Bayesian inference of epidemics on networks via Belief Propagation

Fabrizio Altarelli,1,2 Alfredo Braunstein,1,3,2 Luca Dall’Asta,1,2
Alejandro Lage-Castellanos,1,4 and Riccardo Zecchina1,3,2

1DISAT and Center for Computational Sciences, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
2Collegio Carlo Alberto, Via Real Collegio 30, 10024 Moncalieri, Italy
3Human Genetics Foundation, Via Nizza 52, 10126 Torino, Italy
4Physics Faculty, Havana University, San Lazaro y L, 10400 Habana, Cuba

We study several bayesian inference problems for irreversible stochastic epidemic models on networks from a statistical physics viewpoint. We derive equations which allow to accurately compute the posterior distribution of the time evolution of the state of each node given some observations. At difference with most existing methods, we allow very general observation models, including unobserved nodes, state observations made at different or unknown times, and observations of infection times, possibly mixed together. Our method, which is based on the Belief Propagation algorithm, is efficient, naturally distributed, and exact on trees. As a particular case, we consider the problem of finding the “zero patient” of a SIR or SI epidemic given a snapshot of the state of the network at a later unknown time. Numerical simulations show that our method outperforms previous ones on both synthetic and real networks, often by a very large margin.
Tracing epidemic outbreaks in order to pin down their origin is a paramount problem in epidemiology. Compared to the pioneering work of John Snow on cholera in 1854 [1], modern computational epidemiology can rely on accurate clinical data and on powerful computers to run large-scale simulations of stochastic compartment models. However, like most inverse problems, identifying the origin (or seed) of an epidemic outbreak remains a challenging problem even for simple stochastic epidemic models, such as the susceptible-infected-recovered (SIR) model and the susceptible-infected-recovered (SIR) model.

Several studies have recently proposed maximum likelihood estimators based on various kinds of information: topological centrality [2,3], measures of the distance between observed data and the typical outcome of propagations from given initial conditions [4], or the estimation of the single most probable path [5]. Other estimators are derived under strong simplifying assumptions on the graph structure or on the spreading process [6,7]. Notably, for a continuous time diffusion model with gaussian transmission delays, the estimator in [8] is optimal for trees. An extensive numerical simulations show that it is typically a very good approximation of the state of the system conditioned on some observations. BP only relies on a decorrelation assumption similar to the likelihood function which is not necessarily consistent with the more accurate underlying approximation in [10].

In this Letter we derive the Belief Propagation (BP) equations for the probability distribution of the time evolution of the state of the system conditioned on some observations. BP only relies on a decorrelation assumption similar to the one of [10], and is therefore exact on trees. Extensive numerical simulations show that it is typically a very good approximation on general graphs. BP can be used to identify the origin of an epidemic outbreak in the SIR, SI, and similar models, even with multiple infection seeds and incomplete or heterogeneous information.

The SIR model on graphs. We consider the susceptible-infected-recovered (SIR) model of spreading, a stochastic dynamical model in discrete time defined over a graph $G = (V, E)$. At time $t$ a node $i$ can be in one of three states represented by a variable $x_i^t \in \{S,I,R\}$. At each time step $t$, each infected node $i$ infects each one of its susceptible neighbors $\{j \in \partial i : x_j^t = S\}$ with independent probabilities $\lambda_{ij} \in [0,1]$; then, node $i$ recovers with probability $\mu_i \in [0,1]$. The dynamics is irreversible, as a given node can only undergo the transitions $S \to I \to R$. Two important special cases of SIR are the independent cascades model (obtained when $\mu_i \equiv 1$) [11] and the susceptible-infected-recovered model (obtained when $\mu_i \equiv 0$).

SIR dynamics as a graphical model and Bayesian inference. Assume that a certain set of nodes initiates the infection at time $t = 0$, i.e. with $x_i^0 = I$. A realization of the SIR process can be univocally expressed in terms of a set of independent recovery times $g_i$, distributed as $G_i(g_i) = \mu_i(1 - \mu_i)^{g_i}$, and conditionally independent transmission delays $s_{ij}$ following a geometric distribution $\omega_{ij}(s_{ij}|g_i) = \lambda_{ij}(1 - \lambda_{ij})^{s_{ij}}$ for $s_{ij} \leq g_i$ and $\omega_{ij}(\infty|g_i) = \sum_{s > g_i} \lambda_{ij}(1 - \lambda_{ij})^s$. The infection times $t_i$ can be deterministically computed by imposing the condition $1 = \phi_i = \delta(t_i, I \cup \{x_i^0 = S\}) (\min_{j \in \partial i} \{t_j + s_{ij}\} + 1)$ for every $i$.

