Blockchain Transaction Analysis Using Dominant Sets

Malik Khurram Awan and Agostino Cortesi(✉)

DAIS - Ca' Foscari University, Venice, Italy
cortesi@unive.it

Abstract. Blockchain is an emerging backbone technology behind different crypto-currencies. It can also be used for other purposes and areas. There are different scalability issues associated with blockchain. It is important to know the in depth structure of blockchain by identifying common behaviors of the transactions and the effect of these behaviors on the nodes of the network. Dominant set approach can categorize the blockchain transactions into different clusters without mentioning number of clusters in advance. The experimental evaluation of blockchain transactions shows better clustering accuracy of dominant set approach than existing method of central clustering approach.

Keywords: Transaction behavior analysis · Blockchain technology · Clustering

1 Introduction

Blockchain is a technology that is working behind bitcoin. The rising adoption and promising security of blockchain technology has the potential to reshape the current infrastructure of different business areas. In bitcoin, blockchain is a decentralized ledger containing the complete transaction history which is public at each node. All transactions are locked with the information of time, date, nodes and the amount. Nodes generate transactions and broadcast them to the network. Miners generate new blocks by solving proof-of-work and broadcast them to the network.

Blockchain analysis under different clustering heuristics can help to study its in-depth structure and to know the behavior of nodes and associated transactions. The clustering results obtained from these heuristics can also be useful for backend designers of the blockchain technology. These days achieving scalability in blockchain is a heated topic. Many researchers have presented different approaches and ideas for the scalability of blockchain to get better throughput in minimum latency. The recent research work [3,5,7,10,12,15,20,24] of past two years about scalability issues of blockchain model it is shown that blockchain incur storage, latency, security, processing and bandwidth problems [5,12,15,20,24]. Different researchers have carried out research activities

K. Saeed et al. (Eds.): CISIM 2017, LNCS 10244, pp. 229–239, 2017.
DOI: 10.1007/978-3-319-59105-6_20

by introducing some new proposals and improving the existing mechanism of blockchain in different crypto-curriencies [1,4,8,9,11,13,14,16,19,25]. In [18] the authors discovered that by making the clustering of non-real account names of users according to their shared ownership and associated real names can give the clearer picture of the cryptocurrency. In [17] transaction volume per node is presented with clustering coefficients per node in the bitcoin network.

Inside blockchain network, there are thousands of transactions which are increasing over time. To manage and organize this increasing number of transactions it is important to study their common behavior. Clustering these increasing number of transactions can help to find out the transactions with common behavior and to identify and trace anomalous transactions.

This paper is a step towards it by using the dominant set approach for blockchain transactions that automatically categorize the blockchain transactions into different clusters without specifying number of clusters in advance. The percentage accuracy of experimental results show better performance of dominant set approach than existing method of central clustering approach.

In next Sect. 2 we highlight the basic terminologies and algorithmic structure of dominant set approach. Section 3 shows experimental results, evaluation measures and the results summary. Conclusion is discussed in Sect. 4.

2 Background

Generally clustering can be categorized into two types: central and pairwise. Central clustering is a feature based clustering. K-means algorithm is a type of feature based clustering. Pairwise clustering is a graph based clustering and a more general approach and adaptable in the sense that the algorithm works fine with wider range of input forms.

Dominant set clustering framework has been defined and introduced in [22,23]. The clustering method of choice is dominant sets, a pairwise clustering method that generalizes the idea of maximal clique to weighted graphs. It works with pairwise similarities i.e. metric embedding is not needed. For an undirected unweighted graph $G = (V, E)$, *clique* can be defined as a subset of mutually adjacent nodes where $V = \{1, 2 \ldots, n\}$ denotes vertex set and $E \subseteq V \times V$ represents the edge set. The key notions and definitions of the dominant set approach are discussed below:

