Discrete Applied Mathematics 200 (2016) 108–122

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Reconstructing Markov processes from independent and anonymous experiments

Silvio Micali, Zeyuan Allen Zhu *

MIT CSAIL, United States

ARTICLE INFO

Article history:
Received 21 June 2014
Received in revised form 7 June 2015
Accepted 21 June 2015
Available online 20 July 2015

Keywords:
Graph reconstruction
Random walk
Markov process
Local algorithms

ABSTRACT

We investigate the problem of *exactly* reconstructing, with high confidence and up to isomorphism, the ball of radius r centered at the starting state of a Markov process from *independent* and *anonymous* experiments. In an anonymous experiment, the states are visited according to the underlying transition probabilities, but no global state names are known: one can only recognize whether two states, *reached within the same experiment*, are the same.

We prove quite tight bounds for such exact reconstruction in terms of both the number of experiments and their lengths.

1. Introduction

The problem of reconstructing a large "object" from partial observations is quite fundamental, and arises in many fields, such as system biology [22,29], social networks [38,27], brain networks [36,19], telecommunication networks [10], and many others.

We investigate a more complex type of reconstruction. In essence, our goal is to reconstruct a Markov process from the records produced by *limited* observers acting *independently*, without coordination, and without even sharing a common "name space". Let us explain.

1.1. Our model

Our Markov model. In a Markov process, we denote the underlying transition graph by $G = (V, E)$ and the starting vertex by v. In this paper, the graph G is undirected and has infinitely many vertices, each of finite degree. An infinite sequence of vertices is generated by the following process. The first vertex is v, and, if the ith vertex is u, then the $(i+1)$-st vertex is chosen at random uniformly and independently among the neighbors of u.

A sequence of vertices so generated is called a *random walk*. If $(v =)v_0 \to v_1 \to \cdots$ is a random walk, then $v_0 \to \cdots \to v_\ell$ is a *random walk of length* ℓ.

Note. Assuming that G is undirected and unweighted allows us to present our results in the cleanest way. We shall discuss how to relax both assumptions in Section 1.4. Assuming that G has infinitely many vertices is a simple way to force us to consider only "local" algorithms: essentially, algorithms whose performance does not depend on the size of the whole graph, which may be larger than all the parameters we shall care about.

* Corresponding author.
E-mail addresses: silvio@csail.mit.edu (S. Micali), zeyuan@csail.mit.edu (Z.A. Zhu).

http://dx.doi.org/10.1016/j.dam.2015.06.035
0166-218X/

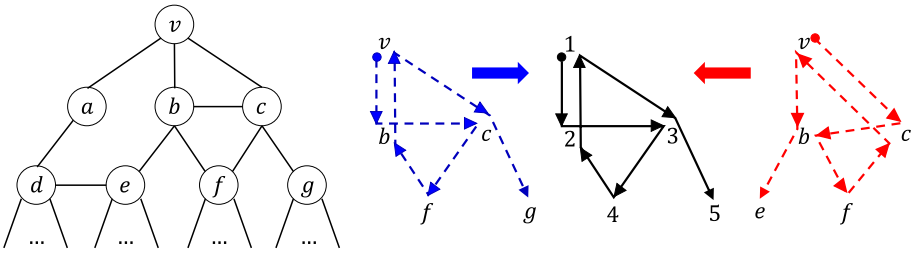

Fig. 1. An example illustrating the definition of *anonymous experiment*.

Our anonymous observation model. If $v_0 \to v_1 \to \cdots$ is a random walk, then its corresponding *(anonymous) experiment* is the sequence of integers $f(v_0) \to f(v_1) \to \cdots$, where $f(v_i) \stackrel{\text{def}}{=} |\{v_0, \ldots, v_{i^*}\}|$ and i^* is the smallest integer j such that $v_j = v_i$. Intuitively, $f(u)$ maps u to an integer indicating that u is the $f(u)$th distinct vertex in this walk.

Example. In the graph of Fig. 1, the length-7 walk $v \to b \to c \to f \to b \to v \to c \to g$ corresponds to the anonymous experiment $1 \to 2 \to 3 \to 4 \to 2 \to 1 \to 3 \to 5$.