The distribution of infection and recovery times $t, g$ given the initial state $x^0$ can thus be written as

$$P(t, g | x^0) = \sum_s P(t | x^0, g, s) P(s | g) P(g)$$

where $t = T$ is the state of every node $x_i^T \in \{S, I, R\}$ is known, and $s_{ij}$ is known for every $i, j$. Using Bayes’ theorem and the posterior can be expressed as

$$P(x^0 | x^T) \propto \sum_{t, g} P(x^T | t, g) P(t, g | x^0) P(x^0)$$

where we exploited the fact that the state $x^T$ at time $T$ given a set $(t, g)$ of infection and recovery times follows a deterministic law $P(x^T | t, g) = \prod_i \delta_i (t_i, g_i, x_i^T)$ where $\delta_i = \mathbb{1} [x_i^T = S, T < t_i] + \mathbb{1} [x_i^T = I, t_i \leq T < t_i + g_i] + \mathbb{1} [x_i^T = R, t_i + g_i \leq T]$. The Belief Propagation equations for the posterior. Finding the marginals of (2) is computationally hard, and we propose to approximate them using BP. We will borrow from graphical models the factor graph representation of the dependence of the factors on their variables in a generalized Boltzmann distribution. It is convenient to introduce the new variables $t_i' = t_i + s_{ij}$ and eliminate the $s_{ij}$ and $s_{ji}$ parameters from the graphical model. By defining
the factors \( \phi_{ij} = \omega_{ij} (t'_i - t_i | g_i) \omega_{ji} (t'_j - t_j | g_j) \) and \( \phi_i = \delta (t_i, \mathbb{I} [x^0_i = S] (\min_{j \in \mathcal{N}_i} \{ t'_j \} + 1)) \), the posterior becomes

\[
\mathcal{P}(x^0 | x^T) \propto \sum_{t', g} \Pi_{i < j} \phi_{ij} \prod_i \phi_i \mathcal{G}_i \mathcal{G}_j
\]

Note that even in the simple deterministic case \( \mu_i \equiv \lambda_i \equiv 1 \) where \( t'_i = t_i \), the graphical model corresponding to the factors \( \phi_i \) has the loopy representation displayed in Fig.1. The representation can be disentangled (see also [12]) by grouping pairs of activation times \( (t_i, t'_i) \) in the same variable node (see Fig.1), which is crucial to make the BP approximation more accurate, and exact on trees. Similarly for the general case (2), we introduce the triplets \( (g_i^{(j)}, t_i^{(j)}, t'_j) \) and group the constraints \( \phi_i \) with compatibility checks into the factor node \( \psi_i = \phi_i(t_i, t'_i) \prod_{j \in \mathcal{N}_i} \delta(t_i^{(j)}, t_i) \delta(g_i^{(j)}, g_i) \) (See Fig.1) to obtain an effective model

\[
Q = \frac{1}{Z} \prod_{i < j} \phi_{ij} \prod_i \psi_i \mathcal{G}_i \mathcal{G}_j
\]

so that \( \mathcal{P}(x^0 | x^T) \propto \sum_{t', g} Q(g, t, t', x^0) \). As the topology of the factor graph now mirrors the one of the original network, this approach allows the exact computation of posterior marginals for the SIR model on acyclic graphs.

We derived the BP equations for (1) (See appendix). A single BP iteration can be computed in time \( O(T \cdot G^2 \cdot |E|) \), where \( G \) is the maximum allowed recovery time, which can be assumed constant for a geometric distribution \( \mathcal{G} \). Once the BP equations converge, estimates for the infection time \( t_i \) of each node are obtained, and nodes can be ranked by the posterior probability of being a seed of the epidemics. The topology of the factor graph now mirrors the one of the original network, this approach allows the exact computation of posterior marginals for the SIR model on acyclic graphs.

Identification of a single seed. We compared the inference performance of BP, of DMP, of a DMP variant we call DMPPr (see appendix) and of the Jordan centrality method [13] on random graphs. We considered random regular graphs (RRG) with degree \( k = 4 \) and preferential attachment scale-free graphs (SFG) with average degree \( \langle k \rangle = 4 \), both with \( N = 1000 \) nodes and homogeneous propagation probability \( \lambda_{ij} \equiv \lambda \) and recovery probability \( \mu_i \equiv \mu \). Simulations summarized in Fig.2 show that BP generally outperforms the other methods by a large margin (see appendix).

Incomplete information. In a more realistic setup, much of the information we assumed to know can be missing. First, a fraction \( \xi \) of the nodes might be unobserved. Figs. 3, 2 show the performance of the four methods in this case. BP finds the true origin in more than 70% of the instances with up to \( \xi = 60\% \), and it outperforms the other three methods for almost all \( \xi \).

Second, the initial time \( T_0 \) and thus the age \( \Delta T = T - T_0 \) of the epidemics could be unknown. For a given upper bound on \( \Delta T \), it suffices to consider the dynamical process to start from the all-susceptible state but to allow nodes to be spontaneously infected at an arbitrary time. This is equivalent to the addition of a fictitious neighbor to every node with no constraint \( \psi_i \) in its activation time but with a prior probability \( \varepsilon_i(g_i^{(j)}, t_i^{(j)}, t_i) = \delta(t_i^{(j)}, \infty)(1 - \gamma) + [1 - \delta(t_i^{(j)}, \infty)] \gamma \) of spontaneous infection (See Fig.3). An example of inference for an epidemics with transmission and recovery

