Theoretical Computer Science 548 (2014) 68–78

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Sublinear-time algorithms for monomer–dimer systems on bounded degree graphs ☆

Marc Lelarge, Hang Zhou *

INRIA, École Normale Supérieure, France

ARTICLE INFO

Article history:
Received 24 September 2013
Received in revised form 21 May 2014
Accepted 29 June 2014
Available online 8 July 2014
Communicated by J. Díaz

Keywords:
Graph algorithms
Matchings
Correlation decay
Approximation algorithms
Random sampling

ABSTRACT

For a graph G, let $Z(G,\lambda)$ be the partition function of the monomer–dimer system defined by $\sum_k m_k(G)\lambda^k$, where $m_k(G)$ is the number of matchings of size k in G. We consider graphs of bounded degree and develop a sublinear-time algorithm for estimating $\log Z(G,\lambda)$ at an arbitrary value $\lambda > 0$ within additive error ϵn with high probability. The query complexity of our algorithm does not depend on the size of G and is polynomial in $1/\epsilon$, and we also provide a lower bound quadratic in $1/\epsilon$ for this problem. This is the first analysis of a sublinear-time approximation algorithm for a $\#P$-complete problem. Our approach is based on the correlation decay of the Gibbs distribution associated with $Z(G,\lambda)$. We show that our algorithm approximates the probability for a vertex to be covered by a matching, sampled according to this Gibbs distribution, in a near-optimal sublinear time. We extend our results to approximate the average size and the entropy of such a matching within an additive error with high probability, where again the query complexity is polynomial in $1/\epsilon$ and the lower bound is quadratic in $1/\epsilon$. Our algorithms are simple to implement and of practical use when dealing with massive datasets. Our results extend to other systems where the correlation decay is known to hold as for the independent set problem up to the critical activity.

1. Introduction

The area of sublinear-time algorithms is an emerging area of computer science which has its root in the study of massive datasets [6,26]. Internet, social networks or communication networks are typical examples of graphs with potentially millions of vertices representing agents, and edges representing possible interactions among those agents. In this paper, we present sublinear-time algorithms for graph problems. We are concerned more with problems of counting and statistical inference and less with optimization. For example, in mobile call graphs, phone calls can be represented as a matching of the graph where each edge has an activity associated with the intensity of the interactions between the pair of users. Given such graphs, with local activities on edges, we would like to answer questions like: What is the size of a typical matching? What is for a given user the probability of being matched? As another example, models of statistical physics have been proposed to model social interactions. In particular, spin systems are a general framework for modeling nearest-neighbor interactions on graphs. In this setting, the activity associated with each edge allows to model a perturbed best-response dynamics [2].

☆ This is the full version of the extended abstract that appears in [18].

* Corresponding author.
E-mail addresses: marc.lelarge@ens.fr (M. Lelarge), hang.zhou@ens.fr (H. Zhou).

http://dx.doi.org/10.1016/j.tcs.2014.06.040
0304-3975/

Again in this setting, it is interesting to compute estimations for the number of agents playing a given strategy or the probability for an agent in the graph to play a given strategy at equilibrium.

There are now quite a few results on sublinear-time approximation algorithms for graph optimization problems: minimum spanning tree weight [5], minimum set cover [23], maximum matching [23,34] and minimum vertex cover [23–25]. There are also a couple of works on sublinear-time algorithms for statistical and counting problems, e.g., approximating the average degree of a graph [7,11] and approximating the number of occurrences of a certain structure (such as a star) in a graph [12]. Our focus in this paper is on the algorithmic problems arising in statistical physics and classical combinatorics [32]. We now present the monomer–dimer problem which will be the main focus of our paper.

Let $G = (V, E)$ be an undirected graph with $|V| = n$ vertices and $|E| = m$ edges, where we allow G to contain parallel edges and self-loops. We denote by $N(G, v)$ the set of neighbors of v in G. We consider bounded degree graphs with $\max_v |N(G, v)| \le \Delta$. In a monomer–dimer system, the vertices are covered by non-overlapping arrangement of monomers (molecules occupying one vertex of G) and dimers (molecules occupying two adjacent vertices of G) [13]. It is convenient to identify monomer–dimer arrangements with matchings; a matching in G is a subset $M \subset E$ such that no two edges in M share an endpoint. Thus, a matching of cardinality $|M| = k$ corresponds exactly to a monomer–dimer arrangement with k dimers and $n - 2k$ monomers. Let $\mathbb{M}$ be the set of matchings of G. To each matching M, a weight $\lambda^{|M|}$ is assigned, where $\lambda > 0$ is called the activity. The partition function of the system is defined by $Z(G, \lambda) = \sum_{M \in \mathbb{M}} \lambda^{|M|}$, and the Gibbs distribution on the space $\mathbb{M}$ is defined by $\pi_{G,\lambda}(M) = \frac{\lambda^{|M|}}{Z(G,\lambda)}$. The function $Z(G, \lambda)$ is also of combinatorial interest and called the *matching polynomial* in this context [21]. For example, $Z(G, 1)$ enumerates all matchings in G. From an algorithmic viewpoint, no feasible method is known for computing $Z(G, \lambda)$ exactly for general monomer–dimer system; indeed, for any fixed value of $\lambda > 0$, the problem of computing $Z(G, \lambda)$ exactly in a graph of bounded degree Δ is complete for the class #P of enumeration problems, when $\Delta \ge 5$ (see [30]). The focus on computing $Z(G, \lambda)$ shifted to finding approximate solutions in polynomial time. For example, the Markov Chain Monte Carlo (MCMC) method yields a provably efficient algorithm for finding an approximate solution. Based on the equivalence between the counting problem (computing $Z(G, \lambda)$) and the sampling problem (according to $\pi_{G,\lambda}$) [16], this approach focuses on rapidly mixing Markov chains to obtain appropriate random samples. A Fully Polynomial-time Randomized Approximation Scheme (FPRAS) for computing the total number of matchings based on MCMC was provided by Jerrum and Sinclair [14,27].

