Bitcoin Criminality

LSDE-group11

You Hu
VUnetID: yhu310
VU University Amsterdam
Student Number: 2631052
adolphus.hu@student.vu.nl

Xiaoyu Yang
VUnetID: xyg230
VU University Amsterdam
Student Number: 2640948
2640948@student.vu.nl

Jiamian Liu
VUnetID: jlu510
VU University Amsterdam
Student Number: 2632301
j10.liu@student.vu.nl

Abstract—**Bitcoin is now famous for its anonymity and become a well-known channel for illegal trade. This paper is meant to find the characteristics of criminality related transaction network distinguished from normal users. The network and the distribution of transaction over time serials were visualized at the first stage with general features extracted in the same time. This provides intuitive view of the difference between two types of users. At the second stage, we analyzed the difference by statistics and case study. The results show that they differ in some aspects which can reflect how illegal business operated.**

Keywords-Bitcoin criminality; statistical analysis; network visualization

I. INTRODUCTION

Bitcoin is a cryptocurrency, a form of electronic cash. It is a decentralized digital currency without a central bank or single administrator that can be sent from user-to-user on the peer-to-peer bitcoin network without the need for intermediaries [1]. It has the characteristics of money (durability, portability, fungibility, scarcity, divisibility, and recognizability) based on the properties of mathematics rather than relying on physical properties or trust in central authorities. The complete history of all transactions ever performed in the Bitcoin network, called "blockchain", is public and replicated on each node. The data contained into the network is difficult to analyze manually, but can yield a high number of relevant information. Because bitcoin is freedom of payment, low service charge, low risk, neutral, transparency and easy to control, it has been the most trustworthy digital currency nowadays.

However, many bitcoin transactions are associated with illegal activity, such as drugs, hacks and extortion emails. Bitcoin, a kind of cryptocurrency, is an excellent choice for criminals. Through blockchain, Bitcoin users do not need a middleman (such as a bank or financial institution) to transfer their bitcoins. Thus, Bitcoin is anonymity and is likely to help criminals to avoid government monitoring. Furthermore, users are able to obtain a large amount of virtual wallet addresses that are only identified by some numbers, which increases the difficulty to find the real users and the real transaction goods behind the bitcoin data. In the current project, we aim to use identified addresses from both criminals and normal users to distinguish the criminal party.

II. RESEARCH QUESTIONS

Since Bitcoin and other 'coins' are used for criminal activities for many years because of the decentralism and anonymity, recognizing the criminality by their transaction pattern is a realistic approach. As we showed on the course, our goal is to find:

- What are the unique characteristics of criminality transaction activities compared to normal transactions?
- Can we distinguish criminality addresses from normal users?
- What would be done when illegal activities happened and whether it can be reflected by transaction records?

Although we have 130GB of raw data, we only focus on few addresses and the transaction records related to criminal transaction records. Besides, because of limitation of the quantity of identified addresses, it is hard to make a strong conclusion on macro level. Thus, we will pay more attention on finding relationships between transactions and reality behaviors.

III. RELEVANT WORKS

Researching bitcoin criminality has been an important and hot research field nowadays. Researchers have done a lot of relevant works in technical level or non-technical level. Our work is done on the basis of many previous work. Spagnuolo and his colleges presented a modular framework, BitIodine, which parses the blockchain, clusters address that are likely to belong to a same user or group of users, classifies such users and labels them, and finally visualizes complex information extracted from the Bitcoin network, but the BitIodine does not adapt to the data from 2017 [2]. By doing statistical analysis work in Silk Road criminal platform, the value evolution, Nicolas evaluated the activations of the bitcoin user in the criminal network. However, the conclusion cannot be concluded into a formula [3]. Alice did the research work in analyzing the account transaction records of Silk Road, such as time and frequency. Only in the theoretical level, they gave advices to establish e-economic law and regulation [4]. Brown established cryptocurrencies to establish the performance of bitcoin in criminal transactions, which proved to be a low-risk currency [5]. Joseph and his colleges proposed Mixcoin, a

protocol to facilitate anonymous payments in Bitcoin and similar cryptocurrencies. They build on the emergent phenomenon of currency mixes, adding an accountability mechanism to expose theft, but the accuracy need to be further increased [6]. Malte provided a first systematic account of opportunities and limitations of anti-money laundering (AML) in Bitcoin, a decentralized cryptographic currency proliferating on the Internet. They used reverse-engineering methods to understand the mode of operation and try to trace anonymized transactions back to our probe accounts. The operation was well analyzed but the transactions is unclear to some extent [7].

IV. Overview Of Work

A. *Piepline*

Our work starts from identified addresses and ends to characteristics. To make it run efficiently we set up a pipeline to handle the data. Fig.1 shows the overview of our work.

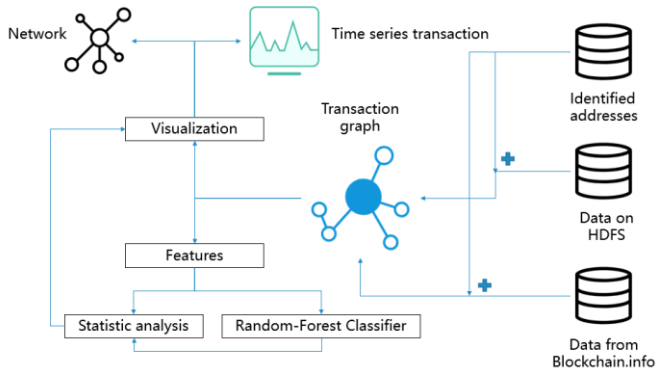


Figure 1. Work Overview

When an identified address come out, we extract transaction records from HDFS or Web service. And then transform them to a transaction graph, which is formed of NetworkX structure.

After that, visualization and analysis work will be commenced.

We visualize the network to find special patterns of relationship between users and charts about distribution of transactions among time series, in order to verify our guess like whether criminality addresses active in very short time.

