COMP20008 Elements of Data Processing

Blockchain data processing

Announcements

- Reminder
  - Phase 4 oral presentations next week
    - Follow the questions asked in the project specification
    - Not expecting to hear all the results, just a taste
    - Submit slides to LMS 30min after presentation
  - We take into account that presenters earlier in the week will have had less time to prepare since Phase 3 completion

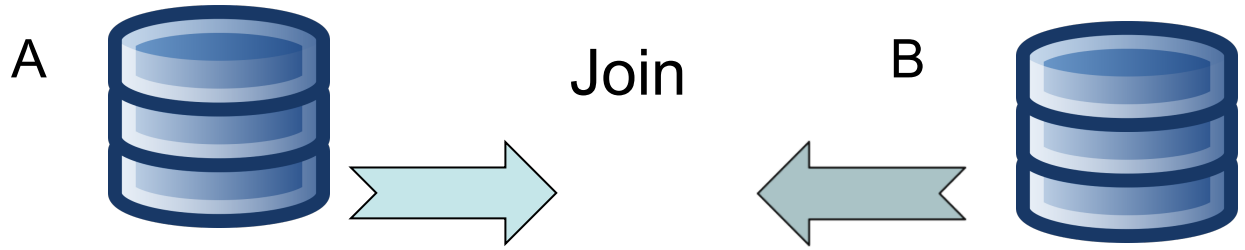
Plan today

- Blockchain
  - Relevance to data processing
  - Benefits
  - Structure and operation (sketch)
  - Applications

- Discussion
  - Blockchain and Unimelb
    - Guest: Dr Sandra Milligan, Director of Assessment Research Centre at The University of Melbourne

Blockchain and data processing

We have studied the integration of separate datasets, from two parties, A and B



- *Suppose we have thousands of parties sharing data?*
- *Suppose we want the data to be easily accessible to all?*

- Blockchain is an infrastructure to support this. Based on some core computing technologies
  - Peer to peer, hashing, public key cryptography, …

Blockchain: what is it?

- Blockchain
  - A distributed database
  - Stored on many computers
  - No central point of control

- *"Notoriously difficult to explain"*

  --Mark Pascall

- *"The technology likely to have the greatest impact on the next few decades has arrived. And it's not social media. It's not big data. It's not robotics. It's not even AI. You'll be surprised to learn that it's the underlying technology of digital currencies like Bitcoin. It's called the blockchain."*

  —Don Tapscott

Blockchain: what is it? Cont.

- A public record, known as a ***ledger***
  - Can record events, facts, asset transfers, …
    - E.g. Graduation certificate, exam marks, medical test results, transfer of money, …
- The ledger is typically public and shared between many parties
  - Some of these parties may be hostile/untrustworthy
- Once data is entered into the ledger, it can't be altered
  - No deletions, revisions, alterations, ….
- The integrity of the ledger can be verified
  - Ensure that it hasn't been tampered with

What's the benefit of blockchain?

- Who typically keeps track of our digital assets?
  - Money: banks have a record
  - Health: Doctors/Hospitals keep our health records
  - Education: Universities record academic results/graduations

- To transfer my money -> involve bank
- To share my health data -> involve doctor/hospital
- To share my academic credentials -> involve university

- In each scenario there is a middleman (trusted party) holding private records
  - Bank
  - Doctor/hospital
  - University

What's the benefit of Blockchain? Cont.

- What if we could remove the middleman?
  - No central, trusted point of control
  - Benefits
    - Less administration, less bureaucracy
    - Less expensive
    - Faster transactions
    - More control over records handed to users
    - Users can verify data in the blockchain
    - More secure solution (maybe)

Blockchain: Where did it come from?

- Proposed in 2008, by Satoshi Nakamoto
  - Software released on Sourceforge the following year. Underpinning of the digital currency bitcoin.

- Who is Satoshi Nakamoto?
  - No-one knows …….

