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,[4] and Riccardo Zecchina[1,3,2]
[1]DISAT and Center for Computational Sciences, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
[2]Collegio Carlo Alberto, Via Real Collegio 30, 10024 Moncalieri, Italy
[3]Human Genetics Foundation, Via Nizza 52, 10126 Torino, Italy
[4]Physics Faculty, Havana University, San Lazaro y L, 10400 Habana, Cuba
(Received 24 July 2013; revised manuscript received 26 November 2013; published 17 March 2014)

We study several Bayesian inference problems for irreversible stochastic epidemic models on networks from a statistical physics viewpoint. We derive equations which allow us 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 susceptible-infected-recovered or susceptible-infected 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.

DOI: 10.1103/PhysRevLett.112.118701 PACS numbers: 89.75.Hc, 89.20.Hh

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 1854 London's cholera hit [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* epidemic 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 (SI) model and the susceptible-infected-recovered (SIR) model.

Several studies have recently proposed maximum likelihood estimators based on various kinds of information: topological centrality [2–4], measures of the distance between observed data and the typical outcome of propagations from given initial conditions [5], or the estimation of the single most probable path [6]. Other estimators are derived under strong simplifying assumptions on the graph structure or on the spreading process [7,8]. Notably, for a continuous time diffusion model with Gaussian transmission delays, the estimator in [8] is optimal for trees. An interesting systematic approach is dynamic message passing (DMP) [9], which consists of a direct approximation and maximization of the likelihood function. DMP makes use of an approximate description of the stochastic process, inspired by statistical physics and relying on some decorrelation assumption, which is very accurate in providing local probability marginals [10]. However, as noted by the authors, it has two drawbacks. First, the space of initial conditions considered must be explored exhaustively. Second, DMP relies on a further assumption of single-site factorization of 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 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 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 $\mathcal{G}_i(g_i) = \mu_i(1 - \mu_i)^{g_i}$, and conditionally independent transmission delays s_{ij}, following a geometric distribution

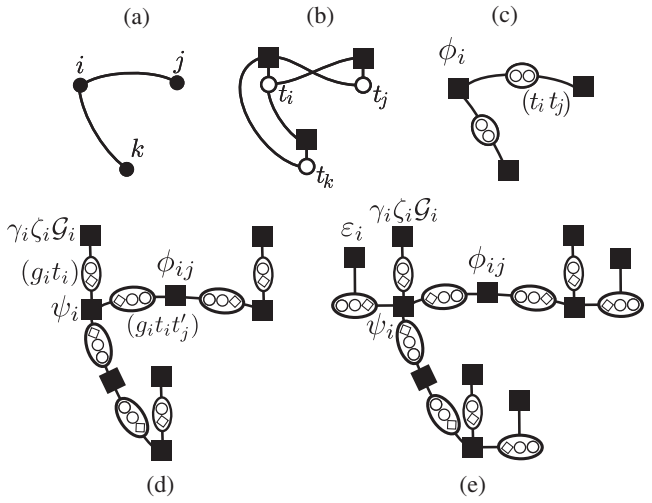

FIG. 1. Factor graph representations for irreversible dynamics: full squares represent the factors of a generalized Boltzman distribution and ellipses the variables on which they depend. (a) Original graph. (b) Loopy, naive factor graph for a deterministic dynamics. (c) Disentangled dual tree factor graph. (d) Factor graph for the SIR model given in (3), with known epidemic age. (e) Factor graph representation with unknown epidemic age.

$\omega_{ij}(s_{ij}|g_i) = \lambda_{ij}(1-\lambda_{ij})^{s_{ij}}$ 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 $1 = \phi_i = \delta(t_i, \mathbb{1}[x_i^0 = S](\min_{j\in\partial i}\{t_j + s_{ji}\} + 1))$ for every i.

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

$$\mathcal{P}(\mathbf{t},\mathbf{g}|\mathbf{x}^0) = \sum_{\mathbf{s}} \mathcal{P}(\mathbf{t}|\mathbf{x}^0,\mathbf{g},\mathbf{s})\mathcal{P}(\mathbf{s}|\mathbf{g})\mathcal{P}(\mathbf{g}) = \sum_{\mathbf{s}} \prod_{i,j}\omega_{ij}\prod_i \phi_i\mathcal{G}_i. \quad (1)$$

