Binary  ing codes – can we prove that someone is guilty?!

Marcel Fernández (1)

Elena Egorova (2)

Grigory Kabatiansky (3)

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

(1) Department of Network Engineering
Technical University of Catalonia
Barcelona, Spain

NATIONAL RESEARCH UNIVERSITY

(2) Faculty of Computer Science,
National Research University Higher School of Economics, *Moscow, Russia*

RUSSIAN ACADEMY OF SCIENCES
INSTITUTE FOR INFORMATION TRANSMISSION PROBLEMS

(3) Institute for Information Transmission Problems
Russian Academy of Sciences
Moscow, Russia

WIFS 2015 – Rome, Italy, 16-19 November, 2015

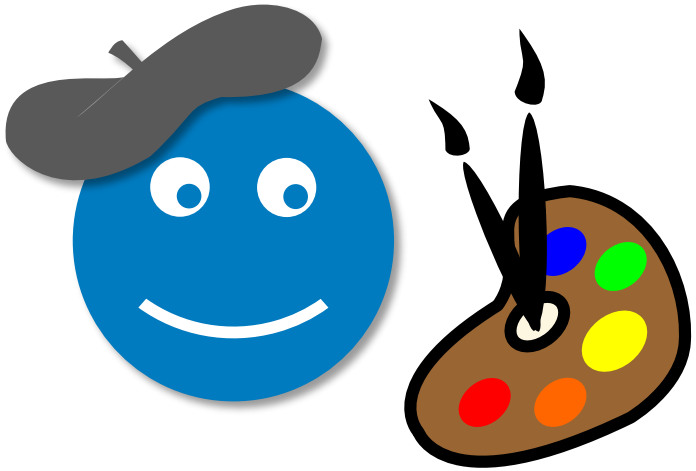
Outline

1. Introduction and background
2. Codes with traceability properties
3. Families of collusion-secure digital fingerprinting codes
4. Families of almost *t*-IPP codes or provably secure family of digital fingerprinting codes
5. Almost *2*-IPP codes (the case of two pirates)
6. More then 2 pirates: negative result for almost t-IPP codes
7. Relaxed version of almost t-IPP codes
8. Summary and conclusion

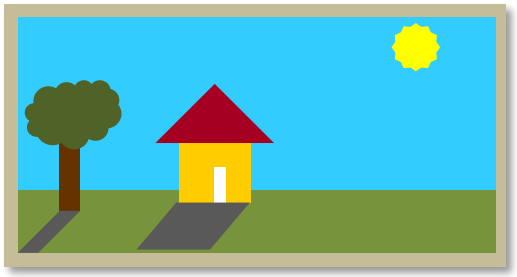
1

Introduction and Background

The Redistribution Problem

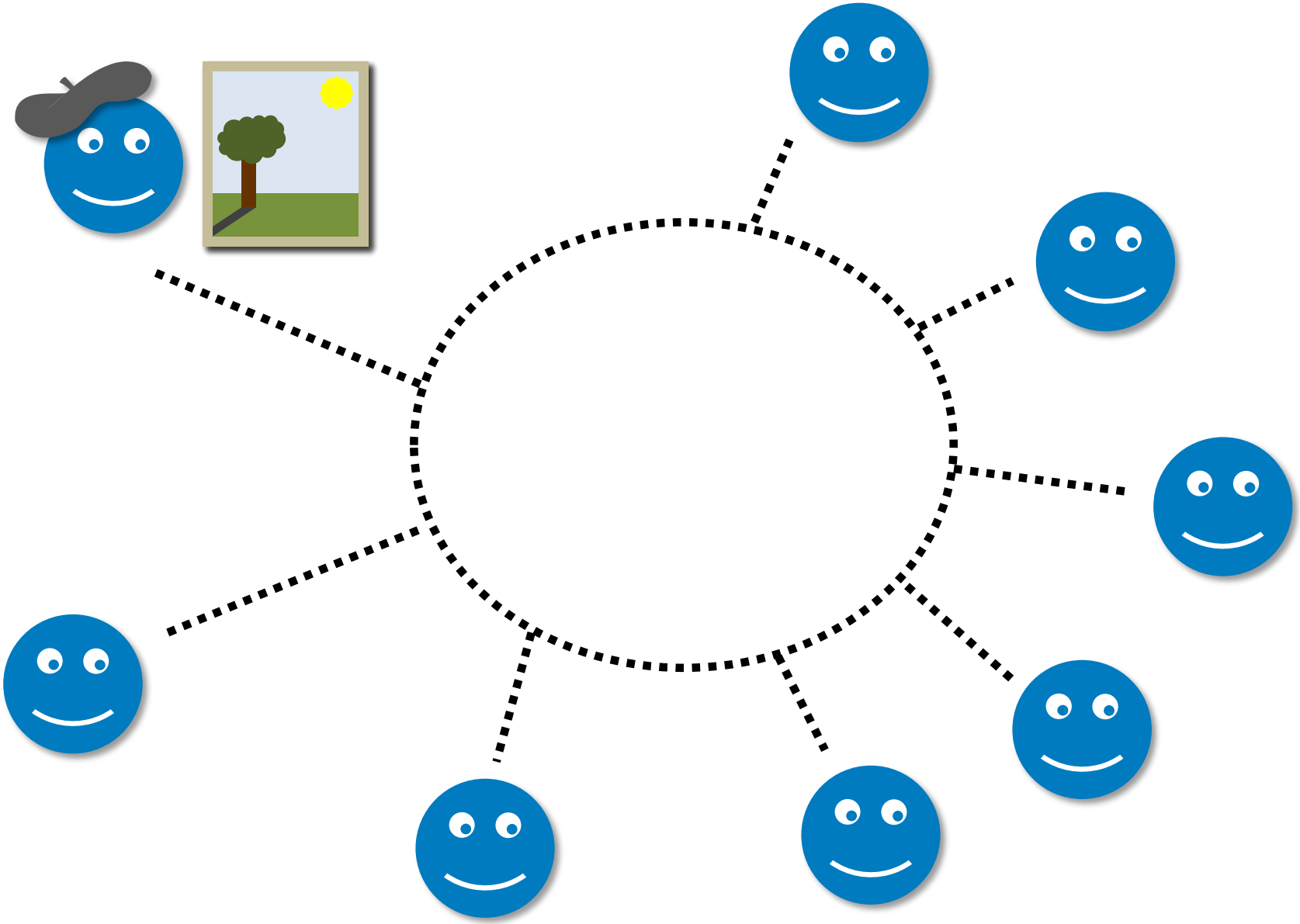


AUTHOR

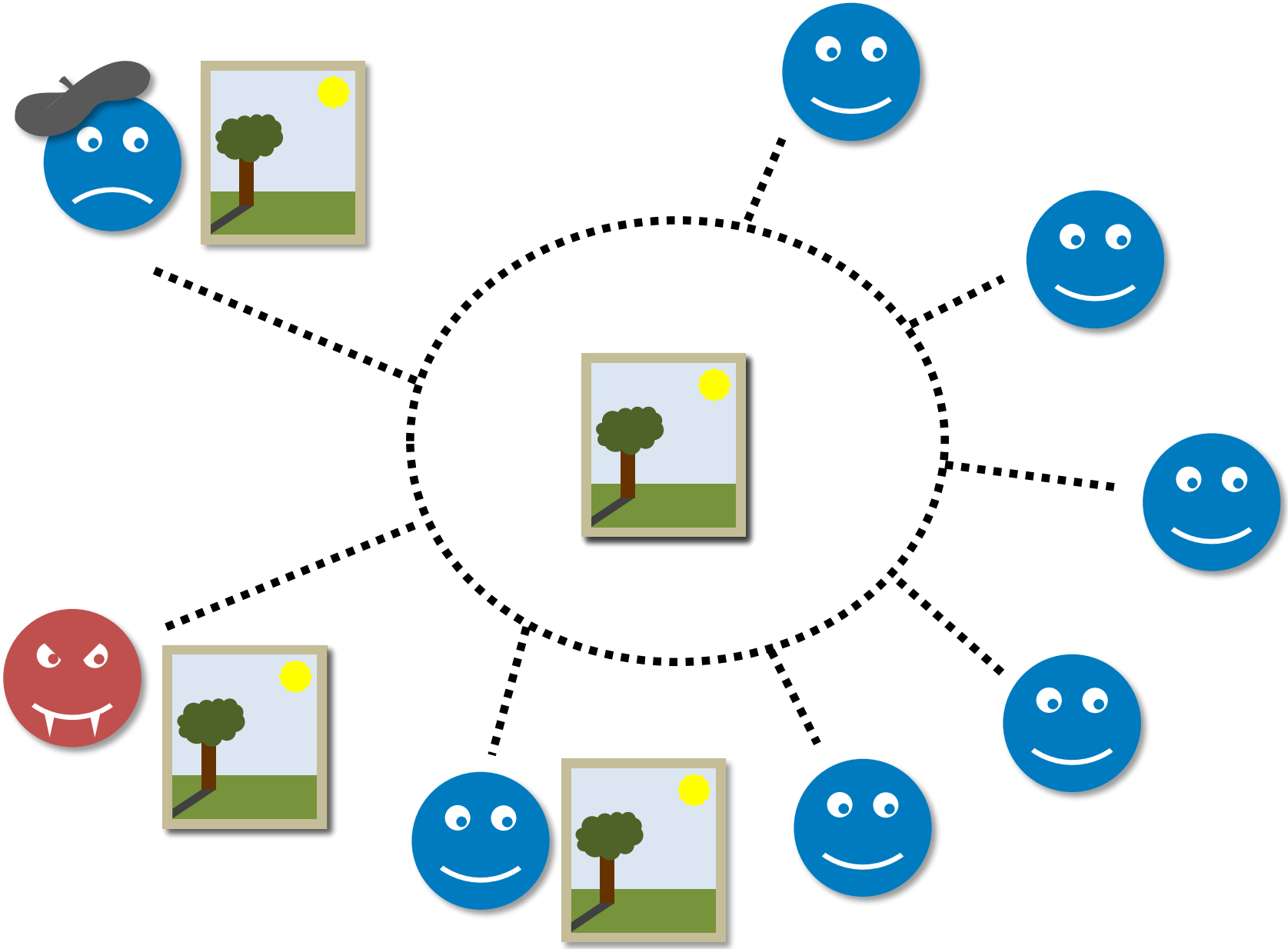


WORK

The Redistribution Problem



The Redistribution Problem



The Redistribution Problem

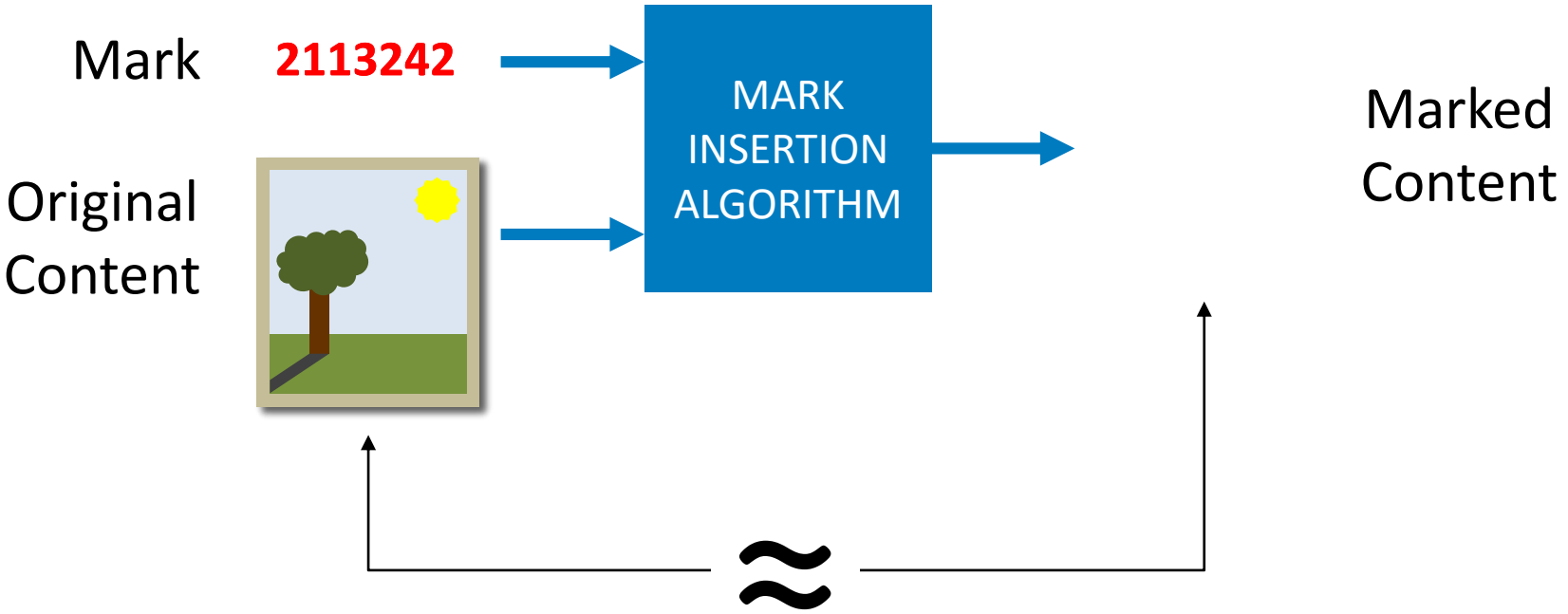
We would like to find the source of **leaked information** when we are dealing with “problematic” data, including, but not limited to:

- personal documents,
- industrial secrets,
- classified information,
- copyrighted material,
- etc.

