefficients of the relaxed risk $\hat{\mathcal{R}}(X, D)$ by $\mathbf{M}^{\mathbf{d}}_{\mathbf{uv}}\left[t_u^i, t_v^i, t_u^r\right]$, $\mathbf{P}^{\mathbf{d}}_{\mathbf{v,i}}$ and $\mathbf{P}^{\mathbf{d}}_{\mathbf{u,i,r}}$ for SIR model. However, $\mathbf{M}^{\mathbf{d}}_{\mathbf{uv}}\left[t_u^i, t_v^i\right]$ becomes a second-order tensor (matrix) for SI due to the disappearance of recovery times, and we use it together with the vectors $\mathbf{P}^{\mathbf{d}}_{\mathbf{v,i}}$ and $\mathbf{P}^{\mathbf{d}}_{\mathbf{u,i}}$ to estimate $\hat{\mathcal{R}}(X, D)$ coefficients by $O(|D||V|^2 \max(|T_d|)^2)$ operations.

### C. CORMIN Speedups

We speed-up *CORMIN* substantially via two improvements: Edge inference for each node is independent of each other, so risk minimization problem for each node can be solved optimally in parallel which makes *CORMIN* scalable to large graphs as in [6]. Secondly, when estimating the tensor multiplication in (27) for traces with large $T_d$, we approximate the resulting coefficients by building the tensors (24)–(25) and (26) for subset of time points by sampling them via MCMC. The coefficients estimated by ignoring subset of time points are good approximations, as well as *CORMIN* can infer graphs reasonably well in several minutes from the traces that are sampled at a high rate.

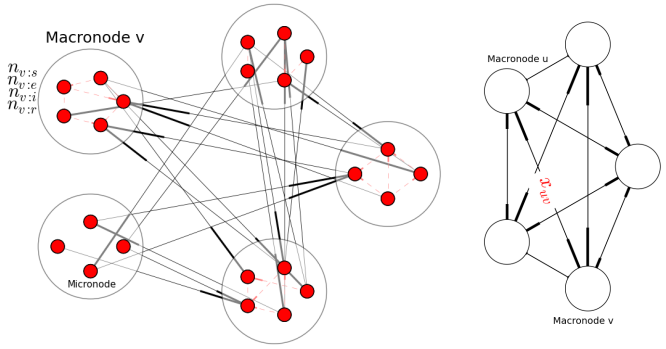
### D. Caveats

In Problem (1), we assume DM parameters between consecutive time steps to be independent and uncorrelated. However, noise dynamics in many realistic scenarios can be better modeled by time-sensitive Dirichlet Mixture model where Dirichlet mixture parameters are also correlated across different time points. These additional dependence constraints further reduce the solution space. We leave improving *CORMIN* to handle such caveats as a future work.

## VI. Macroscale Inference

In Section (IV), we focused on inferring the exact connectivity structure which may be a human-contact network at high school or Facebook friendship network. However, we may not be always interested in inferring the exact network structure since (1) networks we are considering may be massively large, and available diffusion data may not be enough for large-scale inference over them, and (2) it is not worth inferring the every single edge as the connectivity structure at a higher level may be enough for our purpose. For instance, it is impossible to infer the whole human contact network or influenza diffusion network in U.S. from the available influenza diffusion data. Additionally, understanding U.S. influenza network at macroscale, such as inferring the diffusion rates between U.S. states rather than between the humans, may be enough to take preventive measures to stop epidemics.

At macroscale connectivity level, each macronode is composed of micronodes as in Figure (2a), and we are rather interested in estimating the ensemble connectivity rates between and inside the macronodes instead of between the single




(a) (b)

Fig. 2: a) The Original Network, b) The Same Network at macroscale from our perspective

micronodes as in Figure (2b). More formally, we define a fully connected weighted graph $G = (V_m, E_m)$ with self-loops where $V_m$ are macronodes, every edge $(u, v)$ in $E_m$ has an associated macro connectivity rate $x_{uv}$ that the two random micronodes between $u$ and $v$ are connected, and self-loops model the diffusion inside each macronode. Additionally, we assume that the connectivity inside and between every macronode pair is *homogenous* which is quite realistic for large uniform macronodes.

