INFERRING DYNAMICAL SYSTEMS THROUGH ACTIVE QUERIES

Abhijin Adiga[1] *Chris J. Kuhlman*[1] *Madhav V. Marathe*[1]
S. S. Ravi[1,2] *Daniel J. Rosenkrantz*[2] *Richard E. Stearns*[2]
[1]*Biocomplexity Institute & Initiative, University of Virginia,* [2]*University at Albany – SUNY,*
{*abhijin, ckuhlman, mmarathe, ssravi*}*@virginia.edu, drosenkrantz@gmail.com, thestearns2@gmail.com*

SIAM Workshop on Network Science 2019
May 22–23 · Snowbird

Summary

Inferring the parameters of networked dynamical systems is currently a popular research topic. In a typical setting, model parameters are estimated from passive observations over which the user has no control. Here, we consider the problem of determining the local functions of a dynamical system by actively interacting with the system. The user submits queries to the system and infers the model from the outputs. We develop tight bounds on the number of queries needed, complexity results for producing optimal query sets and efficient algorithms that produce near-optimal query sets for several classes of deterministic and stochastic dynamical systems.

Background

Inferring unknown parameters of networked systems is currently a popular research area [5–7]. Here, we consider networked Boolean dynamical systems of the following general form. A **Graph Dynamical System** (GDS) $\mathcal{S}$ over $\{0,1\}$ is a pair $(G,\mathcal{F})$, where (a) $G(V,E)$ is an undirected graph, and (b) $\mathcal{F}=\{f_1,f_2,\ldots,f_n\}$ is a collection of **local functions**, with f_i being associated with node v_i. Each node of G has a state value from $\{0,1\}$. The inputs to function f_i are the state of v_i and those of the neighbors of v_i in G, and its output is the state of v_i. At any time t, the **configuration** $\mathcal{C}$ of a SyDS is the n-vector $(s_1^t,s_2^t,\ldots,s_n^t)$, where $s_i^t\in\mathbb{B}$ is the state of node v_i at time t $(1\le i\le n)$.

In this work, we study inference problems where the user has *control* over what information is extracted from the system through **queries** (Figure 1). The algorithm gives a set of configurations (or queries) to the system and infers properties using the system's responses. We study two query modes, namely **batch** and **adaptive**, that differ in their degrees of control. Under the batch mode, all the queries must be submitted together. Under the adaptive mode, queries can be submitted in several stages, and queries at a stage can depend on the answers to previous queries.

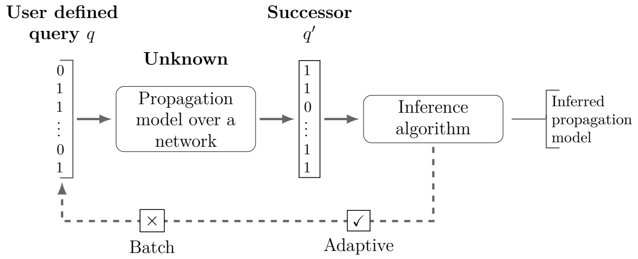

Figure 1: A schematic of the active querying framework.

Motivation for these problems comes in part from the DARPA Next Generation Social Science (NGS2) Program, where experimental data from social networks are used to infer properties of predictive models. Our work is similar in spirit—but quite different with respect to problem domain and results—from some recent work (e.g. [6]) where queries are used to infer users' choices from a finite set of ranked options.

Here, we summarize some of our recent work on inference of GDS by active querying [1–3]. These papers consider several classes of deterministic and stochastic GDSs. In each case, the focus is on developing (i) tight bounds on the number of queries required and (ii) efficient algorithms to obtain near-optimal query sets. To this end, we exploit well-studied graph theoretic problems such as coloring (both vertex and edge) and probabilistic methods such as Lovász local lemma, thus highlighting links between network structure and dynamics.

Deterministic GDS: Threshold and Symmetric local functions

We consider two classes of local functions, namely **threshold** and **symmetric** functions. They are defined below. (i) **Threshold functions:** The local function f_v associ-

ated with node v of a SyDS $\mathcal{S}$ is a t_v-**threshold** function for some integer $t_v \geq 0$ if the following condition holds: the value of f_v is 1 if the number of 1's in the input to f_v is *at least* t_v; otherwise, the value of the function is 0.
(ii) **Symmetric functions:** A local function f_v at node v is **symmetric** if the value of the function depends only on the number of 1's in the input. Clearly, threshold functions are a special case of symmetric functions.

