Partially Collaborative Regenerating Codes for Distributed Storage Systems

Shiqiu Liu

joint work with Frédérique Oggier

INC

23, November, 2016

Outline

- Background
- A Min-cut Bound for Partial Collaboration
- Code Constructions for Partial Collaboration
    - Codes With Minimum Storage
    - Codes With Minimum Repair Bandwidth
- Bounds on Secure File Size against a Passive Eavesdropper
- Conclusion
- Future work

Outline

Background

A Min-cut Bound for Partial Collaboration

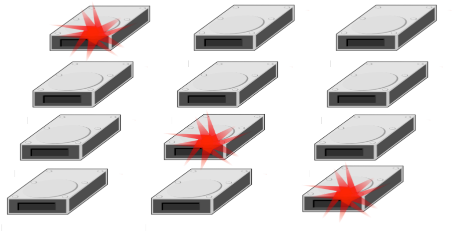
Code Constructions for Partial Collaboration

- Codes With Minimum Storage
- Codes With Minimum Repair Bandwidth

Bounds on Secure File Size against a Passive Eavesdropper

Conclusion

Distributed Storage Systems: A data object either is stored in a single node, or is split into smaller pieces that are stored over several nodes. The goal of the systems is to make sure the data object is not lost despite failures.



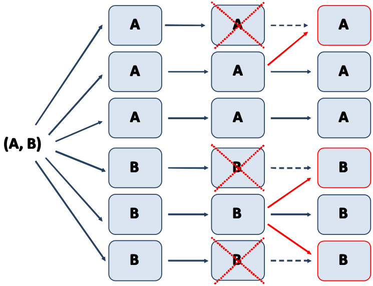
[Kubiatowicz et al., “OceanStore: An Architecture for Global-Scale Persistent Storage”, 2000.]

[Weatherspoon and Kubiatowicz, “Erasure coding vs. replication:a quantitative comparison”, 2002.]

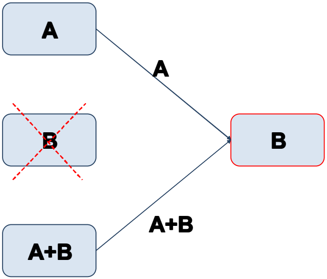
Redundancy: the information contained in the data object is stored several times.

Repairability: the data object can be recovered by contacting live nodes, downloading data from them, and computing the missing codeword coefficients.

- Repetition Codes



- Erasure codes: MDS codes (Reed-Solomon codes)



Good Storage Codes: do not store too much redundancy, repair efficiently (less communication cost, fast, computationally cheap...)

Regenerating Codes: aim at reducing the amount of symbols communicated during repair.

[Dimakis et al.,[1] "Network coding for distributed storage systems", 2008]

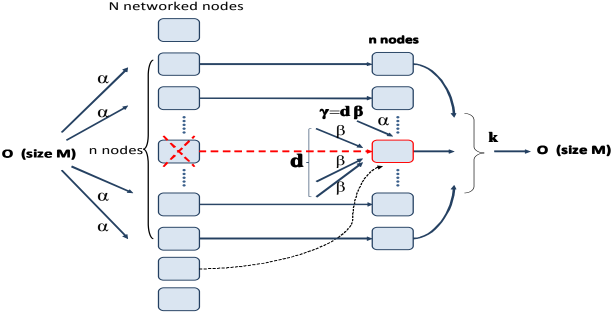
Storage Capacity α: the amount of data stored in one node such that the object can be retrieved by contacting any choice of k nodes.

Repair Bandwidth γ: the amount of symbols communicated during repair.

Parameters of Importance:

- N: total number of nodes in the network
- n: number of nodes storing one object
- k: any choice of k nodes allow the object retrieval

- M: the size of an object
- d: number of live nodes contacted by one repairing node
- α: storage capacity per node
- β: download repair bandwidth
- γ: repair bandwidth per node



- **MSR: Minimum Storage Repair point**: the point where the storage capacity is minimum, that is $\alpha = M/k$.

- **MBR: Minimum Bandwidth Repair point**: the point where the repair bandwidth achieves its minimum, that is $\gamma = \alpha$.

