Distributed Trust and Blockchains		Date:	*31st October 2019*
Instructor: *Sujit Prakash Gujar*		Scribes:	Nuthalakanti Rupa, Ayush Deva

Lecture 14: Distributed Consensus and Differential Privacy

1 Recap

In this lecture, the discussion was mostly on ZKP (Zero Knowledge Proof), ZK SNARK and Cryptographic accumulator and then we discussed a paper on local coin.

1.1 Zero Knowledge Proof

In Cryptogaphy, ZKP is an interactive system in which one party can prove to another party that they know a value x, without conveying any information apart from the fact that they know the value x.

There is a Prover (P), who knows a secret (S) and a Verifier (V) such that,

- P proves to V that he/she knows S.
- V doesn't know S.
- V couldn't prove that P knows S.

For example,we can do this by using the function

$$c = g^r mod p \tag{1}$$

You can also prove a statement such as "I know x such that H(x) belongs to the following set: ...".The proof would reveal nothing about x, nor about which element of the set equals H(x). Zerocoin crucially relies on zero-knowledge proofs and in fact the statements proved are very similar to this. Zerocoin is usually 6 times slower than bitcoin.

1.1.1 ZK SNARK

It stands for Zero-Knowledge Succinct Non-Interactive Argument of Knowledge. It is a non-interactive system. As the name suggests,it does not require an interactive process, avoiding the possibility of collusion, but may require additional machines and programs to determine the sequence of experiments. It is basically an improved version of ZKP without interaction between P and V.

1.1.2 Examples

- Leader selection by proof of Work.
- Chain selection by choosing the longest chain.

1.2 Cryptographic Accumulator

We can think of a crypto accumulator as a super-charged hash function(Quasi-Commutative hash function) that works on sets. A regular hash function, like SHA-3, takes a single message and outputs a fixed-size hash. An accumulator, however, takes a set of values and turns them into a single number, also of constant size. It allows membership proofs for all the elements with in the set and non-membership proofs for elements not in the set.

$$H\ (H(x\ ,y)\ ,z) = H(H(x\ ,z)\ ,y)$$

1.3 Local Coin

- Main motivation of Local coin is that Bitcoin and other crypto-currencies demand high computational power devices and internet connectivity
- It replaces the computational hardness.
- It uses proof of verification.
- It is based on opportunistic networking rather than relying on infrastructure and incorporates characteristics of mobile networks such as users' locations and their coverage radius
- Each user selects a set of trusted users (NFC pairing)
- The receiver of one transaction accepts the transaction iff she received the transaction signed by at least p1 users of her trusted network.
- The transactions are verified in bunches of p2
- at least p3 users are needed to verify each transaction
- the average physical distance between the users that verify the creation of a new block has to be more than p4

2 Introduction

2.1 Distributed Consensus

Distributed consensus is achieving agreement on a single data value among distributed processes or systems.This is the major technical problem to be solved in a distributed e-cash system.

2.1.1 Distributed consensus protocol

- Termination: Every non-faulty process must eventually decide.
- Agreement: The final decision of every non-faulty process must be identical.
- Validity: If every non-faulty process begins with the same initial value v, then their final decision must be v

2.1.2 Distributed Consensus vs Distributed Agreement

The main difference is that

- In Distributed Consensus, all the n nodes have an initial value,of which some are malicious nodes. When all the honest nodes agree on a single value, consensus is reached.
- Whereas in Distributed Agreement, there's only one source node. Consensus is reached when all the honest nodes agree on this particular value.

3 Bitcoin: A Peer-to-Peer Electronic Cash System

3.1 Calculations

We consider the scenario of an attacker trying to generate an alternate chain faster than the honest chain. Even if this is accomplished, it does not throw the system open to arbitrary changes, such as creating value out of thin air or taking money that never belonged to the attacker. Nodes are not going to accept an invalid transaction as payment, and honest nodes will never accept a block containing them. An attacker can only try to change one of his own transactions to take back money he recently spent.

