

Watermarking Relational Databases

Presented by:
Mohamed Shehab
mshehab@uncc.edu

1

Research Interests

- Access Control for Distributed Environments.
- Database Security and Privacy
- Watermarking and Other Fun Stuff

2

Courses (Fall 2007)

- Currently Teaching IT IS 6210/8210, Access Control and Security Architecture.
 - Access Control Models
 - Delegation Management
 - Trust Management
 - Identity Management
 - Web Services Security

3

Talk Outline

- Introductory Material
- General Watermarking Model & Attacks
- WM Technique 1 (Agrawal et al.)
- WM Technique 2 (Sion et al.)
- Future Challenges and References

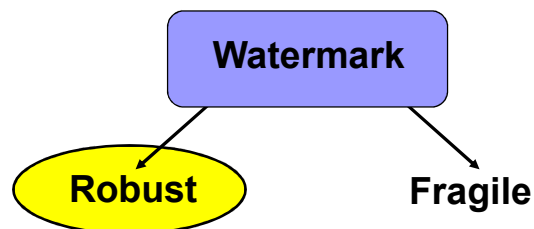
4

What is Watermarking ?

- A “*watermark*” is a signal that is securely, imperceptibly, and “*robustly*” embedded into original content such as an image, video, or audio signal, producing a watermarked signal.
- The watermark describes information that can be used for proof of ownership or tamper proofing.

5

What is Watermarking ? (Cont.)



- Robust Watermark: for proof of ownership, copyrights protection.
- Fragile Watermark: for tamper proofing, data integrity.

6

Why Watermarking ?

- Digital Media (Video, Audio, Images, Text) are easily copied and easily distributed via the web.
- Database outsourcing is a common practice:
 - Stock market data
 - Consumer Behavior data (Walmart)
 - Power Consumption data
 - Weather data
- Effective means for proof of authorship.
 - Signature and data are the same object.
- Effective means of tamper proofing.
 - Integrity information is embedded in the data.

7

Why is Watermarking Possible ?

- Real-world datasets can tolerate a small amount of error without degrading their usability
 - Meteorological data used in building weather prediction models, the wind vector and temperature accuracies in this data are estimated to be within 1.8 m/s and 0.5 °C.
 - Such constraints bound the amount of change or alteration to that can be performed on the data.

8

What defines the usability constraints ?

- Usability constraints are application dependent.
 - Alterations performed by the watermark embedding should be unidentifiable by the human visual system in images/video.
 - For consumer behavior data: watermarking should preserve periodicity properties of the data.

9

What defines the usability constraints ? (Cont.)



Courtesy of <http://maps.google.com>

10

Watermark Desirable Properties

- Detectability (Key-Based System)
 - Can be easily detected only with the knowledge of the secret key.
- Robustness
 - Watermark cannot be easily destroyed by modifying the watermarked data.
- Imperceptibility
 - Presence of the watermark is unnoticeable.
- Blind System
 - Watermark detection does not require the knowledge of the original data.

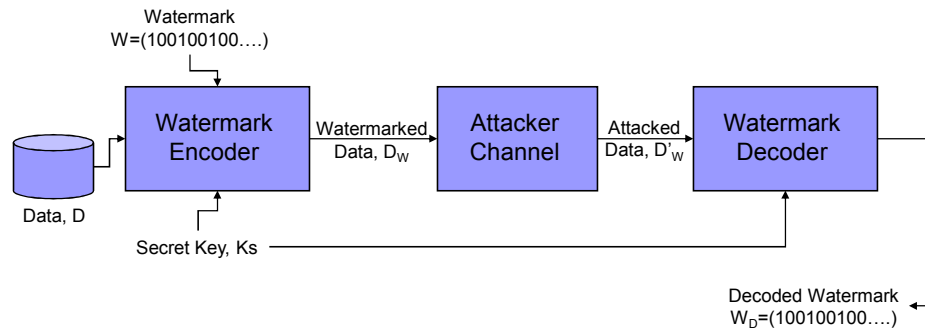
11

Talk Outline

- **Introductory Material**
- General Watermarking Model & Attacks
- WM Technique 1 (Agrawal et al.)
- WM Technique 2 (Sion et al.)
- Future Challenges and References

12

Watermarking Model



13

Relational and multimedia data

- A multimedia object consists of a large number of bits, with considerable redundancy. Thus, the large watermark hiding bandwidth.
- The relative spatial/temporal positioning of various pieces of a multimedia object typically does not change. Tuples of a relation on the other hand constitute a set and there is no implied ordering between them.
- Portions of a multimedia object cannot be dropped or replaced arbitrarily without causing perceptual changes in the object. However, a pirate of a relation can simply drop some tuples or substitute them with tuples from other relations.