Let $n_v^t$ be the number of micronodes inside macronode $v$ at time $t$, and $n_{v:s}^t$, $n_{v:e}^t$, $n_{v:i}^t$, $n_{v:r}^t$ be the number of micronodes inside macronode $v$ belonging to *S*, *E*, *I*, *R* states respectively. Similarly, we define $p_{v:s}^t = \frac{n_{v:s}^t}{n_v^t}$, $p_{v:e}^t = \frac{n_{v:e}^t}{n_v^t}$, $p_{v:i}^t = \frac{n_{v:i}^t}{n_v^t}$ and $p_{v:r}^t = \frac{n_{v:r}^t}{n_v^t}$ as the fractions of micronodes in the corresponding SEIR states at time $t$, and let $\hat{p}_{v:i}^t = p_{v:i}^t - p_{v:i}^{t-1}$, $\hat{p}_{v:e}^t$, $\hat{p}_{v:r}^t$ be the fraction of newly infected, exposed, recovered nodes respectively. In this case, set of $\hat{p}_{v:x}^t$ for each macronode $v \in V$, each state $x \in \{s, e, i, r\}$, and each time step $t$ define the diffusion data for Problem (1) at macroscale where we do not know exactly which micronode got infected or recovered. This diffusion data has a natural interpretation: each $\hat{p}_{v:i}^t$ is the probability that a random micronode in $v$ has transitioned to state *I*. At this scale, we estimate the ensemble connectivity rates $x_{uv}$ by optimally minimizing the non-relaxed version of the objective (18) where negative log-likelihood is modified as follows:

$$-\log\left(\mathcal{L}(X|b)\right) = \sum_{v\in V} -n_{v:e}^{t_{e,v}^b} \log\left(1 - ss_v^d\left(t_{e,v}^b\right)\right) + \sum_{v\in V}\sum_{u\in V} \sum_{t_{i,u}^b \le t < \min(t_{r,u}^b, t_{e,v}^b)} -n_{v:s}^t n_{u:i}^{t_{i,u}^b} \log\left(1 - \mathrm{p}_{uv}^d\left(t-1, t|t_{i,u}^b\right)\right) x_{uv} \tag{36}$$

where log probabilities of each macronode are also multiplied by the number of micronodes since micronodes are independent and multiplications turns into a summation by taking the logarithm. Solution is then obtained without rounding $X$.

## VII. Experiments & Results

### A. Synthetic Networks and Trace Generation

We tested the inference performance over synthetic networks as follows: We generated 10 synthetic networks of 500 nodes and 5000 edges from each of DMC [21], LPA [22], ForestFire (FF) [23] and Erdos-Renyi (RDS) models by sampling uniformly through their parameters space. Each synthetic trace was generated by choosing a source node randomly and running the diffusion over the network until either all nodes become recovered (or infected under the SI model) or until the spread dies out. When a node gets infection from multiple nodes at different times, it is infected at the earliest infection time. Given noise ratio $p$ between 0 and 1, we added synthetic noise as follows: For every node and time step, we assign probabilistic state vector sample obtained from Dirichlet distribution with concentration parameter vector $\alpha = \left[\frac{p}{4}, \frac{p}{4}, \frac{p}{4}, 1 - \frac{3p}{4}\right]$ where $1 - \frac{3p}{4}$ is the concentration parameter for the current state. This parameter vector becomes uniform for higher noise levels, where it becomes almost impossible to recover the original state.

### B. Real Networks

We tested *CORMIN* by modeling influenza spreading over the human contact network, called *Contact-static*, at an American high school [1] as SI, SIR, SEIR. In this network, nodes represent people and an edge exists between two people if they are near each other. We simulated influenza spreading with $s_{uv} = 0.2$, $p_{uv}$ as weibull distribution with ($\lambda = 9.5, k = 2.3$), and $p_v^{ei}$, $p_v^{ir}$ as exponential distributions with $\lambda = 0.5$ and $\lambda = 0.2$ respectively as discussed in [24]. We also inferred the average influenza transmission rates between U.S. states at macroscale by using the Google Flu Trends Data between 2003–2013 treating each influenza season from September through May as an independent trace where each week is modeled by a single time step. In this *Macro-state* network, each node represents a U.S. state, edges model the influenza transmission rates between those states, and the graph has self-loops to model the influenza diffusion inside the states. The probability of infection at each time step at each U.S. state is the percentage of the people affected by influenza in that state in the corresponding week.

