

Ensemble Study for Inference in Epidemic Spreading

I. INTRODUCTION

Epidemic inference is crucial to develop advanced contact tracing strategies in order to mitigate the spreading of an epidemic.

There are regimes in which inference is harder (high transmissibility?)

The goal of this paper is to give a quantitative study, on a simple model, on the hardness of inference tasks, depending on the epidemic's parameters, properties of the contact network, and noise in the observations.

Hardness of SI inference, connection with Steiner tree problem.

II. ENSEMBLE STUDY FOR INFERENCE OF EPIDEMIC TRAJECTORIES

A. Epidemic Inference

SI model on graphs. We consider the Susceptible-Infected (SI) model of spreading, defined over a graph $\mathcal{G} = (V, E)$. At time t a node $i \in V$ can be in two states represented by a variable $x_i^t \in \{S, I\}$. At each time step, an infected node can infect each of its susceptible neighbors ∂i with independent probabilities $\lambda_{ij} \in [0, 1]$. The dynamic is irreversible: a given node can only undergo the transition $S \rightarrow I$. Therefore the trajectory in time of an individual can be parameterized by its infection time t_i . We assume that a subset of the nodes initiate with an infection time $t_i = 0$, i.e. $x_i^0 = I$. A realization of the SI process can be univocally expressed in terms of the independent transmission delays $s_{ij} \in \{1, 2, \dots, \infty\}$, following a geometrical distribution $w_{ij}(s) = \lambda_{ij}(1 - \lambda_{ij})^{s-1}$. Once the initial condition $\{x_i^0\}_{i \in V}$ and the set of transmission delays $\{s_{ij}\}_{(ij) \in E}$ is fixed, the infection times can be uniquely determined from the set of equations:

$$t_i = \delta_{x_i^0, S} \min_{j \in \partial i} \{t_j + s_{ji}\} \quad (1)$$

We assume that each individual has a probability γ to be infected at time $t = 0$, and we assume for simplicity that the transmission probabilities are site-independent: $\lambda_{ij} = \lambda$ for all $(ij) \in E$. The distribution of infection times conditioned on the realization of delays and on the initial condition can be written:

$$P(\{t_i\} | \{x_i^0\}, \{s_{ij}\}) = \prod_{i \in V} \psi^*(t_i, \underline{t}_{\partial i}, x_i^0, \{s_{ji}\}_{j \in \partial i})$$

where ψ^* enforces the above constraint on the infection

times:

$$\psi^* = \mathbb{I}[t_i = \delta_{x_i^0, S} \min_{j \in \partial i} \{t_j + s_{ji}\}] \quad (2)$$

with $\mathbb{I}[A]$ the indicator function of the event A . Once averaged over the transmission delays and over the initial condition, we obtain the following distribution of times:

$$P(\{t_i\}) = \prod_{i \in V} \psi(t_i, \underline{t}_{\partial i}) \quad (3)$$

where:

$$\psi = \sum_{x_i^0} \gamma(x_i^0) \sum_{\{s_{ji}\}_{j \in \partial i}} \psi^*(t_i, \underline{t}_{\partial i}, x_i^0, \{s_{ji}\}_{j \in \partial i}) \prod_{j \in \partial i} w(s_{ji})$$

and $\gamma(x) = \begin{cases} \gamma & \text{if } x = I \\ 1 - \gamma & \text{if } x = S \end{cases}$

Inferring individual's trajectories from partial observations. In the inference problem we assume that some information $\mathcal{O} = \{o_i\}_{i \in \mathcal{S}}$ on the trajectory of a subset $\mathcal{S} \subseteq V$ of individuals is given by the result of medical tests. Most of the time we will take $o_i = x_i^T$, i.e. we observe the state of an individual at time $t = T$. The probability of observations $P(\mathcal{O} | \{t_i\})$ factorizes over the set of individuals:

$$P(\mathcal{O} | \{t_i\}) = \prod_{i \in \mathcal{S}} \rho(x_i^T | t_i). \quad (4)$$

In the simplest case, the state of each individual at time $t = T$ is perfectly known: $\mathcal{O} = \{x_i^T\}_{i \in \mathcal{S}}$, and:

$$\rho(x_i^T | t_i) = \mathbb{I}[x_i^T = r(t_i)]$$

with $r(t) = \begin{cases} I & \text{if } t \leq T \\ S & \text{if } t > T \end{cases}$.

