1 Digital Currencies

1.1 Key Concepts

1.1.1 General Concepts

- Barter and the double coincidence of wants: Two people can only trade if A wants something B has and B wants something A has. What are some other problems with this?
- Currency is transferable, divisible, has value now and will (usually) have value in the future.
- When a currency is on a gold standard, the value of a currency is linked to the value of gold, and usually the currency can be exchanged at a fixed rate for gold.
- With fiat money, a government issues money that has value because the government has declared it a legal medium of exchange. What are some problems with fiat money?
- Digital currencies have cost advantages, privacy advantages, and decentralization advantages over traditional exchanges.
- In credit networks, users print their own money and trust the currencies printed by others to varying degrees.

1.1.2 iOwe

- With iOwe, iotas (similar to IOUs) are transferred between users. One user initiates an iota which can then be given to other users and then eventually redeemed with the issuer.
- Some attacks against iOwe include double-spending attacks, sybil attacks, and step-omission attacks.
- Two drawbacks to iOwe include bootstrapping trust with other agents, and establishing trust with an entire chain of agents to prevent sybil attacks.

1.1.3 Bitcoin

- Bitcoin attempts to limit the problems of iOwe by making it costly to create new currency, bounding the total amount of currency available, and keeping better track of currency ownership (preventing double spending).
- Each coin carries with it a history of all its previous owners.
• Using public/private key encryption, a coin’s owner is able to transfer a coin to another user. The second user is able to use his public key to check that the transfer has in fact occurred.

• Bitcoin uses distributed time-stamping to keep track of who owns which coins and thereby prevent double spending.

• Bitcoin achieves this by maintaining a block chain, which is essentially a distributed record of all bitcoin transactions. A bitcoin is only transferred when that transfer is recorded in the block chain. In order to add a block to the block chain, miners must performs some work (find an appropriate nonce), and are rewarded when they are successful.

• Only hashes with the correct number of 0s will be accepted. On average, a new block is accepted every ten minutes.

• When a user creates an acceptable block, he receives a fixed payment (12.5 Bitcoins, effectively out of thin air) and a transaction fee from all transactions contained in the block he created. Most transaction do include a fee since there is congestion trying to get into a new block.

• Forks happen when two miners accidentally make a block at the same time. It is resolved through whoever has a longer chain. Hard forks could occur when communities have disagreements over policies.

1.2 Exercises

1. A New Cryptocurrency

(a) What are the steps required for you to make a payment to Alice over BTC? How is this transaction encoded?

Solution: You must have a public key and private key. For each coin in the transaction, you take the last transaction involving that coin as well as Alice’s public key, and compute a hash value based on these two inputs. You use your private key to sign the hash value.

(b) How can Alice verify that you really own the BTC that you spent before she gives you the good?

Solution: Alice can verify that you have computed the new hash correctly. Furthermore, Alice can use your public key to verify that you were indeed the last owner of the coins by checking that you have generated a correct signature of the hash with respect to the public key contained in the previous transaction that provided an input to the hash.

2. iOwe versus Bitcoin

(a) Discuss the pros and cons of iOwe.
(b) Discuss the pros and cons of Bitcoin.

Solution:

(a) Iotas are easily created. This is good if there is a trustworthy user who wants to enter a system, needs some work done, and fully intends to redeem the iotas that he has issued at a later date. However, one drawback to iOwe is that when this later date arrives, this user might not make good on his iotas. As a result, only agents who trust this user (perhaps from
real-life) will accept his iotas as payment. This user is new to the system so if there is no one who trusts him (and since he does not have any iotas), he might not succeed in getting the work done that he wants. Iotas are easily transferable, which is good, but a drawback to this is you should only accept iotas where all the agents on the chain are trusted by you. Otherwise you risk the iota not being redeemable or being double spent.

(b) Bitcoin is a fiat currency that maintains its value as long as others believe it has value. Its positives are that bitcoin is a relative democratic and decentralized coin whose values cannot be tanked by an incompetent central bank. Double-spending is very difficult because the counterparty can wait for the transaction to be confirmed and written on the block chain. A concern with Bitcoin is that all transactions are publicized to be tracked in the block chain (transactions are pseudonymous), there is fixed money supply so it’s deflationary, and there’s no central ‘arbitrator’, so if someone steals your money, or doesn’t deliver on the goods that you buy, there is no insurance that covers you.