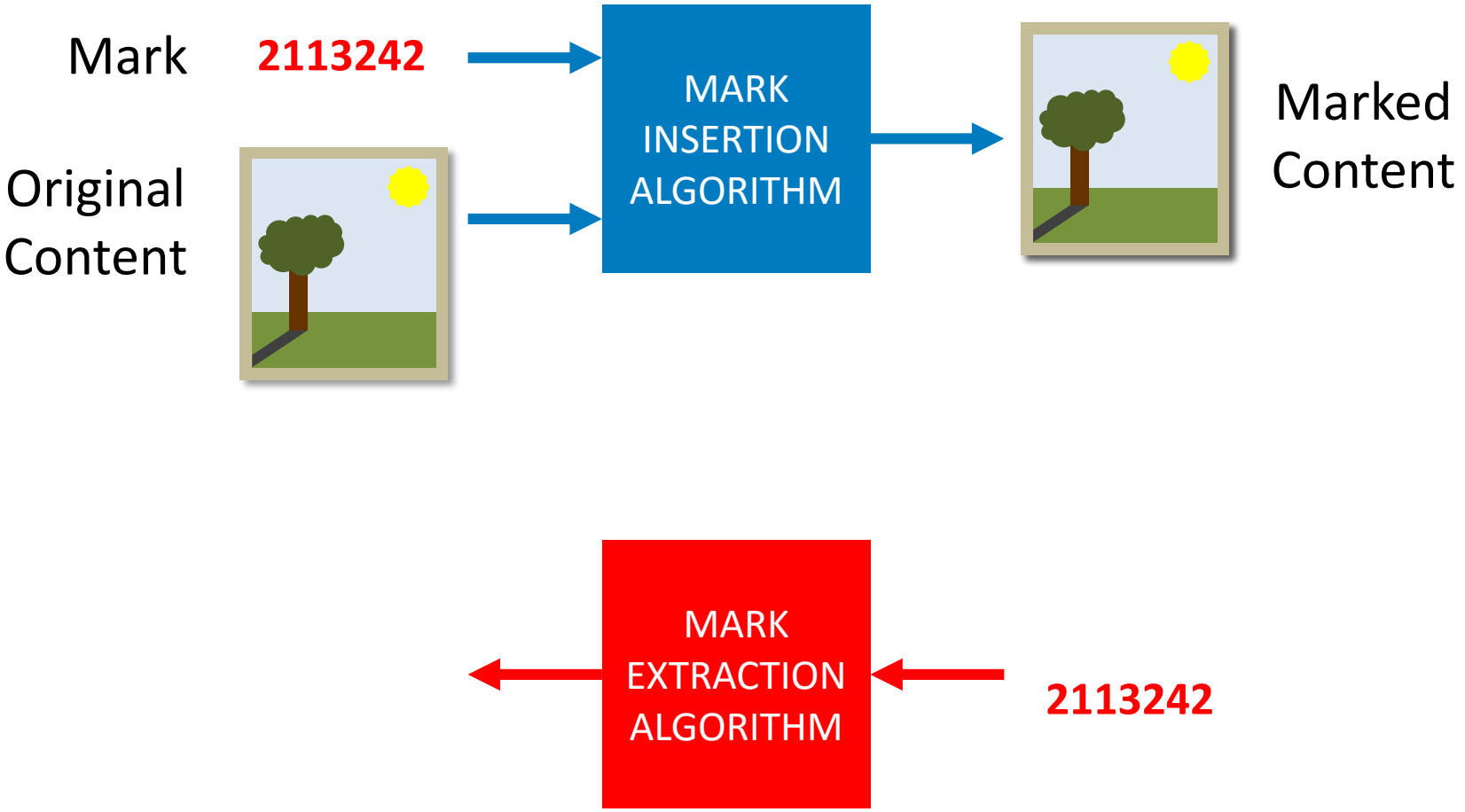
The Redistribution Problem

- Redistribution is very difficult to avoid.
- However, by marking each copy of the content, the distributor can deter **plain redistribution**.
- We remark that we will not focus on the (nontrivial) problem of how marks are inserted in the content.

The Redistribution Problem



The Redistribution Problem



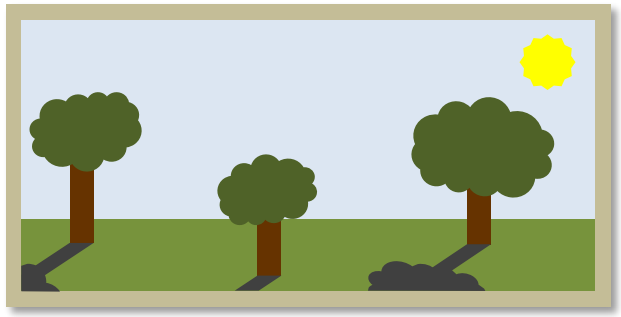
The Redistribution Problem

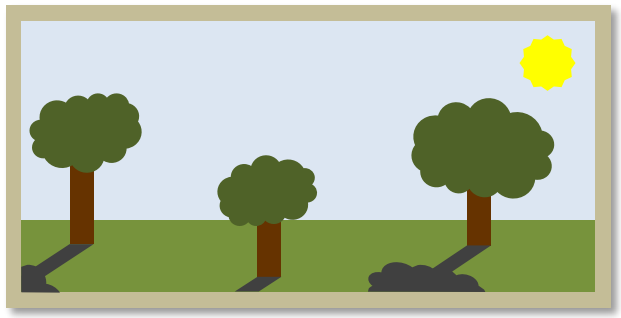
- Redistribution is very difficult to avoid.
- However, by marking each copy of the content, the distributor can deter **plain redistribution**.
- We remark that we will not focus on the (nontrivial) problem of how marks are inserted in the content.
- The focus of our study will be the design of the set of user marks, known as **fingerprinting code**.

Collusion Attacks

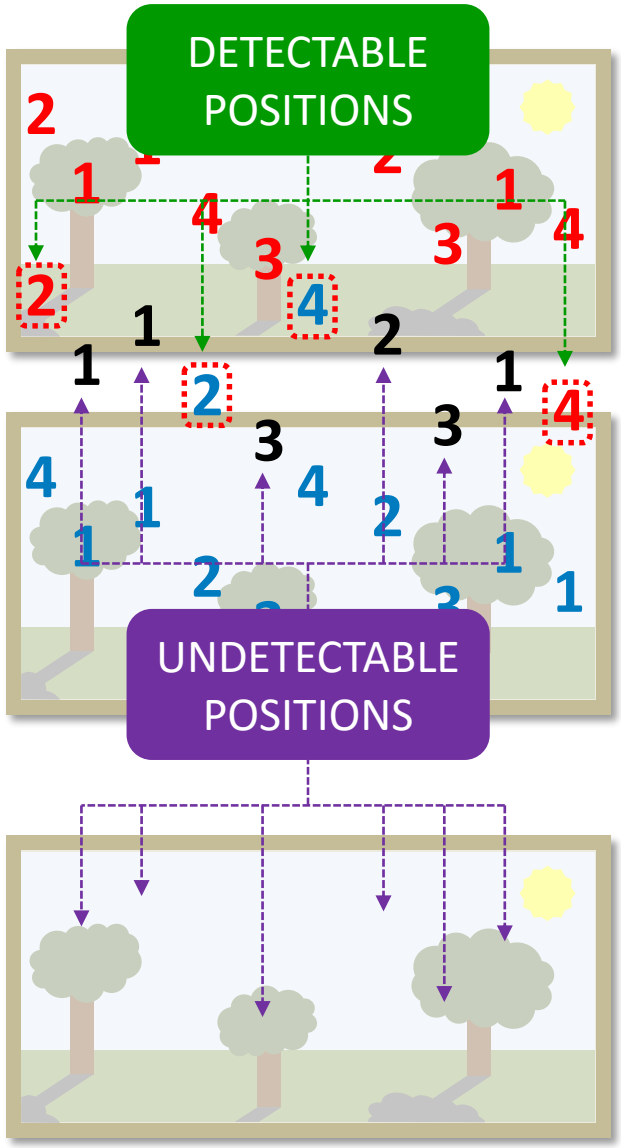
Weakness of the
fingerprinting technique

Collusion Attacks





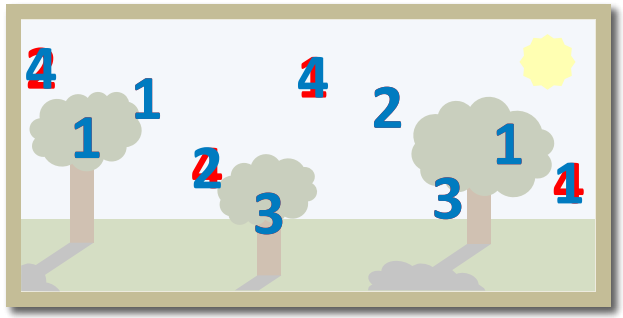
Collusion Attacks

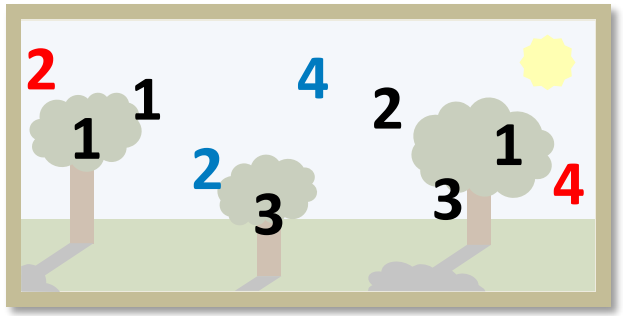


The traitors spot differences where their user's marks are different.

Marking Assumption: The undetectable positions remain unchanged in the pirated content.

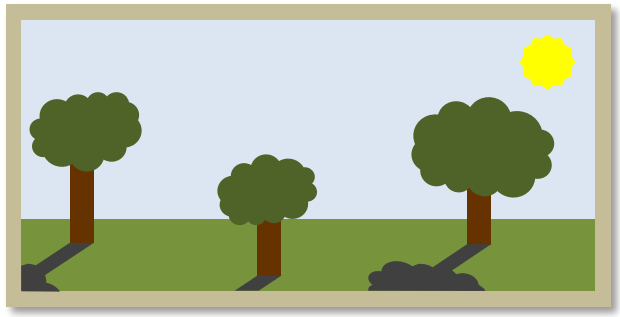
Collusion Attacks

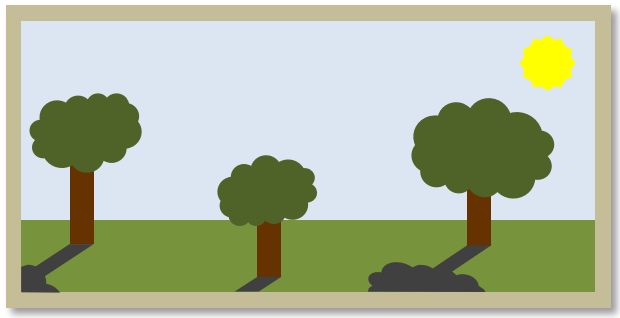


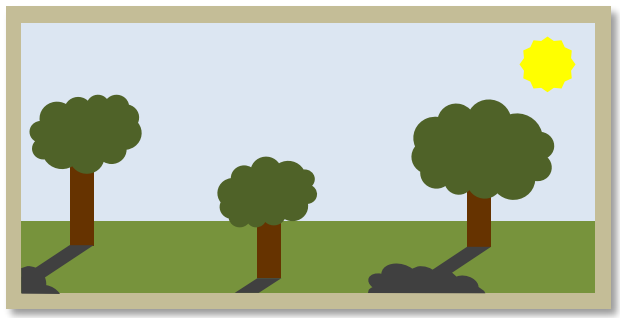


The goal of the **pirated content** is to disguise the identity of the traitors.