14

Attacker Model

- Attacker has access to only the watermarked data set.
- The attacker's goal is to weaken or even erase the embedded watermark and at the same time keep the data usable.
"Attacker's Dilemma"
- Possible Attacks
 - Tuple deletion
 - Tuple alteration
 - Tuple insertion

15

Talk Outline

- **Introductory Material**
- **General Watermarking Model & Attacks**
- WM Technique 1 (Agrawal et al.)
- WM Technique 2 (Sion et al.)
- Future Challenges and References

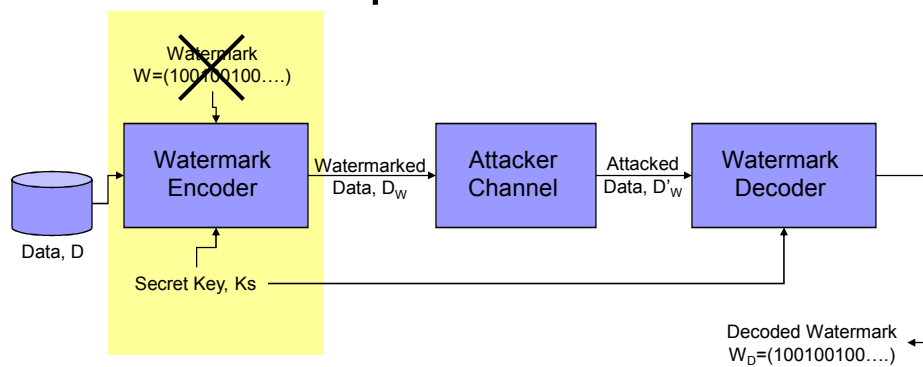
16

WM Technique 1 (Agrawal et. al.)

- Watermarking of numerical data.
- Technique dependent on a secret key.
- Uses markers to locate tuples to hide watermark bits.
- Hides watermark bits in the **least significant bits**.

17

WM Technique 1: Encoder



Instead:
Watermark is a function of the data and the secret key

18

WM Technique 1: Encoder

■ Assumptions

- K , e , m and v are randomly selected by the data owner and are kept secret.
- “ K ” is the secret key.
- “ e ” least significant bits can be altered in a number without affecting its usability. Example, $e=3$, 101101101.1011101
- “ m ” used for marker selection
- “ v ” is the number of attributes used in the watermarking process.

19

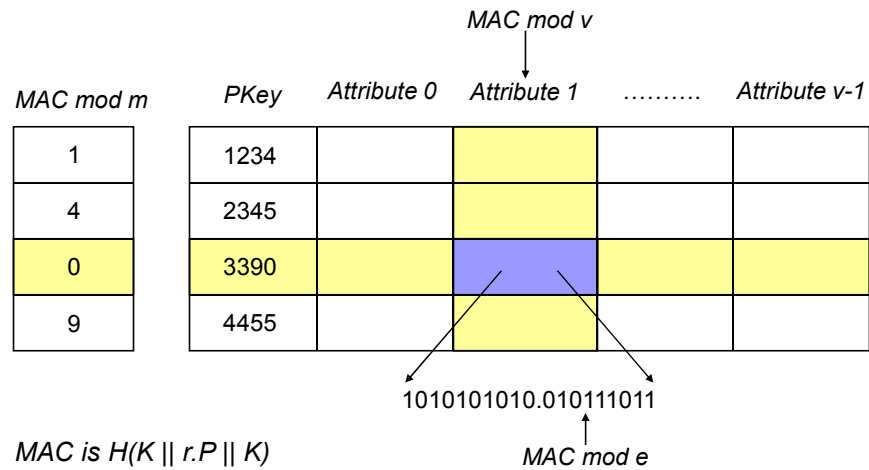
WM Technique 1: Encoder

■ For all tuples r in D

- $r.MAC = H(K||r.P||K)$
- $if(r.MAC \bmod m == 0)$ // Marker Selection
 - $i = r.MAC \bmod v$ // Selected Attribute
 - $b = r.MAC \bmod e$ // Selected LSB index
 - $if(r.MAC \bmod 2 == 0)$ // MAC is even
 - Set bit b of $r.A_i$
 - Else
 - Clear bit b of $r.A_i$

