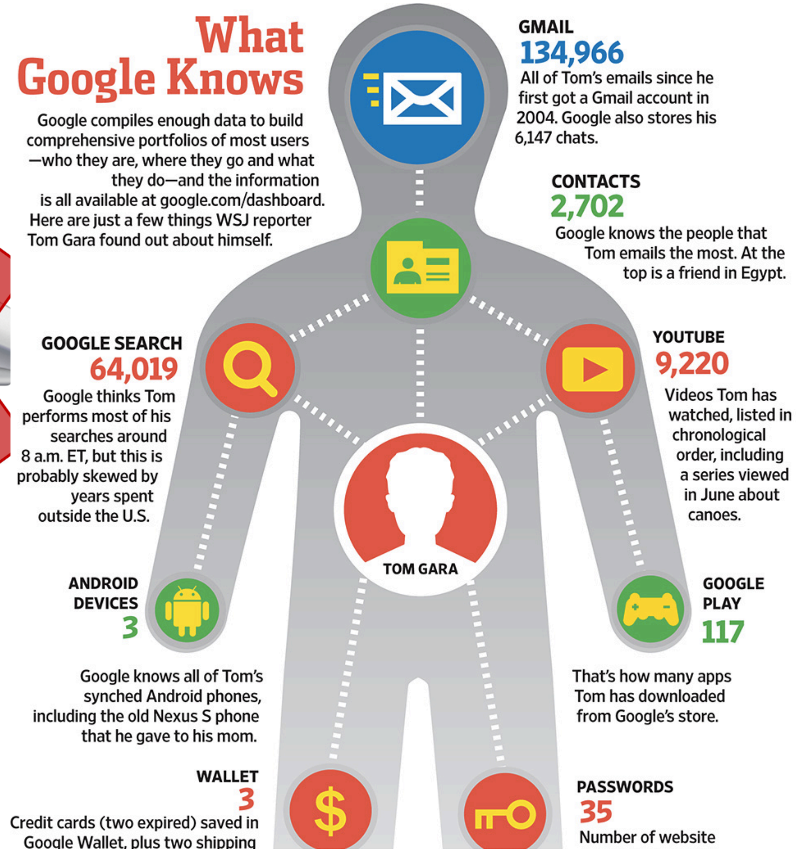
Private Information Retrieval from MDS Coded Data in Distributed Storage Systems

Salim El Rouayheb
ECE Department
Illinois Institute of Technology

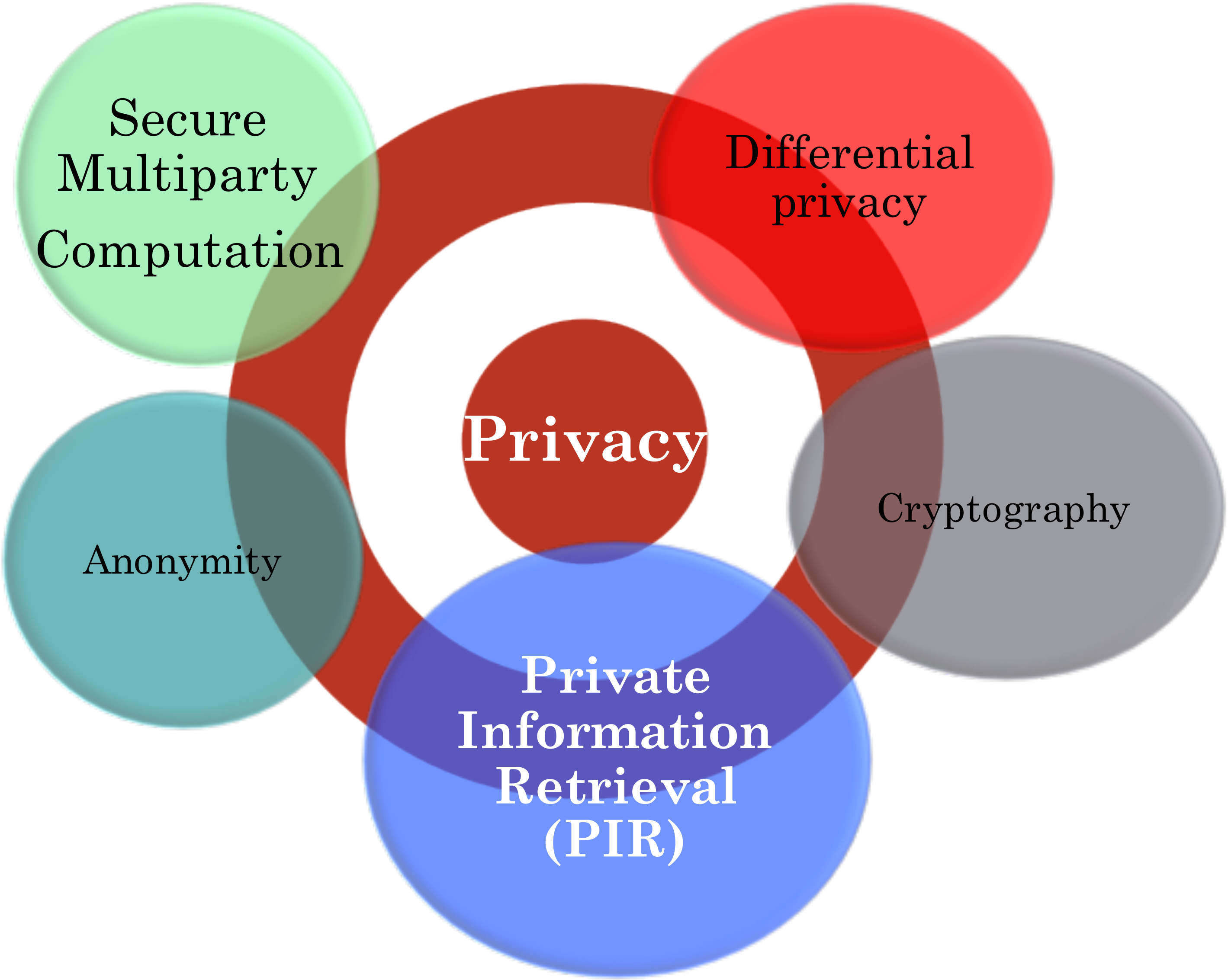
Joint work with Razane Tajeddine

Motivation



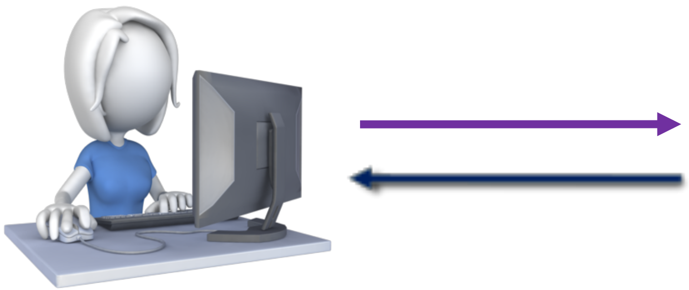
Apple Touts 'Differential Privacy' Data Gathering Technique in iOS 10

