April 2018

Deanonymization of the Bitcoin System

Hongjie Chen
Chongyao Xia

Content

- Background
- Existing Work
- Our work
- Reference

Background

- Basic concepts
- Important relationship
- Bitcoin transaction
- P2P networks
- Bitcoin deanonymization

Background - basic concepts

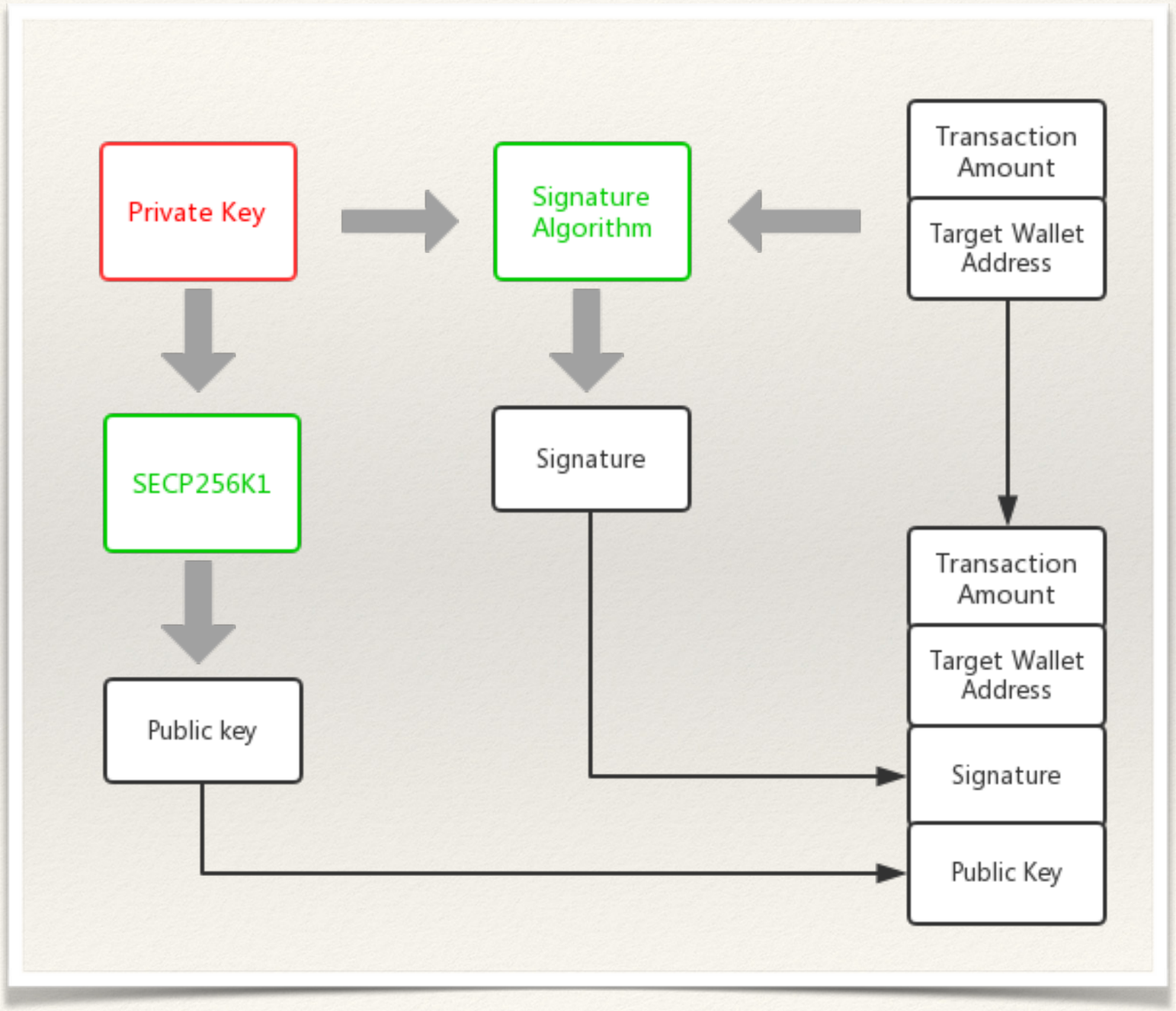
- **Private Key**: Random 256 bits generated by the bitcoin algorithm, only known to yourself. Private key can be regarded as users' account.
- **Public Key**: 512 bits generated by the private key, but it can't be converted to the corresponding private key.
- **Message**: A typical data form consisting of the details of a transaction.
- **Wallet Address**: A random-length data generated by public address used for others to send bitcoins to the corresponding account.
- **Signature**: 512 bits generated by the message and private key to give authorization to this particular transaction.

Background - important relationship

Private key is all that matters to you!

Background - bitcoin transaction

Private key plays a key roll in the transaction like your right hand ready to sign a contract!