One can also introduce some uncertainty in the result of medical tests, with probability p :

$$\rho(x_i^T | t_i) = (1 - p) \mathbb{I}[x_i^T = r(t_i)] + p \mathbb{I}[x_i^T = \bar{r}(t_i)] \quad (5)$$

(with \bar{r} the negation of r : if $r = I$ then $\bar{r} = S$). Here for simplicity we take the same value for FNR and FPR: $p_{\text{FNR}} = p_{\text{FPR}} = p$, but we could generalize to different FNR and FPR values. Using Bayes rule, the posterior probability of infection times is:

$$P(\{t_i\} | \mathcal{O}) = \frac{P(\{t_i\})P(\mathcal{O} | \{t_i\})}{P(\mathcal{O})} \quad (6)$$

with $P(\{t_i\})$ given in (3). In the Bayes optimal setting, the parameters (λ, γ, p) of the true trajectory are

known, this means that the parameters (λ, γ, p) used in the posterior probability (6) are the same than the true parameters used to generate the observations. However in real cases, the value of the parameters is not known, and we denote by $(\lambda^*, \gamma^*, p^*)$ (resp. λ^I, γ^I, p^I) the parameters used to generate the observations (resp. to infer the infection times).

B. A graphical model for the joint distribution over planted and inferred trajectories

Our objective is to estimate how close is the time t_i inferred from the posterior distribution given the observations \mathcal{O} , from the true infection time, that we denote τ_i . We consider the joint distribution over the true (or planted) times $\{\tau_i\}_{i \in V}$ and the inferred times $\{t_i\}_{i \in V}$:

$$\begin{aligned} P(\{t_i\}, \{\tau_i\}) &= P(\{\tau_i\}) \sum_{\mathcal{O}} P(\mathcal{O}|\{\tau_i\}) P(\{t_i\}|\mathcal{O}, \{\tau_i\}) \\ &= P(\{\tau_i\}) \sum_{\mathcal{O}} P(\mathcal{O}|\{\tau_i\}) P(\{t_i\}|\mathcal{O}) \\ &= P(\{\tau_i\}) P(\{t_i\}) \sum_{\mathcal{O}} \frac{P(\mathcal{O}|\{\tau_i\}) P(\mathcal{O}|\{t_i\})}{P(\mathcal{O})} \end{aligned} \quad (7)$$

where in the second line we used $P(\{t_i\}|\mathcal{O}, \{\tau_i\}) = P(\{t_i\}|\mathcal{O})$, i.e. the probability law of $\{t_i\}$ conditioned on the observations $\{\mathcal{O}\}$ and on the planted times $\{\tau_i\}$ depends only on the observations. In the third line we used the Bayes law (6). In the Bayes optimal setting, i.e. when $(\lambda^*, \gamma^*, p^*) = (\lambda^I, \gamma^I, p^I)$, the joint probability $P(\{t_i\}, \{\tau_i\})$ is invariant under the permutation of its two arguments $\{t_i\}, \{\tau_i\}$.

The probability distribution (7) cannot a priori be written as a graphical model because of the sum over the observations \mathcal{O} and of the denominator $P(\mathcal{O}) = \sum_{\{t_i\}} P(\{t_i\}) P(\mathcal{O}|\{t_i\})$. We instead consider the joint probability distribution *conditioned* on the realization of the true initial condition $\{x_i^0\}$, on the delays $\{s_{ij}\}$, and on the realization of the noise in the observations. For the last one, we introduce binary variables c_i , with $c_i = 0$ when the observation is not corrupted ($x_i^T = r(\tau_i)$), and $c_i = 1$ when the observation is corrupted: $x_i^T = \overline{r(\tau_i)}$. In this way, each c_i is a Bernoulli variable of parameter p . We denote $\mathcal{D} = \{\{x_i^0, c_i\}_{i \in V}, \{s_{ij}, s_{ji}\}_{(ij) \in E}\}$ a realization of the disorder. The joint probability of the planted times $\{\tau_i\}$, of the observations $\mathcal{O} = \{x_i^T\}$ and of the inferred times conditioned on the disorder is:

$$\begin{aligned} P(\{t_i\}, \mathcal{O}, \{\tau_i\}|\mathcal{D}) &= P(\{\tau_i\}|\mathcal{D}) P(\mathcal{O}, \{t_i\}|\mathcal{D}, \{\tau_i\}) \\ &= P(\{\tau_i\}|\mathcal{D}) P(\mathcal{O}|\mathcal{D}, \{\tau_i\}) P(\{t_i\}|\mathcal{D}, \{\tau_i\}, \mathcal{O}) \\ &= P(\{\tau_i\}|\mathcal{D}) P(\mathcal{O}|\mathcal{D}, \{\tau_i\}) P(\{t_i\}|\mathcal{O}) \end{aligned}$$

