Detection and Visualization of Bitcoin Anomaly

SUIBIN SUN

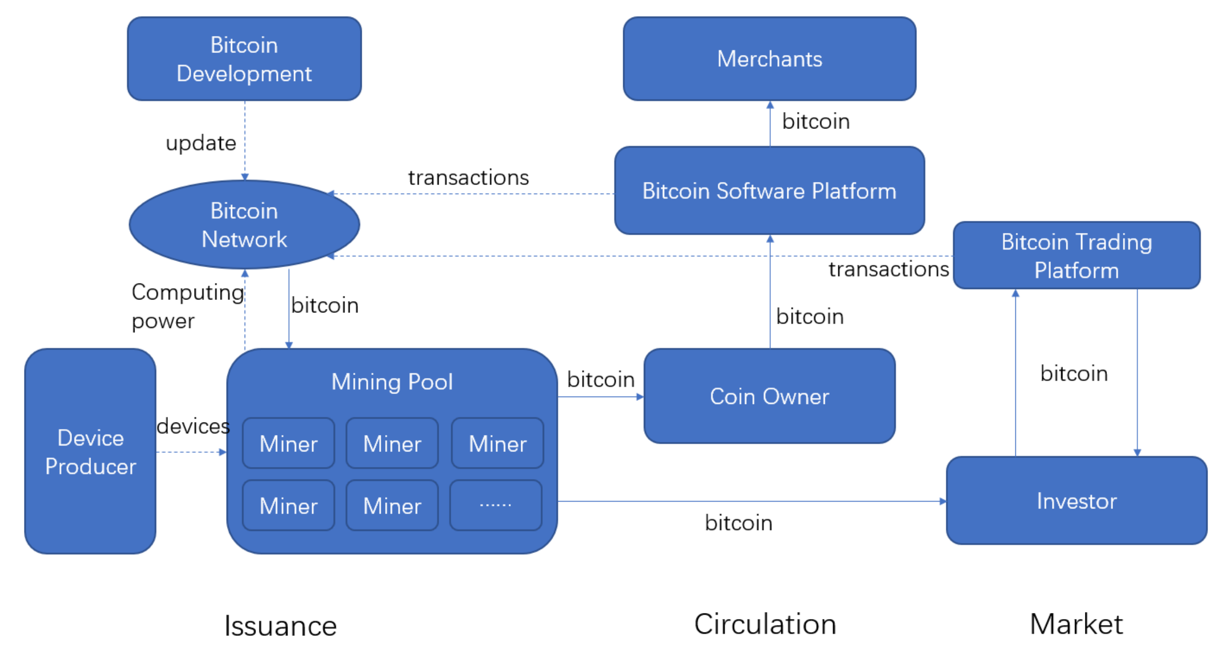
2018 5 26

Contents

- Introduction
- Methods
  - Data collection
  - Unsupervised learning
  - Online learning
  - Visualization
- Result and evaluation
  - Unsupervised learning
  - Online learning
  - Visualization
- References