Bitcoin: A Peer-to-Peer Electronic Cash System

Satoshi Nakamoto
satoshin@gmx.com
www.bitcoin.org

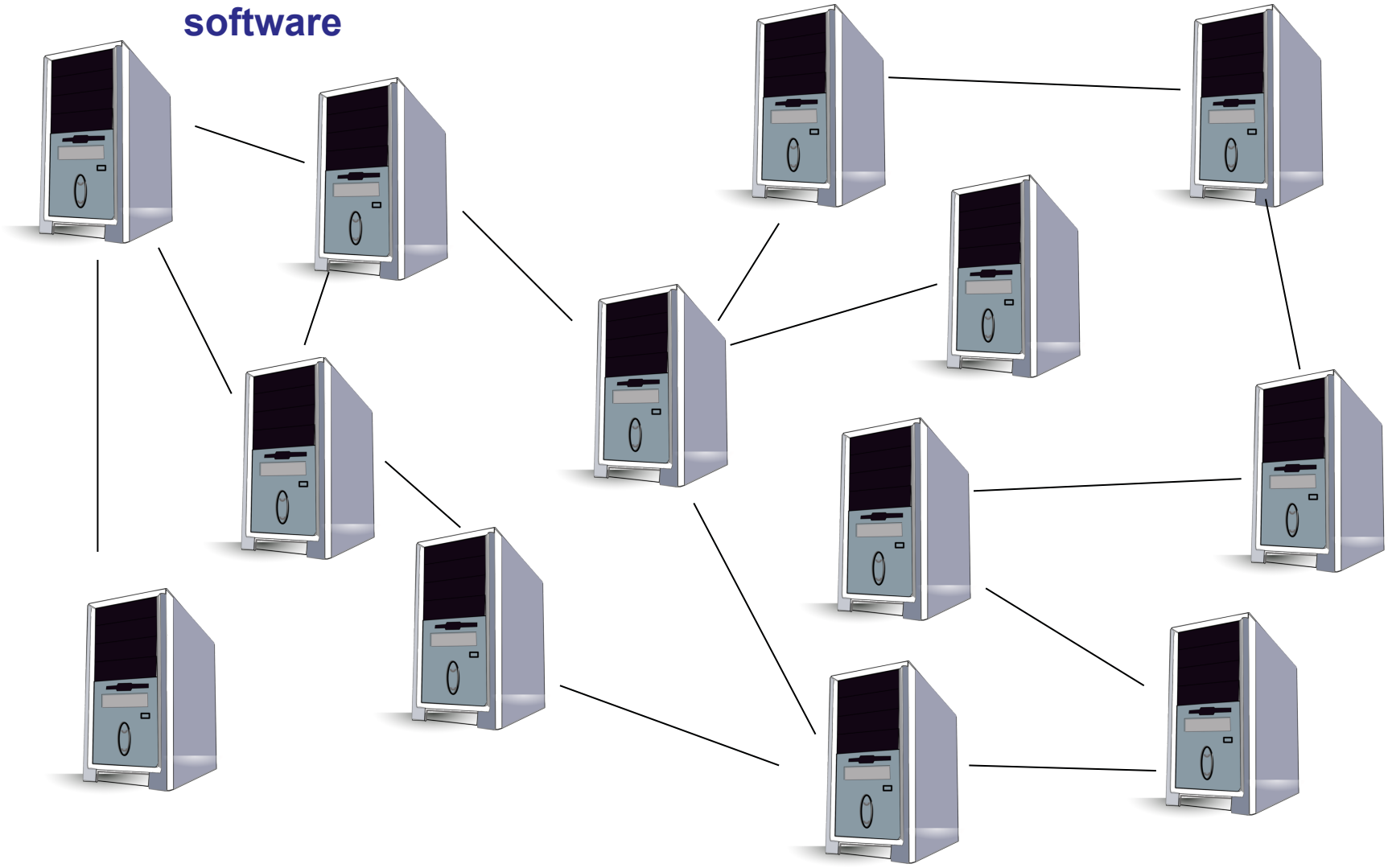
Abstract. A purely peer-to-peer version of electronic cash would allow online payments to be sent directly from one party to another without going through a financial institution. Digital signatures provide part of the solution, but the main benefits are lost if a trusted third party is still required to prevent double-spending. We propose a solution to the double-spending problem using a peer-to-peer network. The network timestamps transactions by hashing them into an ongoing chain of hash-based proof-of-work, forming a record that cannot be changed without redoing the proof-of-work. The longest chain not only serves as proof of the sequence of events witnessed, but proof that it came from the largest pool of CPU power. As long as a majority of CPU power is controlled by nodes that are not cooperating to attack the network, they'll generate the longest chain and outpace attackers. The network itself requires minimal structure. Messages are broadcast on a best effort basis, and nodes can leave and rejoin the network at will, accepting the longest proof-of-work chain as proof of what happened while they were gone.