Tuesday June 14, 2016 4:02 AM PDT by Tim Hardwick

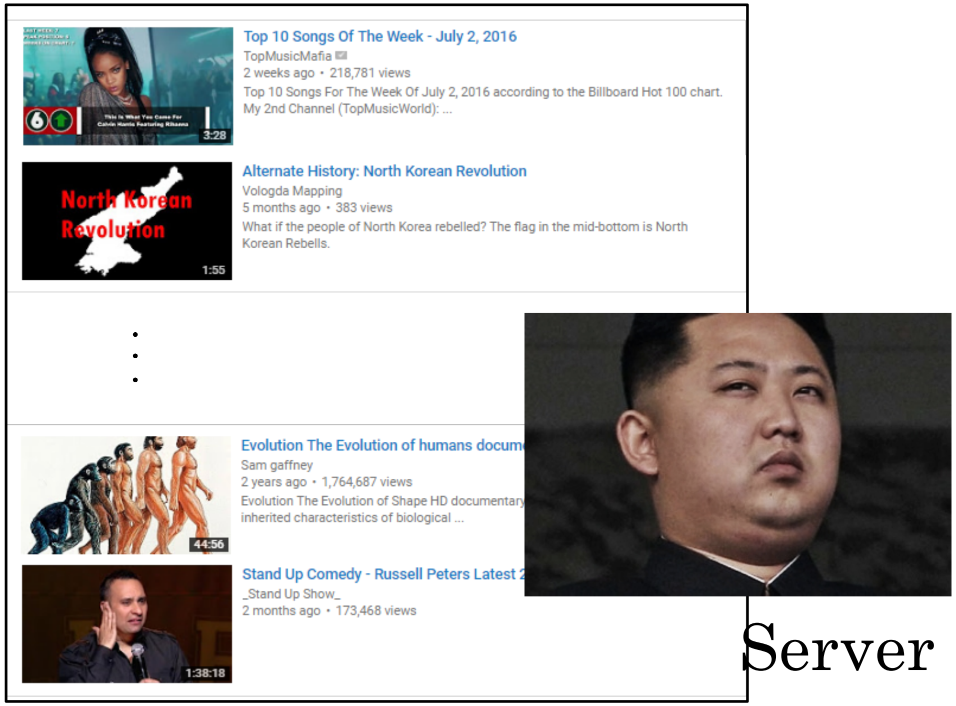

Secure Multiparty Computation
Differential privacy
Privacy
Anonymity
Cryptography
Private Information Retrieval (PIR)

Private Information Retrieval (PIR)

- User wants to retrieve a file from a database without revealing the identity of this file.
- This problem was first introduced by Chor et. al in 1995. [Chor, Goldreich, Kuchilevitz and Sudan '95].



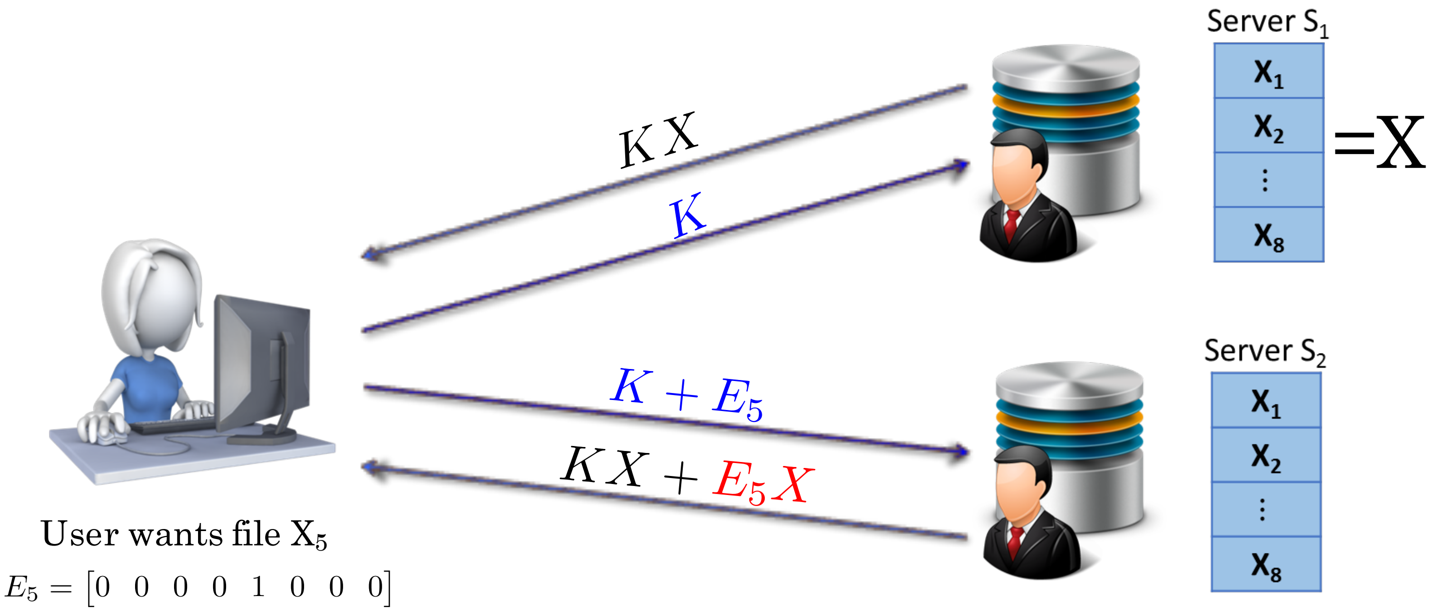
User wants to watch a video on North Korean Revolution



Server

Toy Example

- Database replicated on two non colluding servers.