For a non-empty set of vertices $S \subseteq V$ and $i \in V$. The average weighted degree of node i w.r.t set S is defined as:

$$AWDeg_S(i) = \frac{1}{|S|} \sum_{j \in S} w_{i,j} \tag{1}$$

In case of relative similarity $\phi_S(i, j)$, when node j does not belong to set S as shown in Fig. 1. We can measure the similarity between nodes i and j w.r.t the average similarity between node i and its neighbors in S which is expressed as:

$$\phi_S(i, j) = w_{i,j} - AWDeg_S(i) \tag{2}$$

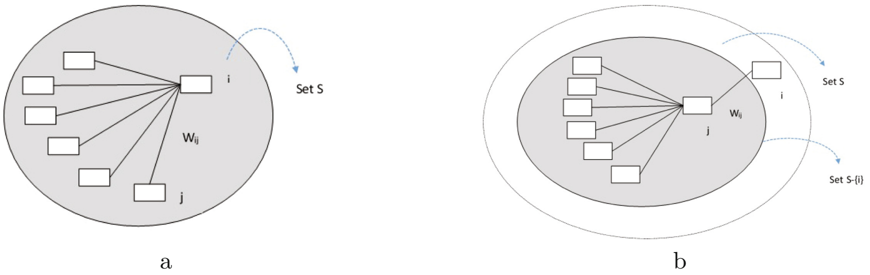

Fig. 1. (a) Average weighted degree of node i (b) Relative similarity between two nodes

For set $S \subseteq V$ and $i \in S$ we can compute the weight of node i w.r.t S:

$$W_S(i) = \begin{cases} 1 & if\,|S| = 1 \\ \sum\limits_{j \in S\setminus\{i\}} \phi_{S\setminus\{i\}}(j,i) W_{S\setminus\{i\}}(j) & \text{otherwise} \end{cases} \tag{3}$$

Total weight of the set S can be calculated by adding up all weights of $W_S(i)$ of the above mentioned recursive function.

$$W(S) = \sum_{i \in S} W_S(i) \tag{4}$$

Further in [22,23] set S is specified as dominant set by defining the internal and external criteria of clustering as:

I. $W_S(i) > 0$ for all $i \in S$

II. $W_{S\cup\{i\}}(i) < 0$ for all $i \notin S$

The authors in [22,23] transformed the combinatorial problem of identifying dominant set in graph into quadratic optimization problem and the use of dynamic system of evolutionary game theory to solve it. Which is a general-form of a problem in graph theory known as Motzkin-Straus problem [21]. The findings in [22,23] reveal that there is a one-to-one correspondence between dominant sets and the strict local maximizer of the problem.

$$max \quad x^T W x \qquad subject.to \quad x \in \Delta \tag{5}$$

where $\Delta_n = \{\bar{x} \in R^n_+ : \bar{x} \geq \bar{0}, \bar{e}^T \bar{x} = 1\}$ is the standard simplex. The weighted characteristic vector of a dominant set S, which is a strict local solution of the problem (5) is defined as:

$$x^S = \begin{cases} \frac{W_S(i)}{W(S)} & if\, i \in\, S; \\ 0 & otherwise. \end{cases}$$

The Motzkin-Straus problem is generalized by creating 1-to-1 correspondence between dominant set and strict local solution of the problem (5) in a conversely way by explaining that if x^* is a strict local solution of the problem (5) then its support $\sigma(x^*) = \{i \in V : x_i^* \neq 0\}$ is a dominant set provided that $w_{\sigma \cup \{i\}} \neq 0$ $\forall i \neq \sigma$.

For finding the local solution of the above mentioned quadratic problem (5) we use replicator dynamics that originates from evolutionary game theory. The following evolutionary step is adopted.

$$x_i(t+1) = x_i(t) \frac{(AX)_i}{X(t)^T AX(t)} \tag{6}$$

For $i = 1 \ldots n$, all trajectories that start within standard simplex $\triangle$ will remain in the simplex $\triangle$ for any number or iterations of (6). Let A be a non-negative symmetric matrix so the objective function will strictly increase along any non-constant trajectory of (6). The asymptotically stable points correspond to dominant set for the similarity matrix A. With an affinity matrix A of graph G the abstract algorithmic structure of dominant set clustering approach is:

```
Partition_into_dominant_sets(G,A)
begin
   repeat
     extract dominant_set
     remove extracted dominant_set from graph G
   until all vertices are clustered
end
Extract_transaction_dominant_set_edges(G,Dominant_sets,K)
begin
 index←1
   repeat
     if(Dominant_sets==index)
      Show it in graph G}
      index←index+1
   until (index <= K)
end
```

3 Results and Evaluation

This section highlights the detail description about performed experiments mentioned in Sect. 3.1 and the results comparison with different evaluation measures in Sect. 3.3. The experimental evaluation summary is mentioned in Sect. 3.4.