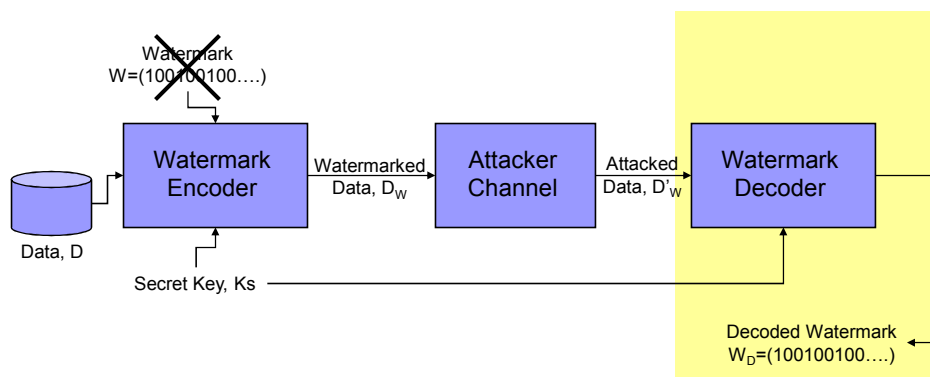
20

WM Technique 1 : Encoder



21

WM Technique 1 : Decoder



22

WM Technique 1 : Decoder

- $Match = Total_Count = 0$
- For all tuples r in D
 - $r.MAC = H(K || r.P || K)$
 - $if(r.MAC \bmod m == 0)$ // Marker Selection
 - $Total_Count++$
 - $i = r.MAC \bmod v$ // Selected Attribute
 - $b = r.MAC \bmod e$ // Selected LSB index
 - $if(r.MAC \bmod 2 == 0)$ // MAC is even
 - if bit b of $r.A_i$ is Set
 - $Match++$
 - Else
 - If bit b of $r.A_i$ is Clear
 - $Match++$
- Compare $(Match/Total_count) > Threshold$

23

WM Technique 1 : Decoder

$MAC \bmod v$
 ↓
 Attribute 1

MAC mod m	PKey	Attribute 0	Attribute 1	Attribute v-1
1	1234				
4	2345				
0	3390				
9	4455				

1010101010.010111011
 ↑
 $MAC \bmod e$

MAC is $H(K || r.P || K)$

24

WM Technique 1 : Strengths

- Computationally efficient $O(n)$
 - Tuple sorting not required.

- Incremental Updatability

25

WM Technique 1 : Weaknesses

- No provision of multi-bit watermark, all operations are dependent only on the secret key.
- Not resilient to alteration attacks. LSB can be easily manipulated by simple numerical alterations
 - Shift LSB bits to the right/left.
- Requires the presence of a primary key in the watermarked relation.
- Does not handle other usability constraints such as:
 - Category preserving usability constraints.

26

Talk Outline

- Introductory Material
- General Watermarking Model & Attacks
- WM Technique 1 (Agrawal et al.)
- WM Technique 2 (Sion et al.)
- Future Challenges and References

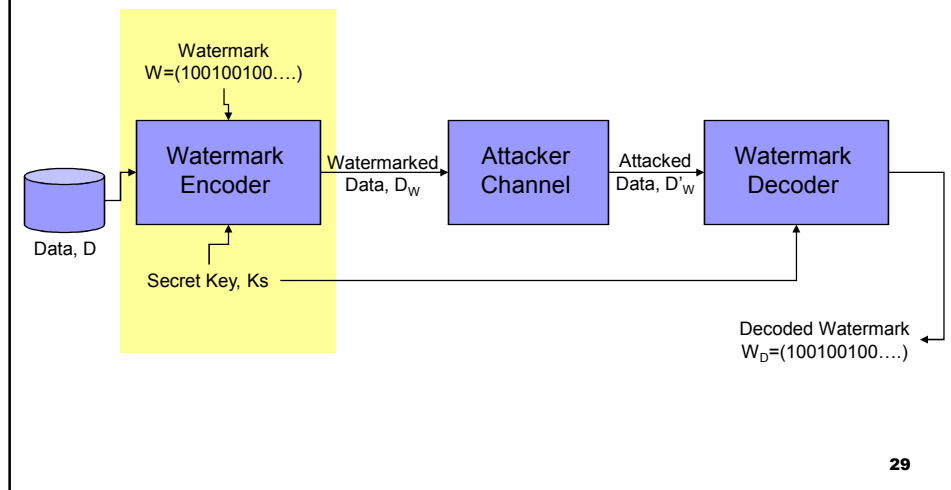
27

WM Technique 2 :(Sion et. al.)