2 Privacy by Design

2.1 Key Concepts

- Public data release - when is this necessary/useful? How has the presence of the Internet changed this?
- Definitions of data user (someone interested in performing queries on a database), query (arbitrary function on the data), counting query (the count of the number of records that satisfy a property).
- Data curators (trusted intermediary between a data user and the data), and interactive (every query goes to the curator) vs. non-interactive approaches (curator publishes a modified database and database summary) to releasing data.
- Auxiliary information (information user already has) vs. direct identifiers (information that identifies an individual) vs. indirect identifiers (e.g. date of birth, ZIP code, income)
- Re-identification attacks (combining indirect identifies to identify individuals). Can you think of some examples?
- A database release satisfies \( k \)-anonymity if identifiers are removed and every quasi-identifier appears at least \( k \) times in the released data, if it appears at all. What are some examples?
- A randomized algorithm \( A \) is \( \epsilon \)-differentially private, for some privacy parameter \( \epsilon > 0 \), if, for all adjacent databases \( D \) and \( D' \), and all sets \( S \subseteq \mathcal{R} \), then

\[
\exp(-\epsilon) \leq \frac{\Pr(A(D) \in S)}{\Pr(A(D') \in S)} \leq \exp(\epsilon),
\]

where the randomization comes from the coin flips of algorithm \( A \).
- Properties of Differential Privacy: bounded effect on beliefs, bounded effect on utility, robustness to composition, robustness to post-processing
- For real-valued vector \( z \in \mathbb{R}^k \), let \( ||z||_1 = \sum_{j=1}^{k} |z_j| \). The sensitivity of a query \( q \) is a measure of the maximum effect that any single participant can have on the result of the query, the maximum taken over all databases.
- The sensitivity of numerical query \( q \) is

\[
\Delta q = \max_{D, D' \text{ s.t. } D \sim D'} ||q(D) - q(D')||_1.
\]

3
Given a multi-valued numerical query \( q : \mathcal{X}^n \rightarrow \mathbb{R}^k \), with \( k \geq 1 \) and sensitivity \( \Delta q \), the effect of Laplace perturbation is algorithm

\[
A_L(D, q, \epsilon) = q(D) + (Z_1, \ldots, Z_k),
\]

where \( \epsilon > 0 \) is the privacy parameter, and \( Z_j \sim \text{Lap}(\frac{\Delta q}{\epsilon}) \) is independently sampled for each \( j \in \{1, \ldots, k\} \). Why is this useful?

### 2.2 Exercises

1. What is the sensitivity of the following queries? Some of them may be best structured as a sequence of two or more queries. State any assumptions.

   (a) A count of the number of people with different medical conditions, when each person can have multiple conditions.

   **Solution:** The number of different medical conditions, because the count of each can change by one.

   (b) A count of the number of people with different medical conditions, when each person can have at most one condition.

   **Solution:** One, because the total effect over the count vector can be to change one entry by at most one.

   (c) The fraction of people in the islands database with cancer (assume the number of people in the database is not known to the analyst).

   **Solution:** For this we would need to do a query for the number of people and then a query for the number of people with cancer. The sensitivity is 1 for each and thus the total sensitivity is 2.

   (d) A count of the number of unique patients admitted to a hospital in Boston on January 1, 2013.

   **Solution:** One.

   (e) A count of the number of patients receiving emergency room treatment in each of a hospital in Copenhagen and a hospital in Sydney on each Monday in January.

   **Solution:** The number of Mondays in January (either four or five), because someone can’t be in Copenhagen and Sydney at the same time and so can only affect one of these counts each time.

   (f) A count of the total number of patients treated in the emergency room during January in a hospital in Boston.

   **Solution:** Depends on the maximum number of times any one patient can be treated in one day. If this is 3 then this is 31 * 3.
(g) A count of the total number of unique people who were re-admitted to at least one hospital in Boston during 2012.

**Solution:** One (since each person can only be re-admitted once)

2. $k$-anonymity and Marginal Table Releases

(a) (Composition attack) Suppose that two $4$-anonymous releases are made of the Alpha and Beta data:

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Smoke</th>
<th>#Idols</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amelia</td>
<td>24</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Ava</td>
<td>24</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Adam</td>
<td>24</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Aubrey</td>
<td>33</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Austin</td>
<td>42</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Ant</td>
<td>53</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Andrew</td>
<td>53</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Anna</td>
<td>77</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Release one: age and name is suppressed, but location of each person is revealed

Release two: location and name is suppressed, cancer is suppressed, and age is generalized (three intervals $<30$, $30-49$ and $\geq 50$).

