Workshop Week 12

COMP20008 2018

Blockchain

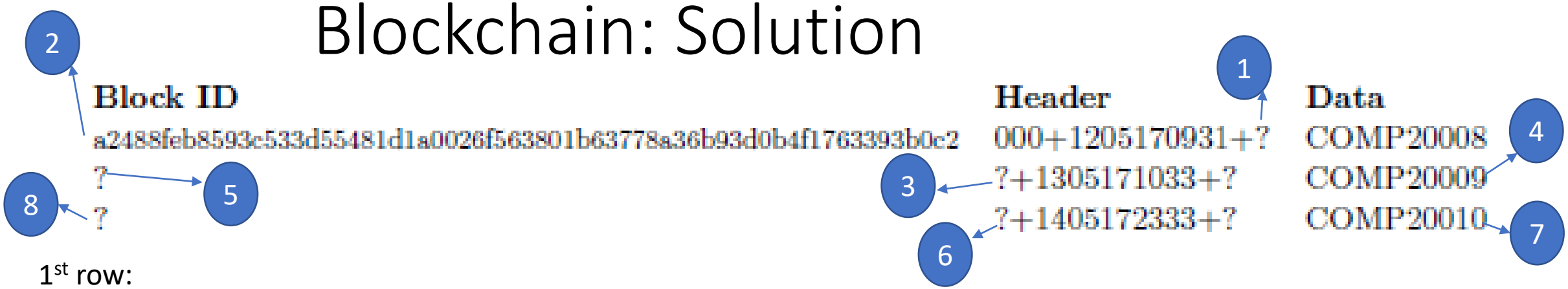
1. Consider a simple blockchain with three blocks. Each block consists of header and some data. The ID of a block is equal to a hash of its header. The header is a string consisting of the following items in sequence, each separated by a '+' character.
   - ID of the parent block
   - Timestamp (DDMMYYHHMM format)
   - Hash of the data stored in the block

The following table gives a partial description of the first three blocks in this blockchain. Row 1 is the first block, row 2 is the second block, row 3 is the third block. Assume there exists a dummy block zero, (not included) with ID 000.

Block ID	Header	Data
a2488feb8593c533d55481d1a0026f563801b63778a36b93d0b4f1763393b0c2	000+1205170931+?	COMP20008
?	?+1305171033+?	COMP20009
?	?+1405172333+?	COMP20010

Each ' ?' in the table represents a 64 digit hexadecimal value. Replace each ' ?' by its appropriate value, to maintain the integrity of the blockchain. Use the SHA-256 function for hashing, available at this website.

Blockchain: Solution



1st row:

1. H(COMP20008) = 23a602232c74b3e00a31ed1eddda091669f77acf24f2db04591e74259047e6ba
2. Show that Block ID = H(000+1205170931+23a602232c74b3e00a31ed1eddda091669f77acf24f2db04591e74259047e6ba)= a2488feb8593c533d55481d1a0026f563801b63778a36b93d0b4f1763393b0c2

2nd row:

3. Parent ID = a2488feb8593c533d55481d1a0026f563801b63778a36b93d0b4f1763393b0c2
4. H(COMP20009) = 611bf18429335a9a544f5734b0c8b3081a1c304247d7b61226bad8777551456c
5. Block ID = H(a2488feb8593c533d55481d1a0026f563801b63778a36b93d0b4f1763393b0c2+ 1305171033+611bf18429335a9a544f5734b0c8b3081a1c304247d7b61226bad8777551456c) = e8e801a51cb7683d81b5ebb5275b4aa12b24a2bc39655bebf4543b3bc25bd2a2

3rd row:

6. Parent ID = e8e801a51cb7683d81b5ebb5275b4aa12b24a2bc39655bebf4543b3bc25bd2a2
7. H(COMP20010) = cd76ca0b4c2691071a40c5b1c4a21486757c3f32ac5b4c5c4ffbf6d1a8d28bc0
8. Block ID = H(e8e801a51cb7683d81b5ebb5275b4aa12b24a2bc39655bebf4543b3bc25bd2a2+1405172333+ cd76ca0b4c2691071a40c5b1c4a21486757c3f32ac5b4c5c4ffbf6d1a8d28bc0) = d6f384aa4bfc506d7a0c345b5c2535e9ac8cc09a25c89b6be617bc0a2b0641ef

Blockchain

2. Blockchain modifications:
   i. Suppose the data for the first block was modified to read COMP20008!, describe (at a high level) how would this would affect the rest of the blockchain.