- Watermarking of numerical data.
- Technique dependent on a secret key.
- Instead of primary key uses the most significant bits of the *normalized* data set.
- Divides the data set into partitions using markers.
- Varies the partition statistics to hide watermark bits.

28

WM Technique 2 : Encoder

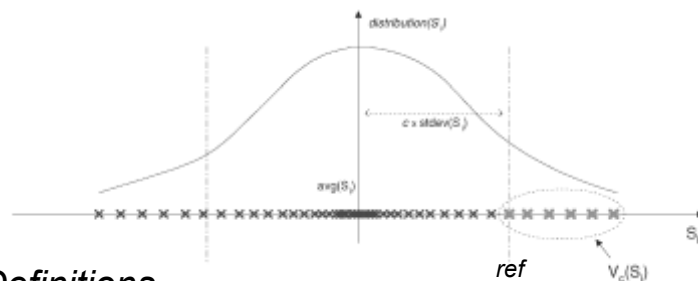


WM Technique 2: How to hide a single bit in a number set ?

■ Problem:

“ Given a number set $S_i = \{s_1, \dots, s_n\}$, how to vary their statistics to embed bit b_i . Subject to the provided usability constraints.”

Paper 2: How to hide a single bit in a number set ?

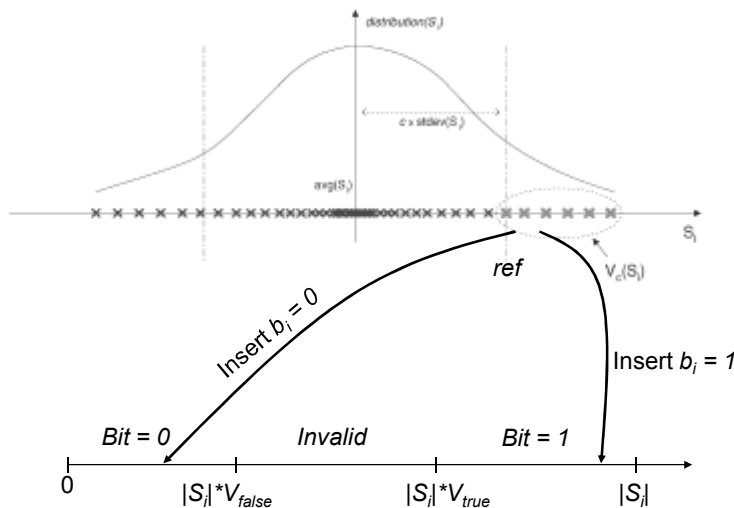


■ Definitions

- $\mu = \text{mean}(S_i)$
- $\sigma = \text{stdev}(S_i)$.
- $\text{ref} = \mu + c\sigma$
- $V_c(S_i) = \text{number of points greater than ref}$. We refer to them as "**positive violators**".

31

Paper 2: How to hide a single bit in a number set ?



32

WM Technique 2: How to avoid using the primary key ?

- Given a number set $S_i = \{s_1, \dots, s_n\}$, generate $Norm(S_i) = S_i / \max(S_i)$.
- For each number in s_k in $Norm(S_i)$ use the first n most significant bits as the primary key for s_k .