3.1 Experimental Settings

We have selected a sample of 2,048 vectors of blockchain transaction data. The pairwise distance between two set of observations has been computed by using standardized euclidean distance similarity function. The transaction data has

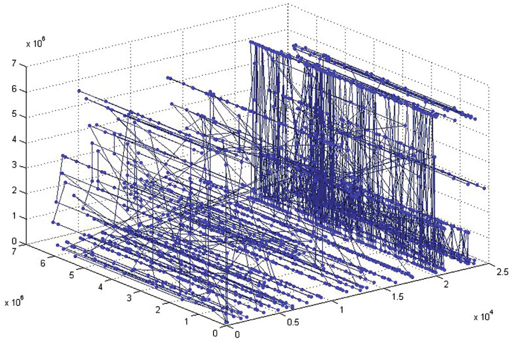

Fig. 2. Sample output of blockchain transaction edges (Color figure online)

been experimented in matlab with different parameters and settings. Some graph structures have been studied with gephi graph visualization tool. The machine used to perform experiments and compute the clustering results is a standard computer with 2.3 GHz Intel Dual-core with 3GB RAM.

3.2 Experimental Results

Dominant set approach and central clustering approach has been applied on standard blockchain transactions data. With K as number of clusters, the sample output of first 2,048 vectors of transaction edges are shown in Fig. 2. With dominant set approach, only 2 required clusters of transaction edges are extracted out of the whole graph. Transactions edges that show 'similar behavior' are in one cluster(blue) and the transaction edges that show 'different behavior' are in another cluster(green) as shown in Fig. 3. With K-means approach, we have extracted 2 clusters(red and blue) from the whole transaction graph which is actually a partitioning of the whole graph as shown in Fig. 4.

3.3 Evaluation Measures and Comparison

There are different measures for comparing clustering results. Clustering results of blockchain transactions are evaluated by computing silhouette value and rand index value. Silhouette values of results are computed to know about well-separatedness of clusters and their average is calculated to get better quantitative analysis view. Silhouette value plots are shown in Figs. 5 and 6.

Table 1 shows average silhouette value computed for range of values of K for both approaches as shown in Fig. 7

Rand index is usually used to see the accuracy or measure of similarity between two clusterings. Mathematically rand index is defined as:

$$R = \frac{\mathbf{2(a+b)}}{\mathbf{n(n+1)}}$$

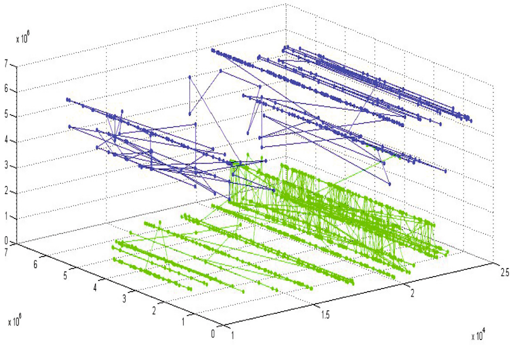

Fig. 3. Sample result of transaction clustered edges after dominant set approach with K = 2 (Color figure online)

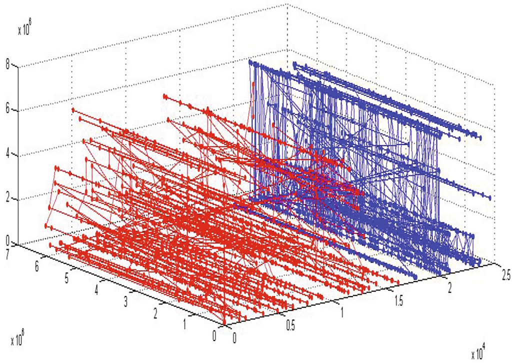

Fig. 4. Sample result of transaction clustered edges after K-means with K = 2 (Color figure online)