### C. Experiment Details

We implemented *CORMIN* using CPLEX. Its code, used datasets and supplementary text are available on the web[1]. Edge inference for each node is independent and can be solved optimally in parallel which makes *CORMIN* scalable to large graphs. *CORMIN* is reasonably fast; it can infer a graph of 500 nodes and 5000 edges from 100 traces in less than a minute on personal laptop. We compared the performance of *CORMIN* with the best performing existing methods MultiTree [7], NetRate [6], NetInf [5] and InfoPath [8]. We run MultiTree and NetInf giving them the exact number of edges in the true graph although such a perfect estimate is not available a priori. When the diffusion data is perfect, we run *CORMIN* nonparametricly by using only the covering constraints, and estimate the sparsity parameter $\lambda$ in (28) by cross-validation when the diffusion data is partially observable. In this case, we performed 5 cross-validation over the diffusion data as follows: We estimate the set of edges from the training part of the diffusion data for 500 $\lambda$ parameters between 0 and 100, and estimate the error of observing the remaining traces over the inferred graph for every $\lambda$. After repeating this for 5 parts, we return the $\lambda$ minimizing the total error.

When estimating the prediction score at microscale, the edges of the unknown graphs are the positive examples and the pairs between which no edge exists are the negatives. Unknown graphs are sparse so we measure the performance by both $F1 = \frac{2\,\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$ and $F0.1 = \frac{1.01\,\text{precision}\times\text{recall}}{0.01\,\text{precision}+\text{recall}}$ to put more weight on precision where precision is the fraction of edges in the inferred network that are also present in the true network, and recall is the fraction of edges in the true network that are also present in the inferred network. We evaluated the performance of *CORMIN* at macroscale by estimating Pearson correlation coefficient between the inferred influenza rates and the transportation rates between U.S. states estimated from Gowalla dataset [25].

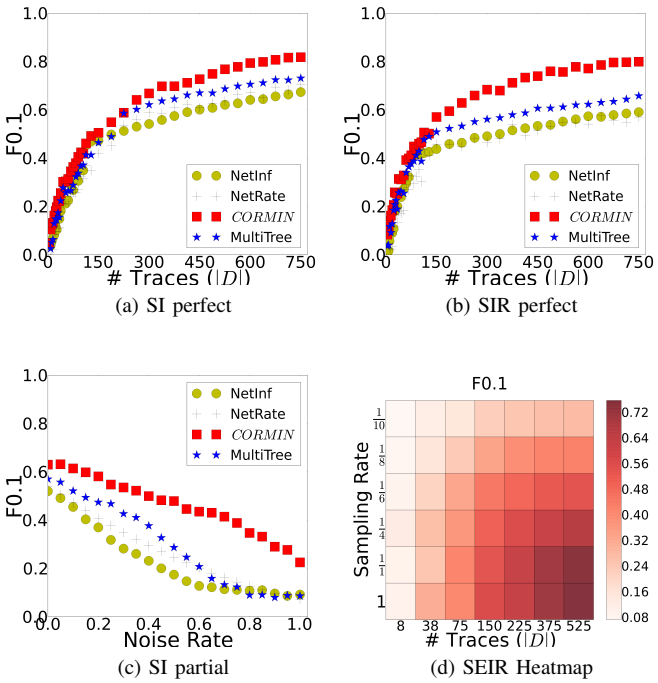
### D. Inferring Static Human Contact Network

We inferred *Contact-static* by synthetic influenza traces on SI, SIR, SEIR that are generated from the real influenza diffusion parameters as discussed above. In these influenza traces, infected state models the human infected with the influenza that is also spreading it to the other people, whereas the exposed state models the human infected with the influenza but has not yet started spreading it to the rest of the school network. When the diffusion data is perfect, *CORMIN* performs the best even though it is nonparametric as in Figure 3a. Similarly, *CORMIN* performs the best under SIR as in Figure 3b, and the performance difference between *CORMIN* and the existing methods are greater than in Figure 3a.

