Data Collection Infrastructure for Location-Location-Unaware Sensor Networks

Distributed coding protocols for data storage

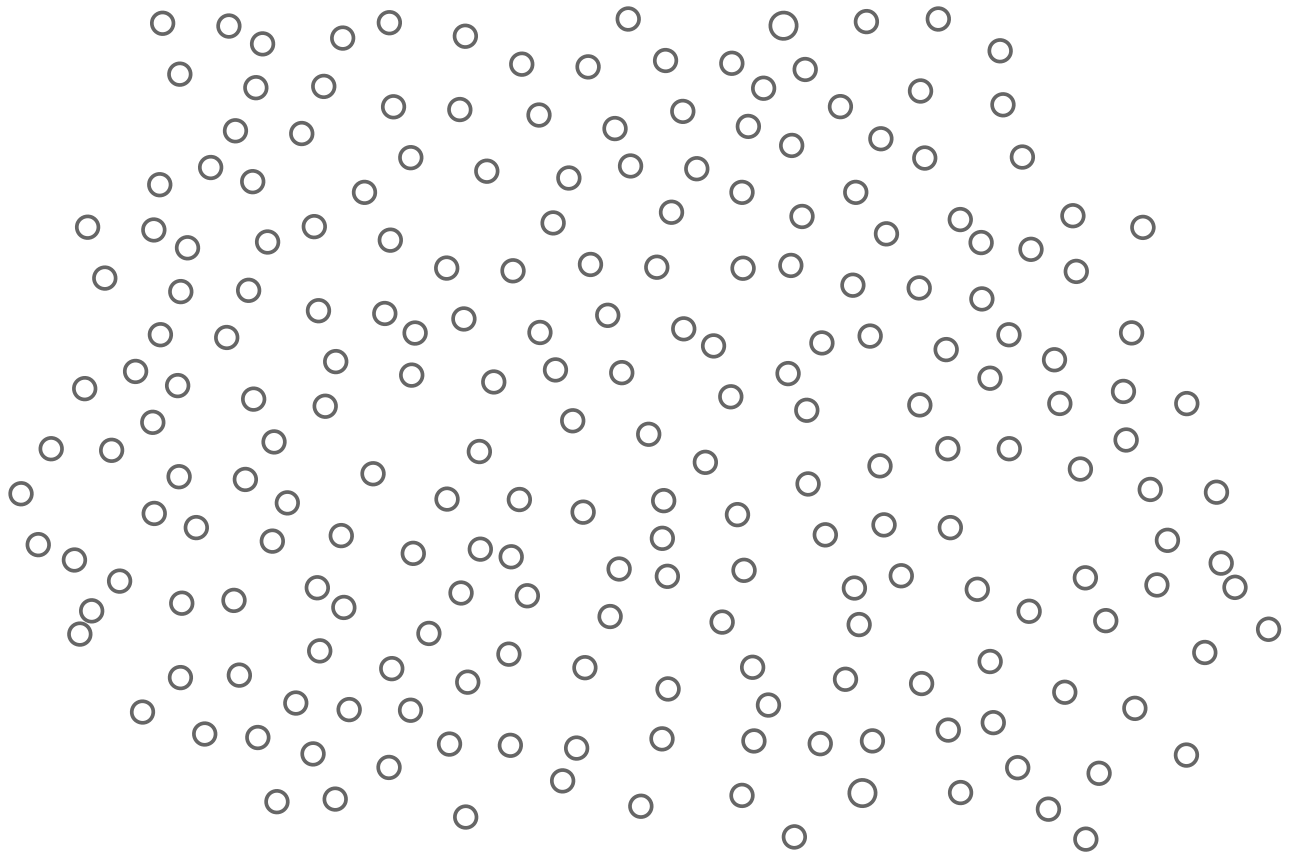
Silvija Kokalj-Filipovic

Predrag Spasojevic Roy Yates

Talk OutLine

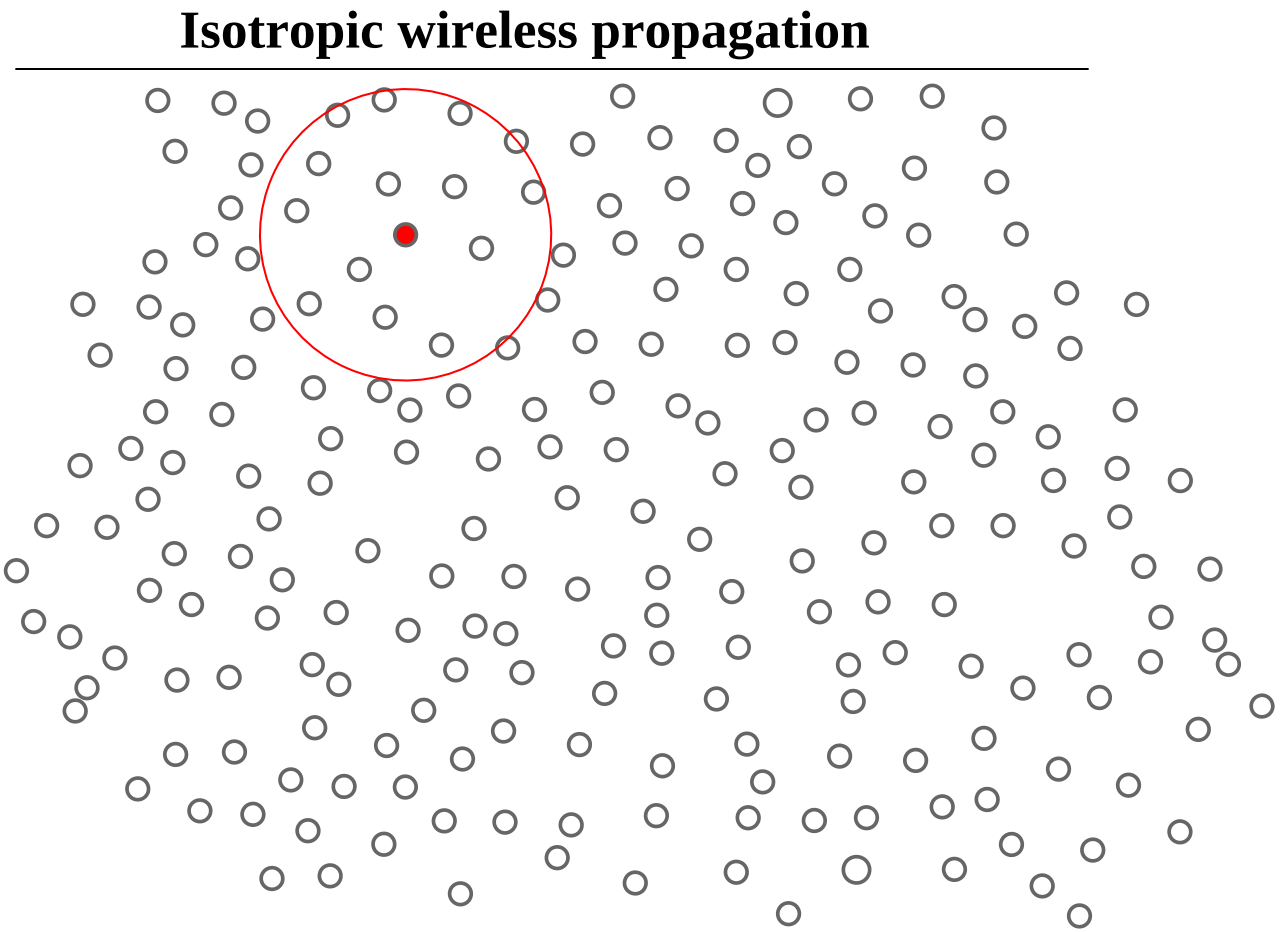
- Data Collection from a Location-Unaware Wireless Sensor Network
  - Network nodes self-organize into a web-like infrastructure of routes
  - Network data is encoded and stored along circular infrastructure routes using a distributed coding protocol
  - A Mobile Data Collector arrives to a random point of the network perimeter
  - Connects to the closest node of each circular route and collects encoded data from the nodes within its immediate neighborhood
  - Up-front collection from the neighborhood combined with polling distant nodes selectively to collect symbols which unlock the decoding process is an energy-efficient solution that allows for full decoding
  - The data collection is completed when the collector decodes all network data