Collusion Attacks

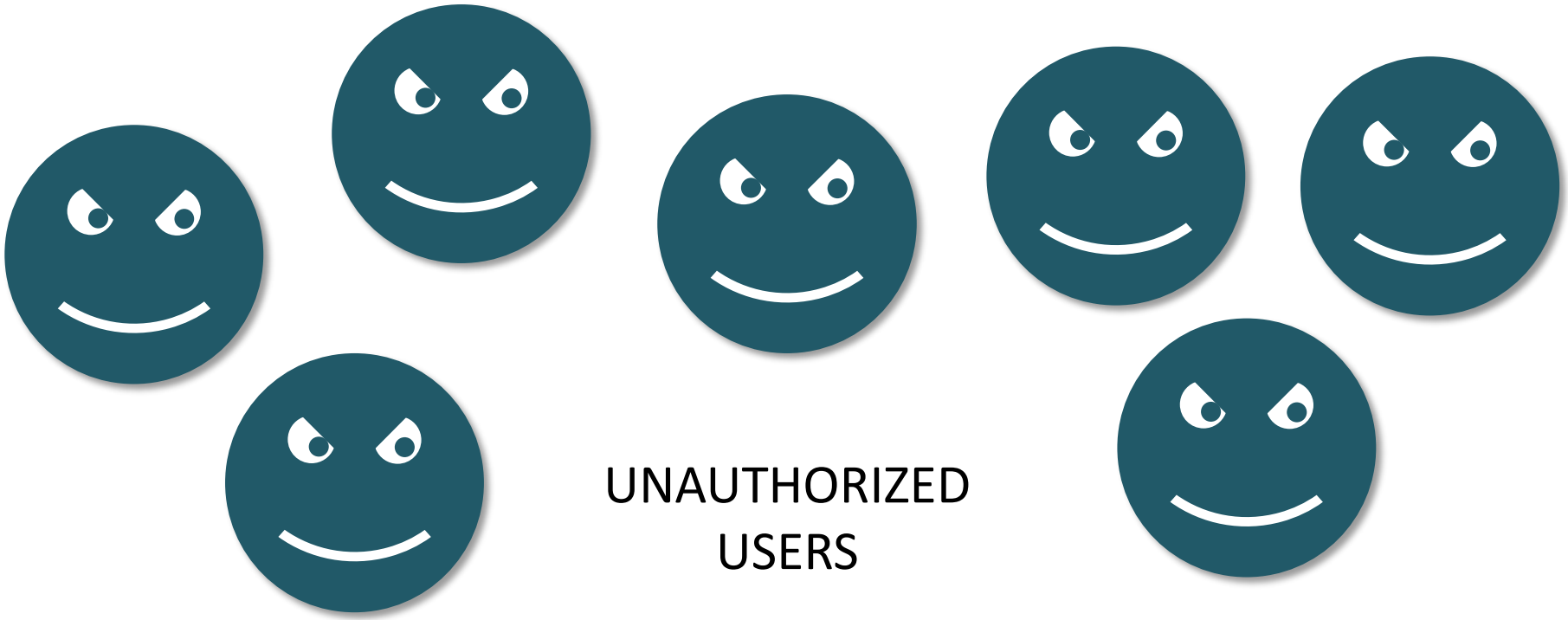




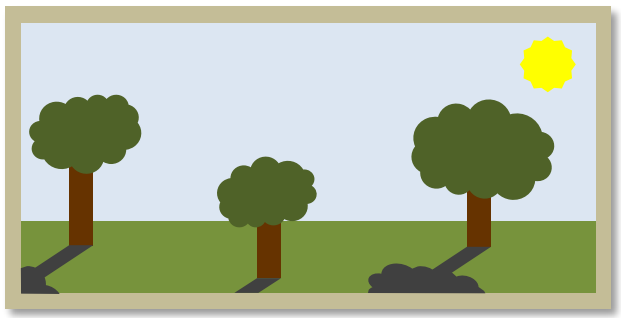


ILLEGITIMATE REDISTRIBUTION

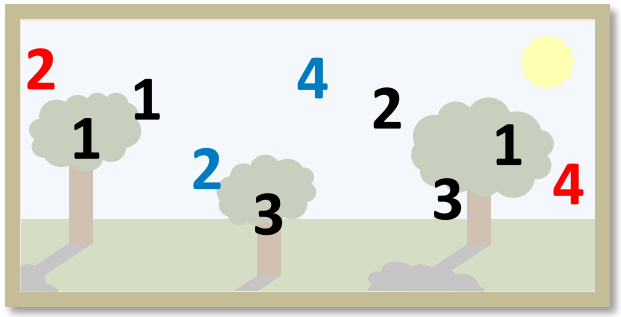
Collusion Attacks



UNAUTHORIZED
USERS



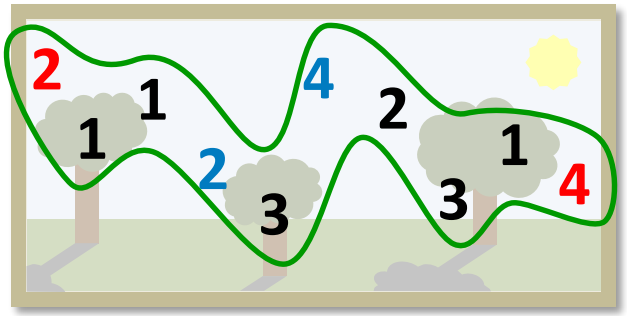
Collusion Attacks



Collusion Attacks

The Fingerprinting Problem:

- Can we identify a traitor using this information?
- What is the error in the identification process?
- How should we design the user marks ?

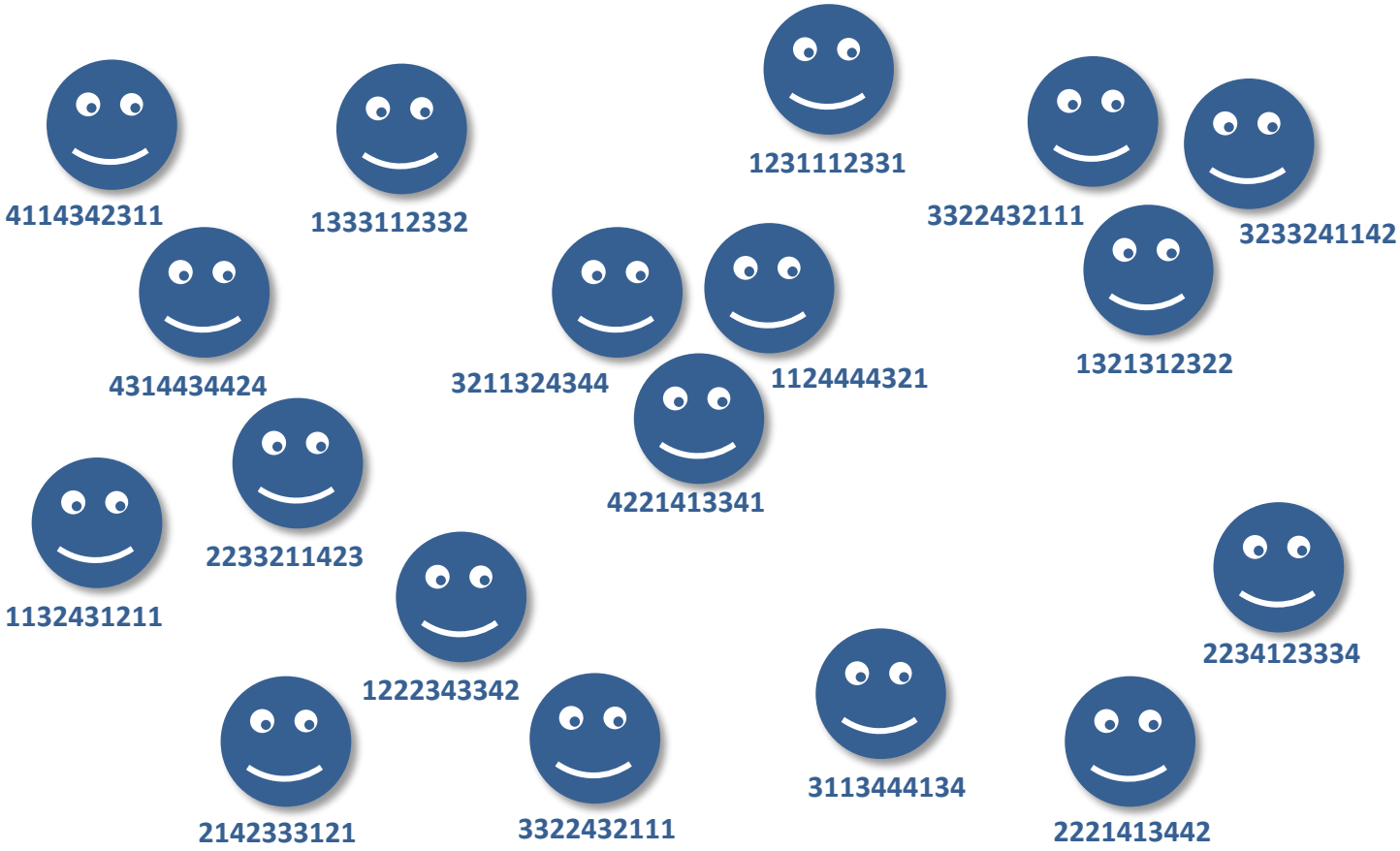


2

Codes with Traceability Properties

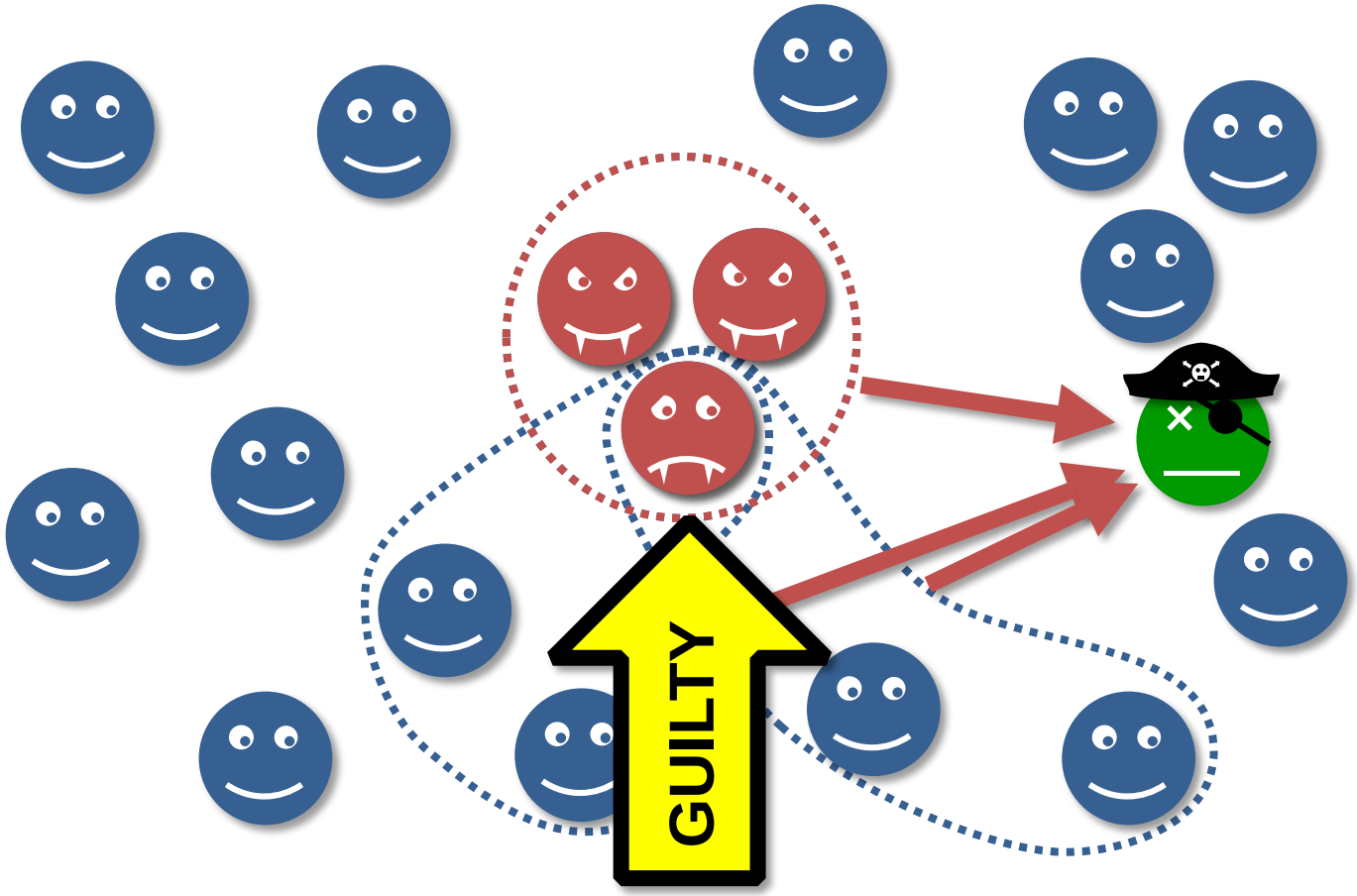
Codes with Traceability Properties

We assume that each codeword of a fingerprinting code identifies an unique user



Codes with Traceability Properties

Identifiable Parent Property Codes (**IPP**): identify, at least, one actual traitor.



t-IPP

IPP-codes over binary alphabet do not exist!

It turns out that in the binary case

a single code is not enough!

