Geometric Computing over Uncertain Data

Subhash Suri

Computer Science
University of California, Santa Barbara

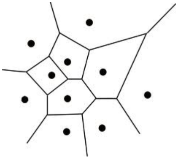
Algo 2012

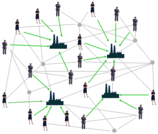
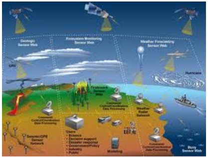
Geometric Computing

- Reasoning about points, lines, polygons, hyperplanes, balls.
- Geometric abstractions, combinatorial algorithms, data structures.

Geometric Computing

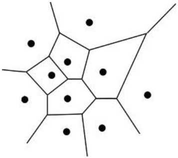
- Reasoning about points, lines, polygons, hyperplanes, balls.
- Geometric abstractions, combinatorial algorithms, data structures.
  - Nearest neighbors, intersections, shortest paths.
  - Voronoi diagram, Delaunay triangulation, search structures.
  - Sensor networks, bio-informatics, spatial DB, vision, robotics.
  - Wonderful algorithms and data structures.

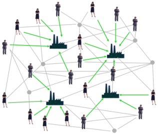
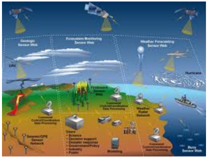


Geometric Computing

- Reasoning about points, lines, polygons, hyperplanes, balls.
- Geometric abstractions, combinatorial algorithms, data structures.
  - Nearest neighbors, intersections, shortest paths.
  - Voronoi diagram, Delaunay triangulation, search structures.
  - Sensor networks, bio-informatics, spatial DB, vision, robotics.
  - Wonderful algorithms and data structures.

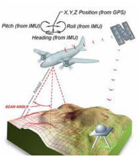


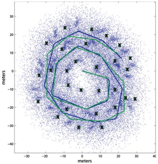



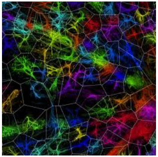
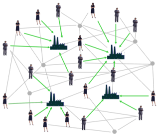
- But typically assume precise, noiseless input data.

Geometric Computing and Uncertainty

- What can we compute when underlying data is uncertain?
- Diverse causes of uncertainty.
  - Positional measurements are inherently noisy (sensing errors).
  - Privacy: many location services deliberately add random noise.
  - Incomplete information: avian flu, sensor awake.
  - Stochastic modeling: customers for a new service, facility.





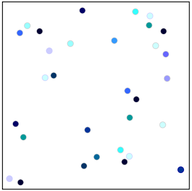



- Complexity of basic geometric questions under imperfect knowledge.
- Preliminary work. More questions than answers. (SoCG, WADS)

Uncertain Point Data: A simple model

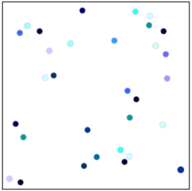
Uncertain Point Data: A simple model

- Uncertainty: each point s_i active with independent prob. p_i.
  - Prob. that node i has flu, is a client, is active sensor.
  - Darker color indicates higher probability.
  - What can we say about the geometric structure of this stochastic set?



Uncertain Point Data: A simple model

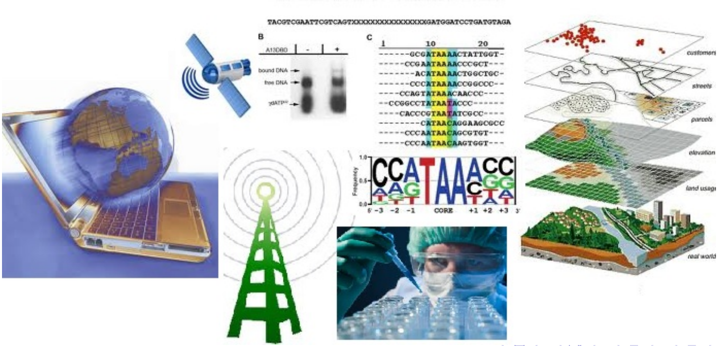
- Uncertainty: each point s_i active with independent prob. p_i.
  - Prob. that node i has flu, is a client, is active sensor.
  - Darker color indicates higher probability.
  - What can we say about the geometric structure of this stochastic set?



