Distributed Trust and Blockchains		Date:	*7th Nov 2019*
Instructor: *Sujit Prakash Gujar*		Scribes:	Samhita, Siddhartha Laguvarapu

Lecture 15: Differential Privacy and Randomized Algorithms

1 Recap

In the previous class we have covered Distributed Consensus, Distributed Agreement, A Paper entitled 'Bitcoin: A Peer to Peer Electronic Cash System' and Bootstrapping.

1.1 Distributed Consensus

There are n nodes each with an input value. Some of these nodes are malicious or faulty. The nodes must somehow put forth their input values, and all the honest nodes must agree on a single consensus value. Reaching an agreement is the fundamental goal in the distributed consensus.

- A distributed consensus protocol tolerating halting failures must satisfy the following properties:
  1. **Termination:** All honest nodes must agree on a single value.
  2. **Agreement:** The decision value by all honest nodes must be identical.
  3. **Validity:** If all the honest nodes proposed the same value v, then the final value agreed by all the honest nodes must be that same value v.
- Distributed Consensus has various applications, and it has been studied for decades as it is the key technical problem to solve in building a distributed e-cash system.

1.2 Distributed Agreement

- In distributed agreement, only single node proposes a initial value. The goal of this is all the honest nodes amongst the others to agree on a common value, the decision value.
  1. **Termination:** All honest nodes must agree on a single value.
  2. **Agreement:** The decision value by all honest nodes must be identical.
  3. **Validity:** : If the source node is honest, then the agreed upon value by all other honest nodes must be the same as the initial value of the source node.

1.3 Bitcoin: A Peer to Peer Electronic Cash System

'Bitcoin: A Peer to Peer Electronic Cash System' is a paper published by *Satoshi Nakamoto* on the cryptography mailing list describing the bitcoin digital currency.

- This paper proposes a solution to the double-spending problem using a peer-to-peer network.
- It presents an interesting evaluation of the dynamic evolution of two competing paths of the network, when faced with the prospect of an attacker building an alternate adversarial chain.
- Using a binomial random walk process to model this situation simplifies the problem, because one need to only consider the probability of the attacker's chain gaining a relative advantage of one block during each time period elapsed, and then playing this race out over time.
- At each step of the random walk, some random variable can assume one of two values, +1 or -1, with probabilities p and q, respectively.

- The probability that an attacker ever catches up with the honest chain is calculated, 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 up from z blocks behind

$$q_z = \begin{cases} 1, & \text{if } p \leq q \\ (\frac{q}{p})^z, & \text{if } p > q. \end{cases}$$

- The attacker will be z steps ahead of the honest chain is a probability distribution with expected value:

$$\lambda = z(\frac{q}{p})$$

- The probability with which the attacker could still catch up the honest node is:

$$P = 1 - \sum_{i=0}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!}(1 - (\frac{q}{p})^{(z-i)})$$

1.4 Bootstrapping

There is tricky interplay between three different and important ideas in Bitcoin. They are:

- The security of the blockchain
- The health of the mining ecosystem
- The value of currency

We want the blockchain to be secure for the bitcoin to be a viable currency. For the blockchain to be secure, an adversary must not be able to create a lot mining nodes and take over major percent of new block creation. This requires a mining ecosystem made up of honest nodes. The honest nodes should have trust in security of blockchain for them to put a lot of computational power.

The existence of each of these is predicated on the existence of the others, because of cyclical nature of this three-way dependence.

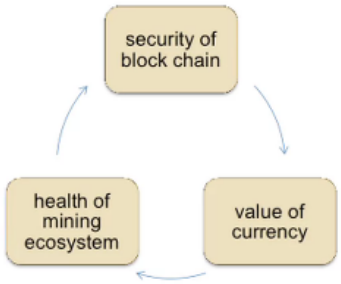

Figure 1: BootStrapping

Image credits: www.coursera.org/cryptocurrency

2 Differential Privacy

2.1 Motivation

If **I** derive a utility U by bearing cost C. The contribution of **I** here is based on U and C.

Free-Riding: If "We" derive utility U by everybody contributing cost C. **I** may opt-out and still enjoy the utility U.

2.1.1 Privacy issues in Data Sharing

Data analytics and machine learning are hugely valuable, providing insights and spurring advancements in many industries including IoT, healthcare, and financial services. Unfortunately, the threat here is the data that powers these advancements is often highly sensitive. In many cases this data cannot be accessed or shared due to privacy concerns. This results in data silos in which data is not used for its full potential value.

2.1.2 Alternative Solutions and their Limitations

i. **Anonymize the data:** The data is being shared after removal of personal information. However, **linkage attacks** are possible. A linkage attack attempts to re-identify individuals in an anonymized dataset by combining that data with background information. The 'linking' uses quasi-identifiers (such as postcode, gender, salary, etc..) that are present in both sets to establish identifying connections.

ii. **Low vs High Resolution Data:** The data privacy can be achieved by limiting the analysts to query low resolution data that gives aggregated outputs instead of those on specific individuals. But, **differencing attacks** are possible in this case. differencing attack attempts to retrieve information about specific individual by cleverly choosing the queries.

iii. **Query Auditing:** One other solution can be to audit the sequence of queries and responses, with the goal of interdicting any response if, in light of the history, answering the current query would compromise privacy. This leads to following difficulties:

- it is possible that refusing to answer a query is itself disclosive.
- query auditing can be computationally infeasible.

iv. **Summary Statistics:** This is not safe. Problems with summary statistics include a variety of reconstruction attacks against a database in which each individual has a "secret bit" to be protected.