where in the last line we have again noted that the posterior distribution on the inferred times $\{t_i\}$ depends only on the observations: $P(\{t_i\}|\mathcal{D}, \{\tau_i\}, \mathcal{O}) = P(\{t_i\}|\mathcal{O})$. The first term in the product is:

$$P(\{\tau_i\}|\mathcal{D}) = \prod_{i \in V} \psi^*(\tau_i, \underline{\tau}_{\partial i}; x_i^0, \{s_{ji}\}_{j \in \partial i})$$

with ψ^* given in (2). The second term in the product is the probability of having observation $\mathcal{O} = \{x_i^T\}$ given the planted times $\{\tau_i\}$ and de disorder \mathcal{D} . Each x_i^T is a deterministic function of τ_i and of the corruption variable c_i :

$$P(\mathcal{O}|\mathcal{D}, \{\tau_i\}) = \prod_{i \in V} \mathbb{I}[x_i^T = (1 - c_i)r(\tau_i) + c_i\overline{r(\tau_i)}]$$

The third term is expressed using Bayes's law:

$$P(\{t_i\}|\mathcal{O}) = \frac{P(\{t_i\}) P(\mathcal{O}|\{t_i\})}{P(\{\mathcal{O}\})}$$

with $P(\{t_i\})$ given in (3) and $P(\mathcal{O}|\{t_i\})$ given in (4) (with $\rho(x_i^T|\tau_i)$ given in (5)). Finally, the denominator

$$P(\mathcal{O}) = \sum_{\{t_i\}} P(\{t_i\}) P(\mathcal{O}|\{t_i\})$$

can be seen as a complicated function of the observations \mathcal{O} , but since we have fixed the disorder $\mathcal{D} = \{\{x_i^0, c_i\}_{i \in V}, \{s_{ij}, s_{ji}\}_{(ij) \in E}\}$, the observations are a deterministic function of the disorder: $x_i^T = c_i\overline{r(\tau_i)} + (1 - c_i)r(\tau_i)$, and τ_i is itself a function of the initial condition $\{x_i^0\}$ and of the delays $\{s_{ij}\}$. So we can re-write it as a normalization that depends only on the disorder:

$$P(\mathcal{O}) = Z(\mathcal{D})$$

We obtain a joint-probability on $\{\tau_i\}$, \mathcal{O} and $\{t_i\}$ that is factorized (graphical model):

$$\begin{aligned} P(\{t_i\}, \mathcal{O}, \{\tau_i\}|\mathcal{D}) &= \frac{1}{Z(\mathcal{D})} \prod_{i \in V} \psi^*(\tau_i, \underline{\tau}_{\partial i}; x_i^0, \{s_{ji}\}_{j \in \partial i}) \\ &\quad \times \psi(t_i, \underline{t}_{\partial i}) \mathbb{I}[x_i^T = (1 - c_i)r(\tau_i) + c_i\overline{r(\tau_i)}] \rho(x_i^T|t_i) \end{aligned}$$

Summing over the observations $\mathcal{O} = \{x_i^T\}$ is harmless since only one configuration $\{x_i^T\}_{i \in V}$ is accepted due to the indicator function above (x_i^T is fixed by the disorder). We obtain the joint probability distribution of planted and inferred times $\{\tau_i\}, \{t_i\}$ conditioned on the disorder:

$$\begin{aligned} P(\{\tau_i\}, \{t_i\}|\mathcal{D}) &= \frac{1}{Z(\mathcal{D})} \prod_{i \in V} \psi^*(\tau_i, \underline{\tau}_{\partial i}; x_i^0, \{s_{ji}\}_{j \in \partial i}) \\ &\quad \times \psi(t_i, \underline{t}_{\partial i}) \xi(\tau_i, t_i; c_i) \end{aligned} \quad (8)$$

with:

$$\begin{aligned} \xi(\tau_i, t_i; c_i) &= \rho(x_i^T | t_i) \\ \text{where } x_i^T &= (1 - c_i)r(\tau_i) + c_i \overline{r(\tau_i)} \end{aligned} \quad (9)$$

with $\rho(x|t)$ given in (5).

