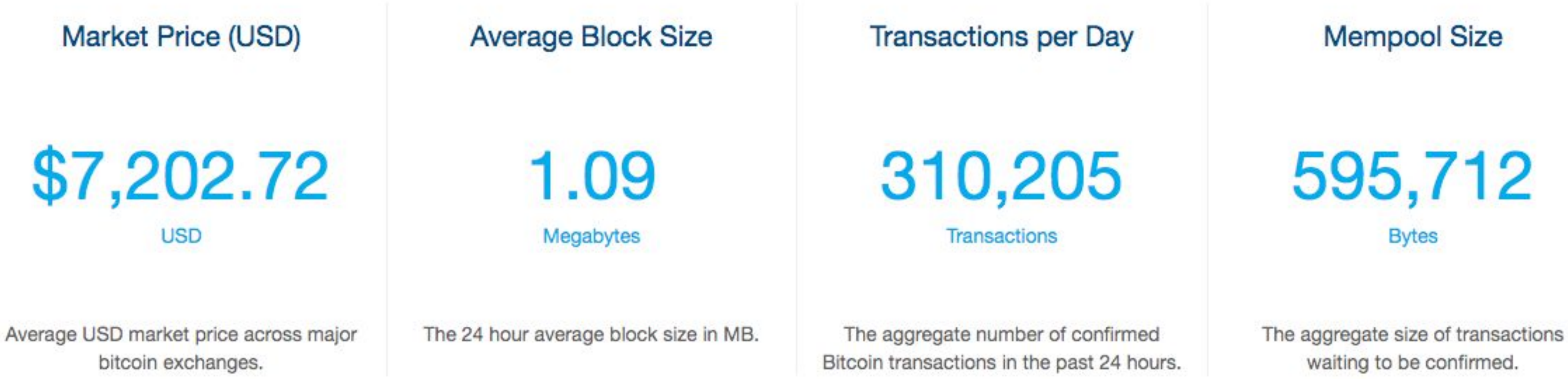
Bitcoin User Identification Research & Transaction Data Visualisation

CS-8903-L03 Special Problem with Prof. Ling Liu

Shubhi Agarwal

Bitcoin



Market Price (USD)

$7,202.72

USD

Average USD market price across major bitcoin exchanges.

Average Block Size

1.09

Megabytes

The 24 hour average block size in MB.

Transactions per Day

310,205

Transactions

The aggregate number of confirmed Bitcoin transactions in the past 24 hours.

Mempool Size

595,712

Bytes

The aggregate size of transactions waiting to be confirmed.

Motivation

- 480 million transactions in last 8 years and around 3 million transactions per day
- Widely considered anonymous due to use of pseudonyms
- Makes fraud detection hard for regulators
- All transaction data publicly available
- Dire need for analysis of the data for link prediction and classification
- Analysis of privacy by user identification

Part - I

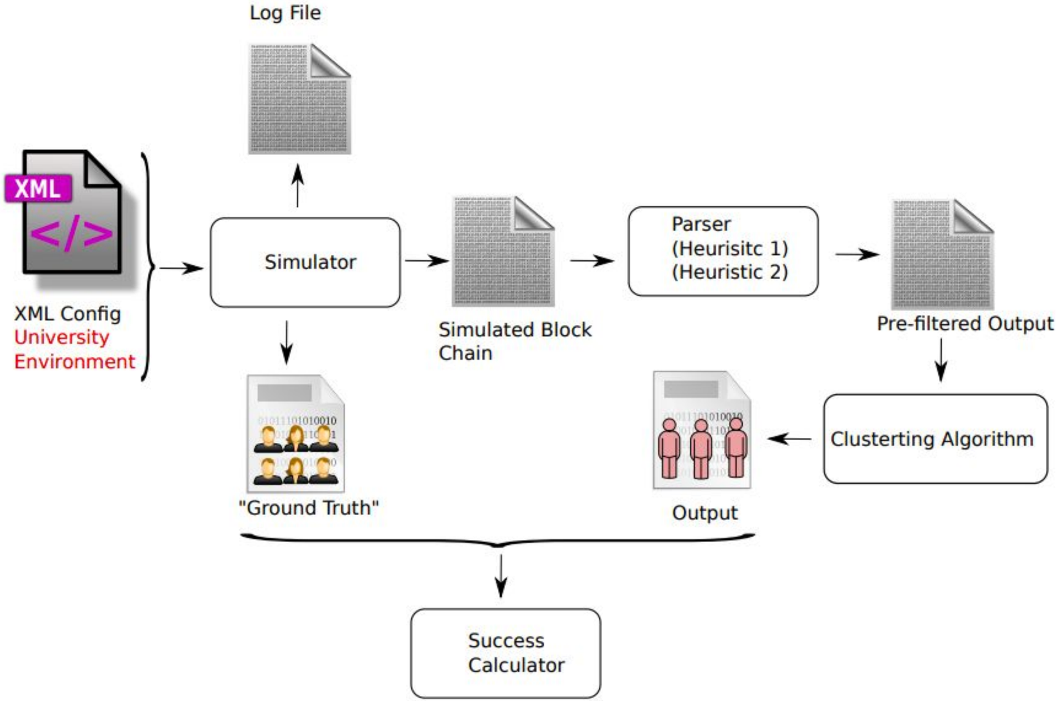
Research on user identification problem

Related Work

- Works related to statistical characterization of Bitcoin transaction logs
- Mainly focusing on the problem of user identification