Figure 1. Factor graph representations for irreversible dynamics: full squares represent the factors of a generalized Boltzmann distribution and ellipses the variables on which they depend. (a) Original graph (b) Loopy, naive factor graph for a deterministic dynamics (c) Disentangled dual tree factor graph (d) Factor graph for the SIR model given in (1.1) with known epidemic age (e) Factor graph representation with unknown epidemic age.
Figure 2. (Color online) Comparison between BP and DMP, DMPr and Jordan methods. Each point is an average over 1000 epidemics in random graphs with $N = 1000$. Panels (a)–(b) are for SFG with average degree $\langle k \rangle = 4$, recovery probability $\mu = 0.5$ and observation time $T = 5$. Panels (c)–(f) are for RRG with $k = 4$, $\mu = 0.5$ and $T = 10$. Panels (a) and (c): probability of finding the true origin $i_0$ of the epidemics as a function of the average epidemics size $N_{IR}$. Panels (b) and (d): normalized rank of the true origin $(rank_{i_0})/N_{IR}$ as a function of the transmission probability $\lambda$. The normalized epidemics size $N_{IR}/N$ is also plotted (purple, right axis) versus $\lambda$. For very large epidemics BP may fail to converge in (a) and (b) within the specified number of iterations (the convergence probability is plotted in green in (a)), but relevant information is still present in the (unconverged) marginals. Panel (e): probability of finding the true origin of the epidemic as a function of the fraction of unobserved sites $\xi$ for transmission probability $\lambda = 0.5$ and recovery probability $\mu = 1$. Panel (f): Absolute rank given by each algorithm to the true origin as a function of $\xi$.

probabilities $\lambda = 0.7$ and $\mu = 0.6$ is shown in Fig. 3. The plot shows large correlation between true and inferred infection times, and also that the true origin (which was not observed) corresponds to the individual with largest inferred probability.

Finally, the proposed approach can be also used to estimate the epidemic parameters. Indeed, the partition function $Z$ in (A1) is proportional to the likelihood of the unknown parameters. The log-likelihood $\log Z$ is well-approximated by the opposite of the Bethe free energy, which can be computed easily as a function of the BP messages at the fixed point. We show results for two different realizations of epidemics in Fig. 3. BP equations are iterated for equally spaced parameters $\mu$ and $\lambda$ in $[0, 1]$, and the Bethe free energy is computed after convergence. In both cases the epidemic age and the origin are correctly inferred and the parameters are recovered with good accuracy. In a real setting the search for the point of maximum likelihood can be performed with an expectation-maximization scheme rather than with an exhaustive search.

Multiple seeds. If the epidemics initiates at multiple seeds, methods based on the exhaustive exploration of initial states like DMP suffer a combinatorial explosion. This problem does not affect BP, as the trace over initial conditions
Figure 3. (Color online) Examples of inference with incomplete information on a RRG with $N = 1000$, $k = 4$. Left: inference with $\lambda = 0.7$, $\mu = 0.6$, $\gamma = 10^{-4}$, from observations fraction $1 - \xi = 0.6$ at time $T = 0$ for an epidemics with $T_0 = -8$ (unknown to BP). Each colored block (R: recovered, I: infected and X: unknown) corresponds to a vertex, ordered in the horizontal axis by its real infection time given in the vertical axis. The mean and standard deviation of their BP posterior marginal distribution of infection time is plotted (black dots and error bars) along with the marginal posterior probability of spontaneous infection (orange, circles, right axis). Right: Inference of epidemic parameters. Heat-plot of the likelihood density of the parameters for two virtual epidemics. The first one with $\lambda = 0.5$, $\mu = 0.5$, $\Delta T = 8$ (size $N_{IR} = 653$) shows a maximum of the estimated likelihood at $\hat{\lambda} = 0.695$ and $\hat{\mu} = 0.605$, and the second with $\lambda = 0.5$, $\mu = 0.5$, $\Delta T = 10$ (size $N_{IR} = 462$) shows a maximum at $\hat{\lambda} = 0.5$ and $\hat{\mu} = 0.52$.

is performed directly within the framework. Fig. 4 displays experiments with multiple seeds on RRG, showing that effective inference can also be achieved in this regime (see appendix).

**Evolving networks.** We studied the case of time-dependent transmission probabilities $\lambda_{ij}$. This scenario can be analyzed by considering a distribution of transmission delays $\omega_{ij}(s_{ij}|g_{ij}, t_i)$ depending explicitly on infection times $t_i$ (see appendix). We considered two real-world datasets of time-stamped contacts between pairs of individuals, which we aggregated into $\Delta T$ effective time steps: (i) A dataset describing 20 seconds face-to-face contacts in an exhibition [13]. We employed the following parameters: probability of contagion in a 20 second interval $\lambda_{ij}^{20\text{sec}} = 0.2$, recovery probability $\mu^{20\text{sec}} = 0.0014$. (ii) A dataset of sexual encounters self-reported on a website [14]. We set the probability of transmission in a single contact as $\lambda_{ij}^{\text{contact}} = 0.2$ (within the range considered in [13]) and choose $\mu^{\text{sexual}} = 0.5$. Results on the inference of simulated epidemics on both datasets are summarized in Table 4 showing a striking difference in
| ΔT | Proximity | Sexual |
|-----|-----------|--------|
|     | samples   |        | 5    | 10   | 15   | 20   |
|     | 4082      | 1649   | 2636 | 2611 | 2592 | 2597 |
| BP  | 64/1.4    | 58/2.2 | 91/0.1 | 83/0.2 | 80/0.2 | 78/0.3 |
| DMP | 60/1.6    | 54/2.2 | 79/0.2 | 61/0.6 | 58/0.9 | 55/1.2 |
| DMP | 56/1.8    | 53/2.4 | 79/0.2 | 59/0.9 | 55/1.6 | 53/2.4 |

