55th Annual Allerton Conference on Communication, Control, and Computing

Correcting **Localized** Deletions Using Guess & Check Codes

Salim El Rouayheb

Rutgers University

Joint work with

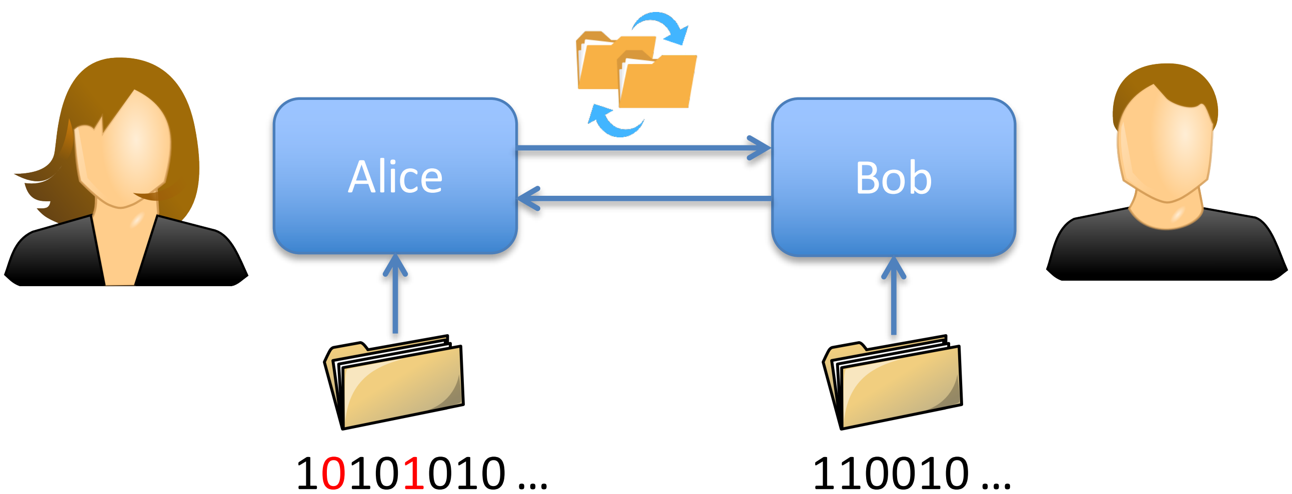
Serge Kas Hanna and Hieu Nguyen

Motivation

- Deletions: $\overset{\text{Transmitted}}{10101010} \longrightarrow \overset{\text{Received}}{110010}$

- Deletions were first studied by Varshamov-Tenengolts ('65) and Levenshtein ('66)

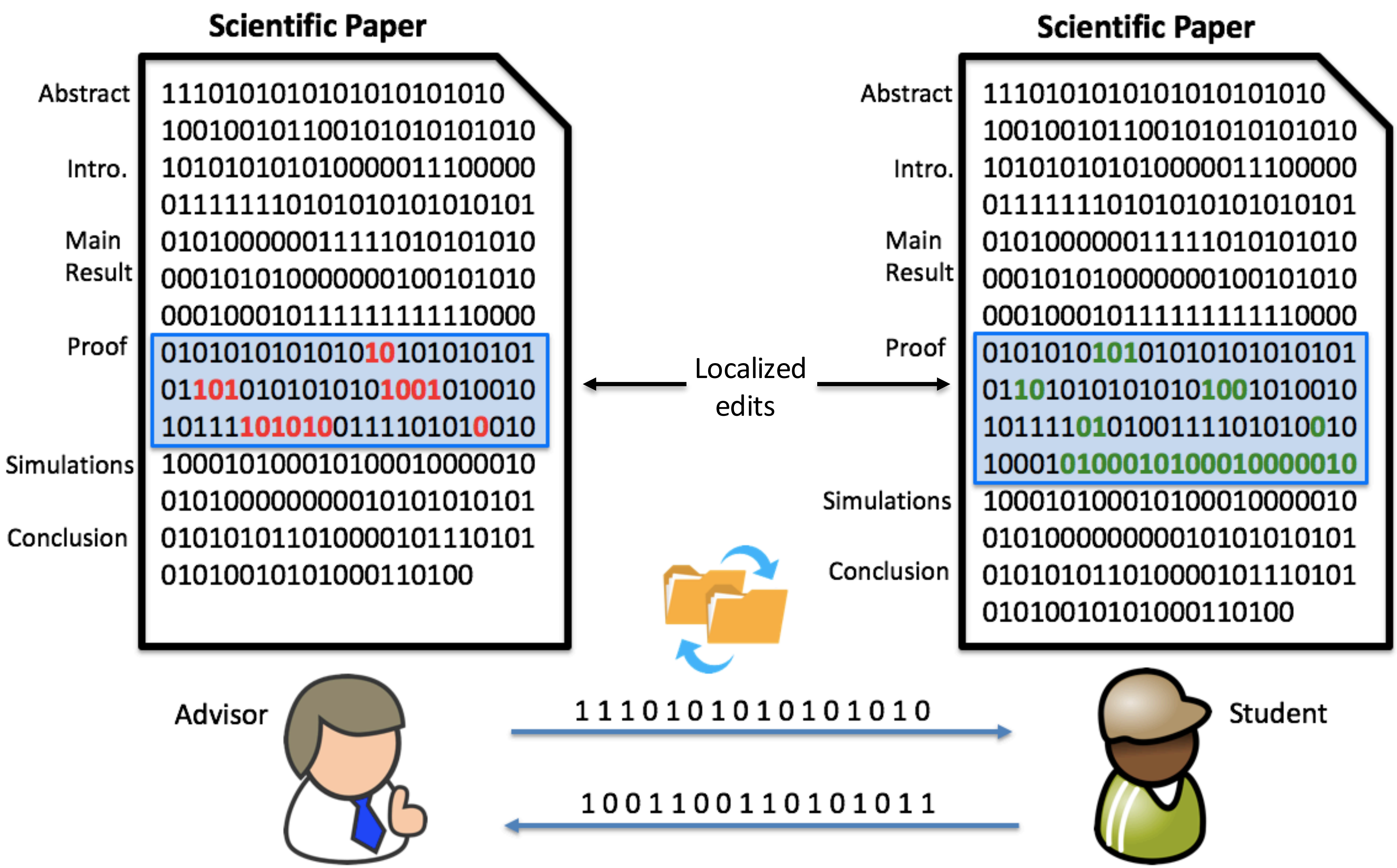
- Our motivation: file synchronization, E.g. Dropbox