Related Work

- Simulator mimics the use of Bitcoin as a primary currency to support daily transactions in a university
- Results showed that almost 40% user profiles could be recovered



Evaluating User Privacy in Bitcoin, Elli Androulaki et al

Heuristics

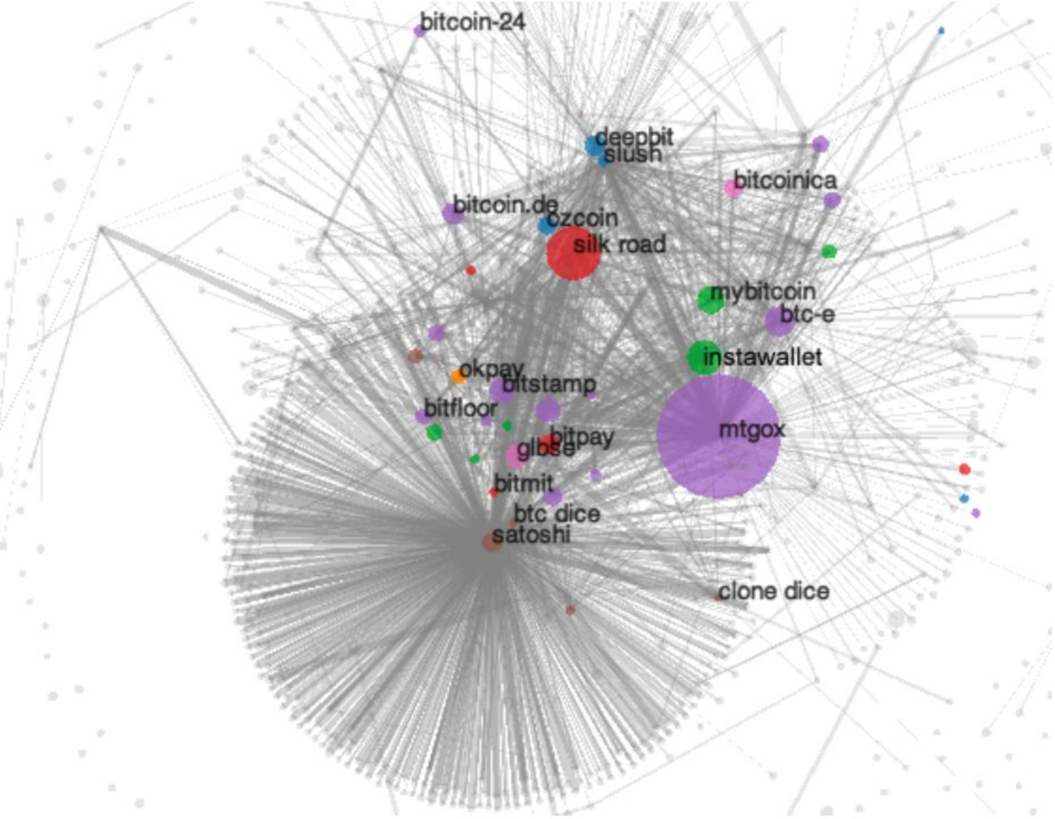
I - Common Spending

If two or more addresses are inputs of the same transaction with one output, then all these addresses are controlled by the same user. This heuristic is expected to be very accurate as common spending by different users should be based on high degree of trust between them

Evaluating User Privacy in Bitcoin, Elli Androulaki et al

Related Work

- Heuristics to perform address clustering on tagged data
- Tagged users by simply transacting with them (e.g., depositing into and withdrawing bitcoins from Mt. Gox) and then observing the addresses they used
- Known (or assumed) addresses from various forums and websites,.



A Fistful of Bitcoins: Characterizing Payments Among Men with No Names, Sarah Meiklejohn et al, 2013

Heuristics

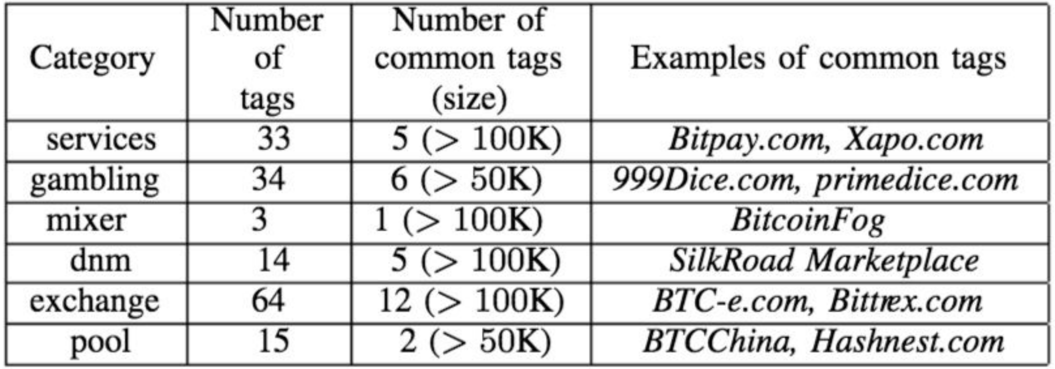
II - One time Change (OTC)

It is based on the standard Bitcoin mechanism where the change from the transaction is returned to a new address. If the transaction satisfies the conditions of a one-time change transaction, then the OTC output and all the inputs of the transaction are controlled by the same user.

