Cointopia: Blockchain Analysis using Online Forums

George He
Stanford University
georgehe@stanford.edu

Haithem Turki
Stanford University
hturki@stanford.edu

Tony Hu
Stanford University
tjhu@stanford.edu

December 10, 2017

Abstract

In this paper we detail the results of exploratory data analysis against the Bitcoin blockchain, and compare various communities identified through mention and usage in online forums and datasets. By comparing these disparate communities, we manage to identify trends in convergence and adoption, cluster different addresses into groups with similar characteristics, and uniquely identify individuals that may be engaged in suspicious activity and merit further investigation.

1. Introduction

Interest and participation in the blockchain ecosystem has surged in recent years, spearheaded in particular by the explosion of Bitcoin into the mainstream collective consciousness. As the original and oldest blockchain community, Bitcoin offers a unique view into the history and evolution of the blockchain phenomenon. Having been once limited to a core community of developers and hardcore enthusiasts, Bitcoin is now traded by individuals from all walks of life, with the estimated USD transaction volume having increased almost twenty-fold in the past year alone. We examine how that usage has evolved over time, both within the original developer/enthusiast ecosystem and the increasingly heterogeneous overall community present, and investigate whether they diverge in significant aspects.

To this effect, our data collection efforts focus on scraping data (namely public key hashes) from forums frequented by the former group, and some other general interest Bitcoin-related forums that serve as a baseline proxy for the network at large. We further augment this data with a list of tagged addresses from Blockchain.info and reputation data gleaned from bitcoin-otc's web of trust network in order to identify an initial set of nodes that may be of particular interest. We then apply various address linking/clustering techniques to build out the corresponding Bitcoin transaction graph and compare them using theoretical frameworks such as power law distribution of node degrees and small world phenomenon.

2. Background and Related Work

The inherently public nature of the Bitcoin ledger allows for unique insights into previously inscrutable transaction flows, and has attracted a significant amount of interest from researchers across the academic spectrum. A particularly active subtopic of research has centered around address linking and resolving disparate addresses to the same entity. Two of the most widely used heuristics involve are the multi-input transaction heuristic proposed by Androulaki et al in Evaluating User Privacy In Bitcoin [6] and the one-time change address detector suggested by Meiklejohn et al in The Fistful of Bitcoins [13]. Further work by Nick in Data-Driven De-Anonymization in Bitcoin [10] expands the surface area of these heuristics in promising new directions.

Lischke and Fabian [9] explore the structure of the Bitcoin user network, examining metrics such as average clustering coefficient, average shortest path, whether the degree distribution of nodes follows a power law, whether the Bitcoin network exhibits the small world phenomenon, and various measures of centrality (degree, betweenness, and closeness). The paper also ties in IP and business tag data to examine the geographic and industry/sector characteristics of the Bitcoin user network. Some main results include that certain industries and even specific entities, like gambling via SatoshiDICE, account for a large portion of transaction volume; the variation in business distribution by geography, with transactions dominated by gambling in some countries and mining in other countries; small world phenomenon was evidenced by high average clustering coefficient and low average shortest path among examined subgraphs; and the Bitcoin graph showed a power law distribution of degrees indicating a scale-free network.

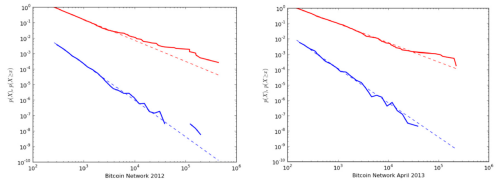


Figure 21. Degree Distribution (probability density function (PDF), complementary cumulative distribution function (CCDF)) over Time.

Figure 1.

It is however worth noting that many of the most commonly cited research publications originate from almost five years ago (2012 and 2013 for Androulaki and Meiklejohn's respective contributions). This equates to an eternity in the Bitcoin world, with the number of confirmed transactions having increased by more than a factor of ten since mid-2012 (and the estimated USD transaction value having increased by more a thousand times over). The overall

size of blockchain has also grown exponentially since (over 140 GB as of August 2017). Most of the existing human-parseable data available on the web is either rate-limited via API calls, with publicly accessible bulk downloadable csv files only going up to 2014 at the latest. We believe that there is significant value in directly parsing an up-to-date blockchain obtained from the client itself that justifies the effort we spend into processing the raw files.