  - Length of the expected MST or TSP?
  - Size of the expected Convex Hull?
  - Expected distance between the Closest Pair?
  - Similar questions for positional uncertainty.

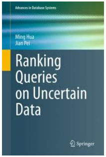
Data-Driven Science

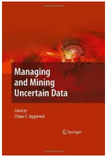
- Age of inexpensive, ubiquitous sensing and Big Data.
  - Scanners (3D, LiDAR, medical, satellites), Biology, GPS, social graphs
- Enables modeling of complex phenomena (ecology, biology, social).
- But invariably, these data are "ambiguous":
  - Noisy, inaccurate, approximate, incomplete

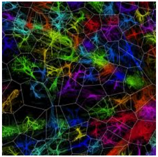


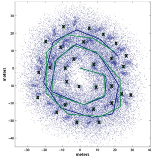
Computing with Uncertain Data

- Many computer science areas are focussed on uncertainty:
  - Databases, Data mining
  - Machine Learning
  - Computer Vision, Sensor Networks, Optimization etc.









- Design of *uncertainty-aware* geometric algorithms?
  - Gracefully cope with uncertainty of input.

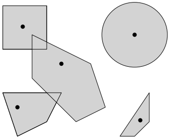
Related Work: Geometry

Related Work: Geometry

- Classical "stochastic geometry:" limit theorems [BHH, F, S]
  - Expected length for n random points etc.
  - Computational complexity and worst-case distributions.

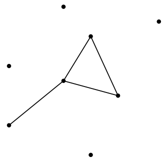
Related Work: Geometry

- Classical "stochastic geometry:" limit theorems [BHH, F, S]
  - Expected length for n random points etc.
  - Computational complexity and worst-case distributions.
- Imprecise Points [Loffler-van Kreveld]
  - Each point can be anywhere inside a simple region
  - Max or Min measures (bounding box, diameter, convex hull, etc)
  - Different point positions give different answers
  - Analysis of robustness, sensitivity, finite precision



Related Work: Optimization

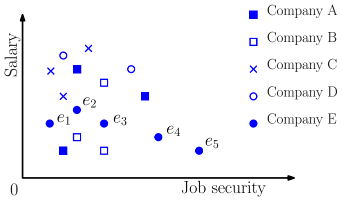
- 2 Stage Optimization (Erdös' Random Race)
  - Planning under uncertainty: Network Design.
  - Cheaper to buy in stage 1, but future demand unknown
  - Demand becomes known in stage 2, but more expensive to buy



- A priori optimization [Bertsimas, Jaillet].

Related Work: Databases

- Alternative Worlds
  - Incomplete information
  - Probability distribution over values
  - Few (discrete) possible values for each datum

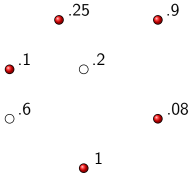


- Example problems.
  - Ranking, Top-k, Indexing, Range Searching
  - Clustering, Skyline (maxima), etc.

Uncertain Minimum Spanning Tree

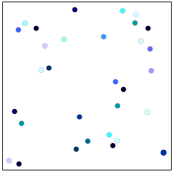
Uncertain Minimum Spanning Tree

- A *master* set $M = \{s_1, s_2, \ldots, s_n\}$ of points in d dimensions.
- Each s_i is *active* with an independent probability p_i.
- What is the expected MST length of M?



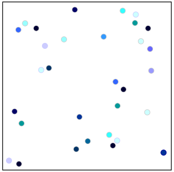
- Equivalently, the expected MST of a random sample of M?

Uncertain Minimum Spanning Tree



- Outcome $A \subseteq M$ occurs with prob. $\Pr[A] = \prod_{s_i \in A} p_i \prod_{s_i \notin A} (1 - p_i)$
- The sample space has 2^n outcomes (sets of active points).
- Compute $\mathbb{E}[MST] = \sum_{S \subseteq M} p(S) MST(S)$.