[Y. Wu et al., “Deterministic regenerating codes for distributed storage”, 2007.]

[Cullina et al., “Searching for minimum storage regenerating codes”, 2009.]

[Rashmi et al., “Exact regenerating codes for distributed storage”, 2009.]

Collaborative Repair:

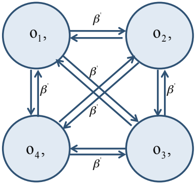
(Hu et al.[2], Kermarrec et al.[3], 2010)

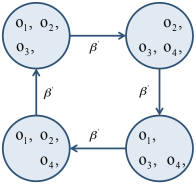
- Repair several (t) failures simultaneously.
- One repairing node exchanges β' data with every other repairing nodes (after the download phase).

Partially Collaborative Repair:

(Liu & Oggier [4], 2014)

- Collaboration among small subsets of repairing nodes.
- Gives $t - s$ levels of collaboration, which ranges from 0 to $t - 1$ corresponding to the level from none to full.





[Shum et al., "Exact Minimum-Repair-Bandwidth Cooperative Regenerating Codes for Distributed Storage Systems", 2011.]

[N. Le Scouarnec, "Exact Scalar Minimum Storage Coordinated Regenerating Codes", 2012.]

Outline

Background

A Min-cut Bound for Partial Collaboration

Code Constructions for Partial Collaboration
- Codes With Minimum Storage
- Codes With Minimum Repair Bandwidth

Bounds on Secure File Size against a Passive Eavesdropper

Conclusion

A Min-cut Bound for Partial Collaboration

Theorem

An object of size M stored in a distributed storage system, where every node has a storage capacity of α, and repairs are performed by a group of t nodes. Suppose that the data collector connects to a subset of k nodes which were all involved in different phases of repairs, where each phase involves a group of u_i nodes, $1 \leq u_i \leq t$, and $k = \sum_{i=0}^{g-1} u_i$. Then the size M is bounded by

$$M \leq \min_{\mathbf{u} \in P} \Big(\sum_{i \in I} u_i \min\{\alpha, (d - \sum_{j=0}^{i-1} u_j)\beta + (t - s + 1 - u_i)\beta'\} + \sum_{i \in \bar{I}} u_i \min\{\alpha, (d - \sum_{j=0}^{i-1} u_j)\beta\} \Big)$$

where $I = \{i,\ t - s + 1 - u_i \geq 0\}$, $\bar{I} = \{i,\ t - s + 1 - u_i < 0\}$ and $P = \{\mathbf{u} = (u_0, \ldots, u_{g-1}),\ 1 \leq u_i \leq t$ and $\sum_{i=0}^{g-1} u_i = k\}$.

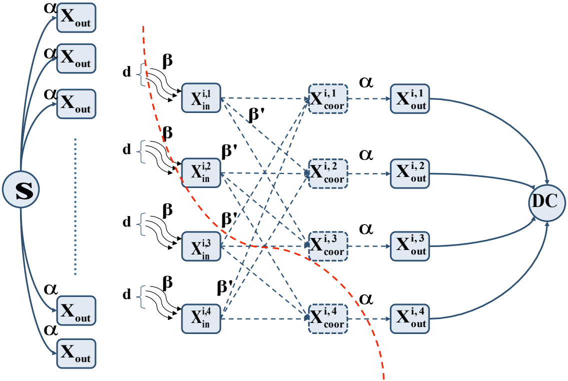


Figure: The capacity of a cut in a repair group is $2d\beta + \alpha + 4\beta'$.

Corollary

When $s = t$, there is no collaboration, and

$$M \leq \sum_{i=0}^{k-1} u_i \min\{\alpha, (d-i)\beta\},$$

for $u_i = 1$ and $g = k$, which is the known bound from [1].

Corollary

When $s = 1$, the collaboration phase involves all the other $t - s = t - 1$ nodes, and

$$M \leq \sum_{i=0}^{g-1} u_i \min\{\alpha, (d - \sum_{j=0}^{i-1} u_j)\beta + (t - u_i)\beta'\},$$

for $1 \leq u_i \leq t$ such that $\sum_{i=0}^{g-1} u_i = k$ which is the known bound from [3,5].