- Recent application: DNA-based storage

Localized Deletions

- Motivation: file synchronization, E.g. Dropbox



Previous Work on Deletions

- **Unrestricted deletions**
  - *Information theoretic approach:* [Gallager ’61], [Dobrushin ‘67]; lower and upper bounds on the capacity: [Mitzenmacher and Drinea ‘06], [Diggavi et al. ‘07], [Kanoria and Montanari ‘13], [Venkataramanan et al. ’13] …
  - *Recent file synchronization algorithms:* [Yazdi and Dolecek ‘14], [Venkataramanan et al. ‘15], [Sala et al. ’17] …
  - *Code constructions and fundamental limits:* [Varshamov and Tenenglots ‘65], [Levenshtein ‘66], [Schulman and Zuckerman ‘99], [Helberg and Ferreira ‘02] , [Cullina and Kiyavash ‘14], [Gabrys et al. ‘16], [Brankensiek et al. ’16], [Thomas et al. ’17] …
- **Bursty deletions**
  - File synchronization: [Ma et al. ‘11]
  - Code constructions: [Levenshtein ‘67], [Cheng et al. 14], [Schoeny et al ‘17]

Existence of codes for localized model **w**=3,4

Model and Contribution

δ deletions (in red)

$$x = 1010101110001001000101\boxed{1101011110000}0101000110011111000010$$

window of size $\boldsymbol{w}$

- $\delta \leq \boldsymbol{w}$ deletions localized in a window of size $\boldsymbol{w}$
- Hard problem for $\boldsymbol{w} = n$
- [Schoeny et al. '17]: existence of codes for $\boldsymbol{w} = 3,4$
- Our assumptions: 1) positions of the deletions are independent of the codeword; 2) information message is uniform iid
- Contribution: Explicit codes with deterministic polynomial time encoding and decoding that can correct localized deletions whp

Guess & Check (GC) Codes

- Logarithmic redundancy: $n - k = c \log k + \boldsymbol{w} + 1$
- Polynomial time encoding and decoding
- Asymptotically vanishing probability of decoding failure
- Can be generalized to multiple windows

GC Codes Example

Deletions occur in one of these windows

- Encoding the message of length k=16: 1 1 1 0 | 0 0 0 0 | 0 0 1 1 | 0 0 0 1

GF(17)

1110 | 0000 | 0011 | 0001

14 | 0 | 3 | 1 $\xrightarrow{\text{(6,4) MDS}}$ 14 | 0 | 3 | 1 | 1 | 0

2 MDS parities

GF(17)

- Assume that the deletions (in red) affect only one systematic block

$\underbrace{\text{1110 | 0000 | 0011 | 0001}}_{\text{16 bits}}$ → $\underbrace{\text{1110000000001}}_{\text{13 bits}}$

- Decoding

Decoded using 1st parity

	?	12	0	1		
➢ Guess 1:	~~1~~	1100	0000	0001	→	5 | 12 | 0 | 1
➢ Guess 2:	1110	~~0~~	0000	0001	→	14 | 3 | 0 | 1
➢ Guess 3:	1110	0000	~~0~~	0001	→	14 | 0 | 3 | 1
➢ Guess 4:	1110	0000	0000	~~1~~	→	14 | 0 | 0 | 4

Check with 2nd parity

0

[Kas Hanna and El Rouayheb ISIT 17']

Generalizing to Any Window Position

- Assume that 3 deletions (in red) affect systematic bits, w=log k=4 bits

window

1110 0100 0011 0001 → 1100000110001

16 bits — 13 bits

- Same encoding with one extra parity

log k bits

1110 | 0100 | 0011 | 0001 $\xrightarrow{(7,4)\text{ MDS}}$ 14 | 4 | 3 | 1 | 5 | 8 | 12