A Fistful of Bitcoins: Characterizing Payments Among Men with No Names, Sarah Meiklejohn et al, 2013

Related Work

- Off-chain information (tags) show contradictory tags in clusters created by common behavior pattern analysis (heuristics mentioned earlier)
- Tags used as votes against address union using a probabilistic clustering model



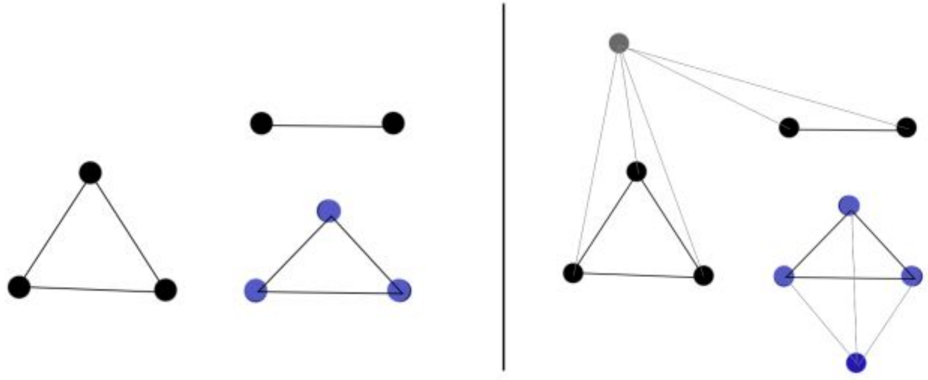
Category	Number of tags	Number of common tags (size)	Examples of common tags
services	33	5 (> 100K)	*Bitpay.com, Xapo.com*
gambling	34	6 (> 50K)	*999Dice.com, primedice.com*
mixer	3	1 (> 100K)	*BitcoinFog*
dnm	14	5 (> 100K)	*SilkRoad Marketplace*
exchange	64	12 (> 100K)	*BTC-e.com, Bittrex.com*
pool	15	2 (> 50K)	*BTCChina, Hashnest.com*

TABLE IV: Tags of the biggest cluster in case of clustering without constraints.

Automatic Bitcoin Address Clustering, Dmitry Ermilov et al,

Related Work

- Network attributes have information like script types, relaying IPs, timestamps, etc
- Constructed a hybrid structure-attribute graph
- Clustering using HDBSCAN
- Used chainalysis data as ground truth. Shows good results in the test cases analyzed in the paper

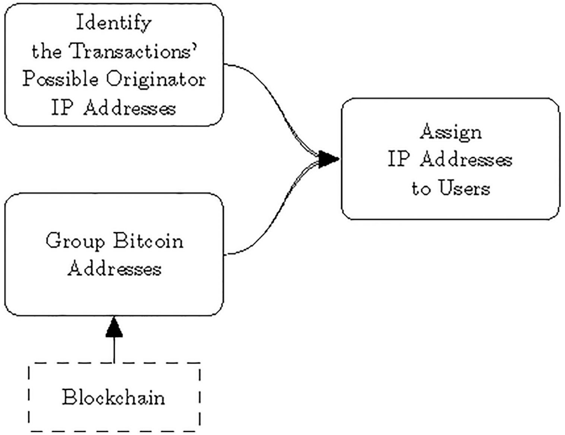


Figures 5, 6: In the left image, we have the original graph. Beyond its structure, it has various attributional qualities that should also be taken into account (represented by the colors). On the right, we create points to represent the 'black' and 'blue' attributes and draw the appropriate edges into the graph. Now, the colors no longer matter – we have represented that information into the structure itself!

De-Anonymizing the Bitcoin Blockchain, Bharath Srivatsan et al. 2016

Related Work

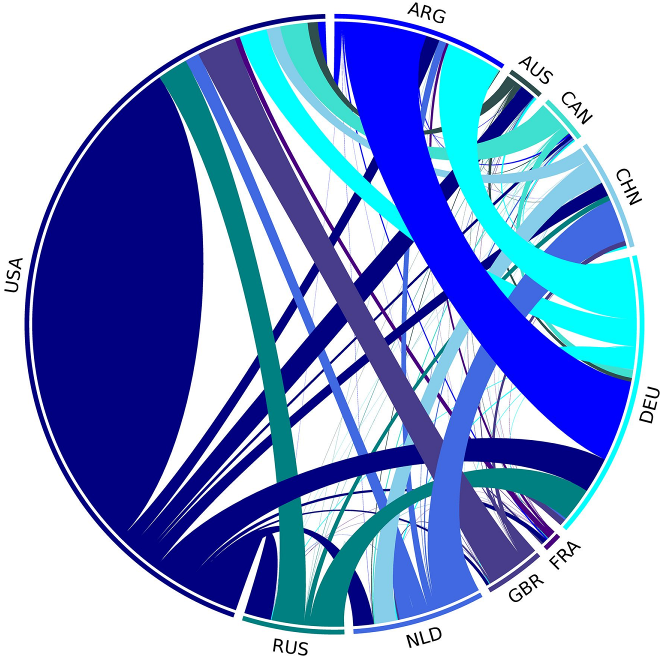
- Installed modified Bitcoin clients distributed in the network to observe as many bitcoin transaction messages as possible.
- Assigned probabilities of originator IP addresses to messages observed by the clients based on the relay time
- Combined these with address clustering heuristics to get final probability assignments for user identification


Identify the Transactions' Possible Originator IP Addresses
Group Bitcoin Addresses
Blockchain
Assign IP Addresses to Users

A Bayesian approach to identify Bitcoin users, Péter L. Juhász, József Stéger, 2018

Related Work

- Traced geographic location from originating IP addresses
- Were able to identify several thousand Bitcoin clients and bind their transactions to geographical locations during a two month observation period.