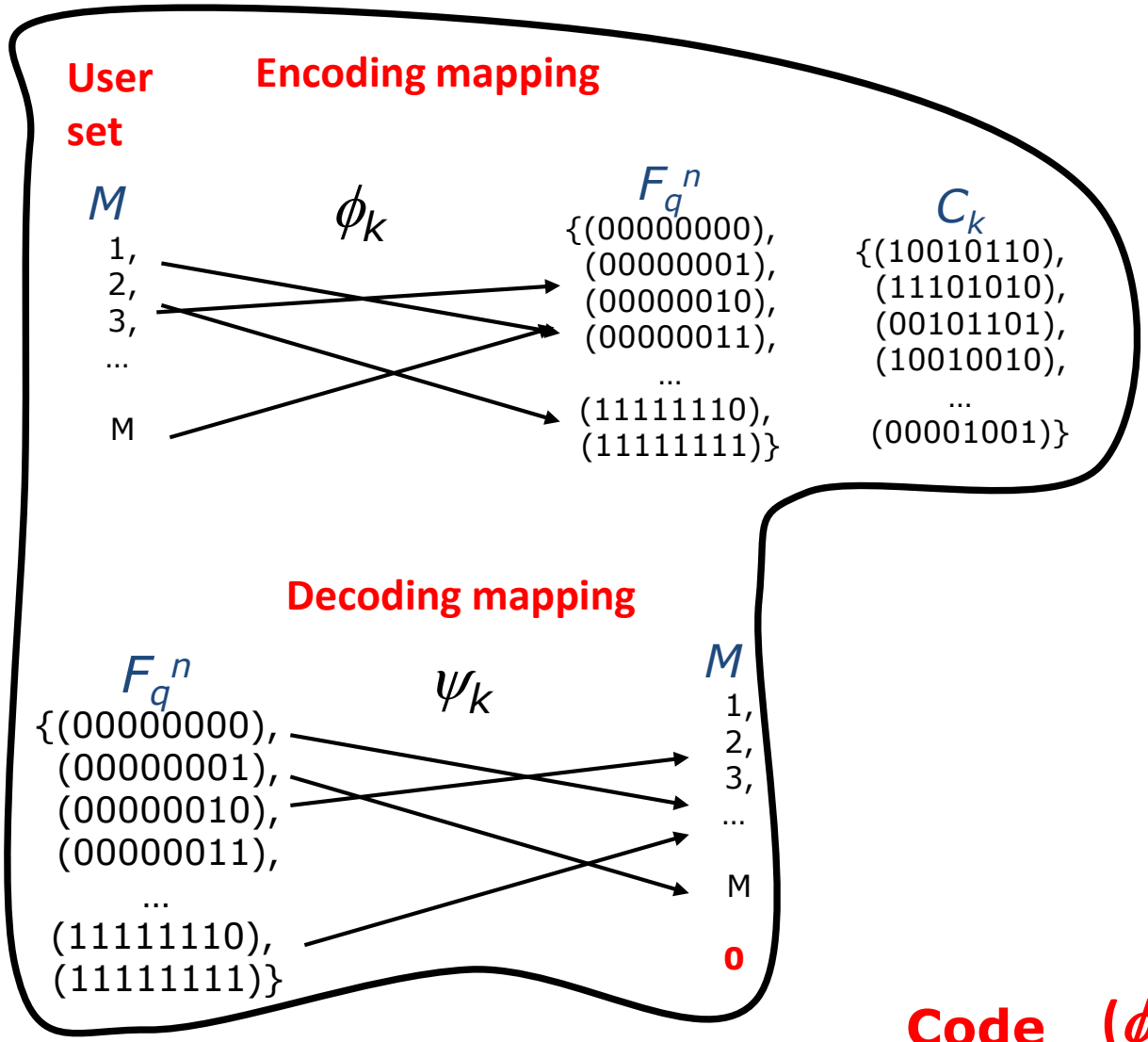
We need a family of codes !

3

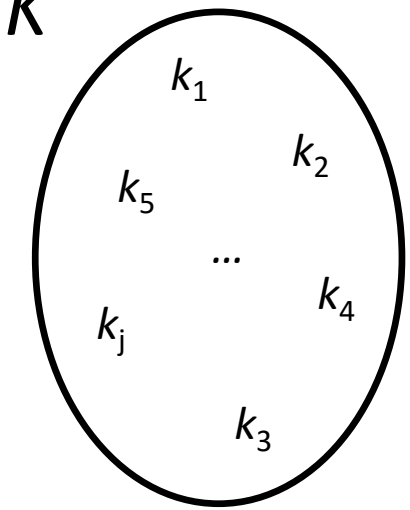
Families of Collusion-Secure Binary Digital Fingerprinting Codes

Families of Digital Fingerprinting Codes

For every key k



Family of codes

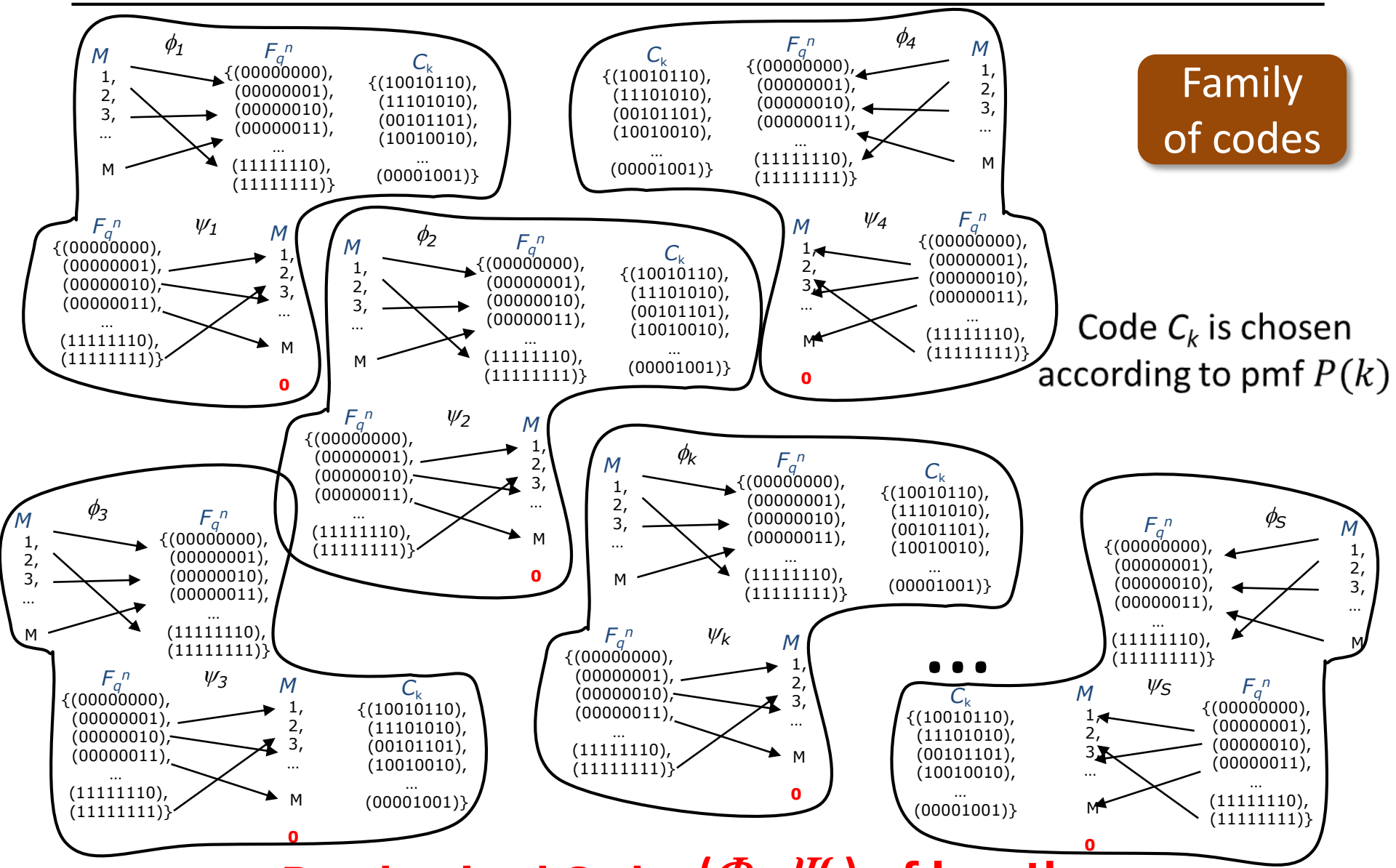


Probability Distribution

$P_K(k)$

Code (ϕ_k, ψ_k) of length n

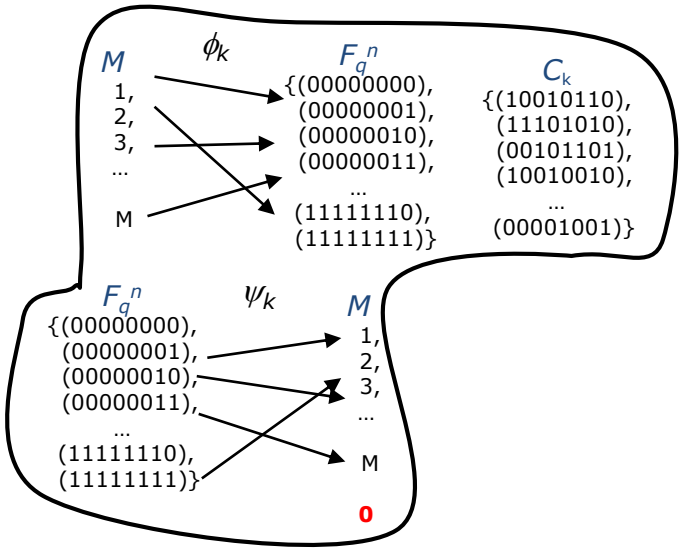
Families of Digital Fingerprinting Codes



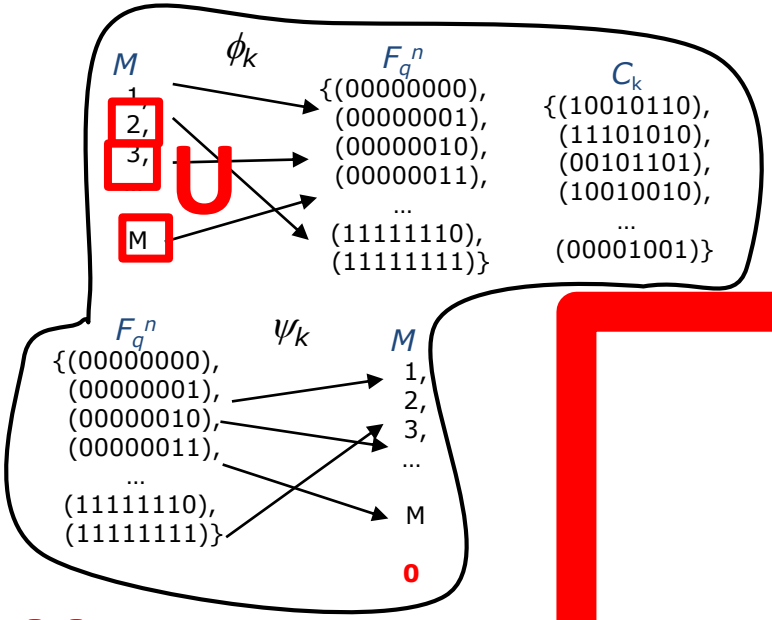
Families of Digital Fingerprinting Codes

Family of codes

Code C_k is chosen according to pmf $P(k)$



Families of Digital Fingerprinting Codes



Code C_k is chosen according to pmf $P(k)$

Family of codes

A family of codes (Φ_k, Ψ_k) is said to be a t-collusion secure digital fingerprinting code with ε error if

$$\max_{\text{coalitions } U} \left(\text{Average Probability}_{\text{codes } C_k} \left(\text{identification error for a coalition } U \text{ and any strategy } P \right) \right) < \varepsilon$$

U

ϕ_k

2 ⟶ (10010110) = $\boldsymbol{v}_1$

3 ⟶ (00101101) = $\boldsymbol{v}_2$

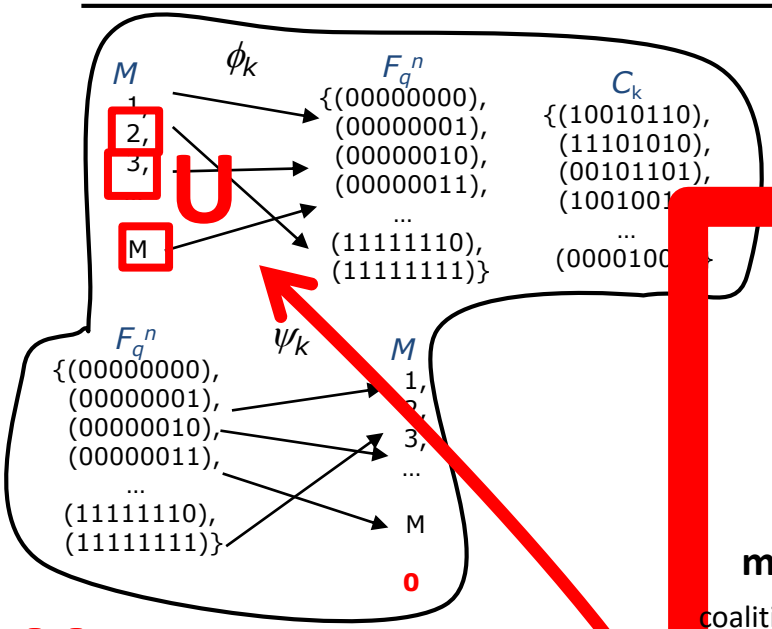
... ...

M ⟶ (00001001) = $\boldsymbol{v}_t$

Strategy $\mathbf{P}(\cdot \mid \cdot, \ldots,$

(10010110) = ***pirate x***

Families of Digital Fingerprinting Codes



Code C_k is chosen according to pmf $P(k)$

Family of codes

A family of codes (Φ_k, Ψ_k) is said to be a t-collusion secure digital fingerprinting code with ε error if

$$\max_{\text{coalitions } U} \left(\text{Average Probability}_{\text{codes } C_k} \left(\begin{array}{c} \text{identification error} \\ \text{for a coalition } U \\ \text{and any strategy } P \end{array} \right) \right) < \varepsilon$$

U

ϕ_k

2 ⟶ (10010110) = $\boldsymbol{v}_1$

3 ⟶ (00101101) = $\boldsymbol{v}_2$

... ...