Sensor Network Example



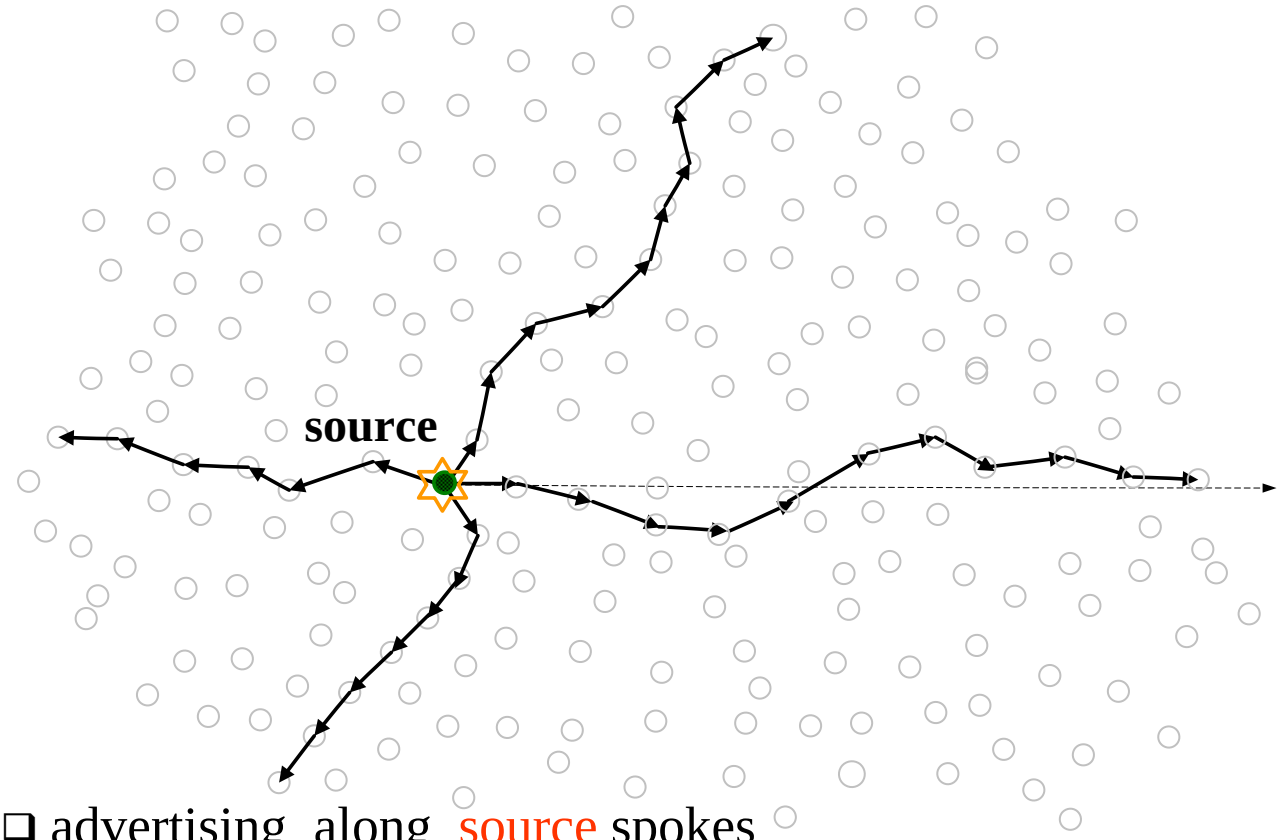
location-unaware sensor nodes **randomly scattered** in a plane

Sensor Network Example



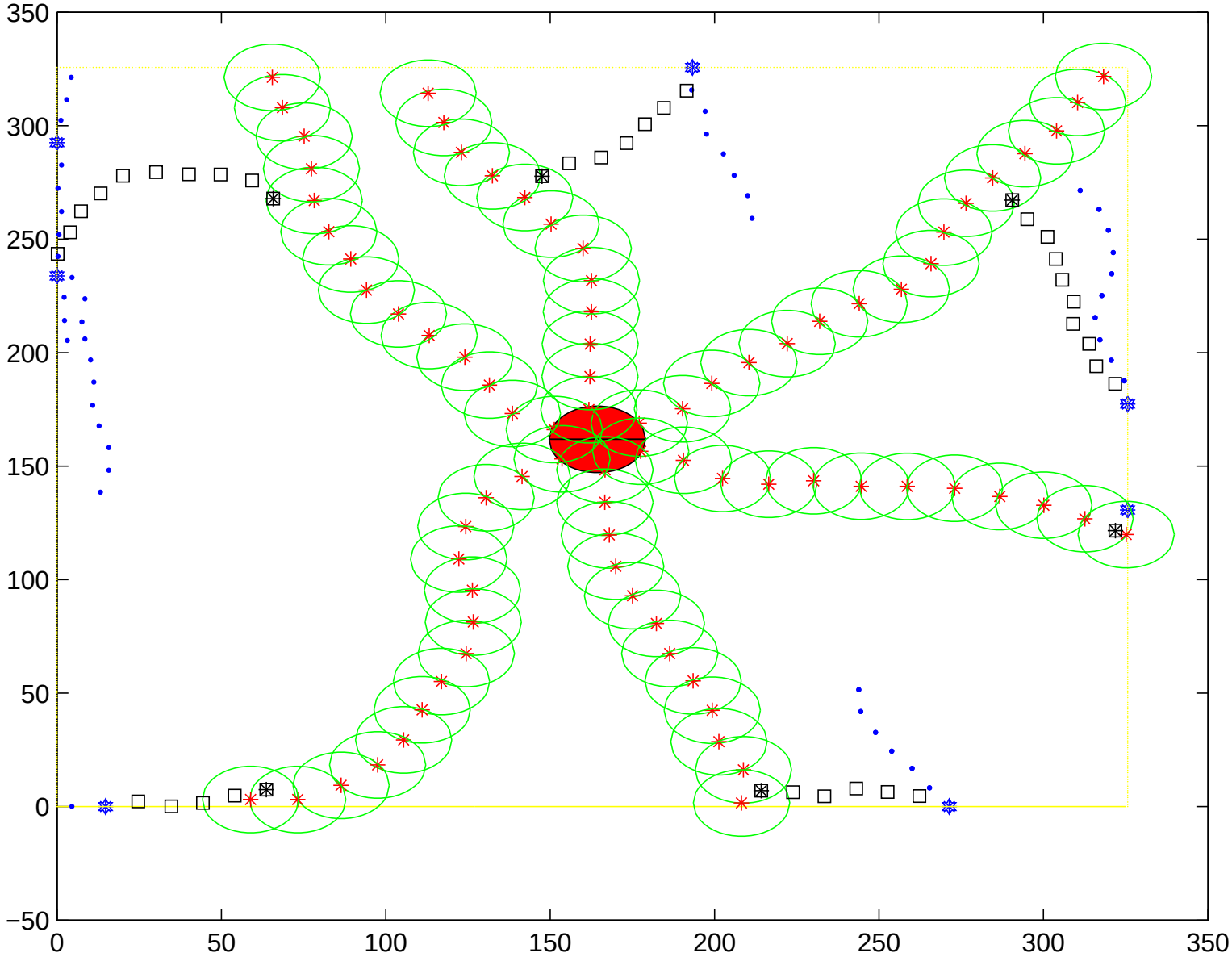
location-unaware sensor nodes **randomly scattered** in a plane