Solution

- This modification will require a change to the header of first block, why?
  - Because header includes a hash of the data.
- A change to the header of first block, will require a change to its block id, why?
  - Because block ID is a hash of the header.
- A change to the block ID of first block will require a change to the header of second block, why?
  - Because header records its parent's id.
- A change to the header of second block will require a change to its block id (which is a hash of its header).
- A change to id of second block will require a change to the header of third block (which records its parents id).
- A change to the header of the third will require a change to its block id (which is a hash of its header) ……

The bottom line: a cascading series of changes to headers of all the blocks after the first block.

Blockchain

2. Blockchain modifications:
   ii. Suppose the timestamp for the second block was modified, describe (at a high level) how this would affect the rest of the blockchain.

Solution

- If timestamp for second block was modified, this would cause a change to its …
  - block id (which is a hash of its header).
- A change to id of second block will require a change to the …
  - header of third block (which records its parents id).
- A change to the header of the third will require a change to …
  - its block id (which is a hash of its header) ……

[so similar behavior to part i), except that cascade begins at second block]

Blockchain Data Modification Conclusion: THM

- If an adversary wants to change information recorded a long time ago (in early blocks), it is very hard, because it requires complete recomputation of all subsequent blocks.

- In practice, this will be infeasible, since the blockchain is replicated across computers in a P2P network, and it would require a node to convince all other nodes in the network to substantially revise their list of blocks. (i.e. consensus mechanisms)