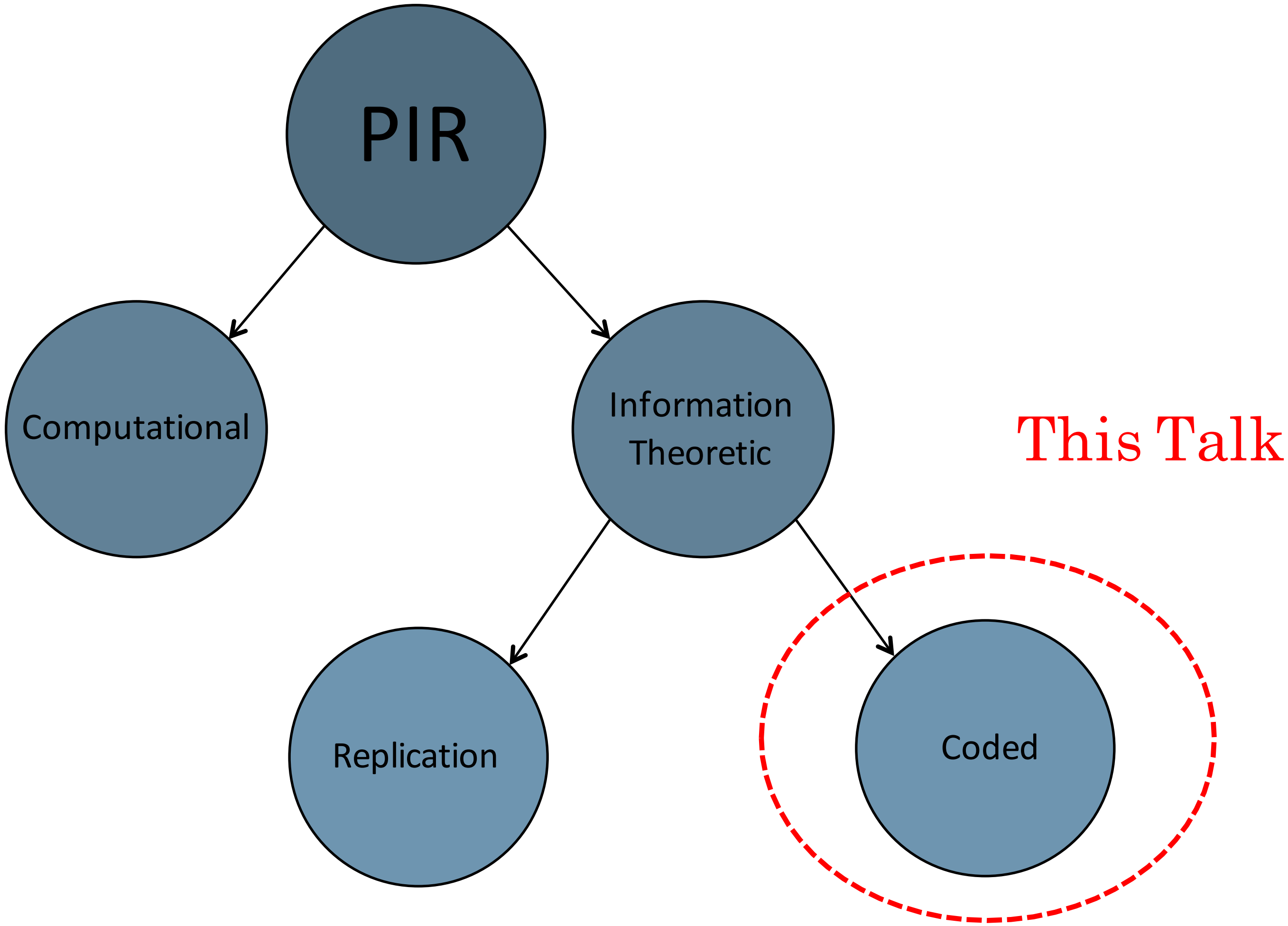
[Chor et al. ‘1995]

Computational vs. Info Theoretic Privacy

- Relaxation: **Computational PIR**

- Can achieve privacy on one server without downloading whole database. [Kushilevitz and Ostrovsky, '97], [Chor and Gilboa, '97], [Cachin, Micali, and Stadler, '99], …

- High computational complexity. [Sion and Carbunar, '07]


PIR
Computational
Information Theoretic
Replication
Coded
This Talk

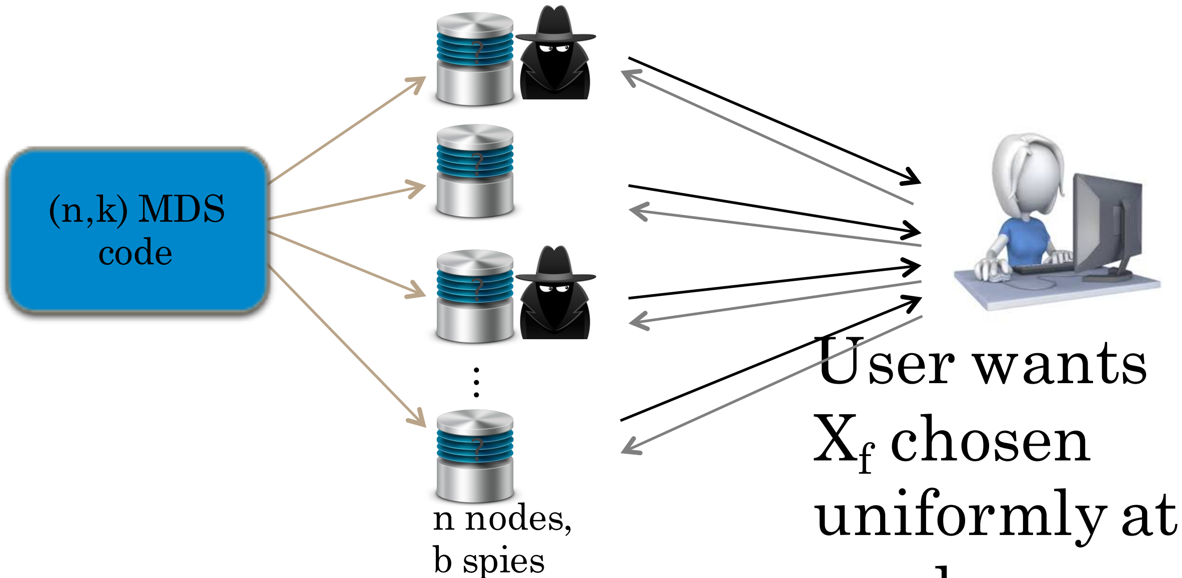
Model

- A distributed storage system with n nodes storing files $X_1, \ldots, X_m$
- (n,k) MDS code is given and not design parameter.
- b passive spy nodes

Goal: Design PIR scheme with min download cost

Perfect privacy:

$$H(f \mid \text{queries to } b \text{ servers}) = H(f)$$



User wants X_f chosen uniformly at random

Related work: Replicated Data

- PIR scheme on replicated non-colluding nodes with total, upload and download, communication cost of $\mathcal{O}((n^2 \log n) m^{1/n})$ and $\mathcal{O}(m^{1/3})$for the case when n=2 [Chor, Goldreich, Kuchilevitz and Sudan '95]

- PIR scheme on replicated non-colluding nodes with communication cost $\mathcal{O}(m^{1/(2n-1)})$ [Ambainis, '97] and $\mathcal{O}(m^{\frac{c \log \log n}{n \log n}})$ [Beimel et al, '02]
- PIR protocols with total communication cost that is subpolynomial in the size of the database [Yekhanin '08], [Efremenko '12] and [Dvir and Gopi '15]

- Fundamental limits and achievable schemes on ***download cost*** for replication. [Sun and Jafar '16]