The performance difference between *CORMIN* with the sparsity parameter $\lambda$ estimated by cross-validation and the existing methods becomes more significant when the diffusion data is noisy. This noisy data case is realistic: it may be too costly to track the influenza dynamics exactly since influenza symptoms may be confused with other symptoms, and the diffusion data may be limited especially for novel influenza types such as H5N1 [26] when they first appeared. According to Figure 3a, *CORMIN* achieves $F0.1$ score of 0.7 from 350 perfect traces, and it can achieve the same score from approximately 700 noisy traces. In contrast to this performance, the existing methods can only achieve $F0.1$ score of 0.5 from the same noisy traces. When plotted against increasing noise levels as in Figure 3c, *CORMIN* can achieve $F0.1$ score greater than 0.4 even from highly corrupted traces whereas the existing methods are significantly affected by the increasing noise levels, as $F0.1$ for all of them quickly drop below 0.2.

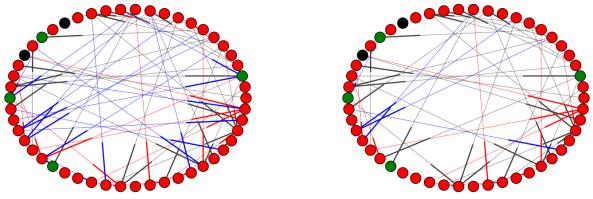
Diffusion data sampled at a lower rate provides less information, and this leads to a overall decrease in *CORMIN*'s performance as in Figure 3d where $\frac{1}{x}$ rate means we only observe 1 time point in every $x$-length interval. *CORMIN*'s performance is affected by the lower sampling rates, but its performance is still reasonable for sampling rates higher than $\frac{1}{5}$ for SEIR across various numbers of diffusion traces. In summary, *CORMIN* performs well on both perfect and partially observable data, and its performance is less affected by the

[1] http://www.cs.cmu.edu/~ckingsf/software/cormin/




(a) SI perfect (b) SIR perfect (c) SI partial (d) SEIR Heatmap

Fig. 3: $F0.1$ vs. number of traces for *Contact-static* under (a) SI, (b) SIR from perfect data; c) $F0.1$ vs. noise ratio for *Contact-static* under SI from 250 traces, d) $F0.1$ Heatmap of number of traces vs. sampling rate under SEIR



a) 50 traces b) 200 traces

Fig. 4: 50 node subgraph of True and the Estimated *Contact-static* From *CORMIN* under SI from a) 50 traces, b) 200 traces

noise in the diffusion data which is not the case for the existing methods.

*CORMIN* can reasonably reveal the human contacts as in Figure 4 which shows the random 50 node subgraph of both estimated and the true contact networks from 50 and 200 diffusion traces respectively. In Figure 4, gray edges represent the edges that are correctly predicted by *CORMIN*, red edges represent the edges that are in the true contact network but not in the estimated network, and the blue edges represent the edges that are in the estimated network but not in the true contact network. In terms of nodes, red nodes represent the students, green and black nodes represent the teachers and the school staff respectively. According to Figure 4, students are densely connected with each other, and the most of the mispredicted connections are between the students instead of between the rest of the people.

Networks inferred by *CORMIN* closely mimick the full range of properties of the true network even from a limited number of traces. Comparison of some of the metrics of *Contact-static* estimated from 50 traces, and the true *Contact-static* can be seen in Table III. For instance, we know that human contact network has scale-free degree distribution with exponent 2.254, and the network estimated by *CORMIN* has similar exponent 2.072.

| | Estimated | Truth |
|---|---|---|
| Modularity [27] | 0.67 | 0.73 |
| Scale-free exponent | 2.072 | 2.254 |
| Assortativity | 0.141 | 0.121 |
| Avg. Clustering Coefficient | 0.23 | 0.261 |
| Diameter | 10 | 8 |

TABLE III: Metrics of true and estimated *Contact-static* networks from 50 traces

### *E. Estimating Influenza Diffusion Rates Between U.S. States*

We estimated the average influenza diffusion rates between U.S. states at macroscale from Google Flu Trends data as described in Section VII-B without rounding the resulting $x_{uv}$. Google Flu Trends data shows the number of weekly infections at each U.S. state between 2003-2013. Here, each node in the network represents a U.S. state, and we treat each influenza season from September to May as an independent trace. Google Flu Trends data is incomplete so we completed the missing data for states at each week as the average of the neighbouring states.

