Approximately Uniform Random Sampling in Sensor Networks

Boulat A. Bash, John W. Byers and Jeffrey Considine

Computer Science

Motivation

- Data aggregation
  - Approximations to COUNT, SUM, AVG, MEDIAN
  - Existing work does not use sampling
    - TAG (Madden et al. 2002)
    - State of the art: FM sketches (Considine et al. 2004)
- Randomized algorithms
  - e.g. randomized routing

Introduction

Computer Science

- What is this talk about?
  - Selecting (sampling) a random node in a sensornet
- Why is sampling hard in sensor networks?
  - Unreliable and resource-constrained nodes
  - Hostile environments
  - High inter-node communication costs
- How do we measure costs?
  - Total number of fixed-size messages sent per query

Outline

Computer Science

- Exact uniform random sampling
  - Previous work
- Approximately uniform random sampling
  - Naïve biased solution
  - Our almost-unbiased algorithm
  - Experimental validation
  - Heuristics for improving samples
- Preliminary simulations
- Conclusions and future work

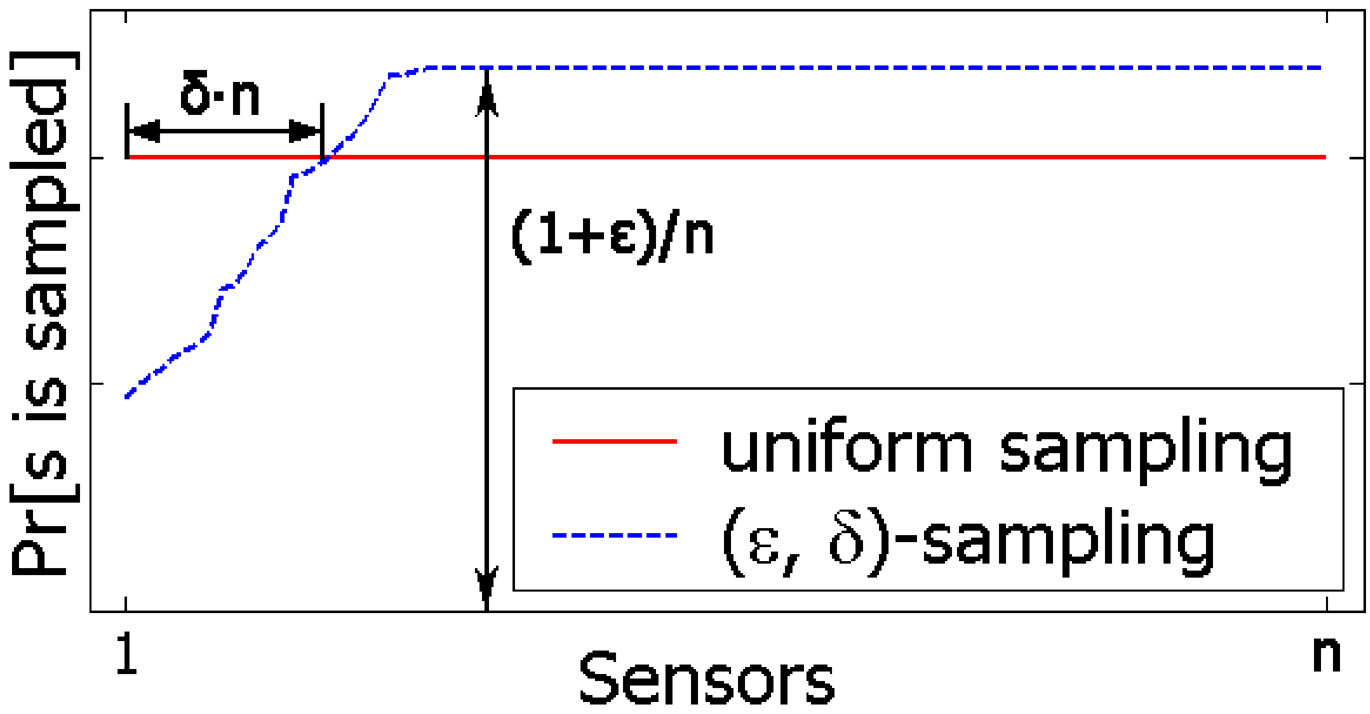
Sampling Problem

- Exact uniform random sampling
  - Each sensor s is returned from network of n reachable sensors with probability 1/n
- Existing solution (Nath and Gibbons, 2003)
  - Each sensor s generates (r_s, ID_s) where r_s is a random number
  - Network returns ID of the sensor with minimal r_s
  - Cost: $\Theta(n)$ transmissions

Relaxed Sampling Problem

Boston University
Computer Science

- (ε, δ)-sampling

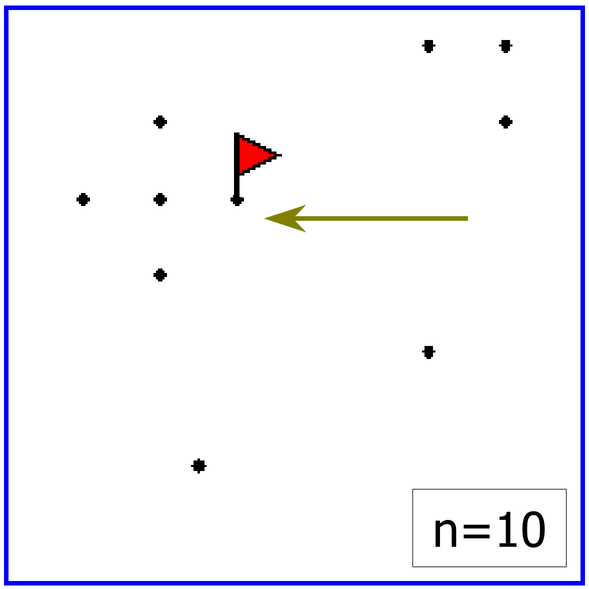


  - Each sensor s is returned with probability no greater than (1+ε)/n, and at least (1-δ)·n sensors are output with probability at least 1/n

