Bitcoin addresses to the user's identity. However, based on assumptions (multi-inputs in a transaction are owned by the same owner) or limited annotation data (voluntary disclosure or accidental disclosure through online forums), de-anonymity cannot achieve large-scale applications. Regarding the analysis of darknet markets, researches focus on statistical descriptions, including types of items being sold, the number of active sellers, sales volume, and the use of darknet market in different countries (Dolliver 2015; Soska and Christin 2015). In addition, the applications of Bitcoin and its blockchain technology in smart health focus on the use of blockchain technology to build a decentralized database to store electronic health records and medical research data (Ekblaw et al. 2016; Kim et al. 2017; Linn and Koo 2016), and no work on data mining for the Bitcoin transaction network has been carried out yet.

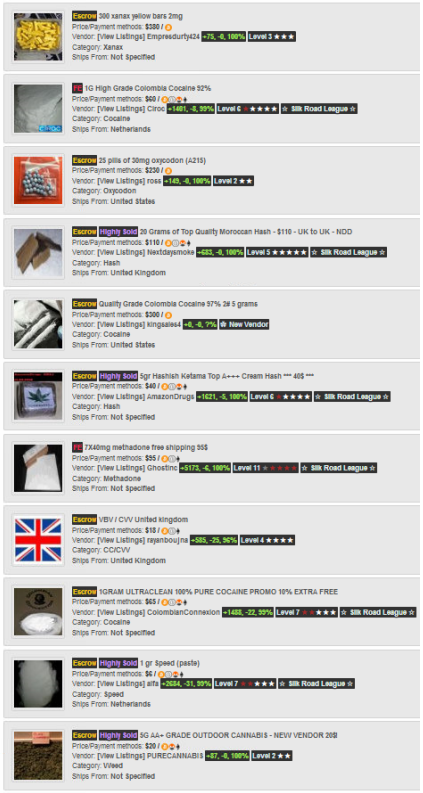

Figure 1. Part of Silk Road 3.1 homepage

Thus, there is a strong demand for techniques that detect illegal behaviors in the Bitcoin transaction network based on transaction and user patterns instead of user identities. Bitcoin exchanges, trading platform for bitcoins and fiat moneys, play an important role as they provide the only channel that links people with virtual Bitcoin addresses. Identifying the exchange addresses in the network is crucial for regulation, as they can be used to help analyze the transactions and addresses of interest, observe when they join or exit the network. It can be used as a basis for analyzing illegal drug transaction in the Bitcoin transaction network, thereby providing insights into smart health.

In this paper, by statistical analysis, we found that the exchange addresses are significantly different from the general addresses. Then, the network embedding representation is used to represent the feature space to find more comprehensive features. Based on the features, several binary classifiers are built to identify exchange addresses. Finally, the transaction network data of one week is used to analyze the performance and it is found that the proposed method is effective in the exchange addresses identification task. The contribution of this paper includes two points: first, we propose a de-anonymity method based on network embedding representation learning to identify the Bitcoin exchange addresses. Second, we explore the potential applications of Bitcoin transaction network in smart health, as the identified exchange addresses can provide a basis for subsequent analysis of illegal drug transactions in online darknet markets.

Data Collection and Analysis

In this section, we first introduce the transaction data we collected, then we show how to build a transaction network with these data, we also conduct some quantitative analyses on the behaviors of exchange addresses.

As illustrated in Liang et al. (2018), cryptocurrency transaction network is changing all the time: new nodes are added by creating new addresses and some old nodes are no longer active, and the same as the behavior of edges. Also, Liang et al. (2018) found that the monthly repetition ratio is low. Here we further calculate weekly repetition ratio defined as the ratio of the number of the same nodes or edges to the total number between adjacent weeks, and find that either edges or nodes are nearly 0.1, thus we choose one-week transactions as the research object. A summary of the dataset of different frequencies is provided in Table 1. Here, the specific number is the number of the year/month/week, and the repetition ratio is the average value.