Kalodner et al have recently released a modern analysis platform detailed in BlockSci: Design and applications of a blockchain analysis platform [7], and cluster using their relatively recent implementation of the multi-input and change address heuristics that also supports more recent developments such as the CoinJoin input mixing technique proposed by Gregory Maxwell. We make extensive of their tooling as part of our data exploration.

3. Algorithms

3.1. Graph Clustering

The following clustering algorithms of particular interest to us and are explored as part of our analysis:

- Multi-input transactions: In Evaluating User Privacy In Bitcoin [6], Androulaki et al. propose that if two or more addresses are inputs to the same transaction, they are controlled by the same user. This technique is used to group clusters of individual Bitcoin addresses into manageable groups for computation.

- One-time change address: The Fistful of Bitcoins paper [13] by Meiklejohn et al. suggests that if a transaction pays out to a one-time change address (ie: that address only ever receives one input), that address is likely controls that the initiator of the transaction.

3.2. PageRank

In the original PageRank whitepaper [11], network links between different websites represent edges in a network, whereas websites represent nodes. We adopt a similar approach tailored for the Bitcoin transaction community, where clusters of individuals identified through the Multi-input transaction and One-time change address technique represent individual nodes, and transactions between clusters are considered unweighted directed edges.

3.3. Paranoid Distrust Propagation (PDP)

We propose algorithm 1 to identify all nodes that can be flagged as potentially "untrustworthy" in the network, provided a ground base truth for individuals that are not trustworthy. This approach follows the outlinks from untrustworthy nodes, and considers the number of incoming edges from untrustworthy nodes. Additionally, this algorithm considers the proximity of nodes with respect to the nodes identified in the ground truth.

Given the set of untrustworthy individuals U_i, untrustworthy clusters U_c are identified as follows:

$$U_c = \{\forall U_i : Cluster[U_i]\}$$

Algorithm 1 Paranoid Distrust Propagation

procedure PDP($U_c, threshold$) ▷ Paranoid Distrust Propagation
  U_c
  $Distrust \leftarrow Counter([(Cluster, 1) for (Cluster \in U_c)])$
  $Length \leftarrow len(Distrust)$
  while $Length = len(Distrust)$ **do**
    $NewDistrust \leftarrow Counter()$
    for $node \in Distrust$ **do**
      $NewDistrust[node] \leftarrow 1$
    $Distrust.add(NewDistrust)$
  return $Distrust.filter(key, value \rightarrow value > threshold)$

3.4. Estimating Average Shortest Path Length

Computing exact average shortest path on such large graphs, up to 635,575 nodes in the case of the largest weakly-connected component of the general Bitcoin community 2017 snapshot, was very computationally intensive and ultimately infeasible to scope within the resources available to our project. However, since we are interested in an average, we are able to apply statistical tools to produce a good estimate.

The average shortest path for a graph is defined as

$$L = \frac{1}{n(n-1)} \sum_{i \neq j} d(v_i, v_j)$$

where n is the number of nodes and $d(v_i, v_j)$ is the shortest path length between the nodes v_i and v_j. Implementation-wise, the SNAP [8] code library provides a convenient method for computing $\sum_{i \neq j} d(v_i, v_j)$ by computing the tree of shortest path lengths from a particular source node s. But since

$$L = \frac{1}{n(n-1)} \sum_{s \neq j} d(s, v_j)$$

is very computationally intensive, we instead compute an estimate

$$\hat{L} = \frac{1}{|S|(n-1)} \sum_{s \neq j} d(s, v_j)$$

where sample node $s \in S$.

We aim to select a sample size that is more computationally tractable while still providing an accurate estimate. The minimum sample size $|S|$ required for an estimate with a margin of error e at confidence level α is

$$|S| = \frac{z_\alpha^2 \tilde{\sigma}^2}{e^2}$$

where z_α is the z-score associated with confidence level α and $\tilde{\sigma}$ is a prior estimate of population standard deviation for the quantity of interest (average shortest path length for a single source node s).

