The Value of Temporally Richer Data for Learning of Influence Networks

Munther A. Dahleh, John N. Tsitsiklis, and Spyros I. Zoumpoulis

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge MA 02139
{dahleh,jnt,szoumpou}@mit.edu

Working Paper
July 25, 2014

Abstract. We infer local relations of influence between networked entities from data on outcomes and assess the value of temporally richer data by characterizing the speed of learning when knowing the set of entities who take a particular action, versus when knowing the order that the entities take an action. We propose a parametric model of influence which captures directed pairwise interactions, formulate different variations of the learning problem, and provide theoretical guarantees for correct learning based on sets and sequences. The asymptotic gain of having access to richer temporal data for the speed of learning is thus quantified in terms of the gap between the derived asymptotic requirements under different data modes. Experiments on real data on mobile app installations quantify the improvement due to the availability of richer temporal data, and show that our maximum likelihood methodology recovers the underlying network well.

Keywords: Network inference, influence, social network, ordered/unordered data, mobile apps

1 Introduction

Consumers adopting a new product (Kempe, Kleinberg, and Tardos, 2003); an epidemic spreading across a population (Newman, 2002); a sovereign debt crisis hitting several countries (Glover and Richards-Shubik, 2013); a cellular process during which the expression of a gene affects the expression of other genes (Song, Kolar, and Xing, 2009); an article trending in the blogosphere (Lerman and Ghosh, 2010), a topic trending on an online social network (Zhou, Bandari, Kong, Qian, and Roychowdhury, 2010), computer malware spreading across a network (Kephart and White, 1991); all of these are temporal processes governed by local interactions of networked entities, which influence one another. Due to the increasing capability of data acquisition technologies, rich data on the outcomes of such processes are oftentimes available (possibly with time stamps), yet the underlying network of local interactions is hidden. In this work, we infer who influences whom in a network of interacting entities based on data of their actions/decisions, and quantify the gain of learning based on sequences of actions versus sets of actions. We answer the following question: how much faster can we learn influences with access to increasingly informative temporal data (sets versus sequences)?

Motivation. Clearly, having access to richer temporal information allows, in general, for faster and more accurate learning. Nevertheless, in some contexts, the temporally poor data mode of sets could provide almost all the information needed for learning, or at least suffice to learn key network relations. In addition, collecting, organizing, storing, and processing temporally richer data may require more effort and more cost. In some contexts, data on times of actions, or even sequences of actions, is noisy and unreliable; for example, the time marking of epilepsy seizure events is done by physicians on an empirical basis and is not exact. In some other contexts, having access to time stamps or sequences of actions is almost impossible. For example, in the context of retailing, data exist on sets of purchased items per customer (and are easily obtained by scanning the barcodes at checkout); however, no data exist on the order in which the items a customer checked out were picked up from the shelf (and obtaining such data would be practically hard). In this light, the question of quantifying the gain of learning with increasingly informative temporal data, and understanding in what scenarios learning with temporally poor data modes is good enough, is highly relevant in various contexts.

Background and related literature. Untangling and quantifying local influences in a principled manner, based on observed outcomes, is a challenging task, as there are many different confounding factors that may lead to seemingly similar phenomena. In recent work, inference of causal relationships has been possible from multivariate time-series data (Lozano and Sindhwani, 2010; Materassi and Salapaka, 2012; Kolar, Song, Ahmed, and Xing, 2010). Solutions for the influence discovery problem have been proposed, which, similarly to this work, treat time explicitly as a continuous random variable and infer the network through cascade data, e.g., Du, Song, Smola, and Yuan (2012); Myers and Leskovec (2010); Gomez-Rodriguez, Leskovec, and Krause (2010); Gomez-Rodriguez, Balduzzi, and Schölkopf (2011). However, the focus of our work is not just to infer the underlying network, but rather to quantify the gain in speed of learning, due to having access to richer temporal information.

Most closely related to this work are Amin, Heidari, and Kearns (2014); Abrahao, Chierichetti, Kleinberg, and Panconesi (2013); Netrapalli and Sanghavi (2012); Daneshmand, Gomez-Rodriguez, Song, and Schoelkopf (2014), which all derive sample/trace complexity results for the network inference problem. Amin et al. (2014), but also Gripon and Rabbat (2013), share with us the question of reconstructing a graph from traces defined as sets of unordered nodes. Similarly to Abrahao et al. (2013), we assume exponentially distributed infection times. Nevertheless, our scope differs from the works mentioned above, as we wish to compare explicitly the speed of learning when having access to datasets with times or sequences of actions, versus just sets of actions. Furthermore, the models assumed by the works mentioned above differ from the model we study, mainly in that we allow for self-induced infections (not just in the initial seeding), which makes the inference problem harder.

Another strand of recent research has focused on learning graphical models (which subsumes the question of identifying the connectivity in a network), either allowing for latent variables (e.g., Chandrasekaran, Parrilo, and Willsky, 2012; Choi, Tan, Anandkumar, and Willsky, 2011) or not (e.g., Anandkumar, Tan, Huang, and Willsky, 2012). Instead of proposing and learning a general graphical model, we focus on a simple parametric model that can capture the sequence and timing of actions naturally, without the descriptive burden of a standard graphical model.

Of relevance is also Shah and Zaman (2011), in which knowledge of both the graph and the set of infected nodes is used to infer the original source of an infection. In contrast, and somewhat conversely, we use knowledge of the set, order, or times of infections to infer the graph.

Last, economists have addressed the problem of identification in social interactions (e.g., Manski, 1993; Brock and Durlauf, 2001; Blume, Brock, Durlauf, and Ioannides, 2011; Bramoullé, Djebbari, and Fortin, 2009; Durlauf and Ioannides, 2010) focusing on determining aggregate effects of influence in a group; they classify social interactions into an endogenous effect, which is the effect of group members' behaviors on individual behavior; an exogenous (contextual) effect, which is the effect of group members' observable characteristics on individual behavior; and a correlated effect, which is the effect of group members' unobservable characteristics on individual behavior. In sharp contrast, our approach identifies influence at the individual, rather than the aggregate, level.

Overview. The overarching theme of our work is to quantify the gain in speed of learning of parametric models of influence, due to having access to richer temporal information. We seek to compare the speed of learning under three different cases of available data: (i) the data provides merely the set of agents/entities who took an action; (ii) the data provides the (ordered) sequence of agents/entities who took an action, but not the times; and (iii) the data provides the times of the actions. It is clear that learning is no slower with times than it is with sequences, and no slower with sequences than with sets; yet, what can we say about *how much* faster learning is with times than with sequences, and with sequences than with sets? This is, to the best of our knowledge, a comparison that has not been studied systematically before.[1] In this paper[2], we focus on the comparison between learning with sets and learning with sequences.

We propose a parametric model of influence which captures directed pairwise interactions and provide theoretical guarantees on the sample complexity

[1] Netrapalli and Sanghavi (2012) find such a comparison highly relevant.

[2] Most of the material in this paper is presented in the Ph.D. thesis of Zoumpoulis (2014).

for correct learning with sets and sequences. Our results characterize the sufficient and necessary scaling of the number of i.i.d. samples required for correct learning. The asymptotic gain of having access to richer temporal data à propos of the speed of learning is thus quantified in terms of the gap between the derived asymptotic requirements under different data modes. We first assume prior knowledge of a "super graph" that includes all the candidate edges, and we infer which edges of the super graph truly exist; restricting to each edge having either very large or no influence, we provide sufficient and necessary conditions on the graph topology for learnability, and we come up with upper and lower bounds for the minimum number of i.i.d. samples required to learn the correct hypothesis for the star topology, for different variations of the learning problem: learning one edge or learning all the edges, under different prior knowledge over the hypotheses, under different scaling of the horizon rate, and learning with sets or with sequences. We then study more general networks and relax the assumption that each edge carries an influence rate that is either very large or zero; we provide a learning algorithm and theoretical guarantees on the sample complexity for correct learning in the hard problem of telling between the complete graph and the complete graph that is missing one edge.