Background - bitcoin transaction

Height	Age	Transactions	Total Sent	Relayed By	Size (kB)	Weight (kWU)
524318	6 minutes	577	3,977.44 BTC	F2Pool	376.73	1,309.85
524317	9 minutes	1375	5,716.06 BTC	Unknown	855.88	3,043.88
524313	22 minutes	430	428.63 BTC	SlushPool	1,262.89	3,992.99
524311	29 minutes	848	1,941.06 BTC	BitClub Network	1,200.73	3,992.73

Height	Time	Relayed By	Hash	Size (kB)
524318 (Main Chain)	2018-05-25 11:49:17	F2Pool	00000000000000002abe57e410d5bb654980080ee63fb1c075852b0d529e9d	376.73
524317 (Main Chain)	2018-05-25 11:46:02	BTC.TOP	00000000000000000010e7bc57d4cc9999cea5c88391ceb3b5b4cdebbc58d4c2	855.88
524316 (Main Chain)	2018-05-25 11:38:18	SlushPool	00000000000000000000172298ab32f10ddc0029b8f60b0a234b1bf15bbdf58dc9	983.54
524315 (Main Chain)	2018-05-25 11:37:22	BTCC Pool	00000000000000000000b13e1b7992633bfe35525d1ab5c5a4dabc1b6968a0432	1,013.33
524314 (Main Chain)	2018-05-25 11:34:19	Unknown	00000000000000000003fbaf8b31af3b7d0d7d3665ff70c8d5949357f11c45f52	1,129.95
524313 (Main Chain)	2018-05-25 11:32:57	SlushPool	00000000000000000027a3c07fa674be944c784d3c87de9524cea45725161ecd	1,262.89
524312 (Main Chain)	2018-05-25 11:31:25	BTC.com	00000000000000000000a6750a4678d7d5f0317b71852166676773121e777c4b0	1,182.37
524311 (Main Chain)	2018-05-25 11:26:17	BitClub Network	00000000000000000017f1501a85c0e633210159abb643110e325769f881378b	1,200.73

A glimpse of recently produced blocks

Background - bitcoin transaction

	addr_ID	user_ID
100	101	101
101	102	102
102	103	103
103	104	104
104	105	105
105	106	106
106	107	107
107	108	108
108	109	109
109	110	110

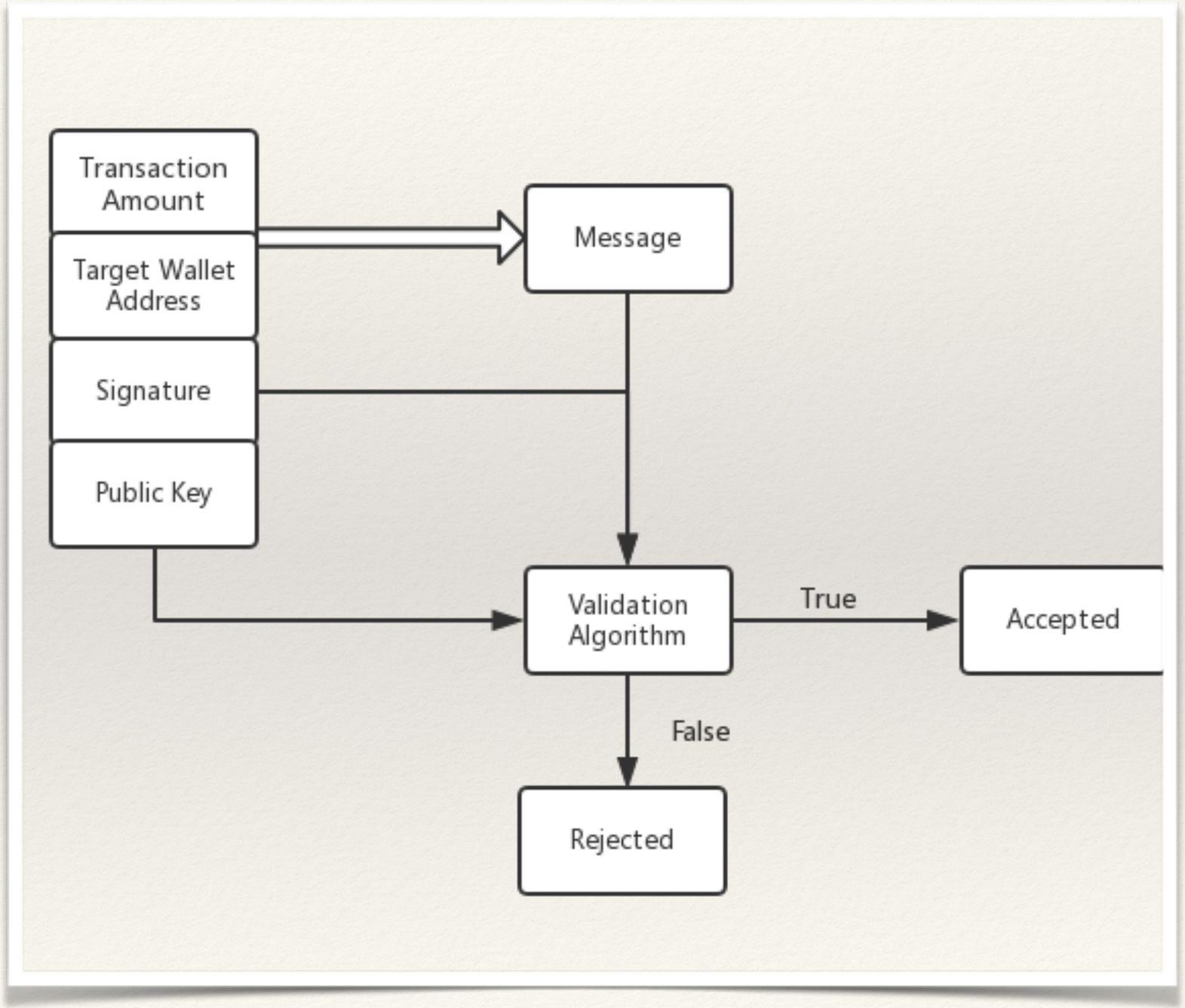
	addr_ID	user_ID
900	901	98
901	902	902
902	903	903
903	904	904
904	905	905
905	906	906
906	907	64
907	908	908
908	909	783
909	910	910