For a prior estimate of population standard deviation, we compute population standard deviation from the two time snapshots of data that are feasible for an exact computation: 2010 for the general bitcoin community and for the developer community. The results are $\sigma_{general,2010} = 0.856456053321$ and $\sigma_{dev,2010} = 0.283577591248$, respectively; for the overall prior estimate we use the average $\tilde{\sigma} = \frac{\sigma_{general,2010}+\sigma_{dev,2010}}{2} = 0.5700168223$.

Z-scores can be obtained from standard tables of values; for $\alpha = 99\%$, $z_\alpha = 2.58$. Z-scores represent the number of standard deviations from the mean a data point is for a reference distribution, in this case the normal distribution. Without making any assumptions about the distribution of the average shortest path length for one source node s, we are able to refer to the normal distribution since applying the Central Limit Theorem shows that for random and independent samples of $|S|$ observations each, the distribution of sample means, in this case the estimator $\hat{L}$, approaches normality as $|S|$ increases, regardless of the shape of the population distribution. The general rule of thumb that is often employed is that the normality of the distribution of sample means arrives around $|S| = 30$; we will see in this case that the computed minimum sample size is far larger than 30 so z-scores can safely be used.

Finally, we decide on a margin of error e and confidence level α. For margin of error, we return to the preliminary analysis done on the 2010 snapshots. We computed the range of values r for the average shortest path for a source node s; the results are $r_{general,2010} = 7.515021459 - 2.090128755 = 5.424892704$ and $r_{dev,2010} = 3.567567568 - 1.310810811 = 2.256756757$. For a more conservative estimate we use the smaller range $r_{dev,2010}$, and take 2% of that range as the margin error, which yields $e = 0.045$. For confidence level, 95% is a standard level to use but we decide to use $\alpha = 99\%$ for a better estimate.

Thus we define a good estimate to have a margin of error $e = 0.045$ at confidence level $\alpha = 99\%$, resulting in the sample size

$$|S| = \frac{z_\alpha^2 \tilde{\sigma}^2}{e^2} = \frac{2.58^2(0.5700168223^2)}{0.045^2} \approx 1068$$

which we use to compute the estimate $\hat{L}$ for average shortest path length for all time snapshots, except for 2010 where the graph was small enough for all nodes to be used in computing average shortest path length.

4. Mathematical Background

We examine the nature of the two communities by analyzing the degree distribution of the nodes in the context of scale-free networks, and by determining the average shortest path length and clustering coefficient which may indicate small world phenomenon.

4.1. Power Law

Empirical studies of real world networks show degree distributions to often follow a heavy-tailed distribution such as a power law. A random variable x is considered heavy-tailed if

$$\forall \lambda > 0, \lim_{x\to\infty} \left[\frac{T(x)}{e^{-\lambda x}}\right]$$

for tail distribution $T(x) = P(X \geq x)$. A common example of a heavy-tailed distribution is a power law distribution, which has the form

$$p(x) = \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-\alpha}$$

Power law distributions are characterized by the exponent α. One approach to estimating the value of α is to fit a linear regression equation to a log-log plot of the degree distribution, with an appropriate amount of pruning of outliers. While this may be a reasonable method for a quick preliminary analysis, this method suffers from ambiguity in defining which outliers should be excluded.

A more robust method of estimation is using maximum likelihood. In this approach, the log-likelihood of observing data value x_i is

$$L(\alpha) = \ln\left(\Pi_i^n p(x_i)\right)$$

In order to find the α that maximizes this likelihood equation, we take the derivative $\frac{\partial L(\alpha)}{\partial \alpha}$ and set it to zero to obtain the α estimator

$$\hat{\alpha} = 1 + n\left[\Sigma_i^n \ln\left(\frac{x_i}{x_{min}}\right)\right]^{-1}$$

from which α can be estimated from the data.

A power law distribution, or a scale-free network, would imply various properties such as resilience to node failure and susceptibility to contagion. Furthermore, the value of α being within certain ranges is associated with particular properties of the network. For example, the average node degree $\bar{k}$ is finite if $\alpha > 2$; or, the network behaves like a random network if $\alpha > 3$.