Table I. Summary of results for virtual epidemics involving 10 or more individuals on real evolving networks. Each cell is in the format $p/r_0$, where $p$ is the probability of perfect inference (%) and $r_0$ is the average ranking given to the real zero-patient.

favour of BP (see full results in the appendix). We tried to determine if the performance of the inference process was favored or hindered by temporal and spatial correlations present in the dataset, that are known to affect significantly the size of the outbreak [15] in some cases. However, after destroying the correlations in one of such cases, we found that the performance of BP was essentially unchanged (see appendix).

Conclusions. We introduced a systematic, consistent and computationally efficient approach to the calculation of posterior distributions and likelihood of model parameters for a broad class of epidemic models. Besides providing the exact solution for acyclic graphs, we have shown the approach to be extremely effective also for synthetic and real networks with cycles, both in a static and a dynamic context. More general epidemic models such as the Reed-Frost model [16] that include latency and incubation times, and other observation models [6, 8] can be analyzed with a straightforward generalization by simply defining appropriate recovery $\mu_i$ and transmission probabilities $\lambda_{ij}$ that depend on the time after infection and by employing modified observation laws $\zeta_i$.

Acknowledgments. The authors acknowledge the European Grants FET Open No. 265496 and ERC No. 267915 and Italian FIRB Project No. RBFR10QUW4.

Appendix A: Belief propagation equations

We show the derivation of the Belief Propagation equations (also known as replica-symmetric cavity equations in statistical physics) for the posterior distribution of the SIR model

$$Q = \frac{1}{Z} \prod_{i<j} \phi_{ij} \prod_i \psi_i \gamma_i \zeta_i G_i$$  \hspace{1cm} (A1)

where

$$\psi_i \left( t_i, g_i, \left\{ \left( t_{ij}^{(j)}, t_{ij}^{(j)} \right), g_j \right\} \right) = \left( \delta \left( t_i, 0 \right) + \delta \left( t_i, 1 + \min_{j \in \partial i} \left\{ t_j^{(j)} \right\} \right) \right) \prod_{j \in \partial i} \delta \left( g_i, g_j \right) \delta \left( t_i, t_j \right)$$  \hspace{1cm} (A2)

$$\phi_{ij} \left( \left( t_{ij}, t_{ij}', g_i \right), \left( t_j, t_j', g_j \right) \right) = \omega_{ij} \left( t_i' - t_i | g_i \omega_{ji} \left( t_j' - t_j | g_j \right) \right)$$  \hspace{1cm} (A3)

$$\gamma_i \left( t_i \right) = \gamma \delta \left( t_i, 0 \right) + \left( 1 - \gamma \right) \left( 1 - \delta \left( t_i, 0 \right) \right)$$  \hspace{1cm} (A4)

$$\zeta_i \left( t_i, g_i, x_i^T \right) = \mathbb{1} \left[ x_i^T = I, t_i \leq T < t_i + g_i \right] + \mathbb{1} \left[ x_i^T = S, T < t_i \right] + \mathbb{1} \left[ x_i^T = R, t_i + g_i \leq T \right]$$  \hspace{1cm} (A5)

$$G_i \left( g_i \right) = r_i \left( 1 - r_i \right)^{g_i}$$  \hspace{1cm} (A6)

Belief propagation consists in a set of equations for single-site probability distributions labeled by directed graph edges. These equations are solved by iteration, and on a fixed point give an approximation for single-site marginals and other quantities of interest like the partition function $Z$.

We recall the general form of the BP equations in the following. For a factorized probability measure on $z = \{z_i\}$.

$$M(z) = \frac{1}{Z} \prod_a F_a(z_a)$$  \hspace{1cm} (A8)
where $z_{2a}$ is the subvector of variables that $F_a$ depends on, the general form of the equations is

$$p_{F_a 	o i} (z_i) = \frac{1}{Z_{ia}} \sum_{\{z_j\}_{j \in \partial a \setminus i}} F_a (\{z_i\}_{i \in \partial a}) \prod_{j \in \partial a \setminus i} m_{j \to F_a} (z_j)$$

(A9)

$$m_{i \to F_a} (z_i) = \frac{1}{Z_{ia}} \prod_{b \in \beta \setminus a} p_{F_b \to i} (z_i)$$