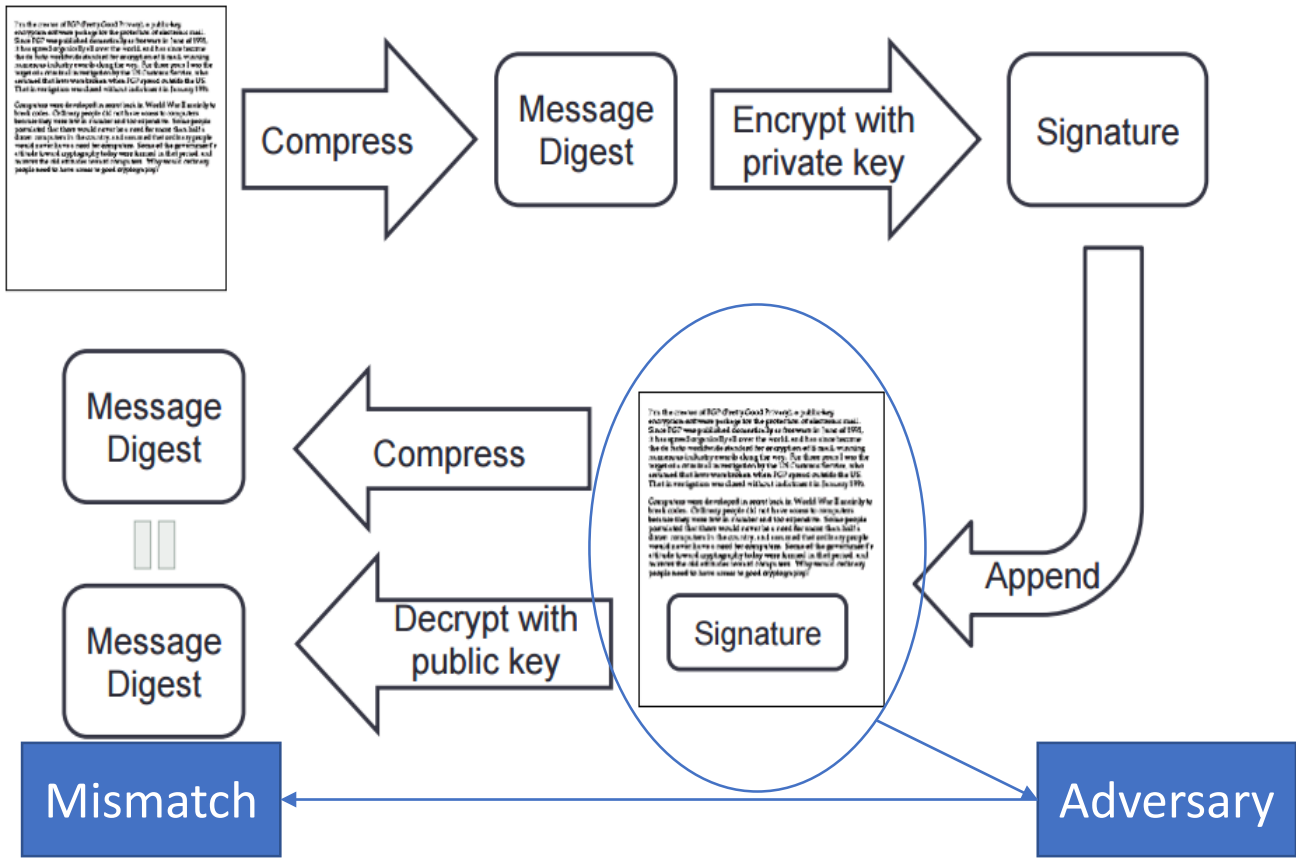
Blockchain

3. Suppose Bob signs a document with his digital signature. Fred receives the document and changes its contents, but leaves the digital signature unchanged. How could a third party (Alice), know that the document has been modified from its original version, by someone other than Bob?

- If the document is modified by an adversary, then this will change its hash value (i.e. message digest).

- This hash value will then not match the contents of the decrypted digital signature.

- The way public key cryptography works, no-one else knows Bob's private key.



Quasi-identifier

4a. Consider the quasi-identifier {job, birth, postcode}. Is data in the following table 1- anonymous? Is it 2-anonymous? Is it 3-anonoynous? Is it 4-anonymous?

Solution

- Is it 1-anonymous?
  - Yes
- Is it 2-anonymous?
  - Yes
- Is it 3-anonymous?
  - No
- Is it 4-anonymous?
  - No

Job	Birth	Postcode	Illness
Cat1	*	4350	HIV
Cat1	*	4350	HIV
Cat1	1955	5432	flu
Cat1	1955	5432	fever
Cat2	1975	4350	flu
Cat2	1975	4350	fever

Quasi-identifier

4b. Consider the quasi-identifier {gender, date of birth, zipcode}. Apply generalisation to the following table to make it 3 anonymous.

Name	Gender	Date of birth	ZIP code	Disease
Alice	F	01/01/1981	11111	Flu
Anne	F	02/02/1981	11122	Flu
Sonia	F	12/03/1981	11133	Flu
Bob	M	12/01/1982	33311	Heart disease
Shunsuke	M	10/04/1982	33322	Cold
Carl	M	02/03/1982	33333	Flu

Solution

QI attributes			PI attribute
Gender	Date of birth	ZIP code	Disease
F	1981	111*	Flu
F	1981	111*	Flu
F	1981	111*	Flu
M	1982	333*	Heart disease
M	1982	333*	Cold
M	1982	333*	Flu

Note: We would first get rid of name attribute before doing any further processing, we then just work with the quasi identifiers and sensitive (attribute)

Quasi-identifier

4c. Consider the quasi identifier {Age, Zip} for the table below. Is the data 1-anonymous? Is it 2-anonymous? Is it 3 anonymous? Is it 4 anonymous? Is it 5 anonymous? With respect to the sensitive attribute Diagnosis - is it 1-diverse? Is it 2-diverse? Is it 3-diverse? Is it 4-diverse?

Solution

- Is it 1-anonymous?
  - Yes
- Is it 2-anonymous?
  - Yes
- Is it 3-anonymous?
  - Yes
- Is it 4-anonymous?
  - No

- Is it 1-diverse?
  - Yes
- Is it 2-diverse?
  - Yes
- Is it 3-diverse?
  - No
- Is it 4-diverse?
  - No

Age	Zip	Diagnosis
[21–28]	9****	Measles
[21–28]	9****	Flu
[21–28]	9****	Flu
[48–55]	92***	Cancer
[48–55]	92***	Obesity
[48–55]	92***	Obesity

Differential Privacy

5. In the context of providing differential privacy:

- What is global sensitivity G? What is the privacy budget k?
- How does the G/k ratio affect the noise level?

Solution

- *Global sensitivity is evaluating the maximum possible change in query output due to a presence of a single record.*
  - How much difference the presence of absence of an individual could make to the result.