	addr_ID	user_ID
12900	12901	8704
12901	12902	12902
12902	12903	12903
12903	12904	12904
12904	12905	12905
12905	12906	979
12906	12907	64
12907	12908	4378
12908	12909	12909
12909	12910	1998

Three snapshots of results of heuristic clustering. The first column is address ID. The second column is the user ID.

Background - P2P Networks

- The validation work is done by “miners”.
- The one who notified you the transaction message may be an intermediary in the P2P network, not the payer.
- The validation work of the decentralized system makes miners important.



Background - deanonymization

- Anonymity = pseudonymity + unlinkability
  - Different interactions of the same user with the system should not be linkable to each other
- Unlinkability is bitcoin system
  - Hard to link different addresses of the same user
  - Hard to link different transactions of the same user
  - Hard to link sender of a payment to its recipient

Background - deanonymization

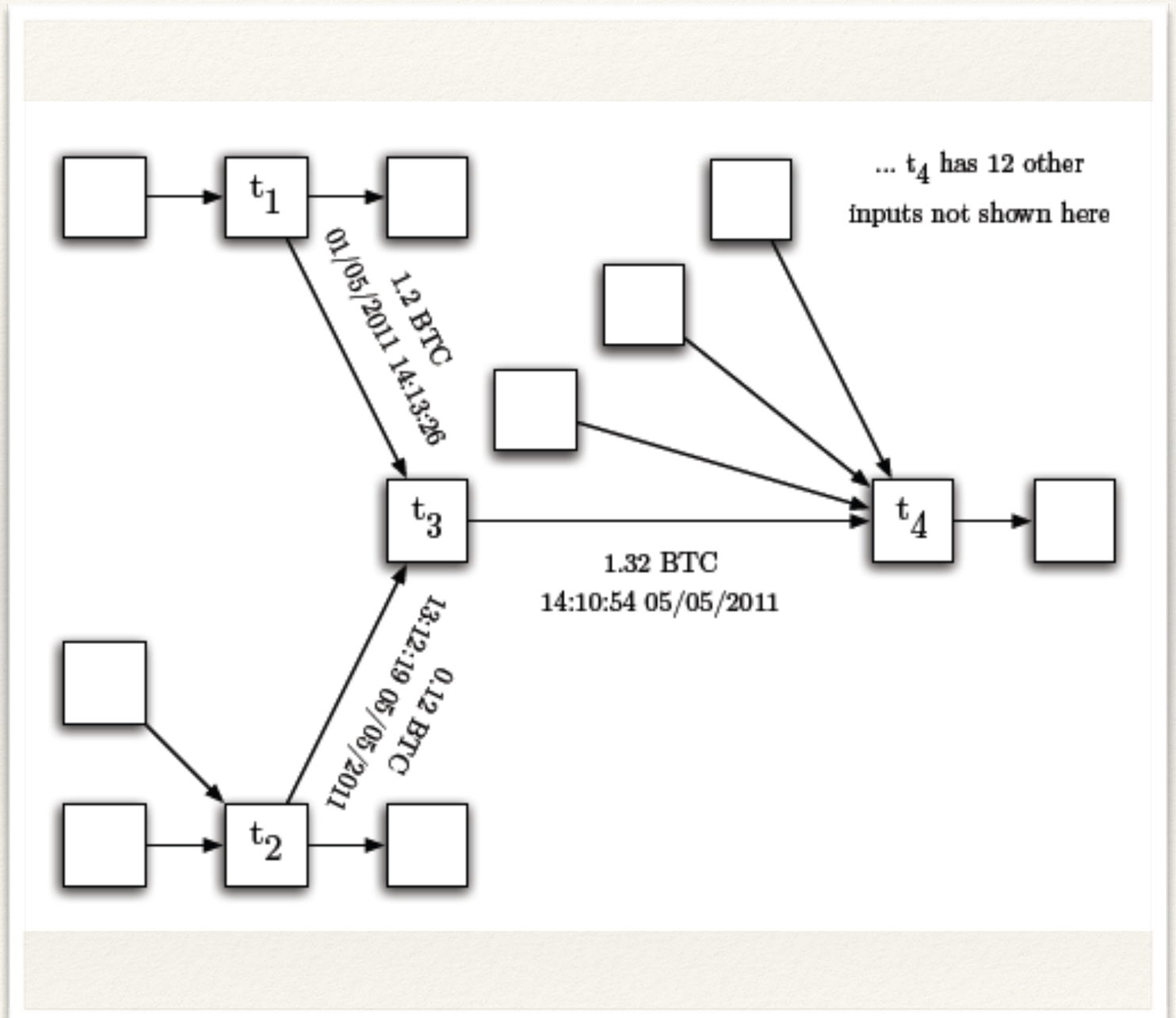
- **Clustering of the Public Keys**
  - A user may possess multiple public keys, which makes it important to link the different public keys belonging to the same user together.
- **IP Address**
  - Link the public key of a certain transaction to the IP address which initiates it.
- **Exact Personal Profile**
  - Link the public key to a specific user with his self-profile, such as accounts of social website