Naïve Solution

Computer Science

- Spatial Sampling
  - Return the sensor closest to a random (x,y)

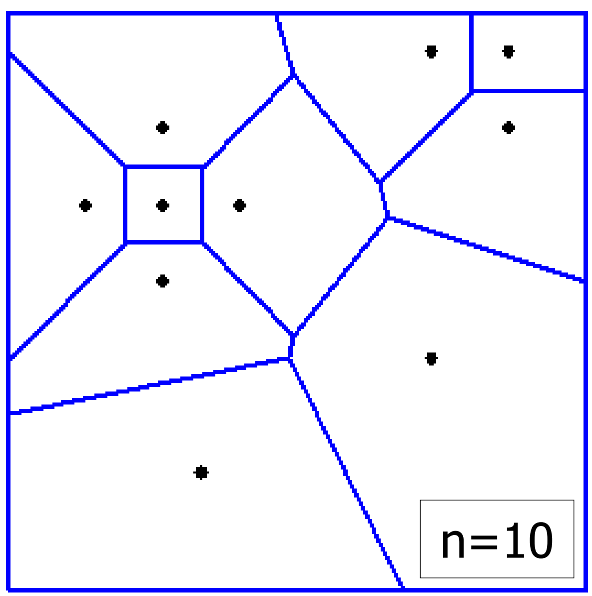


  - Possible with geographic routing (GPSR 2001)
    - Nodes know own coordinates (GPS, virtual coords, pre-loading)
    - Fully distributed; state limited to neighbors' locations
  - <u>Cost</u>: Θ(D) transmissions, D is network diameter

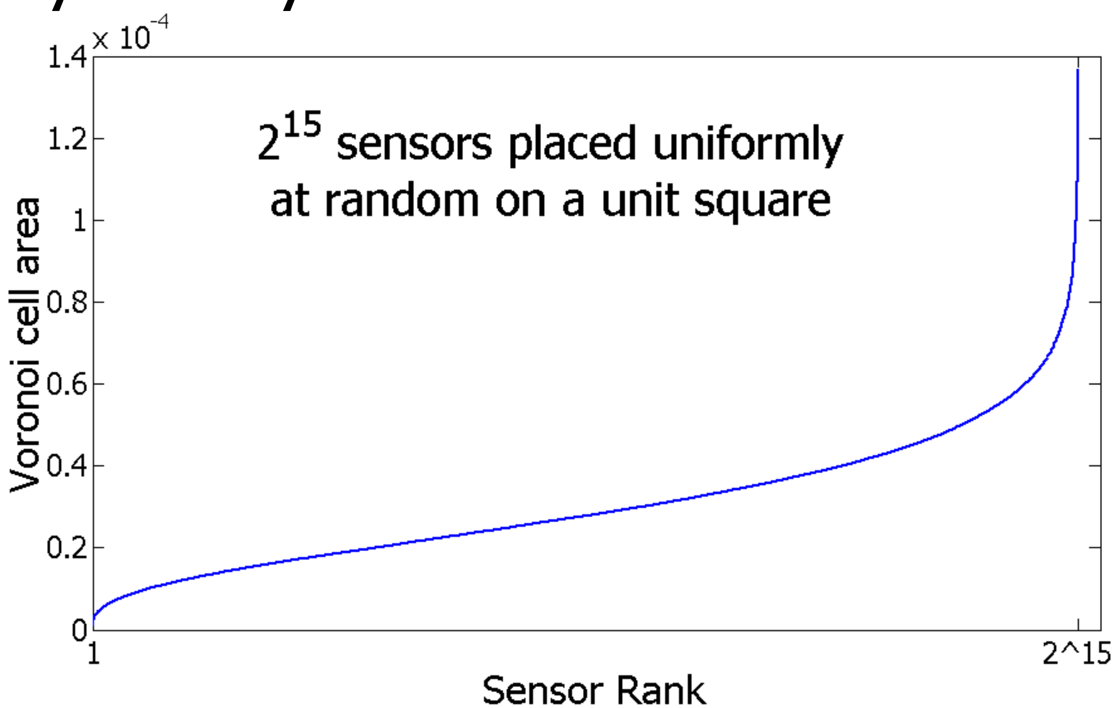
Pitfall in Spatial Sampling

Boston University Computer Science

- Bias towards large Voronoi cells
  - Definition: Set of points closer to sensor s than any other sensor (Descartes, 1644)
  - Areas known to vary widely