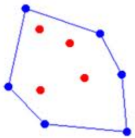
Uncertain Minimum Spanning Tree

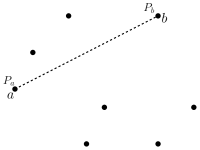


- Outcome $A \subseteq M$ occurs with prob. $\Pr[A] = \prod_{s_i \in A} p_i \prod_{s_i \notin A} (1 - p_i)$
- The sample space has 2^n outcomes (sets of active points).
- Compute $\mathbb{E}[MST] = \sum_{S \subseteq M} p(S) MST(S)$.
- Sum over exponentially many terms worrisome, but...

Computational Geometry under Uncertainty

- Geometric structure can help.
- Consider the expected size (perimeter) of convex hull
- A (directed) pair (a, b) forms an edge of CH iff
  - both a and b active
  - no point on the negative side of the line ab active
- Weighted sum of ab lengths with their prob (linearity of expectation)



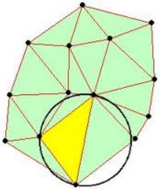


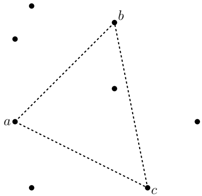
- At worst, $O(n^3)$ time. Similarly, for the CH area.

Expectation for Proximity Graphs

Expectation for Proximity Graphs

- A triple (a, b, c) forms a Delaunay triangle iff
  - a, b, c are all active
  - no point inside circumcircle of $\triangle abc$ is active
- Weighted sum of triangles with their prob (linearity of expectation)
- Subtract the (expected) perimeter

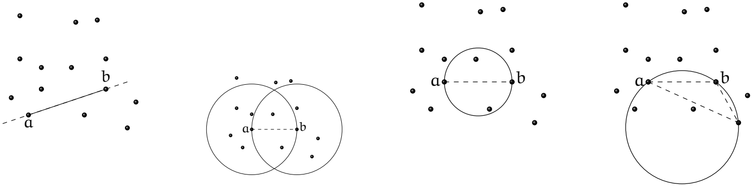




Back to MST and Proximity Graphs

Back to MST and Proximity Graphs

- A *master* set $M = \{s_1, s_2, \ldots, s_n\}$ of points in d dimensions.
- Each s_i is *active* with an independent probability p_i.
- What is the expected MST length of M?
- MST is part of a family: NN, RNG, GG, DT.



(a) (b) (c) (d)

MST and Proximity Graphs

- $\mathrm{NN} \subseteq MST \subseteq \mathrm{RNG} \subseteq \mathrm{GG} \subseteq \mathrm{DT}$
- Expected lengths of NN, GG, RNG, and DT in poly-time.
- Unfortunately, none of them good approximations of MST.
- In worst-case, DT is $\Omega(n) \times MST$, and NN arbitrarily smaller.

Results on Stochastic MST

- Complexity:
  - $\mathbb{E}[MST]$ is #P-Hard for dim $d \geqslant 2$.
  - Trivial in one dimension.
- Approximation of Expectation:
  - A simple *randomized* FPTAS in all dimensions.
  - A deterministic $O(1)$ factor approximation for $d = 2$.
  - A PTAS based on shifted quadtrees and dynamic programming.
- Probability Distribution:
  - Tail bounds inapproximable to any multiplicative factor.
- Hardness and approximation for locational uncertainty model.

Hardness: Reduction from Network Reliability

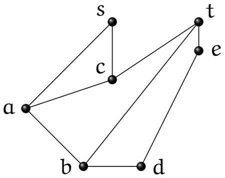
- 2-Terminal Network Reliability Problem (2NRP).
  - $G = (V, E)$, nodes s, t, and failure prob. p_i for each $e_i \in E$.
  - Compute the *probability* that s and t are connected.

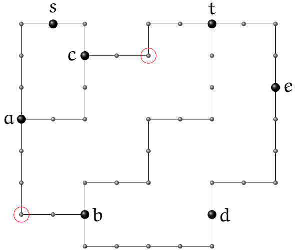
Hardness: Reduction from Network Reliability

- 2-Terminal Network Reliability Problem (2NRP).
  - $G = (V, E)$, nodes s, t, and failure prob. p_i for each $e_i \in E$.
  - Compute the *probability* that s and t are connected.
- An (s, t)-*planar graph* is one that admits a planar embedding with s and t on the boundary.
- 2NRP is $\#$P-Hard for (s, t)-planar graphs of maximum degree 3 even if all edge failure probabilities are the same p [Provan 83].

