More efficient Algorithms for Stochastic Diameter and Some Unapproximated Problems in Metric Space

Daogao Liu[1]

Department of Physics, Tsinghua University, Beijing China
liudg16@mails.tsinghua.edu.cn

Abstract. Dealing with data on uncertainty has appealed to many researchers as there may be many stochastic problems in a realistic situation. In this paper, we study two basic uncertainty models: Existential Uncertainty Model where the location of each node is fixed while it may be absent with some probability, and the Locational Uncertainty Model where each node must be present, but the situation is uncertain. We consider the problem of estimating the expectation and the tail bound distribution of the diameter, and obtain an improved FPRAS(Fully Polynomial Randomized Approximation Scheme) which requires much fewer samples. In the meanwhile, we prove some problems in the two uncertainty models can't be approximated within any factor unless NP⊆ BPP by simple reductions.

Keywords: FPRAS · Stochastic Diameter · Hardness for approximation

1 Introduction

Models: As mentioned before, we focus on two stochastic geometry models, the existential uncertainty model and locational uncertainty model. We'll show the precise definition of these two models below:

Definition 1. *(Locational Uncertainty Model):We are given a metric space P. The location of each node $v \in V$ is a random point in the metric space 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 $V = \{v_1, ..., v_n\}$ and the set of points as $P = \{s_1, ..., s_m\}$, where $n = |V|$ and $m = |P|$. A realization r can be represented by an n-dimensional vector $(r_1, ..., r_n) \in P_n$ where point r_i is the location of node v_i for $1 \leq i \leq n$. Let R denote the set of all possible realizations. We assume that the distributions of the locations of nodes in the metric space P are independent, thus r occurs with probability $Pr[r] = \prod_{i\in[n]} p_{v_i r_i}$, where p_{vs} represents the probability that the location of node v is point $s \in P$.*

Definition 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 r can be represented by a subset $S \subset P$ and $Pr[r] = \prod_{s_i \in S} p_i \prod_{s_i \notin S}(1-p_i)$.*

Problem Formulation The natural problems in the above models are to estimate the expectation and the tail bound of distribution of certain combinatorial objects, denoted by E(Obj) and P(Obj$\geq$ 1)(or the form P(Obj$\leq$ 1)). More accurately, take the expectation of diameter(the longest distance between two realized points) as an example. Note the expectation E(D), and D($\mathbf{r}$) be the longest distance between two points in the realization $\mathbf{r}$. The precise definition of the E(D) is:

$$\mathrm{E(D)}{=}\sum_{\mathbf{r}\in R} Pr[\mathbf{r}]D(\mathbf{r})$$

Similarly, note the probability that the diameter is no less than the given threshold, i.e. P(D$\leq$ 1). And what we estimate in this paper is E(D) and P(D$\leq$ 1). The P(D$\geq$ 1) has been shown unapproxiable [13].

Preliminaries The most useful techniques in the estimation are the straightforward Monte Carlo strategy. We repeat the experiment and obtain the average of the experiment results, and use the average as the estimation of the true value. The number of samples required by this algorithm is suggested by the following standard Chernoff bound.

Lemma 1. *(Chernoff bound)Let random variables $X_1, X_2, ..., X_N$ be independent random variables taking on values between 0 and U. Let $X = \frac{1}{N}\sum_{i=1}^{N} X_i$ and $\mu = E(X)$.Then for any $\epsilon > 0$, we have $P((1-\epsilon)\mu \leq X \leq (1+\epsilon)\mu) \geq 1 - 2e^{-N\frac{\mu}{U}\epsilon^2/4}$*

Then if we want to get an $(1 \pm \epsilon)$ approximation with probability 1-$\frac{1}{poly(N)}$, the number of samples needs to be $O(\frac{U}{\mu\epsilon^2} lnN)$.

Call one realization of all nodes in both models **one sample**. So the main target of our algorithm in this paper is to bound the value $\frac{U}{\mu}$ and use as fewer samples as possible.

Take the Locational Uncertainty Model as an example. To simplify the argument of the running time, we assume the running time of experimenting with one node is nearly the same whatever its locational distribution is. However, it's difficult to argue that how much time it will take to do an experiment with one node. So we take one realization as one sample and use the necessary number of samples as the evaluation criterion of our algorithms.

Our Contributions Recall that the *fully polynomial randomized approximation scheme*(FPRAS) for a problem f is randomized algorithm A that takes an input instance x, a real number $\epsilon > 0$, returns A(x) such that $P[(1-\epsilon)f(x) \leq$

$A(x) \leq (1+\epsilon)f(x)] \geq \frac{3}{4}$ and its running time is polynomial in both the size of the input n and $1/\epsilon$.

We designed the FPRAS for E(D) and $P(D \geq 1)$ in both models which are the best of our knowledge.

Huang et al.[13] gives the FPRAS for E(closet pair) and P(Closet pair≤ 1)(denote by P(C≤ 1) later). The FPRAS for $P(C \leq 1)$ can be used to estimate $P(D \geq 1)$ with some trivial operations. In the existential uncertainty model, suppose there are m nodes, we improve the FPRAS from needing $O(\frac{m^6}{\epsilon^4} lnm)$ independent samples to $O(\frac{m}{\epsilon^4} lnm)$. As for a locational uncertainty model with m nodes and n points, we improve the FPRAS from needing $O(\frac{m^6}{\epsilon^4} lnm)$ samples to $O(\frac{m^3}{\epsilon^4} lnm)$.

As for the E(D), note that we can get an FPRAS for E(D) by the FPRAS for $P(D \geq 1)$, but it will need $O(\frac{m^8}{\epsilon^4} lnm)$ independent samples. We give the first direct FPRAS for E(D) in this paper which only needs $O(\frac{m^2}{\epsilon^2} ln^2 m)$ samples in the worst case and need $O(\frac{m^2}{\epsilon^2} lnm)$ samples in the best situation for both models. This direct FPRAS doesn't need to estimate $P(D \geq 1)$ anymore.

