Exploratory Analysis of Bitcoin Data

Shaun Giudici

Abstract—The process of analyzing a set of data of which one is not previously familiar is an open-ended challenge that presents many potential options to the explorer. We present a method that emphasizes the explorer's self-education in the domain via original story creation. The creation of a metaphorical story whose elements mimic real-life scenarios relevant to the data can helps in two ways: 1) as a method to aid the explorer's understanding of different aspects of their dataset, 2) to improve the communication and following discussions of the data analysis.

To help develop this process, we conducted an exploratory data analysis on a subset of bitcoin activity. We followed a basic high-level process for the analysis: 1. Getting to know the domain, 2. Creating an overview of the data, and 3. Diving deep on select questions formed during the first two steps. We hope that revisiting elements of the story throughout the investigation will reaffirm the strength of our metaphor, and to aid in the discussion and creation of new and interesting avenues to explore.

Index Terms—exploratory data analysis, storytelling, bitcoin, visualization

Introduction

The bitcoin peer-to-peer cryptocurrency network is an open system for maintaining and transacting stores of value between anonymous owners. Many cryptographers and computer scientists have found the bitcoin protocol to be very secure. Still in its early days, the system is technically complex, gaining adoption among users in the technological elite.

In the context of this paper we are interested in the data generated by the open bitcoin network, as there is a lot of it. Among the many potential sources of data (network node traffic, live p2p transaction pool, confirmed transaction record, exchange rates, overall external discussion and sentiment) we will focus our analysis on the confirmed transaction record, thenceforward referred to as "the blockchain".

Exposition

We perform an exploratory data analysis on the bitcoin blockchain, which we break down into three major steps: getting to know the domain, building an overview of the dataset, diving deep on select 'interesting' anomalies and data-driven questions formed during steps 1 and 2. Here is a brief overview of the investigation, followed by an explanation of each major step.

Transactions & fees: A search for any strong correlations to transaction fees. We think it would be interesting to be able to predict fees reasonable fees, potentially recommending optimal fees to choose based on current activity

Inputs & outputs: The ability to combine and redistributed coin value between so many different locations is unique to bitcoin transactions. We are curious about usage of this bitcoin-specific feature as a potential predictor for how bitcoin is changing the traditional payments system. We seek to find out the typical distributions of inputs and outputs in a transaction, and analyze trends of this activity.

Double Spends: Bitcoin is famous for being a system that answers this important problem of digital currency. We investigate double spend activity. This requires monitoring the network in real time as failed attempts are not stored on the blockchain.

Exploring connections: Another nuance of the bitcoin system is that all occurrences of coin have an origin that can be seen by all. Therefore we can collect and analyze the connections between transactions.

Getting to know bitcoin

For our bitcoin case study we drew a storyboard of what happens behind the scenes of a bitcoin transaction. The storyboard is progressively disclosed one scene at a time (**figure 1**) and leverages the visualization techniques of zooming and filtering [1] to introduce new concepts. For some scenes we used low-fidelity animation of small cutouts to visually relate subjects of previous scenes. As we move on to the data analysis, we will revisit scenes from our storyboard to reinforce the metaphorical relationships. In Chart 1 below, you can review a chart that links elements of our storyboard to real-world bitcoin vocabulary.

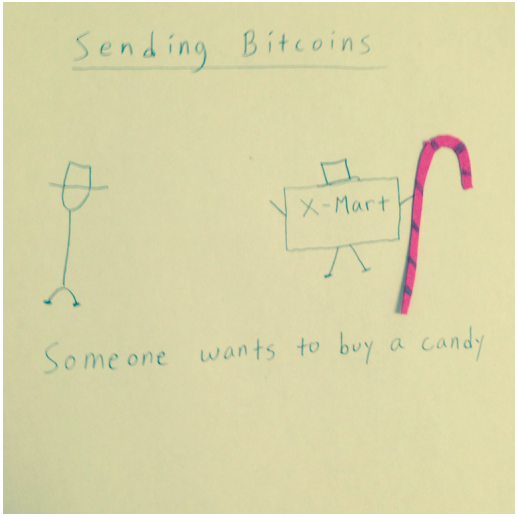

Figure 1: Explaining how a bitcoin transaction works with a storyboard

• *Shaun Giudici, Masters candidate at UC Berkeley School of Information. E-mail: shaun@ischool.berkeley.edu.*

Storyboard	Bitcoin
postcard	transaction
From: field	inputs
To: field	bitcoin address / outputs
postage stamp (optional)	fee collected by miner
drop in mailbox	post transaction to network
dropping two instructions that spend the same coins	double spend attempt
workers	miners
anyone can be a worker	decentralized model
signature verification	public-key cryptography
race to submit verified postcards	block creation
payment for verifying postcards	block subsidy - new coins enter supply
etched glass wall	the blockchain
robot laser etching	new block posted to block chain
people viewing etched wall	all activity is transparent

Chart 1: Terminology conversion from our story to bitcoin technical jargon

One week of bitcoin activity

For our bitcoin case study we began with overall counts. **Figure 2** shows the overall bitcoin supply versus the amount currently in circulation. We quickly realized a unique aspect of the bitcoin blockchain dataset: it holds a massive amount of data (millions of transactions and addresses). Scoping this down was necessary in order to conduct an efficient analysis. Therefore for the remainder of the overview we used a week's worth of bitcoin data, which is a much more manageable chunk to deal with.