4.2. Small World

The small world phenomenon refers to an empirical observation that in some real world networks, any node is able to reach any other node in surprisingly few steps. More precisely, a small world network is one in which the average shortest path L grows proportionally to the logarithm of the number of nodes N in the network, i.e. $L \propto \log N$. Watts and Strogatz [5] found that such networks

are characterized by short average shortest path length and a clustering coefficient higher than would be expected by random chance. For example, the clustering coefficient of an Erdos-Renyi random graph with the same number of nodes and edges could be the "by random chance" baseline for comparison.

5. Data Collection Process

5.1. Reddit [General Community]

Reddit (http://reddit.com/Bitcoin), a social news aggregation, web content rating, and discussion website, is used widely for open discussion where users who have created accounts can post questions and answers to communities, known as subreddits. Reddit exposes an API in python called PRAW (https://praw.readthedocs.io/en/latest/) to query and interact with the website. The API allows for eventual traversal of all forum posts. By focusing on the Bitcoin subreddit, we were able to query over 250,000 posts over a 150 hour collection period, and retrieve over 1100 uniquely identified bitcoin addresses.

5.2. Bitcoin Talk [General Community]

Bitcoin Talk is a forum dedicated to Bitcoin and organized into subforums dedicated to specific topics. Scrapy (https://scrapy.org/), a publicly available webcrawler, was used to traverse and index topics where individuals were able to connect to. After crawling through over 4 million posts, and 500 thousand topics over a 50 hour collection period, approximately 1500 unique Bitcoin addresses were recovered.

5.3. Stackexchange [Developer Community]

Stackexchange, an online forum geared towards developers, hosts a Bitcoin-themed subdomain at https://bitcoin.stackexchange.com/. Using their provided API, over 16,500 questions and 200,000 questions and comments were queried. In total, over 2,000 unique Bitcoin public key hashes were collected from the website. Efforts to collect addresses from the entire Stackexchange community warrant further investigation, as a crawl of 6 million posts from the site resulted in approximately 70 uniquely identifiable addresses being surfaced.

5.4. Bitcoin blockchain [Core]

Bitcoin is available as a public ledger, and the entire blockchain is downloadable using the official Bitcoin client. After downloading the blockchain, we primarily used the BlockSci library [https://github.com/citp/BlockSci] to analyze the blockchain from its inception in 2009 up until August 1st, 2017. We further segmented the transactions by timestamp in order to analyze snapshots of the blockchain at certain time intervals.

5.5. Bitcoin OTC [Auxiliary]

Bitcoin OTC (https://www.bitcoin-otc.com/) is an over-the-counter marketplace for trading with Bitcoin that offers a web of trust service to mitigate the inherent counterparty risk. We collected approximately 1,500 addresses and their associated trustworthiness scores. By matching these scores to the clusters their associated addresses belong to, we identify a seed list of "suspicious" clusters that we can then cross-reference with the other datasets. This is used as a group truth to identify individuals who may be untrustworthy.

5.6. Blockchain.info [Auxilary]

Blockchain.info stores a set of user-specified tags at http://blockchain.info/tags that can be created or updated by anyone. At the time of writing they list over 37,000 tagged addresses. By cross-checking tags with the set of collected individuals and their associated clusters, we are able to further identify clusters of interest during our analysis

6. Results

6.1. Data Summarization

The addresses scraped from the external social data sources are validated against the Bitcoin blockchain to determine the addresses valid as of August 1st, 2017. This process is able to identify 4,104 unique addresses. Of these addresses, 109 of them can be trivially located on the list of tagged addresses collected from blockchain.info (i.e.: their hash is present in the list of tags).

The entire set of addresses (387,936,389 entries) in the blockchain are clustering using the heuristics and implementation currently present in BlockSci (the multi-input transactions and one-time change address transaction). BlockSci defines a cluster as any group where two individuals are the same input in a transaction. This ultimately resulted in 144,975,330 different clusters.

We then map our scraped addresses to those clusters. By reconciling the addresses within these clusters to the same tag, we manage to map 67 additional addresses to a Blockchain.info tag.