In the meanwhile, we extract general features about the graph like centrality, value of transactions, in order to see what the differences among graphs of these two types of addresses are. After acquiring features, we carry on statistical analysis and build random-forest classifier. Although we don't have too many samples, the classifier can return a list of importance of features. The importance list can denote which feature contribute most, which can save time on analyzing features one by one.

This pipeline excluding visualization part has good scalability, when get more identified addresses or more useful features.

B. *Data Collection*

Our bitcoin transaction data come from two sources. One is the compressed data saved on HDFS of SURFsara cluster. This data is raw data of blockchain which need extra tool to decode each block. This data collection is over 130GB and the transactions spinning from 2009 to 2016 [8]. Another is data from BlockChain.info. This website provides APIs to get transaction records of any active addresses. The web site separate block information and transaction records. Since we only concern trade data, it is more flexible to get data from web service than cluster.

C. *Technologies*

Since our data collection is too large, the computation overhead is unaffordable for any single machine. We began to extract transaction records by using Spark with Scala. Because Spark is developed by Scala, Spark API for Scala is the most complete and related problems can be easily found on the internet. Spark provides SQL syntax and Dataframe type to filter data, and MapReduce to process data in parallel way. Besides of Spark, we also use Python library to collect data from web. As we only need to get transaction records and the APIs provide them directly, we choose Requests instead of Scrapy, a light solution.

Because each address could be represented as the node of a network, each transaction can be represented as the edge of a network, other characteristics can be added to the nodes and edges, the dataset we acquired hereby can be considered as a network. By mainly taking use of NetworkX, we not only realized the work of feature extraction from each address, each transaction records and the total transaction network, but also did the calculation and analysis work for the features.

NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. It is suitable for operation on large real-world graphs, such as the bitcoin network. Due to its dependence on a Pure-Python "dictionary of dictionary" data structure, NetworkX is a reasonably efficient, very scalable, highly portable framework for network and social network analysis [9]. It includes many standard graph algorithms and could generate different kinds of graphs, which adapts to different situations. It is well tested with over 90% code coverage according to the GitHub test [10]. Besides, it additional benefits from Python include fast prototyping, multi-platform and so on. In summary, NetworkX is the preferred and irreplaceable tool in network processing work.

In order to visualize our results and give a vivid display, we used an open-sourced and declarative library, named Echarts to build a web-based visualization. Echarts is an interactive and componentized library, so Fig.2 shows it use a streaming architecture [11]. We only need to give some raw data (json format) to Echarts and easily use some function of Echarts in order to get the visualized output which we want. Also, Echarts supports user interaction which can help users easily choose particularly data by themselves.

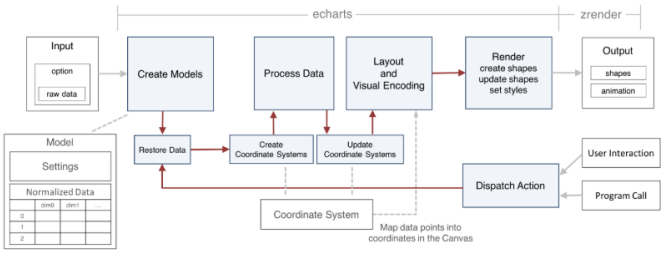


Figure 2. Echarts Streaming Architecture

For statistical analysis, we use Pandas to process our data in Notebook workspace. There is no doubt that Pandas and Numpy have become de facto standard library for data analysis in Python platform. The processed results can be directly used by Matplotlib for drawing graph and Sklearn for machine learning.

To find important features more efficiently as well as classify the normal and criminality addresses to some extent, we used Random forest classifier which is provided by Sklearn. Since our dataset is not big and there are many features of graph should be taken into consideration, Random forest classifier, ensemble learning model based on decision tree, is a good choice. And the most import feature of this model is that it can provide importance of each feature. When the model performs well in test data, we can check which feature contributes most, then we can check how it classify positive and negative samples.

V. EXPERIMENTS AND RESULTS

A. *Collect identified addresses*

Because the Bitcoin transaction and address are anonymous, it is extremely difficult to identify the true owner of a Bitcoin address only through the Bitcoin blockchain dataset. Therefore, we got the identified address from two databases (Bitcoin Abuse Database: https://www.bitcoinabuse.com/ and Bitcoin Who's Who: https://bitcoinwhoswho.com/). From the two databases, we choose about 150 identified addresses with two groups. One group of addresses is the criminals' addresses and the other group of addresses in the normal users' addresses. The reason is that we want to find the difference features between the Bitcoin criminal addresses and normal addresses. For criminals, we chose the 120 identified addresses from 5 kinds of crime, such as extortion email, blackmail, sextortion, FBI and ransomware. However, many criminal addresses do not have any transaction records, and we choose 29 typical criminal addresses. For normal users, we found 20 addresses from charities. Moreover, then, we choose 24 normal users who donate bitcoin to the open source project (ReactOSProject). As a result, we total have 36 typical identified criminals' addresses and 24 normal users' addresses as our Bitcoin identified addresses collections.

B. *Extract transactions to form a graph*

In blockchain, the minimum unit is address, which means that one can own many addresses. We can assume that addresses come up together in input or output of one transaction may belong to same person (especially in input). Besides of direct connection, those who transfer bitcoin to or receive from our central identified addresses, the addresses indirectly connect to central addresses are also important. If we can figure out what is the role of these addresses, it can help us know more about transaction activities of identified addresses.

Therefore, the addresses we concerned, and their relationships are shown in Fig.3:

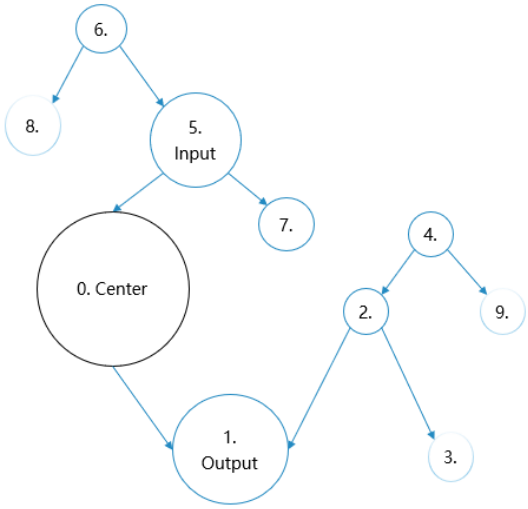


Figure 3. Relationships of addresses

In the graph, it contains these 10 types of node, and we label them with 0 to 9.