Existing work

- 3 ways to model bitcoin transaction data
  - Transaction network
  - Ancillary network
  - User network

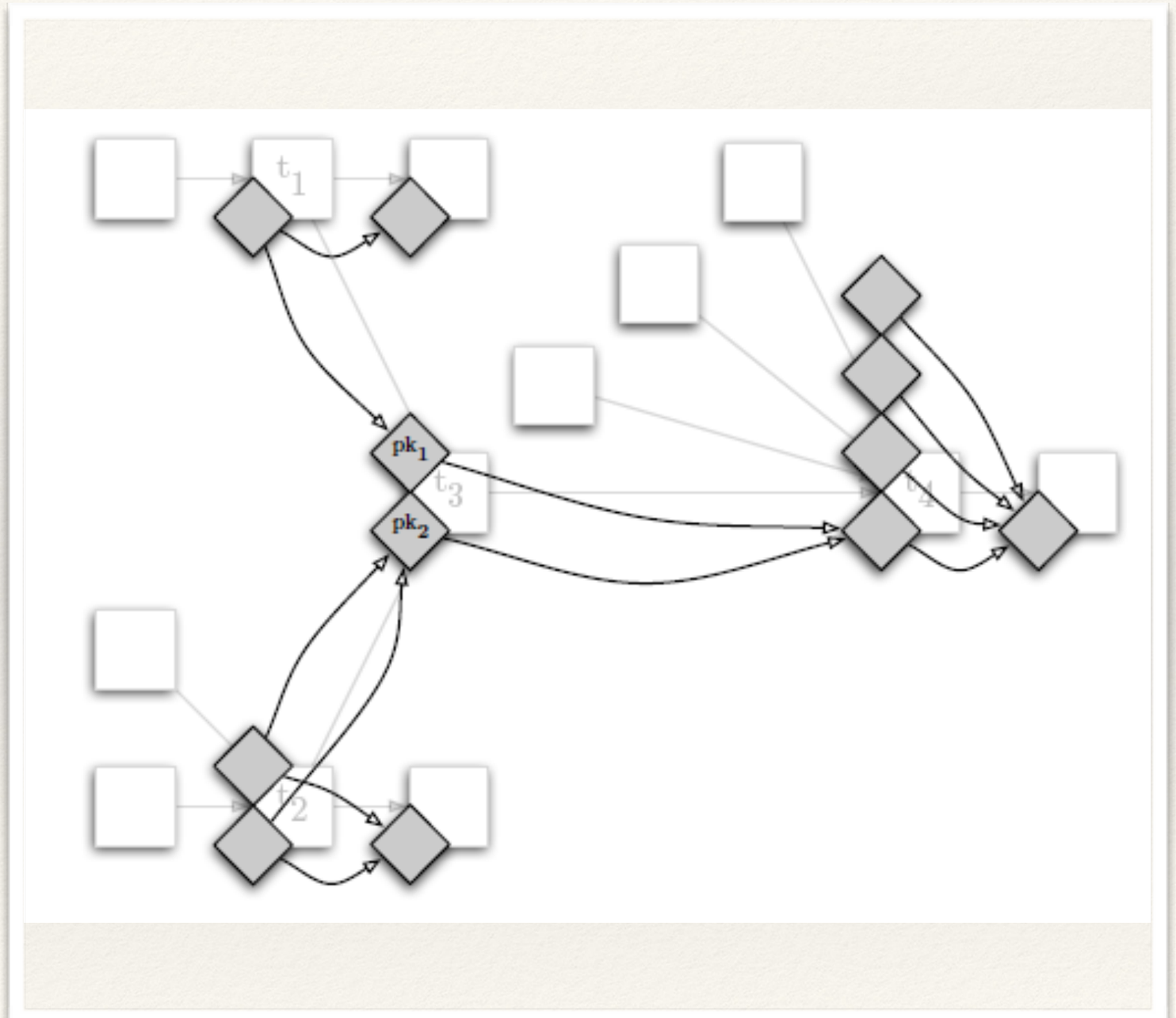
Existing work - transaction network

- **Node**: each transaction in the bitcoin systems
- **Edge**: bitcoin flow in the network
- **Explanation**: the output of one transaction is the input of another