- *The privacy budget determines how close the query result for a database with the record is expected to be compared to query result for a database without the record.*
  - How hard for the attacker to guess the true result

- *For smaller k, or larger global sensitivity, more noise will be added to the query result.*
  - Released result = True result + noise
  - Where noise is a number randomly sampled from a distribution having – average value=0 – standard deviation (spread)= G/k

Differential Privacy

6. Consider a survey that collects two values from the respondents, e.g., marital status and gender.

- Consider a query that takes the survey database as input and outputs a pair of counts (CountNumberFemale, CountNumberMarried).
  - How much can adding or removing an individual affect the output?
    - ***Solution***: *Adding or removing any individual can affect the count of each column by maximum 1.*
  - What is the global sensitivity?
    - ***Solution***: *The maximum difference a single record can make query is 1+1=2*

- Consider a query that takes the survey database as input and outputs the quadruplet of counts (CountMaleMarried, CountMaleSingle, CountFemaleMarried, CountFemaleSingle).
  - How much can adding or removing an individual affect the output?
    - ***Solution***: *Adding or removing any individual can affect the count of each column by maximum 1. If it affects the count of some column by one, then it will affect the counts of other columns by 0 (since the columns are mutually exclusive, you can't have a 1 in both columns)*
  - What is the global sensitivity?
    - ***Solution***: *The maximum difference a single record can make To F is thus 1+0+0+0=1*

COMP20008 – Exam 2016

Age	Weight
20	40
30	50
25	25

7. Consider the following dataset D which describes 3 people:

a) (2 marks) Using a Euclidean distance function, construct the dissimilarity matrix for D. Show all working. (You may leave any square root terms un-simplified).

b) (2 marks) A dissimilarity matrix can be visualised as a heat map. Explain why it can be useful to reorder a dissimilarity matrix, using a technique such as the VAT algorithm. What information can this provide and why?

c) (2 marks) Would there be a benefit of applying principal components analysis to D to assist in visualisation? Explain.

COMP20008 – Exam 2016

7. Consider the following dataset D which describes 3 people:

a) (2 marks) Using a Euclidean distance function, construct the dissimilarity matrix for D. Show all working.

Instance id	1	2	3
1	0	Sqrt(200)	Sqrt(250)
2	Sqrt(200)	0	Sqrt(650)
3	Sqrt(250)	Sqrt(650)	0

b) (2 marks) A dissimilarity matrix can be visualised as a heat map. Explain why it can be useful to reorder a dissimilarity matrix, using a technique such as the VAT algorithm. What information can this provide and why?

Answer: Visualizing as a reordered heat map using algorithm such as VAT – identify the cluster structure of the data. i.e. How many clusters there are and which objects are in each cluster. Might also help identify anomalies – which objects are not similar to other objects.

c) (2 marks) Would there be a benefit of applying principal components analysis to D to assist in visualisation? Explain.

Answer: Little apparent benefit in applying PCA – the dataset is only 2 dimensions and already easy to visualize.

COMP20008 – Exam 2016

8. Bob and Alice are having a debate about imputation methods for missing values. Bob argues that imputation using the mean of a feature works best. Alice argues that imputation using matrix factorization is better. To decide who is correct, they select a dataset with 10 features F1, . . . , F10 and 100 instances x1, . . . , x100. For each instance, they randomly make one of its feature values to be missing (yielding 100 missing values overall). They then separately apply each of the two imputation methods and assess which performs better in recovering the missing values.

i) (2 marks) What metric would be appropriate here for assessing performance of an imputation method? Provide a formula for your metric and justify it.

ii) (3 marks) Bob and Alice are joined by a third student, Barbara. Barbara says "You are both wrong. If an instance has any missing values, it is better to discard it entirely rather than trying to perform imputations." Describe three scenarios which support Barbara's reasoning.

COMP20008 – Exam 2016

i. (2 marks) What metric would be appropriate here for assessing performance of an imputation method? Provide a formula for your metric and justify it.

Answer:

- Use a metric that finds aggregate deviation from the true answers. E.g. something like Mean squared error:

$$\text{MSE} = \frac{1}{100} \text{x} \sum_{i=1}^{i=100} (\text{true\_value}(x_i) - \text{imputed\_value}(x_i))^2$$

- Where i is an index that ranges over the 100 missing values.
- Could also use the mean absolute error as well (average of the absolute values of the deviations)

COMP20008 – Exam 2016

ii. (3 marks) Bob and Alice are joined by a third student, Barbara. Barbara says "You are both wrong. If an instance has any missing values, it is better to discard it entirely rather than trying to perform imputations." Describe three scenarios which support Barbara's reasoning.

Answer:

Reasons it might be better to discard:

- We already have a large dataset, that contains sufficient information even when examples are discarded.
- If imputation method is likely to be computationally expensive (e.g. matrix factorization), then might choose discard instance if efficiency is important
- If we believe imputation is likely to cause problems or contaminate later analysis (due to its unreliability)
- Scenarios where each instance is either complete (has nothing missing), or has mostly missing values.