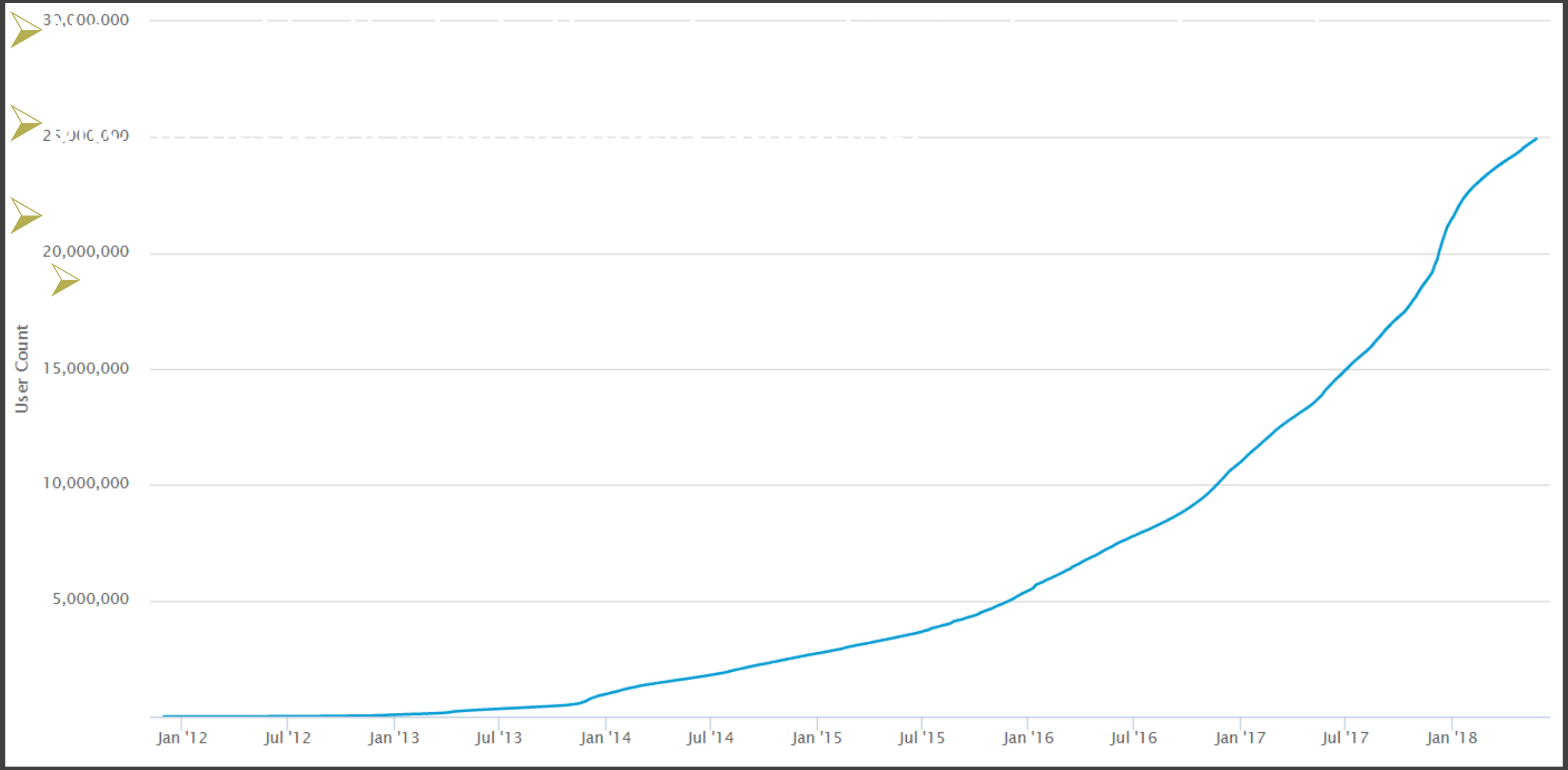
Introduction

- What is Bitcoin?
  - The first decentralized cryptocurrency.
  - Based on blockchain (distributed database system).
  - Using consensus mechanism(PoW) to keep correctness.
  - A wallet address represent an account.



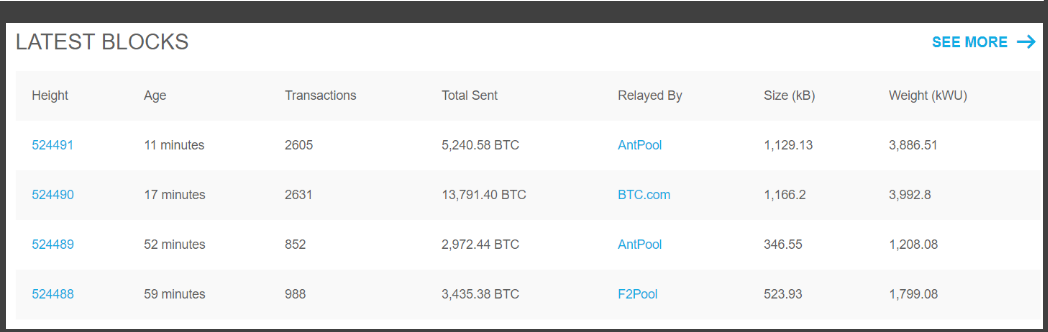
Introduction

- Number of Bitcoin users is growing rapidly.