Moreover, we'll show some problems can't be approximated unless NP⊆BPP, which answers one of the open problems given in [13]. The main results of unapproximation are shown in the below table:

Table 1. Results for unapproximated problems

Unapproximable value	model	NPC problem for reduction
P(k-th closest pair≤ 1)	Loc	max 2-SAT
P(k-th longest m-nearest neighbor≤ 1)	Loc	Maximum clique
P(k-clustering≥ 1)	Loc	3-coloring
P(Minimum cycle cover≥ 1)	Loc	3-coloring
P(Minimum spanning tree≥ 1)	Loc	3-coloring
E(k-th longest m-nearest neighbor)	Loc	vertex cover
P(k-clustering≥ 1)	Exis	Independent Set

We will show the non-approximation of E(k-th longest m-nearest neighbor) in Locational model and P(k-clustering≥ 1) in Existential model in Section 4. The exact definition and brief proof for other problems will be shown in the appendix due to space constraints.

Related Work The uncertain or imprecise data has been studied extensively recently [7, 9]. Consider the locational data collected by the Global-Positioning Systems (GPS), there are always some random measurement errors[28]. For another example, if we use a sensor network to monitor the living habits or migration of certain animals, there will also be some noise among the data we collected as the sensors won't be perfect[23, 8, 17]. Some people study the imprecise data in a model where each point may be in some region[4, 24, 30, 27].

The existential uncertainty model and the locational uncertainty modes we mentioned before have been studied extensively in recent years(eg. [1, 17, 3, 2, 19]). It's worth mentioning that when all the points follow the same distribution, it's a classic topic in stochastic geometry literature [5, 28, 6]. The asymptotics expectation for certain combinatorial problems(such as MST) is the main interest in that topic. The general locational uncertain model is also of fundamental interest in the area of wireless networks. There is a survey[12] and you can see more references about the stochastic model and wireless networks there.

There have been many works under the term *stochastic geometry* in the above uncertainty model and many other different stochastic models. For example, Huang et al.[14] initiate the study of constructing ϵ-kernel coresets for uncertain points in the above two models. The convex hull[11, 21], minimum enclosing ball problem[25], shape fitting[20], MST[5] and many other problems have also been studied on the imprecise data.

The study of estimating the expectation of objects in the model is started by Kamousi, Chan and Suri[15, 16].They showed that the expectation of some values, such as nearest neighbor (NN) graph, the Gabriel graph (GG) and so on, can be solved in polynomial time. And they designed FPRAS for E(MST) and E(the closest pair) in the existential uncertainty model.

Huang et al.[13] gives the FPRAS for the expected values of closest pair, minimum spanning tree, k-clustering, minimum perfect matching, and minimum cycle cover in both models by several powerful techniques. And they also consider the problem of estimating the probability that the length of closest pairs at most, or at least, a given threshold.

Most recently, Li and Deshpandeb [18] observe that the expected value is inadequate in some problems and study the maximization of the expected utility of the solution for some given utility function. The initial motivation for the study is the stochastic shortest-path problem, which has been studied extensively[29, 22, 26].

2 The Expectation of Diameter

Existential Uncertainty Model: Let's show the FPRAS of E(D) in the Existential Uncertainty Model. First, let us show the meaning of the signals we use. Let U be the complete set of the points. And $S\langle \geq j\rangle$ means that there are at least j points realized in the set of points S. Suppose we have m points in total(we use m to describe the complexity of samples we need later). Then there are $l = \binom{m}{2}$ different pairs of points. W.l.o.g, suppose the lengths of the l pairs are distinct. And we sort them in ascending order of their length and index them. Let e_i represent the i-th pair, and d_i is its length. We have $d_1 < d_2 < ... < d_l$. And for a pair $e_i = (u, v)$, $P(e_i|\alpha)$ represent the probability that both u and v are realized conditioning on $event(\alpha)$.

What we want to estimate is indeed $E(D|U\langle \geq 2\rangle)$, because the diameter doesn't make sense if there is only one or zero point realized. Now let us introduce the algorithm.

First,for pair $e_i = (u, v)$, we can calculate $P(e_i|U\langle\geq 2\rangle) = \frac{P(e_i,U\langle\geq 2\rangle)}{P(U\langle\geq 2\rangle)} = \frac{P(e_i)}{P(U\langle\geq 2\rangle)} = \frac{P_u P_v}{P(U\langle\geq 2\rangle)}$, which can be calculated easily by the following lemma:

Lemma 2. *For a set of points C and $j \in \mathbf{Z}$, we can compute $P(C\langle\geq j\rangle)$ in polynomial time. Moreover, there exists a poly-time sampler to sample present points from C conditioning on $C\langle\geq j\rangle$ (Or $C\langle j\rangle$).*

Proof. The idea is essentially from [10]. W.l.o.g, we assume that the points in C are $x_1, x_2, ..., x_n$.We denote the event that among the first a points, at least b points are present by E(a,b) and denote the probability of E(a,b) by P(a, b). Note that our goal is to compute P(n,j), which can be solved by the following dynamic program:

1. If $a < b$, P(a,b)=0. If a=b, P(a,b)=$\prod_{1\leq l\leq a} P_l$. If b=0, P(a,b)=1.
2. For $a > b$ and $b > 0$, P(a,b)=$P_a P(a-1, b-1) + (1 - P_a)P(a-1, b)$.

We can also use this dynamic program to construct an efficient sampler. Consider the point x_n, with probability $P_n P(n-1, j-1)/P(n, j)$, we make it present and then recursively consider the point x_{n-1}, conditioning on the event E(n-1,j-1). With probability $(1 - P_n)P(n-1, j)/P(n, j)$, we discard it and then recursively sample conditioning on the event E(n-1,j).

The proof of $P(C\langle j\rangle)$(ie. there are exactly j points present in C) is similar and we skip it.

Now continue our algorithm.