Related Work on Coded PIR

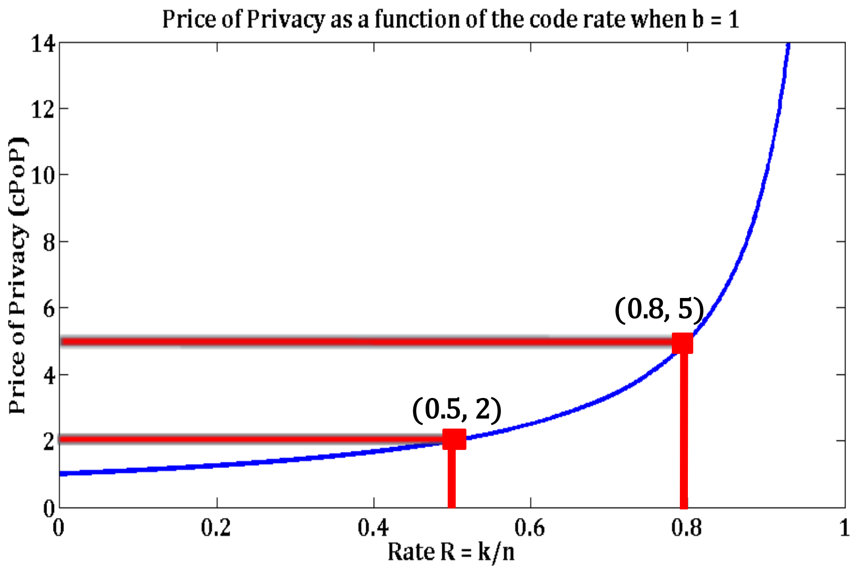
- Batch codes [Ishai et al. ‘04]
- One extra bit of download is sufficient to achieve PIR. [Shah, Rashmi, and Ramchandran, ISIT ’14]
- Methods for transforming a replication-based PIR scheme into a coded-based PIR scheme with the same communication cost [Fazeli, Vardy, and Yaakobi, ISIT ‘15]
- Bounds on the the tradeoff between storage and download communication cost. [Chan, Ho, and Yamamoto, ISIT ‘15]
- PIR array codes [Blackburn & Etzion ‘16]

Our Results: Single Spy

Theorem 1: Consider a DSS using an (n,k) MDS code over $GF(q)$, with $b=1$ spy node. Then, there is an explicit linear PIR scheme for *any number of files* m over $GF(q)$ with download communication cost (price of privacy):

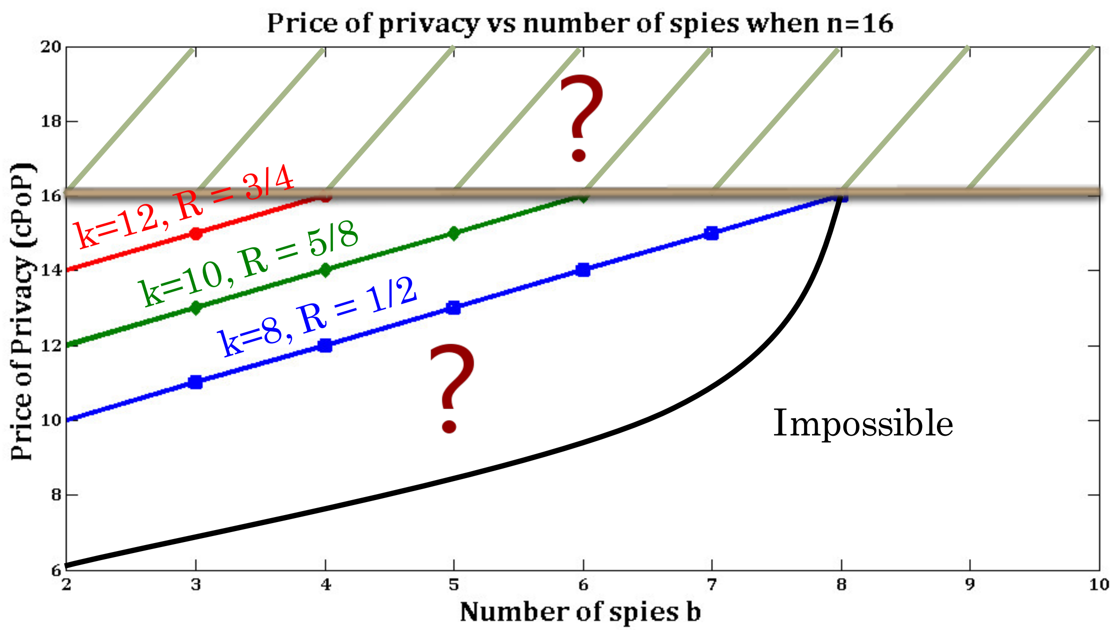
$$PoP = \frac{n}{n-k} = \frac{1}{1-R}.$$

- Achieves the information theoretic optimum given in [Chan et al, ISIT '15] for linear PIR scheme.
- Achieve the bound given in [Sun et. al, ISIT '16] when applying replication.*
- The PIR scheme is universal, i.e. does not depend on the MDS code.



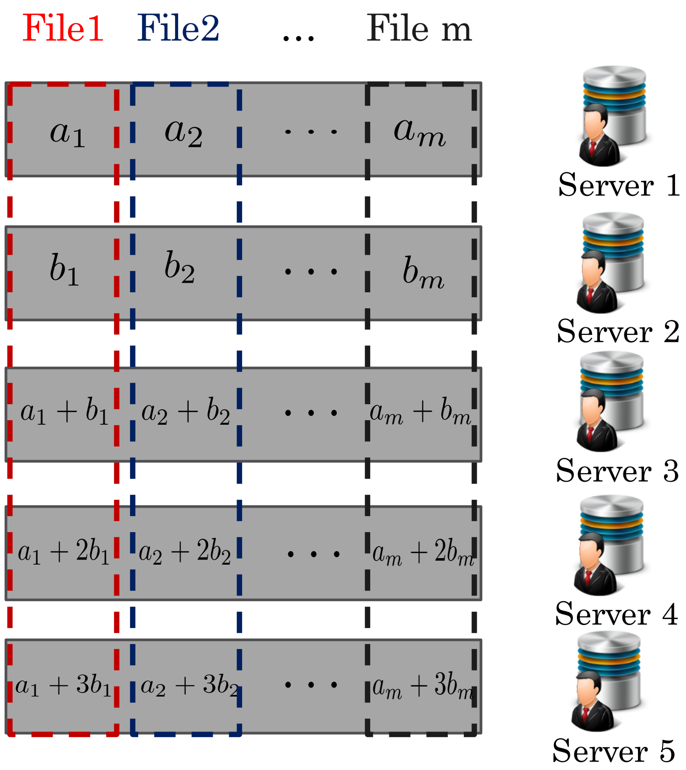
Our Results: Multiple Spies

Theorem 2: Consider a DSS using an (n,k) MDS code over $GF(q)$, with $b \leq n-k$ spy nodes. Then, there is an explicit linear PIR scheme over $GF(q)$ with download communication cost, $PoP = b + k$.



