Approximating the Expected Values for Combinatorial Optimization Problems over Stochastic Points

Lingxiao Huang(✉) and Jian Li

Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
huanglingxiao1990@126.com

Abstract. We consider the stochastic geometry model where the location of each node is a random point in a given metric space, or the existence of each node is uncertain. We study the problems of computing the expected lengths of several combinatorial or geometric optimization problems over stochastic points, including closest pair, minimum spanning tree, k-clustering, minimum perfect matching, and minimum cycle cover. We also consider the problem of estimating the probability that the length of closest pair, or the diameter, is at most, or at least, a given threshold. Most of the above problems are known to be #P-hard. We obtain FPRAS (Fully Polynomial Randomized Approximation Scheme) for most of them in both the existential and locational uncertainty models. Our result for stochastic minimum spanning trees in the locational uncertain model improves upon the previously known constant factor approximation algorithm. Our results for other problems are the first known to the best of our knowledge.

1 Introduction

Background: Uncertain or imprecise data are pervasive in applications like sensor monitoring, location based services, data collection and integration [12,14,33]. Consider a temperature monitoring system which collects measures of humidity and wind speed. Since we do not have the perfect sensing instruments, the data obtained are often contaminated with noises[13]. For another example, the locational data collected by the Global-Positioning Systems (GPS) often contains measurement errors [29]. Moreover, many machine learning and prediction algorithms also produce a variety of stochastic models and a large volume of probabilistic data. Thus, managing, analyzing and solving optimization problems over stochastic models and data have recently attracted significant attentions in several research communities (see e.g., [30,33,34]).

In this paper, we study two stochastic geometry models, the locational uncertainty model and the existential uncertainty model, both of which have been studied extensively in recent years (see e.g., [2–4,7,20,21,24–26],some of which will be discussed in the related work section). In fact, a special case of the locational uncertainty model where all points follow the same distribution is a classic topic in stochastic geometry literature (see e.g., [8–10,22,31]). The main interest there has been to derive asymptotics for the

Research supported in part by the National Basic Research Program of China Grant 2015CB358700, 2011CBA00300, 2011CBA00301, the National Natural Science Foun- dation of China Grant 61202009, 61033001, 61361136003.

M.M. Halldórsson et al. (Eds.): ICALP 2015, Part I, LNCS 9134, pp. 910–921, 2015.
DOI: 10.1007/978-3-662-47672-7_74

expected values of certain combinatorial problems (e.g., minimum spanning tree). The stochastic geometry model is also of fundamental interest in the area of wireless networks. In many applications, we only have some prior information about the locations of the transmission nodes (e.g., some sensors that will be deployed randomly in a designated area by an aircraft). Such a stochastic wireless network can be captured precisely by this model. See the recent survey [19] and more references therein.

Stochastic Geometry Models: In this paper, we focus on two stochastic geometry models, the locational uncertainty model and existential uncertainty model.

1. (Locational Uncertainty Model) We are given a metric space $\mathcal{P}$. The location of each node $v \in \mathcal{V}$ is a random point in the metric space $\mathcal{P}$ and the probability distribution is given as the input. Formally, we use the term *nodes* to refer to the vertices of the graph, *points* to describe the locations of the nodes in the metric space. We denote the set of nodes as $\mathcal{V} = \{v_1, \ldots, v_n\}$ and the set of points as $\mathcal{P} = \{s_1, \ldots, s_m\}$, where $n = |\mathcal{V}|$ and $m = |\mathcal{P}|$. A realization $\mathbf{r}$ can be represented by an n-dimensional vector $(r_1, \ldots, r_n) \in \mathcal{P}^n$ where point r_i is the location of node v_i for $1 \leq i \leq n$. Let $\mathcal{R}$ denote the set of all possible realizations. We assume that the distributions of the locations of nodes in the metric space $\mathcal{P}$ are independent, thus $\mathbf{r}$ occurs with probability $\Pr[\mathbf{r}] = \prod_{i\in[n]} p_{v_i r_i}$, where p_{vs} represents the probability that the location of node v is point $s \in \mathcal{P}$. The model is also termed as the *locational uncertainty model* in [20].
2. (Existential Uncertainty Model) A closely related model is the *existential uncertainty model* where the location of a node is a fixed point in the given metric space, but the existence of the node is probabilistic. In this model, we use p_i to denote the probability that node v_i exists (if exists, its location is s_i). A realization $\mathbf{r}$ can be represented by a subset $S \subset \mathcal{P}$ and $\Pr[\mathbf{r}] = \prod_{s_i\in S} p_i \prod_{s_i\notin S}(1 - p_i)$.

Problem Formulation: We are interested in following natural problem in the above models: estimating the expected values of certain statistics of combinatorial objects. In this paper, we study several combinatorial or geometry problems in these two models: the closest pair problem, minimum spanning tree, minimum perfect matching (assuming an even number of nodes), k-clustering and minimum cycle cover. We take the minimum spanning tree problem for example. Let MST be the length of the minimum spanning tree (which is a random variable) and $\mathsf{MST}(\mathbf{r})$ be the length of the minimum spanning tree spanning all points in the realization $\mathbf{r}$. We would like to estimate the following quantity:

$$\mathbb{E}[\mathsf{MST}] = \sum_{\mathbf{r}\in\mathcal{R}} \Pr[\mathbf{r}] \cdot \mathsf{MST}(\mathbf{r}).$$

However, the above formula does not give us an efficient way to estimate the expectation since it involves an exponential number of terms. In fact, computing the exact expected value (for the problems considered in this paper) are either NP-hard or #P-hard. Following many of the theoretical computer science literatures on approximate counting and estimation, our goal is to obtain fully polynomial randomized approximation schemes for computing the expected values.