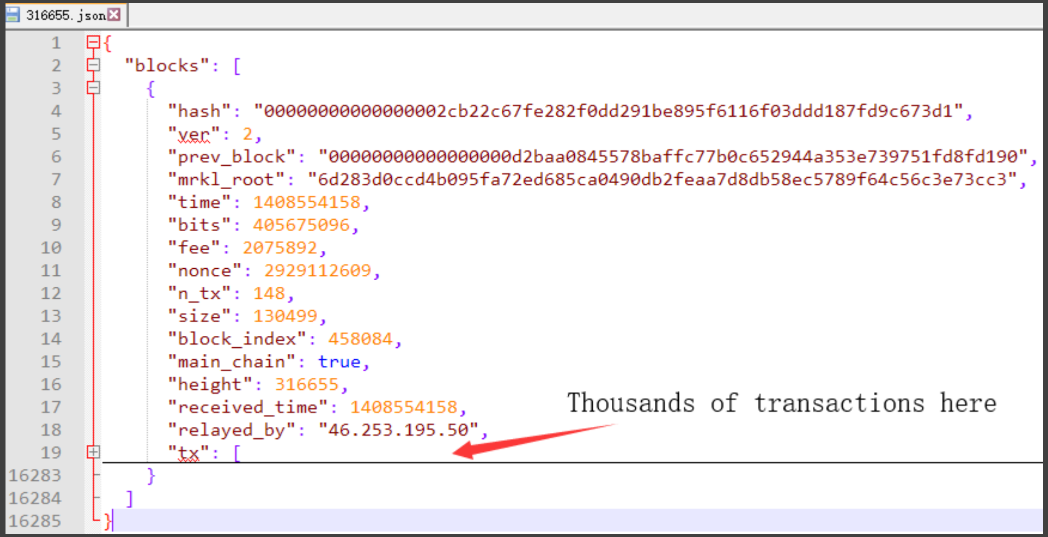


Methods – Data Collection

- All blocks are available on Internet.
- Totally 522137 blocks until May 18 2018.
  - Take 0.4MB as an average size, totally ~200GB!
  - Here I download block height 300000~522137.
  - Each block record thousands of transactions.
  - Extract useful data in transactions.
- Useful data in a transaction:
  - In-degree
  - Out-degree
  - Total fee of this transaction



Height	Age	Transactions	Total Sent	Relayed By	Size (kB)	Weight (kWU)
524491	11 minutes	2605	5,240.58 BTC	AntPool	1,129.13	3,886.51
524490	17 minutes	2631	13,791.40 BTC	BTC.com	1,166.2	3,992.8
524489	52 minutes	852	2,972.44 BTC	AntPool	346.55	1,208.08
524488	59 minutes	988	3,435.38 BTC	F2Pool	523.93	1,799.08



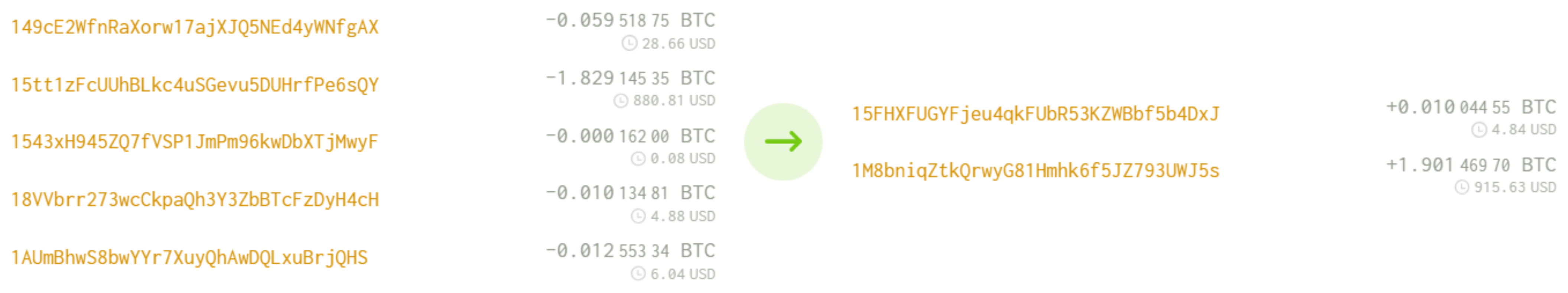
```
{
  "blocks": [
    {
      "hash": "00000000000000002cb22c67fe282f0dd291be895f6116f03ddd187fd9c673d1",
      "ver": 2,
      "prev_block": "00000000000000000d2baa0845578baffc77b0c652944a353e739751fd8fd190",
      "mrkl_root": "6d283d0ccd4b095fa72ed685ca0490db2feaa7d8db58ec5789f64c56c3e73cc3",
      "time": 1408554158,
      "bits": 405675096,
      "fee": 2075892,
      "nonce": 2929112609,
      "n_tx": 148,
      "size": 130499,
      "block_index": 458084,
      "main_chain": true,
      "height": 316655,
      "received_time": 1408554158,
      "relayed_by": "46.253.195.50",
      "tx": [
    }
  ]
}
```