The case of perfect observations. In that case the probability of error is zero: $p = 0$ so the corrupted variables are always $c_i = 0$ (no corruption), and $\rho(x_i^T | t_i) = \mathbb{I}[x_i^T = r(t_i)]$. The coupling term $\xi(\tau_i, t_i; c_i)$ between inferred and planted times in the joint probability becomes:

$$\begin{aligned} \xi(\tau_i, t_i) &= \mathbb{I}[r(\tau_i) = r(t_i)] \\ \text{where } r(t) &= \begin{cases} I & \text{if } t_i \leq T \\ S & \text{if } t_i > T \end{cases} \end{aligned}$$

$$\Psi(\{T_{il}\}_{l \in \partial i}; \mathcal{D}_i) = \xi(\tau_i^{(j)}, t_i^{(j)}; c_i) \psi^*(\tau_i^{(j)}, \underline{\tau}_{\partial i}^{(i)}; \{s_{li}\}_{l \in \partial i}, x_i^0) \psi(t_i^{(j)}, \underline{t}_{\partial i}^{(i)}) \prod_{l \in \partial i \setminus j} \delta_{t_i^{(j)}, t_i^{(l)}} \delta_{\tau_i^{(j)}, \tau_i^{(l)}} \quad (11)$$

where $j \in \partial i$ is a given neighbour of i . The factor graph associated with this probability distribution now mirrors the original graph $\mathcal{G} = (V, E)$ of contact between individuals. The variable vertices live on the edges $(ij) \in E$, and the factor vertices associated with

C. Belief-Propagation equations for the joint-probability

The factor graph associated with the probability distribution (8) contains short loops which compromise the use of BP. [\(add a figure\)](#) In order to remove these short loops, we introduce the auxiliary variables $(\tau_i^{(j)}, \tau_j^{(i)}, t_i^{(j)}, t_j^{(i)})$ on each edge $(ij) \in E$ of the factor graph, which are the copied times $\tau_i^{(j)} = \tau_i$, and $t_i^{(j)} = t_i$ for all $j \in \partial i$. Let $T_{ij} = (\tau_i^{(j)}, \tau_j^{(i)}, t_i^{(j)}, t_j^{(i)})$ be the tuple gathering the copied time on edge $(ij) \in E$. The probability distribution on these auxiliary variables is:

$$P(\{T_{ij}\}_{(ij) \in E}) = \frac{1}{\mathcal{Z}(\mathcal{D})} \prod_{i \in V} \Psi(\{T_{il}\}_{l \in \partial i}; \mathcal{D}_i) \quad (10)$$

where $\mathcal{D}_i = \{\{s_{li}\}_{l \in \partial i}, x_i^0, c_i\}$ is the disorder associated with vertex $i \in V$, and with:

the function Ψ live on the original vertex set V . We introduce the Belief Propagation (BP) message $\mu_{i \rightarrow \Psi_j}$ on each edge $(ij) \in E$ as the marginal probability law of T_{ij} in the amputated graph in which node j has been removed. The set of BP messages obey a set of self-consistent equations:

$$\mu_{i \rightarrow \Psi_j}(T_{ij}) = \frac{1}{z_{\Psi_i \rightarrow j}} \sum_{\{T_{il}\}_{l \in \partial i \setminus j}} \Psi(\{T_{il}\}_{l \in \partial i}; \mathcal{D}_i) \prod_{k \in \partial i \setminus j} \mu_{k \rightarrow \Psi_i}(T_{ik}) \quad (12)$$

were $z_{\Psi_i \rightarrow j}$ is a normalization factor. These equations are exact when the contact graph $\mathcal{G} = (V, E)$ is a tree. In practice, the BP method is also used as a heuristic on random sparse instance. Introducing a horizon time θ , the random variable $T_{ij} = (\tau_i^{(j)}, \tau_j^{(i)}, t_i^{(j)}, t_j^{(i)})$ lives in a space of size $O(\theta^4)$. We see in supplemental how to simplify the BP equations 12, and obtain a set of equivalent equations defined over modified BP messages living in a smaller space.

III. RESULTS IN THE BAYES-OPTIMAL CASE

A. Check against simulations

B. Varying epidemic's parameters

C. Varying the false positive/negative rates

And also dilution

D. Varying graph properties

Vary graph connectivity and graph ensembles (fat tail)

IV. DEPARTING FROM BAYES-OPTIMAL CONDITIONS

Nishimori equalities for epidemic spreading. Explain that the Nishimori argument for Replica Symmetry still holds in this model.

A. Hint of a RSB transition

V. INFERRING EPIDEMIC'S PARAMETERS

VI. CONCLUSION