Note the walk $v \to c \to b \to f \to c \to v \to b \to e$ *also* corresponds to the same experiment.

Rationale. Markov processes naturally model physical systems. In essence, the possible "states" of the system are the vertices of the transition graph G, and, when put in its "initial state" v, the system evolves (i.e., new states are generated) according to the transition rules.

When the system is "new" – better said, studied for the first time – no one initially has any idea about the underlying graph G. However, each individual can, on his own, experiment with the system by putting it into its initial state, and independently observe its "evolution": that is, a random walk in G.

In sum, each individual will observe and record the states encountered in a random walk. Since the system is new, no global names exist for the states. Thus each individual may very well use his own name space for the states encountered, and thus his record is an anonymous experiment as defined above.

Of course, an individual observer might consider writing down a full description of every state he sees. However, this may not be possible due to – say – memory limitations [15], or privacy reasons [26]. Also, an observer may not know how many details are sufficient to identify each encountered state. In any case, an anonymous experiment is a most compact and meaningful record.

1.2. Our results for the basic reconstruction problem

Whether human or not, a realistic observer has a bounded lifetime, and thus cannot visit more than ℓ nodes in his random walk.[1] Thus, even with an unlimited number of such observers, one can at most learn $\mathcal{D}_\ell$, the distribution over the anonymous experiments of length ℓ (that are induced from the random walks of the same length). Since our G has infinitely many vertices, $\mathcal{D}_\ell$ cannot suffice to reconstruct the entire graph G. However, one may be able to use $\mathcal{D}_\ell$ in order to reconstruct $B(v, r)$, the *ball of center v and radius r* (i.e., the subgraph of G induced by all vertices whose distance from v is at most r). More precisely, one may be able to compute a graph $G' = (V', E')$ and a distinguished vertex $v' \in V'$, such that G' is isomorphic to $B(v, r)$ and the isomorphism maps v' to v. Thus, our basic reconstruction problem can be formulated as follows:

For every r, is there a length ℓ such that $B(v, r)$ is reconstructible (up to isomorphism) from $\mathcal{D}_\ell$?

Notice that, given access to the distribution $\mathcal{D}_\ell$, one can also simulate access to the distributions $\mathcal{D}_1, \ldots, \mathcal{D}_{\ell-1}$. Of course, although for now we are ignoring the complexity of learning these distributions, it would be nice if, given $(\mathcal{D}_1, \ldots, \mathcal{D}_\ell)$ as *oracles*, the reconstruction algorithm is efficient. Here, we say that $(\mathcal{D}_1, \ldots, \mathcal{D}_\ell)$ are given as oracles, if the algorithm is allowed to ask for the precise probability of $\mathcal{D}_t(P)$ for any anonymous experiment P of length $t \in [\ell]$.

Notice too that, in principle, our basic reconstruction problem may be impossible. For instance, could there exist two different Markov processes, (G_1, v_1) and (G_2, v_2), having the same distribution $\mathcal{D}_\ell$ for all $\ell \geq 0$? If this were the case, the two processes would be indistinguishable by any number of anonymous experiments, of any length, which immediately implies a negative answer to the above question. Yet, we provide a constructive proof showing that our basic reconstruction problem is indeed possible, when the underlying graph G is undirected.

Theorem 1. *Let n be the number of vertices in $B(v, r)$ and m the number of edges. One can reconstruct $B(v, r)$ in time $O(n^2)$ and with $O(n^2)$ oracle accesses to $(\mathcal{D}_1, \ldots, \mathcal{D}_\ell)$, where $\ell = O(m)$. Moreover, the reconstruction algorithm only makes membership queries to* $\text{supp}(\mathcal{D}_i)$ *for $i \in [\ell]$.*

In contrast, as we shall see in Section 1.4, this reconstruction becomes impossible when the underlying graph is directed but not strongly connected.

Is this algorithm tight? To answer this question we must refine our reconstruction problem.

[1] For concreteness, if he lives for at most 100 years, and each transition from node to node takes 1 s of time, then $\ell = 100 \times 366 \times 24 \times 60 \times 60$.