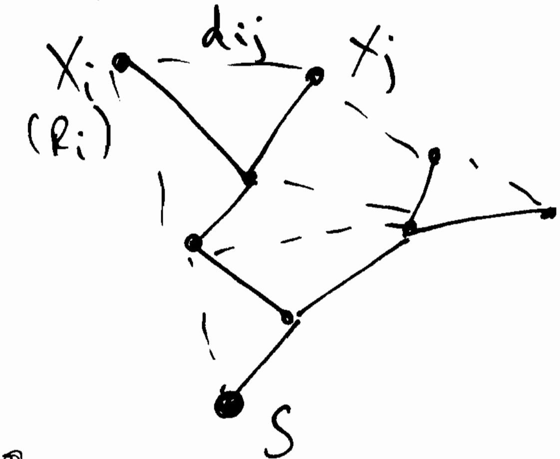
Trees for correlated data gathering

1. Defining Data gathering

- Graph (V, E) connected



- network of nodes with a connectivity graph.

- Sink S needs to reconstruct data at nodes $1 \ldots N$. (a snapshot)
  - Data correlated => use aggregation (compression)

Assumptions

- no capacities on links
- unsplittable flows

Data field X_i $X_1 \ldots X_N$

- temporally i.i.d.
- spatially correlated ; finely quantized (so discrete r.v.)

e.g. Gaussian: $\cancel{X} \sim \mathcal{N}(0, K)$ K – correlation matrix

$$K_{ij} = \exp(-\alpha d_{ij}^{E\,\beta}) \quad \beta \in \{1, 2\}$$

Link weights

$$d_{ij} = f(d_{ij}^E) \qquad f(\cdot) = \begin{cases} \cdot\,^2 \\ \cdot\,^4 \\ \exp(\cdot) \end{cases}$$

<u>Task</u>: – find rate allocation
– Spanning tree T

s.th. to minimize

$\boxed{\text{i.i.d.} \Rightarrow \text{SPT}}$

$$\left(\{ R_i^* \}_{i=1}^{N}, T^* \right) = \arg\min_{\{R_i, T\}} \sum_i R_i \cdot d_T(i, S)$$

– u.c. Sink is able to reconstruct data
(Power consumption)

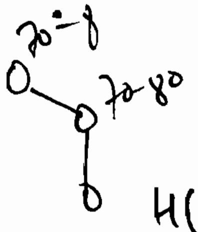
<u>Two approaches</u>:

– Explicit communication (EC)
– only use data for compression "as it comes"
– Slepian-Wolf (no need of inter-node comm.)

---

1. EC.

Approximation: – distance dependent correlation ⇒
→ condition on only 1 incoming node



– conditioning is cst. irrespective of distance
– packets cannot be regen

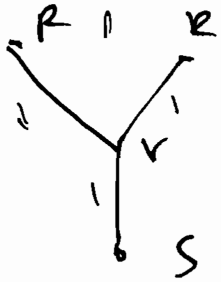
$H(X_i) = R$ if no side info
$H(X_i | X_j, \dots) = H(X_i | X_j) = r < R$ if side info <u>is</u> available

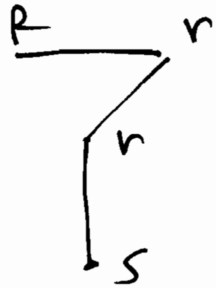
$\rho = 1 - \frac{r}{R}$ – corr. coefficient

$r = 0 \Rightarrow \rho = 1$, <u>very</u> correlated
$r = R \Rightarrow \rho = 0$, data <u>independent</u>

Examples

1.

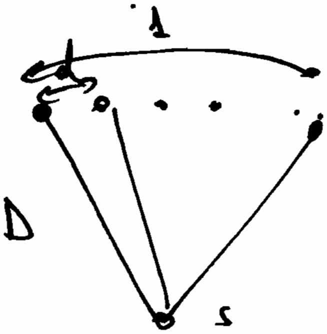




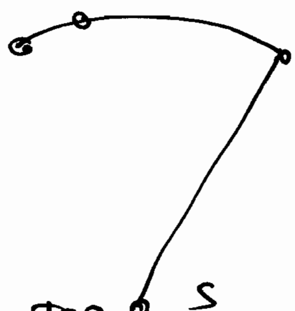
$4R+r$ (SPT) $\quad$ $3R+3r$ (TSP)

$R > 2r \Rightarrow$ TSP is better

2.