We finally crawl the blockchain in its entirety up until August 2017 to surface transaction flows between the clusters, and construct the corresponding transaction graph.

In the resulting graph, the nodes are clusters of addresses that our heuristics indicate as belonging to the same entity and the directed edges represent at least one transfer of Bitcoin between addresses in different clusters (transactions between individuals). A preliminary analysis of this graph shows some familiar characteristics expected of real world networks. For example, a log-log plot of the degree distribution appears to show the degree distribution following a power law; a least squares estimate of alpha, after excluding outliers, suggests an al-

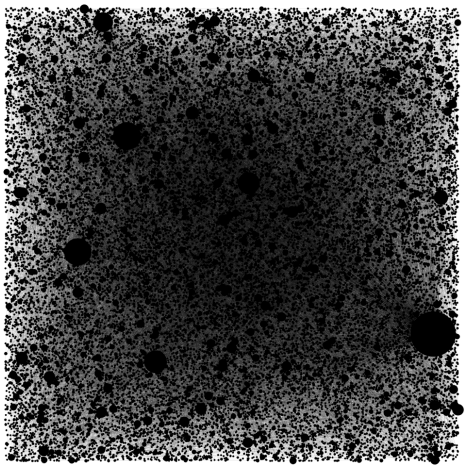
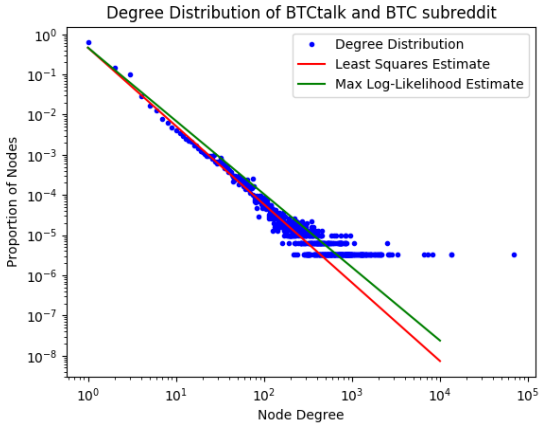


Figure 2. Visualized Bitcoin Clusters and Transactions (using Gephi[2])

pha of around 1.95. The maximum log-likelihood estimate indicates an alpha of around 1.82. This increases the confidence that the transaction graph that was constructed from the scraped addresses and through address clustering is representative of a real world broad network, and can serve as a reasonable point of comparison for what might be the more idiosyncratic network of software developers.



6.2. Network properties

Analyzing the network reveals a variety of convergences and disparities between the developer community and the broader Bitcoin ecosystem. By taking snapshots of the Bitcoin communities across time, we can see the changes to the graph structure and determine whether the communities are converging. Each snapshot includes all of the transactions from the beginning of time until the the start of the end of the indicated year, with the exception of 2017 where the dataset ends on August 1, 2017.

- Average node degree[figure 3]: Historically, the development has been more insular than the general bitcoin community. However, we see an overall decrease in the average degree distribution of the general bitcoin community, leading to a convergence in average degree distribution in 2017.

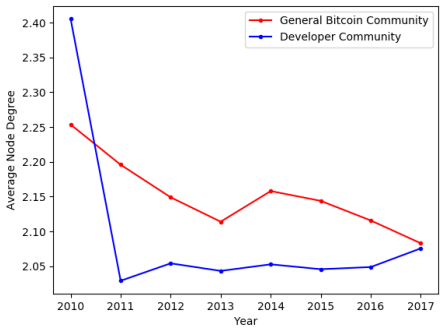


Figure 3. Average Node Degree over Time

- Average clustering coefficient[4]: We notice a similar trend when looking at the average clustering coefficient between both communities, and a similar convergence. Note that for both the general bitcoin and developer communities, the average clustering coefficient is almost always well above that of an Erdos-Renyi random graph with the same number of nodes and edges.