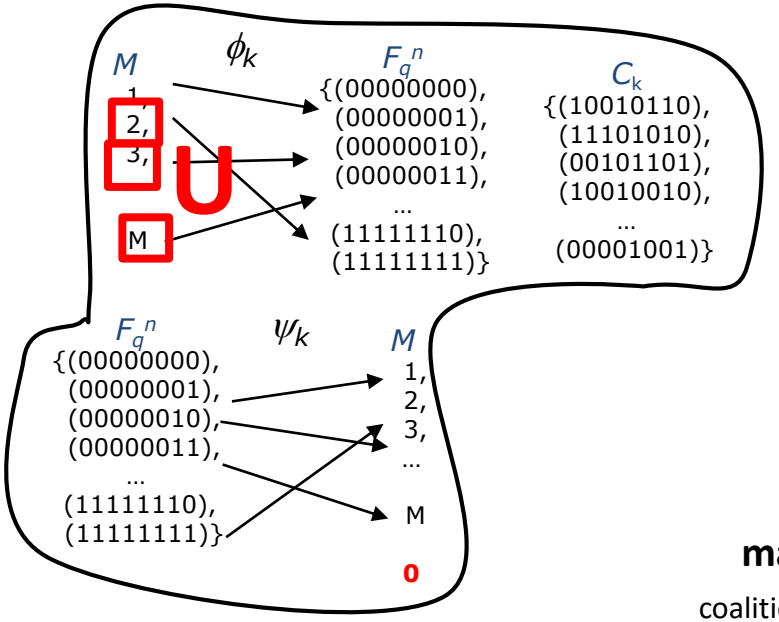
M ⟶ (00001001) = $\boldsymbol{v}_t$

Strategy $P(\cdot \mid \cdot, \ldots, \cdot)$

(10010110) = ***pirate x***

With high probability identifies a member of coalition U that created *x*

Families of Digital Fingerprinting Codes



Code C_k is chosen according to pmf $P(k)$

Family of codes

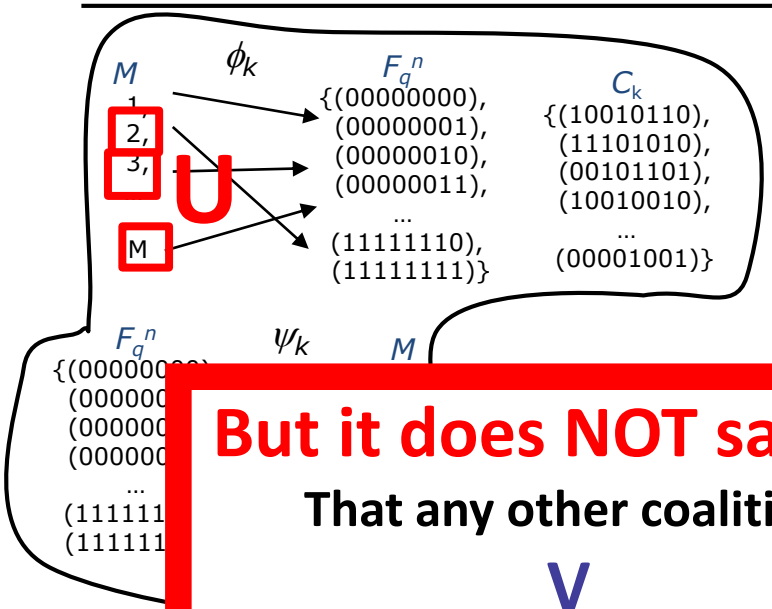
A family of codes (Φ_k, Ψ_k) is said to be a t-collusion secure digital fingerprinting code with ε error if

$$\max_{\text{coalitions } U} \left(\text{Average Probability}_{\text{codes } C_k} \left(\text{identification error for a coalition } U \text{ and any strategy } P \right) \right) < \varepsilon$$

A number **R** is an **ε-achievable rate** for t-fingerprinting if for every $\delta > 0$ there exists a randomized (Φ_k, Ψ_k) code of (sufficiently large) length n with

$$\frac{1}{n}\log_q M > R - \delta \quad \text{and}$$

Families of Almost *t*-IPP Codes



But it does NOT say that

That any other coalition

V

that can also create x

Should contains u

Family of codes

ϕ_k

U

2 → (10010110) = $\boldsymbol{v}_1$

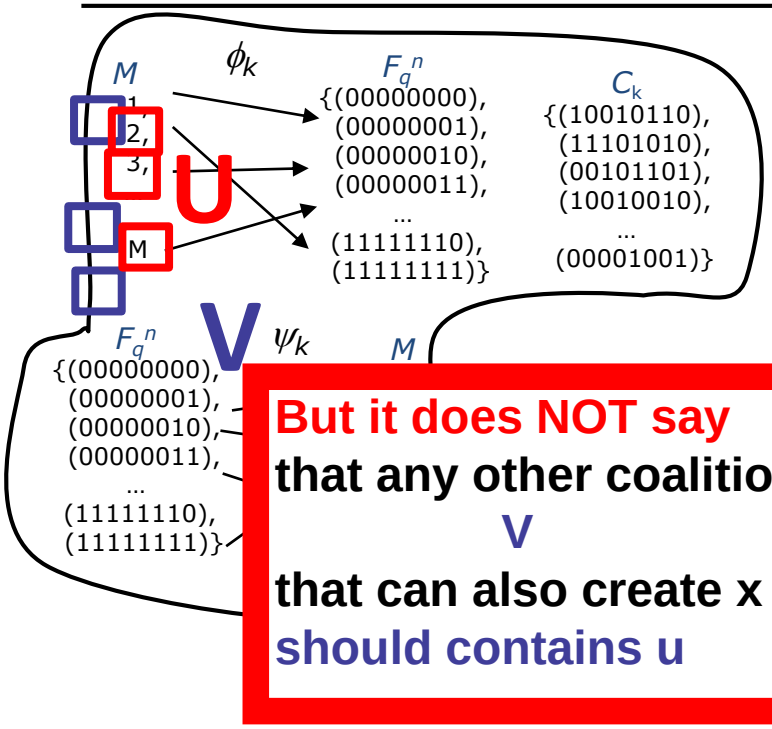
3 → (00101101) = $\boldsymbol{v}_2$

6 → (00001001) = $\boldsymbol{v}_6$

Strategy $\mathbf{P}(\cdot \mid \cdot, \ldots, \cdot)$

(10010110) = ***pirate x***

Families of Almost *t*-IPP Codes



Family of codes

ϕ_k

U

2 → (10010110) = $\boldsymbol{v}_1$

3 → (00101101) = $\boldsymbol{v}_2$

6 → (00001001) = $\boldsymbol{v}_6$

Strategy $\mathbf{P}(\cdot \mid \cdot, \ldots, \cdot)$

(10010110) = ***pirate x***

Strate … $\cdot, \ldots$

V

4 →

8 → = $\boldsymbol{v}_8$

10 →

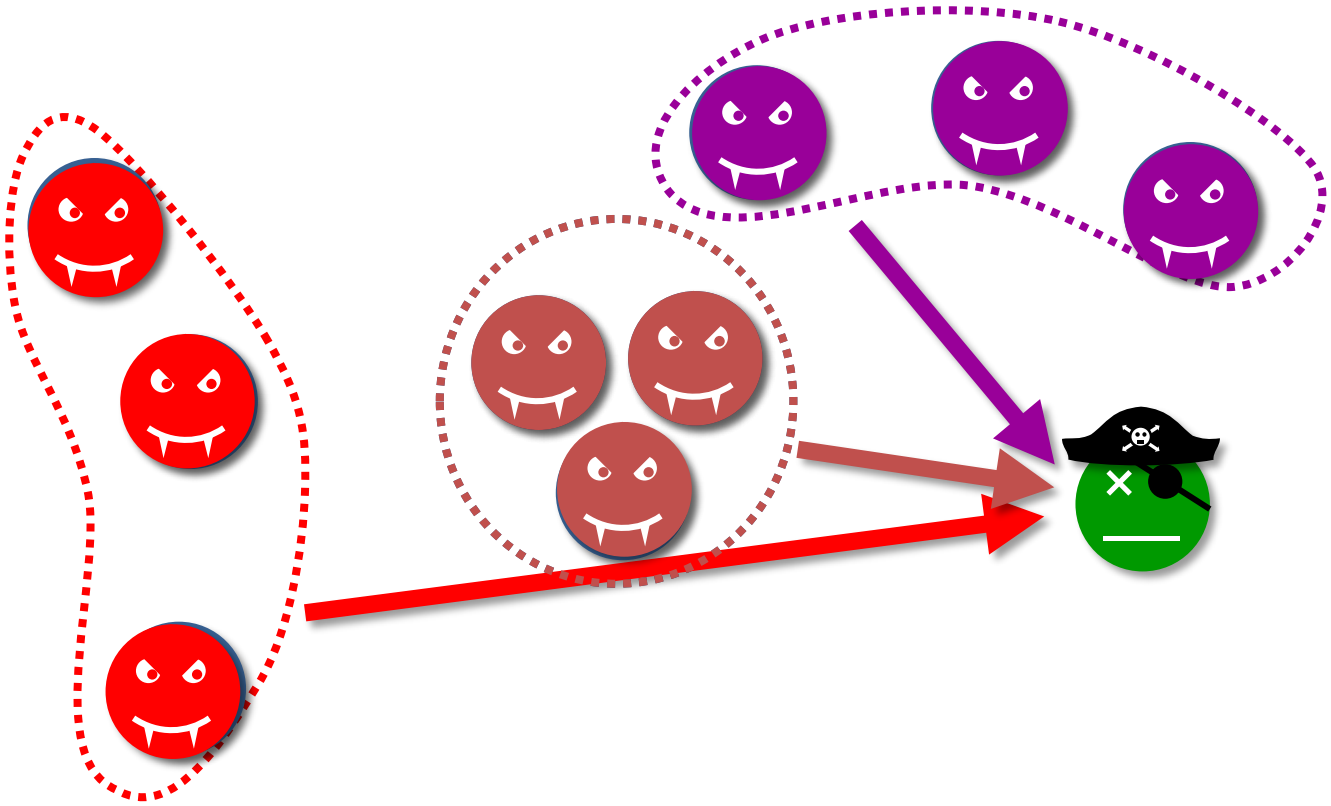
The code is NOT an IPP code!!

4

Almost *t*-IPP-codes

Families of Almost *t*-IPP Codes

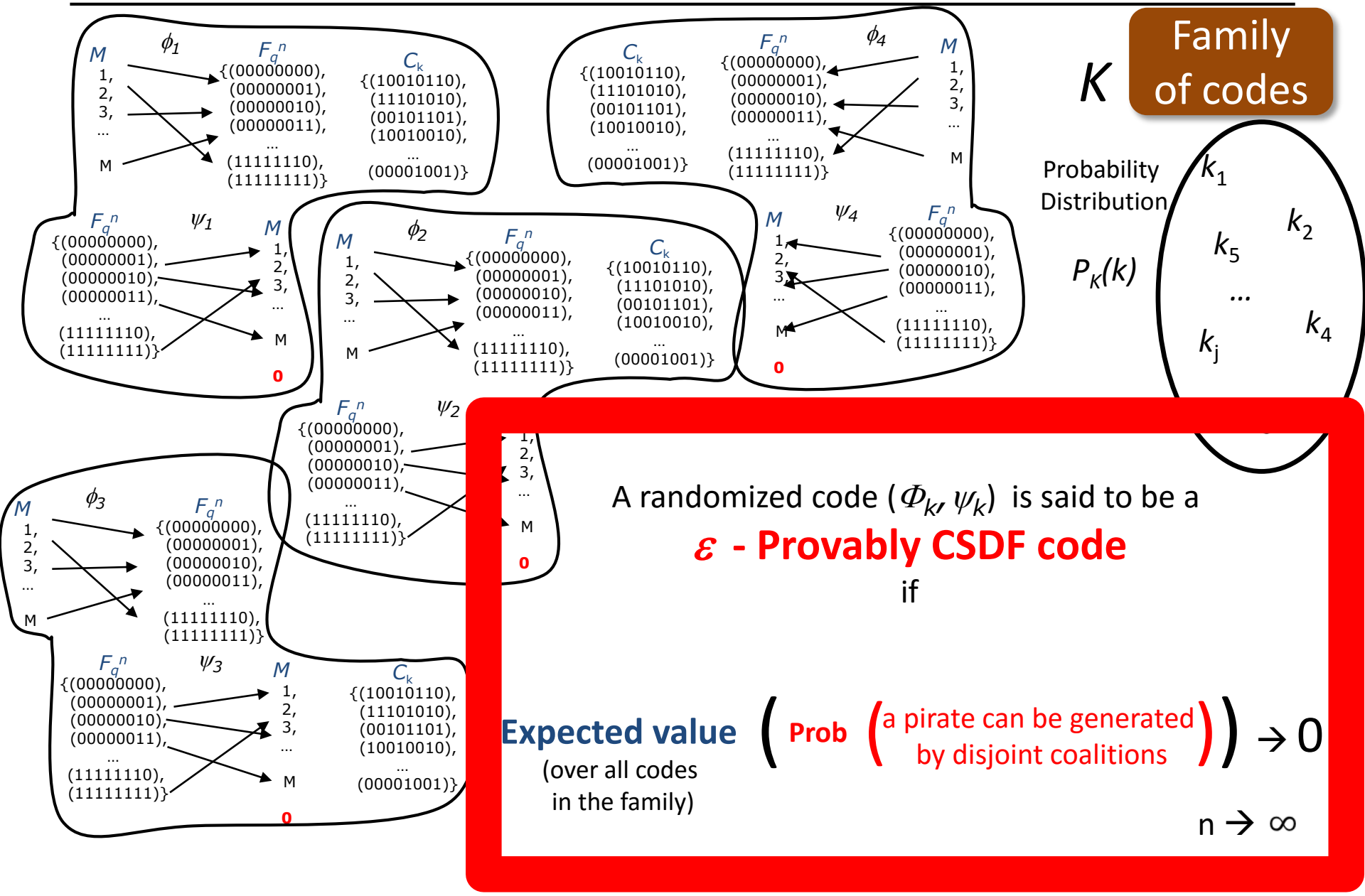
Almost Identifiable Parent Property Codes (Almost IPP)



