Peer to Peer Systems and Blockchains
Academic Year 2021/2022

Mid Term
Analysis of Bitcoin Transactions - DHT Assessment
Deadline 06-05-2022

First Part

Consider the Bitcoin data set that has been published on the Moodle site of the course. The data set contains three CSV files, Transactions, Inputs and Outputs, that represent an abstraction of the Bitcoin transactions. The dataset starts from the genesis block and ends at height 100,001, which is the block mined on 29-12-2010. The data is slightly modified with respect to the original Bitcoin blockchain, and some transactions have been removed. Other than that, this data is an almost entirely *accurate Bitcoin dataset*, so the results that you will obtain would provide you some insights on a real cryptocurrency.

To keep the dataset small, the 256-bit hashes have been replaced with numeric ids. The Transactions file contains a sequence of rows, each row corresponds to a transaction and contains the unique id of the transaction and the *block_id* that transaction appeared in. Each transaction has one or more input(s) and output(s). The Inputs file has a row for each input appearing in any transaction and reports the unique identifier of the input, the transaction id that input appeares in, a *sig_id* denoting the public key used in the *scriptSig*, and the identifier of the output, *ouput_id*, which it is "spending". The file Output has a row for each output appearing in any transaction and reports the unique identifier of the output, the *transaction_id* that output appeared in, a *pk_id* denoting the public key used in the *scriptPubKey* and the value spent.To simplify the analysis, consider the public key as the address to which the bitcoin are sent (this was true in the time period of the data set, i.e. the first period of Bitcoin where the most frequent script was *PayToPublicKey*).

The format of the CSV file is reported in Fig. 1 end the meaning of the fields are reported in Table 1.

You are required to elaborate the dataset and develop the following points:

- describe how a real Bitcoin transaction is abstracted by a transaction in the dataset (which fields are eliminated, which are abstracted and how).
- check if all the data contained in the dataset is consistent, and if some data is invalid, describe what is the problem of that data and remove it from the dataset.
- compute the total amount of UTXOs (Unspent Transaction Outputs) existing as of the last block of the data set, i.e. the sum of all the

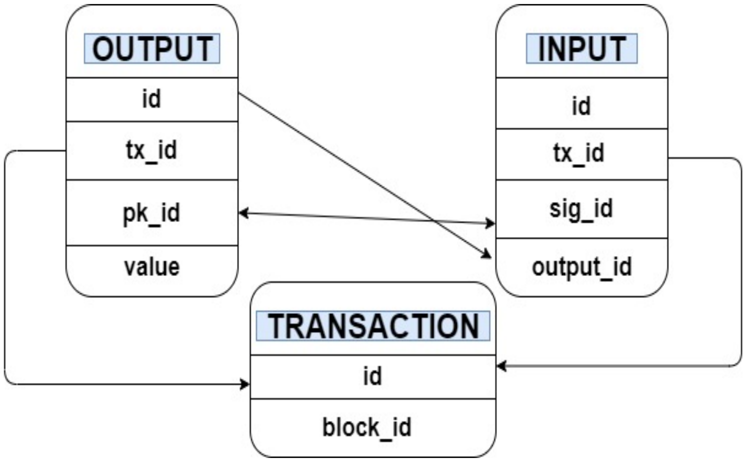

Figure 1: Format of CSV Transaction File

transactions	
id	identifier of the transaction
block_{id}	id of the block containing this transaction
inputs	
id	unique id of this input
tx_{id}	transaction this input is part of
sig_{id}	scriptSig public key id, 0 if coinbase tx, - 1 if nonstandard scripts used
$\text{output}_{\text{id}}$	id of the previous output being referenced, - 1 if coinbase tx
outputs	
id	unique id of this output
tx_{id}	transaction this output is part of
pk_{id}	scriptPubKey public key id, - 1 if nonstandard scripts used

Table 1: Meaning of the Data Fields

Transaction outputs balances on the UTXO set of the last block. Which UTXO (TxId, blockId, output index and address) has the highest associated value?

- compute the following statistics
  - the distribution of the block occupancy, i.e. of the number of transactions in each block in the entire period of time. Furthermore, show the evolution in time of the block size, by considering a time step of one month.
  - the total amount of bitcoin received from each public key that has received at least one COINBASE transaction, in the considered period, and show a distribution of the values
  - the distribution of the fees spent in each transaction in the considered period.
- propose one further analysis of your choice. Give a brief description of the analysis and report the results.

Hint: interesting properties can be studied by analysing the *address graph* that represents the flow of bitcoins between public keys (addresses) over time. The graph has a node for each different public key contained in the dataset and an edge represents a flow of bitcoin between the two public keys (addresses). Since each transaction in Bitcoin represents an $m : n$ relationship between addresses, the address graph corresponding to a transaction links every input address to every output address. Furthermore, it is possible to quantify the flow of bitcoin between an input I_i and an output O_j of a transaction having k input, using the following formula:

$$Est(I_i, O_j) = A(O_j) \times \frac{A(I_i)}{\sum_k A(I_k)}$$

where $A(I_i)$ is the amount paired with the public key corresponding to input I_i, $A(O_j)$ is the amount transmitted on output O_j ,

The analysis of the data can be conducted by using any programming language or data management tools that you desire. For instance you can use non relational databases like *Neo4j* or package for graph analysis like *NeworkX* or *Networkit*. Your solution will not be graded on efficiency, you can use, for instance, an interpreted language.

Second Part

Assume a Kademlia network with ID size of 8 bits. The bucket size is k = 4. The k-buckets of the peer with ID 11001010 are as follows:

k-Bucket 7: 01001111, 00110011, 01010101, 00000010
k-Bucket 6: 10110011, 10111000, 10001000
k-Bucket 5: 11101010, 11101110, 11100011, 11110000
k-Bucket 4: 11010011, 11010110
k-Bucket 3: 11000111
k-Bucket 2:
k-Bucket 1:
k-Bucket 0:

You are asked to answer the following questions:

- Messages from following nodes arrive in this given order: 01101001, 10111000, 11110001, 10101010, 11100011, 11111111 How do the buckets, the orderings in the buckets and the waiting lists change?
- Now the node detects that peer 11101110 cannot be reached anymore, what is the reaction?
- Which addresses would the peer reply to a lookup looking for ID 11010010?

Assignment Submission

The assignment must be done individually and its deadline is 05 May 2022. If the evaluation of both the mid and of the final term will be positive, the student will be relieved from the oral exam. Submit the assignment through Moodle. Its evaluation will be notified through the Moodle as well.

The assignment is not mandatory, if it is not presented, the student will be required to pass the oral exam on this part of the course.