True diffusion rates are unknown but we compared the inferred influenza rates with the transportation rates estimated from Gowalla dataset [25]. Estimated ensemble diffusion rates between the most populated 16 U.S. states are shown in Figure 5. Diagonal entries are the diffusion rates inside U.S. states, and we found influenza diffusion rates inside the most populated states such as New York, Illinois and Texas to be the highest as well as between the nearby states. We estimated the diffusion rates between the northern states to be higher than the diffusion rates for the southern states. However, one may approach these results with caution since the diffusion rates estimated over Google Flu Trends data may be a slight overestimate as discussed in [28].

We estimated the Pearson correlation coefficient between the estimated influenza diffusion rates and transportation rates to be 0.32 which shows that the transportation is one of the major contributors in influenza transmission between U.S. states as discussed previously [29]. We also found influenza diffusion between U.S. states to be fairly asymmetric where we define the asymmetry of the rate matrix as the average of the absolute differences between diffusion rates of every pair of entries the symmetric entries, and estimated it as 0.15 for our rate matrix.

To quantify the degrees of importance of U.S. states in influenza diffusion, we estimated the hubs and authorities values (HITS) for U.S. states on the inferred network by [30]. In general, a good hub represents a U.S. state that diffuses influenza to many other U.S. states, and a good authority represents a U.S. state that gets influenza from other states

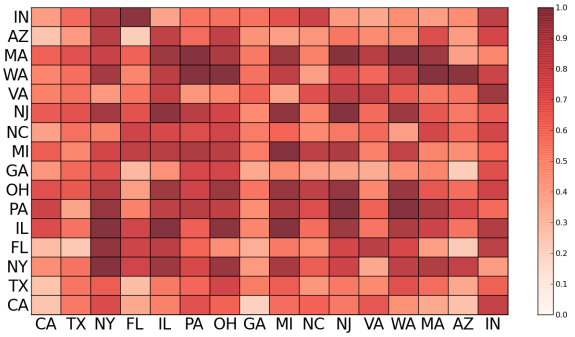



Fig. 5: Estimated Influenza transmission rates between the most populated 16 U.S. states over *Macro-state*

without much spreading it to the other states. Table IV shows the hubs and authorities scores of some U.S. states.

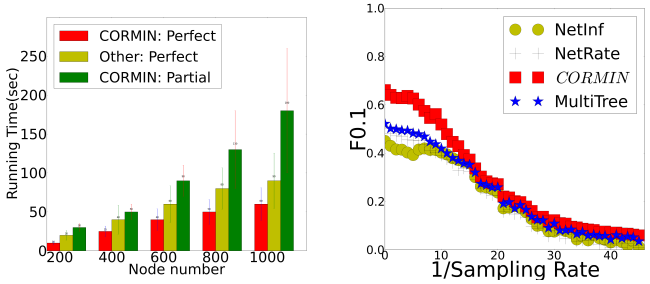
| | Hubs | Authorities | | Hubs | Authorities |
|---|---|---|---|---|---|
| California | 0.058 | 0.060 | Illinois | 0.068 | 0.071 |
| Texas | 0.052 | 0.052 | Pennsylvania | 0.073 | 0.071 |
| Michigan | 0.066 | 0.070 | Massachusetts | 0.071 | 0.068 |
| Florida | 0.060 | 0.056 | Washington | 0.074 | 0.068 |

TABLE IV: Hubs and Authorities scores of some U.S. states on *CORMIN* estimated macroscale network