Corollary

MSR $(\alpha < \gamma)$	α (storage capacity)	γ (repair bandwidth)	β (download bandwidth)	β' (exchange bandwidth)
No	$\frac{M}{k}$	$\frac{M}{k}\frac{d}{d-k+1}$	$\frac{M}{k}\frac{1}{d-k+1}$	0
Full	$\frac{M}{k}$	$\frac{M}{k}\frac{d+t-1}{d-k+t}$	$\frac{M}{k}\frac{1}{d-k+t}$	$\frac{M}{k}\frac{1}{d-k+t}$
Partial	$\frac{M}{k}$	$\frac{M}{k}\frac{d+t-s}{d-k+t-s+1}$	$\frac{M}{k}\frac{1}{d-k+t-s+1}$	$\frac{M}{k}\frac{1}{d-k+t-s+1}$

MBR $(\alpha = \gamma)$	α (storage capacity)	γ (repair bandwidth)	β (download bandwidth)	β' (exchange bandwidth)
No	$\frac{M}{k}\frac{2d}{2d-k+1}$	$\frac{M}{k}\frac{2d}{2d-k+1}$	$\frac{M}{k}\frac{2}{2d-k+1}$	0
Full	$\frac{M}{k}\frac{2d+t-1}{2d-k+t}$	$\frac{M}{k}\frac{2d+t-1}{2d-k+t}$	$\frac{M}{k}\frac{2}{2d-k+t}$	$\frac{M}{k}\frac{1}{2d-k+t}$
Partial	$\frac{M}{k}\frac{2d+t-s}{2d-k+t-s+1}$	$\frac{M}{k}\frac{2d+t-s}{2d-k+t-s+1}$	$\frac{M}{k}\frac{2}{2d-k+t-s+1}$	$\frac{M}{k}\frac{1}{2d-k+t-s+1}$

M: size of an object, k: any choice of k nodes allow the object retrieval, d: number of live nodes contacted by one repairing node, $t-s$: level of collaboration.

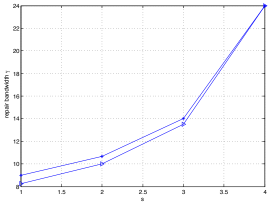


Figure: The total repair bandwidth γ at MSR. A data object of size $M = 24$, $t = 4$ with $1 \leq s \leq 4$, for $k = 6$ (upper line) and $k = 8$ (lower line).

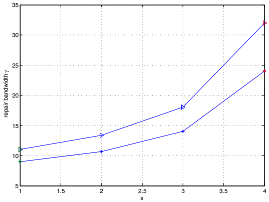


Figure: The total repair bandwidth γ at MSR. Storage capacity $\alpha = 4$, $t = 4$ with $1 \leq s \leq 4$, for $k = 8$ (upper line) and $k = 6$ (lower line).

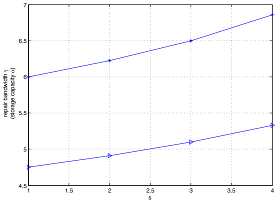


Figure: The total repair bandwidth γ at MBR with $\gamma = \alpha$. A data object of size $M = 24$, $t = 4$ with $1 \leq s \leq 4$, for $k = 6$ (upper line) and $k = 8$ (lower line).

Outline

Background

A Min-cut Bound for Partial Collaboration

Code Constructions for Partial Collaboration

- Codes With Minimum Storage
- Codes With Minimum Repair Bandwidth

Bounds on Secure File Size against a Passive Eavesdropper

Conclusion

Codes With Minimum Storage

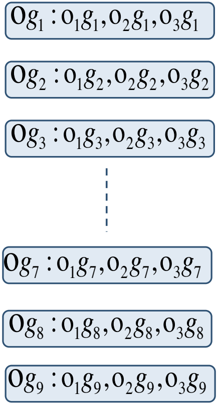
An object $\mathbf{o}$ of length $M = k(t-s+1)$ with coefficients in the finite field $\mathbb{F}_q$, which can be represented by a matrix $\mathbf{O}$ in $\mathbb{F}_q^{(t-s+1)\times k}$, that is

$$\mathbf{O} = \begin{bmatrix} o_{1,1} & \cdots & o_{1,k} \\ \vdots & & \vdots \\ o_{t-s+1,1} & \cdots & o_{t-s+1,k} \end{bmatrix} = \begin{bmatrix} \mathbf{o}_1 \\ \vdots \\ \mathbf{o}_{t-s+1} \end{bmatrix}.$$