In the inference problem we initially assume that (i) at time $t = T$ the state of every node $x_i^T \in \{S, I, R\}$ is known, and (ii) at $t = 0$ the state of every node was extracted independently from the prior distribution $\gamma_i(x_i^0) = \gamma\mathbb{1}[x_i^0 = I] + (1-\gamma)\mathbb{1}[x_i^0 = S]$. Using Bayes' theorem and (1), the posterior can be expressed as

$$\mathcal{P}(\mathbf{x}^0|\mathbf{x}^T) \propto \sum_{\mathbf{t},\mathbf{g}} \mathcal{P}(\mathbf{x}^T|\mathbf{t},\mathbf{g})\mathcal{P}(\mathbf{t},\mathbf{g}|\mathbf{x}^0)\mathcal{P}(\mathbf{x}^0) = \sum_{\mathbf{t},\mathbf{g},\mathbf{s}} \prod_{i,j}\omega_{ij}\prod_i \phi_i\mathcal{G}_i\gamma_i\zeta_i, \quad (2)$$

where we exploited the fact that the state $\mathbf{x}^T$ at time T given a set $(\mathbf{t}, \mathbf{g})$ of infection and recovery times follows a deterministic law $\mathcal{P}(\mathbf{x}^T|\mathbf{t},\mathbf{g}) = \prod_i \zeta_i(t_i, g_i, x_i^T)$, where $\zeta_i = \mathbb{1}[x_i^T = S, T < t_i] + \mathbb{1}[x_i^T = I, t_i \leq T < t_i + g_i] + \mathbb{1}[x_i^T = R, t_i + g_i \leq T]$.

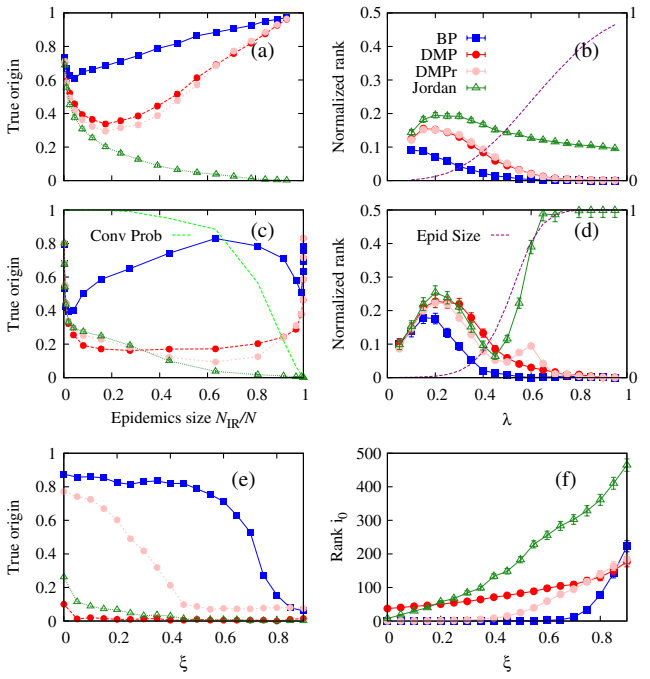

FIG. 2 (color online). Comparison between BP and DMP, DMPr, and Jordan methods. Each point is an average over 1000 epidemics in random graphs with $N = 1000$. Panels (a)–(b) are for SFG with average degree $\langle k\rangle = 4$, recovery probability $\mu = 0.5$ and observation time $T = 5$. Panels (c)–(f) are for RRG with $k = 4$, $\mu = 0.5$, and $T = 10$. (a) and (c) Probability of finding the true origin i_0 of the epidemics as a function of the average epidemics size N_{IR}. (b) and (d) Normalized rank of the true origin $(\text{rank}\, i_0)/N_{IR}$ as a function of the transmission probability λ. The normalized epidemics size N_{IR}/N is also plotted (purple, right axis) vs λ. 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. (e) Probability of finding the true origin of the epidemic as a function of the fraction of unobserved sites ξ for transmission probability $\lambda = 0.5$ and recovery probability $\mu = 1$. (f) Absolute rank given by each algorithm to the true origin as a function of ξ.

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 $t_j' = t_j + s_{ji}$ 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 $\phi_i = \delta(t_i, \mathbb{1}[x_i^0 = S](\min_{j\in\partial i}\{t_j'\} + 1))$, the posterior becomes $\mathcal{P}(\mathbf{x}^0|\mathbf{x}^T) \propto \sum_{\mathbf{t},\mathbf{t}',\mathbf{g}} \prod_{i<j}\phi_{ij}\prod_i \phi_i\mathcal{G}_i\gamma_i\zeta_i$.