(A10)

$$m_i (z_i) = \frac{1}{Z_i} \prod_{b \in \beta} p_{F_b \to i} (z_i)$$

(A11)

where $F_a$ is a factor (i.e. $\psi$, $\phi_{ij}$, $\gamma$, $\zeta$, $G_t$ in our case), $z_i$ is a variable (i.e. $(t_i, g_i), (t_i^{(j)}, t_k^{(k)}, g_i^{(j)})$ in our case), $\partial a$ is the subset of indices of variables in factor $F_a$ and $\partial i$ is the subset of factors that depend on $z_i$. Terms $Z_{ia}, Z_{ai}$ and $Z_i$ are normalization factors that can be calculated once the rest of the right-hand side is computed. While equations (A10)-(A11) can be always computed efficiently in general, the computation of the trace in (A9) may need a time which is exponential in the number of participating variables. The update equations (A9) for factors $\phi_{ij}$, $\gamma$, $\zeta$ and $G_t$ can be computed in a straightforward way because they involve a very small (constant) number of variables each. We show the derivation of an efficient version of equation (A9) for factor $\psi$ that can be computed in linear time in the degree of vertex $i$:

$$p_{\psi \to j} (t_i^{(j)}, t_j^{(j)}, g_i^{(j)}) \propto \sum_{g_{i_1}, t_{i_1}} \sum_{\{t_k, j_k\} \in \{t_i^{(k)}, j_i^{(k)}\}} m_{i \to \psi} (t_i, g_i) \times$$

$$\prod_{k \in \partial a \setminus j} m_{k \to \psi} (t_i^{(k)}, t_k^{(k)}, g_i^{(k)}) \psi_i (t_i, g_i, \{t_i^{(k)}, j_i^{(k)}\}) \times$$

$$m_{\psi \to j} (0, g_i^{(j)}) \prod_{k \in \partial a \setminus j} m_{k \to \psi} (0, t_k^{(k)}, g_i^{(j)}) +$$

$$m_{i \to \psi} (t_i^{(j)}, g_i^{(j)}) \mathbb{I} (t_i^{(j)} \leq t_j^{(j)} + 1) \sum_{k \in \partial a \setminus j, t_k^{(k)} \geq t_j^{(j)} - 1} m_{k \to \psi} (t_i^{(j)}, t_k^{(k)}, g_i^{(j)})$$

$$- m_{i \to \psi} (t_i^{(j)}, g_i^{(j)}) \mathbb{I} (t_i^{(j)} < t_j^{(j)} + 1) \sum_{k \in \partial a \setminus j, t_k^{(k)} > t_j^{(j)} - 1} m_{k \to \psi} (t_i^{(j)}, t_k^{(k)}, g_i^{(j)})$$

(A13)

where in (A14) we use the fact that $\delta \left( t_i \left( 1 + \min_{j \in \partial a} \{t_j^{(j)}\} \right) \right) = \prod_{j \in \partial a} \mathbb{I} \left( t_i \leq t_j^{(j)} + 1 \right) - \prod_{j \in \partial a} \mathbb{I} \left( t_i < t_j^{(j)} + 1 \right)$. The last equation (A14) can be computed efficiently. Similarly,

$$p_{\psi \to i} (t_i, g_i) \propto \mathbb{I} (t_i, g_i) \prod_{k \in \partial a} m_{k \to \psi} (0, t_k^{(k)}, g_i) +$$

$$\prod_{k \in \partial a, t_k^{(k)} \geq t_i - 1} m_{k \to \psi} (t_i, t_k^{(k)}, g_i)$$

(A15)

A more efficient parametrization of the equations can be derived by noting that in the right-hand side of (A14), incoming distributions $m_{i \to \psi, j}$ are aggregated in a simple way. In this way the BP update of $\psi$ can be computed in time $O (G \cdot T \cdot |\partial i|)$, where $G$ is the maximum allowed information delay, and the one of $\phi_{ij}$ in time $O (G^2 \cdot T)$. In practice, $G$ can be taken constant for a geometric distribution $G$. A single BP iteration can be thus computed in time $O (T \cdot G^2 \cdot |E|)$. We remark that the BP equations for the posterior distribution are exact (and have a unique solution) on tree factor graphs [17]. As the topology of the factor graph mirrors the one of the original graph, (A1) allows the exact computation of posterior marginals for the SIR model on tree graphs (at difference with the DMP method).
where recovered state at observation time. In this expression, \( \log Z \) suggests that the desired probability over all nodes in the graph, to compute the terms \( \sum_{i} \prod_{k \in \partial a} p_{F_a \rightarrow i}(z_i) \). Of course, this add an overall factor set is fixed; in those cases the estimation of the BP log-likelihood (the free energy of our model) helps to identify the cases. A second fact we observed in simulations is that the equations seem to converge almost always when the seed probability of being the origin of the epidemics on a number of BP iterations (e.g. 100), gives excellent results in most very useful information is still present on the unconverged marginals. A simple strategy, consisting in averaging the good inference on the initial conditions. We found out that this limitation is not showstopping; in almost all cases this is often (but not always!) the case when the information present in the observation is insufficient to perform This is often (but not always!) the case when the information present in the observation is insufficient to perform good inference on the initial conditions. We found out that this limitation is not showstopping; in almost all cases this is often (but not always!) the case when the information present in the observation is insufficient to perform good inference on the initial conditions. We found out that this limitation is not showstopping; in almost all cases this is often (but not always!) the case when the information present in the observation is insufficient to perform good inference on the initial conditions. We found out that this limitation is not showstopping; in almost all cases this is often (but not always!) the case when the information present in the observation is insufficient to perform good inference on the initial conditions. We found out that this limitation is not showstopping; in almost all cases this is often (but not always!) the case when the information present in the observation is insufficient to perform good inference on the initial conditions. We found out that this limitation is not showstopping; in almost all cases this is often (but not always!) the case when the information present in the observation is insufficient to perform good inference on the initial conditions. We found out that this limitation is not showstopping; in almost all cases good inference on the initial conditions. We found out that this limitation is not showstopping; in almost all cases good inference on the initial conditions. We found out that this limitation is not showstopping; in almost all cases good inference on the initial conditions. We found out that this limitation is not showstopping; in almost all cases this is often (but not always!) the case when the information present in the observation is insufficient to perform good inference on the initial conditions. We found out that this limitation is not showstopping; in almost all cases this is often (but not always!) the case when the information present in the observation is insufficient to perform good inference on the initial conditions. We found out that this limitation is not showstopping; in almost all cases this is often (but not always!) the case when the information present in the observation is insufficient to perform good inference on the initial conditions. We found out that this limitation is not showstopping; in almost all cases
probabilities $P_s^i(T, i)$, $P_f^i(T, i)$ and $P_R^i(T, i)$ come only from epidemic paths that are consistent with the observed epidemic. In other words, the effective graph for the epidemic transmission is only the part of the original graph that got infected, and its boundary. The susceptible boundary is not used in the message passing procedure, but it is evaluated in the first factor of equation (D2). In the case in which there are unobserved nodes, they are logically considered as possible infected/recovered nodes and, therefore, the DMP equations are iterated over the connected component of infected, recovered and unobserved nodes. We found that this approach, that we will call Restricted Dynamic Message Passing (DMP), in many situations gives better estimates, as is shown in Fig. 4 of the manuscript, and also in figures 6, 7 and 8.