The flow of Bitcoin between countries

A Bayesian approach to identify Bitcoin users, Péter L. Juhász, József Stéger, 2018

Part-II

Data Visualization API for link prediction

Related Work

- Works that have successfully used network representation learning approaches on Bitcoin data for purposes like community detection and price prediction

Related Work

- Refined address linking heuristics by also contracting output side addresses

- Following clustering techniques were used on the contracted transaction network
  - K-means on handpicked features
  - K-means on vector embeddings generated by Node2Vec
  - Spectral clustering by using the Fiedler vector method

- Different communities and user categories were identified from all the methods

Related Work

- Representation learning approach based on the geometric structure and topological features of the Bitcoin transaction graph
- Betti numbers and sequences generated from the graph in a predictive learning model
- 40% improvement from previous baselines in 7-day ahead price prediction

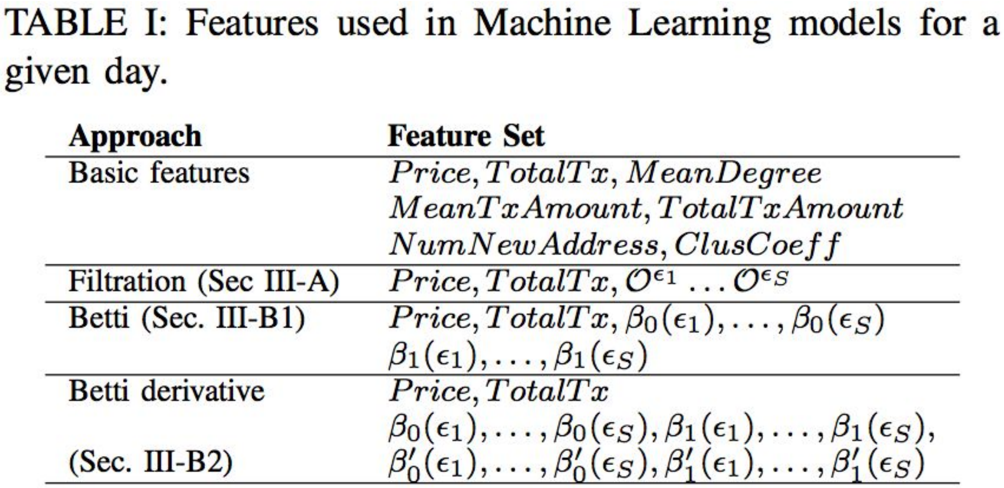


TABLE I: Features used in Machine Learning models for a given day.

Approach	Feature Set
Basic features	$Price, TotalTx, MeanDegree$ $MeanTxAmount, TotalTxAmount$ $NumNewAddress, ClusCoeff$
Filtration (Sec III-A)	$Price, TotalTx, \mathcal{O}^{\epsilon_1} \ldots \mathcal{O}^{\epsilon_S}$
Betti (Sec. III-B1)	$Price, TotalTx, \beta_0(\epsilon_1), \ldots, \beta_0(\epsilon_S)$ $\beta_1(\epsilon_1), \ldots, \beta_1(\epsilon_S)$
Betti derivative (Sec. III-B2)	$Price, TotalTx$ $\beta_0(\epsilon_1), \ldots, \beta_0(\epsilon_S), \beta_1(\epsilon_1), \ldots, \beta_1(\epsilon_S),$ $\beta_0'(\epsilon_1), \ldots, \beta_0'(\epsilon_S), \beta_1'(\epsilon_1), \ldots, \beta_1'(\epsilon_S)$

ChainNet: Learning on Blockchain Graphs with Topological Features, Nazmiye Ceren Abay et al

Proposed Methodology

Use network representation learning techniques on the bitcoin transaction graph to solve problems like -

- Link Prediction
- User Identification

Proposed Methodology

Various models can be used according to the graph properties-

- Dynamic - Online Node2Vec
- Sparse - DeepWalk
- Behavioral/Structural similarity - Struct2Vec

Problem

- No real ground truth for Bitcoin user data
- All previous results are essentially unreliable
- Hard to explain the results of any study on Bitcoin analysis
- Need for a data visualization API that makes it possible to observe and analyze subgraphs and node activities for explanation purposes

Revised Objective

Create bitcoin data visualizations that can help get intuitive explanations for the results from AI models.