Note that even in the simple deterministic case $\mu_i \equiv \lambda_{ij} \equiv 1$, where $t_i' = t_i$, the graphical model corresponding to the factors ϕ_i has the loopy representation displayed in Fig. 1(b). The representation can be

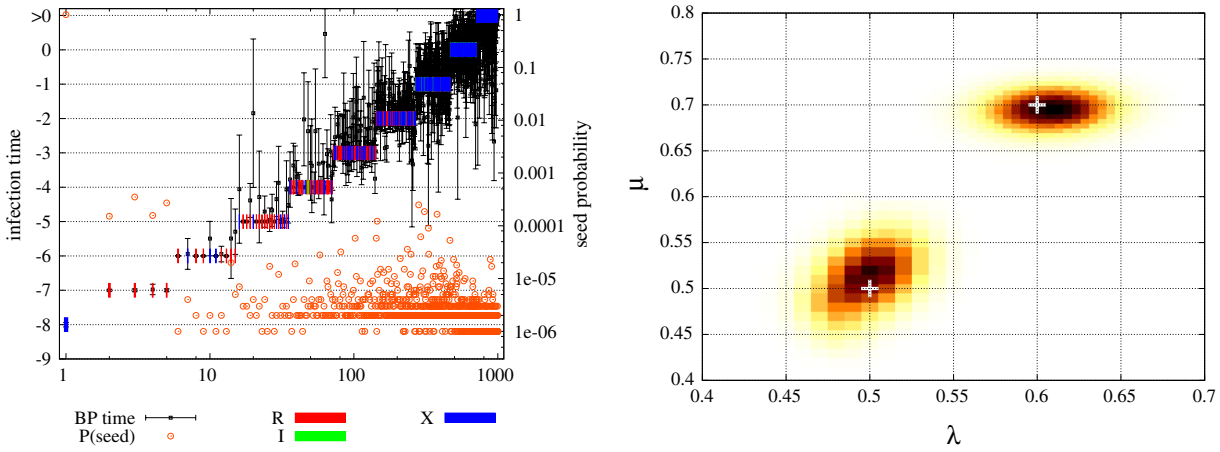

FIG. 3 (color online). Examples of inference with incomplete information on a RRG with $N = 1000$, $k = 4$. Left: inference with $\lambda = 0.7, \mu = 0.6, \gamma = 10^{-6}$, from observations fraction $1 - \xi = 0.6$ at time $T = 0$ for an epidemic with $T_0 = -8$ (unknown to BP). Each colored block (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 $\lambda = 0.7$, $\mu = 0.6$, $\Delta T = 8$ (size $N_{\mathrm{IR}} = 653$) 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 $N_{\mathrm{IR}} = 462$) shows a maximum at $\hat{\lambda} = 0.5$ and $\hat{\mu} = 0.52$.

disentangled (see also [12]) by grouping pairs of activation times (t_i, t_j) in the same variable node [see Fig. 1(c)], 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 ϕ_i with compatibility checks into the factor node $\psi_i = \phi_i(t_i, \mathbf{t}'_{\partial i}) \prod_{j \in \partial i} \delta(t_i^{(j)}, t_i) \delta(g_i^{(j)}, g_i)$ [see Fig. 1(d)] to obtain an effective model

$$\mathcal{Q} = \frac{1}{Z} \prod_{i<j} \phi_{ij} \prod_i \psi_i \mathcal{G}_i \gamma_i \zeta_i, \tag{3}$$

so that $\mathcal{P}(\mathbf{x}^0 | \mathbf{x}^T) \propto \sum_{\mathbf{t},\mathbf{t}',\mathbf{g}} \mathcal{Q}(\mathbf{g}, \mathbf{t}, \mathbf{t}', \mathbf{x}^0)$. 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 (3) [13]. A single BP iteration can be computed in time $O(TG^2|E|)$, where G is the maximum allowed recovery time, which can be assumed constant for a geometric distribution $\mathcal{G}$. Once the BP equations converge, the marginal of 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 $\mathcal{P}(x_i^0 = I | \mathbf{x}^T)$ in decreasing order.

*Identification of a single seed.*We compared the inference performance of BP, of DMP, of a DMP variant we call DMPr [13] and of the Jordan centrality method [2–4] 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 Figs. 2(a)–2(d) show that BP generally outperforms the other methods by a large margin [13].

Incomplete information.—In a more realistic setup, much of the information we assumed to know can be missing. First, a fraction ξ of the nodes might be unobserved. Figures 2(e) and 2(f) 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 ξ.