We also evaluate learning with sets and sequences *experimentally.* Given real data on outcomes, we learn the parametric influence model by maximum likelihood estimation. The value of learning with data of richer temporal detail is quantified, and our methodology is shown to recover the underlying network structure well. The real data come from observations of mobile app installations of users, along with data on their communications and social relations.

2 The Model

A product becomes available at time $t = 0$ and each of $n + 1$ agents may adopt it or not. (In this paper the word "product" is used throughout, but could be interchanged by any of the following: information, behavior, opinion, disease, etc., depending on the context.) Agent i adopts it at a time that is exponentially distributed with rate $\lambda_i \geq 0$. After agent i adopts, the rate of adoption for all other agents $j \neq i$ increases by $\lambda_{ij} \geq 0$. The overall time horizon of the adoption and infection process is modeled as an exponentially distributed random variable with rate λ_{hor}. No adoptions are possible after the end of the horizon.[3]

We study the adoption decisions for a collection of products, assuming that the parameters are static across products, and adoptions across products are independent.[4]

[3] The proposed cascade model suggests a recursive definition for the times of adoption for each agent given a product c, which we denote $\{T_c^i\}_{i=1}^{n+1}$. We define $T_c^i = \infty$ if agent i does not adopt product c.

[4] Given product c, we consider the following three data modes:

- learning with *sets* of adoptions: the learner observes vector $\left(\mathbb{1}_{\{T_c^1<\infty\}}, \dots, \mathbb{1}_{\{T_c^{n+1}<\infty\}}\right)$;

3 Theoretical Guarantees for Learning Influence in Networks with Zero/Infinity Edges

A directed[5] graph $G = (\mathcal{V}, \mathcal{E})$ is a priori given and $\lambda_{ij} = 0$ if edge (i, j) is not in $\mathcal{E}$. In this section, we provide theoretical guarantees for learning for the case where each edge in $\mathcal{E}$ carries an influence rate of either zero or infinity, casting the decision problem as a hypothesis testing problem. We restrict to influence rates that are either zero or infinite in order to simplify the analysis and derive crisp and insightful results. Given a graph G, lower and upper bounds for the number of i.i.d. products required to learn the correct hypothesis can be sought for different variations of the problem, according to the following axes:

- **Learning one edge versus learning all edges:** We pose two decision problems: learning the influence rate λ_{ij} between two specified agents i, j; and learning *all* the influence rates $\lambda_{ij}, i \neq j$.
- **Different prior knowledge over the hypotheses:** We study this question in the Bayesian setting of assuming a prior on the hypotheses, in the worst case over the hypotheses, as well as in the setting in which we know how many edges carry infinite influence rate. In general, a high prior probability of each edge carrying infinite influence rate, or knowing that a high number of edges carry infinite influence rate, correspond to the case of dense graphs; a low prior probability of each edge carrying infinite influence rate, or knowing that a low number of edges carry infinite influence rate, correspond to the case of sparse graphs, with few influence relations.
- **Different data modes:** We characterize the growth of the minimum number of i.i.d. products required for learning with respect to the number of agents n, when the available data provides information on sets or sequences of adoptions.
- **Different scaling of the horizon rate with respect to the idiosyncratic rates:** We consider different scalings of λ_{hor} with respect to the idiosyncratic rates $\lambda_1, \ldots, \lambda_n$. Small values of λ_{hor} correspond to large horizon windows, during which many agents get to adopt; large values of λ_{hor} correspond to small horizon windows, during which only few agents get to adopt.

We first discuss conditions on the graph topology that guarantee learnability, and then we carry out the proposed program for the star topology. The star topology is one of the simplest non-trivial topologies, and is illustrative of the difference in the sample complexity between learning scenarios with information of different temporal detail.

- learning with *sequences* of adoptions: the learner observes vector $\left(R_c^1, \ldots, R_c^{n+1}\right)$, where R_c^i denotes the rank of T_c^i in $\{T_c^j\}_{j=1}^{n+1}$. If $T_c^i = \infty$, define $R_c^i = \infty$.
- learning with *times* of adoptions: the learner observes vector $\left(T_c^1, \ldots, T_c^{n+1}\right)$.

[5] We allow bi-directed edges.

3.1 Conditions on Topology for Learnability

We say that a graph is *learnable* if there exists an algorithm that learns all edges with probability of error that decays to zero in the limit of many samples. We show what graphs are learnable when learning with sets, assuming all edges carry influence zero or infinity. Adopting a Bayesian approach, we assume that if an edge exists in the sets of edges $\mathcal{E}$ of graph G, then the edge carries infinite influence with probability $q, 0 < q < 1$, and zero influence with probability $1 - q$. We also assume that the realization of each edge is independent of the realizations of all other edges. Last, we assume all idiosyncratic rates and the horizon rate to be equal to some $\lambda > 0$, which can be known or unknown.

Proposition 1. *When learning with sets, if the graph G has distinct nodes i, j, h such that*

(i) $(i, j) \in \mathcal{E}$, *and*
(ii) there exists a directed path from i to j through h,

then the graph is not learnable.[6] *If such triplet of distinct nodes does not exist, then the graph G is learnable, using $O(n^2 \log n)$ products. In particular, any polytree*[7] *is learnable with sets, using $O(n^2 \log n)$ products.*

3.2 Learning Influence in the Star Network

We consider the hypothesis testing problem in which each of n agents influence agent $n + 1$ either with rate zero or infinity. (A rate of $\lambda_{i,n+1} = \infty$ signifies that agent $n + 1$ adopts right when agent i adopts.) Each of agents $1, \ldots, n$ adopts with rate $\lambda > 0$, which can be known or unknown, while agent $n + 1$ does not adopt unless she is triggered to. There is no influence from agent $n + 1$ to any of the agents $1, \ldots, n$, or from any of the agents $1, \ldots, n$ to any of the agents $1, \ldots, n$. Figure 1 illustrates this model.

We consider two settings for the horizon rate: the setting in which the horizon rate is equal to the agents' idiosyncratic rate of adoption, that is, $\lambda_{hor} = \lambda$, and the setting $\lambda_{hor} = n\lambda$. We pose two decision problems: learning the influence rate between a specified agent i and agent $n + 1$; and learning all the influence rates $\lambda_{1,n+1}, \ldots, \lambda_{n,n+1}$. We study the Bayesian setting of assuming a prior on the hypotheses, the setting of the worst case over the hypotheses, as well as the setting in which we know how many agents have infinite influence rate and how many have zero influence. We characterize the growth of the minimum number of i.i.d. products required for learning with respect to n, both when the available data provides information on sets of adopters, and when the available data provides information on sequences of adopters. (Of course, knowledge of times of adoptions will not induce a gain over knowledge of sequences, because of our assumption that the influence rates are either zero or infinite, and $\lambda_{n+1} = 0$.)

[6] All proofs are relegated to the Appendix.

[7] A *polytree* is a directed acyclic graph (DAG) whose underlying undirected graph is a tree.