There exists a set of pairs $S = \left\{e_i|P(e_i|U\langle\geq 2\rangle) \geq \frac{1}{m^2}\right\}$. S is non-empty, or $P(U\langle\geq 2\rangle|U\langle\geq 2\rangle) = 1 = P(\cup_i e_i|U\langle\geq 2\rangle) \leq lP(e_i|U\langle\geq 2\rangle) < 1$. Let Y be the largest index among all the pairs in S. Recall that the l-th pair is the longest one. If $d_Y \geq \frac{1}{lnm} d_l$, then $E(D|U\langle\geq 2\rangle) \geq P(D \geq d_Y|U\langle\geq 2\rangle)d_Y \geq \frac{1}{m^2 lnm} d_l$. By chernoff bound, we only need to take $O(\frac{m^2}{\epsilon^2} ln^2 m)$ independent samples. This is the worst case of our algorithm.

Now consider the other situation, i.e. $d_Y < \frac{1}{lnm} d_l$. Then we have a set of points $H = \{u|\exists v \in U, (u, v) \in S\}$. It's obvious that $\forall u, v \in H, d(u, v) < \frac{3}{lnm} d_n$ due to the Triangle inequality. As if $d(u, v) \geq \frac{3}{lnm} d_l$, suppose $(u, u') \in S, (v, v') \in S$. Let $u'' = u$ if $P_u > P_{u'}$, or $u'' = u'$. And get v'' by the similar approach. Then $P((u'', v'')|U\langle\geq 2\rangle) \geq \frac{1}{m^2}$ and $d(u'', v'') > \frac{1}{lnm} d_l$, which is impossible.(Remark: We can understand this property as those points with a relatively high probability of realization are surrounded by a small sphere.)

Suppose x is one of the points that have the largest probability of realization among U(if there are more than one points with largest probability, choose one arbitrarily), then we have the following property: $d(x, H) < \frac{2}{lnm} d_l$. The definition of $d(x, H)$ is $d(x, H) = \max_{u\in H} d(x, u)$. Let $H' = H \cup \{x\}$. And we can construct a set of points $H'' = \left\{u|u = x \text{ or } d(u, x) < \frac{4}{lnm} d_n\right\}$. It's obvious that $H \subseteq H' \subseteq H''$. If $H'' = U$, we need only $O(\frac{m^2}{\epsilon^2} ln^2 m)$ independent samples. Else, we can use the following algorithm.

For any point t, let $P(\alpha|t)$ represent the probability of $event(\alpha)$ conditioning on that point t is realized, and $P(\alpha|\bar{t})$ correspond to the probability of $event(\alpha)$

Algorithm 1 construct event

```
$S_0 = U/H'', N_0 = \emptyset$,i=0
while $S_i$ is not empty do
    $t_i = \arg\max_{u \in S_i} d(u, H')$
    $S_{i+1} \leftarrow S_i/\{t_i\}$
    $N_{i+1} \leftarrow N_i \cup \{t_i\}$
    i←i+1
Output:$S_i, t_i$ and $N_i$ for all i
```

conditioning on that t is not realized. $P(t|\alpha)$ represents the probability that t is realized conditioning on $event(\alpha)$. Let $|S_0|$ denote the size of S_0. Then we have that $E(D|U\langle\geq 2\rangle) = \sum_{i=0}^{|S_0|-1} E(D|U\langle\geq 2\rangle, t_i, N_{i-1}\langle 0\rangle)P(t_i, N_{i-1}\langle 0\rangle|U\langle\geq 2\rangle) + E(D|N_{|S_0|-1}\langle 0\rangle, H''\langle\geq 2\rangle)P(H''\langle\geq 2\rangle, N_{|S_0|-1} < 0 > |U\langle\geq 2\rangle)$.

All of the probability can be calculated easily. What we need to get is the expected value in each part. $E(D|N_{|S_0|-1}\langle 0\rangle, H''\langle\geq 2\rangle)$ can be seen as we recurse our original problem into a smaller problem. Then for each i, we have the following lemma:

Lemma 3. *We only need $O(\frac{m}{\epsilon^2}logm)$ independent samples to estimate $E(D|U <\geq 2 >, t_i, N_{i-1} < 0 >)$.*

Proof. Let $U_i = U/N_i = U/(\{t_i\} \cup N_{i-1})$. We can rewrite $E(D|U\langle\geq 2\rangle, t_i, N_{i-1}\langle 0\rangle) = E(D|t_i, N_{i-1}\langle 0\rangle, U_i\langle\geq 1\rangle)$. We have a sampler condition on $event(t_i, N_{i-1} < 0 >, U_i <\geq 1 >)$ according to lemma 1. And recall that x is the point that has the largest probability of realization, it's obvious that x$\in U_i$. Then $P(x|U_i\langle\geq 1\rangle) \geq 1/m$. Let D_i represent the maximum value of Diameter condition on $event(t_i, N_{i-1}\langle 0\rangle, U_i\langle\geq 1\rangle)$. We have $d(t_i, x) \geq \frac{2*d(t_i,H')}{6} \geq D_i/6$ condition on $event(t_i, N_{i-1}\langle 0\rangle, U_i\langle\geq 1\rangle)$. Then $E(D|t_i, N_{i-1}\langle 0\rangle, U_i\langle\geq 1\rangle) \geq d(t_i, x)P(x|U_i\langle\geq 1\rangle) \geq d(t_i, x)/m \geq D_i/(6m)$. And by Chernoff bound, we proved this lemma.

Let T(m) represent the independent samples we need to estimate $E(D|U <\geq 2 >)$ with $|U| = m$. We have the following recursive relation in the best case: $T(m) = T(m - |S_0|) + O(|S_0| * \frac{m}{\epsilon^2}lnm)$. Then we have $T(m) = O(\frac{m^2}{\epsilon^2}lnm)$.