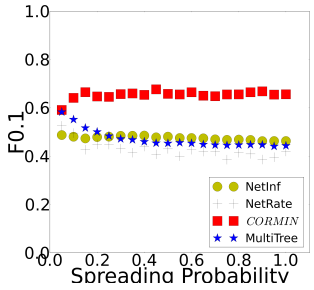
In general, almost all states tend to have close hub and authority scores. We found some of the northern states such as Washington and Massachusetts as well as some mid U.S. states such as Virginia to have higher hub scores whereas the most of the southern states either have slightly higher authority scores or they have close hub and authority scores. Overall, we may think the top-scoring hubs as diffusion accelerators whereas the top-scoring authorities slow down the epidemics. Depending on whether a state is a hub or an authority, we may take different types of measures to prevent or slowdown the epidemics at macroscale.

### F. Inferring Synthetic Networks

*CORMIN* performs consistently better than the existing methods on inferring synthetic networks grown via different growth models from different diffusion models as seen in Table V. Scores in bold represent the cases where *CORMIN* performs reasonably better than the existing methods. *CORMIN* performs significantly better than the existing methods on inferring graphs grown via FF and LPA. All methods perform similar on inferring RDS networks, and they perform the worst on inferring DMC networks. This lower performance can be explained by the loopy structure of DMC networks. In general, *CORMIN* can easily achieve $F1$ score greater than $0.5$ in all models except DMC. Table V shows the performance only from $250$ traces but *CORMIN*'s performance is consistent across different number of traces and conditions.




(a) Running Time (b) Affect of sampling rate




(c) Affect of spreading probability $s_{uv}$

Fig. 6: a) Comparison of running time of *CORMIN* and the existing methods, b) $F0.1$ vs. $\frac{1}{\text{Sampling Rate}}$ from $250$ traces over *Contact-static*, c) Affect of spreading probability $s_{uv}$ on inferring *Contact-static* from $250$ traces

### G. Scalability and Performance under Other Challenging Cases

*CORMIN* infers graphs faster than the existing methods when the data is perfect as in Figure 6a which shows the mean running time as well as the standard deviation from $20$ runs over graphs of different sizes on a single CPU computer. It runs slower when the diffusion data is partial, but this running time is still reasonable considering it is capable of modeling the probabilistic data, and it can infer networks better than the existing methods. *CORMIN* infers *Contact-static* in less than a minute from $500$ traces on a personal laptop.

*CORMIN* is also scalable to very large graphs since convex risk minimization can be done independently for each node. For instance, *CORMIN* can optimally infer graphs having hundred thousands of nodes in less than $3$ minutes by using $100$ processors since it can be parallelized without losing the optimality of the relaxation.

*CORMIN* infers *Contact-static* better than the existing methods when the data is undersampled as in Figure 6b. In this plot, $x$ axis shows the inverse of the sampling rate; $0$ corresponds to the perfectly known case, and $y$ means we only observe $1$ time point in each $y$-length interval. Diffusion data sampled at a lower rate provides less information, but *CORMIN* can tolerate such missing information up to a certain sampling rate as it is still more accurate than the existing methods. However, at sampling rates lower than $\frac{1}{16}$, all methods start to perform similarly and worse since almost all the diffusion information is lost. *CORMIN* is more robust

| | FF | | | LPA | | | DMC | | | RDS | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SI | SIR | SEIR | SI | SIR | SEIR | SI | SIR | SEIR | SI | SIR | SEIR |
| *CORMIN* | **0.62** | **0.57** | **0.61** | **0.59** | **0.5** | **0.61** | 0.45 | 0.44 | 0.49 | 0.52 | **0.53** | **0.55** |
| MultiTree | 0.54 | 0.47 | 0.46 | 0.51 | 0.43 | 0.45 | 0.44 | 0.34 | 0.45 | 0.49 | 0.35 | 0.4 |
| NetInf | 0.52 | 0.45 | 0.46 | 0.50 | 0.42 | 0.47 | 0.4 | 0.41 | 0.47 | 0.47 | 0.33 | 0.38 |
| NetRate | 0.45 | 0.5 | 0.43 | 0.52 | 0.42 | 0.44 | 0.41 | 0.39 | 0.47 | 0.45 | 0.28 | 0.36 |

TABLE V: $F1$ vs. growth and diffusion models for synthetic graphs inferred using 250 traces (No noise added)

to noise as shown previously in Figure 3c, and it consistently performs the best under different probability of diffusion ($s_{uv}$) parameters as in Figure 6c.