Explaining blockchain: context

- Blockchain is a complex technology
- Best known use is for the digital currency bitcoin
  - Ethereum is another well known blockchain
- In what follows, we discuss some core elements of blockchain
  - Leave out many of the details
  - For simplicity, we avoid too much focus on the currency applications (bitcoin)
- Analogy
  - Blockchain is like an operating system. Bitcoin is like an application that can run on that operating system.

Blockchain: What does it look like?

- Peer to peer network of computers each running blockchain software



Before we go further

- We are going to need to use secure hash functions
  - A core tool used in blockchain

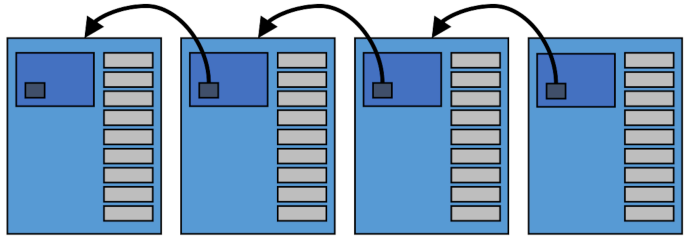
- A quick reminder ..

Reminder: secure hash algorithms

- Secure hash algorithm
  - Takes a string as input and creates an output of fixed length
- SHA-256 hash function (256 bit output)
  - SHA-256(COMP20008 is good) = 2160e01e8a6cdf4b515c04b657bc1aea4c3c99171c17ab8b8bcb2281e5038598
  - SHA-256(COMP20008 is good!) = be0b946ea5bfbcc5fd45fef461f207404266947d9d589ed3b74f2b731f8b4fbc
- Input that are similar give very different outputs
- Given the output, it is infeasible to recreate the input
- Input can be of any length
- Input can be any file (a text file, a song, a movie, …)