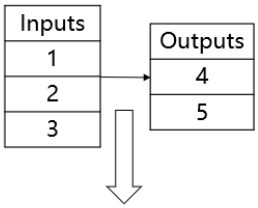
- Type 0: central addresses owned by normal users or criminals
- Type 1: addresses that central addresses transfer bitcoin to
- Type 2: addresses that transfer bitcoin to type 1 along with type 0, maybe belong to the owner of type 0
- Type 3 & Type 4: addresses send or receive bitcoin to or from type 2, reflect the network of type2
- Type 5: probably hold by victims or middlemen in criminality network, maybe other normal users in graphs of negative sample
- Type 6: addresses that send bitcoin to type 5, indirectly connect to central addresses
- Type 7: addresses that receive bitcoin from type 5
- Type 8 & type 9: addresses that receive bitcoin from type 6/type 4 along with type 5/type 2

In general, we collect all transaction records of type 0,2 and 5, part of transaction records of other types of node.

It is easy to collect these data by calling APIs of web service. Firstly, collect all records of central addresses, and then filter addresses of type 2 and type 5. After that, collect their data by the same way.

Collecting data from cluster is much complicated. First, we need to transform the raw data to RDD that can be transformed to Spark Dataframe later. We use Hadoopcryptoledger library to parse the CryptoLedgers of bitcoin. This library provides formats for blocks and transactions. As we only consider the transactions and partial information of them, we directly used the interfaces the library provides to get specific fields of data.

To save the time, we filtered and saved the transaction records as intermediate data because parsing the whole dataset will spend about 20 minutes. In raw transaction data, there are multiple inputs and multiple outputs, we saved it with the form of source-to-destination, of course values, transaction hash and timestamps along with source and destination addresses.



Source	Destination
1	4
2	4
3	4
1	5
2	5
3	5

Figure 4. Source and destination data

After that, each query starts from this dataset, and it will cost about 20 minutes to extract all data same as the data by using web APIs.

These two methods both have advantages and disadvantages. Collecting data from web is very convenient but there are requests' limitations because of the processing ability of server, although we have API key to bypass the request blocking. The time overhead of this method varies from few seconds to hours or even days depending on the transaction amount. Compared to this light method, extracting data from cluster costs much time but steadier, because reading parquet data from HDFS will cost much more time. Maybe we can query multiple addresses in one time to reduce the ratio of reading overhead. Another problem is that the dataset in HDFS is not up to date. However, extracting data from HDFS is more scalable but for the scope of this project, collecting data from web service is a better approach as we don't have too much identified addresses to analyze.

We chose NetworkX to form a structural graph. First, we transform transaction records to the form of source-to-destination. Each source-to-destination record is an edge in the graph. In NetworkX library, edges can be added directly, and nodes will be added automatically if it does not exist in the graph. Then we transform records to edges to build up a graph.

C. Visualize the transaction graph

Fig.5 shows we parser the transaction graph to different format data which is accepted by Echarts. Moreover, we use the Echarts, Javascript and HTML to build the data visualization. We built two different kinds of visualization in order to find the features in Bitcoin addresses(nodes) and transactions(edges).

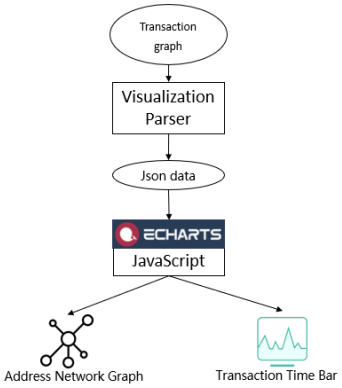


Figure 5. Visualization Method

For Bitcoin addresses, we built a network graph (addresses are the nodes, transactions are the edges) to represent the relationship between the identified addresses and relevant addresses. The network graph can help us to find some special structures to distinguish the criminals' network graphs and the normal users' graphs. We built graph based on the identified address as the central address. And then, we add other addresses which have relevant transaction with the center address. For example, Fig.6 shows the network graph of criminal sex15. The criminals' address (144CDUeBhcwoEUmA2B1cL5p5PqZrhJWCCt) is the big dark blue node, and use different color to represent different distances to center address, such as sky blue, orange and green. For large graph, it is difficult to find the special structures, because there are too many nodes and edges in the graph. Therefore, we choose the nodes whose degree larger than 3 to build subgraphs in order to get a clearer graph.

Through our Bitcoin network graph visualization, we can intuitively find some differences between the criminals and normal users. The Fig.6 shows a sextortion criminal graph whose nodes tend to cluster together and formed into many small clusters. However, Fig.7 show a normal user graph whose network is more chaotic and disordered.

For transactions, we built a time bar (X-axis: time, Y-axis: transaction amount) in order to find the difference of transactions time distribution between the identified addresses and relevant addresses. The lower time bar belongs to the identified address's transactions and the top time bar belongs to the other addresses' transactions. For clearer visualization, we added the data zoom function at the edges of the bar. The data zoom X-axis can help the user choose the transaction amount

that they want. The data zoom Y axis can help the user choose the transaction time that they want in order to avoid some extreme value.