Table 1. Our results for some problems in different stochastic models

Problems		Existential	Locational
Closest Pair (§2)	$\mathbb{E}[\mathsf{C}]$	FPRAS	FPRAS
	$\Pr[\mathsf{C} \leq 1]$	FPRAS	FPRAS
	$\Pr[\mathsf{C} \geq 1]$	Inapprox	Inapprox
Diameter (§2)	$\mathbb{E}[\mathsf{D}]$	FPRAS	FPRAS
	$\Pr[\mathsf{D} \leq 1]$	Inapprox	Inapprox
	$\Pr[\mathsf{D} \geq 1]$	FPRAS	FPRAS
Minimum Spanning Tree (§3)	$\mathbb{E}[\mathsf{MST}]$	FPRAS[20]	FPRAS
k-Clustering	$\mathbb{E}[\mathsf{kCL}]$	FPRAS	Open
Perfect Matching (§4)	$\mathbb{E}[\mathsf{PM}]$	N.A.	FPRAS
kth Closest Pair	$\mathbb{E}[\mathsf{kC}]$	FPRAS	Open
Cycle Cover	$\mathbb{E}[\mathsf{CC}]$	FPRAS	FPRAS
kth Longest m-Nearest Neighbor	$\mathbb{E}[\mathsf{kmNN}]$	FPRAS	Open

1.1 Our Contributions

We recall that a *fully polynomial randomized approximation scheme (FPRAS)* for a problem f is a randomized algorithm A that takes an input instance x, a real number $\epsilon > 0$, returns $A(x)$ such that $\Pr[(1-\epsilon)f(x) \leq A(x) \leq (1+\epsilon)f(x)] \geq \frac{3}{4}$ and its running time is polynomial in both the size of the input n and $1/\epsilon$. Our main contributions can be summarized in Table 1. We need to explain some entries in the table in more details.

1. Closest Pair: We use C to denote the minimum distance of any pair of two nodes. If a realization has less than two nodes, C is zero. Computing $\Pr[\mathsf{C} \leq 1]$ exactly in the existential model is known to be #P-hard even in an Euclidean plane [21], but no nontrivial algorithmic result is known before. So is computing $\Pr[\mathsf{C} \geq 1]$. In fact, it is not hard to show that computing $\Pr[\mathsf{C} \geq 1]$ is imapproximable within any factor in a metric space.
   We also consider the problem of computing expected distance $\mathbb{E}[\mathsf{C}]$ between the closest pair in the same model. We prove that the problem is #P-hard and give the first known FPRAS in Section 2. Note that an FPRAS for computing $\Pr[\mathsf{C} \leq 1]$ does not imply an FPRAS for computing $\mathbb{E}[\mathsf{C}]$ [1].
2. Diameter: The problem of computing the expected length of the diameter can be reduced to the closest pair problem as follows. Assume that the longest distance between two points in $\mathcal{P}$ is W. We construct the new instance $\mathcal{P}'$ as follows: for any two points $u, v \in \mathcal{P}$, let their distance be $2W - \mathrm{d}(u, v)$ in $\mathcal{P}'$. The new instance is still a metric. The sum of the distance of closest pair in $\mathcal{P}$ and the diameter in $\mathcal{P}'$ is exactly $2W$ (if there are at least two realized points). Hence, the answer for the diameter can be easily derived from the answer for closest pair in $\mathcal{P}'$.
3. Minimum Spanning Tree: Computing $\mathbb{E}[\mathsf{MST}]$ exactly in both uncertainty models is known to be #P-hard [20]. Kamousi, Chan, and Suri [20] developed an FPRAS

[1] To the contrary, an FPRAS for computing $\Pr[\mathsf{C} \geq 1]$ or $\Pr[\mathsf{C} = 1]$ would imply an FPRAS for computing $\mathbb{E}[\mathsf{C}]$ since $\mathbb{E}[\mathsf{C}] = \sum_{(s_i,s_j)} \Pr[\mathsf{C} = \mathrm{d}(s_i, s_j)]\mathrm{d}(s_i, s_j) = \int \Pr[\mathsf{C} \geq t]\mathrm{d}t = \sum_{(s_i,s_j)} \Pr[\mathsf{C} \geq \mathrm{d}(s_i, s_j)](\mathrm{d}(s_i, s_j) - \mathrm{d}(s_i', s_j'))$.

for estimating $\mathbb{E}[\mathsf{MST}]$ in the existential uncertainty model and a constant factor approximation algorithm in the locational uncertainty model.
Estimating $\mathbb{E}[\mathsf{MST}]$ is amendable to several techniques. We obtain an FPRAS for estimating $\mathbb{E}[\mathsf{MST}]$ in the locational uncertainty model using the stoch-core technique in Section 3. In fact, the idea in [20] can also be extended to give an alternative FPRAS. It is not clear how to extend their idea to other problems.

4. Clustering (k-clustering): In the deterministic k-clustering problem, we want to partition all points into k disjoint subsets such that the spacing of the partition is maximized, where the spacing is defined to be the minimum of any $\mathrm{d}(u, v)$ with u, v in different subsets [23]. In fact, the optimal cost of the problem is the length of the $(k-1)$th most expensive edge in the minimum spanning tree [23]. We show how to estimate $\mathbb{E}[\mathsf{kCL}]$ using the HPF (hierarchical partition family) technique.
5. Perfect Matching: We assume that there are even number of nodes to ensure that a perfect matching always exists. Therefore, only the locational uncertainty model is relevant here. We give the first FPRAS for approximating the expected length of minimum perfect matching in Section 4 using a more complicated stoch-core technique.