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 ΔT, 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 ψ_i in its activation time but with a prior probability of spontaneous infection given by a new factor $\epsilon_i(g_i'', t_i'', t_i) = \delta(t_i'', \infty)(1-\gamma) + [1 - \delta(t_i'', \infty)]\gamma$ [see Fig. 1(e)]. An example of inference for an epidemic with transmission and recovery probabilities $\lambda = 0.7$ and $\mu = 0.6$ is shown in Fig. 3. The plot shows the 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 (3) 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

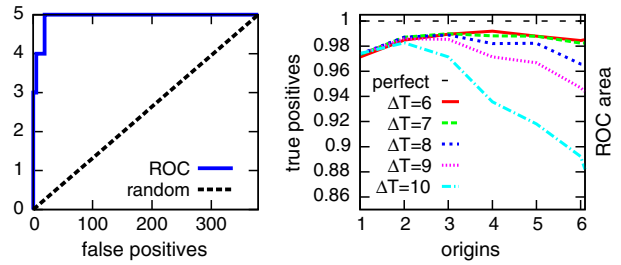

FIG. 4 (color online). BP inference of multiple seeds for virtual epidemics on a RRG with $N = 1000$, $k = 4$, and $\mu = 1$, $\lambda = 0.5$, $\gamma = 0.002$, and different epidemic ages ΔT (unknown to BP). Left: an example receiver operating characteristics (ROC) curve for a sample with 5 origins and $\Delta T = 8$, $N_{\mathrm{IR}} = 379$. The first three ranked nodes are true seeds, and the average normalized area below the curve is 0.975. Right: the average normalized area below the ROC curve vs the number of true origins on 1000 samples. Each ROC curve was computed restricted to the epidemic subgraph.

different realizations of epidemics in Fig. 3. BP equations are iterated for equally spaced parameters μ and λ 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 initiate 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 is performed directly within the framework. Figure 4 displays experiments with multiple seeds on RRG, showing that effective inference can also be achieved in this regime [13].

Evolving networks.—We studied the case of time-dependent transmission probabilities λ_{ij}. This scenario can be analyzed by considering a distribution of transmission delays $\omega_{ij}(s_{ij}|g_i, t_i)$ depending explicitly on infection times t_i [13]. We considered two real-world data sets of time-stamped contacts between pairs of individuals, which we aggregated into ΔT effective time steps. (i) A data set describing 20 s face-to-face contacts in an exhibition [14]. We employed the following parameters: probability of contagion in a 20 s interval $\lambda_{ij}^{20\ \mathrm{s}} = 0.2$, recovery probability $\mu^{20\ \mathrm{s}} = 0.0014$. (ii) A data set of sexual encounters self-reported on a website [15]. We set the probability of transmission in a single contact as $\lambda_{ij}^{\mathrm{contact}} = 0.2$ (within the range considered in [15]) and choose $\mu^{\mathrm{year}} = 0.5$. Results on the inference of simulated epidemics on both data sets are summarized in Table I showing a striking difference in favor of BP (see full results in [13]). We tried to determine if the performance of the inference process was favored or hindered by temporal and spatial correlations present in the data set that are known to affect significantly the size of the outbreak [16] in some cases. However, after destroying the correlations in one of such cases, we found that the performance of BP was essentially unchanged [13].

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.

	Proximity		Sexual			
ΔT	20	30	5	10	15	20
Samples	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
DMPr	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

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 [17] that include latency and incubation times, and other observation models [6,8] can be analyzed with a straightforward generalization by simply defining appropriate recovery μ_i and transmission probabilities λ_{ij} that depend on the time after infection and by employing modified observation laws ζ_i.

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

*Corresponding author.
alfredo.braunstein@polito.it

[1] J. Snow, *On the Mode of Communication of Cholera* (John Churchill, London, 1855).
[2] D. Shah and T. Zaman, SIGMETRICS **38**, 203 (2010).
[3] D. Shah and T. Zaman, IEEE Trans. Inf. Theory **57**, 5163 (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, arXiv:1304.0018.
[6] K. Zhu and L. Ying, in *Information Theory and Applications Workshop (ITA), 2013* (IEEE, New York, 2013), p. 1.
[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, arXiv:1303.5315.

[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, 2003), p. 137.

[12] F. Altarelli, A. Braunstein, L. Dall'Asta, and R. Zecchina, J. Stat. Mech. (2013) P09011; F. Altarelli, A. Braunstein, L. Dall'Asta, and R. Zecchina, Phys. Rev. E **87**, 062115 (2013).

[13] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.112.118701 for details and derivation of the BP equations, description of the DMPr procedure, details on simulations on dynamic contact networks, inference with multiple seeds and inference of dynamic parameters.

[14] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck, J. Theor. Biol. **271**, 166 (2011).

[15] L. E. C. Rocha, F. Liljeros, and P. Holme, Proc. Natl. Acad. Sci. U.S.A. **107**, 5706 (2010).

[16] L. E. C. Rocha, F. Liljeros, and P. Holme, PLoS Comput. Biol. **7**, e1001109 (2011).

[17] N. T. J. Bailey, *The Mathematical Theory of Infectious Diseases and Its Applications* (Griffin, London, 1975).