Scalable Learning of Spreading Models on Networks

Mateusz Wilinski and Andrey Lokhov

Los Alamos National Laboratory, Los Alamos, USA

Abstract

Spreading processes play an important role in modeling diffusion networks and information propagation. Recent real-world events further highlight the need for prediction and control of diffusion dynamics. To tackle these tasks, it is essential to learn the effective spreading model and transmission probabilities across the network of interactions. In most cases not only the transmission rates are unknown, but also full observation of the dynamics is not available. As a result, standard approaches such as maximum likelihood quickly become intractable for large networks. In this work, we study the Independent Cascade model. We introduce an efficient algorithm, based on dynamic message-passing approach, which is able to learn parameters of the effective spreading model given only limited information. Importantly, we show that the resulting model approximates the marginal activation probabilities that can be used for prediction. Additionally, we develop a systematic procedure for learning a mixture of models which further improves the prediction quality.

Learning Framework

We focus on approximating $p_i^c(t)$ - the probability of node i to be active at time t under the cascade c. DMP equations for the IC model allow one approximate this probability:

$$p_i^c(t) = 1 - \left(1 - \bar{p}_i^c\right) \prod_{k \in \partial i} \left(1 - \alpha_{ki} \cdot p_{k\to i}^c(t-1)\right), \tag{1}$$

$$p_{j\to i}^c(t) = 1 - \left(1 - \bar{p}_j^c\right) \prod_{k \in \partial j \setminus i} \left(1 - \alpha_{kj} \cdot p_{k\to j}^c(t-1)\right), \tag{2}$$

where α_{ij} are the transmission rates and $p_{j\to i}^c(t)$ are the probabilities of node j being active at time t, on an auxiliary graph without node i. These equations are further used in our learning scheme, which focuses on optimising a following Lagrange function:

$$\mathcal{L} = \underbrace{\mathcal{O}}_{objective} + \underbrace{\mathcal{C}}_{constrains}, \tag{3}$$

where the objective is to maximise the probability of observed cascades, while the constrains are the DMP equations. The former can be written in a form of log likelihood sum.

$$\mathcal{O} = \sum_{c \in C} \sum_{i \in O} \log \mu_i^c(\tau_i^c), \tag{4}$$

where τ_i^c is the observed activation time of node i, under cascade c and $\mu_i^c(t)$ is the marginal probability of node i being activated at time t. The latter can be computed from the probabilities obtained by DMP. Finally, we use Lagrange equations in a gradient descent approach, to find the optimal transmission rates:

$$\alpha_{ij} = \alpha_{ij} + \varepsilon \cdot \frac{\partial \mathcal{L}}{\partial \alpha_{ij}}. \tag{5}$$

This step is repeated until convergence.

Estimating Model Parameters

We use the above framework to find transmission rates for different types of network with different number of unobserved nodes. Some of the results are shown below.