All of our algorithms run in polynomial time. However, we have not attempted to optimize the exact running time.

Our techniques: Perhaps the simplest and the most commonly used technique for estimating the expectation of a random variable is the Monte Carlo method, that is to use the sample average as the estimate. However, the method is only efficient (i.e., runs in polynomial time) if the variance of the random variable is small (See Lemma 1). To circumvent the difficulty caused by the high variance, a general methodology is to decompose the expectation of the random variable into a convex combination of conditional expectations using the law of total expectation: $\mathbb{E}[X] = \mathbb{E}_Y\big[\mathbb{E}[X \mid Y]\big] = \sum_y \Pr[Y = y]\,\mathbb{E}[X \mid Y = y]$. Hopefully, $\Pr[Y = y]$ can be estimated (or calculated exactly) efficiently, and the random variable X conditioning on each event y has a low variance. However, choosing the events Y to condition on can be tricky.

We develop two new techniques for choosing such events, each being capable of solving a subset of aforementioned problems. In the first technique, we first identify a set $\mathcal{H}$ of points, called the *stoch-core* of the problem, such that (1): with high probability, all nodes realize in $\mathcal{H}$ and (2): conditioning on event (1), the variance is small. Then, we choose Y to be the number of nodes realized to points not in $\mathcal{H}$. We compute the $(1 \pm \epsilon)$-estimates for $Y = 0, 1$ using Monte Carlo by (1) and (2). The problematic part is when Y is large, i.e., many nodes realize to points outside $\mathcal{H}$. Even though the probability of such events is very small, the value of X under such events may be considerably large, thus contributing nontrivially. However, we can show that the contribution of such events is dominated by the first few events and thus can be safely ignored. Choosing appropriate stoch-core is easy for some problems, such as closest pair and minimum spanning tree, while it may require additional idea for other problems such as minimum perfect matching.

Our second technique utilizes a notion called *Hierarchical Partition Family (HPF)*. The HPF has m levels, each representing a clustering of all points. For a combinatorial problem, for which the solution is a set of edges, we define Y to be the highest level

such that some edge in the solution is an inter-cluster edge. Informally, conditioning on the information of Y, we can essentially bound the variance of X (hence use the Monte Carlo method). To implement Monte Carlo, we need to be able to take samples efficiently conditioning on Y. We show that such sampling problems can be reduced to, or have connections to, classical approximate counting and sampling problems, such as approximating permanent, counting knapsack.

Due to space constraints, we omit many details, which can be found in the full version of this paper[2].

1.2 Related Work

Several geometric properties of a set of stochastic points have been studied extensively in the literature under the term *stochastic geometry*. For instance, Bearwood et al. [8] shows that if there are n points uniformly and independently distributed in $[0, 1]^2$, the minimal traveling salesman tour visiting them has an expected length $\Omega(\sqrt{n})$. Asymptotic results for minimum spanning trees and minimum matchings on n points uniformly distributed in unit balls are established by Bertsimas and van Ryzin [10]. Similar results can be found in e.g., [9,22,31]. Compared with results in stochastic geometry, we focus on the efficient computation of the statistics, instead of giving explicit mathematical formulas.

Recently, a number of researchers have begun to explore geometric computing under uncertainty and many classical computational geometry problems have been studied in different stochastic/uncertainty models. Agarwal, Cheng, Tao and Yi [4] studied the problem of indexing probabilistic points with continuous distributions for range queries on a line. Agarwal, Efrat, Sankararaman, and Zhang [5] also studied the same problem in the locational uncertainty model under Euclidean metric. The most probable k-nearest neighbor problem and its variants have attracted a lot of attentions in the database community (See e.g., [11]). Several other problems have also been considered recently, such as computing the expected volume of a set of probabilistic rectangles in a Euclidean space [36], convex hulls [2], skylines (Pareto curves) over probabilistic points [1,7], and shape fitting [27].

Kamousi, Chan and Suri [20] initiated the study of estimating the expected length of combinatorial objects in this model. They showed that computing the expected length of the nearest neighbor (NN) graph, the Gabriel graph (GG), the relative neighborhood graph (RNG), and the Delaunay triangulation (DT) can be solved exactly in polynomial time, while computing $\mathbb{E}[\mathsf{MST}]$ is #P-hard and there exists a simple FPRAS for approximating $\mathbb{E}[\mathsf{MST}]$ in the existential model. They also gave a deterministic PTAS for approximating $\mathbb{E}[\mathsf{MST}]$ in an Euclidean plane. In another paper [21], they studied the closest pair and (approximate) nearest neighbor problems (i.e., finding the point with the smallest expected distance from the query point) in the same model.

The *randomly weighted graph* model where the edge weights are independent non-negative variables has also been studied extensively. Frieze [16] and Steele [32] showed that the expected value of the minimum spanning tree on such a graph with identically and independently distributed edges is $\zeta(3)/D$ where $\zeta(3) = \sum_{j=1}^{\infty} 1/j^3$ and D is the derivative of the distribution at 0. Alexopoulos and Jacobson [6] developed algorithms

[2] http://arxiv.org/abs/1209.5828

that compute the distribution of MST and the probability that a particular edge belongs to MST when edge lengths follow discrete distributions. However, the running times of their algorithms may be exponential in the worst cases. Recently, Emek, Korman and Shavitt [15] showed that computing the kth moment of a class of properties, including the diameter, radius and minimum spanning tree, admits an FPRAS for each fixed k. Our model differs from their model in that the edge lengths are not independent.