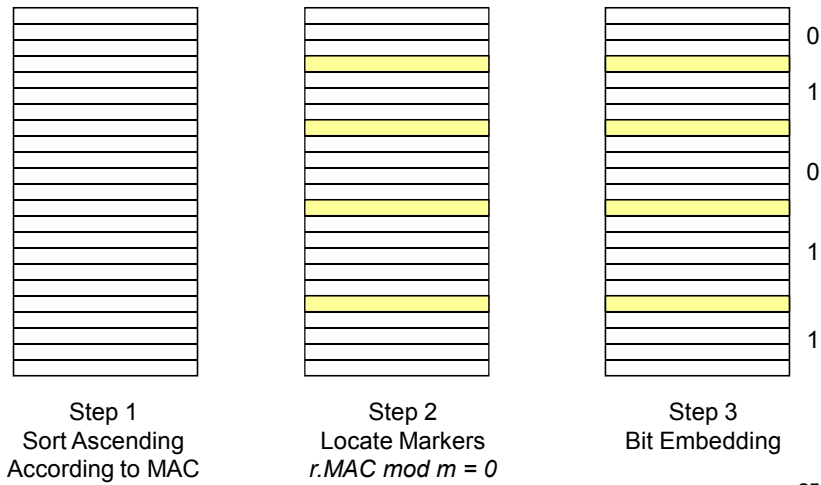
33

WM Technique 2 : Encoder

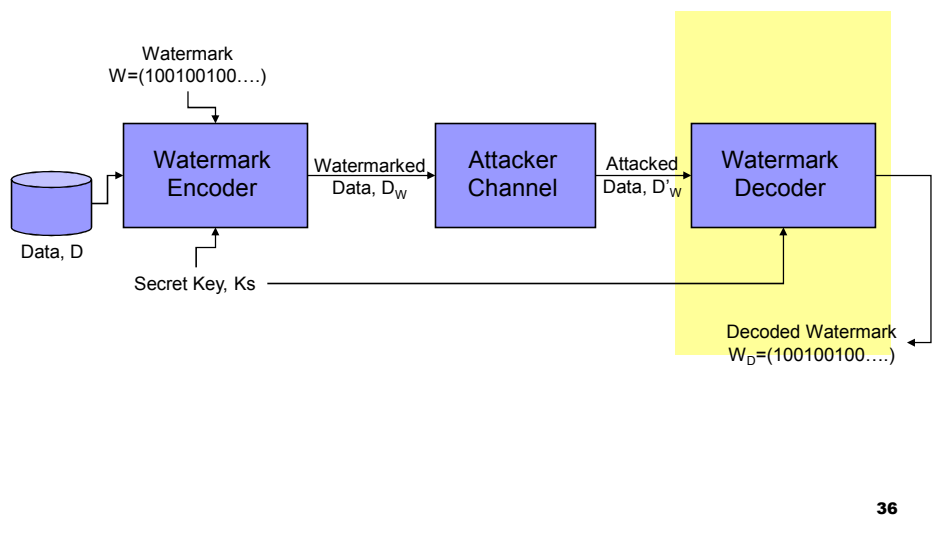
- Step 1: (Sorting)
 - Compute the MAC of each tuple:
 - $r.MAC = H(K || r.P || K)$ // $r.P = MSB(r.A)$
 - Sort tuples in ascending order using the computed MAC.
- Step 2: (Partitioning)
 - Locate markers: tuples with $r.MAC \bmod m = 0$
 - Tuples between two markers are in the same partition.
- Step 3: (Bit Embedding):
 - Embed a watermark bit in each partition using the bit embedding technique discussed earlier.

34

WM Technique 2 : Encoder



WM Technique 2 : Decoder

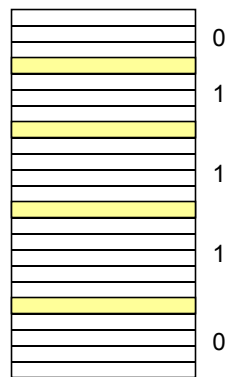


WM Technique 2 : Decoder

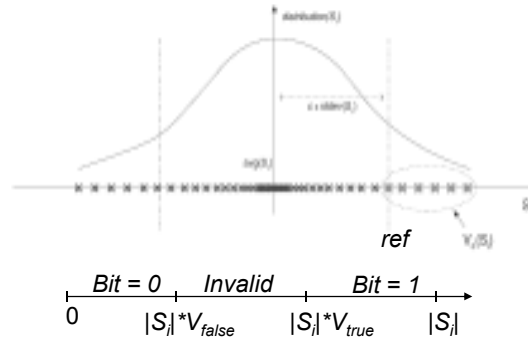
- Step 1: (Sorting & Partitioning)
 - Partition data set using the same approach used in the encoding phase.
- Step 2: (Bit Detection)
 - For each partition S_i compute $V_c(S_i)$ and decode the embedded bit.
- Step 3: (Majority Voting):
 - Watermark bits are embedded in several partitions use majority voting to correct for errors.