## VIII. Conclusion

In this paper, we present a convex risk minimization based approach to infer unknown graphs under SEIR models from probabilistic, partially observable diffusion data. We show improved graph recoverability under both uncertain and perfect node states at multiple scales; our method is capable of recovering the influenza transmission network at microscale and transmission rates at macroscale. The performance advantage of our method can be explained by its better modeling of both edge existence and nonexistence from diffusion data, better handling uncertain data, bounding the number of edges by using covering constraints for perfectly known diffusion data, and its ability to formulate the inference problem at multiple scales. We believe that our model-based inference method can also be extended to the other similar biological network inference problems.

**Acknowledgements.** This work has been partially funded by the US National Science Foundation (CCF-1256087, CCF-1053918) and US National Institutes of Health (R21HG006913 and R01HG007104). C.K. received support as an Alfred P. Sloan Research Fellow.

## References

[1] M. Salathè, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones, "A high-resolution human contact network for infectious disease transmission," *Proceedings of the National Academy of Sciences*, vol. 107, no. 51, pp. 22 020–22 025, 2010.

[2] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10. New York, NY, USA: ACM, 2010, pp. 591–600.

[3] H. W. Hethcote, "The mathematics of infectious diseases," *SIAM Rev.*, vol. 42, no. 4, pp. 599–653, 2000.

[4] S. Myers and J. Leskovec, "On the convexity of latent social network inference," in *Advances in Neural Information Processing Systems 23*, 2010, pp. 1741–1749.

[5] M. Gomez Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10. New York, NY, USA: ACM, 2010, pp. 1019–1028.

[6] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks," in *Proceedings of the 28th International Conference on Machine Learning*, 2011, pp. 561–568.

[7] M. G. Rodriguez and B. Schölkopf, "Submodular inference of diffusion networks from multiple trees," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 2012, pp. 489–496.

[8] M. Gomez Rodriguez, J. Leskovec, and B. Schölkopf, "Structure and dynamics of information pathways in online media," WSDM '13. New York, NY, USA: ACM, 2013, pp. 23–32.

[9] R. M. Anderson and R. M. May, *Infectious Diseases of Humans Dynamics and Control*. Oxford University Press, 1992.

[10] N. Bailey, *The Mathematical Theory of Infectious Diseases and its Applications*. London: Griffin, 1975.

[11] N. Du, L. Song, A. J. Smola, and M. Yuan, "Learning networks of heterogeneous influence." in *NIPS*, 2012, pp. 2789–2797.

[12] A. Defazio and T. S. Caetano, "A convex formulation for learning scale-free networks via submodular relaxation," in *NIPS*, 2012, pp. 1259–1267.

[13] M. S. Elmohamed, D. Kozen, and D. R. Sheldon, "Collective inference on markov models for modeling bird migration," in *Advances in Neural Information Processing Systems 20*, 2007, pp. 1321–1328.

[14] D. Sheldon, T. Sun, A. Kumar, and T. Dietterich, "Approximate inference in collective graphical models," in *Proceedings of The 30th International Conference on Machine Learning*, 2013, pp. 1004–1012.

[15] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Trans. Web*, vol. 1, no. 1, 2007.

[16] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03. New York, NY, USA: ACM, 2003, pp. 137–146.

[17] K. P. Murphy, "Hidden semi-markov models (hsmms," Tech. Rep., 2002.

[18] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[19] C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan, "An introduction to mcmc for machine learning," *Machine Learning*, vol. 50, no. 1-2, pp. 5–43, 2003.

[20] J. Kivinen and M. K. Warmuth, "Exponentiated gradient versus gradient descent for linear predictors," *Information and Computation*, vol. 132, no. 1, pp. 1 – 63, 1997.

[21] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani, "Modeling of Protein Interaction Networks," *Complexus*, vol. 1, no. 1, pp. 38–44, 2003.