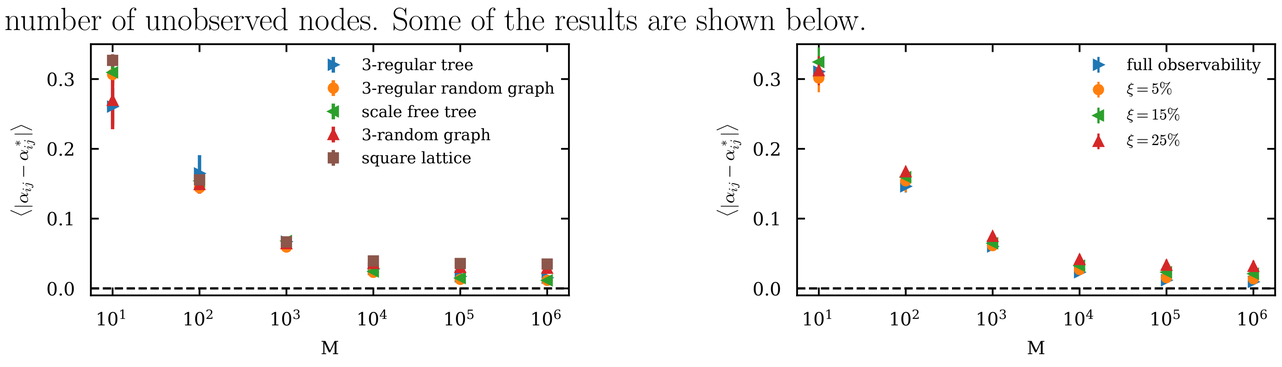

Fig. 1: The difference between inferred and real parameters α_{ij}, as a function of number of available cascades. Regular graph on the right.

We can see that introducing loops leads to a gap between the real and estimated parameters. We further test it by repeating the inference for a large and loopy real-world network, as well as a square lattice in case of long cascades.

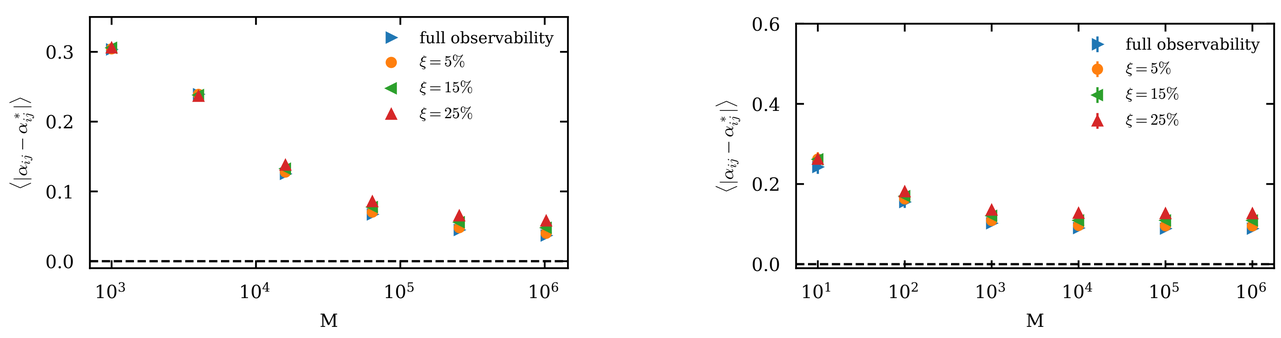

Fig. 2: The difference between inferred and real parameters α_{ij}, as a function of number of available cascades in case of a large ($N = 121,422$) real-world network obtained by a web-crawler on the left and in case of long cascades ($T = 20$) on a square lattice.

As expected, longer cascades increase the size of the gap in case of the square lattice, but as shown in the next part, the gap is created in an attempt to fit the marginals and increase the prediction capabilities of the model. One of the main advantages of the proposed method is its numerical complexity. The reason why we were able to produce results for networks of size $N = 10^5$ and even $N = 10^6$ is that our algorithm scales linearly with number of edges $|E|$, number of cascades M and cascades size T.

Estimating Marginals

The primary source of error in DMP equations is related to loops: Their presence may result in overestimating parameters to account for DMP neglecting information on some spreading paths. To study this effect in more details, we focused on the most challenging case of a square lattice with many short loops, in an adversarial case of very long cascades with $T = 20$. This represents a challenging setting in terms of prediction of marginals using the DMP-based algorithm, which results in a finite error for parameter estimation even for a large number of cascades. The next figure shows relative distance between true and estimated marginals, and demonstrates that DMP equations run with reconstructed parameters produce a better approximation of real marginal probabilities, compared to DMP predictions using the ground-truth parameters α_{ij}^*.