Table 1. Summary of dataset of different frequencies

Data frequency	# nodes	# edges	# exchanges	Repetition ratio Edges/Nodes
Monthly	12,880,147	46,954,541	174,201	0.57% / 4.80%
Weekly	3,100,148	11,135,446	69,224	1.08% / 9.06%

We download Bitcoin transaction histories from July 3 to 9, 2018 UTC from the website Bitcoin Block Explorer[1], including 3,100,148 unique addresses and 1,356,519 transactions. Then, we collect 121 unique exchanges from WalletExplorer.com[2]. For each exchange, we further download their addresses. Based on the hash value of the addresses, we obtain the transaction data with labeled exchanges. Among them, there are 69,224 labeled exchanges accounting for 2.23% of all addresses, and these exchanges are involved in 89,085 transactions, accounting for 6.57% of the total transactions. The frequency of active exchanges during the week is shown in the left of Figure 2.

Using the downloaded transaction data, we construct a corresponding transaction network. Similar to banking transactions, Bitcoin transactions have a natural graphical structure, thus we can use the transaction network to represent the flow of bitcoins between addresses over time. In a transaction network, a node represents an address, and the edge between the source node and the target node represents the input-output relationship in the same transaction. Bitcoin transactions usually have multiple input and output addresses, and the number of bitcoins from inputs to outputs is not clear, thus there exists an edge between any input address and output address in a transaction. For example, a transaction with two inputs and three outputs can form six edges as shown in the right of Figure 2.

Table 2. Proportions of Top 5 Degrees in the Transaction Networks

Degree	1	2	3	4	5	Sum
Exchange	0.135	0.117	0.106	0.042	0.025	0.425
General nodes	0.246	0.187	0.293	0.068	0.037	0.831

In order to find out whether there are differences between nodes corresponding to exchanges and the general nodes in the transaction network, we calculate the degree distribution of the two types of nodes in Figure 3 and show the proportions of the top 5 degrees in Table 2. It can be seen that there exist significant differences as follows: for the general nodes, the frequency of nodes with degree of 3 is the highest, followed by the nodes with degree of 1, while for the nodes corresponding to the exchanges,

[1] Bitcoin Block Explorer—Blockchain, available from: https://blockchain.info/.

[2] WalletExplorer.com: smart bitcoin block explorer, available from: https://www.walletexplorer.com/.

the frequency of nodes with degree of 1 is the highest. In addition, the degree of general nodes is mostly concentrated in the top 5 degrees, while the degree distribution of exchanges is relatively scattered, as the top 5 degrees account for less than half. The above observations reflect different patterns between the two types of nodes to a certain extent, providing a valid basis for subsequent work.