The Construction

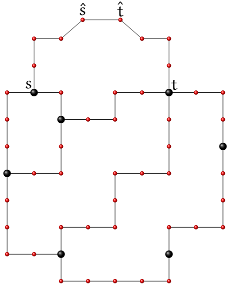
- Given an (s, t)-planar 2NRP, construct a stochastic set of points.
- Compute an orthogonal grid drawing of G [Tamassia '87].

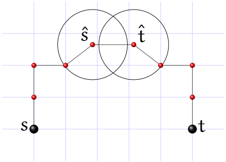




- Edges of G map to "paths" in the grid, using "auxiliary" grid points. Call these paths *virtual* edges.
- Each virtual edge has one special (*representative*) point, which is active with prob. p; all others active with prob. 1.

The Construction

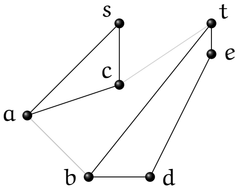


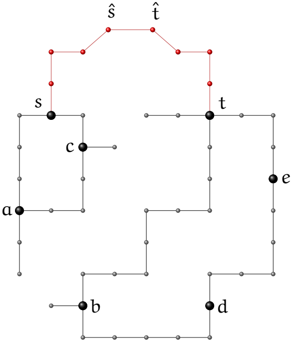


- Add a virtual edge (path) between s and t.
- Add $\hat{s}$ and $\hat{t}$ in the middle with $d(\hat{s}, \hat{t}) = 1.1$ (keeping unit distance to neighboring auxiliary points)
- All interpoint distances 1 (short), 1.1 (medium), or $\geqslant \sqrt{2}$ (long).

Network Reliability to MST

- H: surviving subgraph for 2NRP (an outcome).
- S_H: corresponding point set (without pts. of failed edges).





- Lemma 1: Nodes s and t connected in H iff $\hat{s}\hat{t} \notin MST(S)$.

Finishing the Proof

- Lemma 2: The probability that $ab \in MST$ does not change if $d(\hat{s}, \hat{t})$ changes from 1.1 to 1.2, for any other edge ab.

Finishing the Proof

- Lemma 2: The probability that $ab \in MST$ does not change if $d(\hat{s}, \hat{t})$ changes from 1.1 to 1.2, for any other edge ab.
  - Compute $\mathbb{E}[MST]$ twice, with $d(\hat{s}, \hat{t})$ equal to 1.1 and 1.2.
  - $\mathbb{E}[MST_2] - \mathbb{E}[MST_1] \; = \; 0.1 * p(\hat{s}, \hat{t})$
  - Probability that s, t connected in G equals $1 - p(\hat{s}, \hat{t})$.

Finishing the Proof

- Lemma 2: The probability that $ab \in MST$ does not change if $d(\hat{s}, \hat{t})$ changes from 1.1 to 1.2, for any other edge ab.
  - Compute $\mathbb{E}[MST]$ twice, with $d(\hat{s}, \hat{t})$ equal to 1.1 and 1.2.
  - $\mathbb{E}[MST_2] - \mathbb{E}[MST_1] \; = \; 0.1 * p(\hat{s}, \hat{t})$
  - Probability that s, t connected in G equals $1 - p(\hat{s}, \hat{t})$.

- Computing $\mathbb{E}[MST]$ is #P-Hard for $d \geqslant 2$.

Approximation: $\mathbb{E}[MST]$ by Sampling

- A sample R_j picks each point s_i with probability p_i
- Random variable X_j is length of R_j's MST
- Construct k samples and output the mean length $\sum_{j=1}^{k} X_j/k$.
- How large should k be to get an (ε, δ) approximation?

Approximation: $\mathbb{E}[MST]$ by Sampling

- A sample R_j picks each point s_i with probability p_i
- Random variable X_j is length of R_j's MST
- Construct k samples and output the mean length $\sum_{j=1}^{k} X_j/k$.
- How large should k be to get an (ε, δ) approximation?
- Sample size depends on $\frac{\max|MST|}{\mathbb{E}[MST]}$, the range for the random variable.
- Problematic when point spread is large and probabilities small.