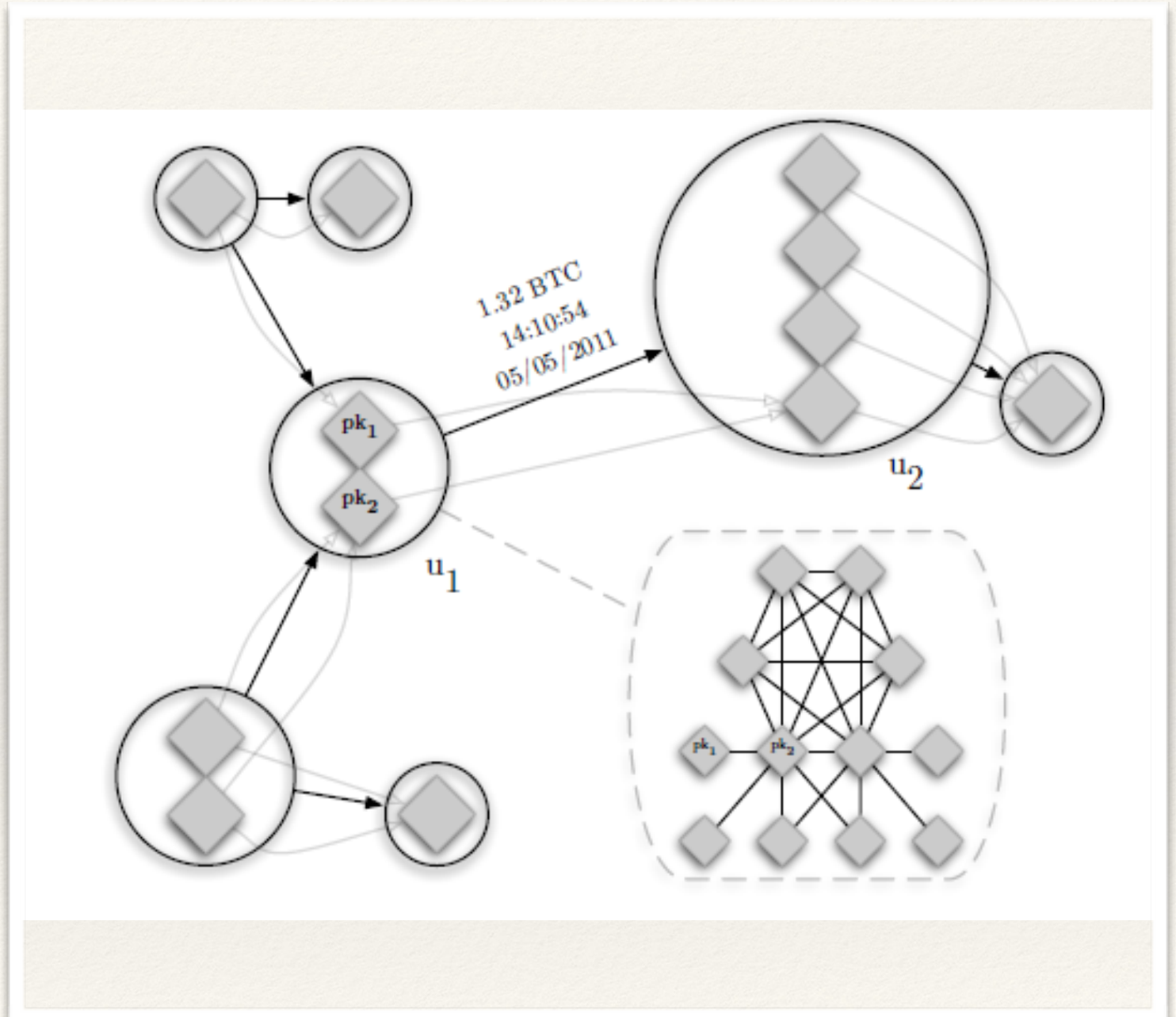
Existing work - ancillary network

- **Node**: each public key in the bitcoin systems
- **Edge**: bitcoin flow in the network
- **Explanation**: pk1 and pk2 serves as the input to another in the same time period, which shows it is very likely that the two public keys belongs to the same user.



Existing work - user network

- **Node**: each user in the bitcoin systems
- **Edge**: bitcoin flow in the network
- **Explanation**: A cluster of public keys is achieved and represented in the user network form



- **Caveat**:
  - Transaction network and ancillary network can be directly derived from bitcoin transaction data.
  - However, user network must be obtained by application of clustering techniques w.r.t nodes (i.e. public keys) in the ancillary network, which is just the core of deanonymization of bitcoins systems.

Existing work - deanonymize bitcoin

- Bitcoin system can be further deanonymized by utilizing leaked users' information, such as public keys they posted on internet.