Probability (a pirate can be generated by disjoint coalitions) $\to 0$

Almost *t*-IPP

Families of Almost *t*-IPP Codes



5

Almost t-IPP codes.
The Case of Two Pirates

The Case of Two Pirates

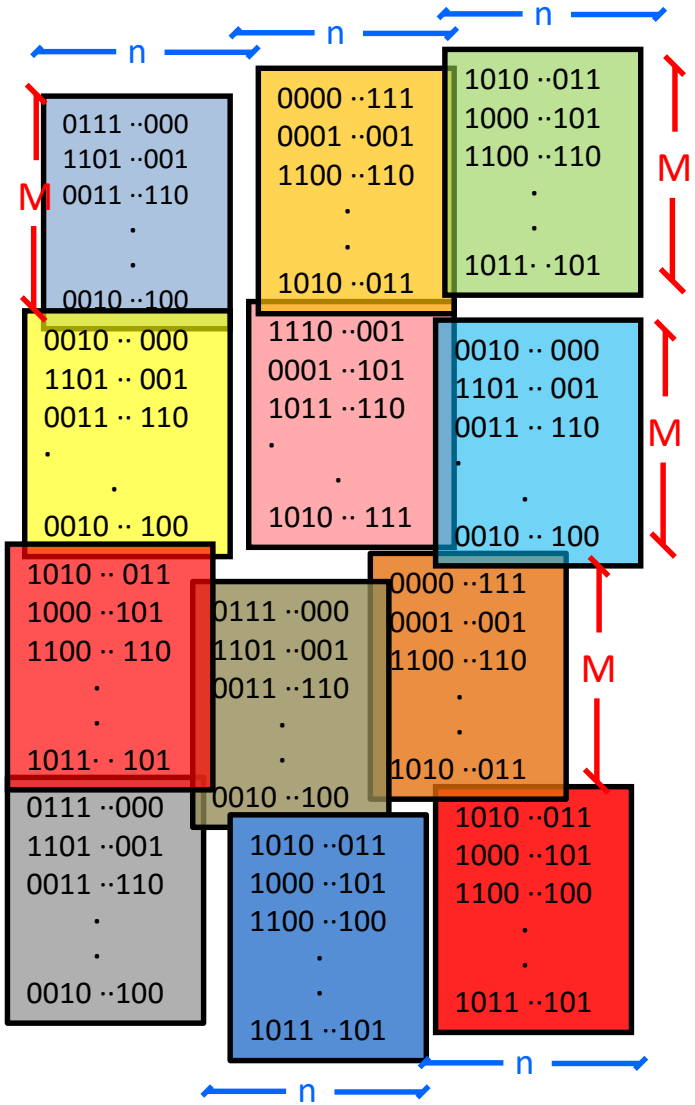
Result

There **exist** almost 2-IPP
or
ε-provably collusion secure digital  ing codes

of rate

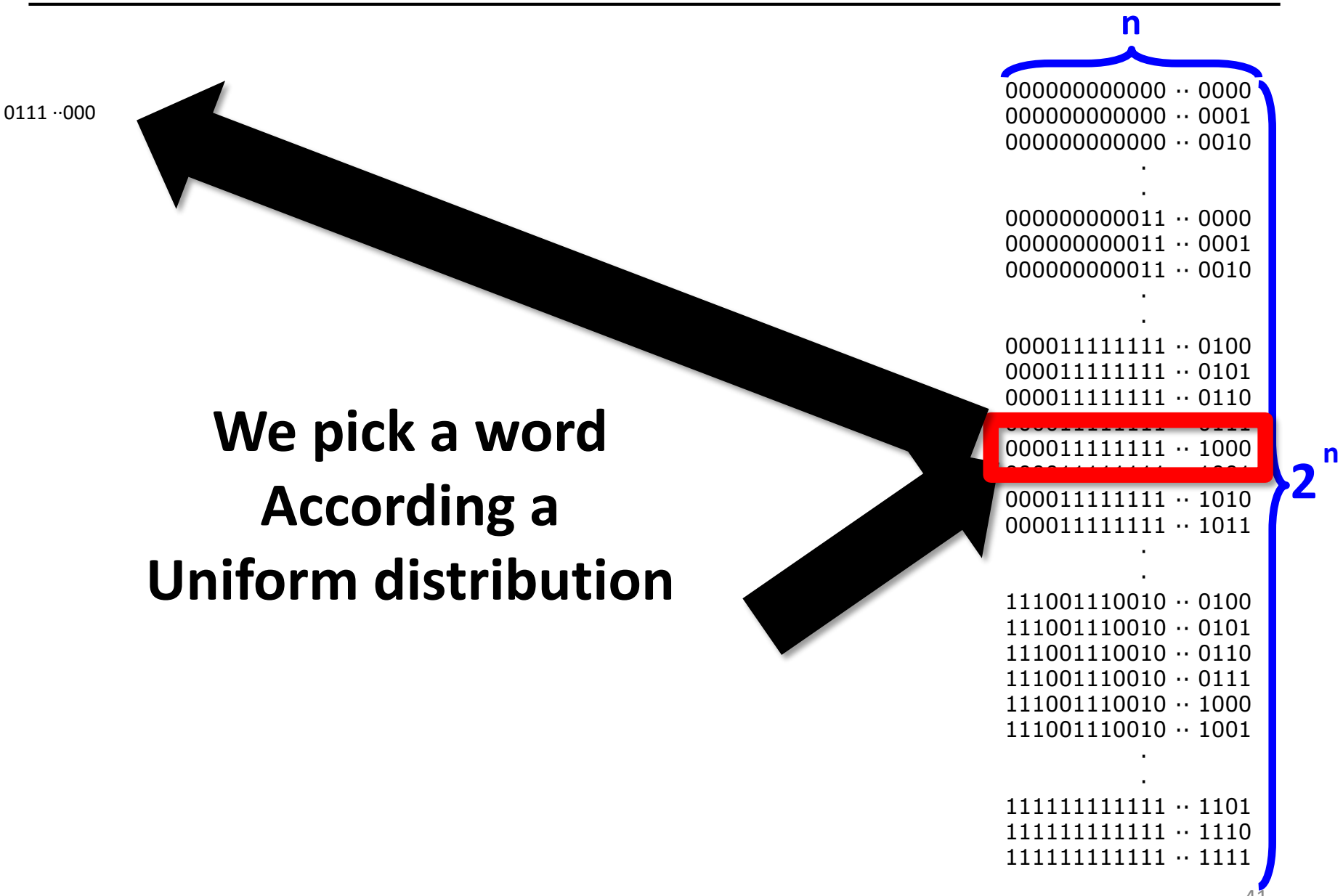
$$R < 0.2075 \ldots$$

The Case of Two Pirates

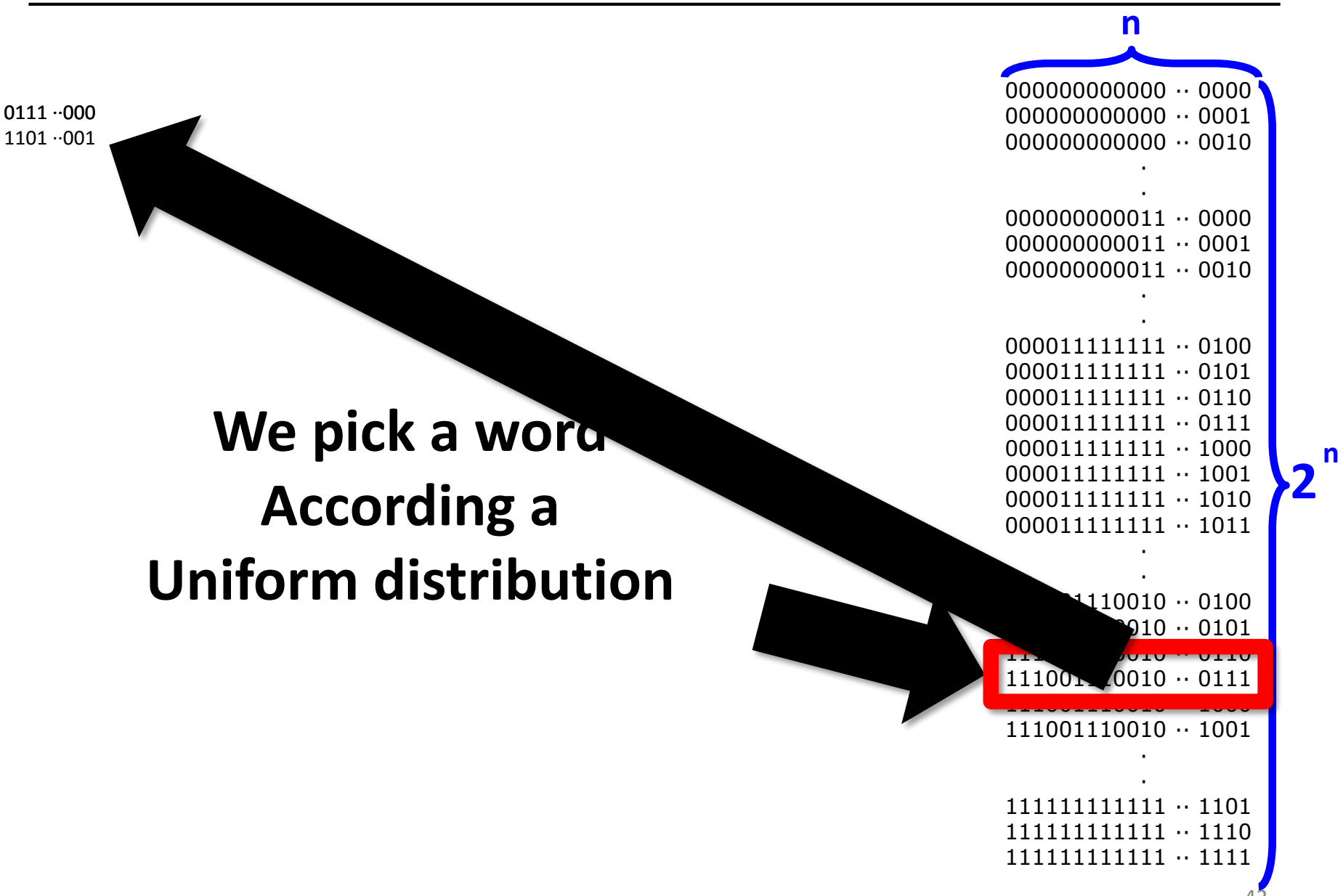


We will construct a code ensemble

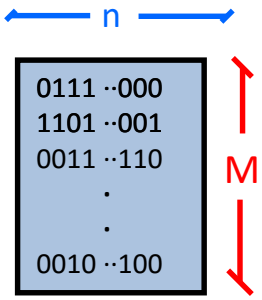
The Case of Two Pirates



The Case of Two Pirates

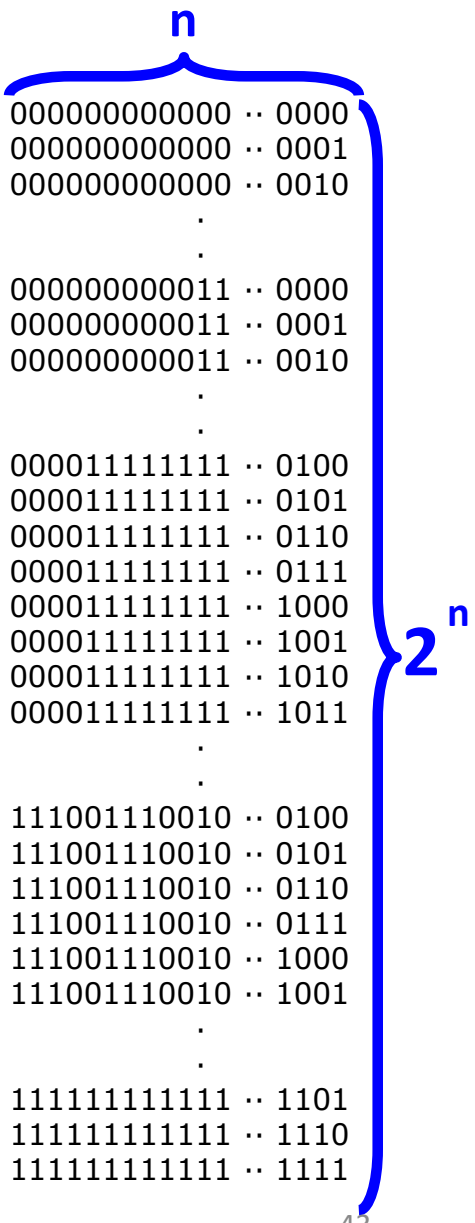


The Case of Two Pirates

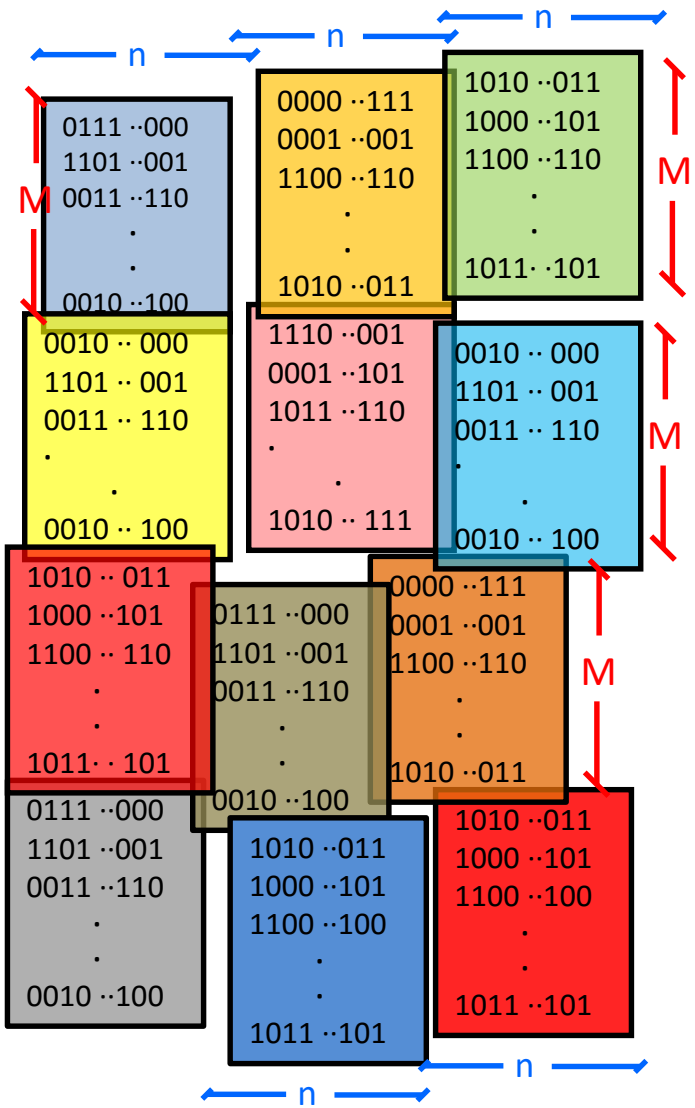


We repeat the procedure
Until we have picked

M codewords



The Case of Two Pirates



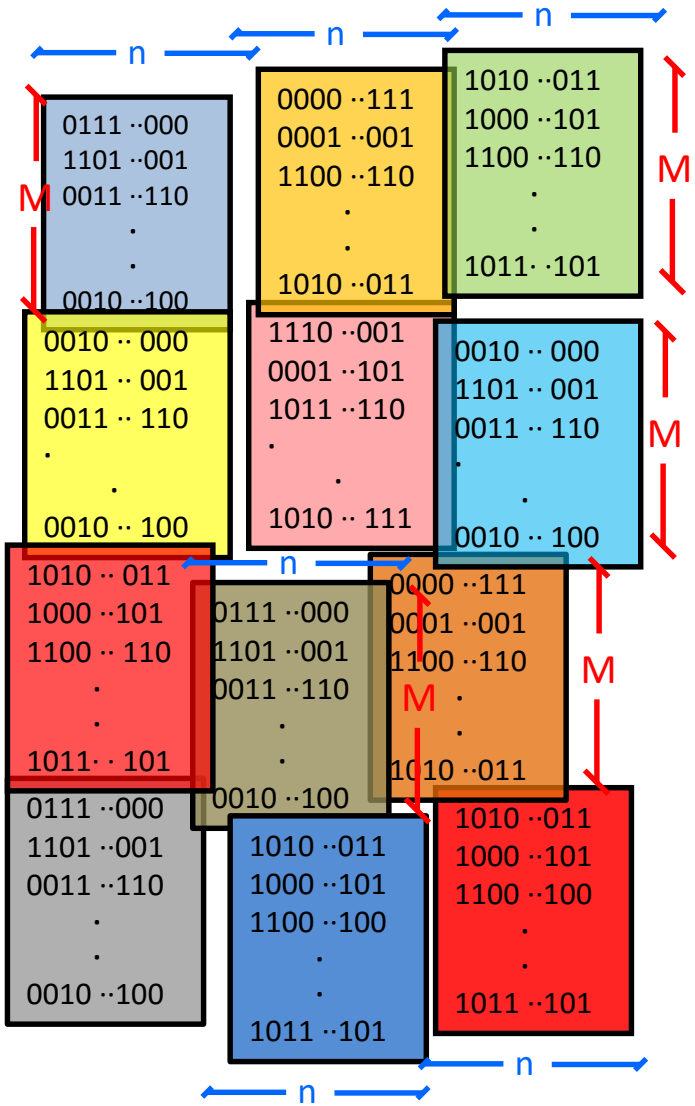
Our ensemble of codes
Is
all 2^{Mn}
binary matrices
with uniform distribution
on them

n

000000000000 ·· 0000
000000000000 ·· 0001
000000000000 ·· 0010
000000000011 ·· 0000
000000000011 ·· 0001
000000000011 ·· 0010
000011111111 ·· 0100
000011111111 ·· 0101
000011111111 ·· 0110
000011111111 ·· 0111
000011111111 ·· 1000
000011111111 ·· 1001
000011111111 ·· 1010
000011111111 ·· 1011
111001110010 ·· 0100
111001110010 ·· 0101
111001110010 ·· 0110