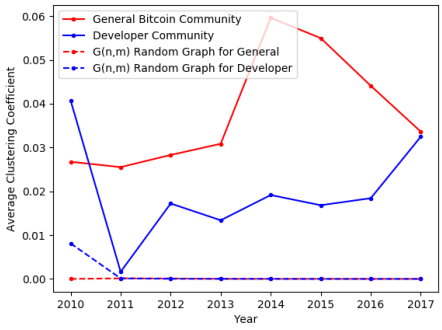


Figure 4. Average Clustering Coefficient over Time

- Average shortest path length[5]: When analyzing average shortest path, there is a consistent difference between the developer community and the broader ecosystem, with the former maintaining a consistent shorter path length throughout. This suggests that the community has maintained some degree of distinctiveness and insularity relative to the general Bitcoin community.

- Power law exponent[6]: A maximum log-likelihood approach to estimating the power law exponent of the degree distributions for both the general Bitcoin and developer networks shows exponents mostly within

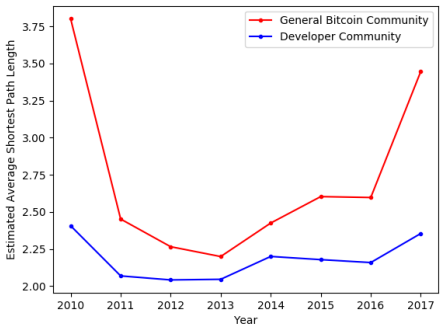


Figure 5. Estimated Average Shortest Path Length over Time

the range of 2.0 to 2.4, similar to values seen for other real life networks such as the internet web graph and online social networks. This corroborates other results such as the observation that the average node degree stays within a narrow range even as the graph size increases rapidly with later snapshots in time. This convergence of the average node degree to a finite value is a known characteristic of power law distributions with exponent $\alpha > 2$. The average clustering coefficient being much higher than an Erdos-Renyi graph with the same number of nodes and edges shows that the general Bitcoin and developer graphs are not behaving like random networks, corroborating the estimate of $\alpha < 3$.

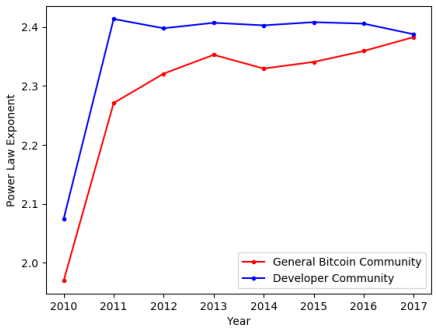


Figure 6. Estimated Power Law Exponent over Time

- Small World Phenomenon[4,5,7]: From the results on average shortest path and average clustering coefficient, we see that both the general Bitcoin and the developer communities exhibit small world phenomenon. In particular, even as the graph size of both communities increase by orders of magnitude as we take later snapshots in time, we observe that the average shortest path lengths do not increase in proportion to the graph size. We note that in this regard, the developer graph shows a nearly constant average shortest path length over time while the general Bitcoin graph seems to have relatively slowly increasing average shortest path length in more recent time snapshots. And as noted earlier, average clustering coefficient for both graphs is, for most of the snapshots in time, well above that of an Erdos-Renyi random graph with the same number of nodes and edges. Thus both the general Bitcoin and developer graphs appear to exhibit small world phenomena and have retained this characteristic over time.

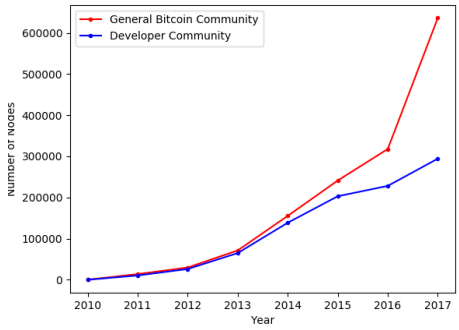


Figure 7. Number of Nodes in Graph Snapshots Over Time

6.3. PageRank

After clustering individuals via the multi-input transaction and one-time change heuristic techniques, and then further examining the edge between cluster across different dampening factors, we apply the PageRank algorithm and observe a stark contrast in the distribution of the Bitcoin network when compared to randomly generated graph models.