Thousands of transactions here

An example block #316655

Methods – Unsupervised Learning

- Feature extraction
  - In-degree
  - Out-degree
  - Total fee



149cE2WfnRaXorw17ajXJQ5NEd4yWNfgAX	-0.059 518 75 BTC ◷ 28.66 USD
15tt1zFcUUhBLkc4uSGevu5DUHrfPe6sQY	-1.829 145 35 BTC ◷ 880.81 USD
1543xH945ZQ7fVSP1JmPm96kwDbXTjMwyF	-0.000 162 00 BTC ◷ 0.08 USD
18VVbrr273wcCkpaQh3Y3ZbBTcFzDyH4cH	-0.010 134 81 BTC ◷ 4.88 USD
1AUmBhwS8bwYYr7XuyQhAwDQLxuBrjQHS	-0.012 553 34 BTC ◷ 6.04 USD

→

15FHXFUGYFjeu4qkFUbR53KZWBbf5b4DxJ	+0.010 044 55 BTC ◷ 4.84 USD
1M8bniqZtkQrwyG81Hmhk6f5JZ793UWJ5s	+1.901 469 70 BTC ◷ 915.63 USD

An example transaction with 5 in-degree, 2 out-degree and 1.91151425 BTC total fee

Methods – K-means

- A classic unsupervised method of clustering
- Given $(x_1, \ldots, x_m)$ where $x_i \in R^3$, find k clusters $S_1, \ldots, S_k$ to solve:

$$\min_S \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

where μ_i is the centroid point of S_i.

Methods – K-means

- How to decide k?
- Calinski Harabaz Index: an efficient strategy to decide which k works better.

$$\max_k S(k) = \frac{tr(B_k)\,(m-k)}{tr(W_k)\,(k-1)}$$

where B_k, W_k are the covariance matrix of two different cluster centroids and the covariance matrix of two inner cluster data points.

- $S(k)$ is greater, k is better.

Methods – K-means

- Detection principle-whether this point is abnormal?

- Mahalanobis distance based method:

Assuming the data points drawn from multivariate normal distribution.

$$p(x;\mu,\Sigma)=\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu\ )\right)$$

where μ,σ can be estimated by

$$\hat{\mu}=\frac{1}{m}\sum_{i=1}^{m}x_i\,,\hat{\Sigma}=\frac{1}{m}\sum_{i=1}^{m}(x_i-\mu)(x_i-\mu)^T$$

Methods – Online Learning

- Bitcoin network is continuously updating, how to update our model respectively?

- Bayesian Probit Regression(BPR)

Suppose the weights of weights w meet Independent Gaussian Distribution.

$$p(w) = N(w|\mu, \Sigma)$$
$$p(y|w) = N(y|x^T w, \beta) = N(y|x^T \mu, x^T \Sigma x + \beta^2)$$

Since we can observe the label y of a new data Y , we can use KL distance to estimate the distribution of y and then the posterior. Finally we get:

$$p(w_d|y) = N(w_d|\widetilde{\mu_d}, \widetilde{\sigma_d})$$
$$\widetilde{\mu_d} = \mu_d + \frac{Y x_{i,d} \sigma_d^2}{\sqrt{x^T \Sigma x + \beta^2}} v\left(\frac{Y x^T \mu}{\sqrt{x^T \Sigma x + \beta^2}}\right)$$
$$\widetilde{\sigma_d} = \sigma_d [1 - \frac{x_i d \sigma_d^2}{x^T \Sigma x + \beta^2} w(Y x^T \mu / \sqrt{x^T \Sigma x + \beta^2})]$$

Methods – Online Learning

- Bayesian Probit Regression(BPR)

Our update algorithm:

(1) initialize $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \dots, \mu_D, \sigma_D^2$,

(2) input a new data y with label Y, for $d = 1, \dots, D$:

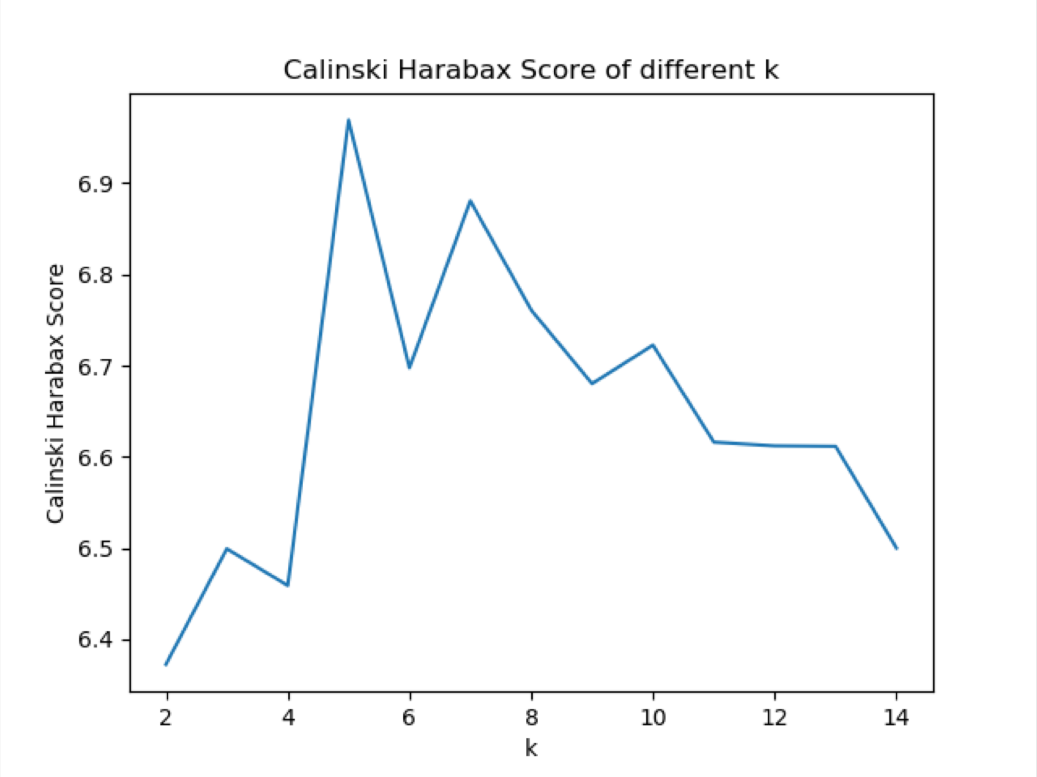
update μ_d and σ_d by the previous formula.

Methods – Visualization

- To show the result more intuitively, we build small web app which plot the realtime data of Bitcoin network.
- Including:
  - Average transaction fee recorded in a block
  - Difficulty of mining
  - Number of transactions per day