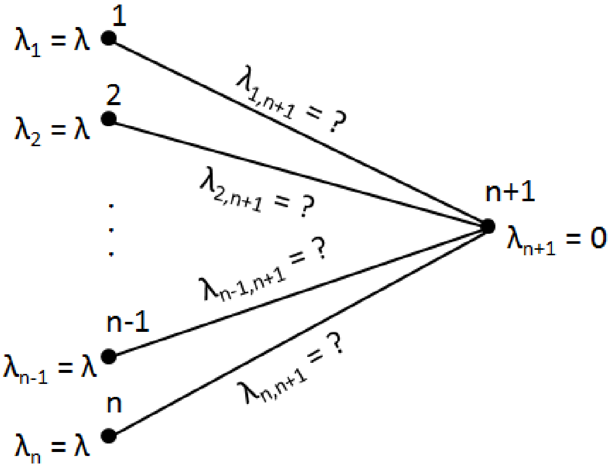

Fig. 1. The hypothesis testing problem: what influence does each link carry to the star agent $(n+1)$: infinite or zero?

The Bayesian setting

In the Bayesian setting, we assume that the influence rate on each link is infinite, with probability p, and zero, with probability $1-p$, and that the selection of the rate for each link is independent of the selection for other links.

The case $p = 1/2$. We assume that the influence rate on each link will be zero or infinite with equal probability. Table 1 summarizes the results on the necessary and sufficient number of i.i.d. products for learning.

Table 1. Matching lower and upper bounds for the minimum number of i.i.d. products required to learn the influence model in terms of n, in the Bayesian setting when $p = 1/2$, for the two cases of learning the influence between one agent and the star agent and of learning the influence between all agents and the star agent, and for the two cases of learning based on sets of adoptions or sequences of adoptions.

	$\lambda_{hor} = \lambda$		$\lambda_{hor} = n\lambda$	
	Sets	Sequences	Sets	Sequences
Learn one	$\Theta(n^2)$	$\Theta(n)$	$\Theta(n)$	$\Theta(n)$
Learn all	$\Theta(n^2 \log n)$	$\Theta(n \log n)$	$\Theta(n \log n)$	$\Theta(n \log n)$

For example, for the case when $\lambda_{hor} = \lambda$, we have the following sample complexity results[8].

Proposition 2. *To ensure correct learning of $\lambda_{1,n+1}, \ldots, \lambda_{n,n+1}$ with probability $1-\delta$ based on sets of adopting agents, it is sufficient for the number of i.i.d. products to be $O(n^2 \log \frac{n}{\delta})$, and necessary for the number of i.i.d. products to be $\Omega(n^2 \log n)$. To ensure correct learning of $\lambda_{1,n+1}, \ldots, \lambda_{n,n+1}$ with probability $1-\delta$ based on sequences of adoptions, it is sufficient for the number of i.i.d. products to be $O(n \log \frac{n}{\delta})$, and necessary for the number of i.i.d. products to be $\Omega(n \log n)$.*

The case $p = 1/n$. We assume that the influence rate on each link will be infinite with probability $p = 1/n$. (In this case, the expected number of agents who can influence agent $n+1$ is $\Theta(1)$.) Table 2 summarizes the results on the necessary and sufficient number of i.i.d. products for learning.

Table 2. Matching lower and upper bounds for the minimum number of i.i.d. products required to learn the influence model, in terms of n, in the Bayesian setting when $p = 1/n$, when learning the influence between all agents and the star agent, for the two cases of learning based on sets of adoptions or sequences of adoptions. Notice that no products are needed to learn just one influence rate; an estimator can just guess that $\lambda_{i,n+1} = 0$.

	$\lambda_{hor} = \lambda$		$\lambda_{hor} = n\lambda$	
	Sets	Sequences	Sets	Sequences
Learn all	$\Theta(\log n)$	$\Theta(1)$	$\Theta(n)$	$\Theta(n)$

The worst-case setting

In the worst-case setting, we assume that each of the influence rates $\lambda_{1,n+1}, \ldots, \lambda_{n,n+1}$ can be either zero or infinity, but we assume no prior over the hypotheses. We provide upper and lower bounds for the minimum number of i.i.d. products required to learn the correct hypothesis assuming that the influence rates on the links are such that the minimum number of i.i.d. products required for learning is maximized (the worst possible). Table 3 summarizes the results on the necessary and sufficient number of i.i.d. products for learning.

The worst-case setting with known scaling of agents with influence rate infinity to agent $n+1$

We denote the number of agents with influence rate infinity to agent $n+1$ by ℓ.

[8] For the sake of brevity, we refrain from providing propositions or proofs for the rest of our results in this section, which are compactly stated in the tables.

Table 3. Matching lower and upper bounds for the minimum number of i.i.d. products required to learn the influence model in terms of n, in the worst-case setting, for the two cases of learning the influence between one agent and the star agent and of learning the influence between all agents and the star agent, and for the two cases of learning based on sets of adoptions or sequences of adoptions.

	$\lambda_{hor} = \lambda$		$\lambda_{hor} = n\lambda$	
	Sets	Sequences	Sets	Sequences
Learn one	$\Theta(n^2)$	$\Theta(n)$	$\Theta(n)$	$\Theta(n)$
Learn all	$\Theta(n^2 \log n)$	$\Theta(n \log n)$	$\Theta(n \log n)$	$\Theta(n \log n)$

Table 4 summarizes the results on the necessary and sufficient number of i.i.d. products for learning.

Table 4. Matching lower and upper bounds for the minimum number of i.i.d. products required to learn the influence model in terms of n, in the worst-case setting when the scaling of agents ℓ with influence rate infinity to agent $n+1$ is known, for the two cases of learning based on sets of adoptions or sequences of adoptions.

	$\lambda_{hor} = \lambda$		$\lambda_{hor} = n\lambda$	
	Sets	Sequences	Sets	Sequences
$\ell = 1$	$\Theta(\log n)$	$\Theta(1)$	$\Theta(n)$	$\Theta(n)$
$\ell = \alpha n, \alpha \in (0,1)$	$\Theta(n^2 \log n)$	$\Theta(n \log n)$	$\Theta(n \log n)$	$\Theta(n \log n)$
$\ell = n-1$	$\Theta(n^2)$	$\Theta(n)$	$\Theta(n)$	$\Theta(n)$

3.3 Discussion

We characterize the scaling of the number of samples required for learning with sets and sequences, thus theoretically quantifying the gain of learning with sequences over learning with sets in regard to the speed of learning. Our inference algorithms look for signature events, and attain optimal sample complexity, as long as the signature events are reasonably chosen. Depending on the setting, learning with sets can take a multiplicative factor of $\Theta(n)$ more samples than learning with sequences, when the horizon rate is moderate (i.e., as large as the idiosyncratic rates of adoption). With much smaller horizon, learning with sequences has no gain asymptotically over learning with mere sets, across all the settings we study; when the observation window (i.e., the horizon) is small, then the sets of adoptions provide asymptotically all the information pertinent to learning that sequences provide. The intuition behind this finding is that, with smaller horizon, only a few adoptions take place; when only a few adoptions take place, sequences do not convey much more information than sets do.

4 Theoretical Guarantees for General Networks

In this section, we provide theoretical guarantees for more general networks than the star topology, which was considered in Subsection 3.2, and we relax the assumption that each edge carries an influence rate of either zero or infinity. In particular, we focus on the question of deciding between the complete graph, and the complete graph that is missing one directed edge, which we cast as a binary hypothesis testing problem[9]. This is a hard hypothesis testing problem, in the sense that the two hypotheses give rise to very similar outcomes. Because of its nature, sample complexity results for this hard problem entail sample complexity results for broader families of networks.