3 MDS parities

GF(17)

- Decoding, window of size log k can affect at most 2 **consecutive** blocks

Decoded using 1st & 2nd parity

- Guess 1: ~~11000~~ (?) | 0011 (3) | 0001 (1) → 14 | 4 | 3 | 1 ✓
- Guess 2: 1100 | ~~00011~~ | 0001 → 12 | 7 | 2 | 1 ✗
- Guess 3: 1100 | 0001 | ~~10001~~ → 12 | 1 | 11 | 15 ✗

Check with 3rd parity

12

Recovering the MDS parities

- How to recover the MDS parity symbols at the decoder?

systematic bits | MDS parities

111000000011000100010000

- Trivial solution: repeat the parity bits

systematic bits | $(\delta + 1)$ repetition code

1110000000110001000000000000111 0 … 0

- Better solution: insert a buffer between systematic and MDS parity bits

systematic bits | buffer | MDS parities

1110000000110001 $\underbrace{\text{00001}}_{\textbf{w}+1\text{ bits}}$ 00010000

  - Buffer: **w** zeros + a single one
  - Deletions cannot affect both systematic and parity bits simultaneously
  - If parity bits get affected -> simply output the first k bits. Else apply Guess & Check decoding

When does decoding fail?

- Encoding the message of length k=16: 1 1 0 0 0 1 0 0 0 1 1 1 0 0 1 0

log k bits

GF(17)

1100 0100 0111 0010

14 4 7 2 $\xrightarrow{\text{(6,4) MDS}}$ 14 4 7 2 8 13

2 MDS parities

GF(17)

- Assume that the deletions (in red) affect only one systematic block

$\underbrace{\text{1100 0100 0111 0010}}_{\text{16 bits}} \longrightarrow \underbrace{\text{0010001110010}}_{\text{13 bits}}$

- Decoding

Decoded using 1st parity

? 4 7 2

- Guess 1: ~~0~~ 0100 0111 0010 → 14 4 7 2 ✓
- Guess 2: 0010 ~~0~~ 0111 0010 → 2 14 7 2 ✗
- Guess 3: 0010 0011 ~~1~~ 0010 → 2 3 1 2 ✗
- Guess 4: 0010 0011 1001 ~~0~~ → 2 3 9 11 ✓