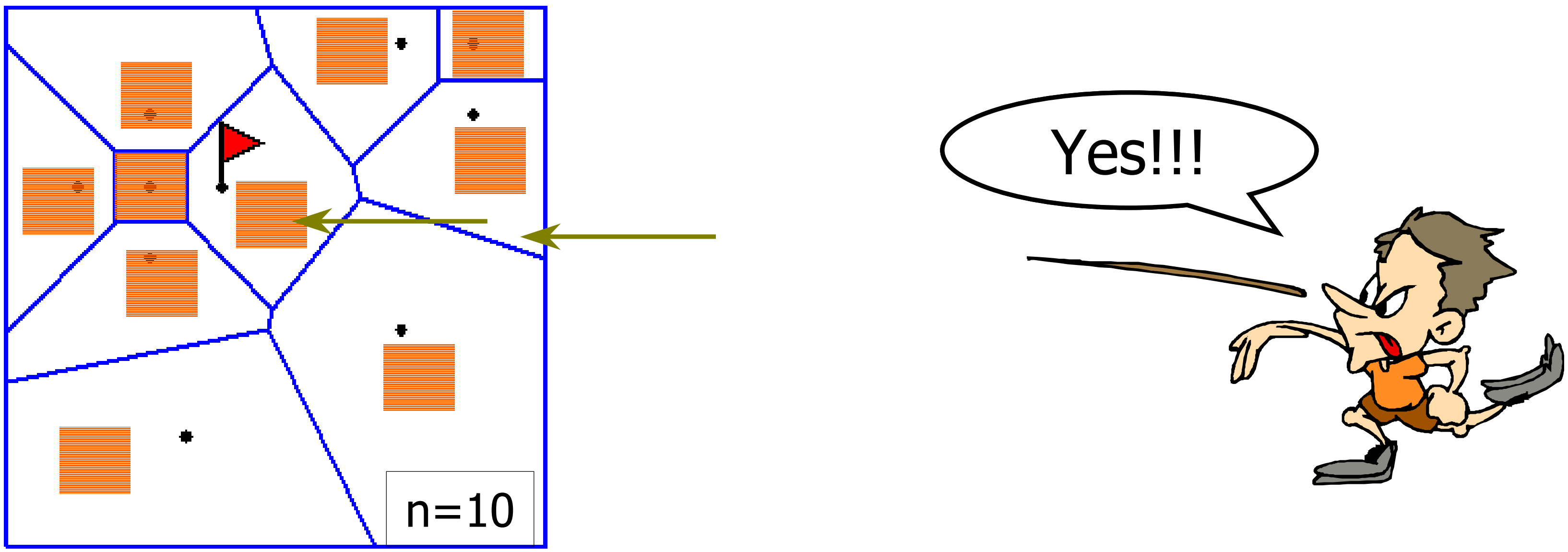


Voronoi diagram



Removing Bias

- Rejection method



  - In each cell, mark area of smallest Voronoi cell
  - Only accept probes that land in marked regions
    - In practice, use Bernoulli trial for acceptance with **P[**accept**]** = A_{min}/A_s (von Neumann, 1951)
    - Find own cell area A_s using neighbor locations
    - Need c = A_{avg}/A_{min} probes per sample on average

Rejection-based Sampling

Computer Science

- Problem: Minimum cell area may be small
- Solution: Under-sample some nodes
  - Let $A_{threshold} \geq A_{min}$ be globally-known cell area

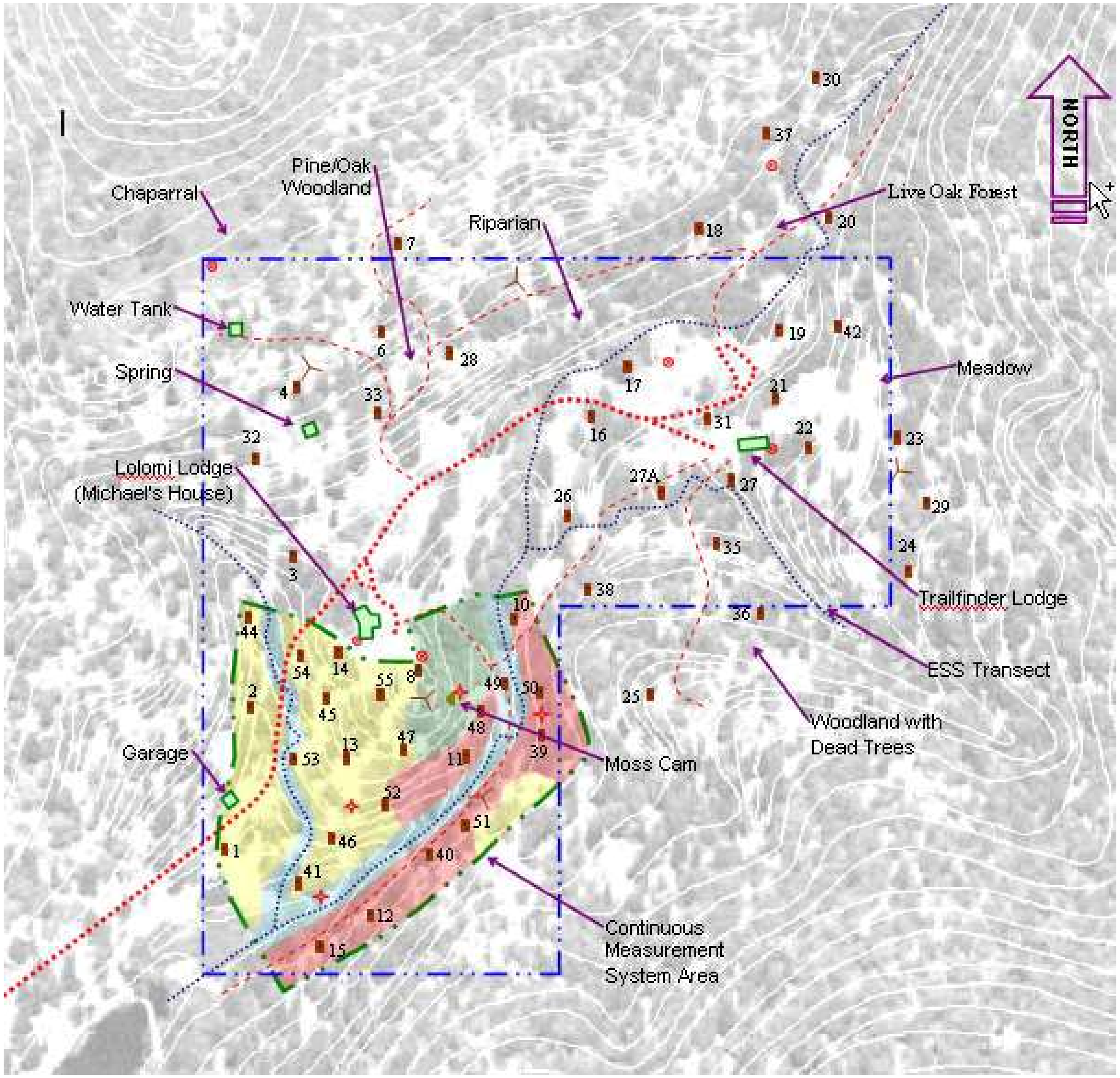
1. Route probe to sensor s closest to random (x,y)
2. If $A_s < A_{threshold}$, then sensor s accepts