An MDS code with generator matrix G in $\mathbb{F}_q^{k\times n}$

$$G = [\mathbf{g}_1, \ldots, \mathbf{g}_n].$$

Let the ith node store the coefficients $\mathbf{Og}_i$, each of node stores $t+s-1$ coefficients.



$Og_1 : o_1g_1, o_2g_1, o_3g_1$
$Og_2 : o_1g_2, o_2g_2, o_3g_2$
$Og_3 : o_1g_3, o_2g_3, o_3g_3$
$Og_7 : o_1g_7, o_2g_7, o_3g_7$
$Og_8 : o_1g_8, o_2g_8, o_3g_8$
$Og_9 : o_1g_9, o_2g_9, o_3g_9$

Figure: An object of size $M = k(t-s+1) = 5\cdot 3$ is encoded by a $(9,5)$-MDS code, and is stored in 9 nodes.

Repair of t failures: Label these t nodes from 1 to t.

- t new comer nodes each contact $d = k$ nodes, the ith node among those t live nodes connect to nodes $i_1, \dots, i_k$ and downloads
$$\mathbf{o}_i \mathbf{g}_{i_j}, \ j = 1, \dots, k.$$
Using the MDS property of G, the ith node can compute one coefficient $\mathbf{o}_i \mathbf{g}_i$.
- The $t - s$ missing coefficients obtained through (partial) collaboration, the ith node contacts nodes $i + 1, \dots, i + (t - s) \pmod{t}$ and gets $\mathbf{o}_{i+1}\mathbf{g}_i, \dots, \mathbf{o}_{i+(t-s)}\mathbf{g}_i$.

Example 1:

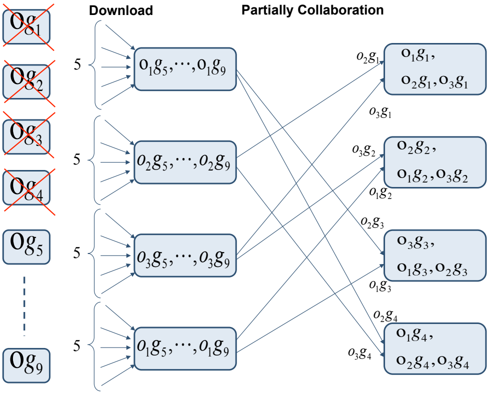
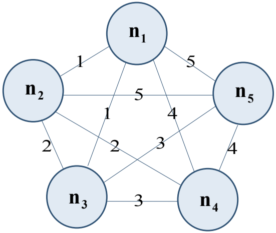


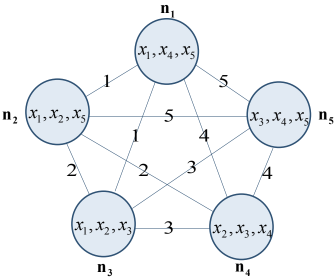
Figure: Minimum Storage Code construction for an object $\mathbf{o}$ of size $M = 15$, storage capacity $\alpha = 3$, $G = [\mathbf{g}_1, \ldots, \mathbf{g}_9]$ be the generator matrix of the $(9, 5)$-MDS code, a threshold of $t = 4$, repair degree $d = k = 5$, collaboration among only $t - s = 4 - 2 = 2$ nodes.

Codes With Minimum Repair Bandwidth

An object **o** of length $M = 2m$ (for some positive integer n) with coefficients in the finite field $\mathbb{F}_q$, which is encoded into the length $2m+1$ codeword **x**, $\mathbf{x} = (o_1, o_2, ..., o_{2m}, o_1 + o_2 + \cdots + o_{2m})$.

