Tracking the trasaction initiator in P2P Bitcoin networks A theory

Lin Yixiao
515021910204
Shanghai Jiao Tong University

May 27, 2018

Abstract

In this project, we discussed a simple source estimator and apply it in the bitcoin transaction network, we also find what we need to do to apply the rumor source estimator in blockchain network. We mainly use the maximum likelihood estimator to find the rumor source in graphs and trees. But the reliable results only gets in specific positions, the futher use in real environment needs more research on the project.

Keywords: maximum likelihood detection, bitcoin network

1 Introduction

The totally anonymous bitcoin transaction networks truly help those people who want to hide themselves. But it also provides oppotunities for someone with evil intensions. They know that they are not absolutely safe. So a simple trade in the real world will include a lot of bitcoin transaction in blockchain.

there should be a picture to show

In this situation, people would like to identify the source or the initiator of the transaction.

The transaction of the bitcoin network is a bit like the rumor spreading model but I make some changes. The common analyse method of the rumor spread model is to use the SIR or susceptible recovered model.

2 Source estimator

2.1 Model of the transaction

The blockchain network can be considered as a directed graph G(V,E), V is a countably infinite set of nodes and E is the set of edges of the form (i,j) for som i and j in V. If there is transaction from i to j, there should be a edge in E of (i,j). The edge should include a nomralized value to present the bitcoin. When the bitcoin is transmit from one to the other, the value of (i,j) is the number of the bitcoin itself. When the bitcoin is transmit from one node to multi-nodes, the value of (i, j_n) is the number of j_n receives. When the bitcoin is transmit from multi-nodes to multi-nodes, assuming that there are n transmitting nodes and each node send some bitcoins represented as $i_1, i_2 ... i_n$k, also there are m receiving nodes and each receives some bitcoins represented as $j_1, j_2, ..., j_m$. The value of (i_q, j_w) edge is $\frac{j_w * i_q}{\sum_{k=1}^{n} i_k}$. Then all the value will be normalized in [0,1]. The formula is $y_{out} = \frac{Y_{max} - y}{Y_{max} - Y_{min}}$.

The use of the value is mainly on choosing the root node in a tree graph, and decide which node is the source node if the maximun likelihood probability of serveral nodes is the same. In the step of the maximun likelihood estimator, the value is not used.

The bitcoin network is a decentralization network and each transaction include a number of bitcoin. The model of the transaction is different than rumor model.

2.2 Maximum likelihood

Firstly, let us suppose that the transaction starts at a node v_*. Although the bitcoin network is a decentralization network, we can still suppose there is a root node for the convenience of our analyse. As for how to choose the root node, the simplest way to choose is to choose the node which has the largest value send out. Then the graph can be transmitted to a undirected node. Then there will be some infected nodes in the network with N numbers. These N nodes construct a part of the graph G. It should be a subgraph of the graph G. We call it G_N. we call the real node of the initiator the source node.

Obviously, the root node is not always the real source node. So it is necessary to make a uniform prior probability of the source node among all nodes of G_N. So the maximun likelihood estimator is to minimize the error probability which means to maximize the detection probability. So the estimator is the same like the rumor spreading model:

$$\hat{v} \in arg\ max\ P(G_N, v), v \in G_n$$

where $P(G_N, v)$ is the probability of observing G_N under the SI model assuming that the v is the source node. What we need to do next is to evaluate the $P(G_N, v)$ to find the maximun value.

2.3 Prior definitions

$Definition1$: For our convenience, we will make some definitions. A notion of rumor centrality is introduced. Rumor centrality is a graph score function. That is, it takes $G = (V, E)$ as input and assigns a positive number or score to each of the vertices. Then the estimated source is the one with maximal score or rumor centrality. The estimated source is called the rumor center. To be short, the rumor centrality is the number of the spreading orders a node can originate. We start with the precise description of rumor centrality for a tree graph G: the rumor centrality of node $u \in V$ with respect to $G = (V, E)$ is

$$R(u, G) = \frac{|V|!}{\prod_{\omega \in V} T_w^u}$$

where T_w^u is the size of a subtree of G that is rooted at w and points away from u.

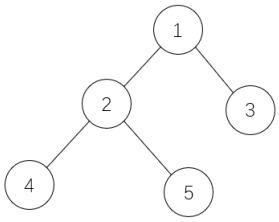


Figure 1 : *a simple tree*

In this figure, let u be the node 1. Then $|V| = 5$, the subtree sizes are $t_1^1 = 5, t_2^1 = 3, t_3^1 = 1, t_4^1 = 1, t_5^1 = 1$, and $R(1, G) = 8$. The rumor centrality on all nodes is based on the relation:

$$\frac{R(u,G)}{R(v,G)} = \frac{T_u^v}{T_v^u}$$

where u, v are neighboring nodes. The properties of the rumor centrality. Proposition1: Given a N node tree, if the node v^* is the rumor center, then any subtree with v^* as the source must have the following property:

$$T_v^{v^*} \leq \frac{N}{2}$$

If there is a node u that for all $v \neq u$

$$T_v^u \leq \frac{N}{2}$$

then u is a rumor center. A tree can have at most 2 rumor centers.