Each computer in the network

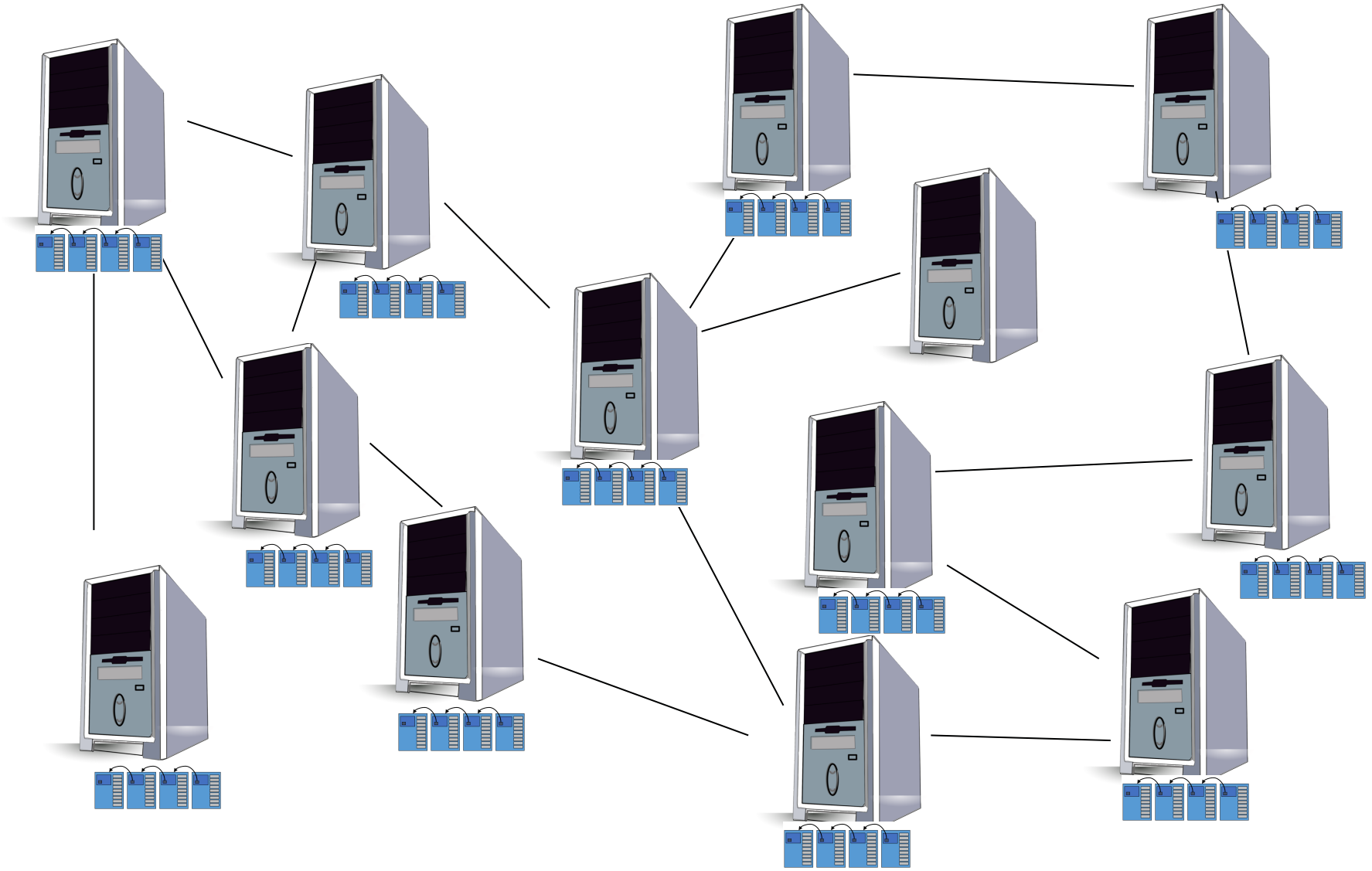
- Each computer in the blockchain network has a complete copy of the public ledger