Locational Uncertainty Model: Our algorithm is almost the same as the existential model with the assumption that at for each point, there is only one node that may be realized at this point. In principle, if more than one node may be realized at the same point, we can create multiple copies of the point co-located at the same place. We can't use the Monte Carlo method directly only when all points with high probability to be realized are 'wrapped in a small ball', we can use the similar algorithm like **Algorithm 1** to do the recursion and get the same required complexity as the existential model. For example, we can run the while loop only once and get a sub-problem with size $m - 1$.

Theorem 1. *There is an FPRAS for estimating the expected distance between the longest pair of nodes both existential and locational uncertainty models. It* *needs $O(\frac{m^2}{\epsilon^2}ln^2m)$ independent samples in the worst case and $O(\frac{m^2}{\epsilon^2}lnm)$ in the best case in both models.*

3 The Tail Bound of Distribution

Existential Uncertainty Model: Now let us introduce the FPRAS of P(D≥1). We can construct a set of points $H' = \left\{u|P_u \geq \frac{\epsilon}{m}\right\}$, and $H = U/H'$. We have that $P(D \geq 1) = \sum_{i=0}^{|H|} P(D \geq 1, H\langle i\rangle)$. We'll show that we only need to esimate $\sum_{i=0}^{2} P(D \geq 1, H\langle i\rangle)$ as the remaining is negligible.

Call a set of points S connected if $\forall u \in S, P_u \geq \frac{\epsilon}{m} \wedge \exists v \in S, d(u,v) \geq 1$. Call a point u$\in$ S is unique in S, if let $S' = S/\left\{u\right\}$, S' is not connected. Let $C = \{u|u \in H' \wedge \exists v \in H', d(u,v) \geq 1\}$. It's obvious that the C we constructed is connected.

Lemma 4. *For connected non-empty set S with all points unique, $P(D \geq 1|S\langle\geq 2\rangle) \geq \frac{\epsilon}{2m}$.*

Proof. It's obvious that S must have even points according to the definition. Call pair (u,v) a match if d(u,v)≥1. Suppose S has 2k points, then S has exactly k matches. Index the 2k points subject to that u_i and u_{k+i} is a match and $P_{u_i} \geq P_{u_{i+k}}$. Let $S_{a,b}$ denote the subset of points with index in [a,b]. Then we have $P(D \geq 1|S\langle\geq 2\rangle) = \sum_{i=1}^{k} P(D \geq 1|u_i, S_{1,i-1}\langle 0\rangle, S_{i+1,2k}\langle\geq 1\rangle)P(u_i, S_{1,i-1}\langle 0\rangle, S_{i+1,2k}\langle\geq 1\rangle|S\langle\geq 2\rangle)$.

When $i \leq k, P(D \geq 1|u_i, S_{1,i-1}\langle 0\rangle, S_{i+1,2k}\langle\geq 1\rangle) \geq \frac{\epsilon}{m}$, and we can get $P(u_i, S_{1,i-1}\langle 0\rangle, S_{i+1,2k}\langle\geq 1\rangle|S\langle\geq 2\rangle) \geq P(u_{i+k}, S_{1,i+k-1}\langle 0\rangle, S_{i+k+1,2k}\langle\geq 1\rangle)$. And notice $\sum_{i=1}^{2k-1} P(u_i, S_{1,i-1} < 0 >, S_{i+1,2k}\langle\geq 1\rangle|S\langle\geq 2\rangle) = 1$. Thus we proved this lemma.

Lemma 5. *We can estimate $P(D \geq 1, H\langle 0\rangle)$ with $O(\frac{m}{\epsilon^4}lnm)$ independent samples.*

Proof. Then in order to prove this lemma, we only need to show that $P(D \geq 1|S\langle\geq 2\rangle) \geq \frac{\epsilon}{2m}$ for any non-empty connected set S by Mathematical induction.

Now prove lemma 5. When $|S| = 2$, then $P(D \geq 1|S\langle\geq 2\rangle) = 1 \geq \frac{\epsilon}{2m}$. Suppose $P(D \geq 1|S\langle\geq 2\rangle) \geq \frac{\epsilon}{2m}$ when $|S| \leq n$ for any connected S and some integer n, then consider the situation when $|S| = n+1$. When all points in C are unique, then we have $P(D \geq 1|S\langle\geq 2\rangle) \geq \frac{\epsilon}{2m}$ by Lemma 4. If there exists some point u that are not unique, we have $P(D \geq 1|S\langle\geq 2\rangle) = P(D \geq 1|u, S\langle\geq 1\rangle)P(u|S\langle\geq 2\rangle) + P(D \geq 1|\overline{u}, S\langle\geq 2\rangle)P(\overline{u}|S\langle\geq 2\rangle)$. Both $P(D \geq 1|u, S\langle\geq 1\rangle)$ and $P(D \geq 1|\overline{u}, S\langle\geq 2\rangle)$are no less than $\frac{\epsilon}{2m}$. And $P(u|S\langle\geq 2\rangle) + P(\overline{u}|S\langle\geq 2\rangle) = 1$, thus we proved $P(D \geq 1|S\langle\geq 2\rangle) \geq \frac{\epsilon}{2m}$ when $|S|$=n+1.

Then by Monte carlo directly, we proved this lemma.

Now let's show how to estimate $P(D \geq 1, H\langle 1\rangle)$. Observe that $P(D \geq 1, H\langle 1\rangle) = \sum_{u\in H} P(D \geq 1, u, H/\left\{u\right\}\langle 0\rangle)$. For point u in H, denote $G_u = \{v|v \in H', d(u,v) \geq 1\}$. We can calculate $P(G_u\langle\geq 1\rangle, u, H/\left\{u\right\}\langle 0\rangle)$ exactly in linear time. We can use the value of $\sum_{u\in H} P(G_u\langle\geq 1\rangle, u, H/\left\{u\right\}\langle 0\rangle)$ as an estimation of $P(D \geq 1, H\langle 1\rangle)$ because of the following claim.