Approximation: $\mathbb{E}[MST]$ by Sampling

- A sample R_j picks each point s_i with probability p_i
- Random variable X_j is length of R_j's MST
- Construct k samples and output the mean length $\sum_{j=1}^{k} X_j/k$.
- How large should k be to get an (ε, δ) approximation?

- Sample size depends on $\frac{\max|MST|}{\mathbb{E}[MST]}$, the range for the random variable.
- Problematic when point spread is large and probabilities small.
- Ways to control this via conditioning.

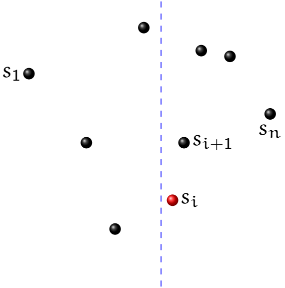
Approximating $\mathbb{E}[MST]$ by Conditioning

- Order the points as $s_1, s_2, \ldots, s_n$.
- L_i be expected MST length of $\{s_i, s_{i+1}, \ldots, s_n\}$.

Approximating $\mathbb{E}[MST]$ by Conditioning

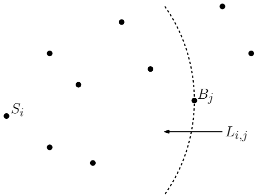
- Order the points as $s_1, s_2, \ldots, s_n$.
- L_i be expected MST length of $\{s_i, s_{i+1}, \ldots, s_n\}$.
- L'_i be expected value of L_i conditioned on s_i being active.

- $L_i = p_i L'_i + (1 - p_i) L_{i+1}$
- Need a recursive formula for L'_i.



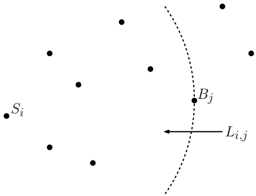
Approximating $\mathbb{E}[MST]$ by Conditioning

- $L_i \;=\; p_i L_i' \;+\; (1 - p_i) L_{i+1}$



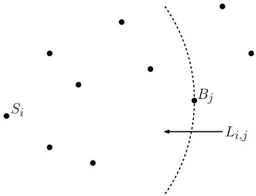
Approximating $\mathbb{E}[MST]$ by Conditioning

- $L_i = p_i L_i' + (1 - p_i) L_{i+1}$
- Now reorder $\{s_{i+1}, \ldots, s_n\}$ in increasing distance order from i. Assume this order is $\{s_{i,i+1}, \ldots, s_{i,n}\}$.

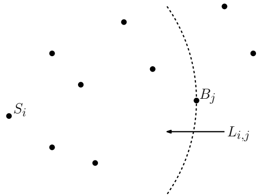


Approximating $\mathbb{E}[MST]$ by Conditioning

- $L_i \;=\; p_i L_i' \;+\; (1-p_i)L_{i+1}$
- Now reorder $\{s_{i+1}, \ldots, s_n\}$ in increasing distance order from i. Assume this order is $\{s_{i,i+1}, \ldots, s_{i,n}\}$.
- L_{ij}' expected MST length of $\{s_i, s_{i,i+1}, \ldots, s_{i,j}\}$ conditioned on s_i being active.

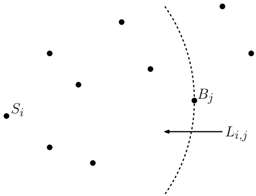


- $L_i = p_i L'_i + (1 - p_i) L_{i+1}$
- Now reorder $\{s_{i+1}, \ldots, s_n\}$ in increasing distance order from i. Assume this order is $\{s_{i,i+1}, \ldots, s_{i,n}\}$.
- L'_{ij} expected MST length of $\{s_i, s_{i,i+1}, \ldots, s_{i,j}\}$ conditioned on s_i being active.
- L''_{ij} expected value of L'_{ij} conditioned on both s_i and $s_{i,j}$ being active.