[22] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[23] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: Densification laws, shrinking diameters and possible explanations," in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05. New York, NY, USA: ACM, 2005, pp. 177–187.

[24] J. Wallinga and P. Teunis, "Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures," *American Journal of Epidemiology*, vol. 160, no. 6, pp. 509–516, 2004.

[25] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11. New York, NY, USA: ACM, 2011, pp. 1082–1090.

[26] R. Liu, V. R. S. K. Duvvuri, and J. Wu, "Spread pattern formation of H5N1-Avian influenza and its implications for control strategies," *Mathematical Modelling of Natural Phenomena*, vol. 3, pp. 161–179, 1 2008.

[27] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[28] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of

Google Flu: Traps in big data analysis," *Science*, vol. 343, no. 6176, pp. 1203–1205, 2014.

[29] D. Balcan, V. Colizza, B. Gonalves, H. Hu, J. J. Ramasco, and A. Vespignani, "Multiscale mobility networks and the spatial spreading of infectious diseases," *Proceedings of the National Academy of Sciences*, vol. 106, no. 51, pp. 21 484–21 489, 2009.

[30] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, Sep. 1999.

## IX. Appendix

**Theorem IV.1.** *Risk $R(X, D)$ with negative log-likelihood loss function in (18) is convex.*

*Proof:*

We need to prove the convexity of the each additive term w.r.t. $X$ to prove the convexity of $R(X, D)$ (18) for negative log-likelihood loss. There are two types of terms involving $X$: $-\log\left(1-\mathrm{p}_{uv}^{d}(t_{e,u}^{b}, t-1, t)\right)x_{uv}$ and $-\log\left(1-ss_{v}^{d}\left(t_{e,v}^{b}\right)\right)$. Among them, $-\log\left(1-\mathrm{p}_{uv}^{d}(t_{e,u}^{b}, t-1, t)\right)x_{uv}$ is convex since it is a linear function of $X$. The other term $-\log\left(1-ss_{v}^{d}\left(t_{e,v}^{b}\right)\right)$ can be explicitly written in (37):

$$-\log\left(1-ss_{v}^{d}\left(t_{e,v}^{b}\right)\right) = -\log\left(1-\exp^{\sum_{u\in V,\, t_{i,u}^{b}<t_{e,v}^{b}} w_u\, x_{uv}}\right) \tag{37}$$

where $w_u$ are defined in (38) for every $u \in V$, $t_{i,u}^{b} < t_{e,v}^{b}$:

$$w_u = \log\left(1-\mathrm{p}_{uv}^{d}(t_{i,u}^{b}, t_{e,v}^{b}-1, t_{e,v}^{b})\right) \tag{38}$$

(37) is convex since its Hessian when expressed in (39):

$$H = \frac{\exp^{\sum_{u\in V,\, t_{i,u}^{b}<t_{e,v}^{b}} w_u\, x_{uv}}}{\left(1-\exp^{\sum_{u\in V,\, t_{i,u}^{b}<t_{e,v}^{b}} w_u\, x_{uv}}\right)^{2}} \begin{bmatrix} w_1^2 & w_1w_2 & w_1w_3 & \dots & w_1w_n \\ w_2w_1 & w_2^2 & w_2w_3 & \dots & w_2w_n \\ w_3w_1 & w_3w_2 & w_3^2 & \dots & w_3w_n \\ \dots & \dots & \dots & \dots & \dots \\ w_nw_1 & w_nw_2 & w_nw_3 & \dots & w_n^2 \end{bmatrix} \tag{39}$$

is Positive semidefinite (PSD) as it can be expressed as $Z\, y^T\, y$ where:

$$y = [\, w_1,\, w_2,\, \dots,\, w_n] \tag{40}$$

$$Z = \frac{\exp^{\sum_{u\in V,\, t_{i,u}^{b}<t_{e,v}^{b}} w_u\, x_{uv}}}{\left(1-\exp^{\sum_{u\in V,\, t_{i,u}^{b}<t_{e,v}^{b}} w_u\, x_{uv}}\right)^{2}} \geq 0 \tag{41}$$

■