4.1 Learning Between the Complete Graph and the Complete Graph that Is Missing One Edge

For ease of exposition, we assume that all from a collection of n agents have the same idiosyncratic rate $\lambda > 0$, and that all directed edges carry the same influence rate, which is equal to λ. λ can be known or unknown. We are learning between two hypotheses for the underlying influence graph: the complete graph, P_1; and the complete graph minus the directed edge (i, j), P_2.

An Algorithm for Learning

We propose a simple algorithm for deciding between the two hypotheses. The sample complexity of our algorithm gives an upper bound for the number of i.i.d. products required for learning. The algorithm is the following:

- We first choose an event of interest A in an appropriate manner.
- For each new product ℓ, we define an indicator variable

$$I_\ell = \begin{cases} 1 & \text{if event } A \text{ obtained in product } \ell \\ 0 & \text{otherwise} \end{cases}$$

- After k i.i.d. products, we compute $\hat{p} = \frac{1}{k}\sum_{\ell=1}^{k} I_k$.
- Choose the hypothesis with the smallest deviation from the empirical probability, $|P_i(A) - \hat{p}|, i = 1, 2$.

By the concentration inequality

$$\mathbb{P}\left(|\hat{p} - \mathbb{E}[I]| \geq t\right) \leq 2e^{-2kt^2},$$

and setting $2e^{-2kt^2} \leq \delta$, $0 < \delta < 1$, we obtain

$$k \geq \frac{\log\left(\frac{2}{\delta}\right)}{2t^2}.$$

[9] Abrahao et al. (2013) study the same binary hypothesis testing problem.

Therefore, setting $t = 0.5 \cdot |\mathbb{E}_1[I] - \mathbb{E}_2[I]| = 0.5 \cdot |P_1(A) - P_2(A)|$, the proposed algorithm learns the true hypothesis correctly with probability at least $1 - \delta$.

The sample complexity of the proposed learning algorithm is given effectively by the inverse square of the distance $|\mathbb{E}_1[I] - \mathbb{E}_2[I]| = |P_1(A) - P_2(A)|$, which scales with the number of agents n.

An alternative derivation of an upper bound is through obtaining a lower bound on the Kullback-Leibler divergence between the two distributions P_1, P_2. In particular, in a Neyman-Pearson setting, the best achievable exponent for the probability of error of deciding in favor of the first hypothesis when the second is true, given that the probability of deciding in favor of the second hypothesis when the first is true is less than ϵ, is given by the negative Kullback-Leibler (KL) divergence, $-D(P_1||P_2)$. In turn, Pinsker's inequality bounds the KL divergence from below:

$$\begin{aligned} D(P_1||P_2) &\geq \frac{1}{2\log 2}||P_1 - P_2||_1^2 \\ &= \frac{1}{2\log 2}\Big(2\big(P_1(B) - P_2(B)\big)\Big)^2, \end{aligned}$$

where $B = \{x : P_1(x) > P_2(x)\}$. Therefore, the larger is the event A of interest in the algorithm proposed above, i.e., the closer it gets to the event B, the tighter upper bound we achieve for the number of i.i.d. products required for learning.

Learning with Sequences

Proposition 3. *To ensure correct learning of the true hypothesis with sequences, it is sufficient for the number of i.i.d. products to be $O(n^2)$.*

We now argue that there is a matching lower bound of $\Omega(n^2)$ for learning with sequences. Indeed, assuming we are learning based on not just sequences, bur rather times of adoptions, then the ratio of the likelihoods for the time of adoption for agent j between hypotheses P_1 and P_2, assuming everybody else has adopted, is

$$\frac{n\lambda e^{-n\lambda t}}{(n-1)\lambda e^{-(n-1)\lambda t}} = \left(1 + \frac{1}{n-1}\right) e^{-\lambda t},$$

resulting in a KL divergence of $\Theta\left(\frac{1}{n^2}\right)$, which in turn implies a $\Omega(n^2)$ complexity for the number of i.i.d. products required for learning. This is in agreement with the $\Omega(\frac{n^2}{\log^2 n})$ lower bound proven in Abrahao et al. (2013), for a model with exponential infection times, but no idiosyncratic adoptions.

Learning with Sets

Proposition 4. *To ensure correct learning of the true hypothesis with sets, it is sufficient for the number of i.i.d. products to be $O(n^6)$.*

4.2 Discussion

We have proposed a simple algorithm for deciding between the complete graph (with all directed edges present), and the complete graph that is missing one directed edge. The algorithm relies on using samples to estimate the probability of an event of interest under each of the two hypotheses. Our algorithm results in sample complexity that is given by the inverse of the square of the difference between the probabilities of the chosen event of interest under each of the two hypotheses. This difference scales with the number of agents n. The dependence on the inverse of the square of the difference can also be derived from a lower bound on the Kullback-Leibler divergence via Pinsker's inequality. One would choose the event of interest in the algorithm so as to maximize the difference between the event's probabilities under each of the two hypotheses, resulting in a tighter upper bound.

When learning with sequences, we propose an implementation of our algorithm that can learn with $O(n^2)$ samples, and we argue that learning is not possible with $o(n^2)$ samples. When learning with sets, we propose an implementation of our algorithm that can learn with $O(n^6)$ samples, and although we are missing a lower bound, we conjecture that $O(n^2)$ samples do not suffice for correct learning with sets.

5 Learning Influence with Real Observational Data

5.1 The Dataset

We use data[10] obtained from an experiment (Pan, Aharony, and Pentland, 2011) for which an Android mobile phone is given to each of 55 participants, all residents of a graduate student dormitory at a major US university campus, and the following information is tracked during the experimental period of four months:

- installations of mobile apps, along with associated time stamps;
- calls among users (number of calls for each pair of users);
- Bluetooth hits among users (number of Bluetooth radio hits for each pair of users);
- declared affiliation (in terms of academic department) and declared friendship among users (binary value denoting affiliation for each pair of users, and similarly for friendship).

5.2 Network Inference

We are interested in recovering patterns of influence among the 55 participants based solely on the mobile app installation data. Under the premise that influence travels through communication and social interaction, we expect a network

[10] We are thankful to Sandy Pentland and the Human Dynamics Laboratory at the MIT Media Lab for sharing the data with us.

inference algorithm that does well to recover a network of influence that is highly correlated with the realized network of social interaction. We therefore separate the available data into the mobile app installation data (i.e., the "actions") and the communication/interaction data (i.e., the social data). We learn the influence network using only the actions, and we then validate using the social data.

We employ maximum likelihood estimation based on sequences of mobile app installations to estimate both the influence rates (i.e., the network structure) and the idiosyncratic rates of adoption. The inferred influence rates are highly correlated with the realized communication networks, providing evidence for the soundness of the proposed inference methodology, as we proceed to show.

For each edge (i, j), we add the inferred rates $\lambda_{ij} + \lambda_{ji}$, and rank all edges based on joint inferred influence. We then choose the top ranked edges based on joint influence, and we report the percentage of edges for which friendship was reported. A friendship edge exists between two randomly selected nodes with probability 0.3508, which is less than the percentage corresponding to the $10, 20, 50, 100$ edges that carry the highest inferred influence. Table 5 shows the results.

Table 5. There is higher probability of friendship in the edges where we detect the highest influence (using sequences) as compared to the random baseline. A friendship edge exists between two randomly selected nodes in the dataset with probability 0.3508.