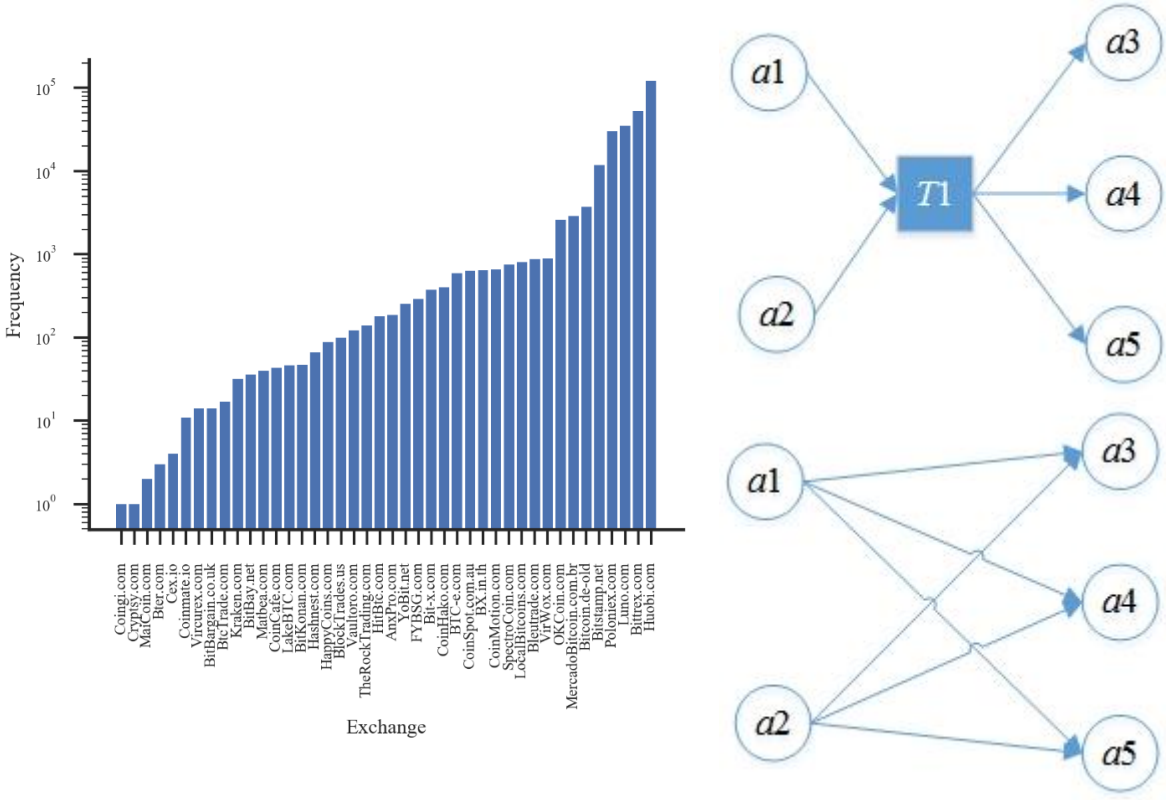

Figure 2. Frequency of exchanges in Bitcoin transactions (left) and illustration of transaction construction (right).

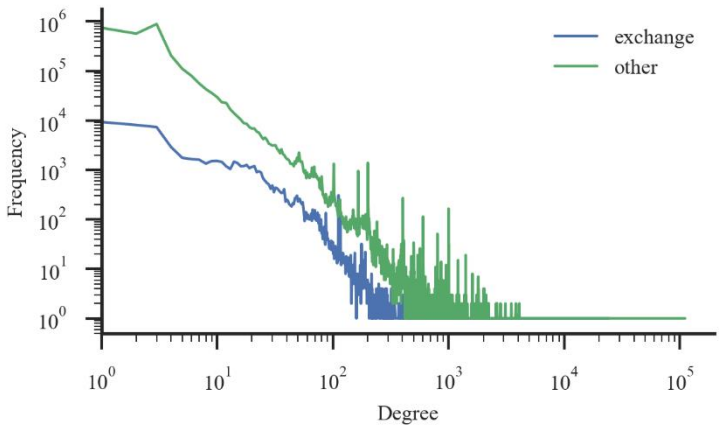

Figure 3. Degree distribution of nodes in the transaction network.

Problem Definition and Solution

We consider the problem of classifying nodes of a transaction network into two categories, one corresponding to the exchange addresses and the other corresponding to the general nodes. Given a transaction network $G = (V; E)$ comprising a set V of nodes and a set E of edges representing the relationships among nodes with $E \subseteq (V \times V)$, let $G_L = (V; E; X; Y)$ be a labeled transaction network with attributes $X \in R^{|V| \times S}$ where S is the size of the feature space for each attribute vector, $Y \in \mathrm{R}^{|V|}$ is the set of labels whose values are 0 (the general nodes) or 1 (exchange addresses). In this paper, our aim is to learn a mapping function $\Phi: X \rightarrow Y$ from the feature space to the set of labels.

Generally, ad-hoc network measurements are selected as the feature space, such as weighted in/out-degree, the number of siblings/successors/predecessors in (Albert and Barabási 2002). Due to the imperfect cognition of the network, these features may overlap each other or may not cover certain attributes of the network. Thus we use DeepWalk (Perozzi et al. 2014), an unsupervised method, to generate a vector representation of each node by capturing the network topology information and obtained features are low dimensional, informative and continuous. To discover the distribution of the nodes, DeepWalk adopts a truncated random walk to generate a set of walk sequences. Formally, the generated random walk sequence of the node v_i of width $2w$ is $v_{i-w},\ldots,v_{i-1},v_{i+1},\ldots,v_{i+w}$. Then learning from Skip-Gram (Rong 2014), DeepWalk aims to learn the latent representation of each node by maximizing the probability of node neighbors for each walk sequence as