   Else, sensor accepts with **Pr[**acc**]** = $A_{threshold}/A_s$

- $A_{threshold}$ set by user
  - For (ε, δ)-sampling, set to the area of the cell that is the k-quantile, where $k = \min(\delta, \varepsilon/(1+\varepsilon))$
- Cost: Θ(cD) transmissions

James Reserve Sensornet

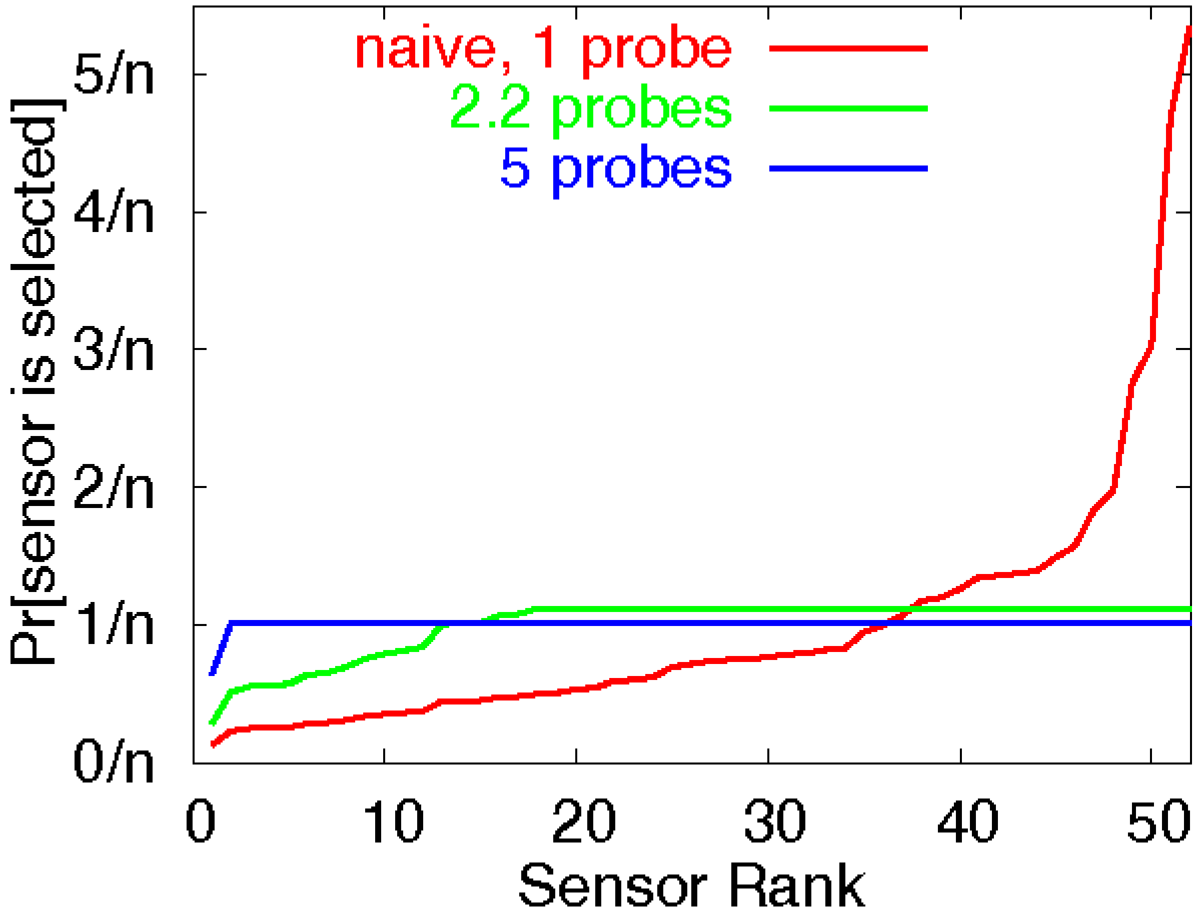
Computer Science



James Reserve Sensornet

Computer Science

- 52 sensors



E[#probes]	ε	δ
1.0 (naïve)	4.3	0.69
1.5	0.48	0.46
2.2	0.12	0.23
3.1	0.041	0.15
4.1	0.012	0.038
5.0	0.0072	0.019

Random topology

Computer Science

- 2^{15} sensors randomly placed on a unit square

E[#probes]	ε	δ
1.0 (naïve)	3.8	0.57
1.3	0.27	0.41
2.1	0.051	0.15
3.1	0.017	0.06
4.0	0.0079	0.029
5.0	0.0042	0.017