2.1.3 Goal

- The goal of a privacy-preserving statistical database is to enable the user to learn properties of the population as a whole, while protecting the privacy of the individuals in the sample.
- We desire a system where the prior probability of the information an adversary has about an individual is equal to posterior probability after the data is being shared.

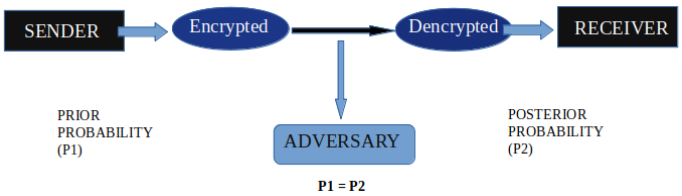

Figure 2

- **Differential Privacy** can help solve this problem.

2.2 Definition

Differential privacy is a system for publicly sharing the information about the dataset by describing the patterns of groups within the dataset while withholding information about individuals in the dataset.

- Informally, it assures that the result of a computation to be similar whether or not any individual is included in the analysis.
- Differential private algorithms are used by some government agencies to publish demographic information or other statistical aggregates while ensuring confidentiality of survey responses, and by companies to collect information about user behavior while controlling what is visible even to internal analysts.
- Differential privacy is a property of some randomized algorithms.

2.3 Randomized algorithm

Definition 1. *A randomized algorithm M with domain A and discrete range S is associated with mapping $M : A \to \Delta(S)$. On input $a \in A$, the algorithm M outputs $M(a) = b$ with probability $(M(a))_b$ for each $b \in B$. The probability space is over coin flips of the algorithm M.*

- $\Delta(S) = \{x_i \in R^{|S|} | x_i \geq 0, \sum_{i=0}^{|S|} x_i = 1\}$. It is called *Probability Simplex.*

2.4 Distance between Databases

Let $X = \{t_1, t_2, ..t_k\}$ be the universe and $x = (x_1, x_2, .., x_n)$, $y = (y_1, y_2, .., y_n)$ be two databases, where x_i is no.of entries of type i, $|x|$ is no.of entries in x and $x_i \in \mathbf{Z}_+^{|\mathbf{x}|}$

- The distance between two databases is x and y is defined as $||x - y||_1$
- The l_1 norm of $x = (x_1, x_2, .., x_n)$ is denoted by $||x||_1$ and is defined to be:

$$||x||_1 = \sum_{i=1}^{|x|} |x_i|$$

- Two databases are adjacent if they are equal or if they differ for the presence or absence of a single individual. i.e. $||x - y||_1 \leq 1$

2.5 Example Model

Let $x_i \in \{0, 1\}$, individual query x_i is either 0 or 1. There is a property P' which is true with probability p. If property P' is true for i then $x_i = 1$, else $x_i = 0$.
The response to the query is made according to the coin flip as follows:

- If the coin flip gives T : respond x_i
- If the coin flip gives H : flip again. If H respond '1', else '0'.

According to the model the expected no.of 1's are: $(\frac{P}{2} + \frac{1}{4})n$, where n is the no.of respondents. we see that $Pr(x_i = o|\widetilde{x_i} = o)$ is the probability with which the responded output is same as the actual value.
$Pr(x_i = 1|\widetilde{x_i} = 1) = \frac{\frac{3}{4}p}{\frac{P}{2} + \frac{1}{4}}$

Some information is being leaked with the above probability. Our intention here is to minimize this probability.

2.6 ϵ - differential private

Definition 2. *A randomized algorithm M is said to be ϵ- differential private if $\forall x, y$ such that $||x - y||_1 \leq 1$, we have*

$$Pr(M(x) \in E) \leq e^{\epsilon} Pr(M(y) \in E) \ \forall E \subset S$$

- if $\epsilon = 0$, then $Pr(M(x) \in E) = Pr(M(y) \in E) \forall E$, privacy is high as x and y are not distinguishable.
- if ϵ is high, there is drastic change in probability which leads to gain of information about the influencing agent. Hence the higher the more is the privacy loss.

2.7 Privacy Loss

In Differential Privacy we make sure that the individual data is not leaked. So, Data is shared only when the privacy loss is minimized. we calculate *Privacy loss* as follows:

$$\text{Privacy Loss} = \ln \frac{Pr(M(x)=O)}{Pr(M(y)=O)} \forall O$$

3 References

- Bitcoin and Cryptocurrency Technologies, Arvind Narayanan, Joseph Bonneau, Edward Felten,Andrew Miller, Steven Goldfeder
- http://homepage.divms.uiowa.edu/ ghosh/16612.week11.pdf
- https://medium.com/coinmonks/a-primer-on-bitcoin-a-peer-to-peer-electronic-cash-system-by-satoshi-nakamoto-95497e249749
- https://www.cis.upenn.edu/ aaroth/Papers/privacybook.pdf
- https://hackernoon.com/what-data-privacy-means-for-the-future-of-blockchain-c0212cd16680
- https://people.eecs.berkeley.edu/ stephentu/writeups/6885-lec20-b.pdf