Claim. $\sum_{u\in H} P(G_u\langle\geq 1\rangle, u, H/\{u\}\langle 0\rangle) \leq P(D\geq 1, H\langle 1\rangle) \leq \sum_{u\in H} P(G_u\langle\geq 1\rangle, u, H/\{u\}\langle 0\rangle) + 2\epsilon P(D\geq 1, H\langle 0\rangle)$.

Proof. Since we only miss the summation probability of these events: there are two points x,y in H'/G_u realized with d(x,y)≥1 and there are no points present in G_u. Write down the expression: $\sum_{u\in H} P(D\geq 1, u, H/\{u\}\langle 0\rangle, G_u\langle 0\rangle)$. Denote the set of realization we may miss by M. Each realization $\mathbf{r}$ in M can be transferred to the $event(D\geq 1, H\langle 0\rangle)$ by making the only present point in H absent. We denote the realization after the transform $\mathbf{r'}$. We have $P(\mathbf{r'})\geq \frac{m}{2\epsilon}P(\mathbf{r})$. And given $\mathbf{r'}$, there are at most m different realizations can be transformed into it. We have $P(D\geq 1, H\langle 1\rangle) = \sum_{u\in H} P(G_u\langle\geq 1\rangle, u, H/\{u\}\langle 0\rangle) + \sum_{\mathbf{r}\in M} P(\mathbf{r}) \leq \sum_{u\in H} P(G_u\langle\geq 1\rangle, u, H/\{u\}\langle 0\rangle) + 2\epsilon\sum_{\mathbf{r'}} P(\mathbf{r'}) \leq \sum_{u\in H} P(G_u\langle\geq 1\rangle, u, H/\{u\}\langle 0\rangle) + 2\epsilon P(D\geq 1, H\langle 0\rangle)$.

Call the argument method of this claim **ARG**, which will be useful later.

As for $P(D\geq 1, H\langle 2\rangle) = \sum_{u,v\in H} P(D\geq 1, u, v, H/\{u,v\}\langle 0\rangle)$. Given u,v∈ H. If d(u,v)≥1, $P(D\geq 1, u, v, H/\{u,v\}\langle 0\rangle) = P_u P_v P(H/\{u,v\}\langle 0\rangle)$ which can be calculated directly. And $\sum_{u,v\in H\wedge d(u,v)<1} P(D\geq 1, u, v, H/\{u,v\}\langle 0\rangle) \leq 2\epsilon(P(D\geq 1, H\langle 0\rangle) + P(D\geq 1, H\langle 1\rangle))$ by the similar argument of **ARG**, which means it's negligible. So we can use the value of $\sum_{u,v\in H\wedge d(u,v)\geq 1} P(D\geq 1, u, v, H/\{u,v\}\langle 0\rangle)$ as an estimation of $P(D\geq 1, H\langle 2\rangle)$.

So far we have shown how to estimate $\sum_{i=0}^{2} P(D\geq 1, H\langle i\rangle)$. The last thing we have to do is to show $\sum_{i=3}^{|H|} P(D\geq 1, H\langle i\rangle)$ is negligible. In fact, we can prove $P(D\geq 1, H\langle 2+i\rangle) \leq (2\epsilon)^i P(D\geq 1, H\langle 2\rangle)$ with the similar method with **ARG**. So $\sum_{i\geq 3} P(D\geq 1, H\langle i\rangle) \leq$ is negligible compared with $\sum_{i=0}^{2} P(D\geq 1, H\langle i\rangle)$.

Theorem 2. *There is an FPRAS for estimating the probability of the distance between the furthest pair of nodes is at least 1 in the existential uncertainty model with only $O(\frac{m}{\epsilon^4} lnm)$ independent samples.*

Locational Uncertainty Model: Please pay attention that the *node* and *point* have different meaning in the locational model. And recall that we assume that at for each point, there is only one node that may be realized at this point. Suppose we have n nodes and m points. Huang et al.[13] has given a FPRAS for $P(D\geq 1)$ which needs $O(\frac{m^6}{\epsilon^4} lnm)$ independent samples. And we improved it and only need $O(\frac{m^3}{\epsilon^4} lnm)$ independent samples.

The thought of FPRAS for $P(D\geq 1)$ in locational uncertainty model is exactly the same as the existential model, while we need a little bit more samples because of the difference of the two models.

Call a point u *not-alone* in a point set H, if $\exists v\in H$, st. $d(u,v)\geq 1\wedge$ u,v correspond to different nodes. And we call the set H single if H doesn't contain any not-alone points.

Let $H = \{u | P_u \geq \frac{\epsilon}{m^2}\}$. $F = V/H$. So similarly, $P(D\geq 1) = \sum_{i=0}^{|F|} P(D\geq 1, F\langle i\rangle)$. And we also only need to estimate $\sum_{i=0}^{2} P(D\geq 1, F\langle i\rangle)$ as $\sum_{i=3}^{|F|} P(D\geq 1, F\langle i\rangle)$ is negligible by the similar argument with **ARG**.

Lemma 6. *We can estimate $P(D \geq 1, F\langle 0\rangle)$ by $O(\frac{m^3}{\epsilon^4} lnm)$ independent samples.*

Proof. 1. It's obvious that if H is single, then $P(D \geq 1, F\langle 0\rangle) = 0$.

2. If H is not single, with the following algorithm 2, we can estimate $P(D \geq 1, F\langle 0\rangle)$ by $O(\frac{m^3}{\epsilon^4} lnm)$ independent samples.

Algorithm 2 Estimate $P(D \geq 1, F\langle 0\rangle)$

```
$S_0 = H, N_0 = \emptyset$,i=0
while $S_i$ not-single do
    find arbitrary not-alone point $t_i$
    $S_{i+1} \leftarrow S_i / \{t_i\}$
    $N_{i+1} \leftarrow N_i \cup \{t_i\}$
    i←i+1
    Estimate $P(D \geq 1 | t_i, N_i\langle 0\rangle, F\langle 0\rangle)$
Output:The summation of $P(D \geq 1 | t_i, N_i\langle 0\rangle, F\langle 0\rangle)$ for all i
```

Note that we can estimate $P(D \geq 1 | t_i, N_i\langle 0\rangle, F\langle 0\rangle)$ with $O(\frac{m^2}{\epsilon^4} lnm)$ for any given i. And $i \leq m$, thus we finish the proof.