The race between the honest chain and an attacker chain can be characterized as a Binomial Random Walk.
The success event is the honest chain being extended by one block, increasing its lead by +1
And the failure event is the attacker's chain being extended by one block, reducing the gap by -1.

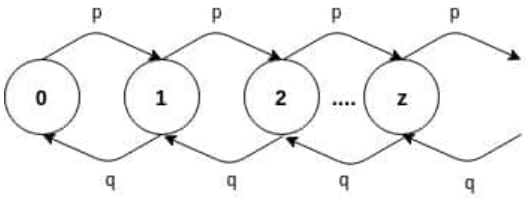

Figure 1: Mining race : Attacker vs Honest Node

We can calculate the probability he ever reaches breakeven, or that an attacker ever catches up with the honest chain, as follows

p : Probability an honest node finds the next block
q : Probability the attacker finds the next block
q_z : Probability the attacker will ever catch the honest node when he is z blocks behind
where p+q=1.

$$q_z = \begin{cases} 1, & \text{if } q \geq p \\ \left(\frac{q}{p}\right)^z, & \text{otherwise.} \end{cases} \tag{2}$$

Suppose the recipient received a coin from attacker, he waits until the transaction has been added to a block and z blocks have been linked after it.
After z blocks are mined, number of blocks mined by attacker will be a poison distribution with expected value:

$$\lambda = \frac{z.q}{p} \tag{3}$$

Probability that the attacker could still catch up or double spend:

$$P = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} \cdot \begin{Bmatrix} (q/p)^{(z-k)}, & \text{if } k \leq z \\ 1, & \text{if } k \geq z \end{Bmatrix} \tag{4}$$

Rearranging (4), we get,

$$P = 1 - \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} \left(1 - (q/p)^{(z-k)}\right) \tag{5}$$

Say p=0.7, q=0.3 and z=5
Then probability of Double spending = 0.1773

Say p=0.9, q=0.1 and z=5
Then probability of Double spending = 0.0009

So we can say that as p increases, probability of Double spending decreases. That means, increase in number of honest nodes(miners) leads to lesser chance of double spending.

4 Boot Strapping

Success of a crypto currency is mainly dependent on the following ideas:

- Number of honest miners
- Security of the block chain
- Value of the currency

As discussed above more honest miners lead to secure block chain. A high and stable value of currency is ensured only when users in general trust the security of the block chain. Any user wants to join the block chain network to a viable currency.
This leads to a cyclic dependence among these 3.

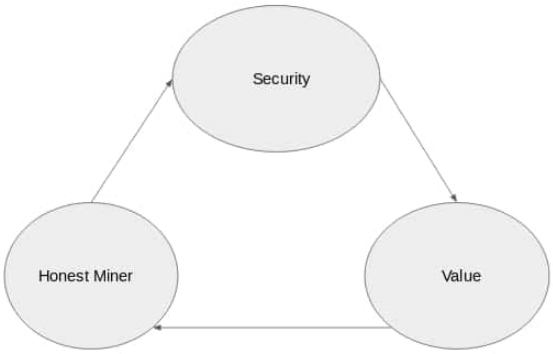

Figure 2: Bootstrapping

Change in one of them leads to corresponding change in other two. Bootstrapping is the process of starting the change in this cycle.

5 Differential Privacy

5.1 Introduction and Motivation

1. **Free Riding:** If I derive a utility of U, by bearing a cost c, I will contribute and/or decide to participate based on the values of U and c. However, in say a resource-shared system, I can enjoy U without having to pay c necessarily, I will opt out and simply enjoy the utility. This

is called free-riding. In game theory, these situations are also referred to as "Tragedy of the Commons".

2. **Data Sharing:** On a daily basis, we are interacting and sharing our data with many systems or interfaces. These include social networks, online services, online markets, search engines, national identification systems like Aadhaar, institutions like bank, schools and hospitals.