$$-\log P_r(\{v_{i-w},\cdots,v_{i+w}\} \setminus v_i \mid \Phi(v_i)) = \prod_{j=i-w, j\neq i}^{j=i+w} \Pr(v_j \mid \Phi(v_i)),$$

where $\Phi(v_i) \in R^d$ is the vector mapping of the node v_i and they are the parameters of the model. The stochastic gradient descent method is used for parameter optimization, and the back propagation algorithm is used to estimate the derivative, so as to learn the implicit representation of nodes. The resulting feature space $X_E \in \mathrm{R}^{|V| \times d}$, where d is the number of vector dimensions, can be used directly for the feature space of the binary classifier.

Experiments

This section analyzes the effectiveness of the proposed special exchange addresses identification algorithms based on network representation learning in the real data set.

Experiment setup

As illustrated in Perozzi et al. (2014), we need to choose an appropriate dimension d considering the efficiency and accuracy factors, that is, too low dimension is hard to express the characteristic of the network, but too high dimensions increase the computational complexity. We have explored a large scale of values for dimension d, and find that the value $d = 32$ is optimal, thus the following experiments are all conducted with the dimension of feature space being 32.

Due to the inherent properties of the Bitcoin transaction network, there exists the imbalance problem of classes—— there are more addresses for general nodes than those for the exchanges. Thus, we use undersampling method (Liu et al. 2009) to randomly remove some general nodes to guarantee the number of the two classes are balanced. Taking the sampled data as the input of the classifier, we briefly introduce the five classifiers employed.

Perceptron (Gallant 1990) is a linear classifier, a function that maps its input x to an output value $f(x)$ by mapping function $f(x) = \mathrm{sign}(w \cdot x + b)$, here w is a vector of real-valued weights, b is the bias, and $\mathrm{sign}(\cdot)$ is the sign function defined as

$$\mathrm{sign}(\mu) = \begin{cases} +1, \mu \geq 0 \\ -1, \mu < 0 \end{cases}.$$

The maximum likelihood estimation method is used to estimate the model parameters, that is, the optimization problem uses the likelihood function as the objective function. And the gradient descent method and the Quasi-Newton method are usually used to solve the optimization problem.

Linear SVM (Vapnik 1995) constructs a hyperplane that has the largest distance to the nearest training data point of any class (the so-called functional margin), since in general the larger the margin, the lower the generalization error of the classifier. And the optimization problem is defined as

$$\begin{aligned} &\min_{w,b} \frac{1}{2}\|w\|^2 \\ &s.t.\ y_i(w \cdot x_i + b) - 1 \geq 0, i = 1,2,\cdots,N \end{aligned}.$$

The resulting hyperplane is $w^*x + b^* = 0$, and the classification function is $f(x) = \text{sign}(w^*x + b^*)$.

Binomial logistic regression (Collins et al. 2002) converts the problem to a conditional probability distribution as $P(Y = 1|x) = \varphi(w \cdot x)$, and $P(Y = 0|x) = 1 - P(Y = 1|x)$, where $\varphi(\cdot)$ refers to the sigmoid function defined as

$$\varphi(\mu) = \frac{1}{1 + e^{-\mu}} = \frac{e^{\mu}}{e^{\mu} + 1}.$$

The solution is the same as that of the perceptron.

Decision tree (Quinlan 1986) is a classifier recursively partitioning the feature space into a flowchart-like tree structure. The tree includes internal nodes and leaf nodes, where each internal node represents a test on an attribute, and each leaf node represents a class label. The paths from the root to leaf nodes represent classification rules. The training process includes the following three steps:

- Feature selection: it is to select features that can be used to classify training data. We use the Gini index to define the purity of the data set and the optimal partition attribute is the attribute that makes the Gini index smallest after partitioning as:

$$\min_{t \in T} Gini(D) = \min_{t \in T} \left\{ 1 - \sum_{k=1}^{K} \left(\frac{|D_k|}{|D|} \right)^2 \right\},$$

  where t is the attribute of attribute set T, and D_k is the sample subset of the category k in D, and k is the number of categories.
- Decision tree generation: the Gini index is calculated recursively from the root node, and the training set is divided into subsets that can be classified correctly.
- Decision tree pruning: regarding the over-fitting problem of the generate decision tree, some leaf nodes or subtrees above leaf nodes are cut off from the tree, and their parent nodes or the root nodes are taken as new leaf nodes to simplify the generated decision tree.