Despite this improvement, DMP is still less accurate than BP. Figs. 7 and 8 show three sets of parameters as benchmarks. In all cases the observation time is known, and epidemics were generated with only one seed. We found that the belief propagation equations sometimes do not converge when the epidemic becomes large compared to the size of the graph (see Figs. 7 and 8). For small observation times ($T = 5$ in Fig. 7) the epidemic remains small and BP always converges. The lack of convergence is associated to a decrease in accuracy and it is in this region where our algorithm performs (slightly) worse. It is important to underline that the amount of information concerning the origin of an epidemic is reduced when the epidemic covers almost all the supporting graph, and not surprisingly all algorithms perform worse in this region (as in Fig. 7).

It is important to notice that the factorization in (D2) is an approximation which can lead to significant errors even when the contact network is a tree (and therefore the solution of the time-forward propagation equations is exact). As an extreme example one can consider the case shown in Fig. 5a. In this case, the most probable position of the seed is at 3, while DMP places the seed at 2. An intuitive explanation of this error is the following: the DMP energy contains a number of contributions from the leaves 6–10 which are larger when the seed is placed at 3 than when it is placed at 2 (because the probability that any of the vertices 6–10 remains $S$ increases with their distance from the seed). This contribution should in fact be ignored, because the probability that the leaves 6–10 are $S$ conditioned on the node 5 being $S$ is 1 independently on the position of the seed, and we know that vertex 5 does not get infected. The factorization assumption in (D2) ignores this. If the number of leaves is large enough, their contribution dominates the energy, and the position at which the seed is placed by DMP is “pushed” to the left. Moreover, if the chain 1–5 is longer, the position at which the seed is placed by DMP can be up to the left-most $I$ seed, and the rank of the actual seed (at the center of the chain) can be arbitrarily large. This example is correctly solved by DMP, which ignores the $S$ leaves to the right of vertex 5. A smaller example based on the same principle is given in Fig. 5b: in this case also DMP gives a wrong answer, as it is unable to eliminate any node. An example with no susceptible nodes is given in Fig. 5c. The most probable seed position is node 1, whereas DMP and DMP predict it to be node 2. Here each of the leaves 6–10 in DMP/DMP “attracts” the seed towards themself, even though for $T$ large enough and conditioned to the fact that node 2 is infected its real contribution to decide between nodes 1 and 2 is negligible, and the most likely position of the seed is mostly influenced by the fact that node 1 is the only $R$ node.

**Appendix E: Evolving networks**

We studied the case of dynamically evolving networks, in which the transmission probability $\lambda_{ij}$ depends on time (representing the time-evolution of interactions between agents). To cope with time dependent transmission probabilities, it suffices to consider transmission delay probabilities that depend on $t_i$, i.e. $\omega_{ij}(s_{ij}|g_{ij}, t_i) = \lambda_{ij}(t_i + s_{ij}) \prod_{k=i}^{t_i-1} (1 - \lambda_{ij}(t + s_{ij}))$ for $s_{ij} \leq g_{ij}$, and $\omega_{ij}(\infty|g_{ij}, t_i) = 1 - \sum_{s=0}^{g_{ij}} \omega_{ij}(s|g_{ij}, t_i)$, the rest of the formalism remaining the same.