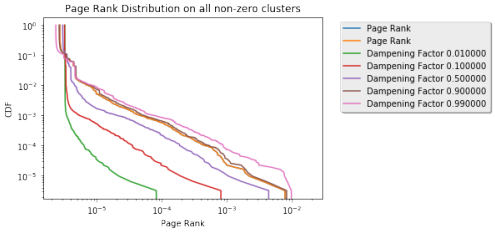


Figure 8. PageRank Distribution

When comparing the PageRank distributions for Bitcoin to the findings noted in Becchetti's [4] analysis of PageRank for large web graphs, we observe a long tail distribution for nodes that have lower PageRank scores. This reveals that the Bitcoin clustered transaction graph is more centralized than models of the web, with a smaller proportion of important nodes relative to the general community. Individual analysis of the top 35 clusters identified through PageRank reveals over 25,000 addresses. By cross-referencing the addresses with the list of tagged addresses available through Blockchain.info, individuals in-

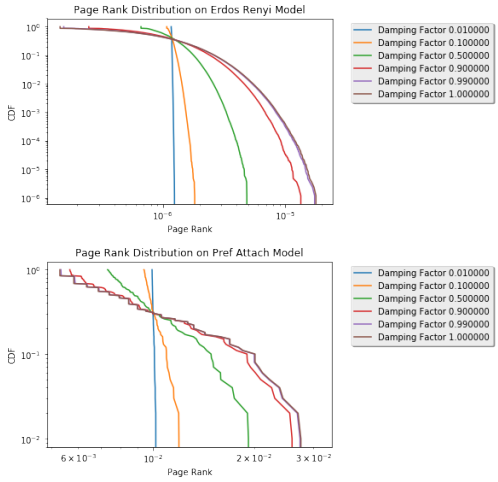


Figure 9. PageRank Distribution of Graphs Generated from the Erdos-Renyi model (above) and Preferential Attachment model (below)

volved in the FBI sale of coins confiscated from SilkRoad [1] were identified.

6.4. Distrust Propagation

We finally examine whether we can identify any suspicious activity in our network. We start with a base set of 174 individuals identified as "untrustworthy" in the bitcoin-otc dataset, using the criteria that any individual with a combined $score < 0$ is not trustworthy, and setting a threshold value of 4. We identify 39 untrustworthy clusters representing 1,234 bitcoin addresses.

In order to then uncover the identity of the clusters, we cross reference the individual addresses of each with the list of tagged addresses scraped from Blockchain.info. We end up uncovering a number of individuals using SatashiDICE (an illicit gambling platform that leverages Bitcoin for anonymity) and escrow/money laundering services. Expanding our analysis would allow us to further identify suspicious activity and de-anonymize illicit behavior.

7. Discussion

7.1. Difficulties encountered

- Quality of the analysis tools: The Bitcoin blockchain format has grown to a non-trivial scale (around 140GB for the timespan we explore), which bounds the possible iteration loop without expensive hardware. Furthermore, the storage format is still constantly changing, making it difficult for parsers to handle the different modifications. For example, few parsers currently handle SegWit, a backward-compatible change in the Bitcoin blockchain to increase block size that was activated in August 2017 (efforts to support it on BlockSci has been ongoing for several months for example). Additionally, many commonly used blockchain parsers are in various states of disrepair (the top hit on Google, https://github.com/znort987/blockparser, has not been updated since December 2015), or are relatively nascent and still have many sharp edges.

- Scraping Large and Poorly Defined Datasets: The unstructured and heterogeneous quality and functionality of the APIs that different forums and websites expose makes it challenging to collect addresses in a robust manner. Each website we explored was structured differently, and understanding and navigating the forum structure for both web crawling and API-scraping presented a non-trivial amount of effort. The number of websites we were able to explore within the scope of this project was thus inherently bounded - casting the net wider in subsequent work would undoubtedly be of interest.

7.2. Clustering results

During the address clustering phase, we observed an outstanding proportion of addresses belong to a giant cluster of over 150 million addresses. Although this merits investigation on its own merits, we largely excluded it from subsequent processing due to computational constraints.