As for the term $P(D \geq 1 | F\langle 1\rangle)$. For point $u \in F$ corresponds to node n_i, then we can either estimate $P(D \geq 1 | u, F / \{u\} \langle 0\rangle)$ with $O(\frac{m^2}{\epsilon^4} lnm)$ independent samples if there are $d(u, v) \geq 1$ for $v \in H \wedge v$ corresponds to a different node n_j, or this value can be neglected by the similar argument with **ARG**, as there will must be a point $u' \in H$ which also corresponds to n_i st. $P(u') \geq \frac{1}{m}$. Thus we can estimate $P(D \geq 1, F\langle 1\rangle)$ with $O(\frac{m^3}{\epsilon^4} lnm)$ samples.

Similarly, we can estimate $P(D \geq 1 | F\langle 2\rangle)$ by enumerating point pairs $(u, v) \in H$, and let $\sum_{u,v \in H, d(u,v) \geq 1} P(u, v)$ be the estimation of $P(D \geq 1 | F\langle 2\rangle)$. There are at most $O(m^2)$ pairs.

Theorem 3. *There is an FPRAS for estimating the probability of the distance between the furthest pair of nodes is at least 1 in the Locational Uncertainty Model with only $O(\frac{m^3}{\epsilon^4} lnm)$ independent samples.*

4 Examples for Unapproximable Values:

k-th Longest m-Nearest Neighbor: The precise description of this problem is under any realization, for each node, find the distance to its m-nearest neighbor, then compute the k-th longest one among these distances. Huang et al.[13] gives a FPRAS for this value in the existential model. And we'll show that this value can't be approximated in the locational uncertainty model unless $NP \subseteq BPP$.

Lemma 7. *Given the undirect graph G, we can construct a Locational Uncertainty Model G'. Then there is a vertex cover of size k iff E((n-k)th Longest (n+m-1)Nearest Neighbor)>0 in G'.*

Proof. Suppose for one point, there may be more than one node realized at it. Now let's show how to construct such an G' according to G. Suppose there are n vertices and m edges in G. Construct n points and n+m nodes in G'. Divide the n+m nodes into two disjoint sets S_1 and S_2, with $|S_1| = n, |S_2| = m$. The i-th node in S_1 can only be present at i-th point with probability equals 1. The n points in G' correspond to the n vertices in G each, the m nodes in S_2 correspond to the m edges in G. Then if vertex v_j is one of the end point of edge e_i in G, the corresponding node can be present at the corresponding point with probability 1/2 in G'. As for the distance of point pairs in G', the distance of each pair is M, which can even be $+\infty$.

Under such a construction, it's obvious that E((n-k)th Longest (n+m-1)Nearest Neighbor)=M*p with $p > 0$ strictly if and only if there exists a vertex cover with size k. Note that when $p > 0$, E((n-k)th Longest (n+m-1)Nearest Neighbor) can be infinitely large

Then if there exists a FPRAS or any other approximation algorithms for E(k-th Longest m-Nearest Neighbor) with finite approximation ratio and guaranteed accuracy, we can construct a Locational Uncertainty Model G' according to *Lemma 7*. Run the algorithm on G', and we can judge if there is a k-vertex cover in G by comparing the output of the algorithm with zero, and get the accurate result with guaranteed accuracy, which means $NP \subseteq BPP$.

Theorem 4. *E(k-th Longest m-Nearest Neighbor) in Locational Uncertainty Model is imapproximable within any finite ratio and guaranteed accuracy unless NP⊆BPP.*

k-clustering problem: Not only in locational uncertainty model, the similar thought can also be used in the existential uncertainty model. In the deterministic kclustering 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 d(u, v) with u, v in different subsets. We have the following lemma:

Lemma 8. *Given an undirect graph G, we can construct a Existential Uncertainty Model G, subject to there exists an independent set of k vertices G iff $P(k - clustering \geq 1) > 0$ in G.*

Proof. Suppose there are n vertices in G, then there will also be n nodes in G'. And there is a bijection between them. Each node will be present with probability 1/2. As for the distance of nodes in G', for pair (n_i, n_j) in G', if there is an edge between the corresponding vertices in G, then $d(n_i, n_j)$=0.9, else $d(n_i, n_j)$=1.8.

Then if there is an independent set of size k in G, the output of the approximation algorithms for the $P(k - clustering \geq 1)$ in G' should be more than 0 strictly with guaranteed accuracy, or the approximation ratio will be ∞.

Theorem 5. *$P(k - Clustering \geq 1)$ in Existential Uncertainty Model is imapproximable within any finite ratio and guaranteed accuracy unless NP⊆BPP.*

5 Conclusion

In this paper, we studied the expectation and the tail bound of distribution of stochastic diameter, and prove some values can't be approximated. One remaining open problem is if there is FPRAS for k-Clustering problem and kth Closest Pair problem, or they are also imapproximable. And studying the threshold probabilities P(Obj$\geq$ 1) and P(Obj$\leq$ 1) for other values is also an interesting topic.

6 Acknowledgements

The author would like to thank Jian Li for several useful discussions and the help with polishing the paper. The research is supported in part by the National Basic Research Program of ChinaGrant2015CB358700, the National Natural Science Foundation of ChinaGrant61822203, 61772297, 61632016, 61761146003.