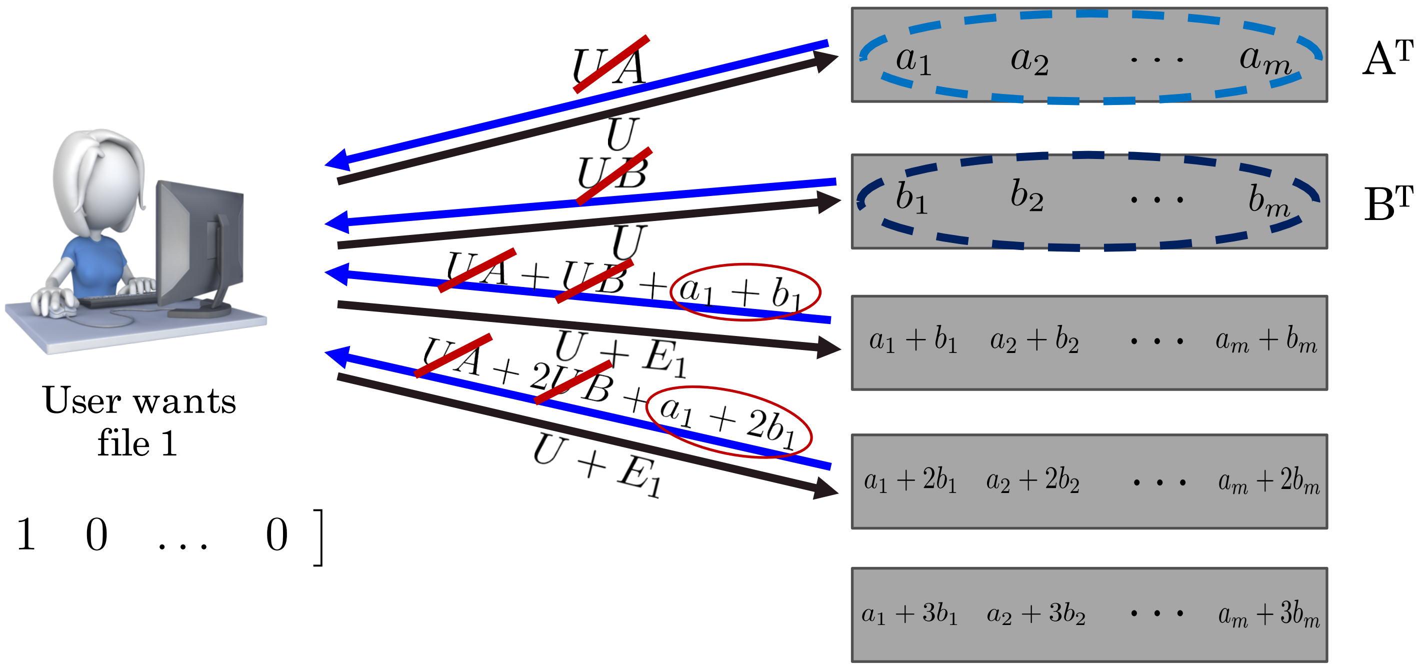
Example on Theorem 1

- Consider a (5,2) MDS code with $b = 1$ spy node.
- Goal is to achieve $PoP = \frac{n}{n-k} = \frac{5}{3}$.



First Attempt

- Generate an iid random vector $U = \begin{bmatrix} u_1 & u_2 & \dots & u_m \end{bmatrix}$.



$$E_1 = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}$$

- This achieves $PoP = 2$ instead of $PoP = \frac{5}{3}$.

Subdivision

- Divide each part into 3,

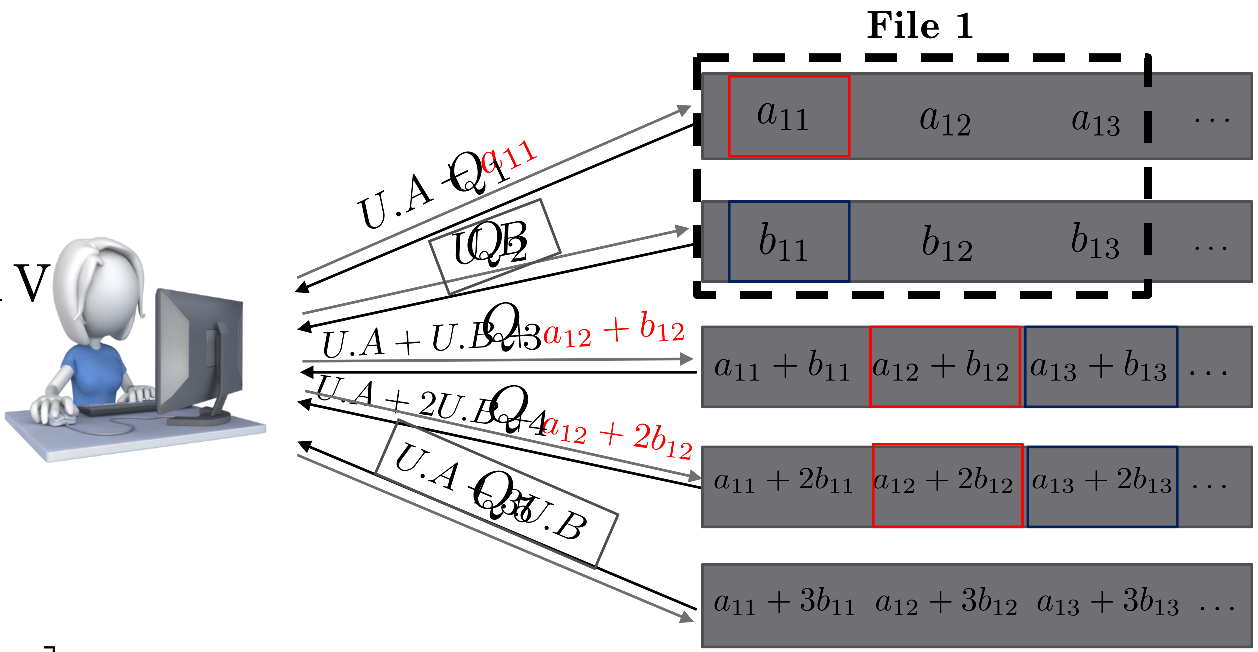
- 2 subqueries.
- 2 random vectors U and V

$$Q_1 = \begin{bmatrix} u_1+1 & u_2 & u_3 & \dots \\ v_1 & v_2 & v_3 & \dots \end{bmatrix}$$

$$Q_2 = \begin{bmatrix} u_1 & u_2 & u_3 & \dots \\ v_1+1 & v_2 & v_3 & \dots \end{bmatrix}$$

