P2P Systems and Blockchains
Final Project
Academic Year 2020/2021

Bitcoin Transactions:
Analysis and De-Anonimyzation

1 Goal of the Project

Goal of this project is to assess the student's understanding of the Bitcoin transactions and to investigate some techniques adopted for de-anonymizing Bitcoin addresses. You will be given a truncated data set of simplified Bitcoin transactions, starting from the genesis block and ending at height 100,001. The block reward was 5,000,000,000 Satoshis (50 BTC) during this period. 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 the student some insights on a real cryptocurrency.

2 Definition of the Analyses

Consider the data set of Bitcoin transactions published on the page of the course and:

- check if all the data contained in the dataset is consistent, and if some data is invalid, remove it.
- 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 Transaction aoutputs balances on the UTXO set of the last block. Which UTXO (TxId, blockId, output index and address) has the highest associated value?
- use the following two common heuristics to cluster the addresses in the data set :
  - *common input heuristic*: addresses used as inputs to a transaction are controlled by the same entity. The rationale behind this heuristics is that all the inputs of the transaction must be signed, so it is likely that they belong to the same user, who knows all the private keys corresponding to those inputs. Refer to [1] for more details on this heuristics.

- *serial control heuristic:* the output address of a transaction with a single input and a single output is usually controlled by the same entity owning the input address.

Define an algorithm to cluster the addresses of the dataset by applying these heuristics. When applying the common input heuristics, take into account transitivity, i.e. two cluster are merged if they contain at least two addresses which appear together as input in at least one transaction.

Each cluster resulting from the application of these heuristics, corresponds to an "entity" which controls all the addresses in that cluster.

You are required to elaborate the following points:

- as of the last block of the data set, find the entity controlling the most total (unspent) bitcoins. What is its lowest address (numerically) and what is the total value of all the bitcoins it controls?
- give the ID of the transaction sending the greatest amount of bitcoins to this entity. Just for fun (this part will not be graded), you can try to deanonymize some addresses corresponding to this entities, by scraping web sites/forum reporting real identities corresponding to Bitcoin's users.
- consider the clusterized transaction graph, i.e. the graph whose nodes correspond to clusters, i.e. entities, and whose edges are such that there is an edge between two clusters iff there is a transaction with an input address in the first cluster and an output address in the second cluster. Find the length of the longest payment path in this graph.
- are the proposed clustering methods accurate? List at least one potential source of false positives (clustering addresses which aren't actually owned by the same entity) and one source of false negatives (failing to cluster addresses which actually are owned by the same entity) in this method. What strategies could you use to make your clustering more accurate?

3 Implementation Details

Download the Bitcoin data set from the course website. The data set contains three CSV files: Transactions, Inputs and Outputs. 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), each given a unique id. 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 appeared in, a sig_id denoting the public key used in the scriptSig, and the

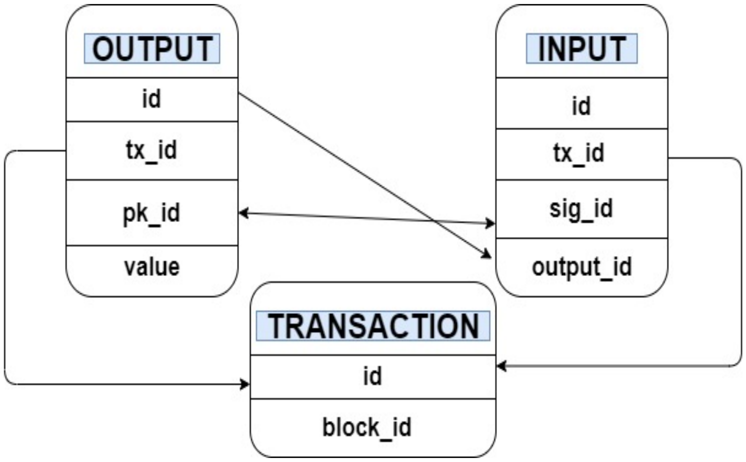

Figure 1: Format of CSV Files

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. The schema of the data is reported in Fig. 1

Note that with real Bitcoin data, the ids would of course be 256bit hashes. To keep the dataset small, they have been replaced with numeric ids.

The analysis of the data can be conducted by using any programming language or data management tools that you desire. Your solution will not be graded on efficiency, you can use, for instance, an interpreted language.

4 Project Submission Rules

The project must be developed individually. The material to be submitted for the evaluation is the following one:

- a report (pdf document) describing the main features of the project. The report should include: a brief summary of the implementation choices and an high level description of the code and the results of the experiments. The results of the experiments must be adequately commented.
- the source code implemented.

The report and the code must be submitted electronically through Moodle. The project will be discussed a week after its submission. The discussion of the project consists in the presentation of a short demo, which can be run on the personal laptop, and a general discussion of the choices made in the implementation of the system and of the reported results.

The oral examination (if required) will regard a review of the topics presented in the course. I remind that the oral examination is waived for the students who have passed the Mid and Final Term.

For any questions or clarifications, do not hesitate to contact us (laura.ricci@unipi.it, andrealisi.12lj@gmail.com) by e-mail, we will fix a meeting in the Teams room of the course.

References

[1] Meiklejohn, S., Pomarole, M., Jordan, G., Levchenko, K., McCoy, D., Voelker, G. M., Savage, S. *A fistful of bitcoins: characterizing payments among men with no names.* Proceedings of the 2013 conference on Internet measurement conference, pp. 127-140.