111001110010 ·· 0111
111001110010 ·· 1000
111001110010 ·· 1001
111111111111 ·· 1101
111111111111 ·· 1110
111111111111 ·· 1111

2^n

The Case of Two Pirates



We want that :

Expected value (over all codes in the family) $\left(\text{Probability that a pirate can be generated by a disjoint coalition}\right) \to 0$

$n \to \infty$

For the case of two pirates it amounts to show that there :

1) is no a “bad pair”

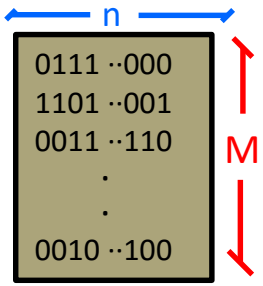
2) is no a “bad triangle”

The Case of Two Pirates

We want that :

Expected value (over all codes in the family) (Probability that a pirate can be generated by a disjoint coalition) $\rightarrow 0$

$n \rightarrow \infty$

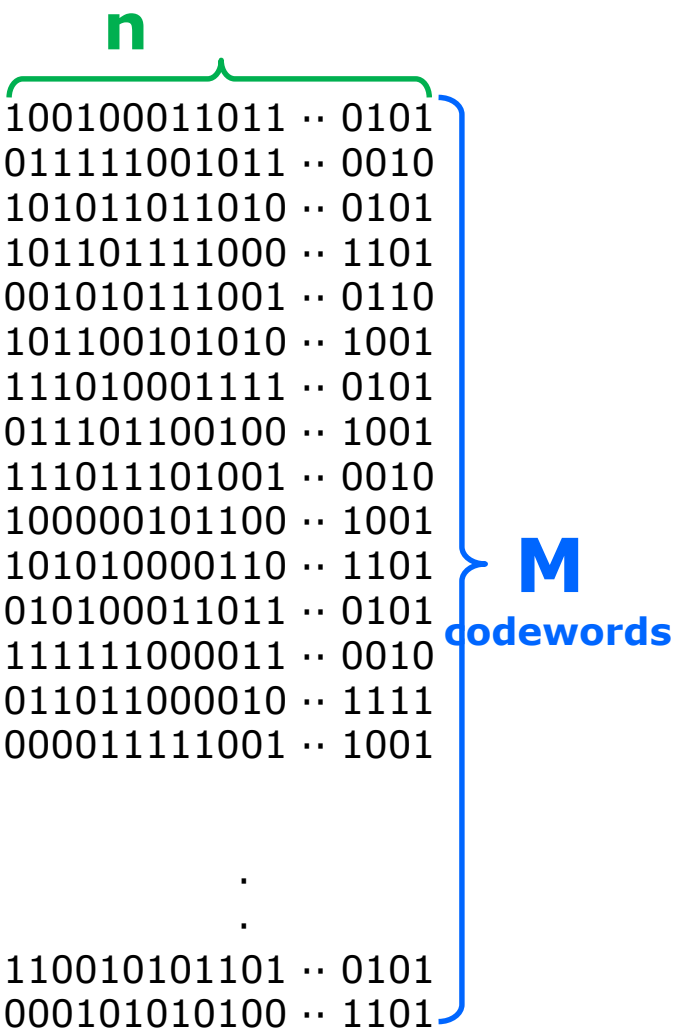


For the case of two pirates it amounts to show that there :

1) is no a “bad pair”

2) is no a “bad triangle”

The Case of Two Pirates



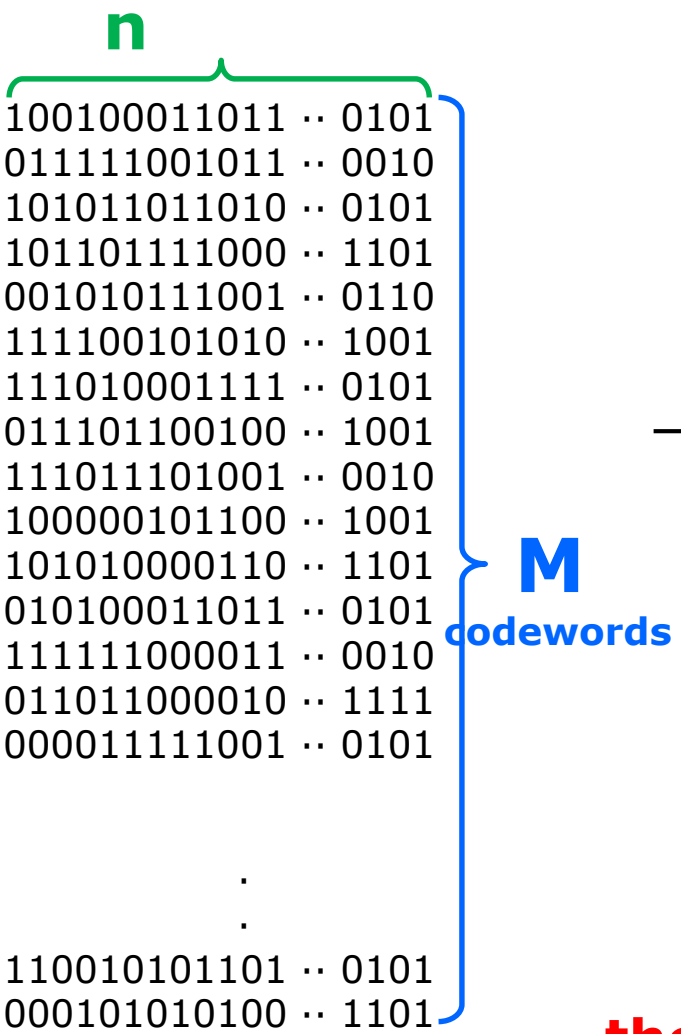
"Bad pair"

0100010 ·· 1111
0111001 ·· 1001

A pair is "bad" if there is another disjoint pair that can generate the same pirate

1101101 ·· 0101
0010100 ·· 1101

The Case of Two Pirates



"Bad pair"

011111001011 ·· 0010
011101100100 ·· 1001

011101101010 ·· 0001

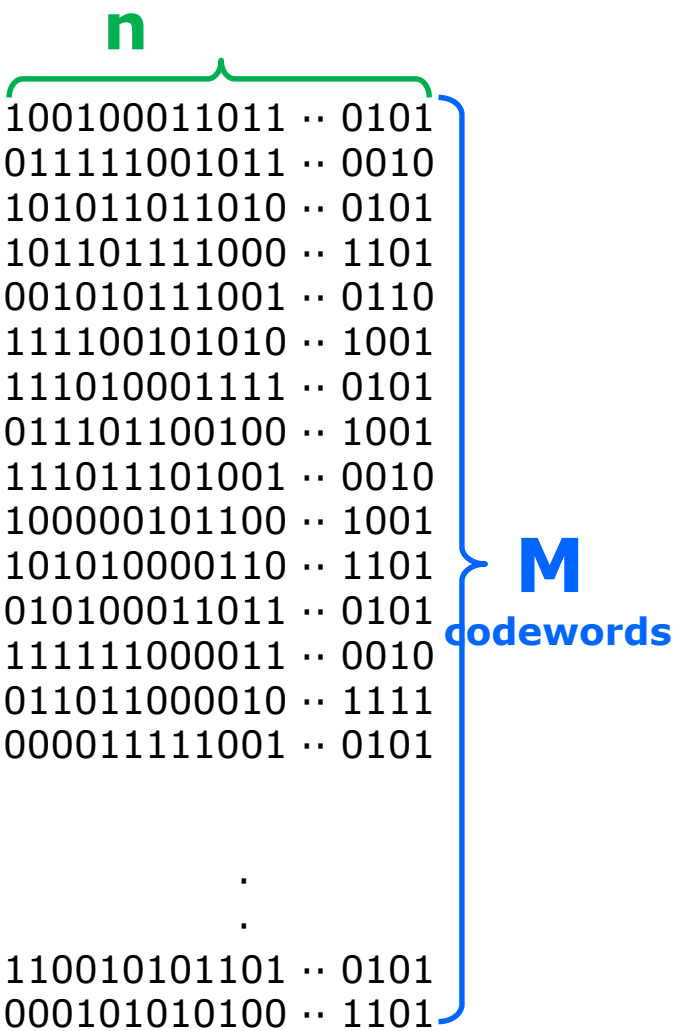
111100101010 ·· 1001
000011111001 ·· 0101

011101101010 ·· 0001

Same Pirate!!