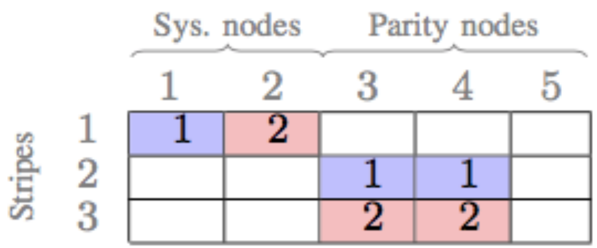
$$Q_3 = Q_4 = \begin{bmatrix} u_1 & u_2+1 & u_3 & \dots \\ v_1 & v_2 & v_3+1 & \dots \end{bmatrix}$$

$$Q_5 = \begin{bmatrix} u_1 & u_2 & u_3 & \dots \\ v_1 & v_2 & v_3 & \dots \end{bmatrix}$$



Proof of Theorem 1

- Scheme:
  - We divide each file into $n - k$ stripes.
  - k sub-queries are made to each node (dimension of code is d).
  - We write $n - k = \beta k + r$.
- Conditions:
  - Decode $n - k$ parts in each sub-query.
  - Parts not on same node.
  - Different parts in each sub-query

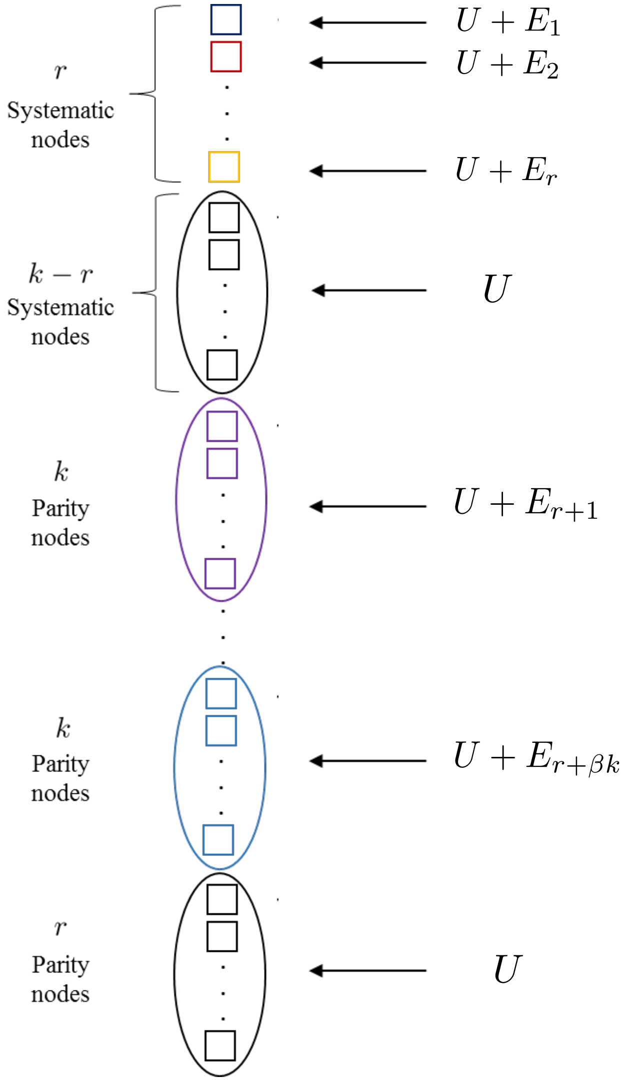


Retrieval pattern

Retrieval pattern for (15,4) MDS

Stripes		Sys. nodes 1	2	3	4	Parity nodes 5	6	7	8	9	10	11	12	13	14	15
r	1	1	2	3	4											
	2	2	3	4	1											
	3	3	4	1	2											
k	4					1	1	1	1							
	5					2	2	2	2							
	6					3	3	3	3							
	7					4	4	4	4							
k	8									1	1	1	1			
	9									2	2	2	2			
	10									3	3	3	3			
	11									4	4	4	4			
						k				k						

Querying



Where the E_is are matrices with 1s at the positions we want to decode.

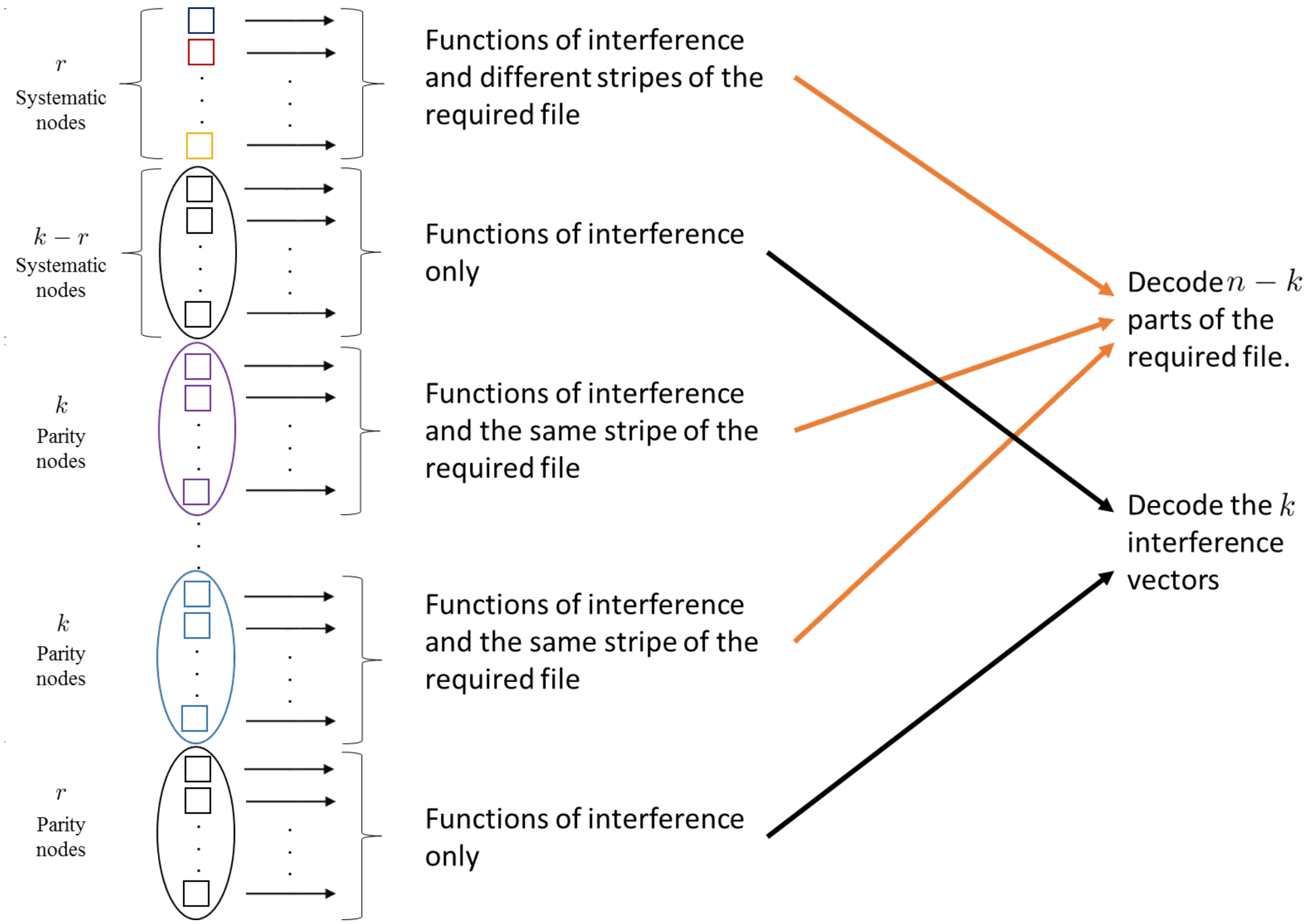
- k equations to decode interference.
- r equations from systematic nodes to decode parts of the first r stripes.
- βk equations from parity nodes to decode βk complete stripes.
- In total, $\beta k + r$ parts decoded.