37

WM Technique 2 : Decoder



Watermarked Data Set



3	1	2	15	04	3	1	2	15	04
1	w_0	11	00	1	w_0	11	00		
1	w_1	11	00	1	w_1	11	00		
0	w_2	11	10	0	w_2	11	10		
w_{total}		11	00	w_{total}		11	00		

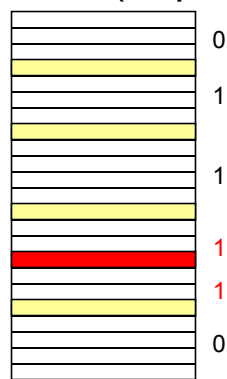
38

WM Technique 2 : Strengths

- Bit embedding technique honors usability constraints.
- Embeds watermark in data statistics which makes technique more resilient to alteration attacks when compared to LSB technique.

39

WM Technique 2 : Watermark Synchronization Error (Tuple Addition)



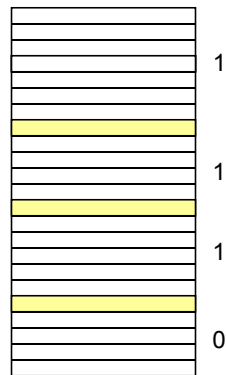
Watermarked Data Set

	5	4	3	2	1	0
W_0	1	0	1	1	1	0
W_1	1	0	1	0	1	0
W_2	1	0	0	0	1	1
W_{result}	1	0	1	0	1	0

	5	4	3	2	1	0
W_0	0	1	1	1	1	0
W_1	0	1	0	1	0	1
W_2	0	0	0	1	1	1
W_{result}	0	1	0	1	1	1

40

WM Technique 2 : Watermark Synchronization Error (Tuple Deletion)



Watermarked
Data Set

	5	4	3	2	1	0
W_0	1	0	1	1	1	0
W_1	1	0	1	0	1	0
W_2	1	0	0	0	1	1
W_{result}	1	0	1	0	1	0

W_0	0	1	0	1	1	1
W_1	1	1	0	1	0	1
W_2	x	1	0	0	0	1
W_{result}	x	1	0	1	0	1

41

Paper 2: Weaknesses

- Watermark suffers badly from watermark synchronization error cause by
 - Tuple deletion attacks.
 - Tuple addition attacks.
- No optimality criteria when choosing the decoding thresholds
 - Errors even in absence of attacker.
- No clear systematic approach for manipulating data
 - Only a very small space of the feasible data manipulations investigated.

42

Talk Outline

- **Introductory Material**
- **General Watermarking Model & Attacks**
- **WM Technique 1 (Agrawal et al.)**
- **WM Technique 2 (Sion et al.)**
- **Future Challenges and References**

43

Challenges

- Investigate watermarking other types of data. Such as data streams.
- Design robust watermarking techniques that are resilient to watermark synchronization errors.
- Design a fragile watermarking technique for relational databases.

44

References

- J. Kiernan, R. Agrawal, "Watermarking Relational Databases," *Proc. 28th Int'l Conf. Very Large Databases VLDB*, 2002.
- R. Sion, M. Atallah, S. Prabhakar, "Rights Protection for Relational Data," *IEEE Transactions on Knowledge and Data Engineering*, Volume 16, Number 6, June 2004
- M. Shehab, E. Bertino, A. Ghafoor, "Watermarking Relational Databases using Optimization Based Techniques," *IEEE Transactions of Knowledge and Data Engineering (TKDE)*, (PrePrint).

45

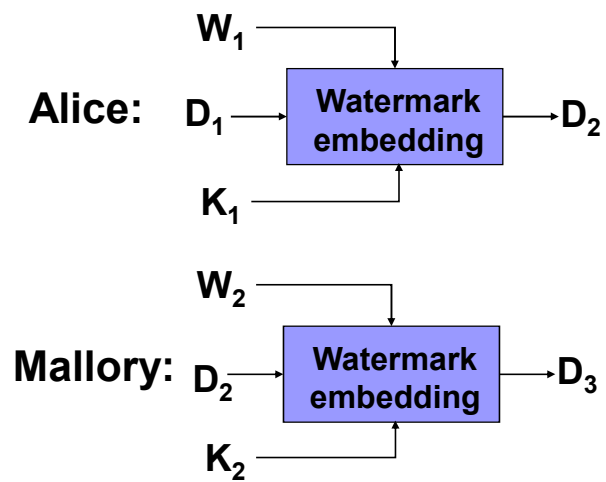
Questions?



mshehab@uncc.edu

46

Problems



47