Our Work - overview

- Learn basics of bitcoin and blockchain
- Collect bitcoin transaction data
- Process collected data
- Design methods
- Do experiments
- Write reports

Our Work - data

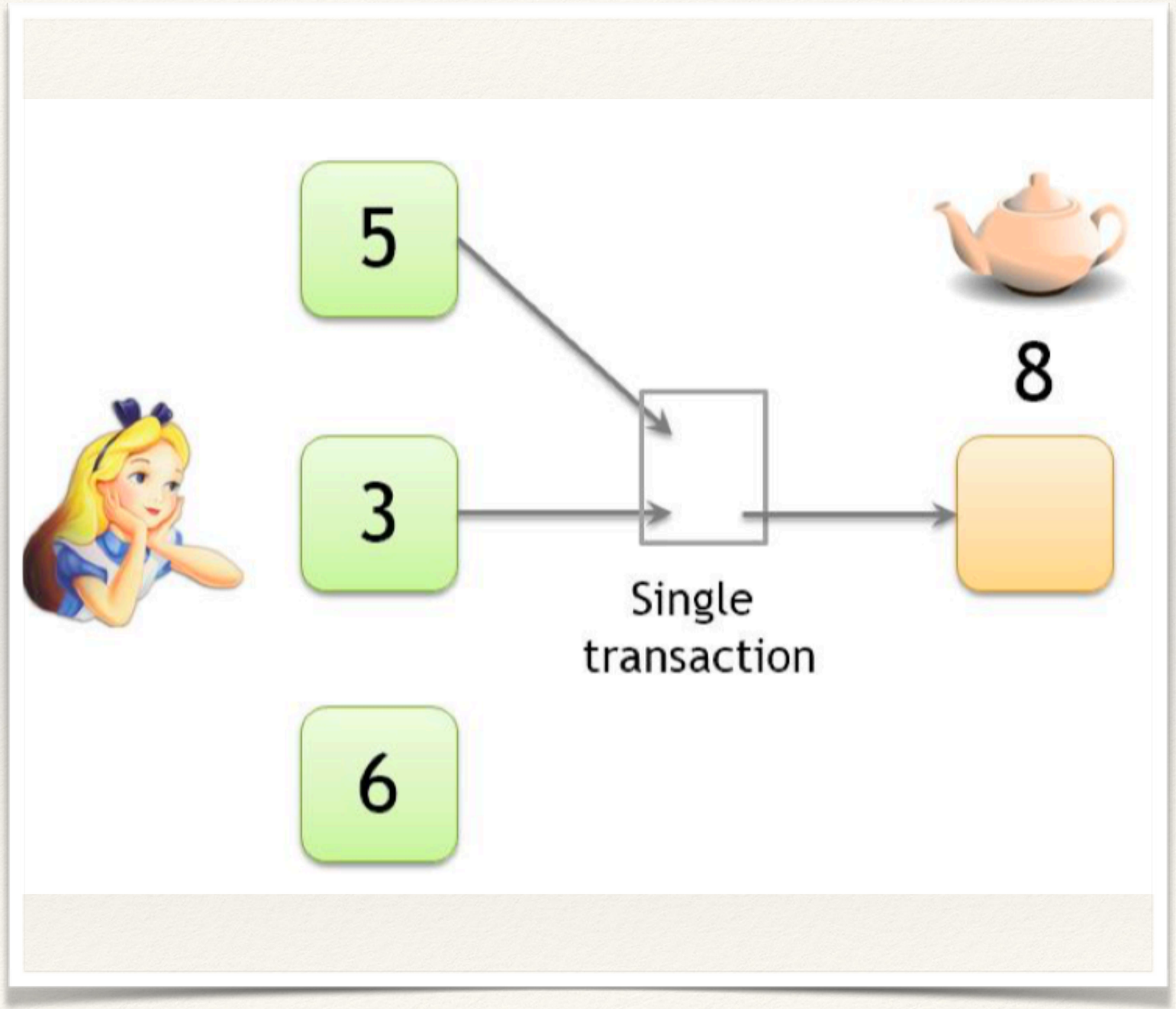
- Whole blockchain up to 2016.02.09. (397,571 blocks).
  - **enumeration of all blocks in the blockchain**, 277443 rows, 4 columns:
    - id used in this database (0 -- 277442, continuous)
    - block hash (identifier in the blockchain, 64 hex characters)
    - creation time (from the blockchain)
    - number of transactions
  - **transaction ID and hash pairs**, 30048983 rows, 2 columns:
    - id used in this database (0 -- 30048982, continuous)
    - transaction hash used in the blockchain (64 hex characters)
  - **BitCoin address IDs**, 24618959 rows, 2 columns:
    - id used in this database (0 -- 24618958, continuous, the address with addrID == 0 is invalid /blank, not used/)
    - string representation of the address (alphanumeric, maximum 35 characters; note that the IDs are NOT ordered by the addr in any way)
  - **enumeration of all transactions**, 30048983 rows, 4 columns:
    - transaction ID (from the txhash.txt file)
    - block ID (from the blockhash.txt file)
    - number of inputs
    - number of outputs