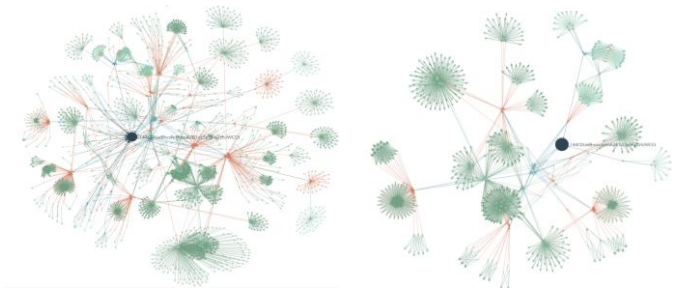


Figure 6. Criminal Sex15 Graph and Sub-Graph

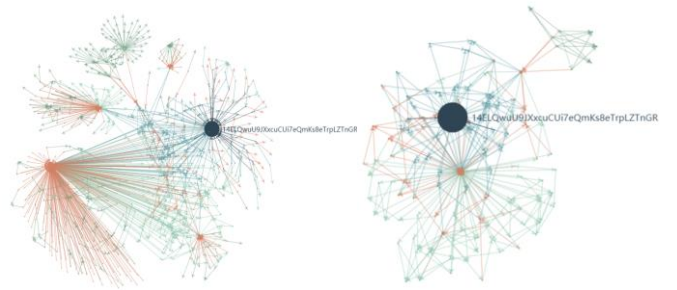


Figure 7. Normal User 236D Graph and Sub Graph

From the time bar of the transaction, we can find the transaction time feature difference between the criminals and normal users. In Fig.8, we can find that criminal user-Sex03's transactions distribute within a short period (10 days). However, Fig.9 shows normal user-236D's transactions distribute in a long period (8 months). The reason is that criminals usually use a new temporary address and will transfer money to another address as soon as possible. Conversely, normal users will use an address for a long time.

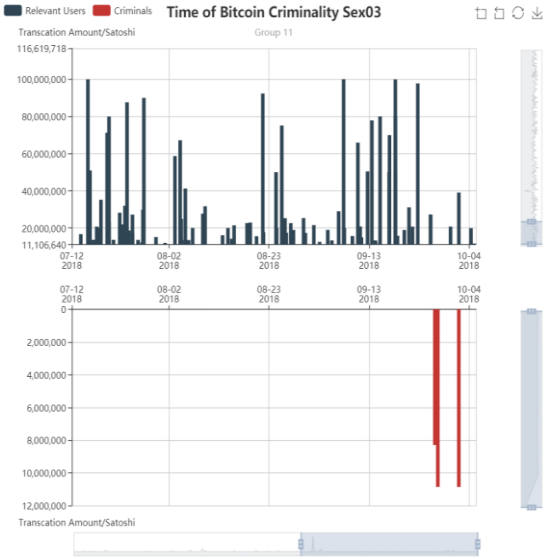


Figure 8. Time Bar of Criminal Sex03

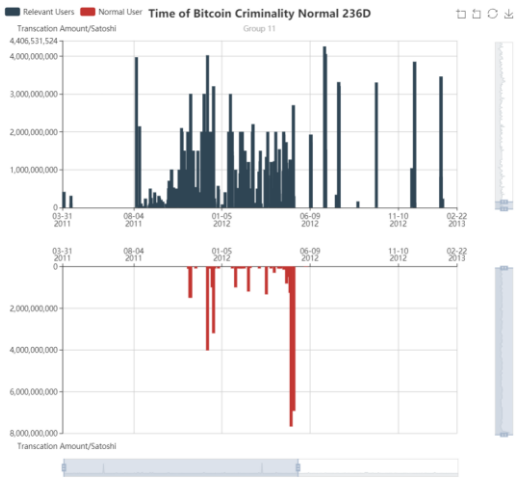


Figure 9. Time Bar of Normal User 236D

D. General Features of Graph

1) -Features and their meanings

In order to analyze the inner special structure and distinguish criminal network from the normal user, just by making use of the visualized transaction graph is unrealistic, we have to first extract the general features of graph, summary the characteristics of each relevant nodes and analyze the data we first processed. Thus, we could get the conclusion in the view of quantity.

By mainly taking use of NetworkX and other analytical tools, we successfully extracted 35 features for each dataset. We select most part of those features to do the deep level analysis. Furthermore, we visualized part of the analyzed results, which could make our easier to understand. Some important general features are to be illustrated as follows.

- Degree. Total degrees, in and out degrees are basic and import characteristics of each node. They reflect the connection situation of the network. We not only summarized the amount and draw the histogram as Fig.10, but also we extract the large out degree nodes to further check the special nodes of the network (the work in next section).
- Time difference and the connection situation. We calculated the time difference of the central address. Hereby the activation period of central addresses could be detected. We also summarized the nodes which is directly connected with central addresses.
- Box plot values. Box plot is a method graphically depicting groups of numerical data through their quartiles. Box plots are non-parametric: they display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution. We summary the box plot values for the transaction values to show the change and regulation.

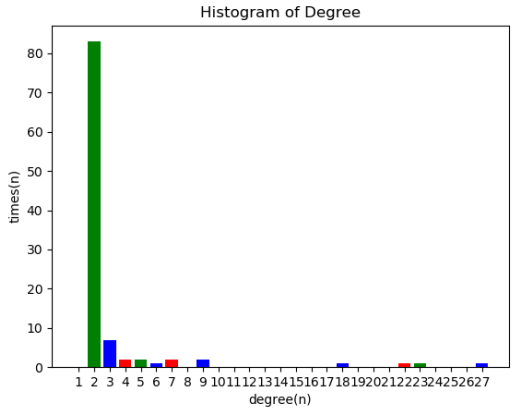


Figure 10. Sample of Degree Histogram

- Centrality. Centrality is given in terms of a real-valued function on the vertices of a graph, where the values produced are expected to provide a ranking which identifies the most important nodes. We select the nodes with in-degree centrality and out-degree centrality larger than 1% and calculated the percentage of each kind of nodes.
- PageRank. PageRank works by counting the number and quality of links to a node to determine a rough estimate of how important the node is. The underlying assumption is that more important websites are likely to receive more links from other websites. We summarized the percentage of PageRank value larger than 1% nodes.
- Clustering coefficient. Clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. We selected clustering coefficient larger than 0 nodes and calculated the percentage of the whole nodes.

2) -Random forest classifier

Since we have got so many general features, we established a random forest model to classify criminality relevant addresses from normal users, and in the meanwhile found which features contribute to the model most. Here, we used 32 out of all features and 30 samples, because some features are strongly correlated to the way we collect identified data. Besides, due to the lack of samples, we used Hold-one cross-validation method to metrics the model.