Querying

- User creates a $d \times m\alpha$ random matrix U.
- Send matrix
  - $Q_l = U + V_l,$ to send to nodes $l = 1, \ldots, n - r$.
  - $Q_l = U,$ to send to nodes $l = n - r + 1, \ldots, n$.
- Where
  - $E_1 = \Big[\underbrace{\mathbf{0}_{k\times(f-1)\alpha}}_{\text{files } 1,\ldots,f-1} \;\vdots\; \underbrace{\begin{matrix} I_{r\times r} \\ \mathbf{0}_{(k-r)\times r} \end{matrix} \quad \mathbf{0}_{k\times\beta k}}_{\text{wanted file } f} \;\vdots\; \underbrace{\mathbf{0}_{k\times(m-f)\alpha}}_{\text{files } f+1,\ldots,m} \Big],$
  - $E_l \quad l = 1, 2, \ldots, k,$ is obtained from matrix E_{l-1} by a single downward cyclic shift of its row vectors.
  - For parity nodes $l = sk + 1, \ldots, sk + k$ for $s = 1, \ldots, \beta$

$$E_l = \big[\underbrace{\mathbf{0}_{k\times(f-1)\alpha}}_{\text{files } 1,\ldots,f-1} \;\vdots\; \underbrace{\mathbf{0}_{k\times r} \;\vdots\; \mathbf{0}_{k\times(s-1)k} \;\vdots\; I_{k\times k} \;\vdots\; \mathbf{0}_{k\times(\beta-s)k}}_{\text{wanted file } f} \;\vdots\; \underbrace{\mathbf{0}_{k\times(m-f)\alpha}}_{\text{files } f+1,\ldots,m} \big].$$

Response and Decoding



Decoding

- Decodability:

  - From systematic node $l = 1, \ldots, k$ and sub-query i

$$\begin{array}{rl} x^f_{i1} + I_l & l = i \\ I_l & l = (i+1)_k, \ldots, (i+k-r)_k \\ x^f_{l(i+k+1-l)_k} + I_l & l = (i+k-r+1)_k, \ldots, (i+k-1)_k \end{array}$$

  where $I_l = U_i^T W_l$

  - From parity nodes $l = n - r + 1, \ldots, n$

$$\lambda_{1l} I_1 + \lambda_{2l} I_2 + \cdots + \lambda_{kl} I_k$$

  - The rest of the parity nodes return:

$$\lambda_{1l} x^f_{1,r+(s-1)k+i} + \cdots + \lambda_{kl} x^f_{k,r+(s-1)k+i} + \lambda_{1l} I_1 + \lambda_{2l} I_2 + \cdots + \lambda_{kl} I_k$$

Part 3 – Privacy

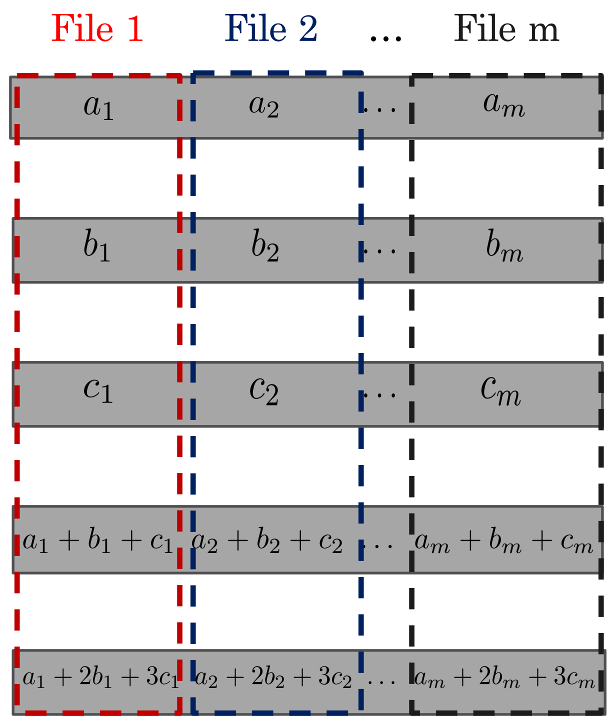
- Privacy:
  - Since b = 1, the only way a node l can learn information about f is from its own query matrix Q_i. By, construction Q_i is statistically independent of f and this scheme achieves perfect privacy.

- cPoP:
  - Every node responds with d = k symbols. Therefore, the total number of symbols downloaded by the user is kn. Therefore,