Data Placement: Consider a network of $n = 2m+1$ nodes, and see it as complete graph, with thus $\frac{(2m+1)2m}{2} = m(2m+1)$ edges. Label the edges of this graph with labels in $\{1, \ldots, 2m+1\}$ so that node i has exactly m edges labeled with i.





$$\begin{bmatrix} 0 & 1 & 1 & 4 & 5 \\ 1 & 0 & 2 & 2 & 5 \\ 1 & 2 & 0 & 3 & 3 \\ 4 & 2 & 3 & 0 & 4 \\ 5 & 5 & 3 & 4 & 0 \end{bmatrix}$$

An object $\mathbf{o} = (o_1, o_2, o_3, o_4)$ of size $M = 4 = 2m$, $m = 2$, which is encoded into the length 5 codeword $\mathbf{x}$ given by $\mathbf{x} = (o_1, o_2, o_3, o_4, o_1 + o_2 + o_3 + o_4)$. A network of $n = 5$ nodes, see as a complete graph with adjacency matrix as shown.

For a network of $n = 2m + 1$ nodes, which corresponds to fill up the adjacency matrix of this graph, done as follows: put zeroes on the diagonal, and then label the ith row, columns $i + 1, \dots, i + m \pmod{2m+1}$ with i:

$$\begin{bmatrix}
0 & 1 & \dots & 1 & & & & \\
 & 0 & 2 & \dots & 2 & & & \\
 & & \ddots & & & & & \\
 & & & 0 & m+1 & & \dots & m+1 \\
m+2 & & & & 0 & m+2 & \dots & m+2 \\
 & & & & & & & \\
2m+1 & \dots & 2m+1 & & & & & 0
\end{bmatrix}$$

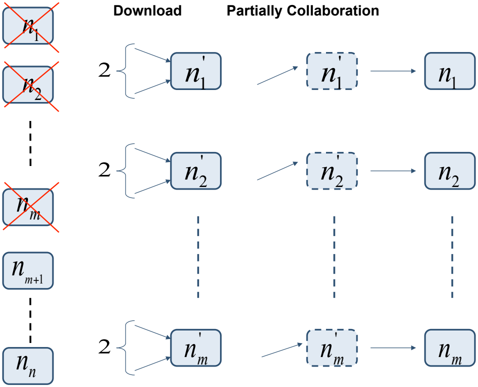
Object Recovery:

Any $k = m$ nodes are enough to retrieve the object.

$$\begin{bmatrix}
0 & 1 & \dots & 1 & & & & \\
 & 0 & 2 & \dots & 2 & & & \\
 & & \ddots & & & & & \\
 & & & 0 & m+1 & & \dots & m+1 \\
m+2 & & & & 0 & m+2 & \dots & m+2 \\
 & & & & & & & \\
2m+1 & \dots & 2m+1 & & & & & 0
\end{bmatrix}$$

Repair of t=m Failures:

It is possible to repair $t = m$ failures, with repair degree $d = 2$, and collaborative repair degree $t - s = 1$. The download and collaboration phases are available by considering the adjacency matrix for $n = 2m + 1$ nodes.



Example 2:

Consider for example $m = 4$, thus an object size of $2m = M = 8$, stored across $n = 9$ nodes, each storing $\alpha = 5$ coefficients of a codeword $\mathbf{x}$.