Ledger

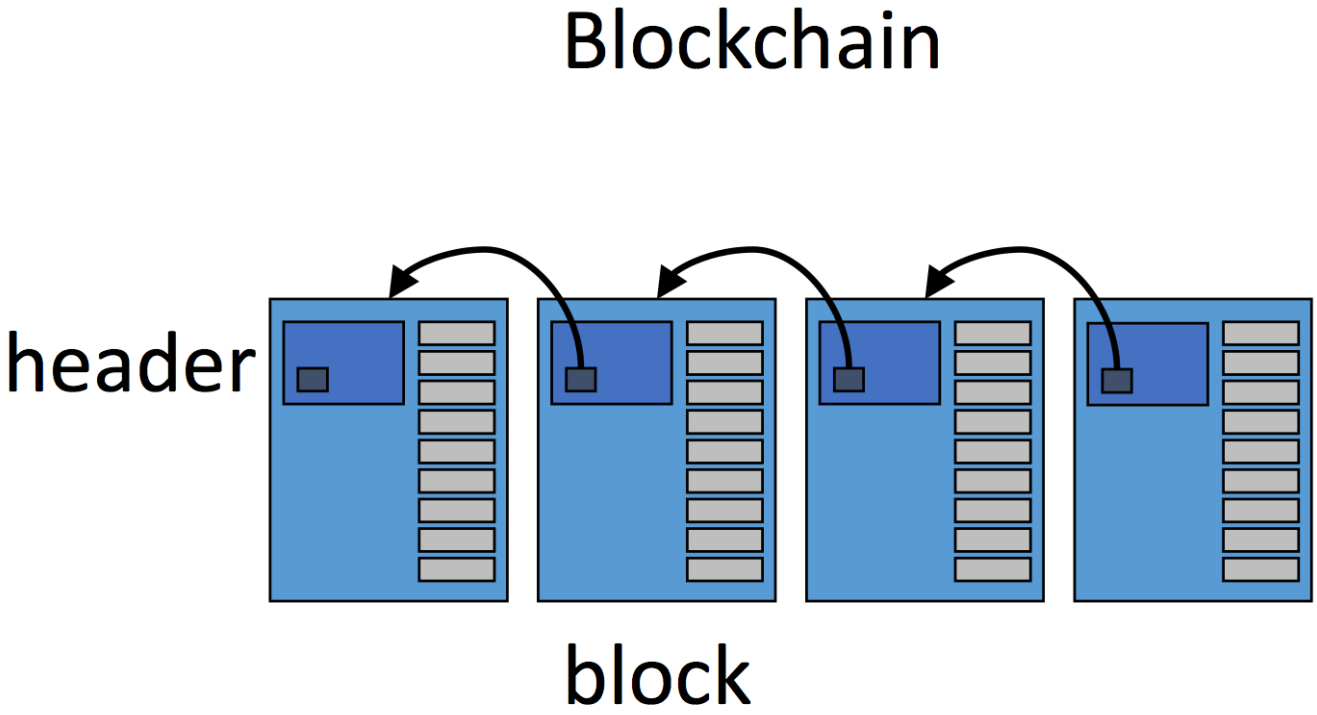
Copies of the ledger

- Each computer in the network has a copy of the ledger



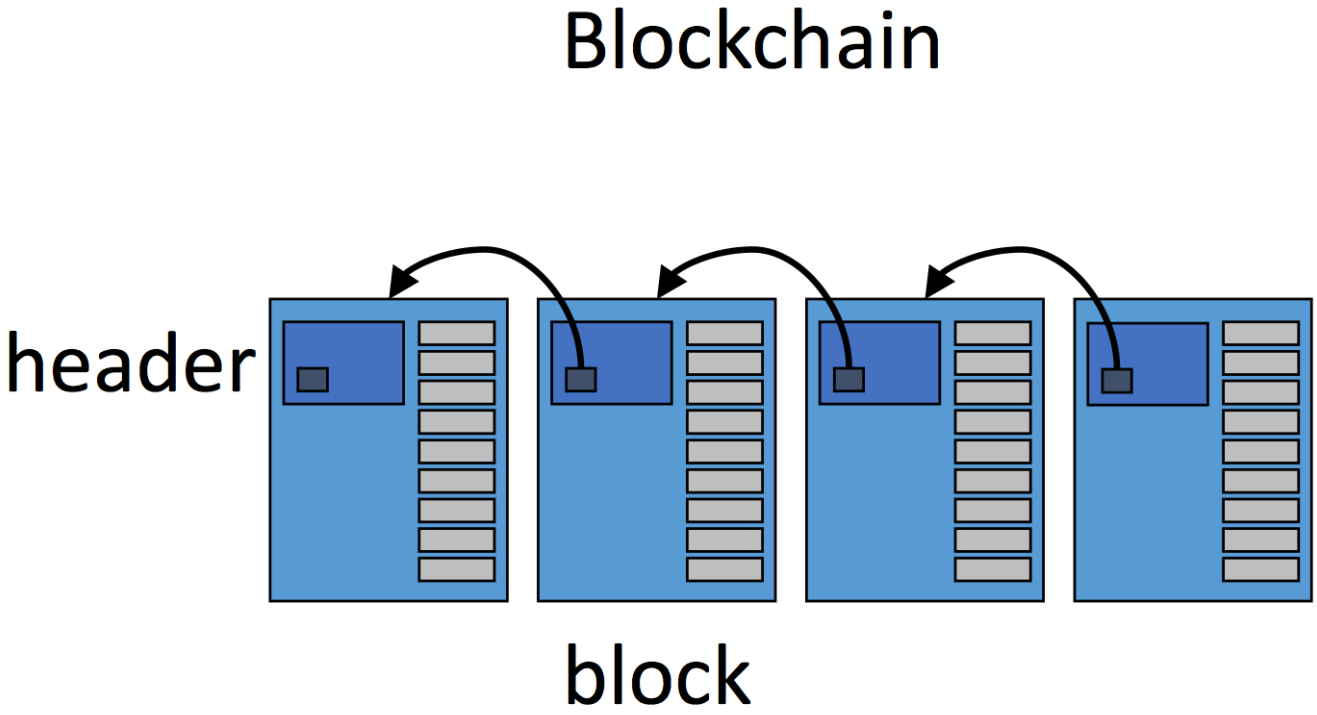
The blockchain

- The public ledger is called the blockchain (a file)
- File is a sequence of blocks, each block contains a header and some data (list of transactions)
- Block ID is equal to a hash of its header
- Each block contains the ID of its parent block



The blockchain

- A block's header typically includes
  - Hash ID of its parent block
  - Timestamp of block's creation
  - Hash of the data (list of transactions) inside a block
  - Example blockchain (bitcoin)
    - http://blockr.io/



Chaining characteristics

- The ID of a block derives from a hash of its header
  - So if a block's header changes, the block's ID changes
- If a parent block is modified, then its header will change and so the parent's ID will change
- If parent's ID changes, then this will change the previous block hash insider the child's header
  - Since the child's ID derives from its own header, the child's ID will therefore change.
  - This in turn will cause a change in the grandchild, which is dependent on the child's ID
  - This in turn will cause change in the great grandchild ….

- *Bottom line: changing a block produces a cascade effect requiring recalculation of all subsequent blocks ……*

Adding data to the blockchain

- Bob has a fact he wants to add to the blockchain
  - His computer broadcasts it to neighbors it is connected to in the blockchain network
  - Those neighbors receive the fact, validate the correctness of its format, then broadcast it to peers they are connected to. These peers recursively follow the same procedure.
  - At some point, a peer will aggregate a collection of facts (transactions) it has received, place them into a block, create an appropriate header and broadcast the block to its neighbors

Adding data to the blockchain cont.

- In practice, it is not so simple
  - Nodes might create blocks simultaneously and propose they be added as the next block in the blockchain
    - Must resolve discrepancies to reach consensus
  - Nodes might try to disrupt the blockchain by creating false facts (E.g. spending the same money multiple times)

- In the currency (bitcoin) blockchain, there are mechanisms that make it hard for a node to create and add blocks
  - Must solve a complex computational puzzle to add a block
  - Nodes are rewarded if they solve the puzzle first

- In a (semi-)trusted scenario, might stipulate that only trusted nodes can create blocks
  - E.g. In an education blockchain, only universities might create blocks

More fundamentals

- Public key cryptography is a fundamental element of blockchain technology, used for promoting security and privacy
  - Public and private keys
  - Digital signatures

- We briefly review these notions next.