Out of top	10	joint influence edges, friendship exists in	70%
	20		65%
	50		54%
	100		38%

The correlation coefficient between the observations of calls and the inferred (joint) influence (using information on sequences of installations) per edge is 0.3381. The positive correlation between calls and (joint) influence inferred from sequences is visualized in Figure 2.

We finally also estimate influence based solely on sets of mobile app installations. We restrict our attention to the most active users out of the 55 participants, and learn influence among them using sets and sequences. When we learn influence based on sets of mobile app installations, as opposed to sequences, the correlation between the inferred rates and the realized communication/interaction networks is only slightly lower (or even higher) than when learning with sequences of mobile app installations. This is an instance where learning with temporally poor data modes is good enough.

6 Conclusion

We have discussed the gain of learning with sequences of actions versus sets of actions under three different angles: in Section 3 we have assumed prior knowl-

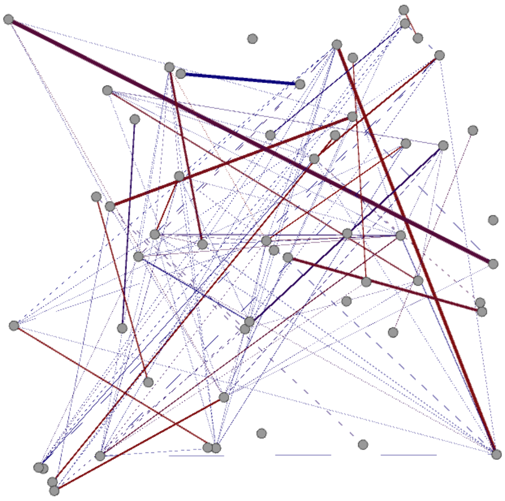

Fig. 2. Concurrent visualization of the realized network of calls (the color of each edge denotes the number of calls between the users: closer to blue for lower number of calls, closer to red for higher number of calls) and the inferred network of influence using information on sequences of adoptions (the thickness of edge $i - j$ is proportional to the sum of the inferred influence rates $\lambda_{ij} + \lambda_{ji}$). We observe that edges with higher number of calls (red) are more likely to carry higher influence (thick).

edge of a "super graph", and under the condition that each edge either carries very large influence or no influence, we infer which edges of the super graph truly exist; in Section 4, we solve a binary hypothesis testing problem which decides the true network topology between two complex candidate hypotheses; in Section 5, we recover the network of influence using a maximum likelihood estimator based on real data. In Zoumpoulis (2014, Chapter 6) we also compare learning with times to learning with sequences and sets: we formulate relevant hypothesis testing problems and characterize the speed of learning of the correct hypothesis via the Kullback-Leibler divergence, under the data modes of sets, sequences, and times; we conclude that when the horizon is small, the sets of decisions provide almost all the necessary information for learning, and there is no value in richer temporal data, which is in agreement with our findings in Section 3 of this paper. Overall, our focus has been on whether having access to data of richer temporal information (such as sequences of actions) has value over having access to mere sets of actions in order to learn the underlying influence network.

In a different formulation, the learner knows that the true network lies within a family of topologies, and the learner is after recovering the true network. For trees, a class of interest, we can prove that $O(\log n)$ samples are sufficient when learning with sets, and $\Omega(\log n)$ samples are necessary when learning with sequences. We thus show that sequences have no value over sets asymptotically when learning the influence network among all trees. Along with trees, we are currently working towards learning results for other classes of networks.

In addition, one can study models of influence other than the exponential delay and random observation window model used for our results so far. Whether employing a different infection and horizon model would alter the results on the value of having access to data of richer temporal information remains unanswered.

Acknowledgments

We thank three anonymous WINE 2014 reviewers for their useful comments and suggestions. This research was partially supported by the following grants: AFOSR FA9550-09-1-0420, AFOSR MURI/GaTech R6756-G2, ONR/GaTech R0897-G1.

Bibliography

Abrahao, B., F. Chierichetti, R. Kleinberg, and A. Panconesi (2013), "Trace complexity of network inference." In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 491–499.

Amin, Kareem, Hoda Heidari, and Michael Kearns (2014), "Learning from contagion (without timestamps)." In *Proceedings of the 31st International Conference on Machine Learning, JMLR W&CP*, volume 32, 1845–1853.

Anandkumar, A., V.Y.F Tan, F. Huang, and A.S. Willsky (2012), "High-dimensional structure learning of Ising models: Local separation criterion." *Annals of Statistics*, 40, 1771–1812.

Blume, L. E., W. A. Brock, S. N. Durlauf, and Y. M. Ioannides (2011), *Identification of social interactions*, volume 1B of *Handbook of Social Economics*, Chapter 18, 853–964. Elsevier B.V., The Netherlands: North-Holland.

Bramoullé, Y., H. Djebbari, and B. Fortin (2009), "Identification of peer effects through social networks." *Journal of Econometrics*, 150, 41–55.

Brock, W. A. and S. N. Durlauf (2001), *Interaction-based models*, first edition, volume 5 of *Handbook of Econometrics*, Chapter 54, 3297–3380. Elsevier, Amsterdam: North-Holland.

Chandrasekaran, V., P. A. Parrilo, and A. S. Willsky (2012), "Latent variable graphical model selection via convex optimization." *Annals of Statistics*, 40, 1935–1967.

Choi, M.J., V. Tan, A. Anandkumar, and A. Willsky (2011), "Learning latent tree graphical models." *Journal of Machine Learning Research*, 12, 1771–1812.

Daneshmand, Hadi, Manuel Gomez-Rodriguez, Le Song, and Bernhard Schoelkopf (2014), "Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm." In *Proceedings of the 31st International Conference on Machine Learning, JMLR W&CP*, volume 32, 793–801.

Du, N., L. Song, A. Smola, and M. Yuan (2012), "Learning networks of heterogeneous influence." *Advances in Neural Information Processing Systems*, 25.

Durlauf, S. N. and Y. M. Ioannides (2010), "Social interactions." *Annual Review of Economics*, 2, 451–478.

Glover, Brent and Seth Richards-Shubik (2013), "Sovereign debt crises and financial contagion." Carnegie Mellon University, preliminary draft.

Gomez-Rodriguez, M., D. Balduzzi, and B. Schölkopf (2011), "Uncovering the temporal dynamics of diffusion networks." In *Proceedings of the 28th International Conference on Machine Learning.*

Gomez-Rodriguez, M., J. Leskovec, and A. Krause (2010), "Inferring networks of diffusion and influence." In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1019–1028.

Gripon, V. and M. Rabbat (2013), "Reconstructing a graph from path traces." In *Proceedings of 2013 IEEE International Symposium on Information Theory.*

Kempe, D., J. Kleinberg, and E. Tardos (2003), "Maximizing the spread of influence through a social network." In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 137–146.

Kephart, J. O. and S. R. White (1991), "Directed-graph epidemiological models of computer viruses." In *Proceedings of IEEE Symposium on Security and Privacy*, 343–359.

Kolar, M., L. Song, A. Ahmed, and E.P. Xing (2010), "Estimating time-varying networks." *Annals of Applied Statistics*, 4, 94–123.

Lerman, K. and R. Ghosh (2010), "Information contagion: An empirical study of the spread of news on Digg and Twitter social networks." In *Proceedings of the International AAAI Conference on Weblogs and Social Media.*

Lozano, A. C. and V. Sindhwani (2010), "Block variable selection in multivariate regression and high-dimensional causal inference." *Advances in Neural Information Processing Systems*, 23.