The computational/algorithmic aspects of stochastic geometry have also gained a lot of attention in recent years from the area of wireless networking. In many application scenarios, it is common to assume that the nodes (e.g., sensors) are deployed randomly across a certain area, thereby forming a stochastic network. It is of central importance to study various properties in this network, such as connectivity [17], transmission capacity [18]. We refer interested reader to a recent survey [19] for more references.

1.3 Preliminaries

Before describing our main results, we first consider the straightforward Monte Carlo strategy, which is an important building block in our later developments. Suppose we want to estimate $\mathbb{E}[X]$. In each Monte Carlo iteration, we take a sample (a realization of all nodes), and compute the value of X for the sample. At the end, we output the average over all samples. The number of samples required by this algorithm is suggested by the following standard Chernoff bound.

Lemma 1. (Chernoff Bound) *Let random variables $X_1, X_2, \ldots, X_N$ be independent random variables taking on values between 0 and U. Let $X = \frac{1}{N}\sum_{i=1}^{N} X_i$ and μ be the expectation of X, for any $\epsilon > 0$,*

$$\Pr[X \in [(1-\epsilon)\mu, (1+\epsilon)\mu]] \geq 1 - 2e^{-N\frac{\mu}{U}\epsilon^2/4}.$$

Therefore, for any $\epsilon > 0$, in order to get an $(1 \pm \epsilon)$-approximation with probability $1 - \frac{1}{\text{poly}(n)}$, the number of samples needs to be $O(\frac{U}{\mu\epsilon^2}\log n)$. If $\frac{U}{\mu}$, the ratio between the maximum possible value of X and the expected value $\mathbb{E}[X]$, is bounded by $\text{poly}(m, n, \frac{1}{\epsilon})$, we can use the above Monte Carlo method to estimate $\mathbb{E}[X]$ with a polynomial number of samples. Since we use this condition often, we devote a separate definition to it.

Definition 1. *We call a random variable X* poly-bounded *if the ratio between the maximum possible value of X and the expected value $\mathbb{E}[X]$ is bounded by* $\text{poly}(m, n, \frac{1}{\epsilon})$.

2 The Closest Pair Problem

2.1 Estimating Pr[C ≤ 1]

As a warmup, we first demonstrate how to use the stoch-core technique for the closest pair problem in the existential uncertainty model. Given a set of points $\mathcal{P} = \{s_1, \ldots, s_m\}$ in the metric space, where each point $s_i \in \mathcal{P}$ is present with probability p_i. We use C to denote the distance between the closest pair of vertices in the realized graph. If the

realized graph has less than two points, C is zero. The goal is to compute the probability $\Pr[\mathsf{C} \le 1]$.

For a set H of points and a subset $S \subseteq H$, we use $H\langle S\rangle$ to denote the event that among all points in H, all and only points in S are present. For any nonnegative integer i, let $H\langle i\rangle$ denote the event $\bigvee_{S\subseteq H:|S|=i} H\langle S\rangle$, i.e., the event that exactly i points are present in H.

The *stoch-core* of the closest pair problem is simply defined to be $\mathcal{H} = \left\{s_i \mid p_i \ge \frac{\epsilon}{m^2}\right\}$. Let $\mathcal{F} = \mathcal{P} \setminus \mathcal{H}$. We consider the decomposition

$$\Pr[\mathsf{C} \le 1] = \sum_{i=0}^{|\mathcal{F}|} \Pr[\mathcal{F}\langle i\rangle \wedge \mathsf{C} \le 1] = \sum_{i=0}^{|\mathcal{F}|} \Pr[\mathcal{F}\langle i\rangle] \cdot \Pr[\mathsf{C} \le 1 \mid \mathcal{F}\langle i\rangle].$$

Our algorithm is very simple: estimate the first three terms (i.e., $i = 0, 1, 2$) and use their sum as our final answer.

We can see that $\mathcal{H}$ satisfies the two properties of a stoch-core mentioned in the introduction:

1. The probability that all nodes are realized in $\mathcal{H}$, i.e., $\Pr[\mathcal{F}\langle 0\rangle]$, is at least $1 - m \cdot \frac{\epsilon}{m^2} = 1 - \frac{\epsilon}{m}$;
2. If there exist two points $s_i, s_j \in \mathcal{H}$ such that $\mathrm{d}(s_i, s_j) \le 1$, we have $\Pr[\mathsf{C} \le 1 \mid \mathcal{F}\langle 0\rangle] \ge \frac{\epsilon^2}{m^4}$; otherwise, $\Pr[\mathsf{C} \le 1 \mid \mathcal{F}\langle 0\rangle] = \Pr[\mathcal{H}\langle 0\rangle \mid \mathcal{F}\langle 0\rangle] + \Pr[\mathcal{H}\langle 1\rangle \mid \mathcal{F}\langle 0\rangle]$. Note that we can compute $\Pr[\mathcal{H}\langle 0\rangle \mid \mathcal{F}\langle 0\rangle]$ and $\Pr[\mathcal{H}\langle 1\rangle \mid \mathcal{F}\langle 0\rangle]$ in polynomial time. We do not consider this case in the following analysis.

Both properties guarantee that the random variable $I(\mathsf{C} \le 1)$, conditioned on $\mathcal{F}\langle 0\rangle$, is poly-bounded, hence we can easily get a $(1 \pm \epsilon)$-estimation for $\Pr[\mathcal{F}\langle 0\rangle \wedge \mathsf{C} \le 1]$ with polynomial many samples with high probability. Similarly, $\Pr[\mathcal{F}\langle i\rangle \wedge \mathsf{C} \le 1]$ can also be estimated with polynomial number of samples for $i = 1, 2$. The algorithm can be found in Algorithm 1.