$Definition2$: For a graph , the distance centrality of node $v \in G$,$D(v, G)$, is defined as

$$D(v, G) = \sum_{j \in G} d(v, j)$$

where $d(v, j)$ is the shortest path from j to v. On a tree, the distance center is equivalent to the rumor center. On a N node tree, if v_D is the distance center, then for all $v \neq v_D$:

$$T_v^{v_D} \leq \frac{N}{2}$$

$Definition3$: For a tree $G(V, E)$ and source node v. We call σ a permitted permutations for tree $G(V, E)$. Let $\Omega(v, G_N)$ be the set of all permitted permutations starting with node v.

$Definition4$: For a tree $G(V, E)$ and node $v \in V$, $W(v)$ is the sum of the value which include all the edges between (v, i).

2.4 ML for trees

2.4.1 Regular trees

The regular trees are trees that each node has the same degree d. For trees with degress less than 2, the source is almost unable to find. For the regular trees with degree $d \geq 3$. The rumor centrality is a exact Maximun likelihood estimator.

We need to determine the probability $P(\sigma|v)$ for each $\sigma \in \Omega(v, G_N)$. Define $G_k(\sigma)$ as the subgraph of G_N. Then

$P(\sigma|v) = \prod_{k=2}^{N} P(k^{th} infected\ node = v_k | G_{k-1}(\sigma), v)$

In summary, the ML estimator for a regular trees becomes

$$\hat{v} = arg\ max\ R(v, G_N) * W(v), v \in G_N$$

2.4.2 General trees

For the general trees, the probability of correct detection of Geometric tree, SI model with spreading time with finite moment goes to 1 as the number of infected nodes increases.

$$\lim_{t\to\infty}\ inf\ P(C^1_{n(t)}) = 1$$

For Generic random tree, SI model with generic spreading time, the probability of correct source detection bounded away from 0.

$$\lim_{t\to\infty}\ inf\ P(C^1_{n(t)}) > 0$$

For Generic random tree, SI model with exponential spreading time, the probability of rumor centrality estimating the k^th infected node as the source decays but remains positive.

$\lim_{t\to\infty} P(C^k_{n(t)}) \leq k(k+1)(\frac{1}{2})^{k-1} \sim exp(-\theta(k))$

The $n(t)$ is the infected node at time t in graph G, and $C^k_{n(t)}$ is the event that the source estimated as per rumor centrality is the k^th infected node.

2.5 ML for graphs

Applying the ML estimator for a general graph needs computing the summation of the likelihoods of all possible permitted permutations given the network structure, which is computationally prohibitive. We could only make some assumptions so that our previous theory could apply into those tree-like graph.

Assume that the rumor spreads along a spanning tree of the observed graph corresponding to the first time each node receives the rumor. Suppose we know which spanning tree was involved in the tree spreading, then we can use the previous tree estimator. If the node v is the source, then the rumor will spreads along a breadth first search tree rooted at v, $T_{bfs}(v)$. Therefore we obtain the estimator:

$$\hat{v} \in\ arg\ max P(\sigma^*_v|v) R(v, T_{bfs}(v)), v \in G_N$$

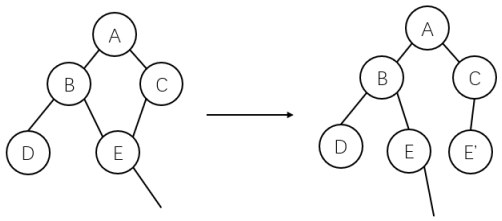
The σ^*_v represents the BFS ordering of nodes in the tree T_{bfs}.

3 Addition results

In the prior section we discuss how to find the rumor in the trees or graphs. Then we are going to discuss how to build this trees or graphs. The transaction is stored in the blocks. The height of the blocks can be regarded as the time of the rumor spreading. Normally all the transactions of a trade will not last for a long time. Thus when we want to build a tree, the selected transaction will be limited in some specific blocks. The futher transaction for some received node is another money trade. These have nothing to do with our research. So some SI model with spreading time is unsuited with the blockchain transaction.

3.1 Build the tree

We know that the Maximun likelihood estimator does not work well in the general random graphs, the ML estimator only works when knowing the spanning tree of the graphs. Therefore, we wish to build a trees. How to choose the root node of the tree is talked before. The leaf node of the tree is the output address of the transaction. If the output address has two or more input address, they will both exists in the tree, and the node is named as E and E'. But the E' node will have no leaf node, which means that although a address may mapping to different nodes. Only on node of them have the subtree.



4 Conclusion and future works

We have analyzed the behavioour of the maximum likelihood estimator in the SI model of the bitcoin transaction. The model only works in few limited positions and mainly concentrate on finding the one source of the graph. Also, finding the source in general trees or graphs needs more research. The multi-source network is still not discussed in the article. Since we know it is possible to detect the source of the transaction, it may be effective to use the machine learning model to make a link prediction of the potential source address to identify which nodes are from the same initiator. That needs us to know which node is finally the output address of the transaction.

REFERENCES

[1]Devavrat Shah,Tauhid Zaman.Rumor Centrality: A Universal Source Detector[J].Performance Evaluation Review,2012,(1):199-210.

[2]D.Shah and T.Zaman. Finding sources of computer viruses in networks: Theory and experiment. In Proc. ACM Sigmetrics, volume 15, pages 52495262, 2010.