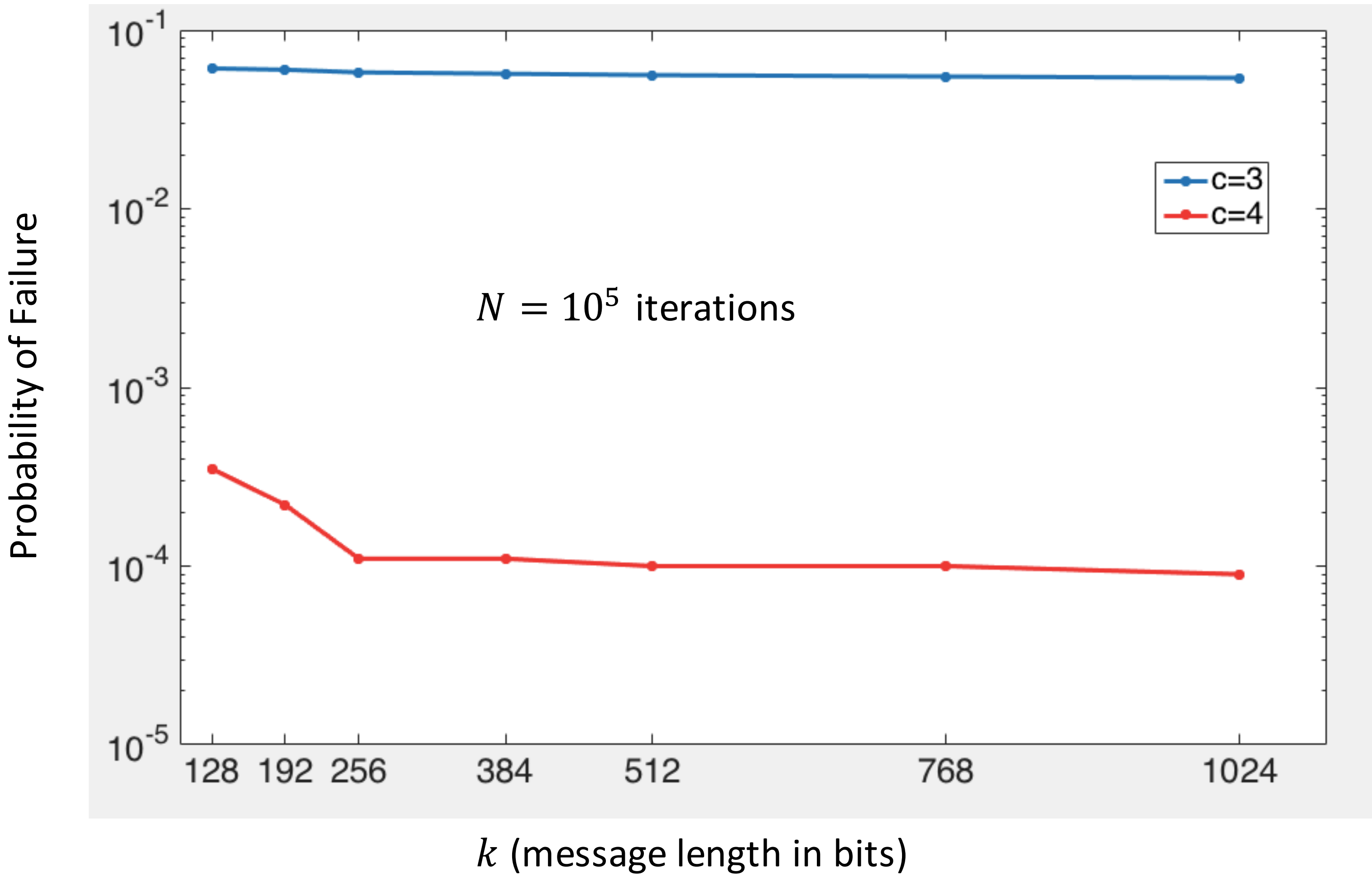
Decoding Failure

Check with 2nd parity

13

Simulations – Decoding Failure

- Simulation results for: $\boldsymbol{w} = \log k$, $\delta = \log k - 1$, $\mathrm{c} = 3{,}4$ MDS parities



Main Results

Theorem 1 (One window): Guess & Check (GC) codes can correct in polynomial time up to $\delta \leq \boldsymbol{w} = O(\log k)$ localized deletions, where $m \log k < \boldsymbol{w} < (m+1)\log k$ for some integer $m \geq 0$. Let $c > m+2$ be a constant integer.

- Redundancy: $c \log k + \boldsymbol{w} + 1$
- Encoding complexity is $\mathcal{O}(k \log k)$, Decoding complexity is $\mathcal{O}(k^{\mathbf{3}} / \log k)$
- Probability of decoding failure: $\Pr(F) \leq k^{m+4-c} / \log k$

Sketch of Theorem 2 ($\boldsymbol{z} > 1$ windows):

- Redundancy: $c(z\boldsymbol{w} + 1)\log k$
- Encoding complexity is $\mathcal{O}(k \log k)$, Decoding complexity is $\mathcal{O}(k^{z+2})$
- Probability of decoding failure: $\Pr(F) \leq k^{z(m+4)-c}$

Sketch of Theorem 3 [ISIT '17] (Unrestricted deletions):

- Redundancy: $c(\delta + 1)\log k$
- Encoding complexity is $\mathcal{O}(k \log k)$, Decoding complexity is $\mathcal{O}(k^{\delta+2} / \log^{\delta} k)$
- Probability of decoding failure: $\Pr(F) = \mathcal{O}(k^{2\delta-c} / \log^{\delta} k)$

Test GC Codes Online

- Test the codes online using the Jupyter notebook

  Go to : https://try.jupyter.org/

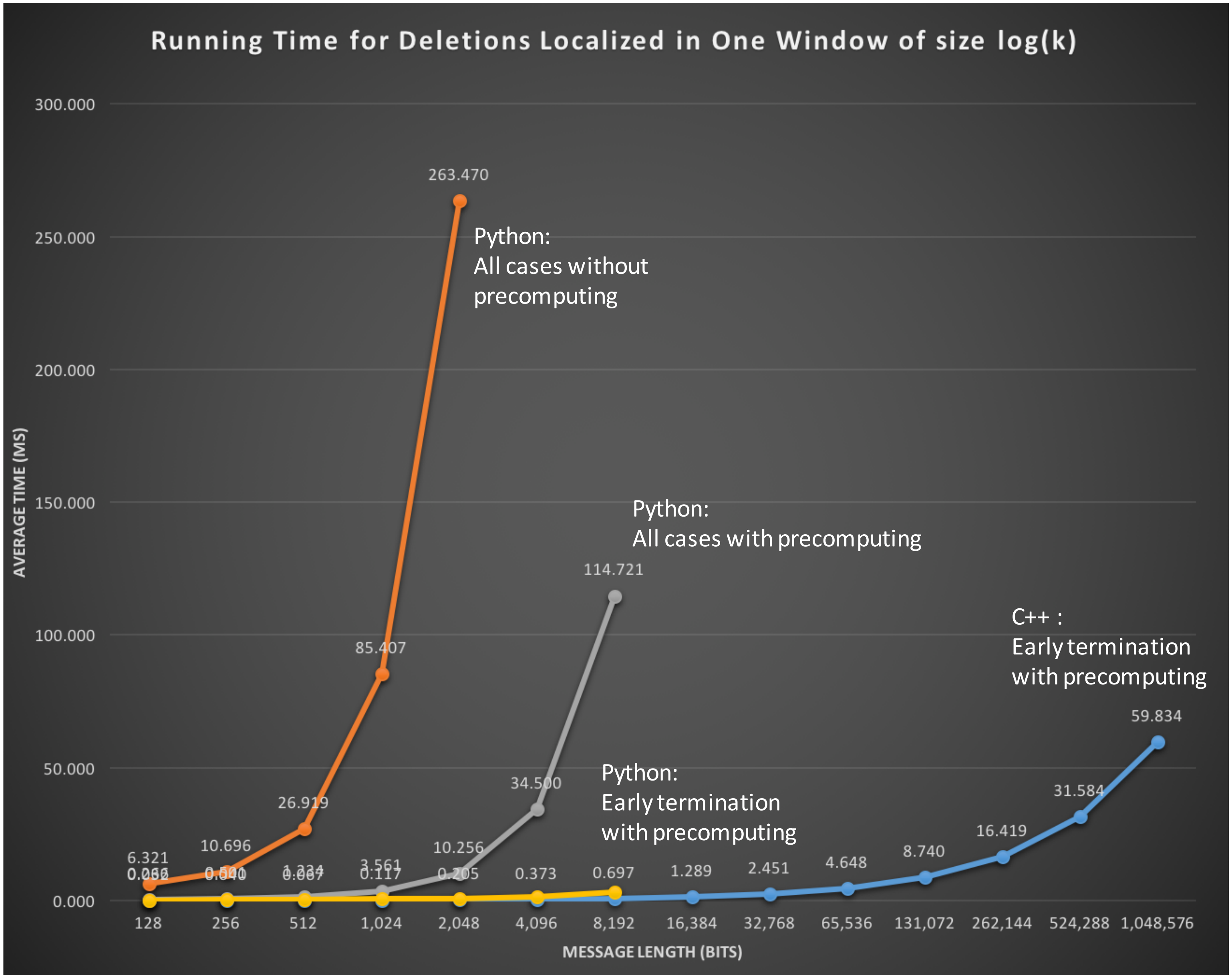
  Upload the notebook files from http://eceweb1.rutgers.edu/csi/software.html