Approximating $\mathbb{E}[MST]$ by Conditioning

- $L_i \;=\; p_i L'_i \;+\; (1 - p_i) L_{i+1}$
- Now reorder $\{s_{i+1}, \ldots, s_n\}$ in increasing distance order from i. Assume this order is $\{s_{i,i+1}, \ldots, s_{i,n}\}$.
- L'_{ij} expected MST length of $\{s_i, s_{i,i+1}, \ldots, s_{i,j}\}$ conditioned on s_i being active.
- L''_{ij} expected value of L'_{ij} conditioned on both s_i and $s_{i,j}$ being active.
- Then, $L'_{i,j} \;=\; p_{i,j} L''_{i,j} \;+\; (1 - q_{i,j}) L'_{i,j-1}$



Approximating $\mathbb{E}[MST]$ by Conditioning

- $L'_{i,j} \ = \ p_{i,j} L''_{i,j} \ + \ (1 - q_{i,j}) L'_{i,j-1}$
- When i and its farthest neighbor are active, and have distance D, then min $|MST|$ is $\Omega(D)$ and max $|MST|$ is $O(nD)$.
- $O(n)$ samples suffice for estimating L''
- Total running time $O(poly(n/\varepsilon) \log(1/\delta))$.

Approximating $\mathbb{E}[MST]$ by Conditioning

- $L'_{i,j} \; = \; p_{i,j} L''_{i,j} \; + \; (1 - q_{i,j}) L'_{i,j-1}$
- When i and its farthest neighbor are active, and have distance D, then $\min |MST|$ is $\Omega(D)$ and $\max |MST|$ is $O(nD)$.
- $O(n)$ samples suffice for estimating L''
- Total running time $O(\text{poly}(n/\varepsilon) \log(1/\delta))$.
- Randomized FPTAS for $\mathbb{E}[MST]$ in any metric space.

Distribution of MST Length

Distribution of MST Length

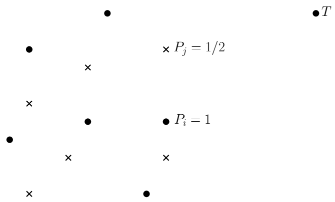
- p_ℓ be Prob. that MST length is at most ℓ.
- c-approximation of p_ℓ:

$$\frac{1}{c}p_\ell \leqslant p' \leqslant cp_\ell$$

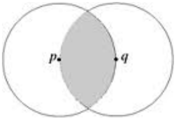
- Not possible assuming $P \neq NP$.
- Reduction from Steiner tree problem.

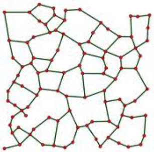
Tail Bound for Probabilistic MST

- A set S of points, and a subset $T \subset S$ called *terminals*.
- NP-complete to decide if Steiner tree of T has length ℓ.
- Set prob. 1 for points of T, and prob. $1/2$ for points of $S \setminus T$.
- The Prob. that $MST(S)$ length is less than ℓ is non-zero if and only if Steiner tree of T has length less than ℓ.
- Thus, $p_\ell = 0$ if Steiner tree answer is no, and positive otherwise.



Deterministic Approximation of $\mathbb{E}[MST]$ in 2D

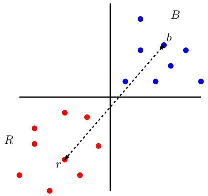




- Relative Neighborhood Graph length can be computed but a poor approximation of $\mathbb{E}[MST]$.
- Apply a pruning rule to RNG that
  - Must be close to MST weight, and
  - Must admit a probabilistic estimation
- Pruning Rule:
  - Delete an edge $uv \in$ RNG if there is a pair $a, b \in S$ such that uv is the longest edge of 4-cycle (u, v, a, b).
- Complicated analysis but raises another fundamental problem.

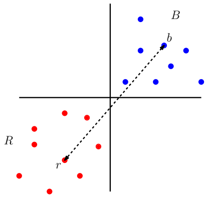
Stochastic Closest Pair

- Stochastic point sets R and B.
  What is the probability that closest R-B pair has distance > 1?



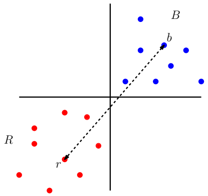
Stochastic Closest Pair

- Stochastic point sets R and B.
  What is the probability that closest R-B pair has distance > 1?
- Points fail with ind. prob. but need to analyze "edges."



