Computational Geometry 47 (2014) 214–223

Contents lists available at ScienceDirect

Computational Geometry: Theory and Applications

www.elsevier.com/locate/comgeo

# Closest pair and the post office problem for stochastic points ☆

Pegah Kamousi [a,*], Timothy M. Chan [b], Subhash Suri [a]

[a] *Department of Computer Science, UC Santa Barbara, CA 93106, USA*
[b] *Cheriton School of Computer Science, University of Waterloo, Ontario N2L 3G1, Canada*

ARTICLE INFO

*Article history:*
Received 7 November 2011
Accepted 29 October 2012
Available online 1 November 2012
Communicated by F. Dehne and J. Iacono

*Keywords:*
Computational geometry
Probabilistic optimization
Approximation algorithms
Data structures

ABSTRACT

Given a (master) set $M$ of $n$ points in $d$-dimensional Euclidean space, consider drawing a random subset that includes each point $m_i \in M$ with an independent probability $p_i$. How difficult is it to compute elementary statistics about the closest pair of points in such a subset? For instance, what is the *probability* that the distance between the closest pair of points in the random subset is no more than $\ell$, for a given value $\ell$? Or, can we preprocess the master set $M$ such that given a query point $q$, we can efficiently estimate the *expected* distance from $q$ to its nearest neighbor in the random subset? These basic computational geometry problems, whose complexity is quite well-understood in the deterministic setting, prove to be surprisingly hard in our *stochastic* setting. We obtain hardness results and approximation algorithms for stochastic problems of this kind.


## 1. Introduction

Many years ago, Knuth [12] posed the now classic *post-office* problem, namely, given a set of points in the plane, find the one closest to a query point $q$. The problem, which is fundamental and arises as a basic building block of numerous computational geometry algorithms and data structures [7], is reasonably well-understood in small dimensions. In this paper, we consider a *stochastic* version of the problem in which each post office may be *closed* with certain probability. In other words, a given set of points $M$ in $d$ dimensions includes the locations of all the post offices but on a typical day each post office $m_i \in M$ is only open with an independent probability $p_i$. The set of points $M$ together with their probabilities form a probability distribution where each point $m_i$ is included in a random subset of points with probability $p_i$. Thus, given a query points $q$, we now must talk about the *expected* distance from $q$ to its closest neighbor in $M$. Similarly, instead of simply computing the closest pair of points in a set, we must ask: how *likely* is it that the closest pair of points are no more than $\ell$ apart?

In this paper, we study the complexity of such elementary proximity problems and establish both upper and lower bounds. In particular, we have a finite set of points $M$ in a $d$-dimensional Euclidean space, which constitutes our *master* set of points, and hence the mnemonic $M$. Each member $m_i$ of $M$ has probability $p_i$ of being present and probability $1 - p_i$ of being absent. (Following the post-office analogy, the $i$th post office is open with probability $p_i$ and closed otherwise.) These probabilities are independent, but otherwise arbitrary, and lead to a sample space of $2^n$ possible subsets, where a sample subset includes the $i$th point with independent probability $p_i$. We achieve the following results.

☆ The work of the first and the third author was supported in part by National Science Foundation grants CCF-0514738 and CNS-1035917. The work of the second author was supported by NSERC. A preliminary version of the paper has appeared in: Proc. 12th Algorithms and Data Structures Symposium (WADS), 2011, pp. 548–559.

* Corresponding author.
*E-mail address:* pkamousi@umail.ucsb.edu (P. Kamousi).

0925-7721/$ – see front matter 
http://dx.doi.org/10.1016/j.comgeo.2012.10.010

1. It is #P-hard to compute the probability that the closest pair of points have distance at most a value $\ell$, even for dimension 2 under the $L_\infty$ norm.
2. In the *linearly-separable* and *bichromatic* planar case, the above closest pair probability can be computed in polynomial time under the $L_\infty$ norm.
3. Without the linear separability, even the bichromatic version of the above problem is #P-hard under the $L_2$ or $L_\infty$ norm.
4. Even with linear separability and $L_\infty$ norm, the bichromatic case becomes #P-hard in dimension $d \geqslant 3$.
5. We give a linear-space data structure with $O(\log n)$ query time to compute the expected distance of a given query point to its $(1+\varepsilon)$-approximate nearest neighbor when the dimension $d$ is a constant.