- C++ & Python codes are available on GitHub

  GitHub repository: https://github.com/serge-k-hanna/GC

- For more details: http://eceweb1.rutgers.edu/csi/software.html

Simulations – Running Time



Decoding Failures: What Happened.

- Decoding failure: more than one possible guess, different decoded strings
- Example for one deletion: 16-bit message 0000100011110110

  2 parities

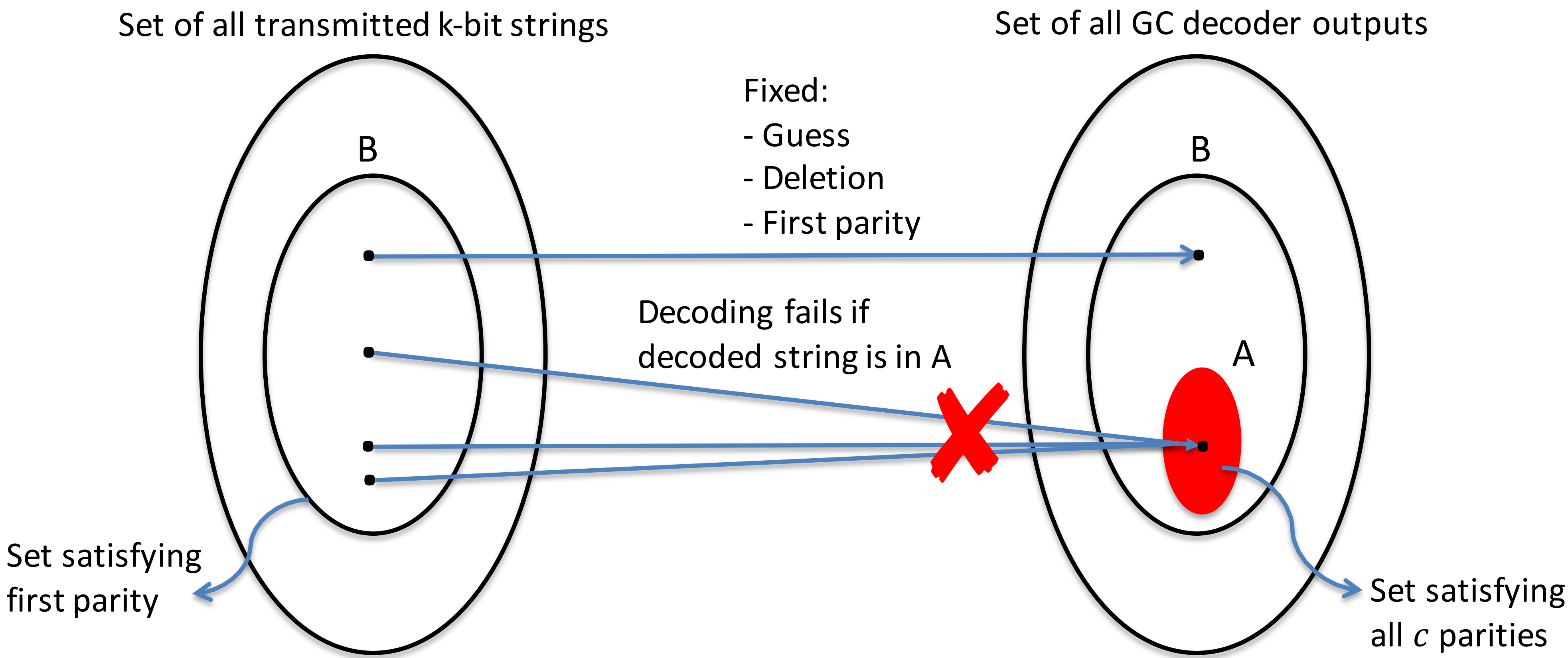
  - (6,4) MDS encoding over GF(17): | 0 | 8 | 15 | 6 | 12 | 5 |
  - Suppose 14^{th} bit gets deleted, decoding:
    - Guess 1: | 8 | 4 | 7 | 10 | ✓
    - Guess 2: | 8 | 4 | 7 | 10 | ✗
    - Guess 3: | 8 | 4 | 7 | 10 | ✗
    - Guess 4: | 0 | 8 | 15 | 6 | ✓