Display

1. Node statistics - (Num transactions, In degree, Out degree, First Active, Last Active)
2. Most recent transactions (with ability to look at a particular subset of transaction ids)
3. Temporal activity of the node
4. Graphical visualization of the recent transactions
5. Vector embedding of the node(s)

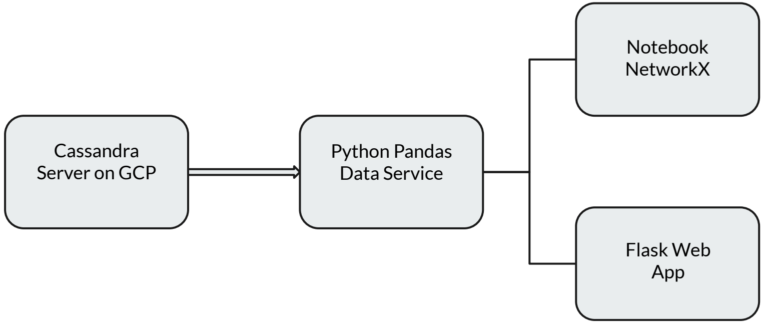
Dataset

- Bitcoin network dataset - D Kondor et al
  https://senseable2015-6.mit.edu/bitcoin/
  - Bh.dat -- Block id, timestamp
  - Th.dat -- transaction id, block id
  - Txin.dat -- transaction id, input addresses
  - Txout.dat -- transaction id, output addresses

- Bitcoin Whos Who - Potential ground truth
  https://bitcoinwhoswho.com/

Framework

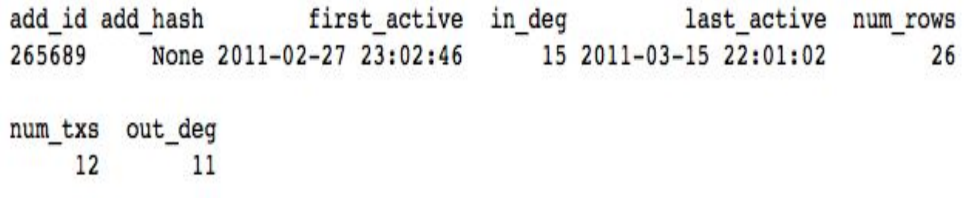
- Google Cloud Platform
- Cassandra 3.x
- Python Pandas Dataframe
- NetworkX
- Flask



Cassandra Tables

Node Info:

- add_id : Integer node id
- first_active : Timestamp of first transaction
- last_active : Timestamp of last transaction
- in_deg : Total number of incoming edges
- out_deg : Total number of outgoing edges
- num_txs : Number of unique transactions



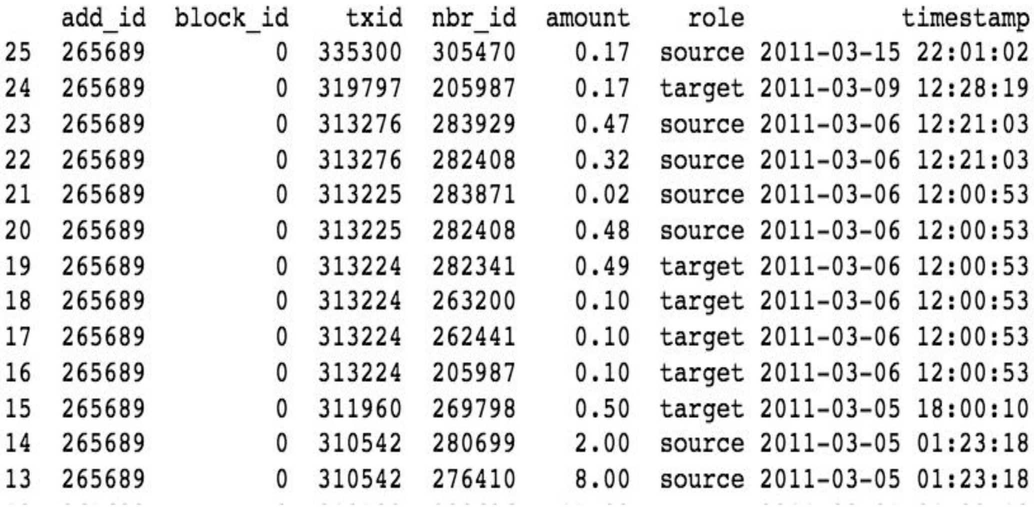
```
add_id add_hash         first_active  in_deg          last_active  num_rows
265689      None 2011-02-27 23:02:46      15 2011-03-15 22:01:02         26

num_txs  out_deg
     12       11
```

Cassandra Tables

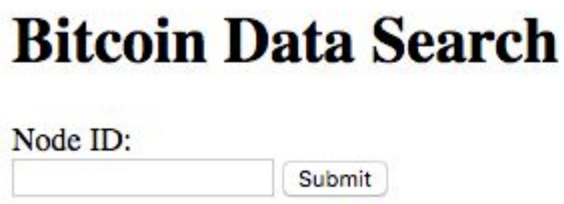
Transactions:

- add_id : Integer node id
- block_id : block number for cassandra partitioning
- txid : Unique transaction id
- role : Role of the add_id in the given transaction
- nbr_id : Node ids on other end of transaction
- timestamp : Time of the transaction



```
    add_id  block_id    txid  nbr_id  amount   role            timestamp
25  265689         0  335300  305470    0.17 source  2011-03-15 22:01:02
24  265689         0  319797  205987    0.17 target  2011-03-09 12:28:19
23  265689         0  313276  283929    0.47 source  2011-03-06 12:21:03
22  265689         0  313276  282408    0.32 source  2011-03-06 12:21:03
21  265689         0  313225  283871    0.02 source  2011-03-06 12:00:53
20  265689         0  313225  282408    0.48 source  2011-03-06 12:00:53
19  265689         0  313224  282341    0.49 target  2011-03-06 12:00:53
18  265689         0  313224  263200    0.10 target  2011-03-06 12:00:53
17  265689         0  313224  262441    0.10 target  2011-03-06 12:00:53
16  265689         0  313224  205987    0.10 target  2011-03-06 12:00:53
15  265689         0  311960  269798    0.50 target  2011-03-05 18:00:10
14  265689         0  310542  280699    2.00 source  2011-03-05 01:23:18
13  265689         0  310542  276410    8.00 source  2011-03-05 01:23:18
```