Manski, C. (1993), "Identification of endogenous social effects: The reflection problem." *Review of Economic Studies*, 60, 531–542.

Materassi, D. and M.V. Salapaka (2012), "On the problem of reconstructing an unknown topology via locality properties of the Wiener filter." *IEEE Transactions on Automatic Control*, 57, 1765–1777.

Myers, S. and J. Leskovec (2010), "On the convexity of latent social network inference." *Advances in Neural Information Processing Systems*, 23.

Netrapalli, P. and S. Sanghavi (2012), "Finding the graph of epidemic cascades." In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, 211–222.

Newman, M. E. J. (2002), "The spread of epidemic disease on networks." *Physical Review E*, 66, 016128.

Pan, W., N. Aharony, and A. "S." Pentland (2011), "Composite social network for predicting mobile apps installation." In *Proceedings of 25th Conference on Artificial Intelligence, AAAI.*

Shah, D. and T. Zaman (2011), "Rumors in a network: Who's the culprit?" *IEEE Transactions on Information Theory*, 57, 5163–5181.

Song, Le, Mladen Kolar, and Eric P. Xing (2009), "Keller: Estimating time-varying interactions between genes." *Bioinformatics*, 25, i128–i136.

Zhou, Z., R. Bandari, J. Kong, H. Qian, and V. Roychowdhury (2010), "Information resonance on Twitter: Watching Iran." In *Proceedings of the First Workshop on Social Media Analytics, ACM*, 123–131.

Zoumpoulis, Spyros I. (2014), *Networks, Decisions, and Outcomes: Coordination with Local Information and the Value of Temporal Data for Learning Influence Networks.* Ph.D. thesis, Massachusetts Institute of Technology.

Appendix

Proof of Proposition 1

We focus on learning edge i, j. We first show the first half of the proposition.

We show that there is a region (which we call BAD) of large probability, where it is not clear what a good estimator should decide. No matter how this event is split between the competing hypotheses, the probability of error will be large.

We use X_p to denote the outcome of product p. We say the outcome X_p of product p is in BAD_p if one of the following happens: both agents i, j adopt; agent i does not adopt. We say the outcome $X_1, \ldots, X_k$ is in BAD if $X_p \in \mathrm{BAD}_p$ for all products $p = 1, \ldots, k$.

We can write

$$\begin{aligned}
\mathbb{P}\left((X_1, \ldots, X_k) \in \mathrm{BAD} \mid \lambda_{ij} = 0\right) &\geq \mathbb{P}\left(\text{path from } i \text{ to } j \text{ realized} \mid \lambda_{ij} = 0\right) \\
&\geq \mathbb{P}\left(\text{paths from } i \text{ to } h \text{ and from } h \text{ to } j \text{ realized} \mid \lambda_{ij} = 0\right) \\
&= q^{\ell_{ih} + \ell_{hj}} \\
&> 0,
\end{aligned}$$

where $\ell_{ih}(\ell_{hj})$ is the number of edges along a path from i to h (from h to j). Note that this is independent of the number of products k.

To show the second half of the proposition, we consider the following estimator: after k products, decide $\hat{\lambda}_{ij} = 0$ if there is a product such that agent i adopts and agent j does not adopt; otherwise, decide $\hat{\lambda}_{ij} = \infty$. Conditioning on the subset of agents $\mathcal{L}$ for which there is a directed path of edges carrying infinite influence to j, we can write

$$\begin{aligned}
\mathbb{P}(\text{error}) &= \mathbb{P}(\lambda_{ij} = 0) \cdot \mathbb{P}\left(\hat{\lambda}_{ij} = \infty \mid \lambda_{ij} = 0\right) \\
&= (1 - q) \cdot \sum_{\mathcal{L} \subseteq \{1, \ldots, n\} \setminus \{i, j\}} \mathbb{P}\left(\hat{\lambda}_{ij} = \infty \mid \lambda_{ij} = 0, \mathcal{L}\right) \mathbb{P}\left(\mathcal{L} \mid \lambda_{ij} = 0\right) . (1)
\end{aligned}$$

Assuming $|\mathcal{L}| = m$, we can write for a given product:

$$\begin{aligned}
\mathbb{P}\left(i \text{ adopts, } j \text{ does not} \mid \lambda_{ij} = 0, \mathcal{L}\right) &\geq \mathbb{P}\left(i \text{ adopts first, } j \text{ does not adopt} \mid \lambda_{ij} = 0, \mathcal{L}\right) \\
&= \frac{\lambda}{n\lambda + \lambda} \cdot \frac{\lambda}{\lambda + m\lambda + \lambda} \\
&= \frac{1}{n+1} \cdot \frac{1}{m+2}.
\end{aligned}$$

Denoting with M the random variable which is the number of agents for which there is a directed path of edges carrying infinite influence to j (i.e., $M = |\mathcal{L}|$), we can now rewrite Equation (1) as

$$\begin{aligned}\mathbb{P}(\text{error}) &\leq (1-q)\cdot\sum_{m=0}^{n-2}\left(1-\frac{1}{(n+1)(m+2)}\right)^k p_{M|\lambda_{ij}=0}(m)\\ &\leq (1-q)\left(1-\frac{1}{(n+1)n}\right)^k\\ &= (1-q)\left(\frac{n(n+1)-1}{n(n+1)}\right)^k \longrightarrow 0 \text{ as } k\longrightarrow\infty.\end{aligned}$$

In fact, assuming $q=\Theta(1)$, to ensure an accurate estimate for λ_{ij} with probability at least $1-\delta$, for given $\delta\in(0,1)$, it suffices that $k\geq\frac{\log\frac{1-q}{\delta}}{\log\frac{n(n+1)}{n(n+1)-1}}=O(n^2)$.

Using the union bound, we relate the probability of error in learning all the edges of the graph, to the probability of error in learning a single coefficient λ_{ij}:

$$\mathbb{P}(\text{error})\leq n(n-1)\cdot(1-q)\left(\frac{n(n+1)-1}{n(n+1)}\right)^k.$$

Again, assuming $q=\Theta(1)$, to ensure accurate estimates for all the edges with probability at least $1-\delta$, for given $\delta\in(0,1)$, it suffices that $k\geq\frac{\log\frac{n(n-1)(1-q)}{\delta}}{\log\frac{n(n+1)}{n(n+1)-1}}=O(n^2\log n)$. □

Proof of Proposition 2

For brevity, we only show the second half of Proposition 2, which is illustrative of the reasoning used to prove the rest of our results. Proofs of all the statements can be found in Zoumpoulis (2014).

To show the upper bound, consider the following estimator: after k products, decide $\hat{\lambda}_{i,n+1}=\infty$ if and only if there exists a product such that agent i adopts and agent $n+1$ adopts immediately after (and decide $\hat{\lambda}_{i,n+1}=0$ otherwise). Using the union bound, we relate the probability of error in learning all of $\lambda_{1,n+1},\ldots,\lambda_{n,n+1}$ to the probability of error in learning $\lambda_{1,n+1}$:

$$\begin{aligned}\mathbb{P}(\text{error}) &\leq n\cdot\mathbb{P}\left(\hat{\lambda}_{1,n+1}=0\mid\lambda_{1,n+1}=\infty\right)\mathbb{P}(\lambda_{1,n+1}=\infty)\\ &= n\cdot\frac{1}{2}\sum_{m=0}^{n-1}\left(1-\frac{\lambda}{m\lambda+\lambda+\lambda}\right)^k\binom{n-1}{m}\left(\frac{1}{2}\right)^{n-1}\\ &\leq n\cdot\frac{1}{2}\left(\frac{n}{n+1}\right)^k.\end{aligned}$$

To ensure accurate estimates with probability at least $1-\delta$, for given $\delta\in(0,1)$, it suffices that $k\geq\frac{\log\frac{n}{2\delta}}{\log\frac{n+1}{n}}=O(n\log n)$.