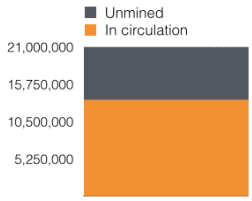

Figure 2: Overview of entire coin supply

Though we completed exercises for getting to know the domain better, the overview is still an important step to improve understanding of the dataset. Here is an example scenario that we encountered while doing an overview for the bitcoin case study:

We plotted out the occurrence of double spends and found 5 over the course of a week (**figure 3**). But wait a minute, I thought our dataset was the bitcoin blockchain (the etched glass wall), and that the nature of bitcoin makes double spending impossible. Have we just uncovered that bitcoin is insecure? On further review of the API documentation we see that the data service provider, BlockCypher, is monitoring the network and reporting "double spend attempts" that didn't make it into the blockchain. We had mistaken the meaning of the true/false values for the column "double spend".

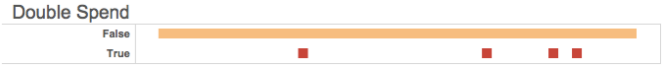

Figure 3: Double spend attempts for one week of bitcoin data

Not all of these aspects of the dataset will be obvious or understood by the investigator at first. This is an example of how an overview exercise can improve the investigator's understanding of the dimensions in play.

A closer look

In our bitcoin case study we expected to see a correlation between total transaction value and total fee collected. This turned out to be false, which led us to conduct a deep-dive to find out why. First we revisited the definition of a transaction (**figure 4**). Then we took a closer look at the specific definitions of everything involved in the comparison and we found that changing transaction value is as cheap and simple as modifying an integer. Therefore it makes sense that a total transaction value wouldn't warrant higher fees.

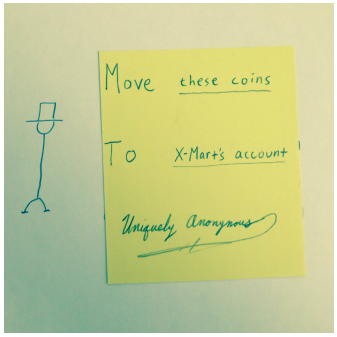

Figure 4: Simple explanation of a bitcoin transaction, relating it to a postcard

However, we took this one step further, considering whether the size (in bytes) might play a role, and indeed it did. Though we didn't have the exact amounts in bytes per transaction available, it is sufficient for the purpose of an overview to estimate by using the number of inputs. More inputs (and theoretically larger transaction sizes) are more complex; we found that complexity of a transaction does have some effect on transaction fee amount (**figure 5**).

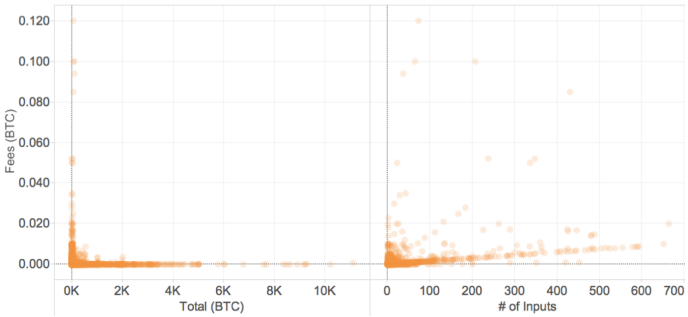

Figure 5: Transaction fees mapped against Value Sent, and then our estimate of Transaction Complexity

Lastly we explored the linking of transactions, and the ability to "follow" money through the blockchain. With one important exception known as the coinbase or "miner reward", Bitcoin transactions reference previously created coins. Thus they are linked backwards. To explore this for the week in question, we created a JavaScript tool that allows user to interactively choose transactions to inspect for connections. Upon selection, all neighbors of selected transaction are highlighted among the rest, making it clear which transactions are linked among the current set. By merging this link inspection ability with a scatter of all our transactions over the week,

we can get a much clearer idea of which transaction patterns are connected, and how the coins splay out. (**see figure 6**)