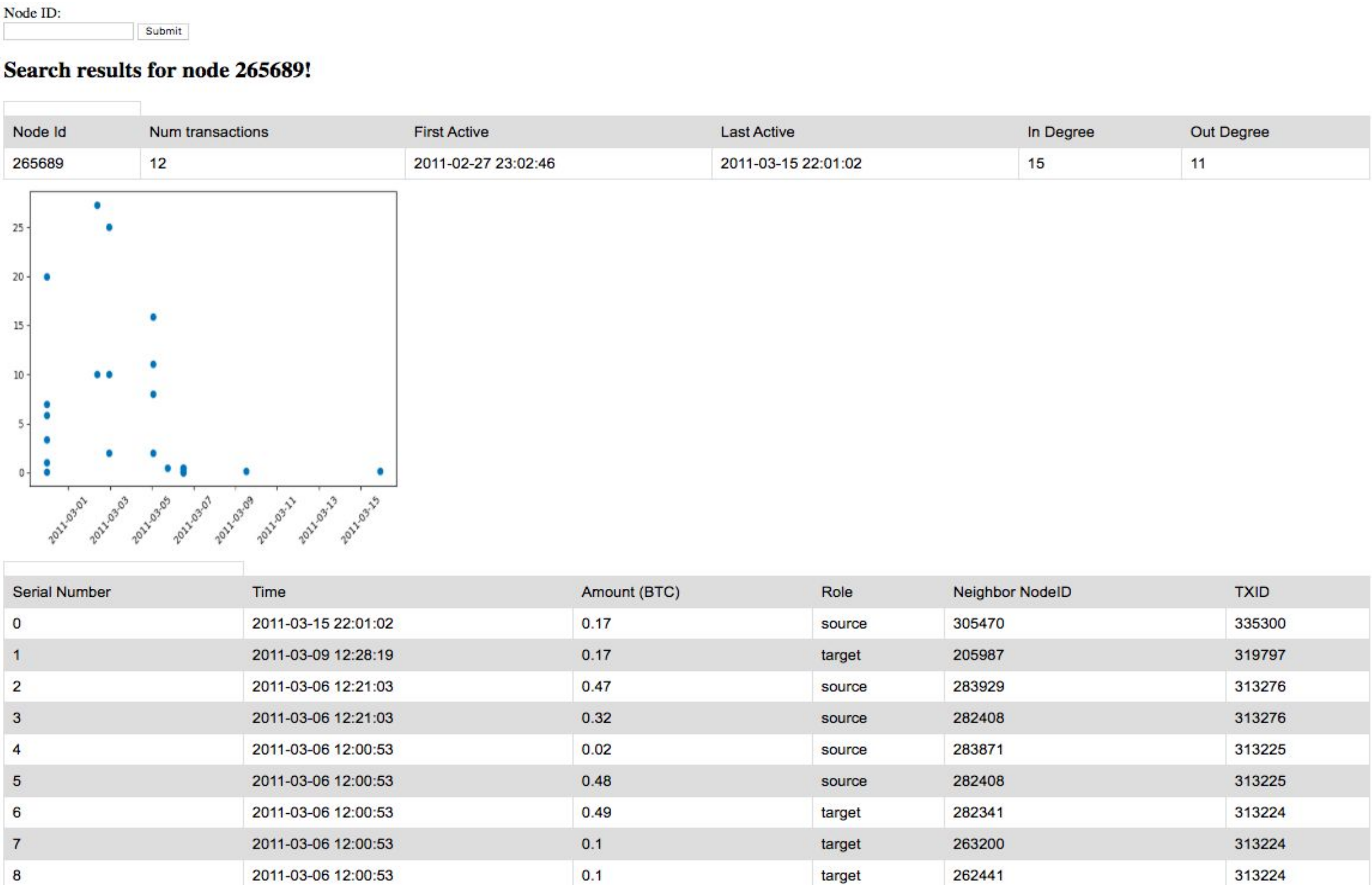
GUI



Bitcoin Data Search

Node ID:

GUI

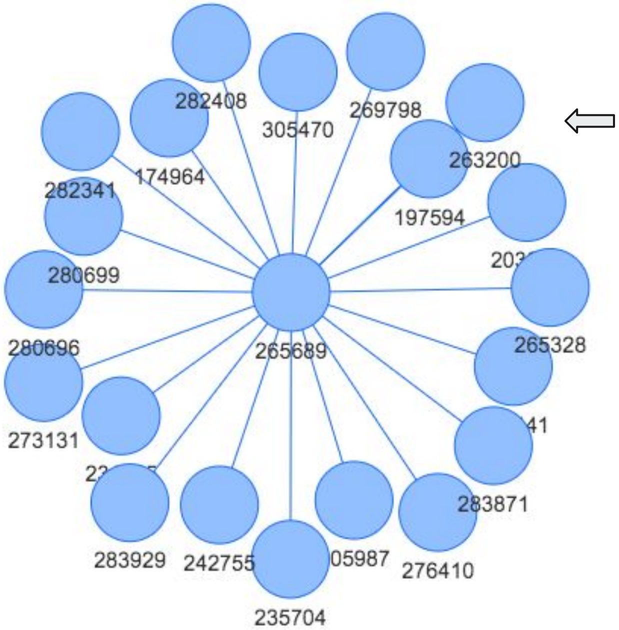


Search results for node 265689!

