HW2: Exploring Bitcoin transactions

March 23, 2020

1 Introduction

In this homework, we will conduct some data analysis on Bitcoin transactions. We will specifically draw on your knowledge of the blockchain structure and de-anonymization techniques. We will be using the Bitcoin dataset from `https://senseable2015-6.mit.edu/bitcoin/`, you can use any tools with which you're comfortable to work on the data (these are big files, so make sure that you allow enough time to download them, and enough storage space too).

Before getting started with this homework, please review the Bitcoin technical guide `https://bitcoin.org/en/blockchain-guide` and `https://bitcoin.org/en/transactions-guide` to familiarize yourself with Bitcoin transactions. In particular, you need to understand:

- Structure of the transaction: inputs, outputs, value. Pay attention on how a transaction relates to previous transactions.
- Bitcoin addresses
- Unspent transaction output (UTXO).
- Coinbase transaction.

Simply speaking, each Bitcoin transaction may have multiple inputs and outputs where each output is assigned with credits and locked to a Bitcoin address. This newly created transaction output is referred to as a UTXO. The UTXOs may then be used as inputs of another transaction, and after this transaction is committed to a block, those UTXOs will be marked as *spent* and cannot be used again (to prevent double-spending). Coinbase transactions are rewards for mining Bitcoin blocks, and they don't have any inputs. Each Bitcoin address is a string of 26-35 alphanumeric characters, beginning with the number 1, 3 or bc1.

Part 1: Transactions analysis

Provide your answer to the following questions:

1. What is the number of transactions and addresses in the dataset?
2. What is the Bitcoin address that is holding the greatest amount of bitcoins? How much is that exactly? Note that the address here must be a valid Bitcoin address string. To answer this, you need to calculate the balance of each address. The balance here is the total amount of bitcoins in the UTXOs of an address.
3. What is the average balance per address?
4. What is the average number of input and output transactions per address? What is the average number of transactions per address (including both inputs and outputs)? An output transaction of an address is the transaction that is originated from that address. Likewise, an input transaction of an address is the transaction that sends bitcoins to that address.
5. What is the transaction that has the greatest number of inputs? How many inputs exactly? Show the hash of that transaction. If there are multiple transactions that have the same greatest number of inputs, show all of them.
6. What is the average transaction value? Transaction value is the sum of all outputs' value.
7. How many coinbase transactions are there in the dataset?
8. What is the average number of transactions per block?

Note: the bitcoin value must be in **Satoshi**, not btc

Part 2: Address de-anonymization

In Bitcoin, a user may possess multiple addresses. In this part, we will apply a simple heuristic to infer the Bitcoin users owning those addresses. The heuristic consists of two phases:

1. Joint control: assume that all input addresses of a transaction are controlled by the same user.

2. Serial control: assume that the output address of a transaction with only a single output is controlled by the same user owning the input addresses.

Implement this heuristic on the dataset by applying both Joint control and Serial control, then answer the following questions (Note that in the dataset that we use, the file **addr_sccs.dat.gz** only applies the Joint control, that means it **CANNOT** be used for this part):

1. How many users are there in the dataset?
2. Answer questions 2, 3, and 4 in part 1 by replacing "address" with "user". Note that each user is identified by the addresses that are owned by him/her. Thus, in answering question 2 (i.e., the user who is holding the greatest amount of bitcoins), you need to list all the user's addresses.
3. Give the hash of the transaction sending the greatest number of bitcoins to the user who is holding the greatest balance.

2 Deliverable

You need to submit a zip file **FirstName_LastName_HW2.zip** that contains the following (please make sure that you name the file correctly):

1. The source code to reproduce your answers. Do NOT include the dataset in the submission.
2. Your report in markdown or pdf.

Your report must include the following material:

- Your name
- Your answer to each of the questions
- Instructions on how to run your code to obtain those answers.