Our Work - data

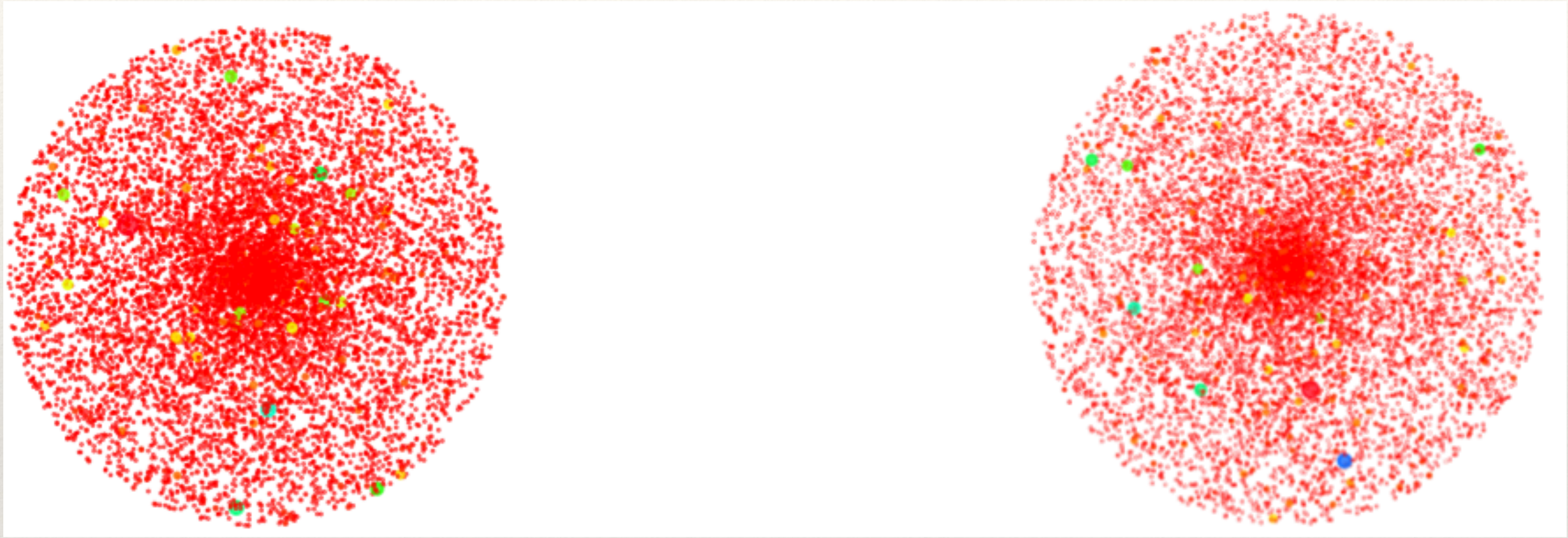
- Whole blockchain up to 2016.02.09. (397,571 blocks).
  - **list of all transaction inputs** (sums sent by the users), 65714232 rows, 3 columns:
    - transaction ID (from the txhash.txt file)
    - sending address (from the addresses.txt file)
    - sum in Satoshis (1e-8 BTC -- note that the value can be over 2^32, use 64-bit integers when parsing)
  - **list of all transaction outputs** (sums received by the users), 73738345 rows, 3 columns:
    - transaction ID (from the txhash.txt file)
    - receiving address (from the addresses.txt file)
    - sum in Satoshis (1e-8 BTC -- note that the value can be over 2^32, use 64-bit integers when parsing)
  - **transaction timestamps** (obtained from the blockchain.info site), 30048983 rows, 2 columns:
    - transaction ID (from the txhash.txt file)
    - unix timestamp (seconds since 1970-01-01)

Our Work - heuristic clustering

- **Heuristic** : shared spending is evidence of joint control of the different input addresses.
- In this case, we can cluster the different addresses described above.



Our Work - heuristic clustering



Left: In this graph, each circle represents a user. And the area of a circle positively proportionally reflects the number of addresses a user owns. From this graph, we can clearly see that most users own just a small number of address, while only few users own a large number of addresses.

Right: In this graph, each circle represents an address. And the area of a circle positively proportionally reflects the number of transactions an address participate. From this graph, we can clearly see that most addresses participate just a small number of address, while only few addresses take part in a large number of transactions.

Our Work - heuristic clustering

	addr_ID	addr
100	100	111ccCf3YzzcXH6G15mukMBQ8rcmo1qCU
101	101	111CH4CEu1PkTMpouRKDTK3yZPHAxz8Vv
102	102	111cphrV8LtixDfWq7HbELtehH5UgRqiA
103	103	111cZqKGQzMyEaPNajPmgGaHS7U9vRSWd
104	104	111D7xaZHpAnJdwKknTXmZzWc6Sw8uwzH
105	105	111Da3uc98pipSvS3mEjNbU12WKs578th
106	106	111DADVu85myVnJ53CzZgB9kapsHwDxW2
107	107	111dDDCBvC598bkNC9qiGPDPnJ1tdBGDe
108	108	111dentifieron1yLettersXXXXasmS9N
109	109	111dentifiersA1waysHaveToXXZPZVdN