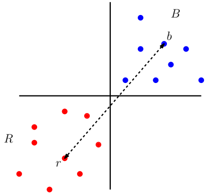
Stochastic Closest Pair

- Stochastic point sets R and B.
  What is the probability that closest R-B pair has distance > 1?
- Points fail with ind. prob. but need to analyze "edges."
- A graph version of the problem:



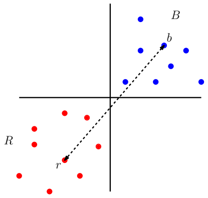
Stochastic Closest Pair

- Stochastic point sets R and B.
  What is the probability that closest R-B pair has distance > 1?
- Points fail with ind. prob. but need to analyze "edges."
- A graph version of the problem:
  - A bipartite graph $G = (U, V, E)$, each node fails with prob. p_i



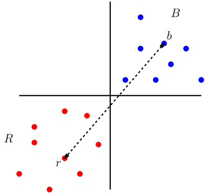
Stochastic Closest Pair

- Stochastic point sets R and B.
  What is the probability that closest R-B pair has distance > 1?
- Points fail with ind. prob. but need to analyze "edges."
- A graph version of the problem:
  - A bipartite graph $G = (U, V, E)$, each node fails with prob. p_i
  - What is the probability that no edge survives?



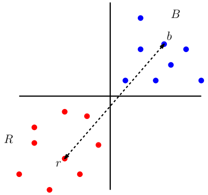
Stochastic Closest Pair

- Stochastic point sets R and B.
  What is the probability that closest R-B pair has distance > 1?
- Points fail with ind. prob. but need to analyze "edges."
- A graph version of the problem:
  - A bipartite graph $G = (U, V, E)$, each node fails with prob. p_i
  - What is the probability that no edge survives?
- Graph problem is NP-Hard: related to counting vertex covers.



Complexity of Stochastic Closest Pair

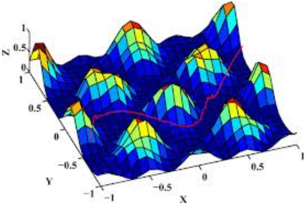
- Computing *Prob[Closest Pair distance in $S \leqslant \ell$]* is $\#\mathrm{P}$-Hard, even in 2D, for either L_2 or L_∞ norm.
- Bi-chromatic version (R, B) also hard.



- Polynomial algorithm if R and B linearly-separable and L_∞ norm.
- Hard if linear separability removed.
- Even linearly-separable and L_∞ hard in 3D.

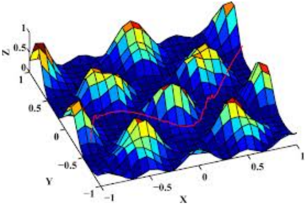
Traveling Salesman Tour Through Uncertain Regions

- Plan a shortest tour visiting geometric neighborhoods.
- Neighborhood are uncertain.
- Each region is a disk, with a known center, but random radius.



Traveling Salesman Tour Through Uncertain Regions

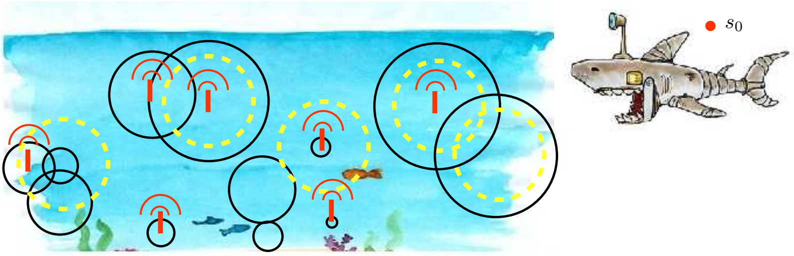
- Plan a shortest tour visiting geometric neighborhoods.
- Neighborhood are uncertain.
- Each region is a disk, with a known center, but random radius.



- Motivation: sensor network data collection.

Traveling Salesman Tour Through Uncertain Regions

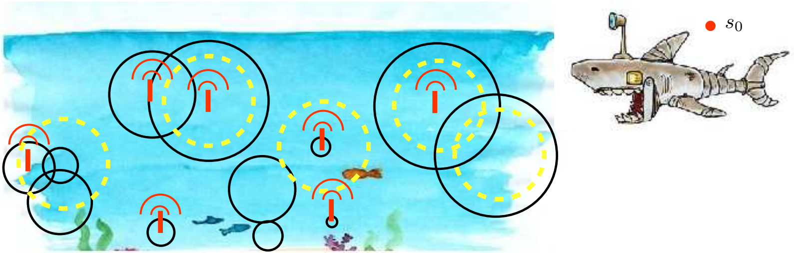
- Buoy-mounted sensors in Southern California Blight.



