Information is localized in growing network models

Till Hoffmann (thoffmann@hsph.harvard.edu) and Jukka-Pekka Onnela (onnela@hsph.harvard.edu)
T. H. Chan School of Public Health, Harvard University

Mechanistic models with simple rules can yield complex networks that capture salient characteristics of real-world data, such as heavy-tailed degree distributions[1] or the small-world effect[2]. However, fitting these models to data is challenging because the order of addition of nodes is typically not known, rendering the likelihood intractable. Existing approaches compare summary statistics[3], develop model-specific likelihood approximations[4], or seek to make the likelihood tractable by assuming the order is known[5,6] or by inferring it[7,8].

While the emergent global properties of these models are complex, real-world networks likely arise from local processes[1] (imagine having to consider billions of people to choose friends). We conjecture that all information required for inference is localized in monotonic growth models, i.e., models that add but do not remove nodes or edges. We offer evidence in support of our conjecture by applying neural posterior density estimators[9] (NPDEs) to data simulated by four network models.

We grow an undirected graph by repeatedly applying the same rule[6]. At each step t, we add a new node t and connect it to the existing network by a set of edges ϵ_t. The rule is fully specified by the conditional distribution $p\left(\epsilon_t \mid G_{t-1}\right)$, where G_t is the graph at time t. It comprises edges $E_t = E_0 \cup \bigcup_{t'=1}^{t} \epsilon_{t'}$, where E_0 is the initial edge set. Consider a restricted, localized model: We sample seed nodes S_t independent of the graph structure and select neighbors for the new node t by exploring the neighborhoods of seeds. We say the rule is *k-localized* if new edges ϵ_t only depend on the subgraphs $B_{S_t}^{(k)}$ induced by the k-neighborhoods of seeds in G_{t-1}. More formally, $p\left(\epsilon_t \mid G_{t-1}\right) = p\left(\epsilon_t \mid B_{S_t}^{(k)}, S_t\right)$, and the likelihood is $p\left(E_t \mid E_0\right) = \prod_{t'=1}^{t} p\left(\epsilon_{t'} \mid B_{S_{t'}}^{(k)}, S_{t'}\right)$.

While we cannot evaluate the likelihood in general, its structure is informative: Neighbors of node t only depend on the k-neighborhoods of the neighbors $u < t$ *it connects to* or its own k-neighborhood for nodes $v > t$ that *connect to it*. All information about the growth process is thus contained in the $k+1$-neighborhood of each node (due to dependence on the k-neighborhood of neighbors). We study four growing network models experimentally: Random attachment with Poisson-distributed number of stubs (0-localized), random attachment with two stubs and probabilistic one-step redirection[1] (2-localized), and two protein interaction models (duplication divergence with random mutation[10] (DMR; 1-localized) or complementation[11] (DMC)). Despite their similarity, the latter has no localization guarantees because edges may be removed. We use a gamma $(2, 1)$ prior for the Poisson rate and independent uniform priors for all other parameters.

As the likelihood remains intractable, we resort to simulation-based inference. We train NPDEs f to approximate the posterior $p\left(\theta \mid G\right)$ by minimizing the negative log probability loss $-\left\langle \log f\left(\theta, G\right)\right\rangle$, where θ are model parameters and $\langle\cdot\rangle$ denotes the expectation under the prior predictive distribution. Importantly, we now have a theoretical foundation for the neural architecture and employ graph isomorphism networks (GINs)[12]. A GIN with ℓ layers yields node representations based on their ℓ-neighborhood. We use a vector of ones as node features and obtain graph-level representations η by mean pooling the concatenated hidden representations of each layer. The final component of the NPDE depends on the specific network model, and we use a gamma distribution parameterized by dense neural networks applied to η for the Poisson rate. For all other parameters, we use beta distributions to approximate the posterior. Our conjecture that information is localized is supported by the results shown in Fig. 1: Performance improves with increasing GIN depth ℓ but saturates when or before $\ell = k + 1$. Even the non-monotonic DMC model does not benefit from deep GINs, suggesting that local features may be sufficient for inference for a broader class of models.

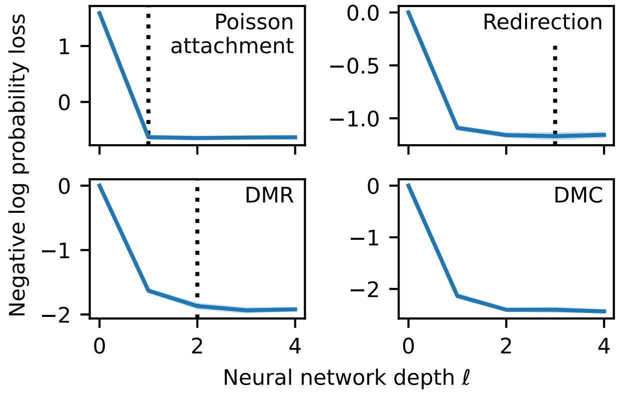

FIG. 1. *The performance of neural posterior density estimators using graph isomorphism networks (GINs) is consistent with theoretical information localization.* Each panel shows the negative log probability loss for a specific model as a function of GIN depth ℓ evaluated on a test set. Dotted lines indicate the predicted depth required for inference based on information localization.

We have not only offered theoretical arguments and empirical evidence for information localization in monotonic growth models but also presented NPDEs for simulation-based inference when the likelihood of mechanistic network models is intractable. In our experiments (results not shown), NPDEs have well-calibrated coverage and satisfy posterior predictive checks–even for non-local statistics such as the spectral gap.

[1] A. Vázquez, Phys. Rev. E **67**, 056104 (2003).
[2] D. Watts and S. Strogatz, Nature **393**, 440 (1998).
[3] L. Raynal and J.-P. Onnela, arXiv **2101**, 07766 (2021).
[4] C. Wiuf *et al.*, Proc. Natl. Acad. Sci. **103**, 7566 (2006).
[5] N. Arnold *et al.*, Sci. Rep. **11**, 5205 (2021).
[6] J. Overgoor *et al.*, in *WWW* (2019) pp. 1409–1420.
[7] G. Cantwell *et al.*, Phys. Rev. Lett. **126**, 038301 (2021).
[8] J. Larson and J.-P. Onnela, arXiv , 2106.09100 (2021).
[9] J.-M. Lueckmann *et al.*, in *AISTATS* (2021) pp. 343–351.
[10] R. Solé *et al.*, Adv. Complex. Syst. **5**, 43 (2002).
[11] A. Vázquez *et al.*, Complexus **1**, 38 (2003).
[12] K. Xu *et al.*, in *ICLR* (2019).