Public-key cryptography: Public and private keys

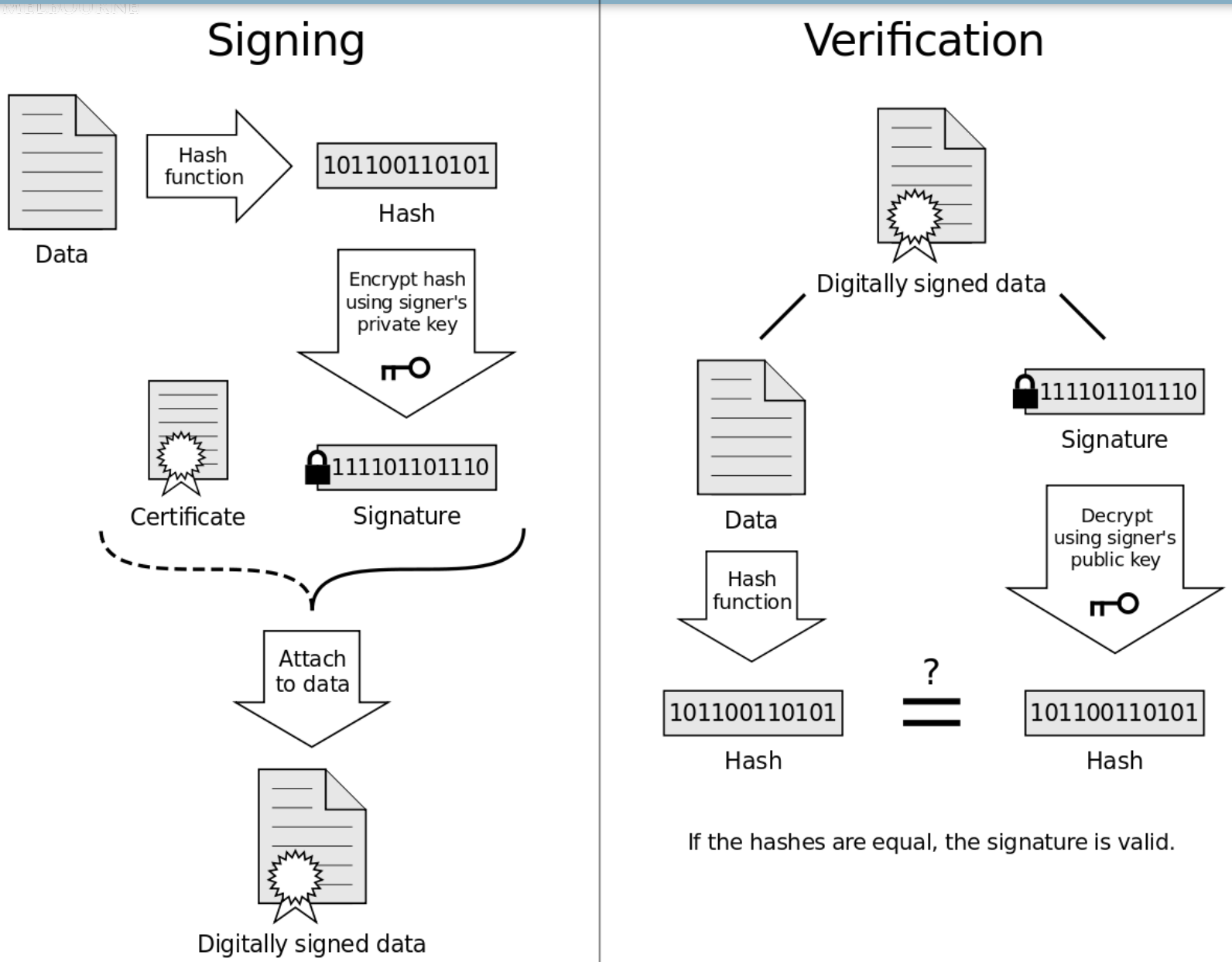
- Using standard software, Bob can create two keys (two long numbers)
  - A private key (information known only by him)
  - A public key (information he shares with other people)

- These keys have special properties
  - Alice can encrypt a message using Bob's public key. Only Bob can decrypt and read the message, using his private key.
  - Bob can encrypt a message using his private key. This message can only be decrypted using his public key.

Digital signatures

- Bob has two keys
  - Private key=….
  - Public key=…..
- Bob wants to place his digital signature on a document
- The digital signature is evidence that Bob authorised the contents of the document
  - Similar to a real signature on a contract
- Alice wants to verify that Bob's signature is real (it could only have come from Bob)

Digital signature (from wikimedia commons)



Digital signatures and the blockchain

- Suppose the University of Melbourne wishes to add a fact to the blockchain
  - "Alice graduated with a BSc in 2017 from the University of Melbourne"

- The University of Melbourne node sends this data to the blockchain. It eventually gets included in a block.

- Everyone can now see this fact, but how do they know it is true?
  - A digital signature from the University of Melbourne is also attached to the fact.
  - Anyone can now verify that University of Melbourne approved this fact.

But what if the data is sensitive?