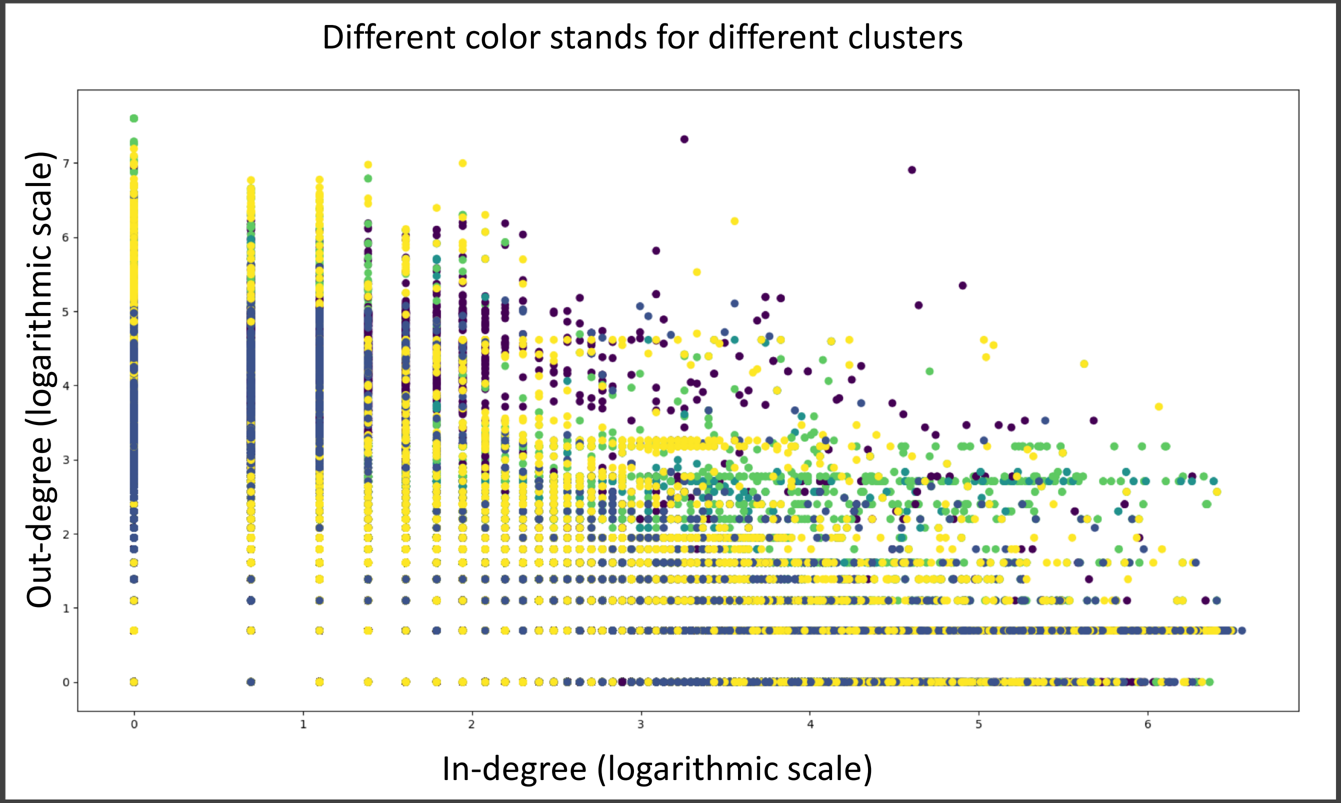
Result and Evaluation

- Using Calinski Hearbaz index to choose best k
- $K = 5$, the score is the greatest
- We choose $k = 5$ for following experiments.



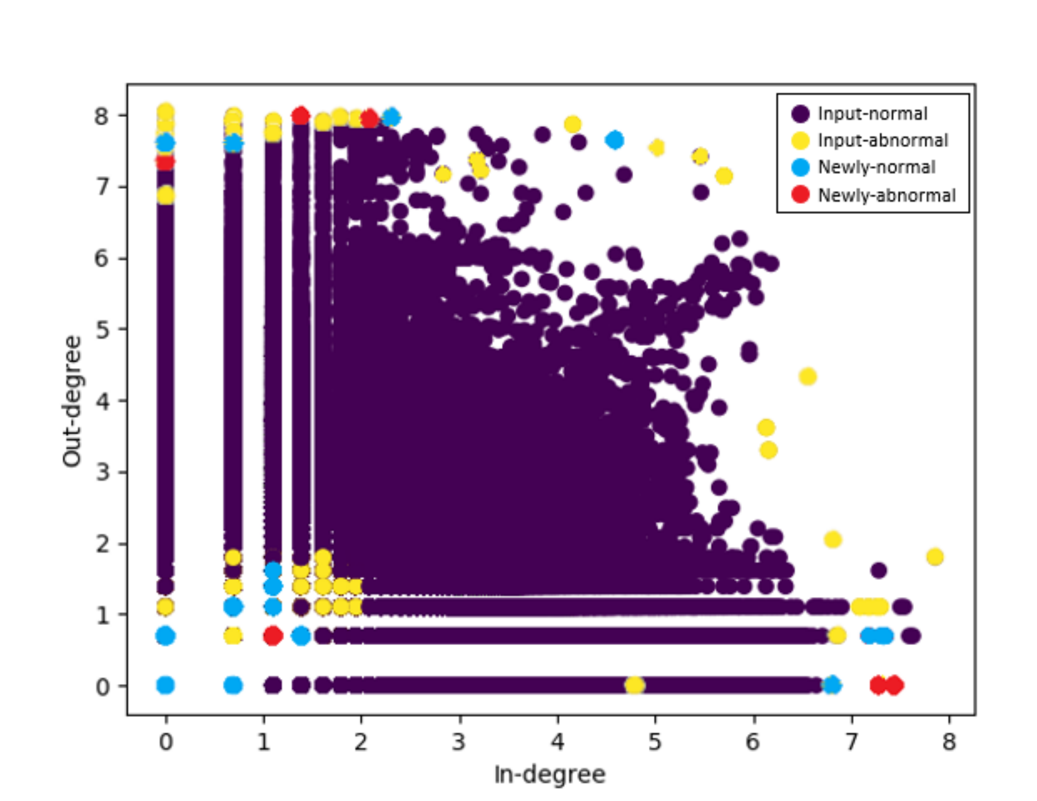
Result and Evaluation

- Result of k-means:
  - Different color means different cluster.
- 3D image is too time costing and not intuitive.