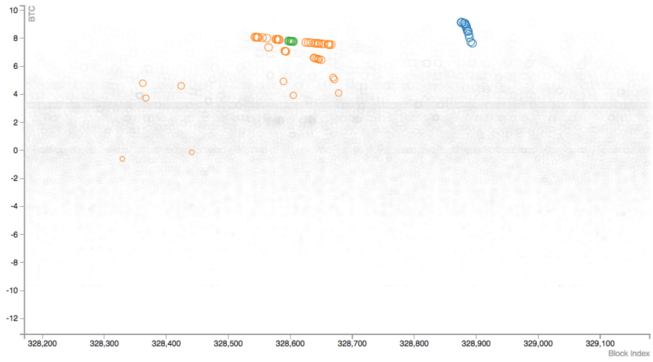

Figure 6: Visualizing links among transactions using D3

Technical Implementation

To compile our dataset we utilized the public API provided by BlockCypher and our own custom python scripts. We then converted this data from JSON to Excel format to support import to tableau. For most plotting and analysis we utilized Tableau, as it is well suited for creating visualizations and managing multidimensional datasets. The one exception to this was our investigation of transaction linking, since tableau does not support node-edge graphs. For this we rebuilt our views in custom D3/javascript.

Discussion

Know The Domain

We argue that a strong exploratory analysis should begin with exercises aimed at improving the investigator's understanding of the domain being explored. We suggest that at this stage, the creation of an original story as a metaphor to the real-world domain can improve understanding of the new space. Similar to how a navigator uses anchors to when setting and checking their bearings, these qualitative additions will improve the investigator's ability to formulate questions and conduct analyses that are logical within the context. It will also empower the investigator to form their data and findings into a context that a greater audience will understand.
Tasks that we find helpful at this stage are:

- Define each dimension (column) in the dataset. This will be especially important later when defining new relationships via comparisons of different dimensions against one another. For our case study on bitcoin, we referenced the official documentation on bitcoin.org to retrieve the necessary definitions. We also reviewed blog posts on related material to see how others were describing the same concepts.
- Find some typical scenarios that occur within the context of the domain that demonstrate how the dataset is influenced. Then translate these scenarios into a story using layman's terms. The use of accompanying images and text can improve how the audience relates to the story. Return to moments within the story throughout the data analysis as a reminder to yourself, and the audience, about the elements of the new territory that we are zoomed in on.
- Create a terminology conversion table, like a glossary between your story and the real-world concepts you are portraying. We found this helpful to be sure that all aspects of our real-world scenario were accounted for in our story.

Build an Overview

Starting the analysis with an overview is an important step to understand all that is contained within a dataset. [1] [2] It helps audience and analyst alike to get a sense of scale, such as checking depth and purity, before diving in headfirst. It is here in the overview that we set our bearings to the anchors defined in the previous step.

- High level goals for this phase include:
- Understanding the magnitude and reach of our dataset. We find it helpful to start with basic counts and distributions to get a sense of size for each independent dimension.
- Question forming and annotation. Throughout our experimentation we keep a keen eye out for surprising or otherwise interesting points or comparisons, and mark them for further analysis. Any questions that we form at this stage should be annotated for later reference.
- Delay deep dives. It will be tempting when observing interesting situation arising in the data, and want to follow it deeper. Without being too prescriptive, we want to warn against going in for deep dives on a particular question or area too soon. The danger here is that additional, potentially helpful data could be yet to be uncovered via the overview that will aid this deep dive.
- Continue through the troves of new information until most all combinations have been exhausted. Then review your results and consider what may be missing or incorrect, on a high level.

Deep Dive on Specific Questions

Following the overview the investigator will ideally be left with a list of interesting questions and annotations on the data for further exploration. Now is the time to focus energy on learning the 'why' for each of these cases, specifying our expected activity and then proving or disproving it with data. We recommend keeping the original story in mind when conducting this activity as it reinforces the frame of reference and improves the questions that one may be asking.

Conclusion

In an open-ended analysis where there are many options and avenues for the investigator to explore, we have found a method that helps to keep a common thread throughout. Multiple iterations of creating a story that accurately portrays a metaphor for a bitcoin transaction involve a deep learning of the details by the investigator. We discovered new meanings and details in the process, perhaps more than our traditional research would have provided.

We presented our exploratory analysis of bitcoin data to an audience of 8, many of who were unfamiliar with technical bitcoin terminology. We received positive feedback from the audience that our method of telling a story, and revisiting that story throughout the presentation of data helped the overall understanding of the new topics. We received great feedback, suggestions and additional questions about the technical aspects of the analysis. Without a control group its unclear whether we succeeded at improving communication of our analysis.

Acknowledgments

Thank you to Marti and Sara for a great seminar, and to the rest of the class for their great feedback as we progressed on this project throughout the semester.

References

[1] B. Schneiderman. The eyes have it: A task by data type taxonomy for information visualizations

[2] S. Few. Exploratory Vistas