Can a data user who knows Ben’s age, that Ben has zero idols, lives on Beta, and is in the data release learn his cancer status?

**Solution:** YES! In the first release, the data user learns that Ben is one of these two records:

<table>
<thead>
<tr>
<th>Island</th>
<th>Age</th>
<th>Smoke</th>
<th>#Idols</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>*</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>*</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

In the second release, the data user learns that Ben is one of these four records, and in particular the fourth one since Ben has 0 idols:

<table>
<thead>
<tr>
<th>Island</th>
<th>Age</th>
<th>Smoke</th>
<th>#Idols</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>30-49</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>*</td>
<td>30-49</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>*</td>
<td>30-49</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>*</td>
<td>30-49</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

The data user now knows that Ben smokes, because only one of these users has zero idols, and therefore concludes that Ben has cancer.

(b) In 2008, the California Department of Education released marginal tables providing test results in Mathematics (M) and English and the Arts (ELA) for Grade 11 students in a particular high school. See Figure 29.14 (Figure 1 in these notes). Entries for cells with less than 10 people were redacted (‘*’ in the table). What went wrong?

**Solution:** We learn that no women passed M, no whites passed M, no whites passed ELA.
Figure 1: Results of Mathematics (M) and English and the Arts (ELA) test results for students in a California High School by the California Department of Education. Entries ‘*’ were hidden.

3. Private $k$-means clustering.

A data analyst wants to run $k$-means clustering while ensuring $\epsilon$-differential privacy. The data records are vectors $x_1, \ldots, x_n$ in $[0, 1]^m$. The output is a set of $k$ cluster centers, each representing the mean vector of a set of associated records, where each record is associated with the closest cluster center (using Euclidean distance).

It is convenient to use a slightly different definition of the adjacency of two databases: we say that $D$ and $D'$ are adjacent if one differs in the values of a single record, rather than by the presence or absence of a record. The theory of differential privacy and Laplace perturbations is unchanged.

A $k$-means clustering algorithm chooses an initial set of centers at random, and repeats the following two steps until the cluster centers remain unchanged in step 2 or for some fixed number of iterations, $T$:

(Step 1) Associate each record with its closest center.

(Step 2) Update each cluster center to the mean of the associated records. Go to step 1.

(a) How is the composition property of differential privacy useful in designing an $\epsilon$-private algorithm for clustering?

**Solution:** It enables an algorithm to be evaluated by computing the eps privacy for each step and summing

(b) Consider a sequence of queries $q_1, q_2, \ldots, q_m$. Prove that the sensitivity of the combined query $q_m = (q_1(D), q_2(D), \ldots, q_m(D))$, satisfies $\Delta q_m \leq \sum_{\ell=1}^{m} \Delta q_\ell$.

**Solution:** Trivial, because this can be formulated as a stacked query

(c) An idea for a private way to compute an approximate version of step 1 is to use the Laplace perturbation on a query that returns, for every record, an integer $\{1, \ldots, k\}$ to indicate its closest center. What is the sensitivity of this query, and why might this approach not be very accurate?

**Solution:** Sensitivity is $k - 1$ because data in the record can change its cluster center from center 1 to center k. No.

(d) How can steps 1 and 2 be completed with a sequence of queries that has total sensitivity $1 + m$? [Hint: you will use two queries.]
**Solution:** Query 1: Count the number of records associated with each cluster center. Sensitivity 1.
Query 2: Sum the vectors associated with each cluster. This query has sensitivity $m$ because each vector is in $[0,1]^m$.
Taken together, a $m+1$-sensitive query is sufficient to obtain new cluster centers.

(e) Assume that the clustering algorithm will be used for $T$ iterations. How much Laplace noise should you use to obtain $\epsilon$-differential privacy for the entire algorithm?

**Solution:** Given that we want to run for $T$ iterations, then adding noise $(m+1)T/\ln(1/\alpha)$ is sufficient to obtain $\epsilon$-differential privacy overall.