Data Dissemination

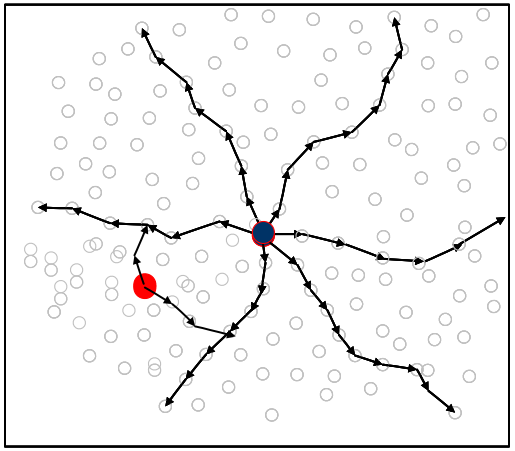


- advertising along source spokes
  increases the likelihood of information discovery
- avoiding flooding-based data publishing
  no redundant transmissions (broadcast storm)

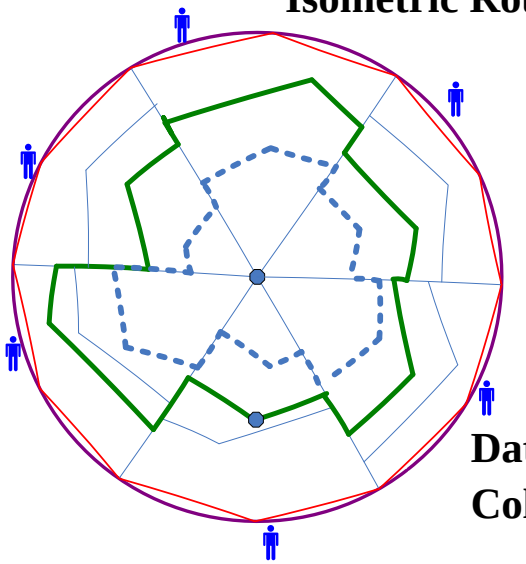
Simulated Dissemination Scenario



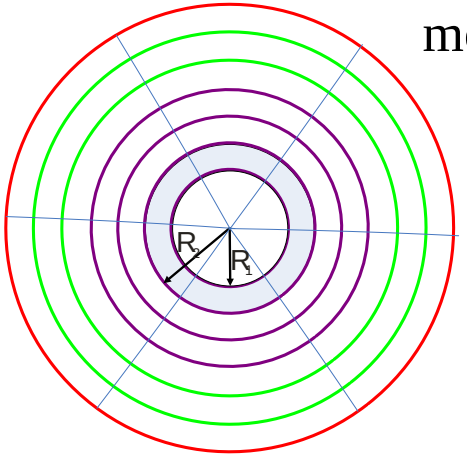
Infrastructure

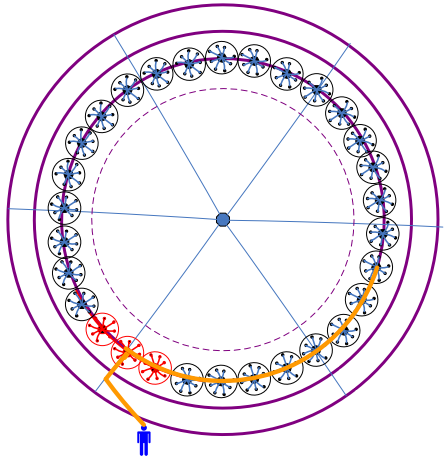


building



modeling



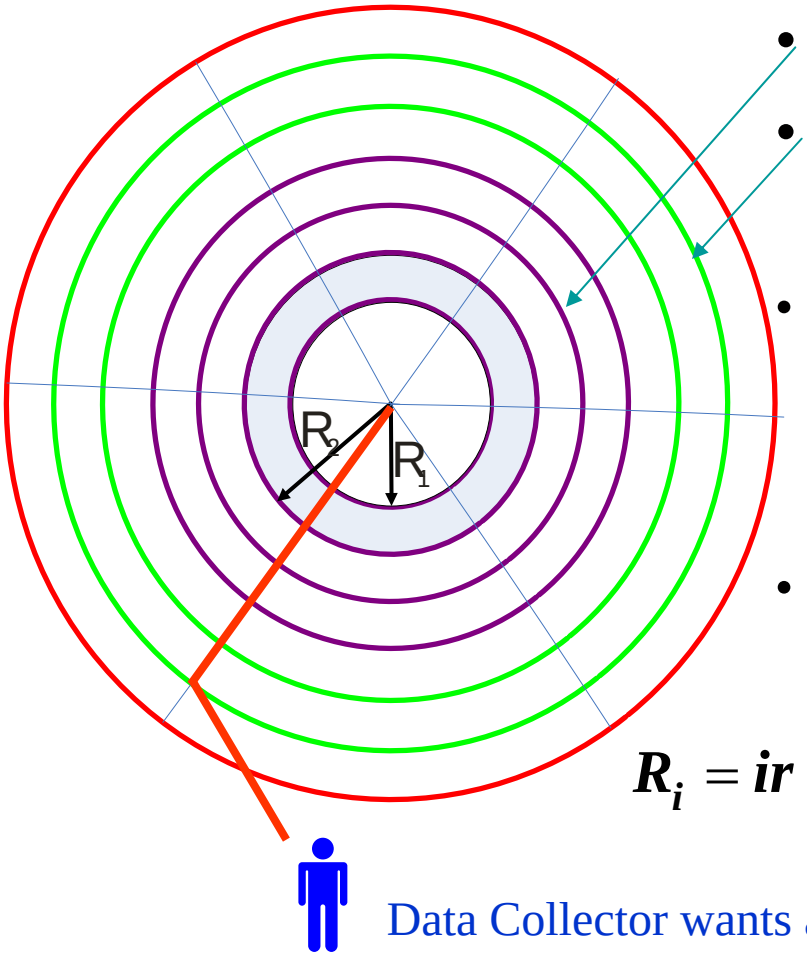


- **coding for distributed data storage**
- **decoding strategies for data collection**