Algorithm 1. Estimating $\Pr[\mathsf{C} \le 1]$

1 Estimate $\Pr[\mathcal{F}\langle 0\rangle \wedge \mathsf{C} \le 1]$: Take $N_0 = O\big((m/\epsilon)^4 \ln m\big)$ independent samples. Suppose M_0 is the number of samples satisfying $\mathsf{C} \le 1$ and $\mathcal{F}\langle 0\rangle$. $T_0 \leftarrow \frac{M_0}{N_0}$.

2 Estimate $\Pr[\mathcal{F}\langle 1\rangle \wedge \mathsf{C} \le 1]$: For each point $s_i \in \mathcal{F}$, take $N_1 = O((m/\epsilon)^4 \ln m)$ independent samples conditioning on the event $\mathcal{F}\langle\{s_i\}\rangle$. Suppose there are M_i samples satisfying $\mathsf{C} \le 1$. $T_1 \leftarrow \sum_{s_i\in\mathcal{F}} p_i M_i / N_1$.

3 Estimate $\Pr[\mathcal{F}\langle 2\rangle \wedge \mathsf{C} \le 1]$: For each point pair $s_i, s_j \in \mathcal{F}$, take $N_2 = O((m/\epsilon)^4 \ln m)$ independent samples conditioning on the event $\mathcal{F}\langle\{s_i, s_j\}\rangle$. Suppose there are M_{ij} samples satisfying $\mathsf{C} \le 1$. $T_2 \leftarrow \sum_{s_i,s_j\in\mathcal{F}} p_i p_j M_{ij} / N_2$.

4 **Output:** $T_0 + T_1 + T_2$

Lemma 2. *Steps 1,2,3 in Algorithm 1 provide* $(1\pm\epsilon)$*-approximations for* $\Pr[\mathcal{F}\langle i\rangle \wedge \mathsf{C} \le 1]$ *for* $i = 0, 1, 2$ *respectively, with high probability.*

Theorem 1. *There is an FPRAS for estimating the probability of the distance between the closest pair of nodes is at most* 1 *in the existential uncertainty model.*

Proof. We only need to show that the contribution from the rest of terms (where more than two points outside stoch-core $\mathcal{H}$ are present) is negligible compared to the third term. Suppose S is the set of all present points such that $\mathsf{C} \leq 1$ and there are at least 3 points not in $\mathcal{H}$. Suppose s_i, s_j are the closest pair in S. We associate S with a smaller set $S' \subset S$ by making 1 present point in $(S \cap \mathcal{F}) \setminus \{s_i, s_j\}$ absent (if there are several such S', we choose an arbitrary one). We denote it as $S \sim S'$. We use the notation $S \in F_i$ to denote that the realization S satisfies $(\mathcal{F}\langle i\rangle \wedge \mathsf{C} \leq 1)$. Then, we can see that for $i \geq 3$,

$$\Pr[\mathcal{F}\langle i\rangle \wedge \mathsf{C} \leq 1] = \sum_{S:S\in F_i} \Pr[S] \leq \sum_{S':S'\in F_{i-1}} \sum_{S:S\sim S'} \Pr[S].$$

For a fixed S', there are at most m different sets S such that $S \sim S'$ and $\Pr[S] \leq \frac{2\epsilon}{m^2}\Pr[S']$ for any such S. Hence, we have that $\sum_{S:S\sim S'} \Pr[S] \leq \frac{2\epsilon}{m}\Pr[S']$. Therefore,

$$\Pr[\mathcal{F}\langle i\rangle \wedge \mathsf{C} \leq 1] \leq \frac{2\epsilon}{m} \cdot \sum_{S':S'\in F_{i-1}} \Pr[S'] = \frac{2\epsilon}{m} \cdot \Pr[\mathcal{F}\langle i-1\rangle \wedge \mathsf{C} \leq 1].$$

Hence, overall we have $\sum_{i\geq 3} \Pr[\mathcal{F}\langle i\rangle \wedge \mathsf{C} \leq 1] \leq \epsilon \Pr[\mathcal{F}\langle 2\rangle \wedge \mathsf{C} \leq 1]$. This finishes the analysis.

□

2.2 Estimating $\mathbb{E}[\mathsf{C}]$

In this section, we consider the problem of estimating $\mathbb{E}[\mathsf{C}]$, where C is the distance of the closest pair of present points, in the existential uncertainty model. Now, we introduce our second main technique, the *hierarchical partition family (HPF)* technique, to solve this problem. An HPF is a family Ψ of partitions of $\mathcal{P}$, formally defined as follows.

Definition 2. *(Hierarchical Partition Family (HPF)) Let T be any minimum spanning tree spanning all points of $\mathcal{P}$. Suppose that the edges of T are $e_1, \ldots, e_{m-1}$ with $\mathrm{d}(e_1) \geq \mathrm{d}(e_2) \geq \ldots \geq \mathrm{d}(e_{m-1})$. Let $E_i = \{e_i, e_{i+1}, \ldots, e_{m-1}\}$. The HPF $\Psi(\mathcal{P})$ consists of m partitions $\Gamma_1, \ldots, \Gamma_m$. Γ_1 is the entire point set $\mathcal{P}$. Γ_i consists of i disjoint subsets of $\mathcal{P}$, each corresponding to a connected component of $G_i = G(\mathcal{P}, E_i)$. Γ_m consists of all singleton points in $\mathcal{P}$. It is easy to see that Γ_j is a refinement of Γ_i for $j > i$. Consider two consecutive partitions Γ_i and Γ_{i+1}. Note that G_i contains exactly one more edge (i.e., e_i) than G_{i+1}. Let μ'_{i+1} and μ''_{i+1} be the two components (called the* split components*) in Γ_{i+1}, each containing an endpoint of e_i. Let $v_i \in \Gamma_i$ be the connected component of G_i that contains e_i. We call v_i the* special component *in Γ_i. Let $\Gamma'_i = \Gamma_i \setminus v_i$.*

We observe two properties of $\Psi(\mathcal{P})$ that are useful later.