Improving Algorithm

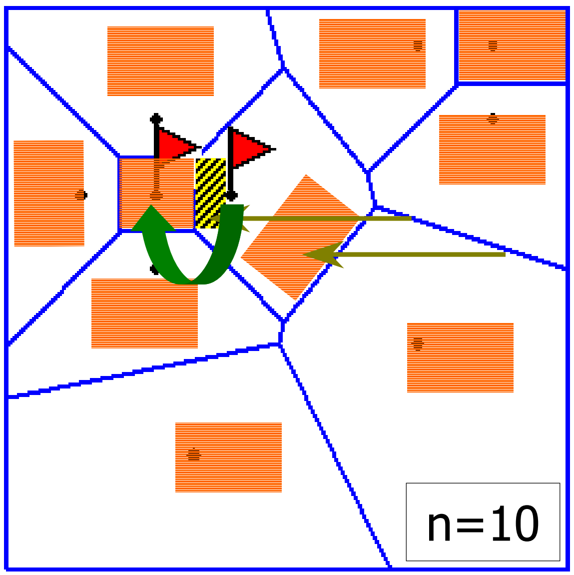
Computer Science

- Put some nodes with small cells to sleep
  - No sampling possible from sleeping nodes
  - Similar to power-saving schemes (Ye et al. 2002)

- Virtual Coordinates
  - Node locations assigned using local connectivity information (Rao et al. 2003)
  - Hard lower bound on inter-sensor distances

Improving Algorithm

- Pointers



  - Large cells donate their "unused" area to nearby small cells
  - When a large cell rejects, it can probabilistically forward the probe to one of its small neighbors

Simulation

Computer Science

- Objectives
  - Avoid the following behavior

  - Show feasibility of implementation
  - Demonstrate scalability
    - Contrast with previous work

Simulation Setup

Computer Science

- Existing simulation tools inadequate
  - Lack of scalability to tens of thousands of nodes
  - No middle-ground “proof of concept” tool
- SGNS – Simple Geographic Network Simulator
  - Implements GPSR
  - Counts messages only, similar to TAG Simulator
- Simulate on 3 topologies
  - James Reserve (JR) – 52 nodes
  - Random placement – 2^{15} nodes
  - Synthetic JR-based – 2^{15} nodes

Synthesizing Topologies

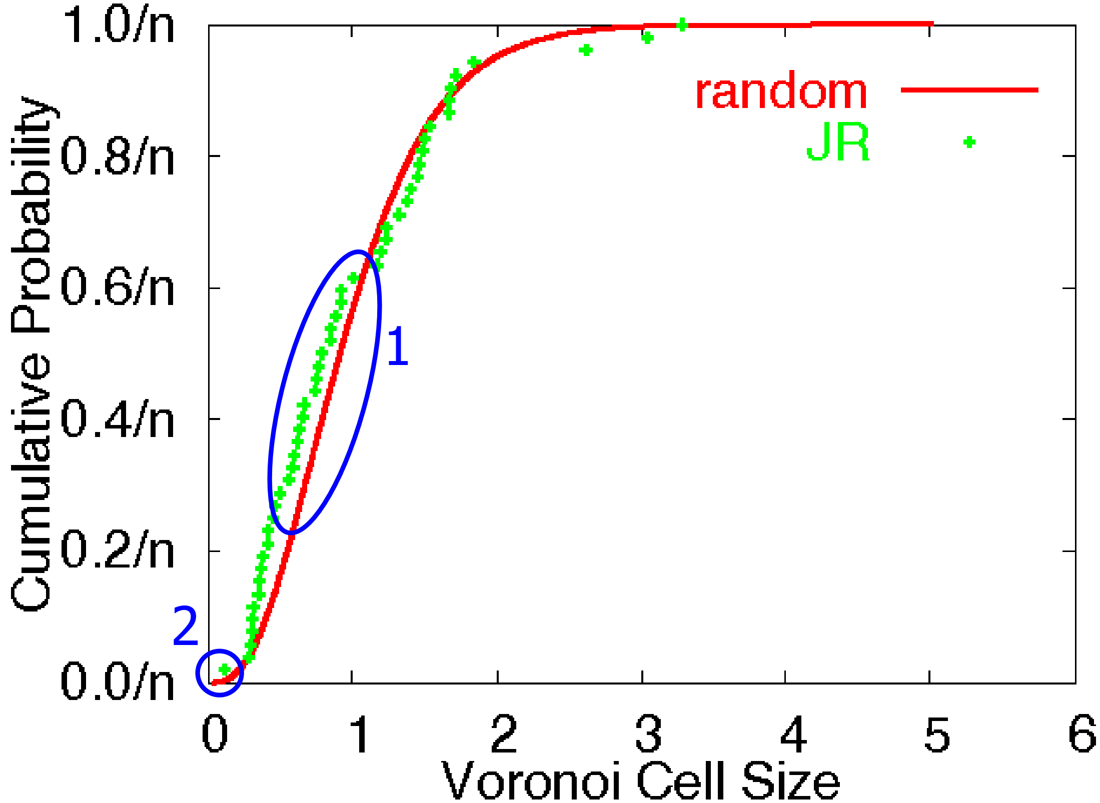
- Trivial random placement is not adequate
  - Humans do not behave randomly!
- First-principles approach to sensor placement
  - Inspired by Li et al. 2004

1. Nodes have two mutually non-exclusive tasks: sensing and routing
   - Areas of higher node densities
2. Humans are rational in sensor placements
   - Minimum inter-nodal distance
3. Sensed environment is not predictable
   - Non-uniformity in placement