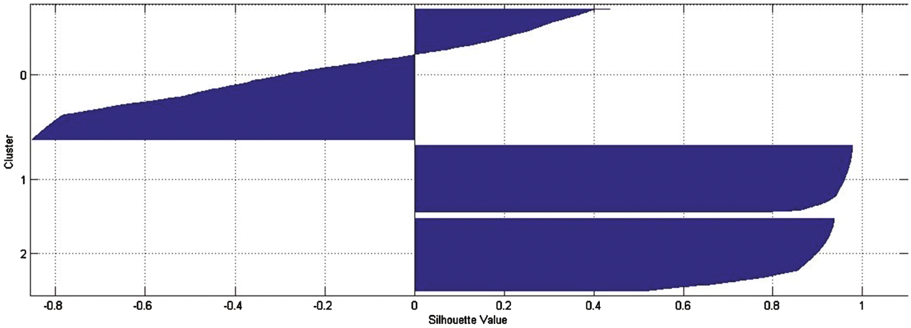

Fig. 5. Dominant Set silhouette value plot of sample with K = 2

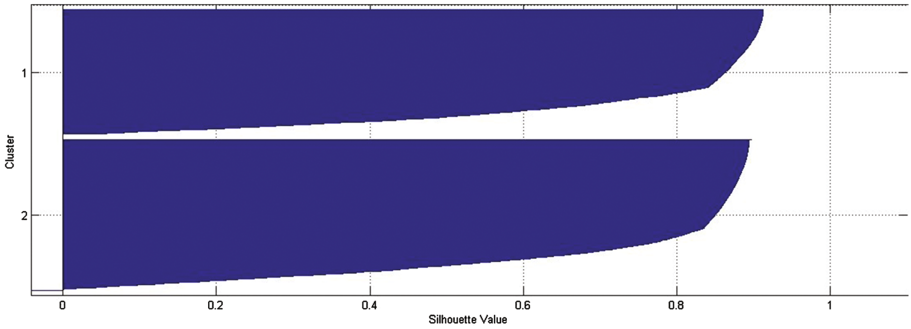

Fig. 6. K-means silhouette value plot of sample with K = 2

Table 1. Average silhouette values for dominant set and K-means clustering from 3 to 7

Clusters	3	4	5	6	7
Dominantset average Silhouette value	0.5001	0.5627	0.5566	0.5566	0.5566
K-means average Silhouette value	0.6678	0.6073	0.6972	0.6632	0.6437

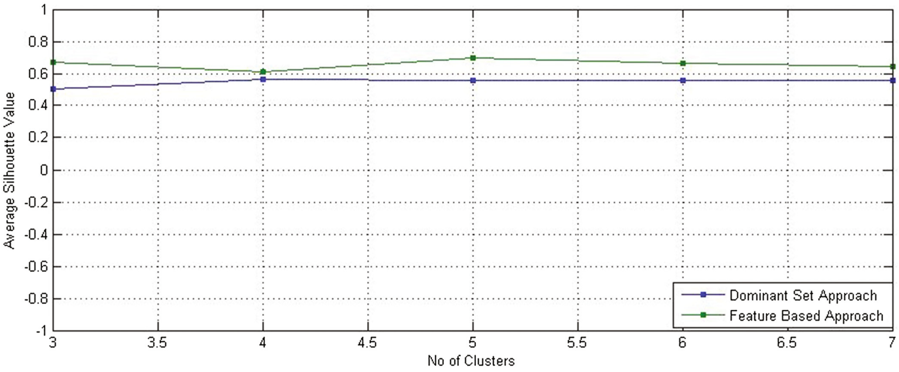

Fig. 7. Average silhouette value plot for dominant set and feature based approach(K-means)

For analysis and evaluation 6 clusters have been selected for the calculation of percentage accuracy and rand index for both approaches as shown in Tables 2 and 3. The plotted results are shown in Figs. 8 and 9.

From Fig. 8 rand index plot shows that at around cluster 3 the slope is getting lower in dominant set approach. The result accuracy plot as shown in Fig. 9 shows that the percentage result accuracy of clustering results of dominant set is better than K-means approach that shows less spike at around cluster 4.

Table 2. Rand index and accuracy for dominant set clustering from 1 to 6

Clusters	1	2	3	4	5	6
%Rand index	62.16	88.25	98.55	99.80	100	100
%Accuracy	24.70	51.70	84.13	91.35	98.63	100
Rand index	0.62	0.88	0.98	0.99	1.0	1.0
Accuracy	0.24	0.51	0.84	0.91	0.98	1.0

Table 3. Rand index and accuracy for K-means clustering from 1 to 6

Clusters	1	2	3	4	5	6
%Rand index	17.96	64.00	70.78	82.08	86.60	99.53
%Accuracy	20.31	38.37	44.62	61.86	71.38	99.36
Rand index	0.17	0.64	0.70	0.82	0.86	0.99
Accuracy	0.20	0.38	0.44	0.61	0.71	0.99