Guesses 1 & 4 satisfy the 2 parities

- Probability of decoding failure for a given string: combinatorial problem that depends on the string and deletion position
- Proof approach: assume message is uniform iid, average over all possible messages

Decoding Failure – 1 Deletion

Assume WLOG that Guess 1 is **correct**, observe the output of decoder at **wrong** Guess $i \neq 1$



$$Pr(\text{decoding failure in guess } i \neq 1) = Pr(\text{decoded string is in } A)$$

Lemma: at most 2 different transmitted sequences can lead to the same decoded string in any Guess $i \neq 1$

Proof of Pr(F) for One Deletion

$$
\begin{aligned}
Pr(F) &\leq Pr\left(\bigcup_{i=2}^{k/\log k} \{\mathcal{Y}_i \in A, \mathcal{Y}_i \neq \mathcal{Y}_1\}\right) && (1)\\
&\leq \sum_{i=2}^{k/\log k} Pr\left(\mathcal{Y}_i \in A, \mathcal{Y}_i \neq \mathcal{Y}_1\right) && (2)\\
&\leq \sum_{i=2}^{k/\log k} Pr\left(\mathcal{Y}_i \in A\right) && (3)\\
&= \sum_{i=2}^{k/\log k} \sum_{Y \in A} Pr\left(\mathcal{Y}_i = Y\right) && (4)\\
&\leq \sum_{i=2}^{k/\log k} \sum_{Y \in A} \frac{2}{|B|} && (5)\\
&= 2\left(\frac{k}{\log k} - 1\right)\frac{|A|}{|B|} && (6)\\
&= 2\left(\frac{k}{\log k} - 1\right)\frac{q^{k/\log k - c}}{q^{k/\log k - 1}} && (7)\\
&< \frac{2}{k^{c-2} \log k}. && (8)
\end{aligned}
$$

Union bound

Lemma

Subspace cardinality

k : length of message

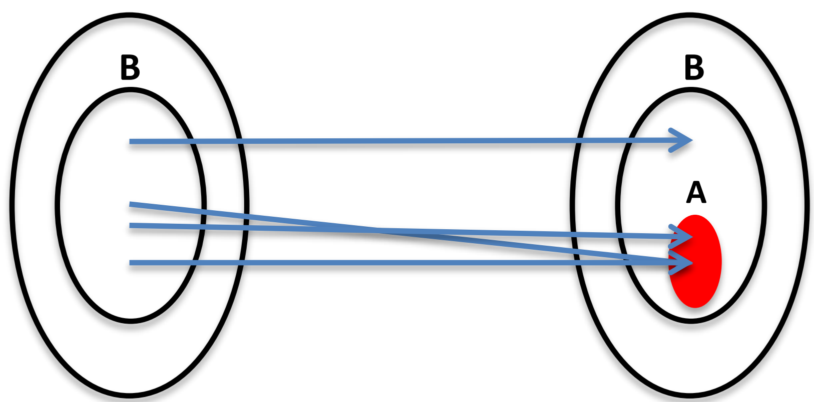
$\mathcal{Y}_i$: string decoded in Guess i