We considered two interesting datasets, each one consisting in a large list of time-stamped contacts between pairs of individuals, which we aggregated into $\Delta T$ effective time steps. We focused on the known initial time scenario $T_0 = 0$ to facilitate the comparison between algorithms. We simulated the progression of many virtual epidemics initiated by single random individuals.
1. Proximity contacts network

The first dataset [13] corresponds to 20 seconds face-to-face contacts in an exhibition, obtained using badges with RFID technology. Here the case $T_0 = 0$ is particularly significant because it corresponds to the situation in which the zero patient was infected before entering the facilities. We concentrated on the day in which the number of individuals and contacts was larger. We employed the following set of parameters: probability of contagion in a 20 second interval $\lambda_{20s} = 0.2$, recovery probability $\mu_{20s} = 0.0014$. We selected the day in the dataset in which the number of individuals and interactions was largest. In Fig. 9 we illustrate the results on a large number of random virtual epidemics. We simulated 5000 virtual epidemics for $\Delta T = 20$ (resp. 2000 virtual epidemics for $\Delta T = 30$) initiated by a random individual. In each case, nodes for which being the seed was impossible due to topological constraints were identified in a fast preprocessing. We then attempted the inference of the zero-patient with BP, DMP and DMPr.

2. Sexual contacts network

The second dataset [14] is a database of sexual encounters between clients and escorts self-reported by clients on a Brazilian website. We fixed the probability of transmission in a single contact as $\lambda_{contact} = 0.2$ and the yearly
recovery probability as $\mu_{\text{year}} = 0.5$. We selected the records of the last two years (slightly over half of the dataset), because the oldest data is sparser and seems relatively incomplete (and leads to very small epidemics). The results are summarized in Fig. 10 and Table I of the manuscript.

We also tried to assess the relevance of temporal correlations in the efficiency of the inference processes. In particular, we replicated the analysis above in a different regime of parameters in which temporal correlation was shown to have a significant effect in the size of the outbreak \[15\]. After reshuffling the edges of the network and reassigning the contact times randomly (as in the method “Configuration Model” of \[? \] but preserving the bipartite structure of the network). Specifically, we employed a time-window of the last 800 days with deterministic recovery times at 90 days and infectivity parameter $\lambda = 0.7$. This is a regime of parameters in which it was shown \[15\] that the “Configurational Model” reshuffled version of the network leads to significantly smaller outbreak size than the original one. We used $\Delta T = 80$ and we obtained indeed that outbreaks involved $437 \pm 82$ individuals for the original dataset and $258 \pm 58$ for the reshuffled one over 500 samples. However, the impact on the inference performance was limited in this case, the probability of inferring the correct zero patient was $93\% \pm 1\%$ against $94\% \pm 1\%$ over 500 samples, $73\% \pm 4\%$ against $75\% \pm 4\%$ when restricted to samples with more than 10 infections. More work would be needed to clarify this aspect with more generality.

Appendix F: Inference with multiple seeds

If the epidemics initiates at multiple seeds, methods based on the exhaustive exploration of initial states like DMP suffer a combinatorial explosion; the straightforward generalization of DMP would require $(N^k)$ different forward propagations to explore all subsets of $k$ seeds in a graph with $N$ vertices. This problem does not affect BP, as the trace over initial conditions is performed directly within the framework. A posterior distribution of seeds is obtained with a finite (non-infinitesimal) prior. We performed simulations by generating epidemic spreads with a (known) uniform prior seed probability. Note that this method does not provide the subset of maximum posterior probability but gives the probability of a single node to belong to a seed subset. As before, nodes are sorted by their seed probability and a ROC curve can be computed. Note that the area above the ROC curve is a natural generalization of the normalized rank for a single origin. An alternative generalization is the average of the ranks of the true origins, with the disadvantage that is bounded below by $\frac{1}{2}(k - 1)$. A comparative plot of both measures is given in Fig. 11

Appendix G: Inference of epidemic parameters with BP and DMP

Among the methods tested in this Letter, Jordan centrality is the least effective, but is also the least demanding in terms of prior information. As presented so far, both Dynamic Message Passing and Belief Propagation require not only the network of contacts, but also the values of the transmission probability $\lambda_{ij}$, and the recovery probability $\mu_i$. 

Figure 8. **Left:** Relative ranking of the true zero patient with respect to the size of the epidemic, as a function of the transmission probability $\lambda$. The size of the epidemic is shown as a dashed line, and corresponds to the right $y$ axis. **Right:** Fraction of the instances in which the true zero patient is found by each algorithm. The forward epidemic is propagated until observation time $T = 10$, with recovery probability $\mu = 0.5$. The figure in the right corresponds to the one shown in the manuscript as Fig 2, but with a different parameterization of the $x$ axis. Simulations were run over 1000 samples of random regular graphs with $N = 1000$ and degree 4.
Figure 9. Performance of BP, DMP and DMP on the proximity contacts temporal network. The cumulative distribution function (CDF) of the size for the 5000 epidemics for $T = 20$ (resp. 2000 epidemics for $T = 30$) is plotted on the left panels.

At least in the case of homogeneous parameters $\lambda_{ij} \equiv \lambda, \mu_i \equiv \mu$, this requirement can be relaxed since both methods can be used to infer the epidemic parameters by comparing the energy or the free energy obtained for different pairs $(\lambda, \mu)$. The probability of parameters $(\lambda, \mu)$ given an observation of an epidemic is

$$P(\lambda, \mu | x^T) = \frac{P(x^T | \lambda, \mu) P(\lambda, \mu)}{P(x^T)} \propto P(x^T | \lambda, \mu)$$

assuming a uniform prior for $(\lambda, \mu)$.