Statistical Modelling and Data Analytic are become ubiquitous with every passing day. With increasing volume and detailing of data on individuals along with powerful collection and analytical tools available today, the need for privacy-preserving algorithms for such analysis has become increasingly important. One of the examples of such data sharing and its use by analysts is shown in Figure 3.

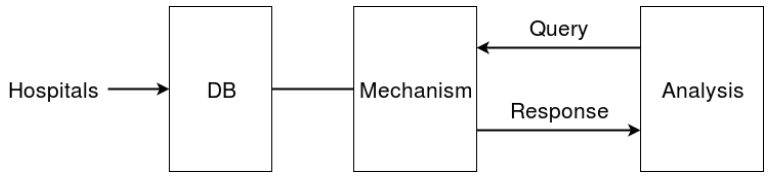

Figure 3: Data Sharing

5.2 Definition

According to the book, The Algorithmic Foundations of Differential Privacy [4],

"Differential privacy" describes a promise, made by a data holder, or curator, to a data subject: "You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available."

Essentially, Differential Private Mechanisms allow data analysis to happen without completely compromising the privacy of the individual. However it is important to note that any mechanism that outputs overly accurate answers can be misused to destroy privacy (Fundamental Law of Information Recovery). We can only trade-off between accuracy of queries and the privacy of individuals and that is what the field of differential privacy hopes to address in a systematic way.

Differential privacy ensures that the same conclusions, will be reached, independent of whether any individual opts into or opts out of the data set.

5.3 Alternative Approaches and Limitations

1. **Anonymize the Data:** Anonymization or removal of personal information refers to removing portions of data records that can be suppressed (usually personal information like name, age, sex etc.) but still allowing the remaining to be published and used for data analysis.

   However, the richness of the data enables "naming" an individual by several attributes, such as the combination of zip code, date of birth, or even the names of three movies and the approximate dates on which an individual watched these movies or the last three items bought off Amazon. This "naming" capability can be used in a **linkage attack** to match "anonymized" records with non-anonymized records in a different dataset. For examples, the medical records of the governor of Massachussetts were identified by matching anonymized medical encounter data with (publicly available) voter registration records.

2. **Low vs High Resolution Data:** One way to prevent privacy attacks is to prevent analysts to query for very specific information (high resolution data) but only limit them to low resolution queries that gives aggregated outputs instead of those on specific individuals.

However, we can still retrieve information about specific individuals by cleverly choosing our low resolution queries. For example, consider the following pair of queries:-

- Q1: Average weight of n people
- Q2: Average weight of above n people and Mr. X

Using the above two queries, we can obtain the weight of Mr. X which might not be desirable. These attacks are called **differencing attack**.

3. **Query Auditing** Another possible solution might be to audit, or filter, the sequence of queries that could lead to differencing attacks (or even other attacks). However, this can be challenging as:-
   (a) Refusing to answer a query is itself disclosive (can lead to leaking of information).
   (b) Auditing might be computationally infeasible or even worse, not even possible to predict algorithmically.
4. **Revealing Only Ordinary Facts:** Revealing only ordinary statistics may also not work. Say, a particular household that buys 5kgs of sugar every month suddenly reduces their consumption to 3kgs a month. Using this information, one can conclude that a family member might have been diagnsed with diabetes.
5. **Summary Statistics:** Allowing only summary statistics is also not safe. One can launch **"reconstructive attacks"** like the differencing attack discussed above against a database to reveal secret information about an individual.

6 References

1. Bitcoin and Cryptocurrency Technologies, Arvind Narayanan, Joseph Bonneau, Edward Felten,Andrew Miller, Steven Goldfeder
2. https://www.cs.uic.edu/ ajayk/Chapter14.pdf
3. https://www.cs.helsinki.fi/group/close/edge-computing-2016/lib/slides/dmitris.pdf
4. The Algorithmic Foundations of Differential Privacy, Cynthia Dwork and Aaron Roth