P1. Consider a component $C \in \Gamma_i$. Let s_1, s_2 be two arbitrary points in C. Then $\mathrm{d}(s_1, s_2) \leq (m-1)\mathrm{d}(e_i)$ (this is because s_1 and s_2 are connected in G_i, and e_i is the longest edge in G_i).

P2. Consider two different components C_1 and C_2 in Γ_i. Let $s_1 \in C_1$ and $s_2 \in C_2$ be two arbitrary points. Then $\mathrm{d}(s_1, s_2) \geq \mathrm{d}(e_{i-1})$ (this is because the minimum inter-component distance is $\mathrm{d}(e_{i-1})$ in G_i).

Let the random variable Y be smallest integer i such that there is at most one present point in each component of Γ_{i+1}. Note that if $Y = i$ then each component of Γ_i contains at most one point, except that the special component v_i contains exactly two present points. The following lemma is a simple consequence of P1 and P2.

Lemma 3. *Conditioning on $Y = i$, it holds that $\mathrm{d}(e_i) \leq \mathsf{C} \leq m\mathrm{d}(e_i)$ (hence, C is poly-bounded).*

Consider the following expansion of $\mathbb{E}[\mathsf{C}]$: $\mathbb{E}[\mathsf{C}] = \sum_{i=1}^{m-1} \Pr[Y = i]\mathbb{E}[\mathsf{C} \mid Y = i]$. For a fixed i, $\Pr[Y = i]$ can be estimated as follows: For a component $C \subset \mathcal{P}$, we use $C\langle j\rangle$ to denote the event that exactly j points in C are present, $C\langle s\rangle$ the event that only s is present in C and $C\langle \leq j\rangle$ the event that no more than j points in C are present. Let μ_i' and μ_i'' be the two split components in Γ_i. Note that

$$\Pr[Y = i] = \Pr[\mu_{i+1}'\langle 1\rangle] \cdot \Pr[\mu_{i+1}''\langle 1\rangle] \cdot \prod_{C \in \Gamma_i'} \Pr[C\langle \leq 1\rangle].$$

The remaining is to show how to estimate $\mathbb{E}[\mathsf{C} \mid Y = i]$. Since C is poly-bounded, it suffices to give an efficient algorithm to take samples conditioning on $Y = i$. This is again not difficult: We take exactly one point $s \in \mu_{i+1}'$ with probability $\Pr[\mu_{i+1}'\langle s\rangle]/\Pr[\mu_{i+1}'\langle 1\rangle]$. Same for μ_{i+1}''. For each $C \in \Gamma_i'$, take no point from C with probability $\Pr[C\langle 0\rangle]/\Pr[C\langle \leq 1\rangle]$; otherwise, take exactly one point $s \in C$ with probability $\Pr[C\langle s\rangle]/\Pr[C\langle \leq 1\rangle]$. This finishes the description of the FPRAS in the existential uncertainty model.

Theorem 2. *There is an FPRAS for estimating the expected distance between the closest pair of nodes in the existential uncertainty models.*

3 Minimum Spanning Trees

We consider the problem of estimating the expected size of minimum spanning tree in the locational uncertainty model. In this section, we briefly sketch how to solve it using our stoch-core method. Recall that the term nodes refers to the vertices $\mathcal{V}$ of the spanning tree and points describes the locations in $\mathcal{P}$. For ease of exposition, we assume that for each point, there is only one node that may realize at this point.

Recall that we use the notation $v \vDash s$ to denote the event that node v is present at point s. Let $p_{vs} = \Pr[v \vDash s]$. Since node v is realized with certainty, we have $\sum_{s \in \mathcal{P}} p_{vs} = 1$. For each point $s \in \mathcal{P}$, we let $p(s)$ denote the probability that point s is present. For a set H of points, let $p(H) = \sum_{s \in H} p(s)$, i.e., the expected number of points present in H. For a set H of points and a set S of nodes, we use $H\langle S\rangle$ to denote the event that all and only nodes in S are realized to some points in H. If S only contains one node, say v, we use the notation $H\langle v\rangle$ as the shorthand for $H\langle\{v\}\rangle$. Let $H\langle i\rangle$ denote the event

$\bigvee_{S:|S|=i} H\langle S\rangle$, i.e., the event that exactly i nodes are in H. We use $\mathrm{diam}(H)$, called the diameter of H, to denote $\max_{s,t\in H} \mathrm{d}(s,t)$. Let $\mathrm{d}(p,H)$ be the closest distance between point p and any point in H.

Finding stoch-core: We find the stoch-core $\mathcal{H} \leftarrow \mathsf{B}(s, \mathrm{d}(s,t)) = \{s' \in \mathcal{P} \mid \mathrm{d}(s',s) \le \mathrm{d}(s,t)\}$, where points s and t are the furthest two points among all points r with $p(r) \ge \frac{\epsilon}{16m}$.