For each iteration, one sample is picked as test set and rest of samples are set as training set. This is very useful for small dataset. In scope of this experiment, we parameter the model with tree number as 10 and max features as all features and default for rest of all other parameters.

After 30 iterations that all samples are set as test set, we got 30 prediction output. Compared to true value, we got ROC score of 0.71 (provide by sklearn) and FPR/TPR of 0.42/0.83 (provided by sklearn). This result shows that random forest classifier is effective for distinguishing criminality from normal users, although FPR is high and it is not convincing because of small dataset.

For another goal of using RF model, we saved importance of features for each iteration. Table x shows the maximum importance of each feature among 30 times of iteration.

As Table I shows, min value and important nodes time difference can reach approximately 50% of importance. But we found that for donors' transaction network the min value of transactions is 1 Satoshi. We think it cannot be considered as an important feature to distinguish negative and positive samples. As to the important nodes time difference, it indicates the length between the first and last transaction of central addresses.

TABLE I. FEATURES AND IMPORTANCE

Assortativity	0.031275	***Percentage of out centrality>0.01 addresses***	0.09809
Average cluster coefficient of the whole graph	0.08885	***Percentage of page Rank >0.01 address***	0.022101
Average value	0.077885	***Reciprocity***	0.113361
Chordal	0.010662	***Significance test(p-value)***	0.193929
Immediate dominator(numbers)	0.253321	***Strongly connected***	0
Important nodes information	0.095397	***The max degree of all addresses***	0.014309
Important nodes time difference	0.494129	***This node as from***	0.181747
Is branching	0.010985	***This node as to***	0.152458
Is simple path	0	***This node connected with(as from)***	0.095978
Max value	0.12362	***This node connected with(as to)***	0.064871
Median value	0.213466	***Three fourth value***	0.115783
Min value	0.526514	***Total Transaction***	0.043424
Normality check(p-value)	0.049208	***Total addresses***	0.0528
One fourth value	0.0775	***Transitivity***	0.047856
Percentage of cluster coefficient>0 addresses	0.0855	***Triangles***	0.028835
Percentage of in centrality>0.01 addresses	0.051411	***Weakly connected***	0

Since we know that the time difference is very important, but we cannot easily get the conclusion that whose transactions are more intensive, because the time difference of one addresses whose first transaction happened last year wouldn't be exceeding 1 year. As is shown in Fig.11, the criminality related transactions happened closer to recent, because we collect them from reports with time order.

To reduce the difference caused by the way we collect data, we choose the probation of time difference of central address among whole transaction network.

$$Percentage = \frac{lastT0 - firstT0}{lastTw - firstTw} \quad (1)$$

In Formula (1), lastT0 means last transaction of Type0, firstT0 means first transaction of Type0. lastTw means last transaction of whole network, firstTw means first transaction of whole network.

As it is shown in Fig.12, the percentages of donors are higher than that of criminalities. This proves our guess in some extent that the difference in transaction persistence and intensity among these two types of users.

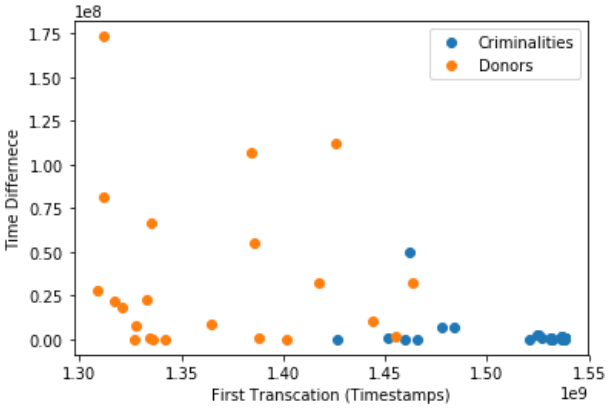


Figure 11. First Transaction & Time difference

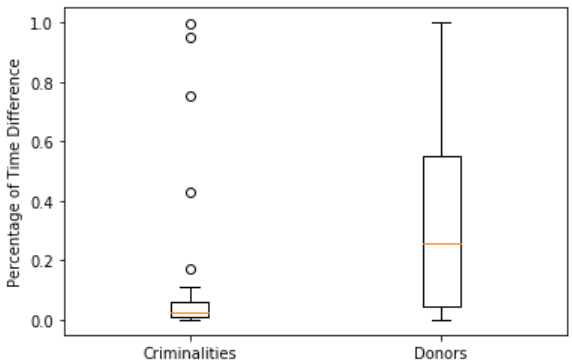


Figure 12. Time difference among the network

E. Analyze special structures

On the whole, we still need to find the difference between them in details. After having glanced up to 20 visualized transaction networks, we found two types of inter structures which indicates two different patterns of transaction.

One is that one address receives bitcoins from or send coins to multiple addresses and the degree of this address is very high. We call it as one2multi. The typical example is shown on Fig 13.

The other is that in one transaction record there are numbers of input addresses and output addresses. We call it as multi2multi. An example is shown in Fig.14

Through our Bitcoin network graph visualization, we can intuitively find some differences between the criminals and normal users. Fig.6 show a sextortion criminal graph whose nodes tend to cluster together and formed into many small clusters. However, Fig.7 show a normal user graph whose network is more chaotic and disordered.

And then we found that the first structure is more common in normal user's networks while multi2multi structures have more chances of appearing in criminality relevant networks. Thus, we analyzed the distribution of these two structures among different types of network.

For one2multi structure, we use the percentage of the number of nodes degree exceed 100 among the number of nodes whose degree exceed 2 to measure density of this structure in one graph. For multi2multi structure, we use the percentage of the number of transaction records whose source and destination addresses are both over 3 among the number of all transactions.

Here we measured 24 negative samples (donors) and 36 positive samples (criminalities) and analyzed them with boxplot.

