Network Representation Learning Based Analysis for Community Detection in Bitcoin User Network

Rajeev Verma

Update: August 25, 2019

Abstract

This report is motivated to detect community structure in Bitcoin user network. Community in network data is defined as the organization of closely connected vertices with dense connections(in the form of edges) within the community and fewer outside it. These are somewhat independent components of the graph with member vertices playing more-or-less similar role. In graph theory, community detection is a challenging problem. In this work, we attempted to uncover the community structure in Bitcoin user network. We used the modern graph representation Learning network in the Bitcoin user network. We found that there are some well-defined smaller communities of users and further study has the potential to give more insights to the behaviour of users in the Bitcoin Blockchain.

1 Bitcoin User Network

For this study, we used the publicly available dataset (1). This dataset consists of transaction information of the Bitcoin Network until July 13, 2011. The users information is used for the study. The statistic of the network in Table 1.

Nodes	881,678
Edges	1,617,212

The dataset consists of two files:

1.1 User Vertices File

This file lists the tab-separated public keys belonging to the user. Since in Bitcoin a single user can have multiple public-private keys, all the identified public keys for a particular user are listed. Each user is identified by the line number.

1.2 User Edges File

This file consists of transaction data for Bitcoin Network. Each line in thisfile represents transaction occurring in the Bitcoin Network. In each line, thefirst column represents the user id of the user who gives bitcoins, the secondcolumn represents the user who receives bitcoins, the third column representsthe amount of bitcoins transferred and the fourth column represents the date-time when this transaction took place. For our study, we did not consider the transaction amount and the available timestamp.

2 Graph Representation Learning

Representation Learning is defined as learning latent vectors of d-dimension of the real-world abstract concepts like words, sentences, entities, etc. Essentially, this transforms real-world abstract concepts to some d-dimensional vector space which will preserve the semantics of the real-world concepts. These vectors are known as embeddings.

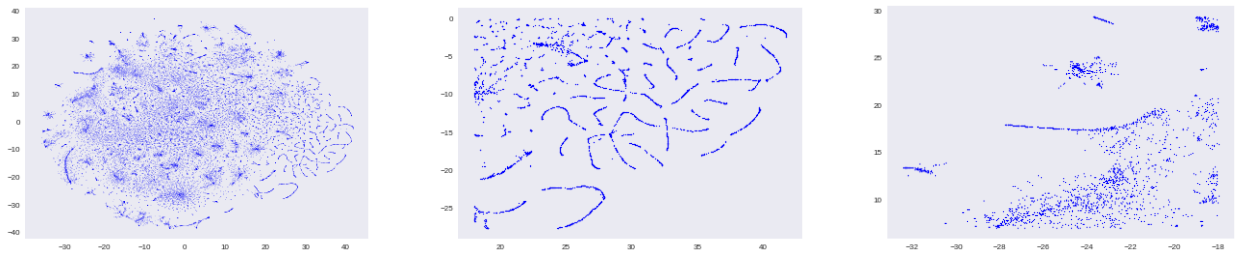
DeepWalk (2) is the first of its kind of work that employed the popular Word2Vec technique to generate graph embeddings. This method is based on modeling local information in a network by generating short truncated random walks. These walks are considered as a sequence of words in a sentence and a Skip-gram model (3) is used to learn words' representations by optimization of neighborhood preserving likelihood objective. In this work, we used the DeepWalk on Bitcoin user network to get the representations of the users. The advantage of such approach is that the learned representations capture the neighborhood structure of nodes and thus, the community structure as well.

2.1 Hyperparameters for DeepWalk

DeepWalk works by first simulating short truncated random-walks in the graph. Starting with every node, it traverses the graph with uniform probability through the neighbors of the nodes. Since the Bitcoin user dataset is large and sparse, we kept the random-walk length to be 100 and 10 walks every node. We also experimented with the default parameters which are 80 and 10 respectively. But increasing the walks per node increases the dataset size as well and due to computational restrictions, the increase in walks per node was unfeasible. Window-size for skip-gram is 5 and the dimension of embeddings is 128.

2.2 Results

We plot the projections of the learned representations through T-SNE to check the community structure. The plot is shown is Figure 1.



(a) **(b)** **(c)**

Figure 1: a:2-Dimensional T-SNE representations of the learned embeddings of users in Bitcoin Network Dataset. Note that smaller interesting close-knit communities are present.**b:** Lower - right region of **(a)** illustrated. **c:** Upper - left region of **(a)** illustrated.

3 Conclusions

As seen in Figure 1, there are interesting groups of users in the Bitcoin User Dataset and probing further into these communities will help in uncovering the users behaviours in the network. Though this is a preliminary analysis to explore the user network structure in Bitcoin Blockchain but it would surely be insightful to explore the behavioral differences among the small clusters as shown in the plot, especially Figure 1.b. Although, K-Means algorithm has been previously used in literature for study of Bitcoin network dataset but this work is unsupervised as it works with only network as input without any manual feature-engineering. We can further use the transactions details among users to study the behaviour and also the public-keys attached with every user to know the active users of the Bitcoin Blockchain. We can also augment the learned representations with classical K-Means Algorithm for further analysis.The large size of Network though put some serious demand on computation and resources.

References

http://anonymity-in-bitcoin.blogspot.com/2011/09/code-datasets-and-spsn11.html

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: Online Learning of Social Representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.pages 701-710. ACM Press, 2014

Thomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in neural information processing systems,* pages 3111-3119, 2013.