- Data periodically collected by AUV robots.
- Communication (acoustic) range a stochastic variable.
- Shortest tour to visit all sensor “neighborhoods”.

Traveling Salesman Tour Through Uncertain Regions

- Buoy-mounted sensors in Southern California Blight.



- Data periodically collected by AUV robots.
- Communication (acoustic) range a stochastic variable.
- Shortest tour to visit all sensor "neighborhoods".
- Online: radii learned only when disk boundary reached.

Stochastic TSP: formal model

- Input: n (fixed) disk centers, i.i.d. random radii, from distribution ϕ, with mean μ.
- Each *draw* is a different instance of the TSPN problem.
- Each instance I (random draw) has an optimal tour $\textsc{Opt}(I)$.
- $\mathbb{E}[L^*]$: *expected* value of $\textsc{Opt}(I)$ over all the instances.

$$\mathbb{E}[L^*] \; = \; \int_0^\infty \cdots \int_0^\infty L^*(x_1, \ldots, x_n) \cdot \Pi_{i=1}^n \, \phi(x_i) \cdot dx_1 \ldots dx_n,$$

- Find a traversal strategy with a provable approximation of $\mathbb{E}[L^*]$.

The Main Idea

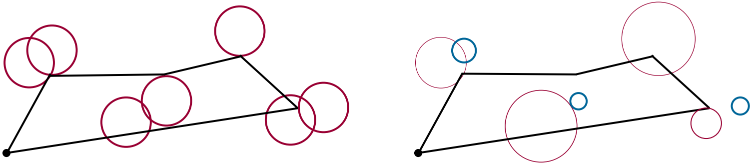
- $\mathcal{M}$: *mean instance* where all the disks have radius μ.
- $\textsc{Opt}(\mathcal{M})$: optimal tour for $\mathcal{M}$.

The Main Idea

- $\mathcal{M}$: *mean instance* where all the disks have radius μ.
- $\textsc{Opt}(\mathcal{M})$: optimal tour for $\mathcal{M}$.
- Theorem: $\textsc{Opt}(\mathcal{M}) \leqslant 2\mathbb{E}[L^*]$.

The Main Idea

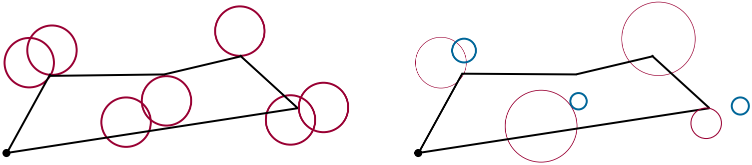
- M: *mean instance* where all the disks have radius μ.
- $\textsc{Opt}(M)$: optimal tour for M.
- Theorem: $\textsc{Opt}(M) \leqslant 2\mathbb{E}[L^*]$.



- Blindly following $\textsc{Opt}(M)$ doesn't work. Only a high level clue about the visit order.

The Main Idea

- M: *mean instance* where all the disks have radius μ.
- $\textsc{Opt}(M)$: optimal tour for M.
- Theorem: $\textsc{Opt}(M) \leqslant 2\mathbb{E}[L^*]$.



- Blindly following $\textsc{Opt}(M)$ doesn't work. Only a high level clue about the visit order.
- $O(1)$ factor approximation if disks in M disjoint.
- Otherwise, $O(\log\log n)$ (offline) and $O(\log n)$ (online) approx.

Conclusions and Future Directions

- Effects of uncertainty on complexity of geometric problems
- Even basic questions (closest pair, MST) become intractable
  - Although many others (DT, RNG, GG, CH etc.) tractable.
- More questions than answers
  - Going beyond the *measure*, what about the *structure*?
  - What to output as 'expected' MST, VoD, DT?
  - Under what conditions, this makes sense?
- Many other geometric questions (clustering, shortest paths).

- Thank you!