Isometric Networks

network partitioned into sub-networks that are customized to handle network storage task according to the number of associated sources

$$E[k_i] = r^2 \pi \lambda_s (2i - 1)$$



$$R_i = ir$$

- Light Isometric Networks
- Heavy Isometric Networks

- infrastructure developed as a side effect of search for specific data items
  - use for storage of network data, through network-network-coding based methods
- inspired by the current work on network coding fo coding for storage in WSN

Data Collector wants a snapshot of network data

Current Work in Network Coding for Data Storage in WSN

Two basic approaches:

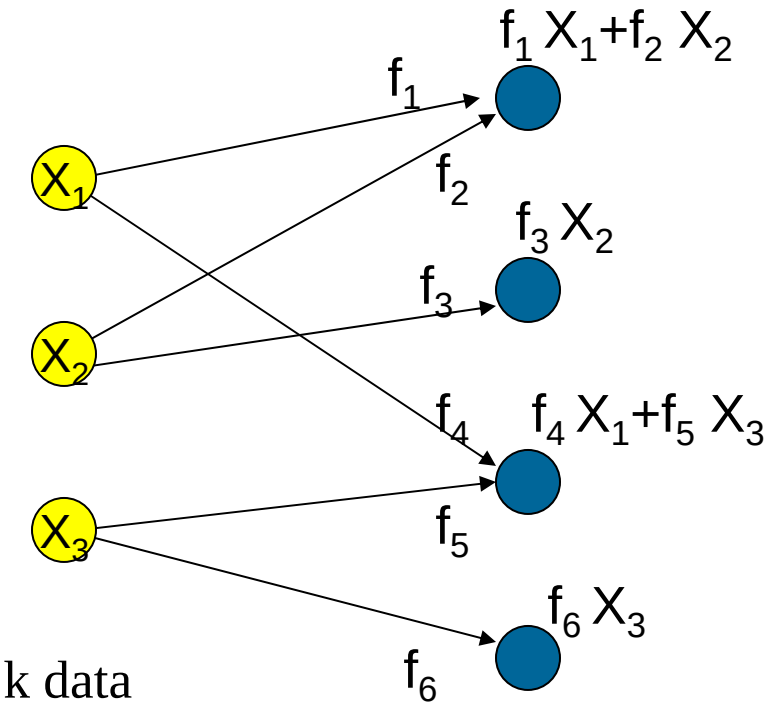
- Decentralized erasure codes [1]
  - encode k symbols into codewords of length n, which can be decoded from any subset of k symbols within the codeword
  - Decoding complexity: $O(k^3)$
- Decentralized fountain codes [2]
  - potentially infinitely many codewords (linear combinations of k data blocks); can be decoded from any k independent combinations
  - Decoding complexity: almost linear in k

- Abstract [1] or overly expensive [2] random routing techniques
- We propose structured approach to decrease cost

[1] Dimakis, Prabhakaran, Ramchandran. Decentralized erasure codes for distributed networked storage ('05/6)

[2] Liu, Liang, Li. Data persistence in large-scale sensor networks with decentralized fountain codes ('07)

Decentralized Erasure Codes



$$[Y_1 \quad Y_2 \quad Y_3 \quad Y_4] = [X_1 \quad X_2 \quad X_3] \underbrace{\begin{bmatrix} f_1 & 0 & f_4 & 0 \\ f_2 & f_3 & 0 & 0 \\ 0 & 0 & f_5 & f_6 \end{bmatrix}}_{G}$$

Want matrix as sparse as possible

(decreases dissemination cost)

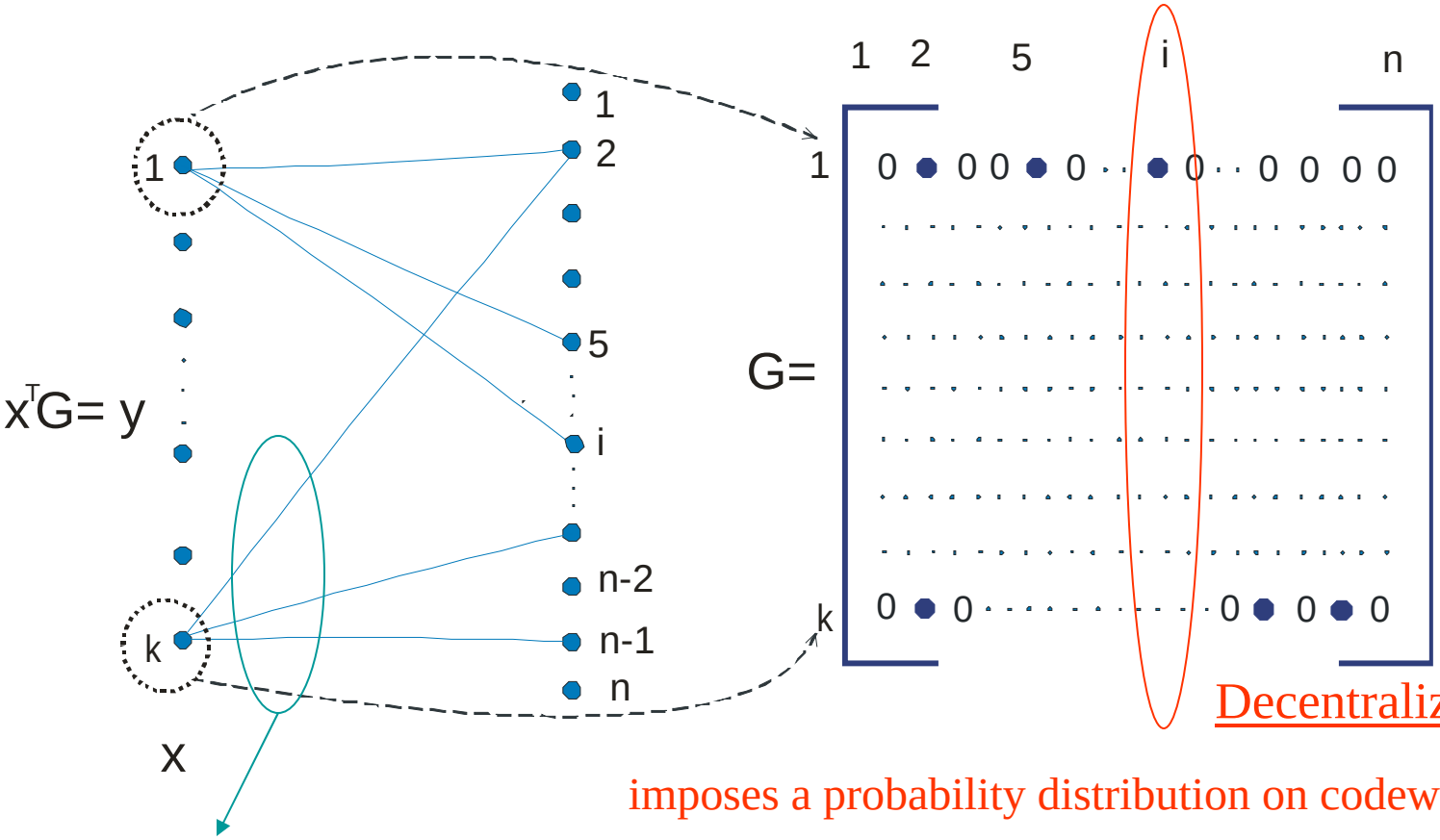
Now assume only storage nodes 1-3 are queried.