c : number of parities

A : set satisfying all c parities

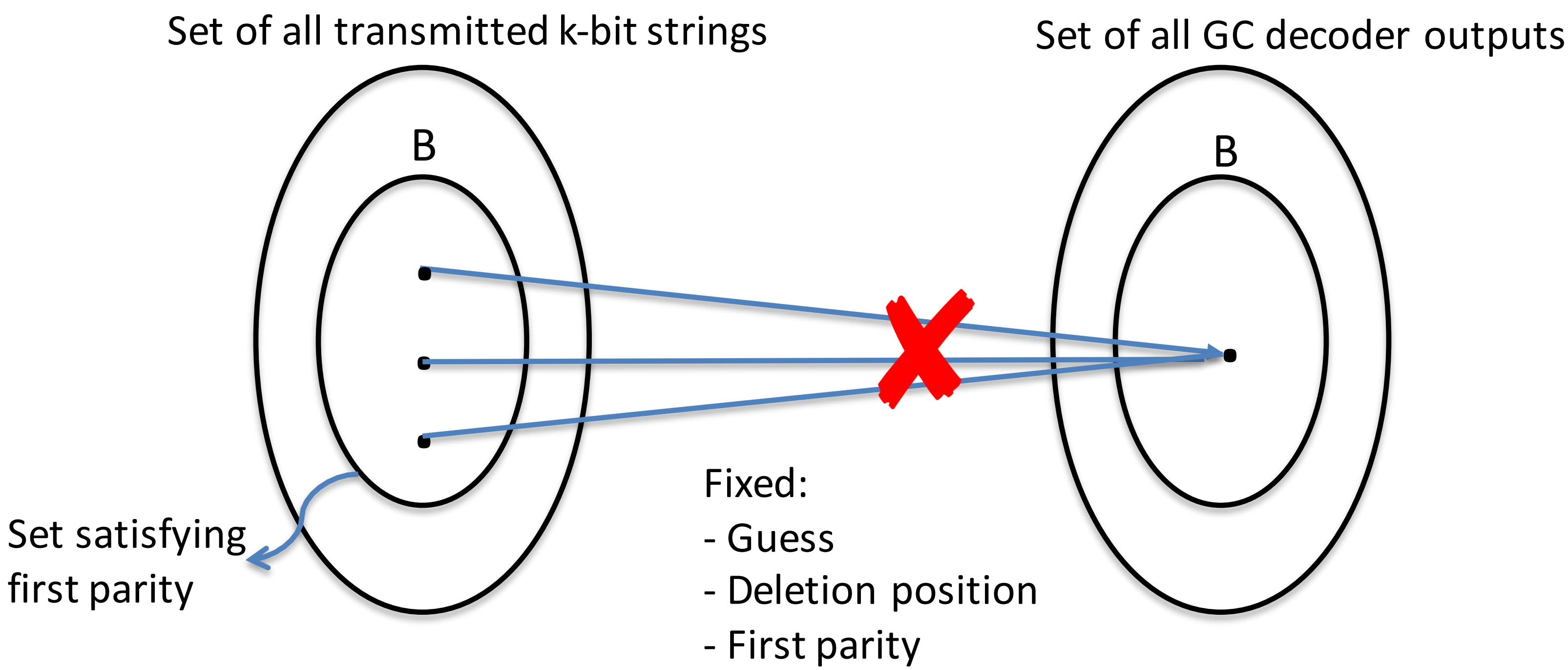
B : set satisfying first parity

q : field size



Claim

- Claim 1 (one deletion): at most 2 different transmitted strings can lead to the same decoded string in any wrong guess



- Claim 2 (δ deletions): a **constant** number of different transmitted strings can lead to the same decoded string in any wrong guess

Claim - Example

- Claim 1 (one deletion): at most 2 different transmitted strings can lead to the same decoded string in any wrong guess
- Example

$$\mathbf{u_1} = 0000000000000000 \xrightarrow{\text{Encoding in GF(16)}} \boxed{0}\ \boxed{0}\ \boxed{0}\ \boxed{0}\ \boxed{0}$$

$$\mathbf{u_2} = 0010000000000010 \xrightarrow{\text{Encoding in GF(16)}} \boxed{\alpha}\ \boxed{0}\ \boxed{0}\ \boxed{\alpha}\ \boxed{0}$$

parity

- 3^{rd} bit deleted; Guess: deletion occurred in 4^{th} block

Received

$$\mathbf{y_1}\ \boxed{0000}\ \boxed{0000}\ \boxed{0000}\ \cancel{000}$$

$$\mathbf{y_2}\ \boxed{0000}\ \boxed{0000}\ \boxed{0000}\ \cancel{010}$$

Decode $\rightarrow$ 0000 0000 0000 0000

- $\mathbf{u_3} = 1111111111111111$ and $\mathbf{u_4} = 0010000000000000$ ✗
- Two conditions: (1) Symmetry constraint; (2) Algebraic linear constraint

Claim

- Suppose 3rd bit is deleted, guess : deletion occurred in 4th block

$$\text{GF(17)}\quad \underbrace{b_1b_2b_4b_5}_{8b_1+4b_2+2b_4+b_5}\quad \underbrace{b_6b_7b_8b_9}_{8b_6+4b_7+2b_8+b_9}\quad \underbrace{b_{10}b_{11}b_{12}b_{13}}_{8b_{10}+4b_{11}+2b_{12}+b_{13}}\quad \underbrace{\cancel{b_{14}b_{15}b_{16}}}_{?}$$