References

1. Pankaj Agarwal, Sariel Har-Peled, Subhash Suri, Hakan Yildiz, and Wuzhou Zhang. Convex hulls under uncertainty. In *European Symposia on Algorithms*, 2014.
2. Pankaj K Agarwal, Siu-Wing Cheng, and Ke Yi. Range searching on uncertain data. *ACM Transactions on Algorithms (TALG)*, 8(4):43, 2012.
3. M.J. Atallah, Y. Qi, and H. Yuan. Asymptotically efficient algorithms for skyline probabilities of uncertain data. *ACM Trans. Datab. Syst*, 32(2):12, 2011.
4. D. Bandyopadhyay and J. Snoeyink. Almost-Delaunay simplices: Nearest neighbor relations for imprecise points. In *Proceedings of the 15th ACM-SIAM Symposium on Discrete Algorithms*, pages 410–419, 2004.
5. J. Beardwood, J. H. Halton, and J. M. Hammersley. The shortest path through many points. In *Proc. Cambridge Philos. Soc*, pages 55:299–327, 1959.
6. D.J. Bertsimas and G. van Ryzin. An asymptotic determination of the minimum spanning tree and minimum matching constants in geometrical probability. *Operations Research Letters*, 9(4):223–231, 1990.
7. Reynold Cheng, Jinchuan Chen, and Xike Xie. Cleaning uncertain data with quality guarantees. *Proceedings of the VLDB Endowment*, 1(1):722–735, 2008.
8. A. Czumaj, F. Ergün, L. Fortnow, A. Magen, I. Newman, R. Rubinfeld, and C. Sohler. Approximating the weight of the euclidean minimum spanning tree in sublinear time. *SIAM Journal on Computing*, 35(1):91–109, 2005.
9. Xin Dong, Alon Y Halevy, and Cong Yu. Data integration with uncertainty. In *Proceedings of the 33rd international conference on Very large data bases*, pages 687–698. VLDB Endowment, 2007.
10. Martin Dyer. Approximate counting by dynamic programming. In *ACM Symp. on Theory of Computing*, pages 693–699, 2003.

11. W. Evans and J. Sember. The possible hull of imprecise points. In *Proceedings of the 23rd Canadian Conference on Computational Geometry*, 2011.
12. M. Haenggi, J.G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti. 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.
13. Lingxiao Huang and Jian Li. Approximating the expected values for combinatorial optimization problems over stochastic points. In *The 42nd International Colloquium on Automata, Languages, and Programming*, pages 910–921. Springer, 2015.
14. Lingxiao Huang, Jian Li, Jeff M Phillips, and Haitao Wang. *epsilon*-kernel coresets for stochastic points. *arXiv preprint arXiv:1411.0194*, 2014.
15. P. Kamousi, T.M. Chan, and S. Suri. Stochastic minimum spanning trees in euclidean spaces. In *Proceedings of the 27th annual ACM symposium on Computational Geometry*, pages 65–74. ACM, 2011.
16. P. Kamousi, T.M. Chan, and S. Suri. Closest pair and the post office problem for stochastic points. *Computational Geometry*, 47(2):214–223, 2014.
17. J. Li and A. Deshpande. Ranking continuous probabilistic datasets. *Proceedings of the VLDB Endowment*, 3(1-2):638–649, 2010.
18. Jian Li and Amol Deshpande. Maximizing expected utility for stochastic combinatorial optimization problems. *Mathematics of Operations Research*, 2018.
19. Jian Li, Jeff M Phillips, and Haitao Wang. ϵ-kernel coresets for stochastic points. *arXiv preprint arXiv:1411.0194*, 2014.
20. M. Löffler and J. Phillips. Shape fitting on point sets with probability distributions. *European Symposia on Algorithms*, pages 313–324, 2009.
21. M. Löffler and M. van Kreveld. Approximating largest convex hulls for imprecise points. *Journal of Discrete Algorithms*, 6:583–594, 2008.
22. Ronald Prescott Loui. Optimal paths in graphs with stochastic or multidimensional weights. *Communications of the ACM*, 26(9):670–676, 1983.
23. Alan Mainwaring, David Culler, Joseph Polastre, Robert Szewczyk, and John Anderson. Wireless sensor networks for habitat monitoring. In *Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications*, pages 88–97. ACM, 2002.
24. J. Matoušek. Computing the center of planar point sets. *Discrete and Computational Geometry*, 6:221, 1991.
25. A. Munteanu, C. Sohler, and D. Feldman. Smallest enclosing ball for probabilistic data. In *Proceedings of the 30th Annual Symposium on Computational Geometry*, 2014.
26. Evdokia Nikolova, Matthew Brand, and David R Karger. Optimal route planning under uncertainty. In *ICAPS*, volume 6, pages 131–141, 2006.
27. Y. Ostrovsky-Berman and L. Joskowicz. Uncertainty envelopes. In *Abstracts of the 21st European Workshop on Comput. Geom.*, pages 175–178, 2005.
28. Dieter Pfoser and Christian S Jensen. Capturing the uncertainty of moving-object representations. In *Advances in Spatial Databases*, pages 111–131. Springer, 1999.
29. C Elliott Sigal, A Alan B Pritsker, and James J Solberg. The stochastic shortest route problem. *Operations Research*, 28(5):1122–1129, 1980.
30. M. van Kreveld and M. Löffler. Largest bounding box, smallest diameter, and related problems on imprecise points. *Computational Geometry: Theory and Applications*, 43:419–433, 2010.

A Appendix

A.1 k-th closest pair:

k-th closest pair means for all pairs of nodes, find the k-th closest one among them. We know that Max 2-SAT is a NP-Complete problem. Given a 2CNF with n clauses and a integer k ¡ n, we would like to ask whether there is an assignment such that at least k clauses are satisfied. Let P(kC≤1) represent P(k-th Closest Pair≤1). We will show that

Lemma 9. *Given the 2CNF and the integer k, we can construct a Locational Uncertainty Model G. Then there is an assignment such that at least k clauses are satisfied iff $P(kC \leq 1) > 0$ in G.*

Proof. Suppose there are n clauses and m variables in the 2-CNF. And there is no clause containing both variable x_i and $\overline{x_i}$ for some i. Corresponding to each variable x_i, there are one node u_i and two possible points A_i and B_i for realization of u_i. We have $P_{u_iA_i} = P_{u_iB_i} = \frac{1}{2}$. Then $P_{u_iA_j} = P_{u_iB_j} = 0$ for $i \neq j$.