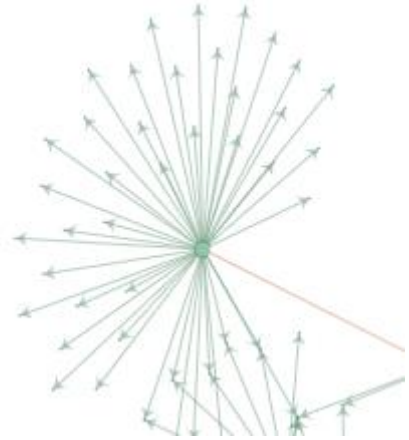


Figure 13. one2multi example

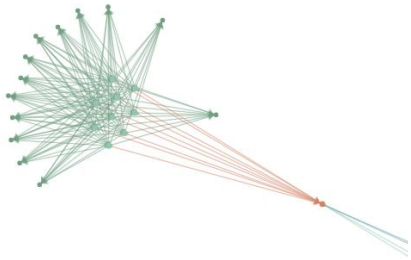


Figure 14. multi2multi example

Comparing the Fig 15 and Fig 16, this illustrates that in donor's or normal user's network there are more one2multi structures and more multi2multi structures in criminality transaction networks. Conversely, we can assume that the addresses with network containing high level of multi2multi are probably criminality related.

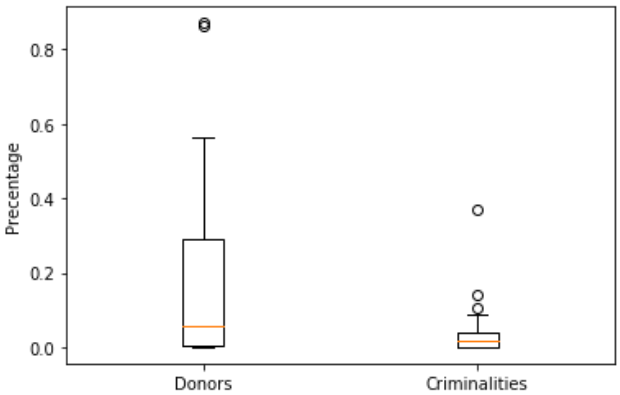


Figure 15. one2multi result

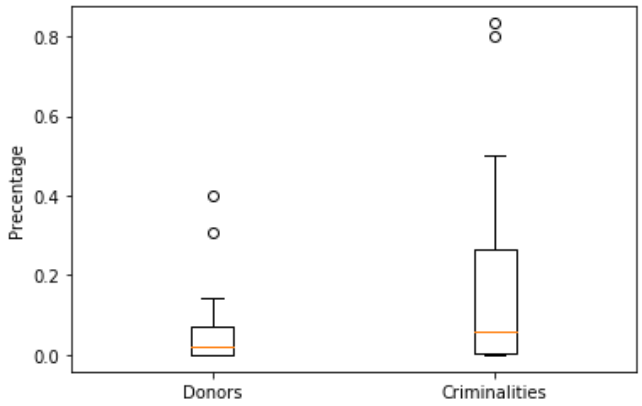


Figure 16. multi2multi example

F. Behavioral pattern analysis

1) Cases Study

We choose some identified addresses to give more detailed analysis. Fig. 17 shows that identified criminal address (the big dark blue node) and the Type.2 node (orange node) both transfer bitcoin to Type.1 address (1E2J2DAFGToqmTxK5H8fG6V3Tp3xqSiAm4). As Fig.18 shown, the nodes around the Type.2 node (1MiKAU9XMbV9QJHfywfRZBpsN2KJRZwEsS) tend to cluster together and formed into a small cluster. We can find the multi2multi structure in the nodes' cluster: one group of nodes in peripheral all transfer bitcoin to another group of nodes in the structure center in Fig. 19. Conversely, the group of nodes in the center always collect bitcoin from transferring group in Fig. 19. From the internet, we can also identify the Type.2 node (1MiKAU9) is also belonged to a criminal, in Fig. 20 [12].

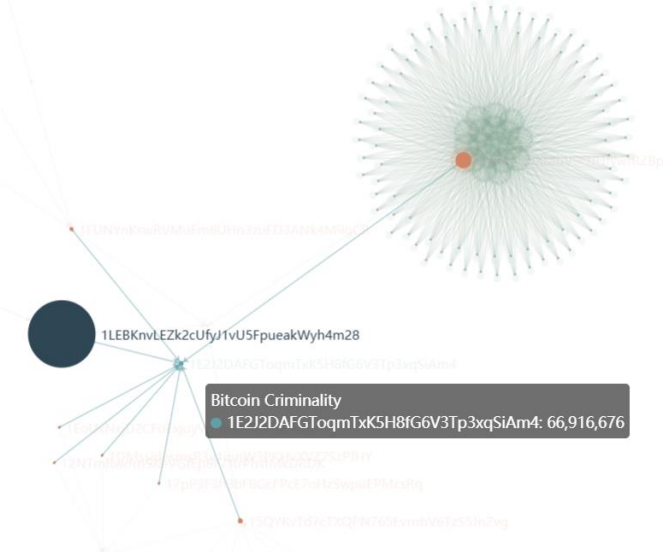


Figure 17. Type.1 Address 1E2J2D

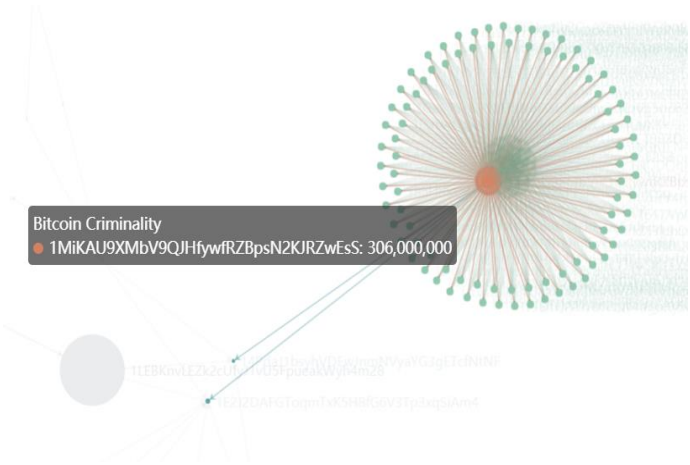


Figure 18. Type.2 Address 1MiKAU (multi2multi)