**Dynamic Message Passing**

The DMP method gives a probability to the observed epidemic starting from every single possible seed, $P(x^T | i, \lambda, \mu)$ (see (D2)), so we can write

$$P(x^T | \lambda, \mu) = \sum_i P(x^T | i, \lambda, \mu)$$

An estimate of the negated log-likelihood of the parameters can therefore be given as the DMP free energy

$$f_{\text{DMP}}(\lambda, \mu) = -\log P(\lambda, \mu | x^T) \simeq -\log \sum_i P(x^T | i, \lambda, \mu).$$

We call this a free energy since the last expression can be written as $-\log \sum_i \exp(-E(i))$. The lower this free energy, the more likely the parameters.
Figure 10. Performance of BP, DMPr and DMP on the sexual contacts temporal network. The cumulative distribution function (CDF) of the size for the 10,000 epidemics for each $T$ is plotted on the left panels.

**Belief Propagation**

Following the approach of the present Letter, the posterior of the parameters can be computed as

$$\mathcal{P}(\lambda, \mu|\mathbf{x}^T) \propto \sum_{t, g, x^0} \mathcal{P}(\mathbf{x}^T|t, g) \mathcal{P}(t, g|x^0) \mathcal{P}(x^0).$$  \hspace{1cm} (G4)

where the trace goes over the infection times $t$, the recovery delays $g$ and the different choices of zero patients $x_0$.

Therefore the desired log-likelihood of the parameters can be computed as the negated free energy of the model in (A1):

$$f(\lambda, \mu) = -\log Z$$  \hspace{1cm} (G5)
Figure 11. Inference of multiple seeds on 1 000 virtual epidemics with $\mu = 1, \lambda = 0.5$, seed probability $0.002$ and different observation times $T - T_0$ (unknown to the inference algorithm). Bottom: average sum of the rankings of the true seeds in the inferred order is plotted vs. the true number of seeds. Top: the average normalized mean ranking, that equals the average area of the ROC curve. Each ROC curve was computed restricted to the epidemic subgraph.

where $Z$ is given in eq. (A1). The exact calculation of $Z$ is computationally hard, but the same Bethe approximation used for the inference of $x_0$ yields an approximate $f_{BP}(\lambda, \mu)$, given in (B1) for the free energy of this model.

In the manuscript is presented an example of the functions $f_{DMP}(\lambda, \mu)$ and $f_{BP}(\lambda, \mu)$ for an epidemic observed at time $T = 8$. The 2-dimensional space of parameters $(\lambda, \mu)$ is discretized and the (standardized) values of the free energies shown in a heat-plot. Note that every single point of the graph corresponds to a run of DMP and DMPr with the corresponding parameters.

[1] J. Snow, On the Mode of Communication of Cholera (John Churchill, 1855).
[2] D. Shah and T. Zaman, SIGMETRICS 38, 203–214 (2010).
[3] D. Shah and T. Zaman, IEEE Trans Inf Theory 57, 5163–5181 (2011).
[4] C. H. Comin and L. da Fontoura Costa, Phys. Rev. E 84, 056105 (2011).
[5] N. Antulov-Fantulin, A. Lancic, H. Stefancic, M. Sikic, and T. Smuc, Statistical inference framework for source detection of contagion processes on arbitrary network structures, arXiv e-print 1304.0018 (2013).
[6] K. Zhu and L. Ying, in Information Theory and Applications Workshop (ITA), 2013 (2013) pp. 1–9.
[7] W. Luo, W. P. Tay, and M. Leng, IEEE Trans. Signal Process. 61, 2850 (2013).
[8] P. C. Pinto, P. Thiran, and M. Vetterli, Phys. Rev. Lett. 109, 068702 (2012).
[9] A. Y. Lokhov, M. Mézard, H. Ohta, and L. Zdeborová, Inferring the origin of an epidemic with dynamic message-passing algorithm, arXiv e-print 1303.5315 (2013).
[10] B. Karrer and M. E. J. Newman, Phys. Rev. E 82, 016101 (2010).
[11] D. Kempe, J. Kleinberg, and É. Tardos, in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD ’03 (ACM, New York, NY, USA, 2003) p. 137–146.
[12] F. Altafelli, A. Braunstein, L. Dall’Asta, and R. Zecchina, J. Stat. Mech. 2013, P09011 (2013). F. Altafelli, A. Braunstein, L. Dall’Asta, and R. Zecchina, Phys. Rev. E 87, 062115 (2013).
[13] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck, Journal of Theoretical Biology 271, 166 (2011).
[14] L. E. C. Rocha, F. Liljeros, and P. Holme, PNAS 107, 5706 (2010). PMID: 20231480.
[15] L. E. C. Rocha, F. Liljeros, and P. Holme, PLoS Comput Biol 7, e1001109 (2011).
[16] N. T. J. Bailey, The mathematical theory of infectious diseases and its applications (Griffin, London, 1975).
[17] J. Pearl, Reverend Bayes on inference engines: A distributed hierarchical approach (Cognitive Systems Laboratory, School of Engineering and Applied Science, University of California, Los Angeles, 1982).
[18] M. Mézard and A. Montanari, Information, Physics, and Computation (Oxford University Press, 2009).