- What about sensitive facts? People don't want these to be publically viewable on the blockchain.
  - "Alice graduated with a BSc in 2017 from the University of Melbourne"

- The University applies a hash to the fact then adds this hash output to the blockchain along with a digital signature.
  - 07ce291c716fd0346d1fb84968062f9a4648af6ba70cae9b4d400f74dee796d2 ….
  - No-one can reverse the hash function to uncover the fact
  - Now stored on the blockchain, but Alice's privacy is preserved

  - Alternatively could encrypt the fact using Alice's public key.

But what if the data is sensitive? Cont.

- Alice provides prospective employer with
  - The fact: Alice graduated with a BSc in 2017 from the University of Melbourne
  - A reference to the blockchain block containing this hashed information

- Employer can apply hash function to the fact and see that it is stored on the blockchain. They can also verify the fact was digitally signed by the University of Melbourne

Blockchain: some possible applications

- Currency
- Health
- Education

Blockchain application: Bitcoin

- Blocks contain information about transactions involving the bitcoin currency
  - Movements of money between individuals
  - Complex protocols for checking whether someone has money available to be spent.
  - Complex consensus protocols to avoid hacking attacks that tamper with the blockchain in order to spend the same money multiple times

- Around 7000 nodes, completely decentralised
- 300k transactions per day, 2k transactions per block
- Total circulation value of bitcoins: ~$USD30 billion

- Can view blocks being created at
  - http://blockr.io/

Blockchain application: health

- Patient data is added to the blockchain
  - Each visit to the doctor
  - Blood tests
  - MRI/CT Scans
  - Vaccinations
  - Prescriptions
  - Physiotherapy, …
  - Fitbit and wearables data???
- Data is encrypted with patient's public key
  - Patient may provide key to insurer or health provider so they can review their medical history. Reduced time, fuller information available and quicker

- Estonia
  - Developing a blockchain system for 1 million medical records

Blockchain application: education

FINANCIAL REVIEW

search the AFR

NEWS BUSINESS MARKETS STREET TALK REAL ESTATE OPINION TECHNOLOGY PERSONAL FINANCE LEADERSHIP LIFESTYLE ALL

Home / Leadership

Apr 30 2017 at 3:16 PM | Updated Apr 30 2017 at 3:16 PM

Save Article | Print | License Article

University of Melbourne first in Australia to use blockchain for student records

- Cheap, secure, shared resource to store credentials and micro-credentials
- Compare with the vision of Institute for the Future
  - https://www.youtube.com/watch?v=DcP78cLPGtE&feature=youtu.be
  - University degrees, MOOCs, Khan academy, corporate training, conference attendance, …..all on the blockchain?

Some open issues

- Open questions
  - Revocation
    - University revokes a degree
  - Permissions and privacy
    - What is appropriate?
  - Who can be trusted to add blocks to the block chain?
  - Smart contracts
    - Make a payment to the University of $100 every day the temperature is less than 10 degrees

THE UNIVERSITY OF MELBOURNE

Discussion with Sandra Milligan

- What is the University of Melbourne doing with blockchain?
- What are the benefits to the University and students?
- What are the barriers/risks?

What you need to know from today

- Motivation for blockchain technology
  - What problems is it trying to solve?
  - Why is it useful?
  - How are blocks chained together?
  - How is hashing used to identify and link blocks?
  - How are digital signatures used to verify data on the blockchain?
  - How is hashing used to make information on the blockchain private?

Further information (optional, only if you want to know more)

- How the blockchain is changing money and business | Don Tapscott
  - https://www.youtube.com/watch?v=Pl8OlkkwRpc
- How does blockchain work?
  - https://medium.com/@micheledaliessi/how-does-the-blockchain-work-98c8cd01d2ae
- Blockchain and education
  - http://hackeducation.com/2016/04/07/blockchain-education-guide

Plan for coming lectures

- Next (final) four lectures
  - Cloaking data: l-diversity and k-anonymity
  - Issues in location privacy
  - Ethics and analytics
  - Subject wrap-up, exam discussion and future directions