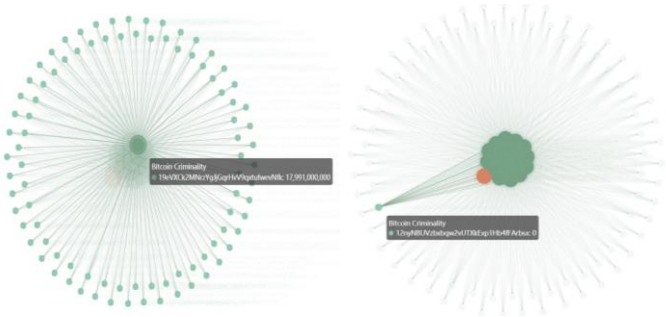


Figure 19. Transferring and Collecting Group of Nodes (multi2multi)

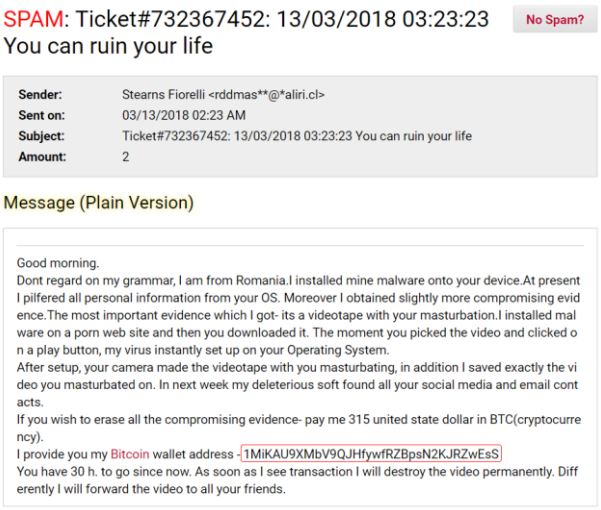


Figure 20. Type.2 Address 1MiKAU Spam Report

This multi2multi structure is usually seen in the criminals' network graph. For instance, we can also see some nodes to cluster together and formed into many clusters in Fig. 21. Specifically, we choose a Type.2 node (115i2xierq98FX3iWmj7yXog4bwQx3ACVg) which has a similar multi2multi structure of criminal in Fig. 22. Also, we

can find this node (115i2xier) belongs to a criminal from Bitcoin Abuse Database, in Fig. 23[13]. The reason for the multi2multi structure of criminal may be that criminals want to distract their bitcoin or hide their main bitcoin address. Therefore, these addresses' transaction is not common which only transfer between same group addresses and form the multi2multi structure.

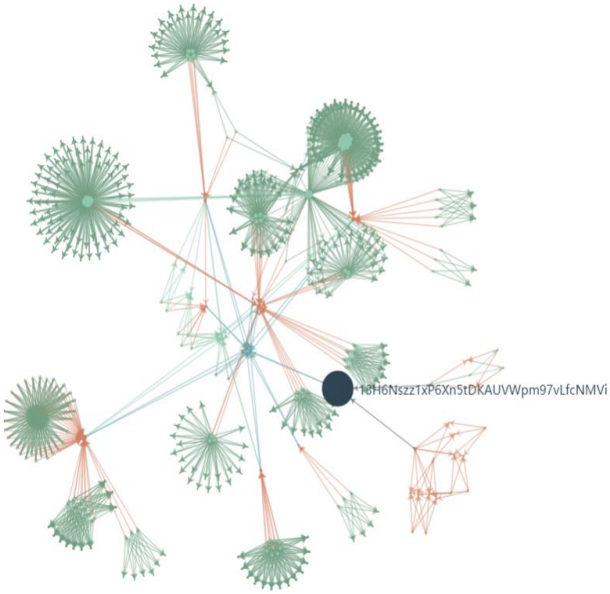


Figure 21. Criminal Sex13 Sub Graph

Figure 22. Type.2 Address 115i2xi (multi2multi)

For normal users, as Fig.24 shown, their networks are more chaotic and disordered. In Fig.25, we can see a Type.5 address (1fv5iQUYLgCCj8tKAQm5cnvZEbW3N5cQP) which both collects and transfers many bitcoins. Fig.26 shows, this address receives many bitcoins from Deepbit and Deepbit is a bitcoin cash mining pool address. Therefore, this address should be a middleman address which is responsible for transferring and collecting bitcoins for other addresses. Normal users' transactions are used for the real field between other addresses, instead of transferring bitcoin between same group addresses.

Address found in database:

Address	*115i2xierq98FX3iWmj7yXog4bwQx3ACVg*
Report Count	57
Latest Report	Thu, 04 Oct 18 18:55:50 +0000 (3 weeks ago)

Figure 23. Type.2 Address 115i2xi in Bitcoin Abuse Database

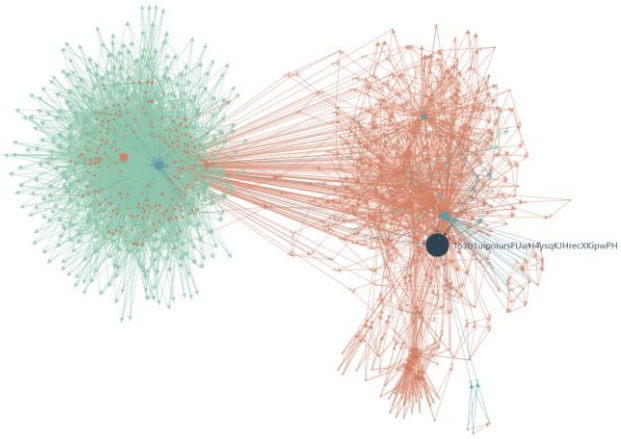


Figure 24. Normal User 2C5D Sub Graph

Figure 25. Type.5 Address 1fv5iQ

Figure 26. 1fv5iQU transaction records

2) - Statistic facts

Apparently, each type of addresses plays different role in the network. Since we have learned some cases that reflect how criminals run their illegal business through bitcoin network, it is still unclear what is the role those special structures we mentioned play. Thus, we counted the number of key addresses of these two structures in each type.