$$cPoP = \frac{kn}{k(n-k)} = \frac{1}{1-R}$$

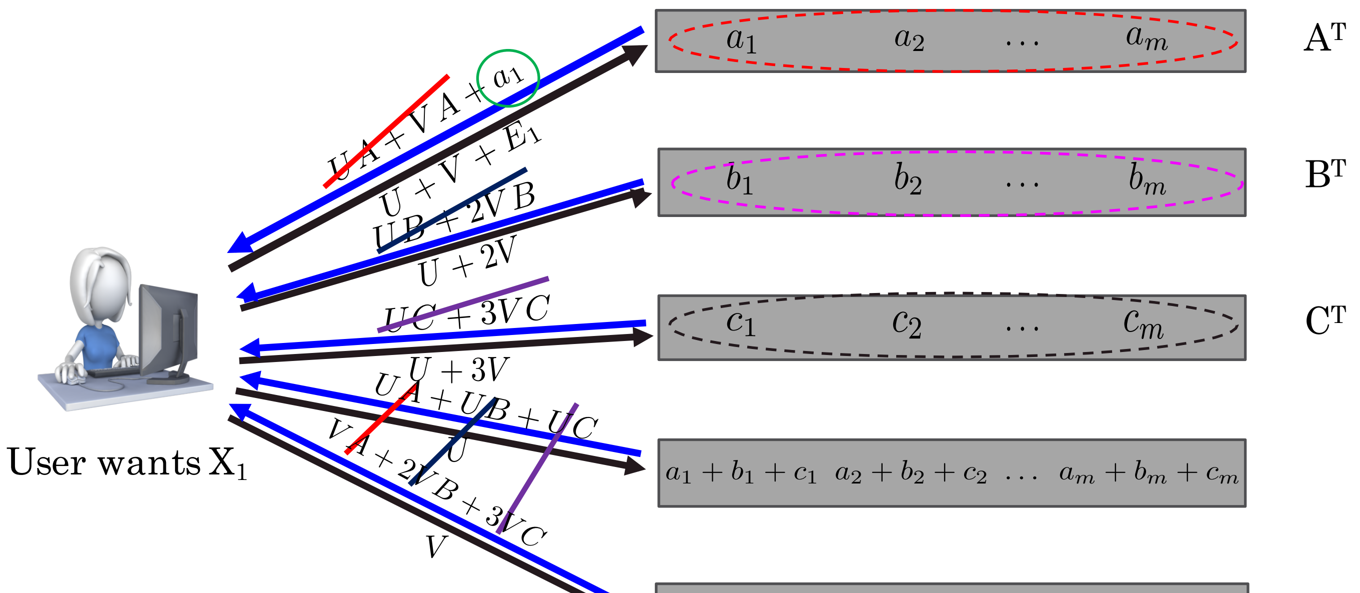
Example on Theorem 2

- Assume a (5,3) MDS code.
- Consider b = 2 colluding nodes.
- W.L.O.G. user wants file 1.



Example on Theorem 2

- User generates 2 random vectors U and V.



5 equations, 7 unknowns:
UA, VA, UB, VB, UC, VC, a_1

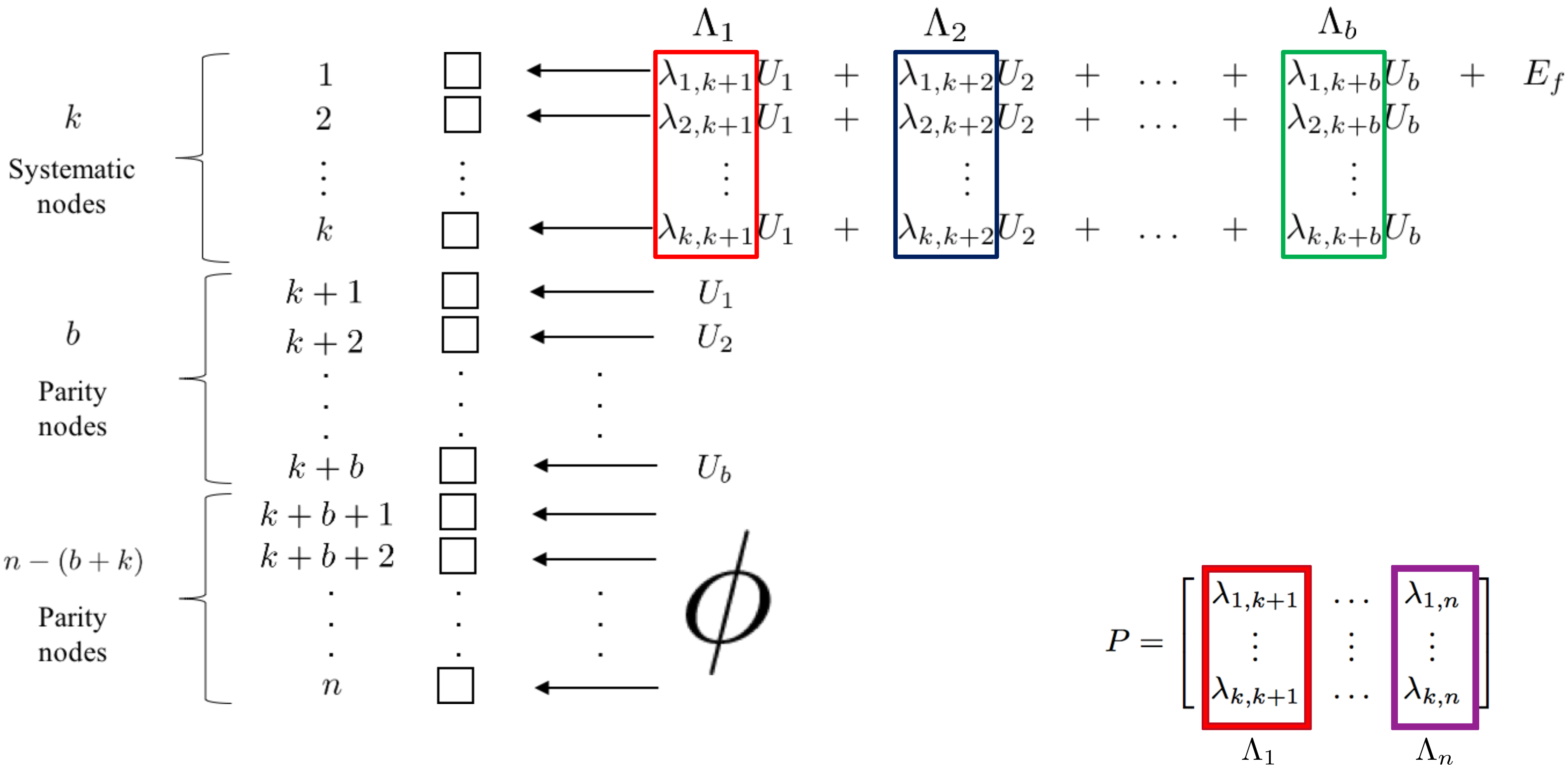
Proof of Theorem 2

Theorem 2: Consider a DSS using an (n,k) MDS code over $GF(q)$, with $b \leq n-k$ spy nodes. Then, there is an explicit linear PIR scheme over $GF(q)$ with download communication cost, $PoP = b+k$.

- Consider an (n,k) MDS code with the following generator matrix:

$$\Lambda = \left[\begin{array}{c|ccc} & \lambda_{1,k+1} & \cdots & \lambda_{1,n} \\ I_{k\times k} & \vdots & \vdots & \vdots \\ & \lambda_{k,k+1} & \cdots & \lambda_{k,n} \end{array} \right]$$

Decodability



$U_1, \dots, U_b :$ b random vectors

Open Problems

- Fundamental information theoretical bounds of the communication cost (cPoP).
- Is joint design of MDS code and PIR scheme necessary to achieve fundamental bounds?
- Partial retrieval of parts of the file.
- Beyond MDS codes, general linear codes, regenerating codes, Locally Recoverable codes, etc.
- General collusion patterns
- Malicious nodes etc…

Thank you!