Lemma 4. *The stoch-core $\mathcal{H}$ satisfies the following properties:*

Q1. $p(\mathcal{H}) \ge n - \frac{\epsilon}{16} = n - O(\epsilon)$

Q2. $\mathbb{E}[\,\mathsf{MST} \mid \mathcal{H}\langle n\rangle\,] = \Omega\left(\mathrm{diam}(\mathcal{H})\frac{\epsilon^2}{m^2}\right)$.

Furthermore, the algorithm runs in linear time.

Estimating $\mathbb{E}[\mathsf{MST}]$: Let $\mathcal{F} = \mathcal{P}\backslash\mathcal{H}$. By the law of total expectation, the expected length of the minimum spanning tree can be expanded as follows: $\mathbb{E}[\mathsf{MST}] = \sum_{i\ge 0} \mathbb{E}[\,\mathsf{MST} \mid \mathcal{F}\langle i\rangle\,] \cdot \Pr[\mathcal{F}\langle i\rangle]$. We only estimate the first two terms $\mathbb{E}[\,\mathsf{MST} \mid \mathcal{F}\langle 0\rangle\,] \cdot \Pr[\mathcal{F}\langle 0\rangle]$ and $\mathbb{E}[\,\mathsf{MST} \mid \mathcal{F}\langle 1\rangle\,] \cdot \Pr[\mathcal{F}\langle 1\rangle]$ and use their sum as our final estimation. Using Properties Q1 and Q2, we can estimate the two terms in polynomial time.

Theorem 3. *There is an FPRAS for estimating the expected length of the minimum spanning tree in the locational uncertainty model.*

4 Minimum Perfect Matchings

In this section, we consider the minimum perfect matching (PM) problem. We use the stoch-core method.

Finding stoch-core: First, we show how to find in poly-time the stoch-core $\mathcal{H}$. See the Pseudo-code in Algorithm 2 for details.

Algorithm 2. Constructing stoch-core $\mathcal{H}$ for Estimating $\mathbb{E}[\mathsf{PM}]$

1 Initially, $t \leftarrow 0$ and each point $s \in \mathcal{P}$ is a component $\mathcal{H}_{\{s\}} = \mathsf{B}(s,t)$ by itself.
2 Gradually increase t; If two different components $\mathcal{H}_{S_1}$ and $\mathcal{H}_{S_2}$ intersect (where $\mathcal{H}_S := \cup_{s\in S}\mathsf{B}(s,t)$); Merge them into a new component $\mathcal{H}_{S_1\cup S_2}$.
3 Stop increasing t while the first time the following two conditions are satisfied by components at t:
  Q1. For each node v, there is a unique component $\mathcal{H}_j$ such that $p_v(\mathcal{H}_j) \ge 1 - O(\frac{\epsilon}{nm^3})$. We call $\mathcal{H}_j$ the stoch-core of node v, denoted as $\mathcal{H}(v)$.
  Q2. For all j, $|\{v \in \mathcal{V} \mid \mathcal{H}(v) = \mathcal{H}_j\}|$ is even.
4 Output the stopping time T and the components $\mathcal{H}_1, \ldots, \mathcal{H}_k$.

Estimating $\mathbb{E}[\mathsf{PM}]$: We use $\mathcal{H}\langle n\rangle$ to denote the event that for each node v, $v \vDash \mathcal{H}(v)$. We denote the event that there are exactly i nodes which are realized out of their stoch-cores by $\mathcal{F}\langle i\rangle$. Again, we only need to estimate two terms: $\mathbb{E}[\mathsf{PM} \mid \mathcal{F}\langle 0\rangle]] \cdot \Pr[\mathcal{F}\langle 0\rangle]$ and $\mathbb{E}[\mathsf{PM} \mid \mathcal{F}\langle 1\rangle] \cdot \Pr[\mathcal{F}\langle 1\rangle]$. Using Properties Q1 and Q2, we can estimate these terms in polynomial time. Our final estimation is simply the sum of the first two terms.

Theorem 4. *Assuming the locational uncertainty model and that the number of nodes is even, there is an FPRAS for estimating the expected length of the minimum perfect matching.*