SPT



TSP

$$\lim_{N\to\infty} \frac{\text{cost}(TSP)}{\text{cost}(SPT)} = (1-\rho)\left(\frac{1}{2D}-1\right)$$

Problem

$$T = \underset{(L \text{ leaves, } I \text{ in-nodes})}{\arg\min} \left(r \sum_{j\in I} d_T(j,s) + R \sum_{\ell \in L} d_T(\ell,s) \right)$$

or

$$T = \underset{L}{\arg\min} \left((1-\rho) \sum_{i\in V} d_T(i,s) + \rho \sum_{\ell\in L} d_T(\ell,s) \right)$$

$\rho = 0 \Rightarrow$ SPT

$\rho = 1 \Rightarrow$ K-TSP

P

NP

$0 < \rho \leq 1$?

NP-hard

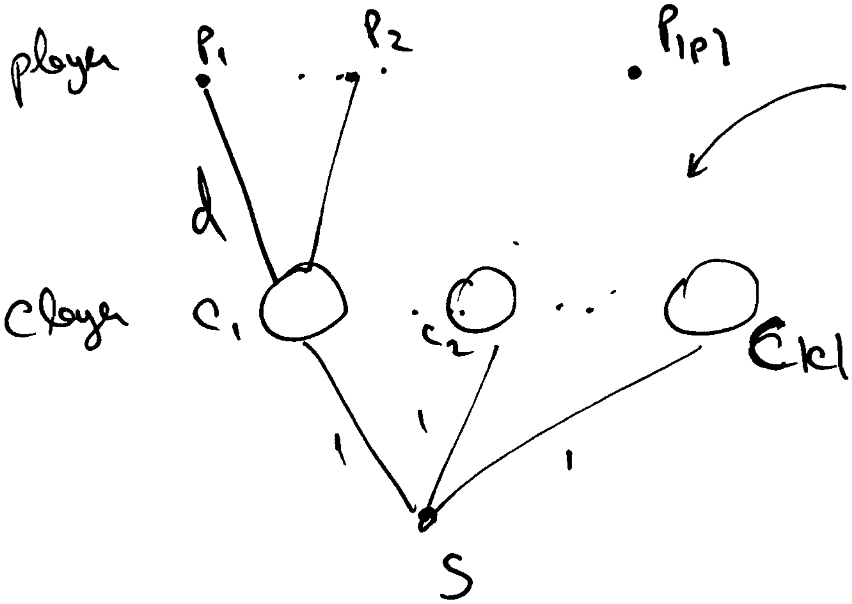
decision is NP complete

reduction from set cover

For every set-cover instance, we build a tree instance where solving the tree $< M$ $\Leftrightarrow$ solving the set cover problem $|C'| = k$

$R > 1$

$r = 1$

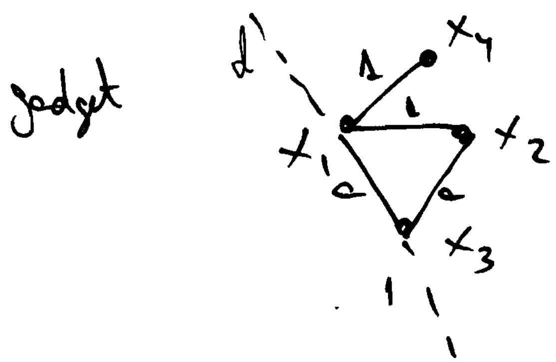


Given k (elements of subcollection C')

cost (Tree) $\leq M$ $\Leftrightarrow$

$\Leftrightarrow$ find $|C'| \leq k$

$M = f(k, ?, R, |P|, |C|)$



- d large, so going back to level P is bad $\Rightarrow$
  $\Rightarrow$ exactly $|P|$ links P to C.

Obs

- All x_3's must be in-tree nodes
- degree of freedom → 2 out of 3 links

Proof idea : Given ~~K~~ k, choose $d, ? \geq M$ s.th. tree with cost $\leq M$ means connecting to at most k nodes of C.

• better than V only if no flow
  (< is always worse)

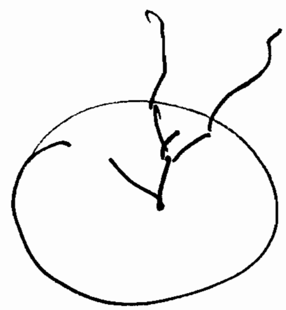
if flow incoming at x_s, V is better than 7

- cost for P nodes always the same
- only cost from C matters ⇒ choose as few ~~little~~ Vs as possible