Node Id	Num transactions	First Active	Last Active	In Degree	Out Degree
265689	12	2011-02-27 23:02:46	2011-03-15 22:01:02	15	11

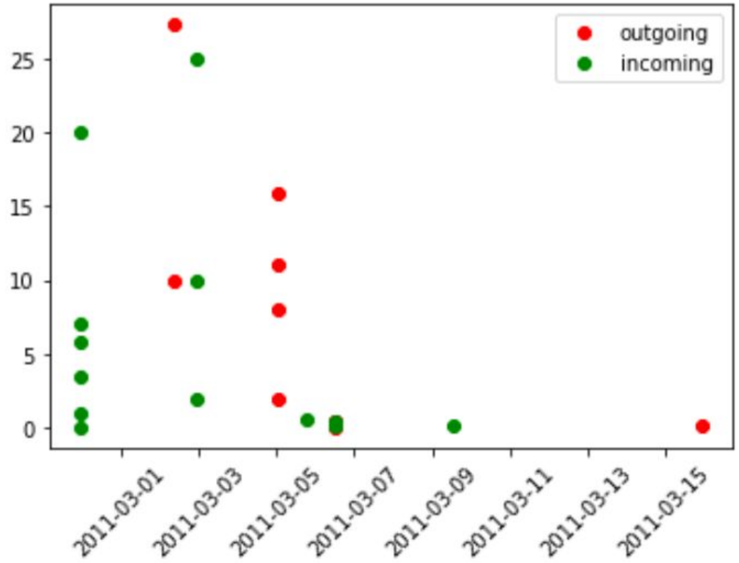
Serial Number	Time	Amount (BTC)	Role	Neighbor NodeID	TXID
0	2011-03-15 22:01:02	0.17	source	305470	335300
1	2011-03-09 12:28:19	0.17	target	205987	319797
2	2011-03-06 12:21:03	0.47	source	283929	313276
3	2011-03-06 12:21:03	0.32	source	282408	313276
4	2011-03-06 12:00:53	0.02	source	283871	313225
5	2011-03-06 12:00:53	0.48	source	282408	313225
6	2011-03-06 12:00:53	0.49	target	282341	313224
7	2011-03-06 12:00:53	0.1	target	263200	313224
8	2011-03-06 12:00:53	0.1	target	262441	313224