To prove the lower bound, we show that if k is small, there is a high probability event, where it is not clear what a good estimator should decide. No matter how this event is split between the competing hypotheses, the probability of error will be large.

We use X_i to denote the outcome of product i. Having fixed agent j, we say the outcome X_i of product i is in BAD_i^j if one of the following happens: (i) agent j adopts, but agent $n+1$ adopts before her; (ii) agent j does not adopt. We say that the outcome $X_1, \dots, X_k$ is in BAD^j if $X_i \in \mathrm{BAD}_i^j$ for all products $i = 1, \dots, k$.

We are interested in the probability that for some agent j, it is the case that $X_1, \dots, X_k \in \mathrm{BAD}^j$. We define A to be the event that each of the agents $1, \dots, n$ adopts some product before all (other) agents $1, \dots, n$ with links of rate infinity to agent $n+1$ adopt that product. We define B to be the event that all agents with links of rate infinity to agent $n+1$ adopt some product first among other agents with links of rate infinity to agent $n+1$. Then, we can write

$$\begin{aligned}\mathbb{P}\left(\exists j : (X_1, \dots, X_k) \in \mathrm{BAD}^j\right) &= 1 - \mathbb{P}(A) \\ &\geq 1 - \mathbb{P}(B).\end{aligned}$$

Let random variable S be the number of i.i.d. products until event A occurs. Let random variable T be the number of i.i.d. products to obtain event B. Then $S \geq T$. The calculation of the expectation of T is similar to the calculation for the coupon collector's problem, after conditioning on the subset of agents $\mathcal{L} \subseteq \{1, \dots, n\}$ whose influence rate on agent $n+1$ is infinite:

$$\begin{aligned}\mathbb{E}[T] &= \sum_{\mathcal{L} \subseteq \{1,\dots,n\}} \mathbb{P}(\mathcal{L}) \mathbb{E}[T \mid \mathcal{L}] \\ &= \sum_{m=0}^{n} \mathbb{E}[T \mid \mathcal{L}] \binom{n}{m} \left(\frac{1}{2}\right)^n \\ &= \frac{1}{2^n} \sum_{m=0}^{n} \left(\left(\frac{m\lambda}{m\lambda + \lambda}\right)^{-1} + \left(\frac{(m-1)\lambda}{m\lambda + \lambda}\right)^{-1} + \dots + \left(\frac{\lambda}{m\lambda + \lambda}\right)^{-1} \right) \binom{n}{m} \\ &= \frac{1}{2^n} \sum_{m=0}^{n} \left(\frac{m+1}{m} + \frac{m+1}{m-1} + \dots + \frac{m+1}{1} \right) \binom{n}{m} \\ &= \frac{1}{2^n} \sum_{m=0}^{n} (m+1) H_m \binom{n}{m} \\ &= \Omega(n \log n),\end{aligned}$$

where H_m is the mth harmonic number, i.e., $H_m = \sum_{k=1}^{m} \frac{1}{k}$ (and we define $H_0 = 0$), and where the last step follows because, defining $f(m) = (m+1)H_m, m \geq 0$, and using Jensen's inequality, we have

$$\begin{aligned}
\frac{1}{2^n}\sum_{m=0}^{n}(m+1)H_m\binom{n}{m} &\geq f\left(\left\lfloor\frac{1}{2^n}\sum_{m=0}^{n}m\binom{n}{m}\right\rfloor\right)\\
&= f\left(\left\lfloor\frac{1}{2^n}n\sum_{m=1}^{n}\binom{n-1}{m-1}\right\rfloor\right)\\
&= f\left(\left\lfloor\frac{1}{2^n}n\sum_{m'=0}^{n-1}\binom{n-1}{m'}\right\rfloor\right)\\
&= f\left(\left\lfloor\frac{1}{2^n}n2^{n-1}\right\rfloor\right)\\
&= f\left(\left\lfloor\frac{n}{2}\right\rfloor\right)\\
&= \left(\left\lfloor\frac{n}{2}\right\rfloor+1\right)H_{\left\lfloor\frac{n}{2}\right\rfloor}\\
&= \Theta(n\log n).
\end{aligned}$$

Similarly, for the variance we can write

$$\begin{aligned}
var\,(T) &= \sum_{\mathcal{L}\subseteq\{1,\ldots,n\}}\mathbb{P}(\mathcal{L})var\,(T\mid\mathcal{L})\\
&= \sum_{m=0}^{n}var\,(T\mid\mathcal{L})\binom{n}{m}\left(\frac{1}{2}\right)^n\\
&= \frac{1}{2^n}\sum_{m=0}^{n}\left(\frac{1-\frac{m\lambda}{m\lambda+\lambda}}{\left(\frac{m\lambda}{m\lambda+\lambda}\right)^2}+\frac{1-\frac{(m-1)\lambda}{m\lambda+\lambda}}{\left(\frac{(m-1)\lambda}{m\lambda+\lambda}\right)^2}+\ldots+\frac{1-\frac{\lambda}{m\lambda+\lambda}}{\left(\frac{\lambda}{m\lambda+\lambda}\right)^2}\right)\binom{n}{m}\\
&= \frac{1}{2^n}\sum_{m=0}^{n}\left(\frac{\frac{1}{m+1}}{\left(\frac{m}{m+1}\right)^2}+\frac{\frac{2}{m+1}}{\left(\frac{m-1}{m+1}\right)^2}+\ldots+\frac{\frac{m}{m+1}}{\left(\frac{1}{m+1}\right)^2}\right)\binom{n}{m}\\
&\leq \frac{1}{2^n}\sum_{m=0}^{n}\left(\frac{1}{\left(\frac{m}{m+1}\right)^2}+\frac{1}{\left(\frac{m-1}{m+1}\right)^2}+\ldots+\frac{1}{\left(\frac{1}{m+1}\right)^2}\right)\binom{n}{m}\\
&= \frac{1}{2^n}\sum_{m=0}^{n}\left((m+1)^2\left(\frac{1}{1^2}+\frac{1}{2^2}+\ldots+\frac{1}{m^2}\right)\right)\binom{n}{m}\\
&\leq (n+1)^2\left(\frac{1}{1^2}+\frac{1}{2^2}+\ldots\right)\\
&= (n+1)^2\cdot\frac{\pi^2}{6}\\
&\leq 2(n+1)^2.
\end{aligned}$$

By Chebyshev's inequality,

$$\mathbb{P}\left(|T-\mathbb{E}[T]| \geq c(n+1)\right) \leq \frac{2}{c^2}.$$

Therefore, with $k = o(n \log n)$ products, there is a very small probability that event B will occur, and therefore a very large probability that the event $\{\exists j : (X_1, \ldots, X_k) \in \mathrm{BAD}^j\}$ will occur, which establishes the $\Omega(n \log n)$ lower bound for the number of products k. □

Proof of Proposition 3

We focus on the event that both i and j adopt, with i adopting before j. We compute the probability for all the cases in this event, under each of the two hypotheses.