References

1. Afshani, P., Agarwal, P.K., Arge, L., Larsen, K.G., Phillips, J.M.: (Approximate) Uncertain skylines. In: Proceedings of the 14th International Conference on Database Theory, pp. 186–196. ACM (2011)
2. Agarwal, P.K., Har-Peled, S., Suri, S., Yıldız, H., Zhang, W.: Convex Hulls under Uncertainty. In: Schulz, A.S., Wagner, D. (eds.) ESA 2014. LNCS, vol. 8737, pp. 37–48. Springer, Heidelberg (2014)
3. Agarwal, P.K., Cheng, S.-W., Yi, K.: Range searching on uncertain data. ACM Transactions on Algorithms (TALG) 8(4), 43 (2012)
4. Agarwal, P.K., Cheng, S.W., Tao, Y., Yi, K.: Indexing uncertain data. In: Proceedings of the Twenty-Eighth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 137–146. ACM (2009)
5. Agarwal, P.K., Efrat, A., Sankararaman, S., Zhang, W.: Nearest-neighbor searching under uncertainty. In: Proceedings of the 31st Symposium on Principles of Database Systems, pp. 225–236. ACM (2012)
6. Alexopoulos, C., Jacobson, J.A.: State space partition algorithms for stochastic systems with applications to minimum spanning trees. Networks **35**(2), 118–138 (2000)
7. Atallah, M.J., Qi, Y., Yuan, H.: Asymptotically efficient algorithms for skyline probabilities of uncertain data. ACM Trans. Datab. Syst **32**(2), 12 (2011)
8. Beardwood, J., Halton, J.H., Hammersley, J.M.: The shortest path through many points. Proc. Cambridge Philos. Soc. **55**, 299–327 (1959)
9. Bern, M.W., Eppstein, D.: Worst-case bounds for suadditive geometric graphs. In: Symposium on Computational Geometry, pp. 183–188 (1993)
10. Bertsimas, D.J., van Ryzin, G.: An asymptotic determination of the minimum spanning tree and minimum matching constants in geometrical probability. Operations Research Letters **9**(4), 223–231 (1990)
11. Cheng, R., Chen, J., Mokbel, M., Chow, C.: Probabilistic verifiers: Evaluating constrained nearest-neighbor queries over uncertain data. In: ICDE (2008)
12. Cheng, R., Chen, J., Xie, X.: Cleaning uncertain data with quality guarantees. Proceedings of the VLDB Endowment **1**(1), 722–735 (2008)
13. Deshpande, A., Guestrin, C., Madden, S.R., Hellerstein, J.M., Hong, W.: Model-driven data acquisition in sensor networks. In: Proceedings of the Thirtieth International Conference on Very Large Data Bases, vol. 30, pp. 588–599. VLDB Endowment (2004)
14. Dong, X., Halevy, A.Y., Yu, C.: Data integration with uncertainty. In: Proceedings of the 33rd International Conference on Very Large Data Bases, pp. 687–698. VLDB Endowment (2007)
15. Emek, Y., Korman, A., Shavitt, Y.: Approximating the statistics of various properties in randomly weighted graphs. In: Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1455–1467. SIAM (2011)
16. Frieze, A.M.: On the value of a random minimum spanning tree problem. Discrete Applied Mathematics **10**(1), 47–56 (1985)
17. Gupta, P., Kumar, P.R.: Critical power for asymptotic connectivity. In: Proceedings of the 37th IEEE Conference on Decision and Control, vol. 1, pp. 1106–1110. IEEE (1998)
18. Gupta, P., Kumar, P.R.: The capacity of wireless networks. IEEE Transactions on Information Theory **46**(2), 388–404 (2000)

19. Haenggi, M., Andrews, J.G., Baccelli, F., Dousse, O., Franceschetti, M.: Stochastic geometry and random graphs for the analysis and design of wireless networks. IEEE Journal on Selected Areas in Communications **27**(7), 1029–1046 (2009)
20. Kamousi, P., Chan, T.M., Suri, S.: Stochastic minimum spanning trees in euclidean spaces. In: Proceedings of the 27th Annual ACM Symposium on Computational Geometry, pp. 65–74. ACM (2011)
21. Kamousi, P., Chan, T.M., Suri, S.: Closest pair and the post office problem for stochastic points. Computational Geometry **47**(2), 214–223 (2014)
22. Karloff, H.J.: How long can a euclidean traveling salesman tour be? In: J. Discrete Math., p. 2(1). SIAM (1989)
23. Kleinberg, J., Eva, T.: Algorithm design. Pearson Education India (2006)
24. Li, J., Deshpande, A.: Ranking continuous probabilistic datasets. Proceedings of the VLDB Endowment **3**(1–2), 638–649 (2010)
25. Li, J., Phillips, J.M., Wang, H.: ϵ-kernel coresets for stochastic points. arXiv preprint arXiv:1411.0194 (2014)
26. Li, J., Wang, H.: Range Queries on Uncertain Data. In: Ahn, H.-K., Shin, C.-S. (eds.) ISAAC 2014. LNCS, vol. 8889, pp. 326–337. Springer, Heidelberg (2014)
27. Löffler, M., Phillips, J.M.: Shape Fitting on Point Sets with Probability Distributions. In: Fiat, A., Sanders, P. (eds.) ESA 2009. LNCS, vol. 5757, pp. 313–324. Springer, Heidelberg (2009)
28. Mainwaring, A., Culler, D., Polastre, J., Szewczyk, R., Anderson, J.: Wireless sensor networks for habitat monitoring. In: Proceedings of the 1st ACM International Workshop on Wireless Sensor Networks and Applications, pp. 88–97. ACM (2002)
29. Pfoser, D., Jensen, C.S.: Capturing the Uncertainty of Moving-Object Representations. In: Güting, R.H., Papadias, D., Lochovsky, F.H. (eds.) SSD 1999. LNCS, vol. 1651, pp. 111–131. Springer, Heidelberg (1999)
30. Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on stochastic programming: modeling and theory, vol. 16. SIAM (2014)
31. Snyder, T.L., Steele, J. M.: A priori bounds on the cuclidean traveling salesman. J. Comput., p. 24(3) (1995)
32. Steele, J.M.: On Frieze's $\zeta(3)$ limit for lengths of minimal spanning trees. Discrete Applied Mathematics **18**(1), 99–103 (1987)
33. Suciu, D., Olteanu, D., Ré, C., Koch, C.: Probabilistic databases. Synthesis Lectures on Data Management **3**(2), 1–180 (2011)
34. Swamy, C., Shmoys, D.B.: Approximation algorithms for 2-stage stochastic optimization problems 37(1), 33–46 (2006)
35. Szewczyk, R., Osterweil, E., Polastre, J., Hamilton, M., Mainwaring, A., Estrin, D.: Habitat monitoring with sensor networks. Communications of the ACM **47**(6), 34–40 (2004)
36. Yıldız, H., Foschini, L., Hershberger, J., Suri, S.: The Union of Probabilistic Boxes: Maintaining the Volume. In: Demetrescu, C., Halldórsson, M.M. (eds.) ESA 2011. LNCS, vol. 6942, pp. 591–602. Springer, Heidelberg (2011)