First Principles in Action

Computer Science

- Voronoi cell size distributions: JR vs. random

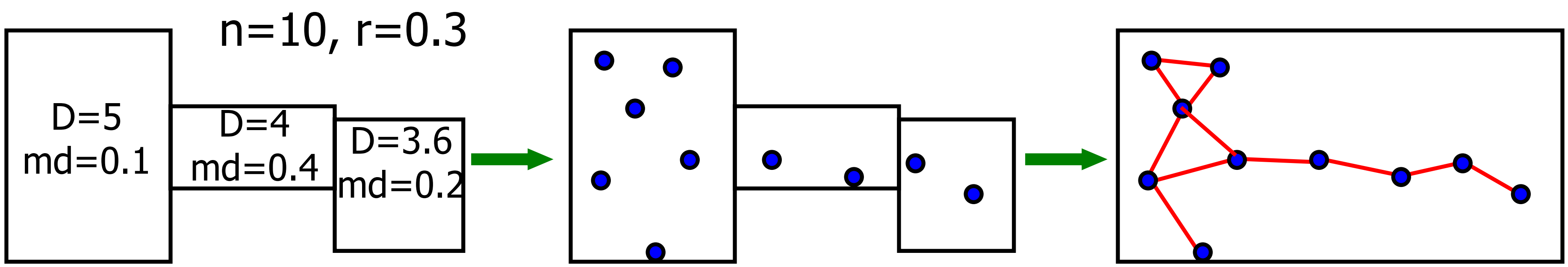


1. Proportionally more nodes with below-average area in JR
2. Smallest cells in random topology much smaller: sensors are too close

Topology Generator

Boston University
Computer Science

- STG – Synthetic Topology Generator



  - Inputs
    - Number of nodes and transmission radius
    - Set of non-overlapping axis-aligned rectangles
      - Relative node density
      - Minimum inter-nodal distance
    - Connectivity requirement
  - Iteratively place nodes on the rectangles at random

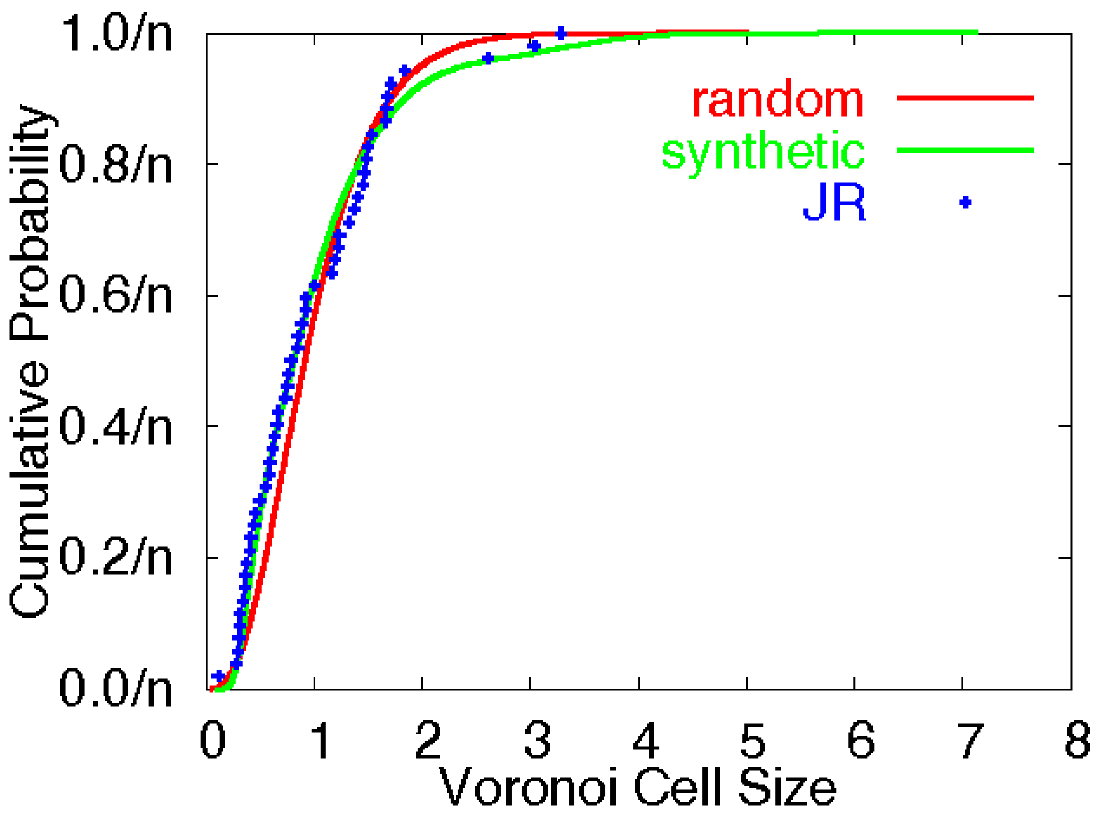
Inflating JR topology

Computer Science

- 2^{15}-node version of 52-node JR topology
  - 15 Rectangles: JR topology split into 3 × 5 grid
    - Node density maintained by increasing area by $2^{15}/52$
      - No zero-node rectangles
    - Minimum inter-nodal distance (MID) set to minimal distance between any two nodes in rectangle
      - One-node rectangles: use probabilistically maximal MID < r
  - Require connectivity
- Problem
  - "Scaling" assumption

Synthetic JR-based topology

- Voronoi cell size distribution



  - Looks better than random!

Simulation Results

	JR	*random*	*synthetic*
n	52	32,768	32,768
Dimensions (sq ft)	2,127x1,306	53,341x32,660	53,341x32,660
Transmission radius (ft)	500	500	500
Number of samples	80 Million	80 Million	80 Million
Expected number of probes	5	5	5
Total transmissions	8.74 Billion	42.38 Billion	25.24 Billion
Transmissions per sample	109	530	315
Greedy mode	14.190 %	67.091 %	64.875 %
Perimeter mode	1.331 %	3.746 %	0.008 %
Closest node determination	**84.480 %**	**29.164 %**	**35.117 %**
Transmissions per sample (excluding closest node determination)	17	375	205
N&G estimated trans. per sample	78	49,152	49,152