Another related problem in the monomer–dimer system is the average size of a matching sampled according to $\pi_{G,\lambda}$, defined by $E(G, \lambda) = \sum_{M \in \mathbb{M}} |M| \, \pi_{G,\lambda}(M)$. Sinclair and Srivastava recently proved in [29] that for any fixed value of $\lambda > 0$, the problem of computing $E(G, \lambda)$ exactly in a bounded degree graph (allowing parallel edges) is #P-hard, for any maximum degree $\Delta \ge 5$. Thus again we are interested in finding approximate solutions to this problem.

In order to study sublinear-time approximation algorithms for these problems, we use the approach based on the concept of correlation decay originating in statistical physics [22] and which has been used to get a deterministic approximation scheme for counting matchings in polynomial time [1]. It follows already from [13] that the marginals of the probability distribution $\pi_{G,\lambda}$ are local in nature: the local structure of the graph around a vertex v allows to compute an approximation of the corresponding marginal. In the computer science literature, this property follows from the so-called *correlation decay property*. Our algorithm is then simple to understand: we need only to sample a fixed number of vertices, approximate the marginals associated with these vertices locally and then from these values output an estimate for the desired quantity. The correlation decay property also holds for other systems such as the independent set problem [31], the coloring problem [8], and the two-state spin system [19,20,28]. In Section 6, we extend our technique to the independent set problem. We believe that similar extensions can be done for other systems as soon as the correlation decay property holds.

A graph G is represented by two kinds of oracles $\mathcal{D}$ and $\mathcal{N}$ such that $\mathcal{D}(v)$ returns the degree of $v \in V$ and $\mathcal{N}(v, i)$ returns the ith (with $1 \le i \le \mathcal{D}(v)$) neighbor of $v \in V$. The efficiency of an algorithm is measured by its query complexity, i.e. the total number of accesses to $\mathcal{D}$ and $\mathcal{N}$. Let VAL denote a real value associated with the graph. We say that $\widehat{\mathrm{VAL}}$ is an *ϵ-approximation* of VAL if $\widehat{\mathrm{VAL}} - \epsilon \le \mathrm{VAL} \le \widehat{\mathrm{VAL}} + \epsilon$, where $\epsilon > 0$ is specified as an input parameter. An algorithm is called an *ϵ-approximation algorithm* for VAL if for any graph G, it computes an ϵ-approximation of VAL with high probability (e.g., at least $\frac{2}{3}$). In our model, we consider the case of constant maximum degree Δ as ϵ tends to zero, i.e., we always first take the limit as $\epsilon \to 0$ and then the limit $\Delta \to \infty$.

Our main contribution (Theorem 6) is an ϵn-approximation algorithm for $\log Z(G, \lambda)$ in a graph G of bounded degree Δ. The query complexity of the algorithm is $\tilde{O}((1/\epsilon)^{\tilde{O}(\sqrt{\Delta})})$, which does not depend on the size of the graph. From the relation between the partition function and the matching statistics, we then obtain ϵn-approximation algorithms for the average size of a matching and the entropy of $\pi_{G,\lambda}$ with the same query complexity as before. We also provide the $\Omega(1/\epsilon^2)$ query lower bound for ϵn-approximation algorithms for $\log Z(G, \lambda)$ and the other two problems.

The main tool of the above algorithms is the approximation of the marginal $p_{G,\lambda}(v)$, which is the probability that the vertex v is not covered by a matching under the Gibbs distribution. We estimate $p_{G,\lambda}(v)$ for an arbitrary vertex $v \in V$ within an error of $\epsilon > 0$ with near-optimal query complexity $\tilde{O}((1/\epsilon)^{\tilde{O}(\sqrt{\Delta})})$.

The rest of the paper is organized as follows. In Section 2, we prove our first main result concerning local computations for matchings. Based on this result, we construct an ϵn-approximation algorithm for $\log Z(G, \lambda)$ in Section 3 and ϵn-approximation algorithms for the average size of a matching and the entropy of $\pi_{G,\lambda}$ in Section 4. We also provide query lower bounds in these two sections. In Section 5, we give some applications of our technique for approximating the permanent of constant degree expander graphs and the size of a maximum matching (in this last case, our algo-