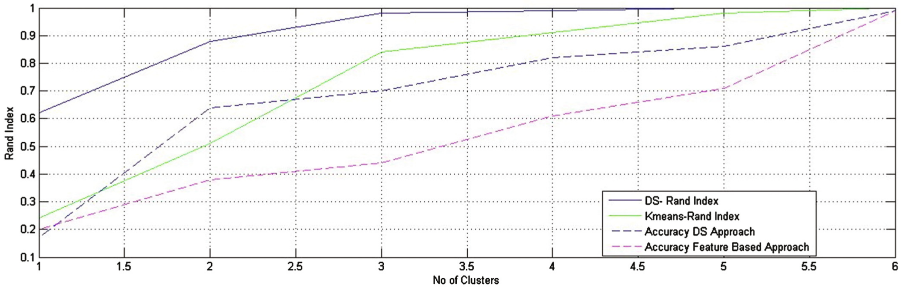

Fig. 8. Rand index plot of dominant set and feature based approach (K-means)

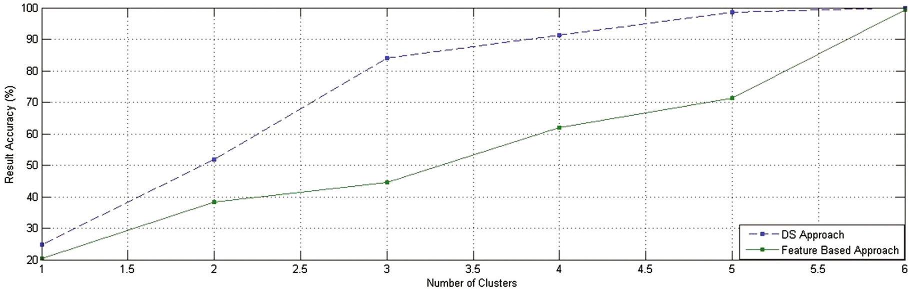

Fig. 9. Result accuracy plot of dominant set and feature based approach (K-means)

3.4 Summary of the Results

The experimental evaluation explains that: (1) dominant set method performs well for 2,048 vectors of blockchain transaction data (2) transactions with 'similar behavior' are clustered without specifying number of clusters in advance (3) with dominant set approach transactions subgraph with 'similar behavior' is extracted out of the main graph, and (4) percentage result accuracy shows better clustering results in dominant set method than feature based approach when number of clusters increases.

3.5 Constraints and Complexity

Some constraints related to experimental results are: (1) smooth running of experiments mainly depend on size of the blockchain transactions selected, and (2) hardware specification. The standard approach used here to find dominant set is replicator dynamics. From theoretical point of view, replicator dynamics has quadratic $O(n^2)$ computational complexity in every step of the dynamics for a dataset with n points.

4 Conclusion

We analyzed blockchain transaction data with pairwise dominant set and central clustering approaches by testing and evaluating it with different measures and settings. The experimental and evaluation results show better clustering accuracy of dominant set approach than existing method of central clustering approach. The in-depth information coming out of the dominant set clustering analysis can be useful for blockchain maintainers and to accurately identify anomalous transactions in the blockchain.

From scalability point of view, further study is needed for the concept of centralized blockchain with distributed chain of dominant set labels of the transactions.

Finally, in would be interesting to investigate the impact of alternative approaches to the analysis of blockchain systems, like for instance semantics-based static analysis techniques, that already provided interesting results in security analysis of software systems [2,6,26].

References

1. Back, A., Corallo, M., Dashjr, L., Friedenbach, M., Maxwell, G., Miller, A., Poelstra, A., Timon, J., Wuille, P.: Enabling Blockchain Innovations with Pegged Sidechains (2014)
2. Barbon, G., Cortesi, A., Ferrara, P., Pistoia, M., Tripp, O.: Privacy analysis of android apps: implicit flows and quantitative analysis. In: Saeed, K., Homenda, W. (eds.) CISIM 2015. LNCS, vol. 9339, pp. 3–23. Springer, Cham (2015). doi:10.1007/978-3-319-24369-6_1