To reconstruct it suffices to have **G** to be full rank

How to build a code if your data is not in one place?

Basic Coding Approach: Random Linear Coding

Node i holds a codeword of degree d equal
to the number of non-zero entries in this column



Decentralized Fountain:

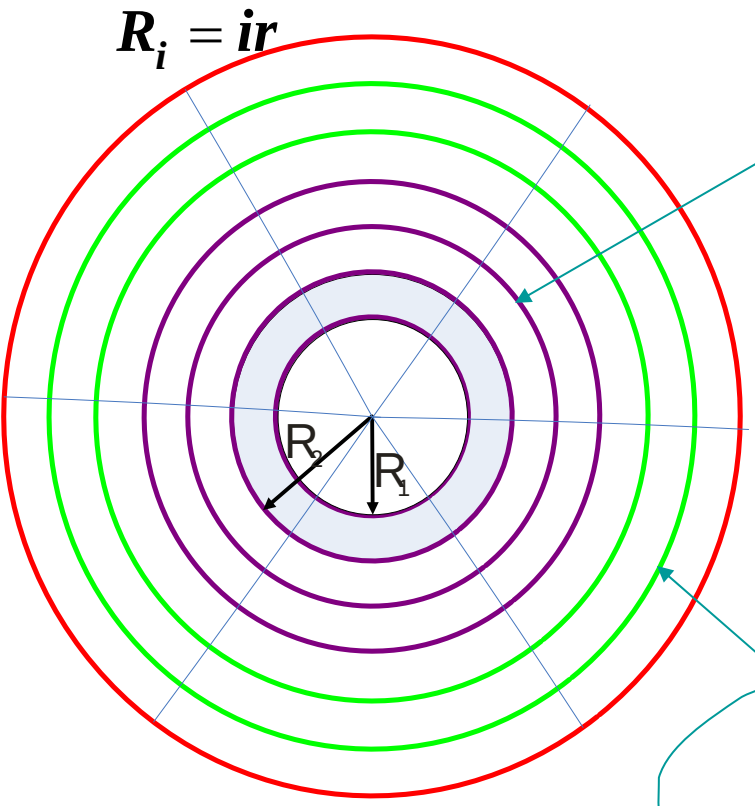
imposes a probability distribution on codeword degrees

Both approaches:

a certain number of packet replicas to be randomly diffused from independent independent sources and stored an random nodes (matrix rows)

Isometric Networks

network coding selected according to the number of associated sources



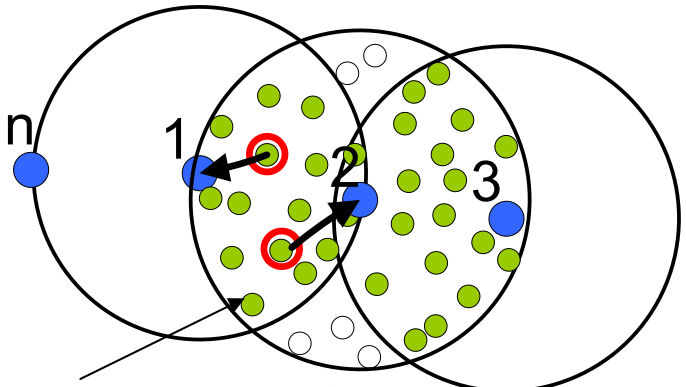
- Light Isometric Networks:
  - Random Linear Codes
- Decoding complexity for L light networks:

$$\sum_{i=1}^{L} k_i^3 \qquad \sum_{i=1}^{L} k_i^3 < \left(\sum_{i=1}^{L} k_i\right)^3$$

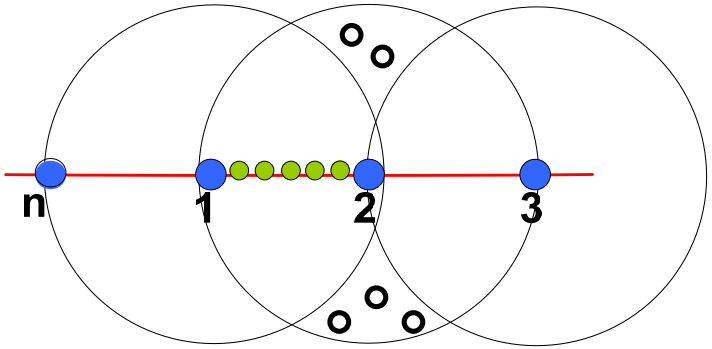
- Heavy Isometric Networks:
  - Decentralized Fountain Codes
- Decoding complexity for H heavy networks:

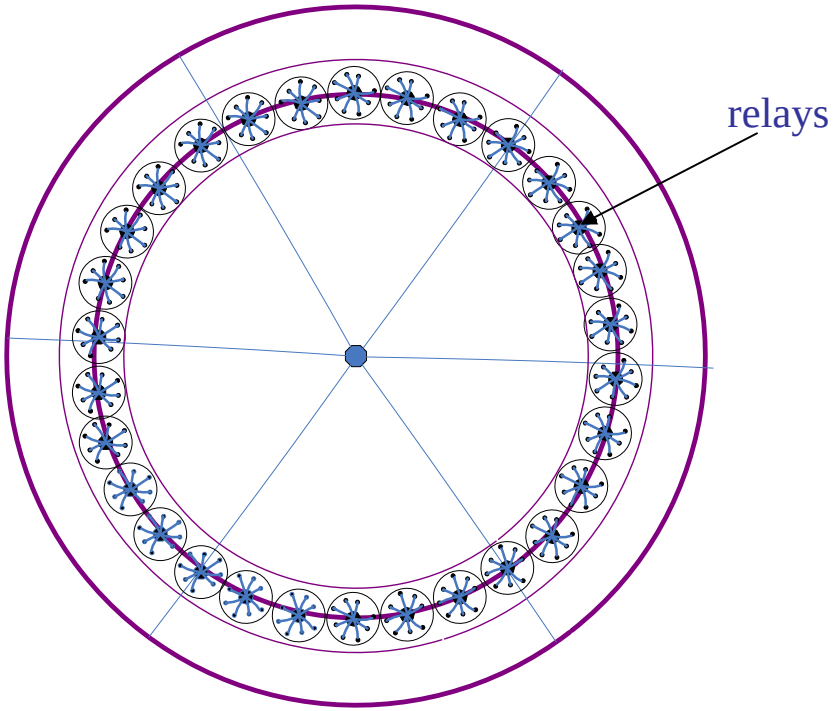
$$\sum_{i=1}^{H} k_i$$

Dissemination and Storage
Relaying and Overhearing/Combining



Storage squad: set of nodes in the range of relay



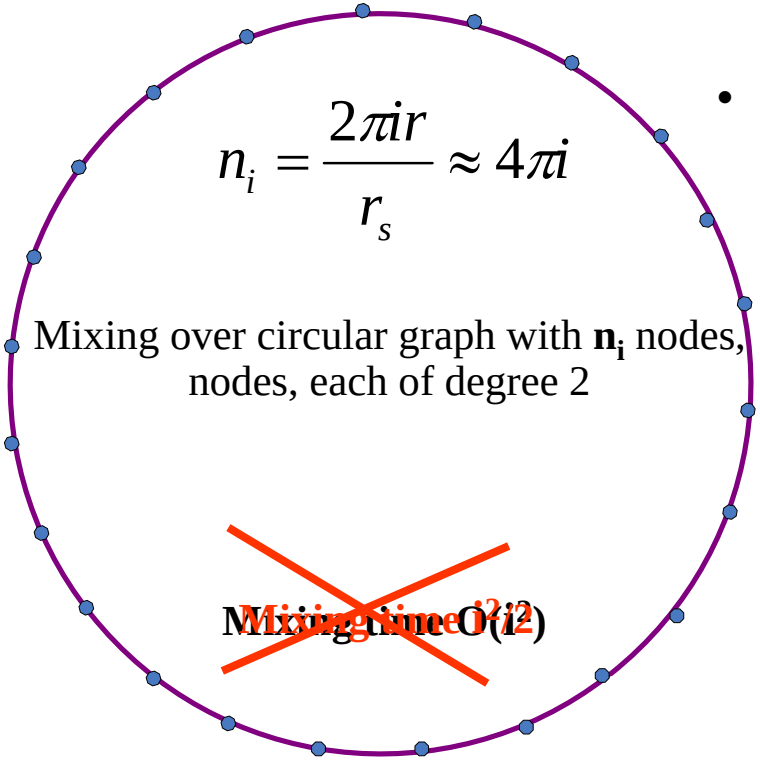


- Sources associate with one of the closest relays
- Relays disseminate (mix data)
- Squad nodes overhear, combine and store data

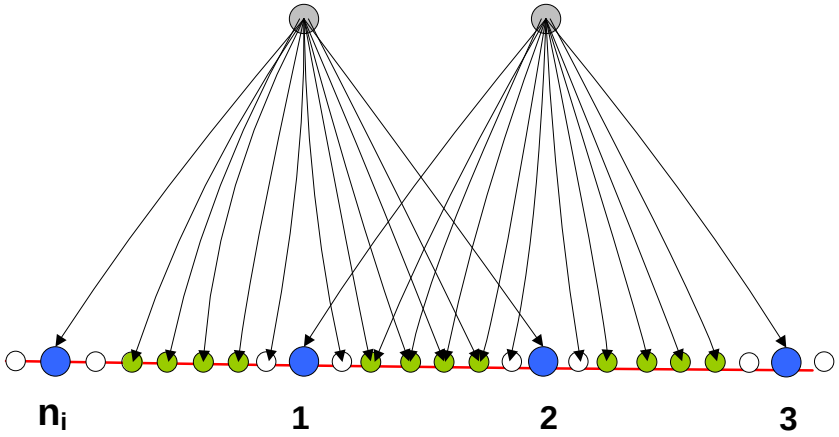
Let us assume that

- there is one source per relay (i.e. per squad)
- number of sources (relays) n larger than squad size h

Storage and Dissemination Graphs



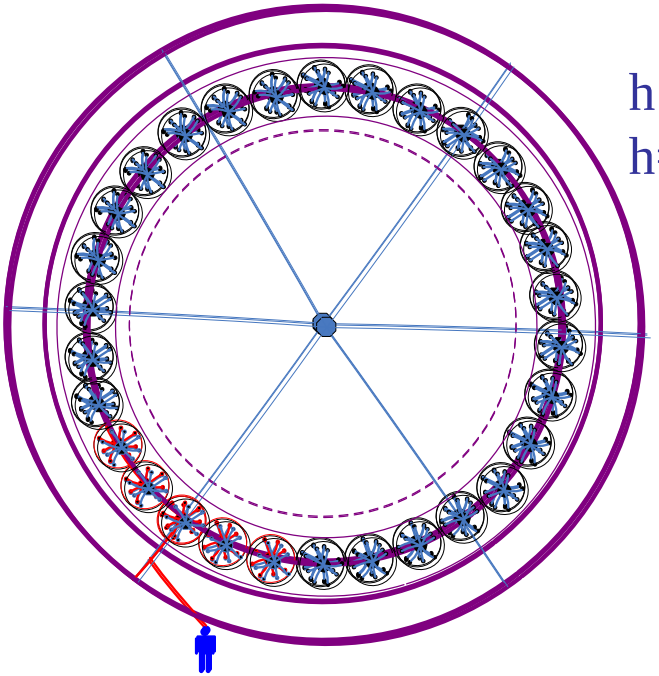
- high energy cost of data dissemination



Super-Squads: sets of adjacent squads

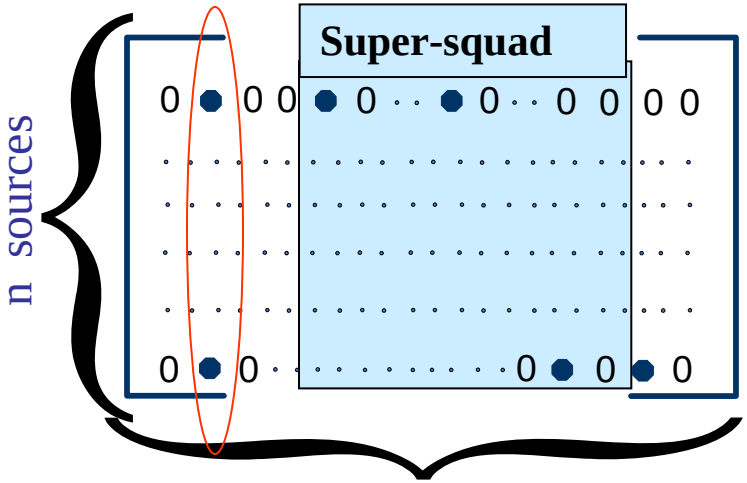
- Circular dissemination with network coding in the context of "wireless multicast multicast advantage"
- Apply network coding for storage in squad nodes that overhear 2 relays