We have developed algorithms for generating query sets under both batch and adaptive modes to infer local functions of dynamical systems [3]. In particular, for threshold dynamical systems, the objective is to infer the thresholds t_v of all vertices. Our results under the batch and adaptive modes are summarized below.

Batch mode. The following results apply to the general case of symmetric functions. We showed that a complete query set must contain at least $\Delta + 2$ queries, where Δ is the maximum degree of the graph. This result uses the fact that any vertex of degree Δ has $\Delta + 2$ possible threshold values. We developed an algorithm based on coloring the vertices of the *square* graph G^2, obtained by adding to graph G, the edges between all distance-two neighbors in G. The algorithm produces a complete query set of size at most $\min\{\Delta^2, n+1\}$, where n is the number of nodes in the graph. Our experiments on more than 20 real-world networks show that in practice, the algorithm yielded query sets whose sizes are very close to the lower bound. Using the Lovász local lemma, we also developed a randomized algorithms with an asymptotically better upper bound of $O\big(\Delta(\log\Delta)^{2.5}\big)$ [2].

Adaptive mode. We developed a greedy adaptive heuristic based on binary search for inferring threshold values of nodes. Our experimental results show that for most cases, it significantly outperforms the batch mode algorithms [3]. We showed that for the special case when the underlying network is a star graph on n nodes, the number of queries required is $\Theta(\log n)$.

Stochastic GDS: Independent Cascade model

Let $G(V, E)$ be a directed network where every edge $e \in E$ is associated with a (transmission or influence) probability $p_e > 0$. In the independent cascade (IC) model, at time t, a node v in state 0 is influenced independently by each in-neighbor u which changed to state 1 at time $t-1$ with influence probability $p_{(u,v)}$. A node in state 1 remains in that state forever. Here, the aim is to obtain provably good estimates of the edge probabilities by active querying. Our results for the IC model are summarized below.

For $0 < \epsilon, \delta < 1$, we presented an (ϵ, δ)-approximation algorithm to infer the edge probabilities of G for the IC model [1]. Formally, our algorithm ensures that for every edge e, the probability that the estimated probability $\hat{p}_e$ differs from the actual probability p_e by more than ϵp_e is at most δ. This approximation relies on two algorithmic ideas. First, it uses a stopping criterion for Monte Carlo sampling developed in [4]. Second, to minimize the number of queries used, it uses a novel edge coloring formulation (which we call *fan-out* edge coloring) for directed graphs.

In practice, edge sets of large social networks are partitioned into classes such that all the edges in the same class have the same transmission probability. We relied on this idea to make our algorithms scalable to very large social networks (with millions of nodes and hundreds of millions of edges). In particular, we formulated a combinatorial problem (which we call **Minimum Cost Covering Subgraph**) to identify a subgraph G' with a small number of nodes such that G' contains at least one edge from each class. We showed that this problem is **NP**-hard but developed an approximation algorithm which provides a performance guarantee of $O(\sqrt{k})$, where k is the number of classes. This allows us to work with the smaller subgraphs generated by the approximation algorithm.

References

[1] A. Adiga, V. Cedeno-Mieles, C. J. Kuhlman, M. V. Marathe, S. Ravi, D. J. Rosenkrantz, and R. E. Stearns. Inferring probabilistic contagion models over networks using active queries. In *Proc. CIKM*, pages 377–386. ACM, 2018.

[2] A. Adiga, C. J. Kuhlman, M. V. Marathe, S. S. Ravi, D. J. Rosenkrantz, and R. E. Stearns. Inferring users' choice functions in networked social systems through active queries. Conference Notes of COMSOC, 2018. (19 pages).

[3] A. Adiga, C. J. Kuhlman, M. V. Marathe, S. S. Ravi, D. J. Rosenkrantz, and R. E. Stearns. Learning the behavior of a dynamical system via a "20 questions" approach. In *Proc. AAAI*, pages 4630–4637, 2018.

[4] P. Dagum, R. Karp, M. Luby, and S. Ross. An optimal algorithm for Monte Carlo estimation. *SIAM J. Comput.*, 29(5):1484–1496, 2000.

[5] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno. The dynamics of protest recruitment through an online network. *Scientific Reports*, 1:7 pages, 2011.

[6] J. Kleinberg, S. Mullainathan, and J. Ugander. Comparison-based choices. arXiv:1705.05735v1 [cs.DS], May 2017.

[7] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proc. WWW*, pages 695–704. ACM, 2011.