3. Beck, R., Czepluch, J.S., Lollike, N., Malone, S.: Blockchain-The Gateway to Trust-Free Cryptographic Transactions. In: ECIS, Turkey (2016)
4. Bamert, T., Decker, C., Elsen, L., Wattenhofer, R., Welten, S.: Have a snack, pay with bitcoins. In: IEEE 13th International Conference on Peer-to-Peer Computing, Italy, pp. 1–5 (2013)
5. Bonneau, J., Miller, A., Clark, J., Narayanan, A., Kroll, J.A., Felten, E.W.: SoK: research perspectives and challenges for bitcoin and cryptocurriencies. In: IEEE Symposium on Security and Privacy (2015)
6. Cortesi, A., Ferrara, P., Pistoia, M., Tripp, O.: Datacentric Semantics for Verification of Privacy Policy Compliance by Mobile Applications. In: D'Souza, D., Lal, A., Larsen, K.G. (eds.) VMCAI 2015. LNCS, vol. 8931, pp. 61–79. Springer, Heidelberg (2015). doi:10.1007/978-3-662-46081-8_4
7. Croman, K., Decker, C., Eyal, I., Gencer, A.E., Juel, A., Koshba, A., Miller, A., Saxena, P., Shi, E., Sirer, E.G., Song, D., Wattenhofer, R.: On Scaling Decentralized Blockchains (2016)
8. Decker, C., Wattenhofer, R.: Information Propagation in the Bitcoin Networks. In: IEEE Thirteenth International Conference on Peer-to-Peer Computing, Italy (2013)
9. Decker, C., Wattenhofer, R.: Bitcoin transaction Malleability and MtGox. In: Kutyłowski, M., Vaidya, J. (eds.) ESORICS 2014. LNCS, vol. 8713, pp. 313–326. Springer, Cham (2014). doi:10.1007/978-3-319-11212-1_18
10. Decker, C., Wattenhofer, R.: A Fast and Scalable Payment Network with Bitcoin Duplex Micropayment Channels. In: 17th International Symposium, Canada, pp. 3–18 (2015)
11. Doll, A., Chagani, S., Kranch, M., Murti, V.: Btctrackr: finding and displaying clusters in bitcoin. Princeton University, USA (2014)
12. Eyal, I., Gencer, A.E., Sirer, E.G., Renesse, R.V.: Bitcoin-NG: a scalable blockchain protocol. In: 13th USENIX Symposium on Networked Systems Design and Implementation, USA (2016)
13. Eyal, I., Sirer, E.G.: Majority is not Enough: Bitcoin Mining is Vulnerable. Cornell University, USA (2014)
14. King, S., Nadal, S.: PPCoin: Peer-to-peer Crypto-Currency with Proof-of-Stake (2012)
15. Kosba, A., Miller, A., Shi, E., Wen, Z.: Hawk: The Blockchain Model of Cryptography and Privacy-Preserving Smart Contracts. In: Cryptology ePrint Archive (2015)
16. Kroll, J.A., Davey, I.C., Felten, E.W.: The economics of bitcoin minning or, bitcoin in the presence of adversaries. In: 12th Workshop on the Economics of Information Security (2013)
17. Lischke, M., Fabians, B.: Analyzing the Bitcoin Network: The First Four Years (2016)
18. Meiklejohn, S., Pomarole, M., Jordan, G., Levchenko, K., McCoy, D., Voelker, G.M., Savage, S.: A fistful of bitcoins: characterizing payments among men with no names. In: Proceedings of Internet Measurement Conference, Spain, pp. 127–140 (2013)
19. Miers, I., Garman, C., Green, M., Rubin, A.D.: Zerocoin: Anonymous Distributed e-Cash from Bitcoin. In: IEEE Symposium on Security and Privacy (2013)
20. Miller, A., Jansen, R.: Shadow-bitcoin: scalable simulation via direct execution of multi-threaded applications. In: IARC Cryptology ePrint Archive (2015)
21. Motzkin, T.S., Straus, E.G.: Maxima for Graphs and a new Proof of a Theorem of Turan (1965)

22. Pavan, M., Pelillo, M.: A New Graph-theoretic Approach to Clustering and Segmentation. In: CVPR (2003)
23. Pavan, M., Pelillo, M.: Dominant Sets and Pairwise Clustering. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 167–172 (2007)
24. Pazmino, J.E., Rodrigues, S.: Simply dividing a bitcoin network node may reduce transaction verification time. In: SIJ Transactions on Computer Networks and Communication Engineering, pp. 17–21 (2015)
25. Ried, F., Harrigan, M.: An Analysis of Anonymity in the Bitcoin System. In: Security and Privacy in Social Networks, pp. 197–223 (2013)
26. Zanioli, M., Ferrara, P., Cortesi, A.: SAILS: static analysis of information leakage with sample. In: Proceedings of the ACM Symposium on Applied Computing, SAC 2012, pp. 1308–1313 (2012)