	trans_ID	addr_ID	val
8800	8773	818085	5000000000
8801	8774	2960363	5000000000
8802	8775	4440252	5000000000
8803	8776	588968	5000000000
8804	8777	6470175	5000000000
8805	8778	1108046	5000000000
8806	8779	2927540	5000000000
8807	8780	2077512	5000000000
8808	8781	1101897	5000000000
8809	8782	878896	5000000000

	trans_ID	addr_ID	val
9100	50563	3418377	4900000000
9101	50564	1008527	80000000000
9102	50564	2518235	236155000000
9103	50564	4990542	72000000
9104	50564	6008494	16000000000
9105	50565	636952	1000000
9106	50568	5211481	1000000
9107	50569	5554320	332160000000
9108	50570	2377352	331660000000
9109	50571	2335172	331160000000

Left: The first column is column ID. The second column is address ID. The third column is address hash, i.e. the real address appearing in a block. **Middle**: The first column is column ID. The second column is address which receives bitcoins. The third column is the amount of 10^−8 bitcoins. **Right**:The first column is column ID. The second column is address which sends bitcoins. The third column is the amount of 10 −8 bitcoins.

Our Work - heuristic clustering

percentile	50%	60%	70%	80%	90%	95%	99%	99.9%
# of addresses	1	1	1	2	2	3	8	51

Table 1: Percentiles of number of addresses owned by one identified user. We note that more than 70% users only own one address, and less than 0.1% users own more than 50 addresses.

percentile	50%	60%	70%	80%	90%	95%	99%	99.9%
# of transactions	3	4	5	6	9	16	59	277

Table 2: Percentiles of number of transactions a single address involved with. We note that more than 50% addresses are only involved in own 3 transactions, and less than 0.1% addresses are involved with more than 277 transactions.

Our Work - machine learning clustering

- Feature extraction of an address
  - in-degree: # of times an address sending bitcoins to others
  - out-degree: # of times an address receiving bitcoins to others
  - mean of in-value: mean of amount of bitcoins an address sending to others
  - mean of out-value: mean of amount of bitcoins an address sending to others
  - variance of in-value: variance of amount of bitcoins an address sending to others
  - variance of out-value: variance of amount of bitcoins an address sending to others

Our Work - machine learning clustering

- Unsupervised learning
  - **K-means**: The k-means algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. This algorithm requires the number of clusters to be specified.
  - **DBSCAN**: The DBSCAN algorithm views clusters as areas of high density separated by areas of low density. Due to this rather generic view, clusters found by DBSCAN can be any shape, as opposed to k-means which assumes that clusters are convex shaped. The central component to the DBSCAN is the concept of *core samples*, which are samples that are in areas of high density. A cluster is therefore a set of core samples, each close to each other (measured by some distance measure) and a set of non-core samples that are close to a core sample (but are not themselves core samples).
  - **Spectral clustering**: Spectral clustering does a low-dimension embedding of the affinity matrix between samples, followed by a K-Means in the low dimensional space. Spectral clustering requires the number of clusters to be specified. It works well for a small number of clusters but is not advised when using many clusters.

Division of Labor

- Learn basic knowledge of bitcoins and blockchains: both
- Literature review: both
- Collect data: Hongjie Chen
- Process data: Chongyao Xia
- Heuristic clustering: Hongjie Chen
- Machine learning clustering: Chongyao Xia
- Reports and PPT: both

Reference

- [1] Meiklejohn, Sarah, Marjori Pomarole, Grant Jordan, Kirill Levchenko, Damon McCoy, Geoffrey M. Voelker and Stefan Savage. “A fistful of bitcoins: characterizing payments among men with no names.” *Commun. ACM* 59 (2013): 86-93.
- [2] Reid, Fergal and Martin Harrigan. “An Analysis of Anonymity in the Bitcoin System.” 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing (2011): 1318-1326.
- [3] Biryukov, Alex, Dmitry Khovratovich and Ivan Pustogarov. “Deanonymisation of Clients in Bitcoin P2P Network.” *ACM Conference on Computer and Communications Security* (2014).
- [4] Jawaheri, Husam Al, Mashael Al Sabah, Yazan Boshmaf and Aiman Erbad. “When A Small Leak Sinks A Great Ship: Deanonymizing Tor Hidden Service Users Through Bitcoin Transactions Analysis.” *CoRR* abs/1801.07501 (2018): n. pag.

Reference

- [5] Narayanan, Arvind, Joseph Bonneau, Edward W. Felten, Andrew Miller and Steven Goldfeder. “Bitcoin and Cryptocurrency Technologies.” .
- [6] Fanti, Giulia C. and Pramod Viswanath. “Deanonymization in the Bitcoin P2P Network.” *NIPS* (2017).
- [7] Goldfeder, Steven, Harry A. Kalodner, Dillon Reisman and Arvind Narayanan. “When the cookie meets the blockchain: Privacy risks of web payments via cryptocurrencies.” *CoRR* abs/1708.04748 (2017): n. pag.

Thanks!

Special thanks to Prof. Wang and Prof. Fu for their constructive advices and supportive help!