Storage in Isometric Networks



h squad nodes

$h=O(r_s^2)$, $h<n$



hn storage nodes

Random Matrix created by Storage Protocol

MDC collects from SUPER-SQUADS

Storage Protocol controls degree distribution of codewords

large number of sources: storage with linear decoding complexity needed

COLLECTION

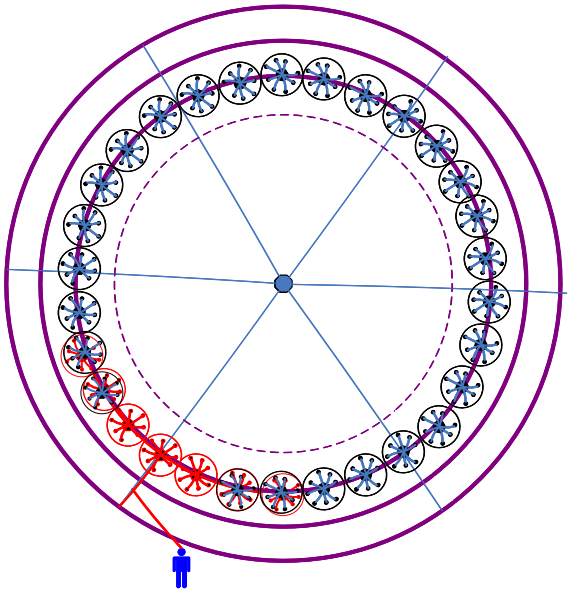
Assumption:

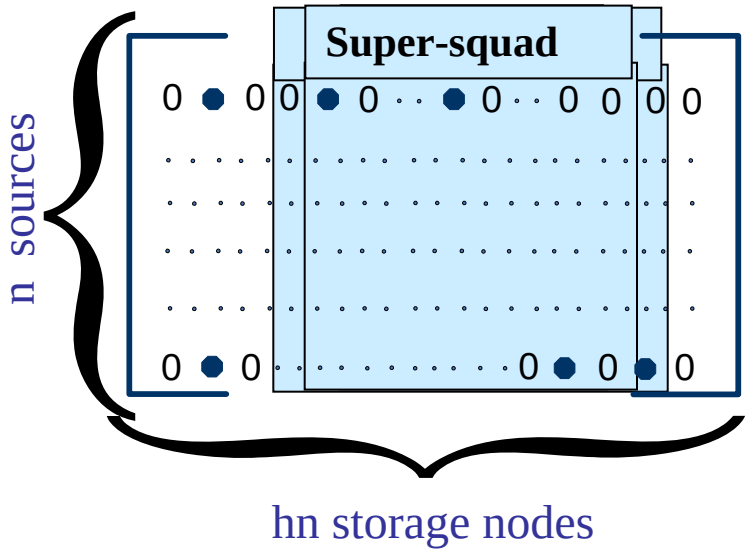
a mobile data collector (MDC) will establish connection with a random relay

Goal:

to have all data of the isometric network available in the vicinity of selected relay

Collection Strategies





Or collect a large number where n independent equations exist with high probability?

Is this matrix invertable?

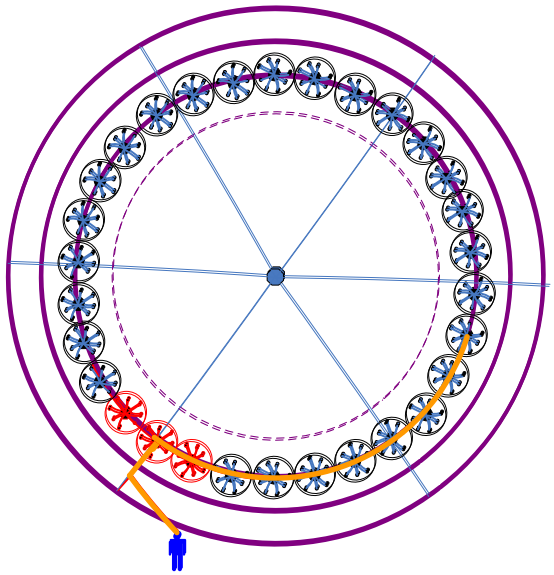
A large collection that guarantees decoding (whp) costs a lot: collection energy constraint

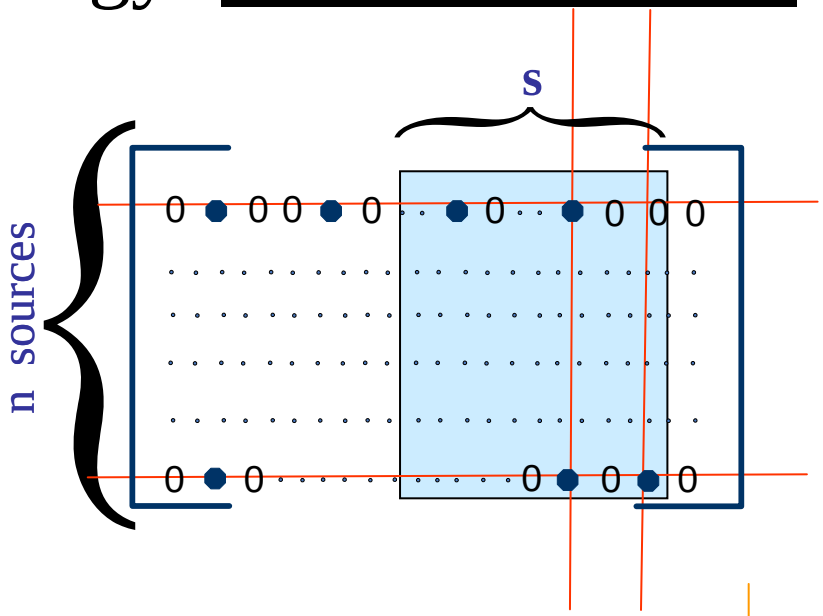
---

TRADE-OFF in collection strategy

- Up-front collecting: Collect small super-squad of code symbols locally
  - **fits the energy budget, but insufficient**
- On-demand collecting : collect selected code symbols
  - **likely to cost more per symbol, but few of them needed during decoding process**

Efficient Collection Strategy: Push-Pull Model

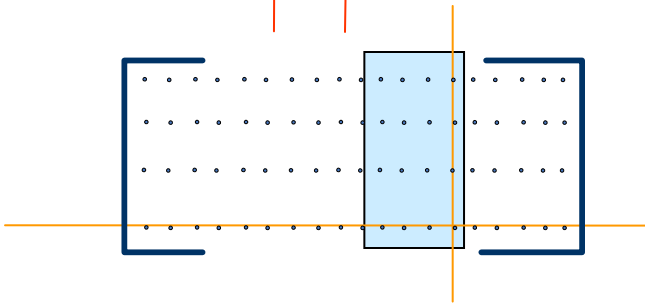




Push:

The closest super-squad of size **s** sends coded packets

- Enough to decode partially (belief propagation decoder)



Doping Cost: $\frac{s+d-n}{n}$

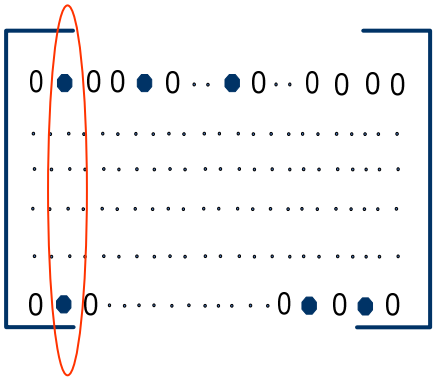
Pull:

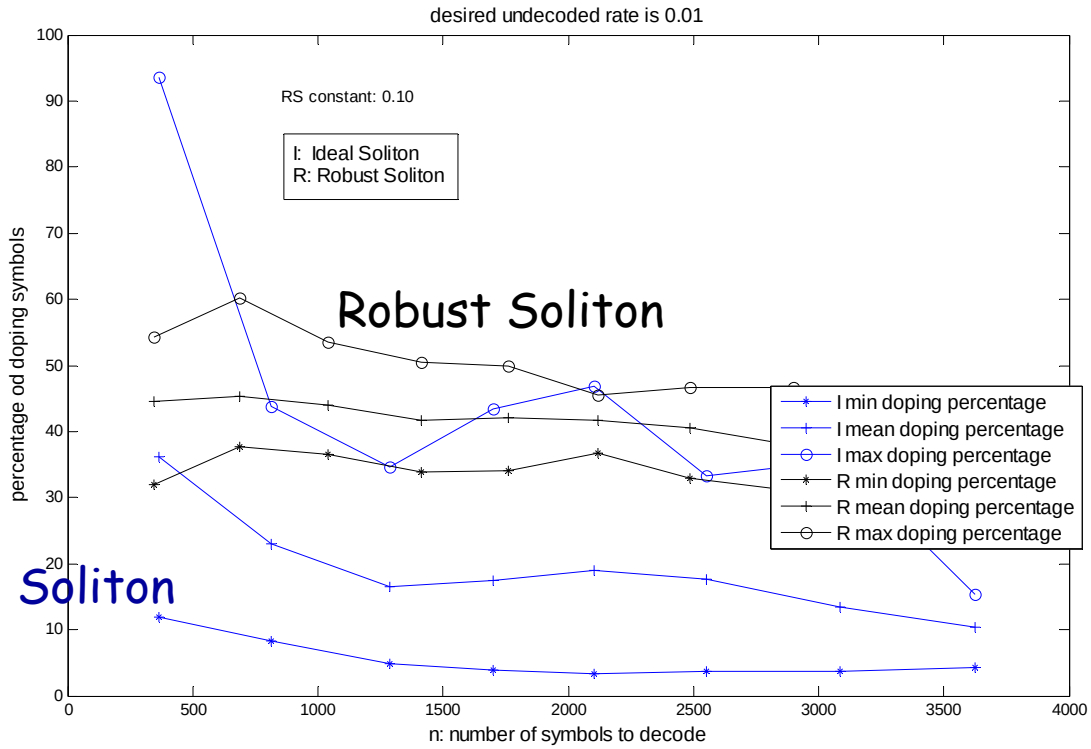
Query for **d** code symbols which can continue belief-propagation decoding (decoder doping)

Random Fountain Encoding of Network Data

Doping Cost Depends on Degree Distribution

How to select degree distribution to collect efficiently?

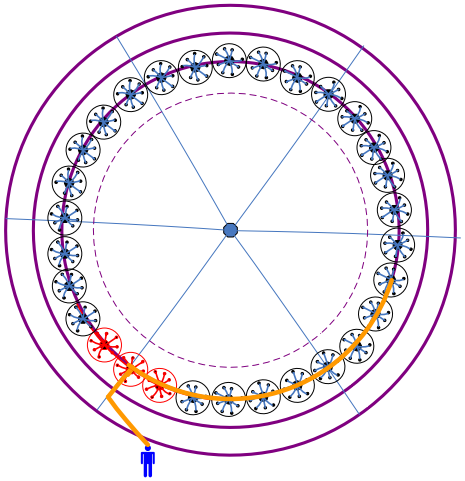


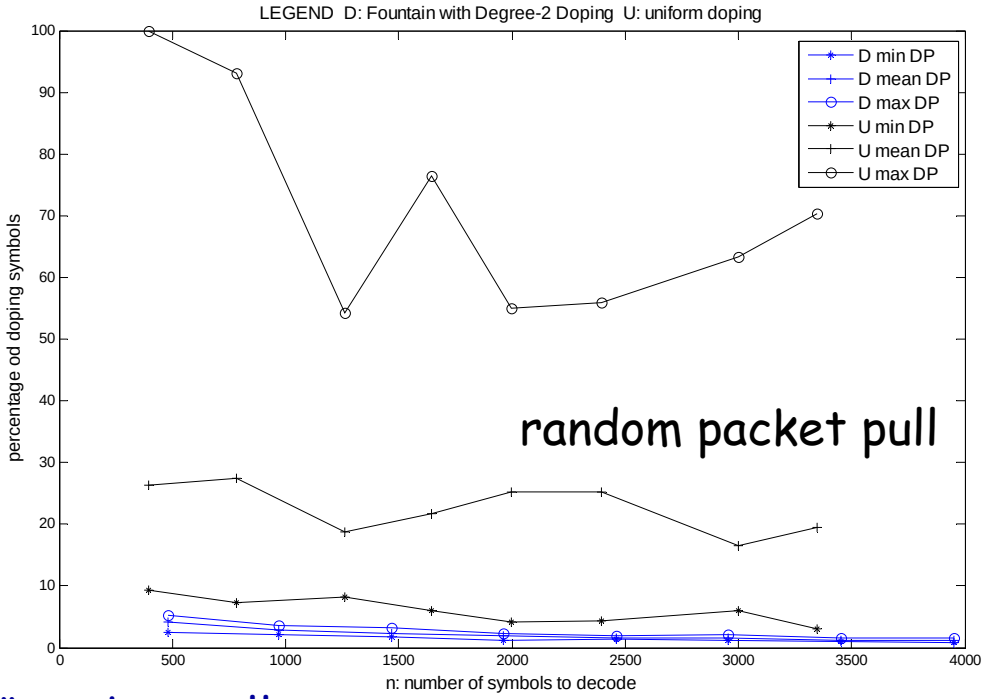


Random-access Push-Pull Data Collection and Decoding

Doping Cost Depends on Doping Mechanism

How many coded packets do I need to pull to decode all data?





"smart" packet pull

Thank You.