### 1.1. Related work

A number of researchers have recently begun to explore geometric computing over *probabilistic* data [1,2,14,19]. These studies are fundamentally different from the classical geometric probability that deals with properties of random point sets drawn from some infinite sets, such as points in unit square [11]. Instead, the recent work in computational geometry is concerned with worst-case sets of objects and worst-case probabilities or behavior. In particular, the recent works in [1,2] deals with the database problem of skyline computation using a multiple-universe model. The work of van Kreveld and Löffler [14,19] deals with objects whose locations are randomly distributed in some *uncertainty* regions. Guibas et al. [8] consider the behavior of fundamental geometric constructs such as the Voronoi diagrams when the point generators are not known precisely but only given as Gaussian distributions. Unlike these results, our focus is not on the locational uncertainty but rather on the discrete probabilities with which each point may appear.

## 2. The stochastic closest pair problem

We begin with the complexity of the stochastic closest pair (SCP) problem in the plane, which asks for the probability that the closest pair of points in the plane has distance at most a given bound $\ell$. We use the notation $d(P,Q)$ for the $L_2$ distance between two sets $P$ and $Q$. (When using the $L_\infty$ or $L_1$ norms, we will use $d_\infty(P,Q)$ and $d_1(P,Q)$.) Consider a set of points $M = \{m_1, m_2, \ldots, m_n\}$ in the plane, called the *master* set. Each point $m_i$ is *active*, or present, with some independent and arbitrary but known (rational-valued) probability $p_i$. The independent point probabilities induce a sample space $\Omega$ with $2^n$ outcomes. Let $S \subset M$ denote the set of active points in a trial. Then an outcome $A \subseteq M$ occurs with probability $\Pr[S = A] = \prod_{m_i \in A} p_i \prod_{m_i \notin A}(1-p_i)$. We ask for the probability that $\min_{m_i, m_j \in S} d(m_i, m_j) \leqslant \ell$. We will show that this basic problem is intractable, via reduction from the problem of *counting minimum vertex covers* in planar unit disk graphs. In order to show that even the *bichromatic* closest pair problem is hard, we also prove that a corresponding vertex cover counting problem is also hard for a bichromatic version of the unit disk graphs.

### 2.1. Counting vertex covers in unit disk graphs

The problem of counting minimum cardinality vertex covers in a graph [13,16,18] is the following. Given a graph $G = (V, E)$, how many node-subsets $S \subseteq V$ constitute a vertex cover of minimum cardinality, where $S$ is a vertex cover of $G$ if for each $uv \in E$, either $u \in S$ or $v \in S$. This problem is known to be #P-hard, even for planar bipartite graphs with maximum degree 3 [16]. This immediately shows that counting the minimum *weight* vertex covers is also hard for this type of graphs, as we can always assign the same weight to all the node, so that the minimum cardinality vertex cover minimizes the weight as well. Throughout this paper, we simply use the term *minimum vertex cover* instead of minimum *weight* vertex cover, and the term minimum *cardinality* vertex cover is used otherwise.

Let **#MinVC** denote the problem of counting minimum weight vertex covers in a graph. We will consider this problem on *unit disk graphs*, which are the intersection graphs of $n$ equal-sized circles in the plane: each node corresponds to a circle, and there is an edge between two nodes if the corresponding circles intersect. The following theorem considers the complexity of **#MinVC** problem on unit disk graphs.

**Theorem 2.1.** *It is #P-hard to compute the number of minimum weight vertex covers in weighted unit disk graphs, even if all the nodes have weight* 1 *or* 2*, and even if the distances are measured in the* $L_\infty$ *metric.*

Throughout this section, we assume that the disks defining unit disk graphs have radius 1. The first step of the proof is the following well-known lemma of Valiant [17].

**Lemma 2.2.** *A planar graph* $G = (V, E)$ *with maximum degree* 4 *can be embedded in the plane using* $O(|V|^2)$ *area in such a way that its nodes are at integer coordinates and its edges are drawn so that they are made up of line segments of the form* $x = i$ *or* $y = j$*, for integers* $i$ *and* $j$*.*