We also explored augmenting the existing clustering implementation with some more exotic heuristics. In Regular payments: In the Conditions of Full Disclosure: The Blockchain Remuneration Model [12], English et al posit that they are able to link together different addresses that pay a fiat-normalized amount of Bitcoin to the same address at regular intervals. This was intriguing to us as being a potentially powerful way of flagging suspicious transactions and further leveraging the reputation data obtained from bitcoin-otc. Investigation into the method however proved to be challenging and gave results that were subject to high false positive rate - we posit that it is better suited for investigating specific nodes of interest instead of relying on it as an indicator during top-down clustering.

7.3. Community Differences

The network analysis indicates that the respective studied communities have been evolving differently over time. The developer community appears to have remained more stable over the different time periods, as the clustering coefficients and average path lengths have held steady. In contrast, there has been decreasing interaction between nodes in the general community graph, as evidenced by the decreasing average node degree and sharply increasing average shortest path length over time. This decreased interaction may reflect the increasing heterogeneity of the broader community over time as more newcomers decide to participate in the Bitcoin economy. We suspect that these trends will continue and lead to further divergence as interest in Bitcoin continues to draw in new mainstream

participants, while the developer community has shown its steadiness in the face of rapid change.

7.4. Further network analysis

The size of our transaction graph made it challenging to apply a wide range of potentially interesting network analysis techniques. In particular, we attempted to run Clauset-Newman-Moore community detection via the implementation present in SNAP [3], but were forced to abort our efforts after four days of compute time. Even relatively simple metrics such as average shortest path were non-trivial to calculate, as shown above, requiring sampling for intents of this paper.

The Bitcoin network itself is also constantly evolving - if anything activity in the blockchain has reached new heights since the end point of our analysis (August 1st, 2017, i.e. right before the introduction of SegWit which no publicly available parser supports at the time of writing). As interest continues to increase, we expect that the number of derivable insights will continue to grow, and the imminent release of SegWit support in BlockSci will allow for even more comprehensive analysis.

Overall we believe that we have only begun to scratch the surface of the insights this corpus of data promises to reveal, and hope that our efforts can serve as a basis for subsequent work to come.

8. Acknowledgment

The authors would like to thank Srijan Kumar and Anthony Kim for their guidance and valuable comments to improve the quality of the paper.

References

[1] Fortune, silkroad. http://fortune.com/2017/10/02/bitcoin-sale-silk-road/.

[2] Gephi graph visualization software. https://gephi.org/.

[3] Snap implementation of cnm clustering. http://snap.stanford.edu/snappy/doc/reference/CommunityCNM.html.

[4] L. Becchetti and C. Castillo. The distribution of pagerank follows a power-law only for particular values of the damping factor. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 941–942, New York, NY, USA, 2006. ACM.

[5] S. H. S. Duncan J. Watts. Collective dynamics of 'small-world' networks. http://snap.stanford.edu/class/cs224w-readings/watts98smallworld.pdf, 1998.

[6] M. R. T. S. S. C. Elli Androulaki, Ghassan O. Karame. Evaluating user privacy in bitcoin. http://www.syssec.ethz.ch/content/dam/ethz/special-interest/infk/inst-infsec/system-security-group-dam/research/publications/pub2012/596.pdf, 2016.

[7] H. A. Kalodner, S. Goldfeder, A. Chator, M. Möser, and A. Narayanan. Blocksci: Design and applications of a blockchain analysis platform. *CoRR*, abs/1709.02489, 2017.

[8] J. Leskovec and R. Sosič. Snap: A general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(1):1, 2016.

[9] B. F. Matthias Lischke. Analyzing the bitcoin network: The first four years. http://www.mdpi.com/1999-5903/8/1/7/htm, 2016.

[10] J. D. Nick. Data-driven de-anonymization in bitcoin. https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/155286/eth-48205-01.pdf, 2015.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia, 1998.

[12] E. N. S. Matthew English. Conditions of full disclosure: The blockchain remuneration model. https://arxiv.org/pdf/1703.04196.pdf, 2017.

[13] G. J. K. L. D. M. G. M. V. S. S. Sarah Meiklejohn, Majori Pomarole. Data-driven de-anonymization in bitcoin. https://cseweb.ucsd.edu/~smeiklejohn/files/imc13.pdf, 2015.