- Large fraction of transmissions is due to closest node determination

Simulation Conclusions

Computer Science

- To do list
  - Test sampling for data aggregation
    - Compare costs with TAG, FM sketches
  - Try inflating Voronoi cells of existing small topologies to obtain large synthetic topologies
    - instead of rectangles
  - Simulate "pointers" improvement
- Wish list
  - Node/link loss models for SGNS
  - Node mobility in SGNS
  - Massive TOSSIM simulation

Conclusions

Computer Science

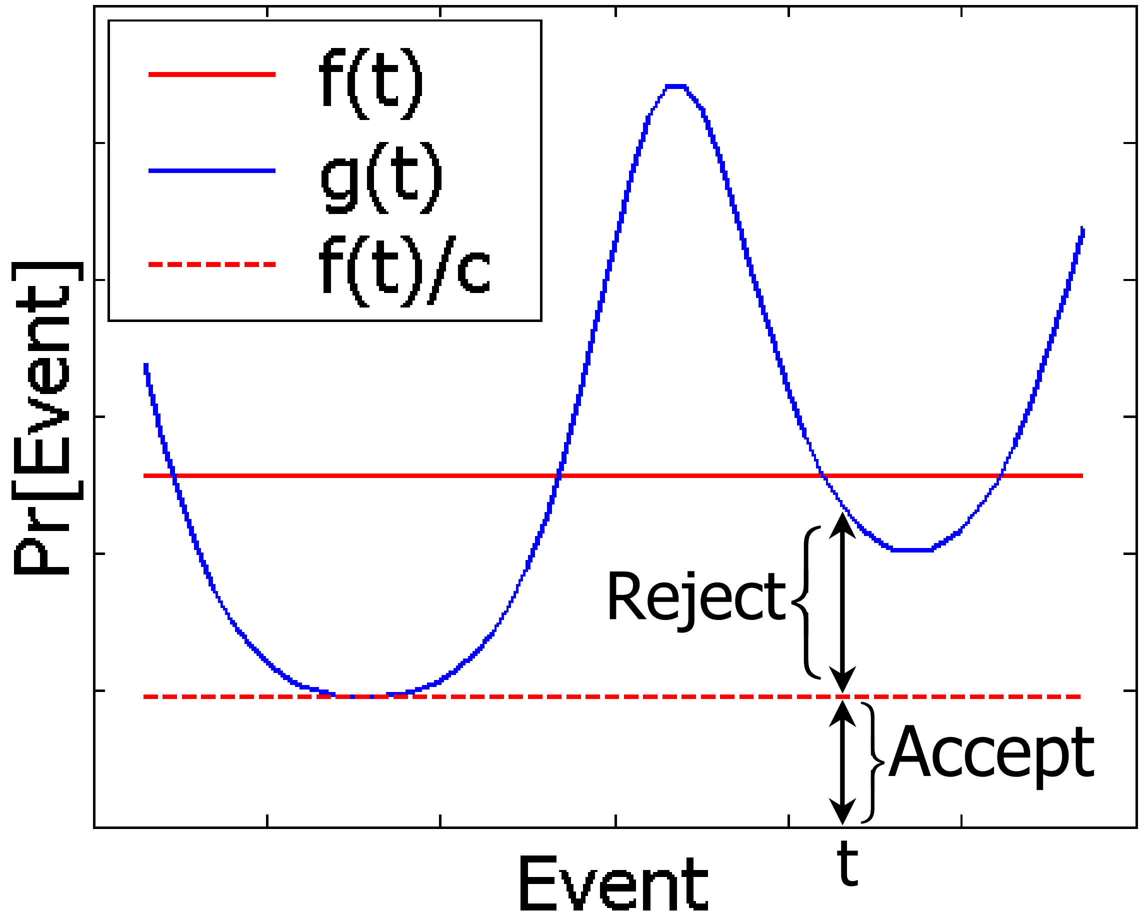
- New nearly-uniform random sampling algorithm
  - Cost proportional to sending a point-to-point message
  - Tunable (and generally small) sampling bias
  - Proof-of-concept simulations show viability
- Future Work
  - Extend to non-geographic predicates
  - Reduce messaging costs for high number of probes
  - Move beyond request/reply paradigm
  - Apply to DHTs like Chord (King and Saia, 2004)

Backup slides

- von Neumann’s Rejection Method
- Geographic Routing

Rejection Method

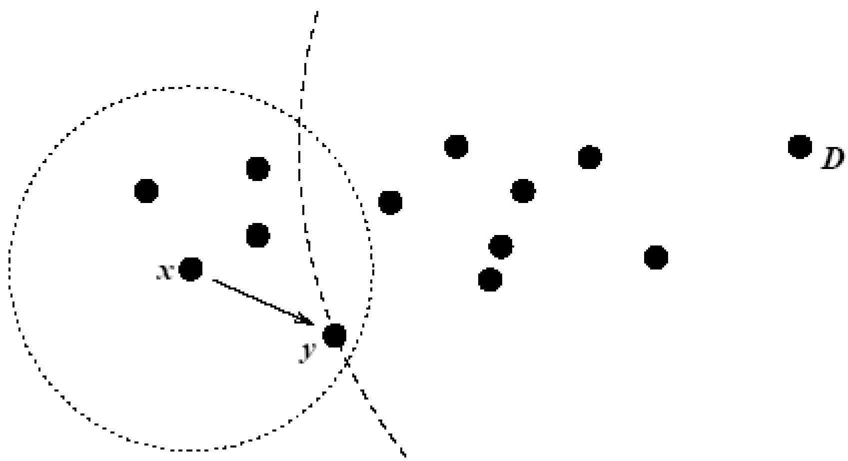
- von Neumann's rejection method (1951)
  - Problem: impossible to sample from PDF *f*
  - Idea: Sample indirectly from *g* and scale to *f*
  - Solution: Draw sample t from *g*, but accept with probability $\frac{f(t)}{c \cdot g(t)}$ where *c* is a positive constant