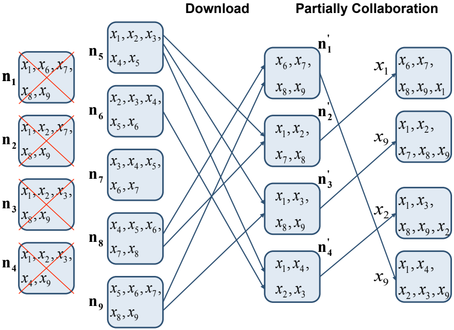

Download
Partially Collaboration
$\mathbf{n}_1$ $x_1, x_6, x_7,$ x_8, x_9
$\mathbf{n}_2$ $x_1, x_2, x_7,$ x_8, x_9
$\mathbf{n}_3$ $x_1, x_2, x_3,$ x_8, x_9
$\mathbf{n}_4$ $x_1, x_2, x_3,$ x_4, x_9
$\mathbf{n}_5$ $x_1, x_2, x_3,$ x_4, x_5
$\mathbf{n}_6$ $x_2, x_3, x_4,$ x_5, x_6
$\mathbf{n}_7$ $x_3, x_4, x_5,$ x_6, x_7
$\mathbf{n}_8$ $x_4, x_5, x_6,$ x_7, x_8
$\mathbf{n}_9$ $x_5, x_6, x_7,$ x_8, x_9
$\mathbf{n}_1'$ $x_6, x_7,$ x_8, x_9
$\mathbf{n}_2'$ $x_1, x_2,$ x_7, x_8
$\mathbf{n}_3'$ $x_1, x_3,$ x_8, x_9
$\mathbf{n}_4'$ $x_1, x_4,$ x_2, x_3
x_1
x_9
x_2
x_9
$x_6, x_7,$ x_8, x_9, x_1
$x_1, x_2,$ x_7, x_8, x_9
$x_1, x_3,$ x_8, x_9, x_2
$x_1, x_4,$ x_2, x_3, x_9

Figure: $t = 4$ failures occur, suppose that nodes $1, 2, 3$ and 4 fail. Repair degree $d = 2$, each repair node exchange coefficients with other $t - s = 1$ node.

Outline

Background

A Min-cut Bound for Partial Collaboration

Code Constructions for Partial Collaboration

- Codes With Minimum Storage
- Codes With Minimum Repair Bandwidth

Bounds on Secure File Size against a Passive Eavesdropper

Conclusion

Bounds on Secure File Size against a Passive Eavesdropper

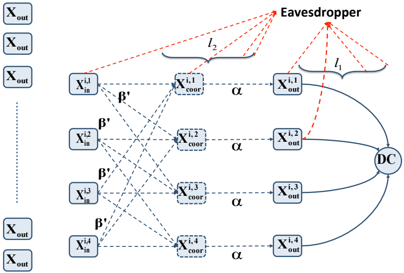


Figure: The eavesdropper spies the information stored in l_1 nodes, the download and collaborated information in l_2 repairing nodes, $l_1 + l_2 \leq k$.

MBR: storage capacity α = repair bandwidth γ, consider $l_2 = 0$, $l_1 \leq k$.
MSR: repair bandwidth $\gamma >$ storage capacity α, consider $l_1 + l_2 \leq k$.

[Pawar et al., "Securing Dynamic Distributed Storage Systems against Eavesdropping and Adversarial Attacks", 2011.]

[Koyluoglu et al., "Secure Cooperative Regenerating Codes for Distributed Storage Systems", 2012.]

Proposition (MBR)

For an (n, k) partially collaborative storage code at the minimum repair bandwidth point (MBR), we have that the secure file size $\mathcal{M}^s$ is upper bounded by

$$\mathcal{M}^s \leq (k - l_1)(2d + t - s - k - l_1 + 1)\frac{\beta}{2} \tag{5.1}$$

in the presence of an l_1-eavesdropper spying the stored content of l_1 nodes with $0 \leq l_1 \leq k$. We further have $\beta = 2\beta'$ and the optimal file size is reached with equality in (5.1) and

$$\beta = \frac{\mathcal{M}^s}{k - l_1} \frac{2}{2d + t - s + 1 - k - l_1},$$

$$\alpha = \frac{\mathcal{M}^s}{k - l_1} \frac{2d + t - s}{2d + t - s + 1 - k - l_1}.$$

A data object of size $\mathcal{M}^s$ is firstly encoded into a file of size M.

Consider the normalized repair bandwidth, that is $\frac{\gamma}{\mathcal{M}^s}$.

Result

When $n = d + t$, collaboration does not reduce the normalized repair bandwidth with respect to no collaboration, that is, the best strategy is to repair one failure at a time.

When $n > d + t$, collaboration indeed reduces the normalized repair bandwidth. If there exists a full collaboration strategy for some t_0, and a partial collaboration strategy for some t, such that $t_0 \leq t - s + 1$, then partial collaboration reduces the normalized repair bandwidth furthermore.