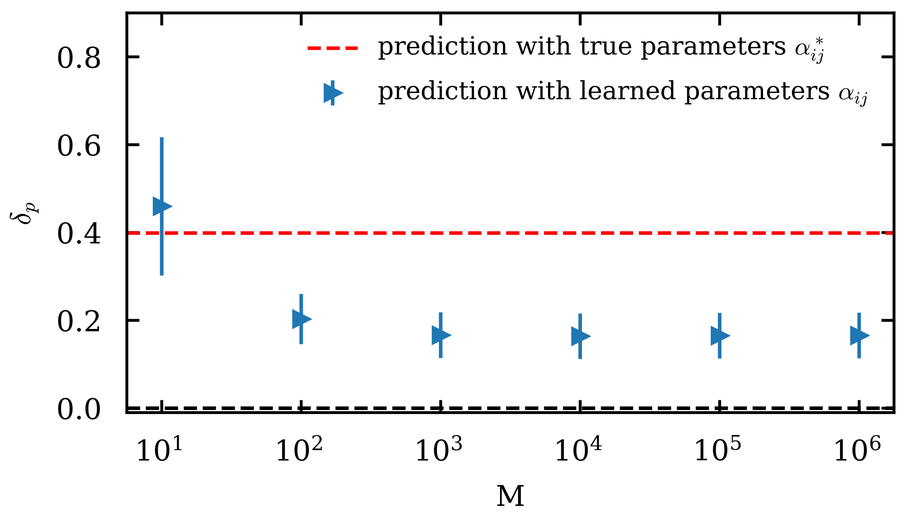

Fig. 3: Relative marginals differences for square lattice. Red line represent marginals for real transmission rates.

Further, we capitalize on the concept of effective models for inference and develop a novel procedure for learning a mixture of spreading models on several replicas of the graph. The new objective in this case is as follows:

$$\mathcal{O}^{\text{mixture}} = \sum_{s \in S} \sum_{i \in O} \sum_{\tau_i^s} m^{\tau_i^s} \log \left(\frac{1}{|R|} \sum_{r \in R} \mu_{i,r}^s(\tau_i^s) \right), \tag{6}$$

where R is the set of replicas. Described approach further improves prediction quality of marginal probabilities, as shown in the next plot.

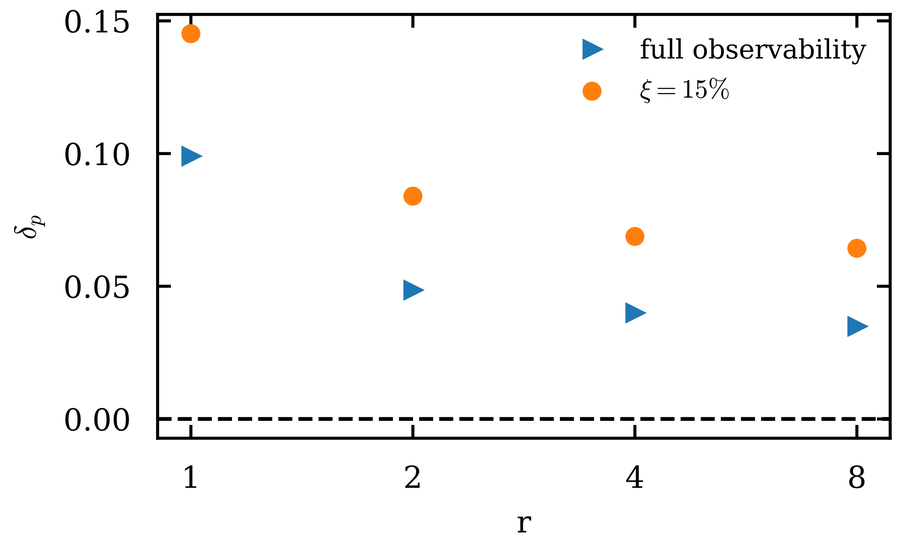

Fig. 4: Relative marginals differences for square lattice as a function of number of replicas used.

As a result, by sacrificing the correctness of parameters, one can achieve better predictions, even in a loopy regime. Importnatly, this result holds also for the partial observation scenario.

References

[1] A. Lokhov, *Neural Information Processing Systems*, 3467-3475 (2016).
[2] A. Lokhov and D. Saad, *PNAS*, 114 (2017).
[3] M. Wilinski and A. Lokhov, *ArXiv*:2007.06557 (2020).