These two pairs of pairs are "bad" because they both can generate the same pirate

The Case of Two Pirates



"Bad pair"

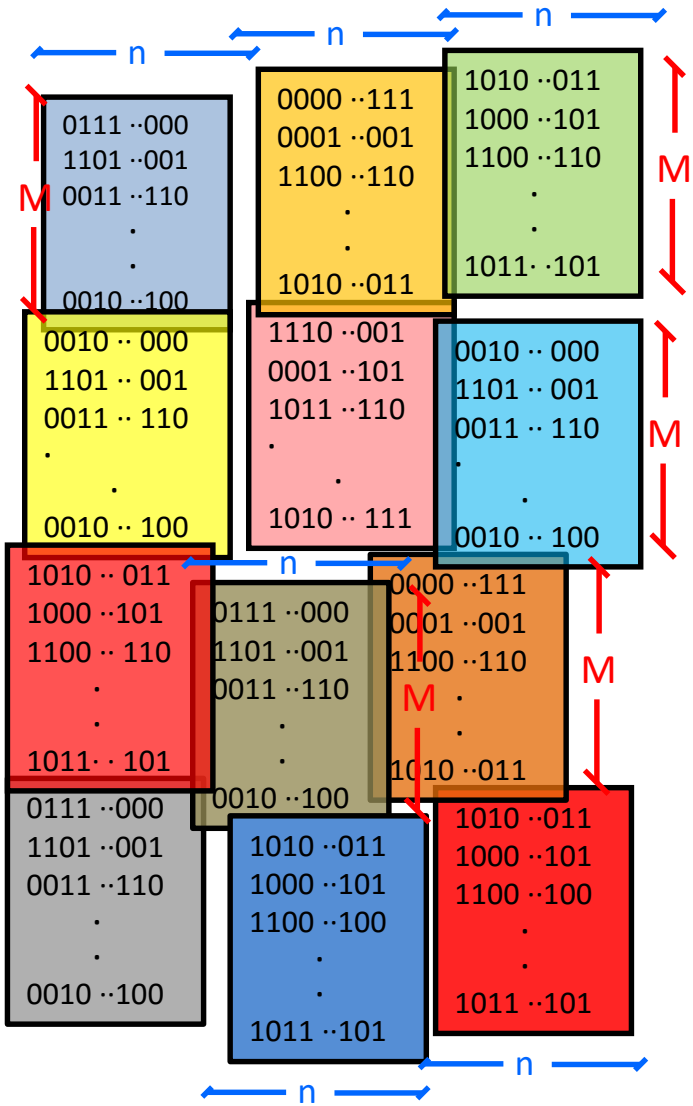
Probability there are no "bad" pairs

$$P_{\{2,2\}} \leq \binom{M-2}{2} \sum_{d=0}^{n} 2^{-n} \binom{n}{d} 2^{d-n} < M^2 \left(\frac{3}{4}\right)^n$$

$$P_{\{2,2\}} \to 0 \quad \text{for} \quad M^2 \leq n^{-1} \left(\frac{4}{3}\right)^n$$

for $R < 0.2075 \ldots$

The Case of Two Pirates



Expected value
(over all codes in the family)

$$\left(\text{Probability that a pirate can be generated by a disjoint coalition}\right) \to 0$$

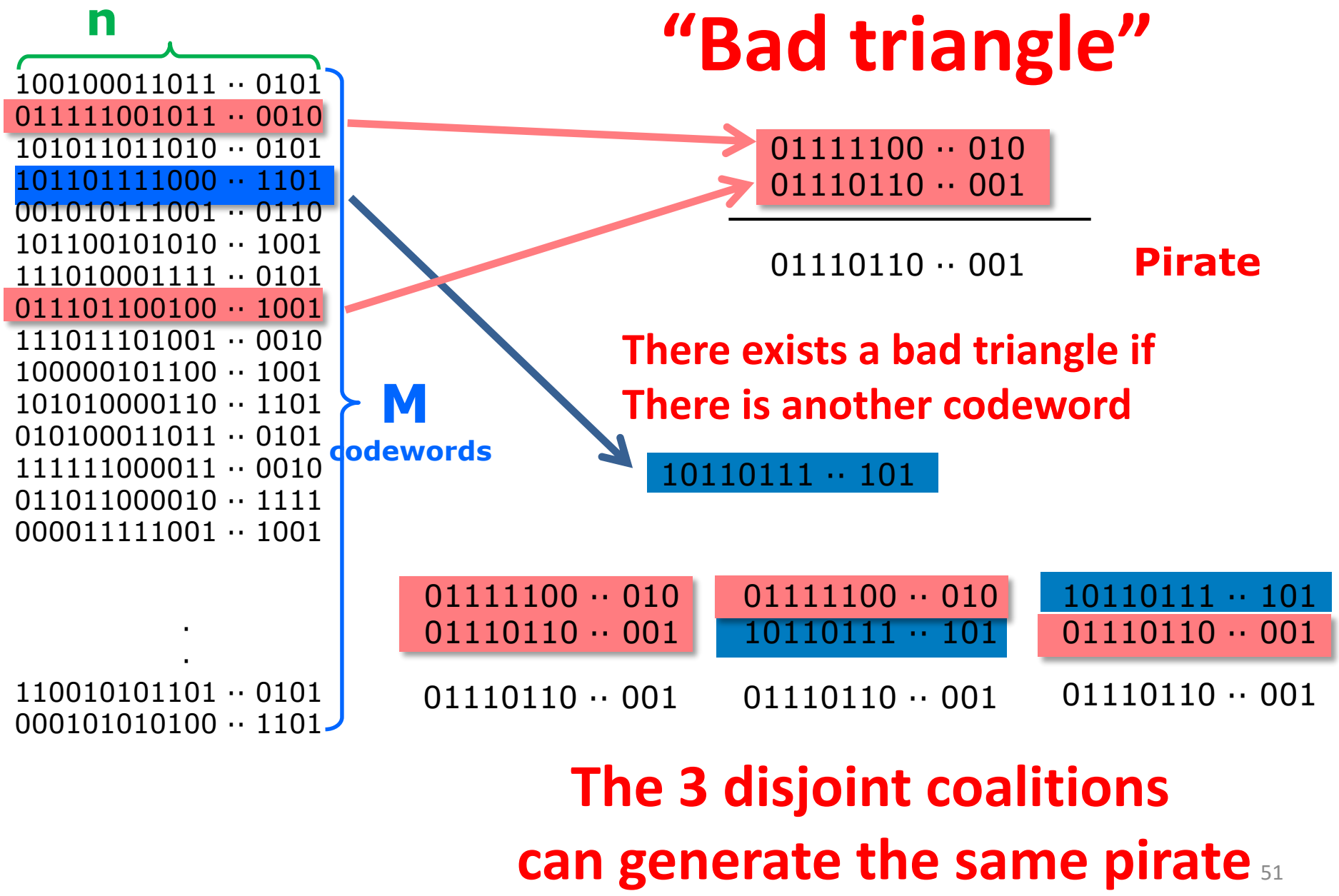
$$n \to \infty$$

For the case of two pirates it amounts to show that there :

1) is no a "bad pair"

2) is no a "bad triangle"

The Case of Two Pirates



The Case of Two Pirates

n

100100011011 ·· 0101
011111001011 ·· 0010
101011011010 ·· 0101
101101111000 ·· 1101
001010111001 ·· 0110
111100101010 ·· 1001
111010001111 ·· 0101
011101100100 ·· 1001
111011101001 ·· 0010
100000101100 ·· 1001
101010000110 ·· 1101
010100011011 ·· 0101
111111000011 ·· 0010
011011000010 ·· 1111
000011111001 ·· 0101

.
.

110010101101 ·· 0101
000101010100 ·· 1101

M codewords

"Bad triangle"

Probability there is no a "bad" triangle

$$P_{\{3\}} < M\left(\frac{3}{4}\right)^n$$

$$P_{\{3\}} \to 0 \qquad \text{for} \qquad M \le n^{-1}\left(\frac{4}{3}\right)^n$$

$$\boldsymbol{R < 0.415\ldots}$$

The Case of Two Pirates

Result

Probability there is not a "bad" pair

Probability there is not a "bad" triangle

$$R < 0.2075\ldots$$

Expected value (over all codes in the family)

$$\left(\text{Prob}\left(\text{a pirate can be generated by disjoint coalitions}\right)\right) \to 0 \quad (n \to \infty)$$

There exist almost 2-IPP or ε-provably collusion secure digital  ing codes of rate $R < 0.2075\ldots$

What is wrong with R=0.25?

The ensemble of random codes according to *N. P. Anthapadmanabhan, A. Barg, and I. Dumer, "On the fingerprinting capacity under the marking assumption", (IEEE-IT, vol.54, no.6, 2008)* has the *fingerprinting property for all rates* $\boldsymbol{R < 1/4}$.

The proof relies on the fact that the attack event is examined only for the high-probability subset of pairs of fingerprint vectors at Hamming distance about $n/2$.

On the other hand, the main contribution to the probability of having *a bad pair* comes from the pairs of vectors at distance about $2n/3$, i.e., from highly atypical pairs.

⟶ A typical pair of users can refute the accusation against them with high probability, showing another pair of users capable of generating the same fingerprint **x**.

6

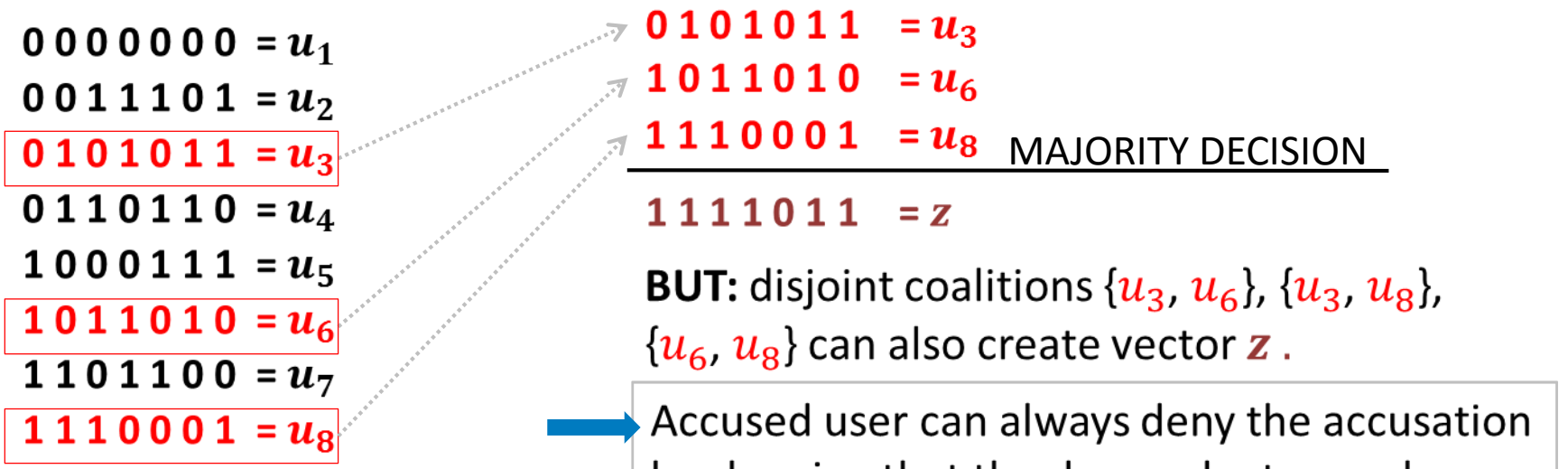
More than 2 pirates: negative result for almost t-IPP codes

More than 2 pirates: negative result

RESULT: there does **not** exist any family of almost t-IPP codes for $t > 2$.

Some intuition:

Consider the following code and collusion attack:



BUT: disjoint coalitions $\{u_3, u_6\}$, $\{u_3, u_8\}$, $\{u_6, u_8\}$ can also create vector z .

→ Accused user can always deny the accusation by showing that the descendant z can be created by the other two users alone.

→ Generalization for the case of arbitrary $t > 2$ can be created in the same way.

7

Relaxed version of almost t-IPP codes

Relaxed version of tracing: definitions

Due to the previous result we introduce a relaxed version of tracing – *net of suspicious users*.

Let $< \varphi(U) >$ denotes the set of all forged vectors that coalition U can create.

Definition 1: For a given code C and a given forged fingerprint z a set L of users is called a ***net of z*** if $L \cap U \neq \emptyset$ for all U such that $z \in < \varphi(U) >$ and $|U| \leq t$.

→ The dealer wish to have minimal possible size of nets.

Definition 2: We call a net of z an $\boldsymbol{\varepsilon}$ $-$ ***net of z*** if the probability that $L \cap U \neq \emptyset$ for all U such that $z \in < \varphi(U) >$ and $|U| \leq t$ is at least $1 - \boldsymbol{\varepsilon}$.

Relaxed version of tracing: results

RESULTS: lower and upper bound on the size of the net.

LOWER BOUND:

LEMMA: For any $0 < \varepsilon < 1$ and any fingerprinting family of codes C_k any ε-net of suspicious users has **size at least** $\lceil t/2 \rceil$.

UPPER BOUND:

THEOREM: For the family of concatenated *t*-CSFC proposed by *A. Barg, G. R. Blakley, and G. Kabatiansky in the paper "Digital fingerprinting codes: Problem statements, constructions, identification of traitors"* **the size of the net of suspicious users is at most t with probability tending to 1** and **almost all users from the net are guilty.**

8

Summary and Conclusion

Summary and Conclusion

We proposed a new paradigm for fingerprinting binary codes: $\varepsilon-$ *provably Collusion Secure Digital Fingerprinting codes*.

A family of fingerprinting codes following this paradigm has the following property:

For any unregistered fingerprint $\boldsymbol{x}$ the tracing algorithm identifies a net of user, say $\widehat{U}$, s.t. the probability of the event that there is a coalition V which is capable of generating the same $\boldsymbol{x}$ but $\widehat{U} \cap V = \emptyset$, can be made arbitrarily small by increasing the code length.