Remark

This bound at MBR includes the parameter $t - s$ (collaboration level), which generalizes the bounds for no collaboration [6] and full collaboration [7] with a passive eavesdropper at MBR.

Proposition (MSR)

For an (n,k) partially collaborative storage code at the minimum storage repair point (MSR), we have that the secure file size $\mathcal{M}^s$ is upper bounded by

$$\mathcal{M}^s \leq \sum_{i=1}^{k-l_1-l_2} (\alpha - I(\mathbf{s}_i; \mathbf{d}_{i,\varepsilon_2})) \tag{5.2}$$

If $I(\mathbf{s}_i; \mathbf{d}_{i,\varepsilon_2}) \geq \beta' = \beta$, then (5.2) becomes

$$\mathcal{M}^s \leq (k - l_1 - l_2)(\alpha - \beta), \tag{5.3}$$

in the presence of an (l_1, l_2) eavesdropper spying the stored content of l_1 nodes and the downloaded and exchanged data of l_2 nodes, where $l_1 + l_2 \leq k$. Furthermore $\beta = \beta'$ and the optimal secure file size is reached with equality in (5.3), with

$$\beta = \beta' = \frac{\alpha}{d - k + t - s + 1}.$$

A data object of size $\mathcal{M}^s$ is firstly encoded into a file of size M.

Result

Consider the upper bound (5.3).
when a passive eavesdropper exists, the secure file size is the smallest for no collaboration, and is the biggest for full collaboration.

If there exists a full collaboration strategy for some t_0, and a partial collaboration strategy for some t, such that $t_0 \leq t - s + 1$, then the secure file size for partial collaboration will be bigger then that for full collaboration.

Outline

Background

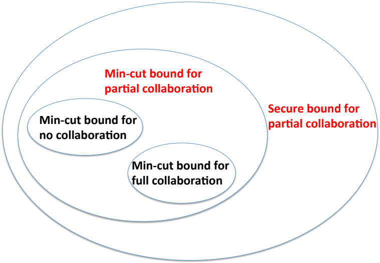
A Min-cut Bound for Partial Collaboration

Code Constructions for Partial Collaboration
- Codes With Minimum Storage
- Codes With Minimum Repair Bandwidth

Bounds on Secure File Size against a Passive Eavesdropper

Conclusion

Conclusion


Min-cut bound for partial collaboration
Secure bound for partial collaboration
Min-cut bound for no collaboration
Min-cut bound for full collaboration

- Design of codes for partial collaboration at MSR point and MBR point.

Future Work

- Optimal MBR codes: our approaches to construct codes can be further improved. In fact, the construction for MBR point varies for different number of live nodes contacted, and the threshold of the failures could vary. A general code construction needs to be considered.
- Code constructions for partial collaboration with a passive eavesdropper.
- Security: the systems need to be robust against adversaries not only passive, but also active.

References

[1] A.G. Dimakis, P.B. Godfrey, Y. Wu, M.J. Wainwright, K. Ramchandran, *"Network Coding for Distributed Storage Systems"*.

[2] Y. Hu, Y. Xu, X. Wang, C. Zhan and P. Li, *"Cooperative Recovery of Distributed Storage Systems from Multiple Losses with Network Coding"*.

[3] A.-M. Kermarrec, N. Le Scouarnec, G. Straub, *"Repairing Multiple Failures with Coordinated and Adaptive Regenerating Codes"*.

[4] S. Liu, F. Oggier, *"On Storage Codes Allowing Partially Collaborative Repairs"*.

[5] K. Shum, *"Cooperative Regenerating Codes for Distributed Storage Systems"*.

[6] A. S. Rawat, O. O. Koyluoglu, N. Silberstein and S. Vishwanath, *"Optimal Locally Repairable and Secure Codes for Distributed Storage Systems"*.

[7]O. O. Koyluoglu, A. S. Rawat and S. Vishwanath, *"Secure Cooperative Regenerating Codes for Distributed Storage Systems"*.

Thank you!