Random forests (Cutler et al. 2004) uses the bagging technique to average the results of many tree learners to reduce the variance of the predictive class label. Also, to reduce the correlation between base learners, a subset of the attribute set is randomly selected, and a subset of data is also randomly chosen as well. The random forests algorithm is simple and easy to implement, which also exhibits powerful performance.

Results and Analysis

We then run 10-fold cross validation 10 times, and use the classic classification evaluation indicators ($F1$ score, precision, and recall) to evaluate the experimental results as shown in Table 3.

All results shown are mean (std) in 10-fold cross validation. First, the perceptron algorithm has the worst performance, indicating that in the Bitcoin transaction network, the network characteristics of nodes and node tags are not simply linear. As a contradictory measure, the precision value of random forests is the highest, and the recall value of linear SVM is the highest. Regarding $F1$ measure, except for the perceptron, the values of the other three weak classifiers are above 85%, and the ensemble classifier random forests raises the value to over 90%. The above results show that the proposed Bitcoin

exchange identification algorithms based on network embedding learning are effective, and the low values of variance show that the proposed algorithms are robust.

Table 3. Results for Exchange Addresses Classification

Model	*F*1	Precision	Recall
Random Forests	0.**9095**(0.0021)	0.**9245**(0.0029)	0.8950(0.0032)
Decision Tree	0.8699(0.0027)	0.8794(0.0038)	0.8607(0.0041)
Logistics Regression	0.8654(0.0023)	0.8258(0.0032)	0.9091(0.0028)
Linear SVM	0.8595(0.0023)	0.8090(0.0031)	0.**9168**(0.0026)
Perceptron	0.7672(0.0668)	0.7893(0.0514)	0.7666(0.1379)

Overall, our experiment results indicate that the addresses of exchanges in the transaction network are identifiable and are significantly different from the general addresses. Also, our proposed algorithms perform well across all random samples, identify the exchanges well with high values of all three evaluation measures, and robust with lower values of variance. Therefore, the node features extracted by the network embedding representation are very effective in the exchange identification task.

Based on the one-week transaction data, we build a transaction network and extract network features using unsupervised learning algorithm, and use undersampling method to guarantee the number of the two classes are balanced. Part of the sampled data is used for training and the rest part is used for testing. The proposed algorithm is based on the assumption that the behavior of the exchange is constant, so the trained classifier in this paper can be applied to the exchange identification in other time periods.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant 2016QY02D0305 and 2017YFC0820105, the National Natural Science Foundation of China under Grants 71621002 and 71702181, as well as the Key Research Program of the Chinese Academy of Sciences under Grant ZDRW-XH-2017-3. Linjing Li is the corresponding author.

Conclusion

In this paper, we proposed methods based on the network embedding representation to identify the exchange addresses in the Bitcoin transaction network. Experiments illustrated that our methods were capable of labeling addresses owned by exchange with high *F*1 scores, which could provide a basis for regulating online darknet markets. Nevertheless, there are still some directions that can be explored in further works. First, to further examine the drug sales in the Bitcoin transaction network and explore applications in smart health. Second, when using the network embedding representation, more information can be considered, including the weight of edges, the time stamps, and the direction of the network.

References

Albert, R., and Barabási, A.-L. 2002. "Statistical Mechanics of Complex Networks," *Reviews of Modern Physics* (74:1), pp. 47-97.

Brenig, C., Accorsi, R., and Müller, G. 2015. "Economic Analysis of Cryptocurrency Backed Money Laundering," *Twenty-Third European Conference on Information Systems (ECIS)*, Münster, Germany.

Christin, N. 2013. "Traveling the Silk Road: A Measurement Analysis of a Large Anonymous Online Marketplace," in: *Proceedings of the 22nd international conference on World Wide Web*. Rio de Janeiro, Brazil: ACM, pp. 213-224.