For one2multi structure, we filtered addresses whose out degree exceed 100, and classified them to different type. Below shows the distribution of total number in each type, here 0 in tag column means the samples of donor/normal user and 1 means criminality samples.

For multi2multi structure, we filtered transaction records with over 5 input addresses and 5 output addresses, and classified by the role of input addresses. The result is shown in fig x, there are only 4 types of addresses can be included in this structure according to the inference.

VI. FUTURE WORK

In the scope of this paper, our identified addresses come from reports of criminality. Because the forms of reports are different case by case, we could not automatically collect them by using web crawler. Therefore, we were only able to collect manually. Limited by the project time schedule, we only got about 50 addresses that reported related to criminalities, and 36 of them are useful. In the future, if more identified addresses provided, we can make more work on classifier and it will be more convincing.

Besides of the number of identified address, our pipeline is very convenient for developing but the computation efficiency is limited by python. In some cases, we had to abandoned part of samples because they are too huge to get some features. And another problem is about transaction data collecting. Although it is very fast when the network we concern is not very big, the time cost is unaffordable when the number of transactions exceeds millions. Therefore, we can transplant the whole pipeline to cluster, although currently we only use it for collecting data and the cost seems higher than that of API calling. In this way, we can reduce the limitation of computation ability and manual work.

TABLE II. ONE2MULTI STRUCTURE

Tag	Normal	Criminality
type 0	0	0
type 1	0	0
type 2	293	8
type 3	0	0
type 4	379	1191
type 5	317	23
type 6	1212	636
type 7	0	0
Total>=100	2201	1858

TABLE III. MULTI2MULTI STRUCTURE

Tag	Normal	Criminality
type 0	397	0
type 2	659	0
type 4	686	459
type 5	2329	3968
type 6	14345	50692
Total	18416	55119

VII. CONCLUSIONS

Our project aims to find the characteristics and trace of criminality related bitcoin addresses. We visualized the transaction network and found some facts which indicate where those networks differ from that of normal users.

By analyzing the time difference of central address, the results show that criminality addresses are active in short time compared to normal users. Their time differences are smaller than that of normal users. We suppose that criminals tend to abandon the address after they sell all bitcoins.

Besides of general features, we found two types of structure. Although we still don't know their identity in real world, it is obvious that they play different roles in different network. In criminality networks, there are more multi2multi structures and less one2multi structures compared to normal users' network. The key addresses of one2multi structures are mainly type 6 in normal users' network and type 4 in criminality network. The criminality addresses won't appear in the input of any multi2multi structures and same as type2 in their networks, while some of donor addresses are in the input side of multi2multi transactions.

The result of case study shows that type2 are probably another criminality addresses and have high possibility of being owned by the same person. This conclusion will help us to find more suspicious addresses.

To sum up, because of lack of identified addresses, our conclusion is not very strong. We are not able to build up an

effective classifier. However, we found and summarized the different structures and their characteristics in criminal and normal networks. These facts of bitcoin criminality can still provide a significant view that will help us go through deeper.

REFERENCES

[1] "Statement of Jennifer Shasky Calvery, Director Financial Crimes Enforcement Network United States Department of the Treasury Before the United States Senate Committee on Banking, Housing, and Urban Affairs Subcommittee on National Security and International Trade and Finance Subcommittee on Economic Policy". fincen.gov. Financial Crimes Enforcement Network. 19 November 2013.

[2] Spagnuolo M., Maggi F., Zanero S. (2014) BitIodine: Extracting Intelligence from the Bitcoin Network. In: Christin N., Safavi-Naini R. (eds) Financial Cryptography and Data Security. FC 2014. Lecture Notes in Computer Science, vol 8437. Springer, Berlin, Heidelberg

[3] Nicolas Christin. "Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace". Proceedings of WWW 2013, May, 2013.

[4] Alice Huang, Reaching Within Silk Road: The Need for a New Subpoena Power That Targets Illegal Bitcoin Transactions,56 B.C.L. Rev.2093(2015)

[5] Brown, Steven David. (2016). Cryptocurrency and criminality: The Bitcoin opportunity. The Police Journal. 10.1177/0032258X16658927.

[6] Bonneau J., Narayanan A., Miller A., Clark J., Kroll J.A., Felten E.W. (2014) Mixcoin: Anonymity for Bitcoin with Accountable Mixes. In: Christin N., Safavi-Naini R. (eds) Financial Cryptography and Data Security. FC 2014. Lecture Notes in Computer Science, vol 8437. Springer, Berlin, Heidelberg

[7] M. Möser, R. Böhme and D. Breuker, "An inquiry into money laundering tools in the Bitcoin ecosystem," 2013 APWG eCrime Researchers Summit, San Francisco, CA, 2013, pp. 1-14.

[8] CWI, https://event.cwi.nl/lsde/2017/results/group21.pdf

[9] Aric A. Hagberg, Daniel A. Schult, Pieter J. Swart, Exploring Network Structure, Dynamics, and Function using NetworkX, Proceedings of the 7th Python in Science conference (SciPy 2008), G. Varoquaux, T. Vaught, J. Millman (Eds.), pp. 11–15.

[10] NetworkX developers, http://networkx.github.io/, 2014-2018

[11] Deqing Li, ECharts: A declarative framework for rapid construction of web-based visualization, Visual Informatics, Pages 136-146, June 2018

[12] MailSpam Info, https://mailspam.info/en/icket732367452-13-03-2018-03-23-23-you-can-ruin-your-life_6379134634749309133-bc8ca4a23f80906b3b0cc55dd9b9065a.htm

[13] Bitcoin Abuse Database, https://www.bitcoinabuse.com/reports/115i2xierq98FX3iWmj7yXog4bwQx3ACVg