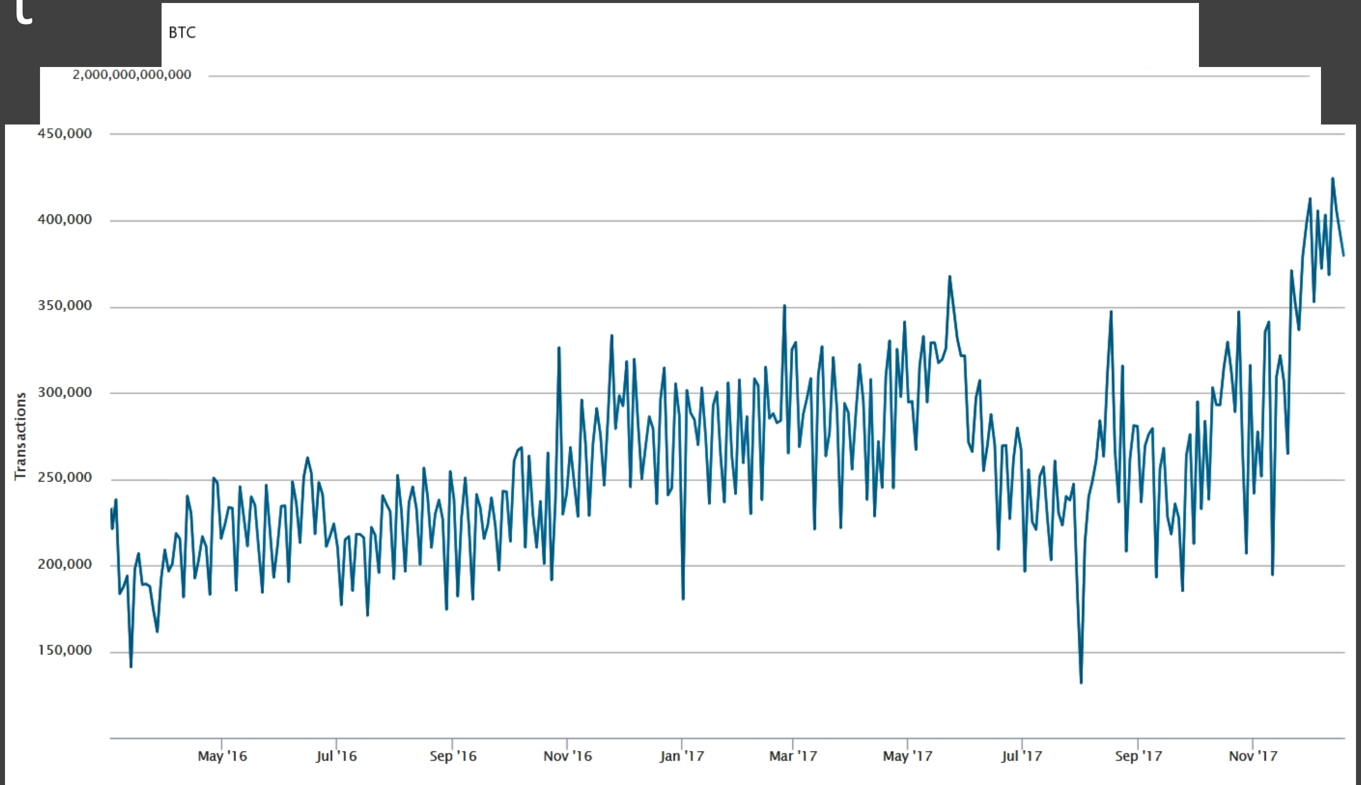
Result and Evaluation

- Input data
  - Purple: normal
  - Yellow: abnormal
- Newly added data:
  - Blue: normal
  - Red: abnormal



Result and Evaluation

- Average transaction fee chart
- Difficulty chart
- Number of transactions per day



Thanks!