We have

$$\mathbb{P}(i \text{ first}, j \text{ second} \mid i \to j) = \frac{\lambda}{n\lambda+\lambda} \cdot \frac{2\lambda}{(n-1)\lambda+(n-1)\lambda+\lambda} = \frac{1}{n+1} \cdot \frac{2}{2n-1},$$

while

$$\mathbb{P}(i \text{ first}, j \text{ second} \mid i \nrightarrow j) = \frac{\lambda}{n\lambda+\lambda} \cdot \frac{\lambda}{(n-1)\lambda+(n-2)\lambda+\lambda} = \frac{1}{n+1} \cdot \frac{1}{2n-2},$$

and thus the difference of the two is

$$\frac{1}{n+1}\left(\frac{2}{2n-1} - \frac{1}{2n-2}\right) = \frac{1}{n+1} \cdot \frac{2n-3}{(2n-1)(2n-2)} \sim \frac{1}{2n^2}.$$

Similarly,

$$\begin{aligned}\mathbb{P}(i \text{ first}, j \text{ third} \mid i \to j) &= \frac{\lambda}{n\lambda+\lambda} \cdot \frac{(n-2)2\lambda}{(n-1)2\lambda+\lambda} \cdot \frac{3\lambda}{(n-2)\lambda+(n-2)2\lambda+\lambda} \\ &= \frac{1}{n+1} \cdot \frac{2n-4}{2n-1} \cdot \frac{3}{3n-5},\end{aligned}$$

while

$$\begin{aligned}\mathbb{P}(i \text{ first}, j \text{ third} \mid i \nrightarrow j) &= \frac{\lambda}{n\lambda+\lambda} \cdot \frac{(n-2)2\lambda}{(n-2)2\lambda+\lambda+\lambda} \cdot \frac{2\lambda}{(n-2)\lambda+(n-3)2\lambda+\lambda+\lambda} \\ &= \frac{1}{n+1} \cdot \frac{2n-4}{2n-2} \cdot \frac{2}{3n-6},\end{aligned}$$

with a difference of

$$\begin{aligned}\frac{2n-4}{n+1}\left(\frac{3}{(2n-1)(3n-5)} - \frac{2}{(2n-2)(3n-6)}\right) &= \frac{2n-4}{n+1} \cdot \frac{6n^2-28n+26}{(2n-1)(3n-5)(2n-2)(3n-6)} \\ &\sim \frac{1}{3n^2}.\end{aligned}$$

Similarly, we can show that

$$\mathbb{P}(i \text{ first}, j \text{ fourth} \mid i \to j) - \mathbb{P}(i \text{ first}, j \text{ fourth} \mid i \nrightarrow j) \sim \frac{1}{4n^2},$$

and in general, for $2 \leq \ell \leq n$

$$\mathbb{P}(i \text{ first}, j \ \ell\text{th} \mid i \to j) - \mathbb{P}(i \text{ first}, j \ \ell\text{th} \mid i \nrightarrow j) \sim \frac{1}{\ell n^2}.$$

We now focus on the events in which i adopts second. We have

$$\mathbb{P}(i \text{ second}, j \text{ third} \mid i \to j) = \frac{(n-2)\lambda}{n\lambda+\lambda} \cdot \frac{2\lambda}{(n-1)2\lambda+\lambda} \cdot \frac{3\lambda}{(n-2)3\lambda+\lambda} = \frac{n-2}{n+1} \cdot \frac{2}{2n-1} \cdot \frac{3}{3n-5},$$

while

$$\mathbb{P}(i \text{ second}, j \text{ third} \mid i \nrightarrow j) = \frac{(n-2)\lambda}{n\lambda+\lambda} \cdot \frac{2\lambda}{(n-1)2\lambda+\lambda} \cdot \frac{2\lambda}{(n-3)3\lambda+2\lambda+\lambda} = \frac{n-2}{n+1} \cdot \frac{2}{2n-1} \cdot \frac{2}{3n-6},$$

and thus the difference of the two is

$$\frac{n-2}{n+1} \cdot \frac{2}{2n-1} \left(\frac{3}{3n-5} - \frac{2}{3n-6} \right) = \frac{n-2}{n+1} \cdot \frac{2}{2n-1} \cdot \frac{3n-8}{(3n-5)(3n-6)} \sim \frac{1}{3n^2}.$$

In general, for $3 \leq \ell \leq n$, the difference is

$$\mathbb{P}(i \text{ second}, j \ \ell\text{th} \mid i \to j) - \mathbb{P}(i \text{ second}, j \ \ell\text{th} \mid i \nrightarrow j) \sim \frac{1}{\ell n^2}.$$

We can sum up all the differences between the two hypotheses for the events in which both i and j adopt with i adopting before j, to get asymptotically

$$\begin{aligned} \frac{1}{n^2} \cdot \Bigg[\frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots\ldots\ldots\ldots\ldots + \frac{1}{n} \\ + \frac{1}{3} + \frac{1}{4} + \ldots\ldots\ldots\ldots\ldots + \frac{1}{n} \\ + \frac{1}{4} + \ldots\ldots\ldots\ldots\ldots + \frac{1}{n} \\ \vdots \\ \vdots \\ + \frac{1}{n} \Bigg] \end{aligned}$$

which can be written as

$$\begin{aligned} \frac{1}{n^2} \left(\frac{1}{2} + \frac{2}{3} + \frac{3}{4} + \ldots + \frac{n-1}{n} \right) &= \frac{1}{n^2} \left((1 - \frac{1}{2}) + (1 - \frac{1}{3}) + (1 - \frac{1}{4}) + \ldots + (1 - \frac{1}{n}) \right) \\ &\sim \frac{1}{n^2} \left(n - 1 - (\log n - 1) \right) \\ &= \frac{n - \log n}{n^2}. \end{aligned}$$

The sample complexity is therefore

$$\frac{1}{\left(\frac{n-\log n}{n^2}\right)^2} = \frac{n^4}{(n-\log n)^2} = \Theta(n^2).$$

□

Proof of Proposition 4

We focus on the event that only i, j adopt. We have

$$\begin{aligned}\mathbb{P}(\text{only } i, j \text{ adopt} \mid i \to j) &= 2\left(\frac{\lambda}{n\lambda+\lambda}\cdot\frac{2\lambda}{(n-1)\lambda+(n-1)\lambda+\lambda}\right)\frac{\lambda}{(n-2)\lambda+(n-2)2\lambda+\lambda}\\ &= 2\cdot\frac{2}{(n+1)(2n-1)}\cdot\frac{1}{3n-5},\end{aligned}$$

while

$$\begin{aligned}\mathbb{P}(\text{only } i, j \text{ adopt} \mid i \nrightarrow j) &= \left(\frac{\lambda}{n\lambda+\lambda}\cdot\frac{\lambda}{(n-1)\lambda+(n-2)\lambda+\lambda}\right.\\ &\qquad\left.+\frac{\lambda}{n\lambda+\lambda}\cdot\frac{2\lambda}{(n-1)\lambda+(n-1)\lambda+\lambda}\right)\frac{\lambda}{(n-2)\lambda+(n-2)2\lambda+\lambda}\\ &= \frac{1}{n+1}\left(\frac{1}{2n-2}+\frac{2}{2n-1}\right)\frac{1}{3n-5},\end{aligned}$$

and thus the difference of the two is

$$\frac{1}{n+1}\cdot\frac{1}{3n-5}\left(\frac{4}{2n-1}-\frac{1}{2n-2}-\frac{2}{2n-1}\right) = \frac{1}{n+1}\cdot\frac{1}{3n-5}\cdot\frac{2n-3}{(2n-1)(2n-2)} \sim \frac{1}{6n^3}.$$

The sample complexity is therefore $\Theta(n^6)$. □