Collins, M., Schapire, R. E., and Singer, Y. 2002. "Logistic Regression, Adaboost and Bregman Distances," *Machine Learning* (48:1), pp. 253-285.

Cutler, A., Cutler, D. R., and Stevens, J. R. 2004. "Random Forests," *Machine Learning* (45:1), pp. 157-176.

Dolliver, D. S. 2015. "Evaluating Drug Trafficking on the Tor Network: Silk Road 2, the Sequel," *International Journal of Drug Policy* (26:11), pp. 1113-1123.

Ekblaw, A., Azaria, A., Halamka, J. D., and Lippman, A. 2016. "A Case Study for Blockchain in Healthcare:"Medrec" Prototype for Electronic Health Records and Medical Research Data," *Proceedings of IEEE open & big data conference*, p. 13.

Foley, S., Karlsen, J., and Putniņš, T. J. 2018. "Sex, Drugs, and Bitcoin: How Much Illegal Activity Is Financed through Cryptocurrencies?," *Review of Financial Studies (Forthcoming)*).

Gallant, S. I. 1990. "Perceptron-Based Learning Algorithms," *IEEE Transactions on Neural Networks* (1:2), pp. 179-191.

Haber, P. S., Demirkol, A., Lange, K., and Murnion, B. 2009. "Management of Injecting Drug Users Admitted to Hospital," *The Lancet* (374:9697), pp. 1284-1293.

Kim, H.-E., Kuo, T.-T., and Ohno-Machado, L. 2017. "Blockchain Distributed Ledger Technologies for Biomedical and Health Care Applications," *Journal of the American Medical Informatics Association* (24:6), pp. 1211-1220.

Liang, J., Li, L., and Zeng, D. 2018. "Evolutionary Dynamics of Cryptocurrency Transaction Networks: An Empirical Study," *PLOS ONE* (13:8), p. e0202202.

Linn, L. A., and Koo, M. B. 2016. "Blockchain for Health Data and Its Potential Use in Health It and Health Care Related Research," *ONC/NIST Use of Blockchain for Healthcare and Research Workshop. Gaithersburg, Maryland, United States: ONC/NIST*.

Liu, X., Wu, J., and Zhou, Z. 2009. "Exploratory Undersampling for Class-Imbalance Learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* (39:2), pp. 539-550.

Nakamoto, S. 2008. "Bitcoin: A Peer-to-Peer Electronic Cash System,").

Ober, M., Katzenbeisser, S., and Hamacher, K. 2013. "Structure and Anonymity of the Bitcoin Transaction Graph," *Future Internet* (5:2), p. 237.

Perozzi, B., Al-Rfou, R., and Skiena, S. 2014. "Deepwalk: Online Learning of Social Representations," in: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, New York, USA: ACM, pp. 701-710.

Quinlan, J. R. 1986. "Induction of Decision Trees," *Machine Learning* (1:1), pp. 81-106.

Reid, F., and Harrigan, M. 2013. "An Analysis of Anonymity in the Bitcoin System," in *Security and Privacy in Social Networks*, Y. Altshuler, Y. Elovici, A.B. Cremers, N. Aharony and A. Pentland (eds.). New York, NY: Springer New York, pp. 197-223.

Reynolds, P., and Irwin, A. S. M. 2017. "Tracking Digital Footprints: Anonymity within the Bitcoin System," *Journal of Money Laundering Control* (20:2), pp. 172-189.

Rong, X. 2014. "Word2vec Parameter Learning Explained," *arXiv preprint arXiv:1411.2738*).

Soska, K., and Christin, N. 2015. "Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem," in: *Proceedings of the 24th USENIX Conference on Security Symposium*. Washington, D.C.: USENIX Association, pp. 33-48.

van Wegberg, R., Oerlemans, J.-J., and van Deventer, O. 2018. "Bitcoin Money Laundering: Mixed Results?: An Explorative Study on Money Laundering of Cybercrime Proceeds Using Bitcoin," *Journal of Financial Crime* (25:2), pp. 419-435.

Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory*.