Geographic Routing

Computer Science

- GPSR – Greedy Perimeter Stateless Routing (Brad Karp and H.T. Kung, 2000)
  - State limited to neighbor location information
- Greedy mode – default protocol state

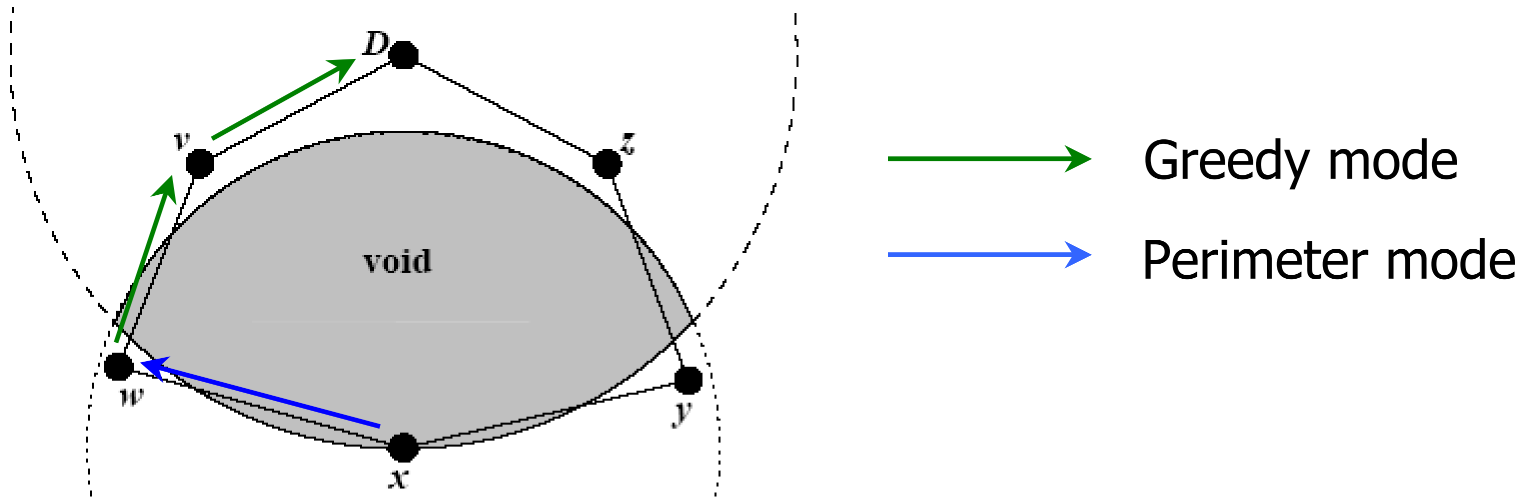


  - Forward to neighbor closest to the destination

Geographic Routing (cont.)

Computer Science

- Voids – where greedy forwarding fails

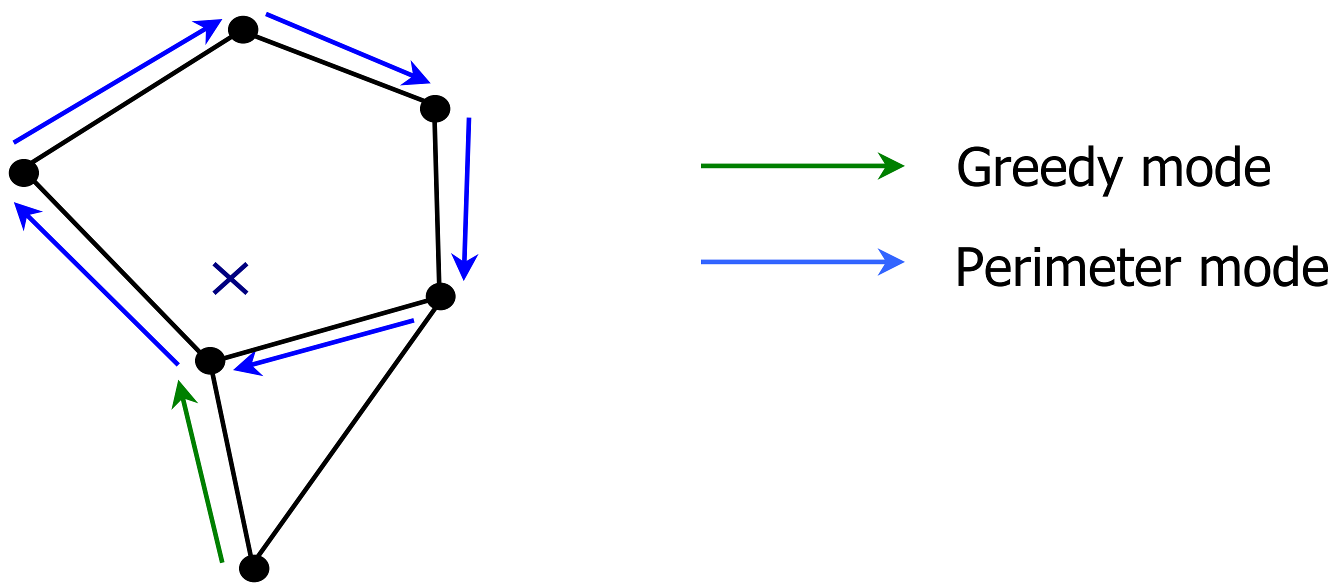


  - No nodes available in transmission range closer to destination then self
- Perimeter mode – circumnavigate voids using the right-hand rule

Geographic Routing (cont.)

Boston University Computer Science

- Closest node determination



  - Visit every node on a perimeter enclosing probe destination
  - Start and finish at the closest node