---

heuristic
SPT/TSP

$\boxed{\text{Figure}}$



radius ↑ with r↑

Approximation algorithm

Lemma $C_{opt} \geq \max\left(\underbrace{r \cdot \text{cost}(SPT)}_{\beta} \;\&\; \underbrace{R \,\text{cost}(MST)}_{\alpha} \right)$

$$\sum_i d_{SPT}(i, s)$$

Proof – each node has to send at least r to the sink

$$C_{opt} \geq r \,\text{cost}(SPT) - \sum_i \ldots$$

– R has to get everywhere, including the sink

$$C_{opt} \geq R \cdot \text{cost}(MST)$$

Shallow-light tree (SLT) – see proof
(Khuller, Raghavachari, Young '93)

figure

Properties

choose $\gamma > 0$

$\text{cost}(SLT) \leq (1 + \sqrt{2}\,\gamma) \,\text{cost}(MST)$ – done on pre-order, thus total additions $< \gamma \,\text{cost}(MST)$

$d_{SLT}(x, s) \leq \left(1 + \frac{\sqrt{2}}{\gamma}\right) d_{SPT}(x, s)$ – by constr.

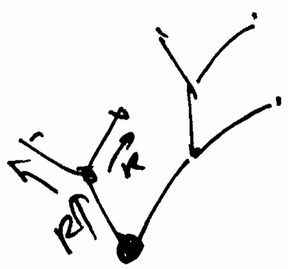
SLT construction

Take $\delta = \frac{\sqrt{2}}{\gamma} + 1$

- Build MST; Initialize $H = MST$.
- Compute its pre-order, starting with s
- Compute $d_{SPT}(s, v)$, $(\forall) v$
- for v in pre-order, if $d_H(s, v) > \delta \, d_{SPT}(s, v)$ then
  - add to H all edges of $SP(s, v)$
- output SPT of H

Back to our problem

Use a broadcast/echo procedure on the SLT
(Note: this is not exactly what we need; but best we could come up with)



Theorem B/E is a $2(1+\sqrt{2})$ approximation of C_{opt}

Proof

$$C_{SLT} = R \cdot \mathrm{cost}(SLT) + \sum_{v \in V} r \cdot d_{SLT}(v,s) \le$$

$\uparrow$ depends on r

$$\le R(1+\sqrt{2}\,\gamma)\,\mathrm{cost}(MST) + r\left(1+\frac{\sqrt{2}}{\gamma}\right)\mathrm{cost}(SPT) =$$

$$\mathrm{cost}(SPT) = \sum_{v} d_{SPT}(v,s)$$

$$\overset{\gamma=1}{\le} \Big[\underbrace{R\,\mathrm{cost}(MST)}_{\alpha} + \underbrace{r\,\mathrm{cost}(SPT)}_{\beta}\Big](1+\sqrt{2})$$

So

$$C_{SLT} \le (1+\sqrt{2}) \cdot 2\max(\alpha,\beta)$$

$$C_{opt} \ge \max(\alpha,\beta) \quad \Rightarrow \quad \frac{C_{SLT}}{C_{opt}} \le 2(1+\sqrt{2})$$

2. S-W coding

- no need of explicit inter-node comm. to code with joint entropy (compressed)

⇒ Obs. Given ~~some~~ any amount of data R_i at a node V, optimum is SP(V, s)

Pb. $$\min_{T, R_i} \sum_i R_i \cdot d_T(i,s)$$

u. constr.

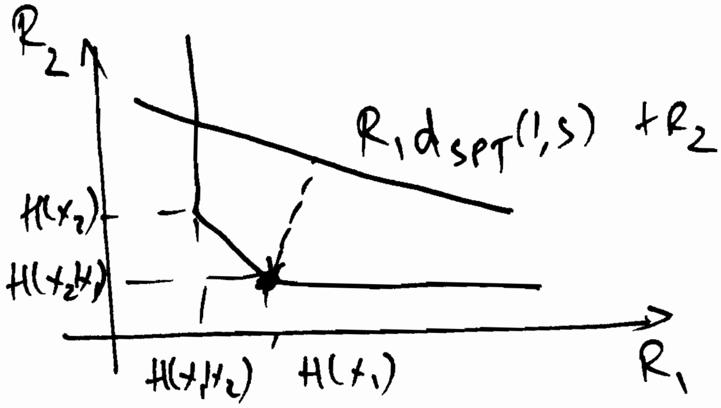
$$\sum_{j \in \mathcal{X}} R_j \geq H(\mathcal{X} \mid \mathcal{X}^c) \quad , (\forall)\ \mathcal{X} \subset V$$

1. T = SPT

2. $$\min_{R_i} \sum_i R_i d_{SPT}(i,s)$$

u. constr. SW

N=2 $\quad d_{SPT}(1,s) < d_{SPT}(2,s)$



order $d_{SPT}(1,s) \leq \dots \leq d_{SPT}(N,s) \Rightarrow$

$\Rightarrow R_1 = H(X_1)$

$\vdots$

$R_N = H(X_N \mid X_{N-1} \dots X_1)$