Symbol 1   Symbol 2   Symbol 3   Erasure

- How many different messages can lead to same decoded string?
- **Symmetry constraint**: same decoded string $\Rightarrow$ same bit values at positions of symbols 1, 2 and 3
- Bits which can be different: b_{14}, b_{15}, b_{16} and b_3 (deleted bit)
- Algebraic constraint: erasure is decoded using first parity

$$4b_{14} + 2(b_3 + b_{15}) + b_{16} = p_1$$

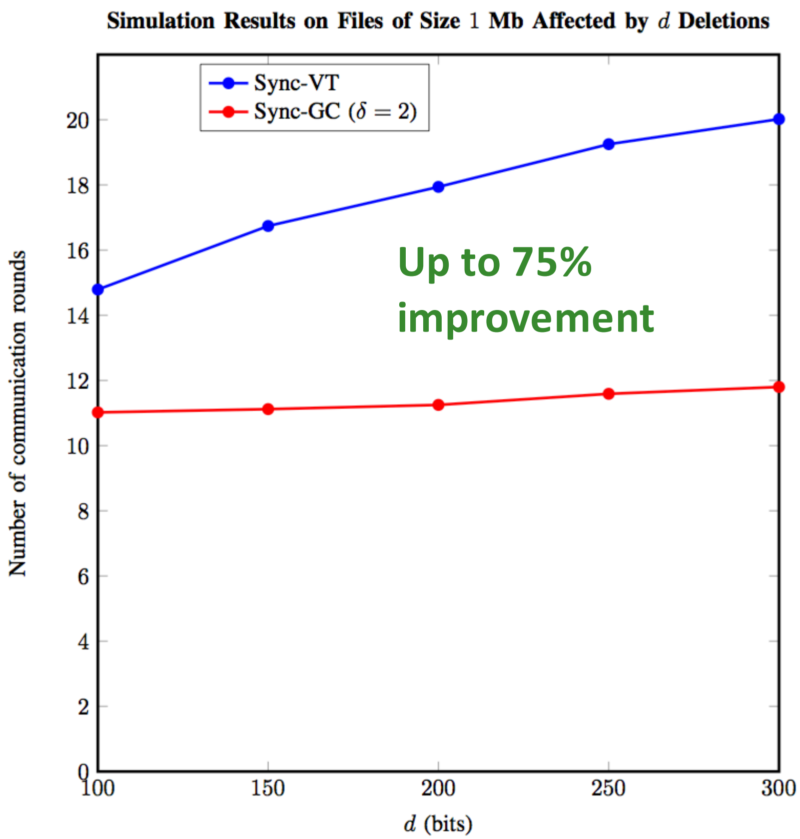
$$b_{14}, b_{15}, b_{16}, b_3 \in GF(2)$$

$$p_1 \in GF(17)$$

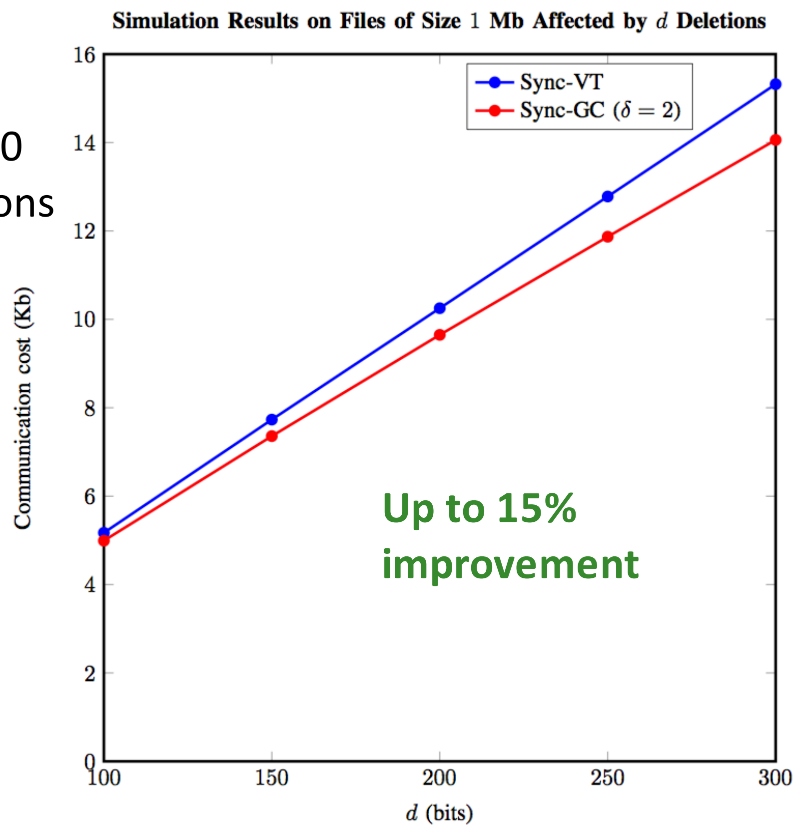
Equation has at most 2 solutions

Application to File Synchronization

- Interactive synchronization algorithm by [Venkataramanan et al. ’15]
  - ➢ Isolate single deletions, use VT codes
  - ➢ Modification: isolate δ or fewer deletions, use GC codes
- Gain: (1) less communication rounds, (2) lower communication cost



N=1000 iterations



Summary

- Guess & Check Codes for localized deletions
  - For single or multiple windows
  - Explicit code construction with logarithmic redundancy
  - Deterministic polynomial time encoding and decoding
  - Asymptotically vanishing probability of decoding failure

- Open problems
- Capacity of deletion channel with localized deletions?
- Codes for adversarial localized deletions
- And of course for “unrestricted” deletion capacity and codes are still open problems

Questions?