GUI



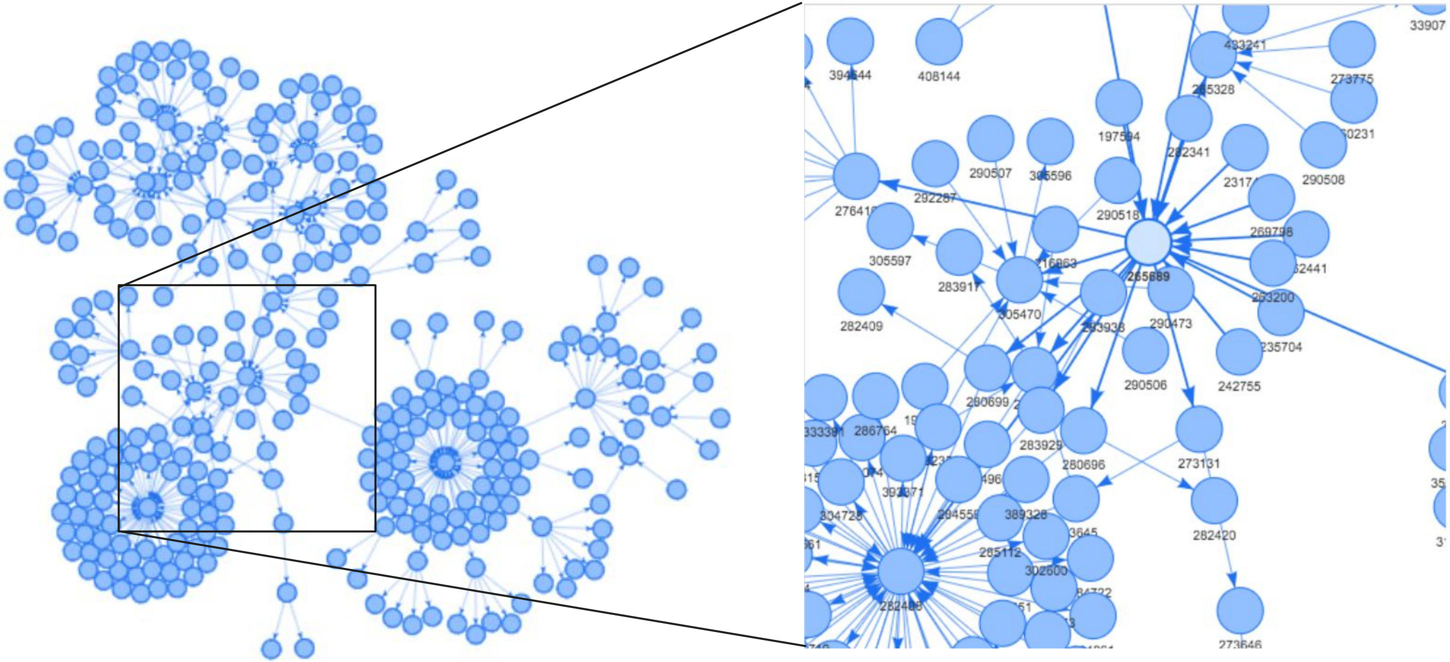
Graphical view of recent transactions

Temporal view of all transactions



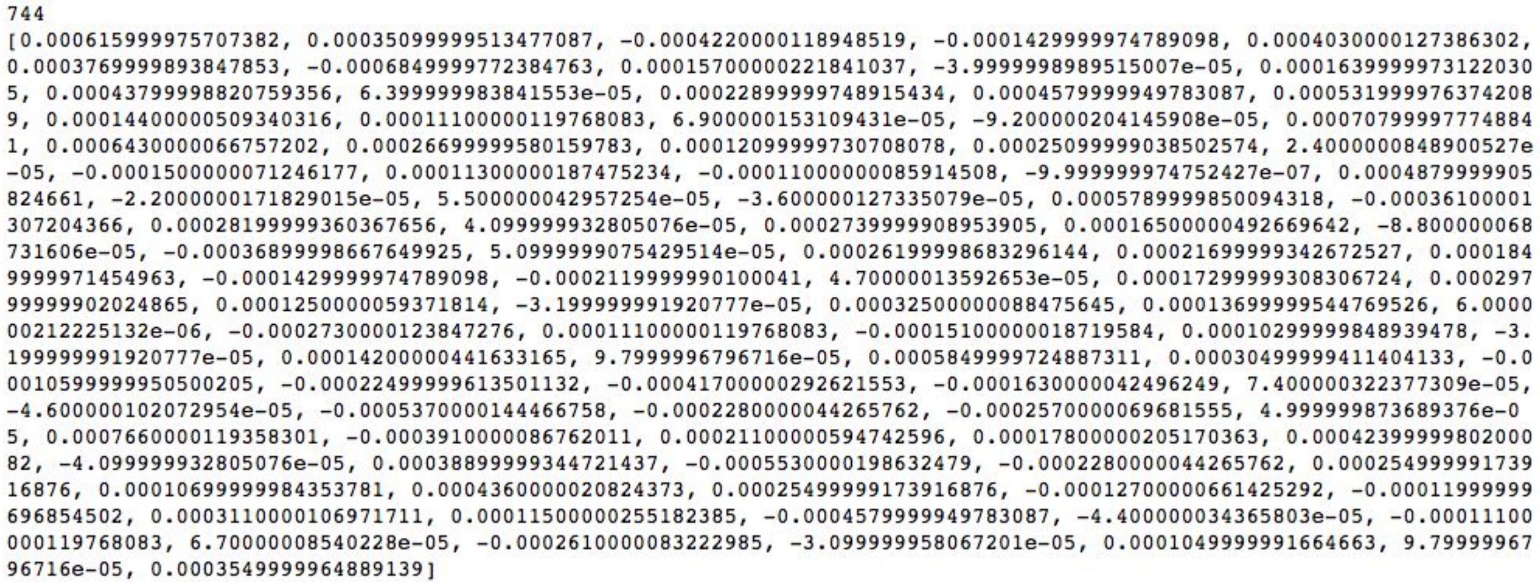
GUI

K-step information (k=3)



GUI

Sample vector embedding of a node



```
744
[0.0006159999975707382, 0.0003509999951347708 7, -0.000422000011894851 9, -0.000142999997478909 8, 0.000403000012738630 2,
```

Future Work

- Utilize data visualization API to explain results of the Link Prediction model
- Expand the database with a dynamic data ingestion pipeline

Conclusion

- Hard to solve the problem of bitcoin user identification as there is no real ground truth. Most previous results are unreliable.
- Good data visualization plays a significant role in making sense of AI models.

References

- Nakamoto S. Bitcoin: A peer-to-peer electronic cash system," http://bitcoin.org/bitcoin.pdf; 2009
- Androulaki E, Karame GO, Roeschlin M, Scherer T, Capkun S. Evaluating User Privacy in Bitcoin. In: Sadeghi AR, editor. Financial Cryptography and Data Security. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 34–51
- Meiklejohn S, Pomarole M, Jordan G, Levchenko K, McCoy D, Voelker GM, et al. A Fistful of Bitcoins: Characterizing Payments Among Men with No Names. Commun ACM. 2016;59(4):86–93
- D. Ermilov, M. Panov, Y. Yanovich, "Automatic Bitcoin address clustering", Proc. Int. Conf. Mach. Learn. Appl. (ICMLA), pp. 461-466, Dec. 2017
- Bharath Srivatsan. 2016. De-Anonymizing the Bitcoin Blockchain. (2016)
- Juhász PL, Stéger J, Kondor D, Vattay G (2018) A Bayesian approach to identify Bitcoin users. PLoS ONE 13(12): e0207000
- D. Mao, Y. Zhang, User Categorization and Community Detection in Bitcoin Network, snap.stanford.edu
- Nazmiye C. Abay et al. "ChainNet: Learning on Blockchain Graphs with Topological Features". In: (2019)