Then for each clause c_i, there will be one node v_i and two possible points C_i and D_i. We also have $P_{v_iC_i} = P_{v_iD_i} = \frac{1}{2}$.

We can set a bijection that $x_i = true$ iff u_i is realized to A_i. Then $\overline{x_i} = true$ iff u_i is realized to B_i.

Then we should give the distance of the pairs of points. We let the distance of any pairs of points be 1.8 for initialization. For the clause $c_i = [x_t \cup x_s]$. The distance of two pairs (C_i, A_t) and (D_i, A_s) should be changed to 0.9.

To see that even if both x_t and x_s are true, the clause $c_i = [x_t \cup x_s]$ can only contributes one pair with distance ≤ 1 in one possible realization. And for another example, if $c_i = [x_t \cup \overline{x_s}]$, we can let the distance of (C_i, A_t) and (D_i, B_s) to be 0.9.

And what we should pay attention is that even if we have two same clauses c_i and c_j, we still need to change the distance of four different pairs of points in G be 0.9, each clause corresponds to two pairs.

Then if there is an assignment such that at least k clauses are satisfied, there will be a realization that each node is realized in the corresponding point according to assignment and bijection. And there will be one possible realization that the k nodes corresponding to the k satisfied clauses are realized in the points whose closest pair =0.9. Then $P(kC \leq 1) > 0$. And the reversal direction is similar.

Having proved this lemma, we can have the theorem below:

Theorem 6. *P(kC≤ 1) in Locational Uncertainty Model is imapproximable within any finite ratio and guaranteed accuracy unless NP⊆BPP.*

A.2 kth Longest m-Nearest Neighbor

Lemma 10. *Given the undirect graph G, we can construct a Locational Uncertainty Model G'. Then there is a clique of size k iff P(longest k-1 nearest neighbor$\leq 1) > 0$ in G'.*

Proof. Let k be the size of clique we want to find. And there are n vertices in G. We will have k nodes in G', denoted by $\{x_1, ..., x_k\}$. And we have k family of points $S_1, ..., S_k$. Each family S_i has n points. And the node x_i can be only realized at the n points in S_i with random probability, ie. $P_{x_i A_j} = \frac{1}{n}$ for point $A_j \in S_i$.

As for the distance of pairs of nodes. For pair (u,v) when u and v are in the same family, let d(u,v)=0.9.(In fact the distance of pair in the same family is not important, as there will be only one node realized in the same family). We want each vertex u in G corresponds to k points in G', and the k points are separated in the k disjoint family. Then there will be a bijection between the n points in one family and the n vertices in G. Then consider pair (u_i, v_j) with u_i in S_i, v_j in S_j and $i \neq j$. Denote the corresponding vertex u of u_i and v of v_j in G, if (u,v) is an edge in G, then let $d(u_i, v_j) = 0.9$, else $d(u_i, v_j) = 1.8$.(Remark: Even if u==v in G, $d(u_i, v_j) = 1.8$). Then all the pairs will have a distance and will meet the triangle inequality.

Theorem 7. *P(kmNN≤ 1) in Locational Uncertainty Model is imapproximable within any finite ratio and guaranteed accuracy unless NP⊆BPP.*

A.3 K-clustering:

We have shown that $P(k-clustering \geq 1)$ is hard to approximate in Existential Uncertainty Model, now we show it's also unapproximated in Locational Model. Note $P(k-clustering \geq 1)$ by $P(kCL \geq 1)$ later.

Lemma 11. *Given an undirect graph G, we can construct a Lacational Uncertainty Model G', subject to G is 3-colorable iff P(kCL≥ 1) > 0 in G'.*

Proof. Suppose there a n vertices in G. We can construct G' with n nodes, and there is a bijection between these n vertices and n nodes. We have 3 family of points, noted by S_1, S_2, S_3. And each family contains n points, where there also is a bijection between n vertices in G and n points in S_i for all i$\in \{1, 2, 3\}$.

For each vertex x_i in G, it has bijection relationships with node u_i and three points A_i, B_i, C_i, where A_i, B_i, C_i are in the three different family. Then we let u_i can only be realized in A_i, B_i, C_i, with probability $\frac{1}{3}$ each.

As for the distance of pairs of points. For pair (u,v) with u and v are in different family, let d(u,v)=1.8. For pair (u_i, u_t) in the same family, let x_i has the bijection relation with u_i and x_t for u_t. If there is an edge (x_i, x_t) in G, then $d(u_i, u_t)$=0.9, else $d(u_i, u_t) = 1.8$.

Theorem 8. *P(kCL≥1) in Locational Uncertainty Model is imapproximable within any finite ratio and guaranteed accuracy unless NP⊆BPP.*

A.4 Minimum Cycle Cover and MST Problem:

In the deterministic version of the cycle cover problem, we are asked to find a collection of node-disjoint cycles such that each node is in one cycle and the total length is minimized. Here we assume that each cycle contains at least two nodes. If a cycle contains exactly two nodes, the length of the cycle is two times the distance between these two nodes. And we still starts from 3-coloring problem to show that P(Minimum Cycle Cover ≥ 1) is imapproximable. We denote Minimum Cycle Cover by MCC below.

With the same construction in A.3, we have following lemmas and theorems:

Lemma 12. *Given an undirect graph G, we can construct a Lacational Uncertainty Model G', subject to G is 3-colorable iff P(MCC$\geq 1.8n) > 0$ in G'.*

Lemma 13. *Given an undirect graph G, we can construct a Lacational Uncertainty Model G', subject to G is 3-colorable iff P(MST$\geq 1.8n) > 0$ in G'*

With this lemma, we can have the following theorem:

Theorem 9. *P(MCC$\geq$1) and P(MST≥ 1) in Locational Uncertainty Model are imapproximable within any finite ratio and guaranteed accuracy unless NP$\subseteq$BPP.*