

PageRank and HodgeRank on Ethereum Transactions: A Measure for Social Credit

Huu-Dung Do, FPT Blockchain Lab, FPT Corporation, Vietnam

Thuat Do, Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong*

ABSTRACT

Mathematical ranking plays a critical role in the era of the internet and bigdata. Google's PageRank is well-known as a trillion-dollar algorithm. Definitely, algorithmic ranking frameworks are found on every search engine. In this paper, the article shall investigate how PageRank can be applied in the blockchain space to build up reliable and verifiable social credit and reputation systems. It is expected to provide a measure of credibility complementary and parallel with FICO, which is not applicable for individuals lacking credit information in financial institutions. Moreover, the approach proposes an unbiased method of interpreting and measuring real social interaction and reputation ranking on a blockchain network. The authors envision a future of payment based on cryptocurrencies (especially stable coins) and digital fiats; thus the proposed credit scoring framework shall be helpful for P2P credit and lending networks, possibly for decentralized finance (Defi) applications.

KEYWORDS

Blockchain, credit score, decentralized finance, HodgeRank, mathematical ranking, PageRank, reputation ranking, social credit

INTRODUCTION

Undoubtedly, getting access to financial instruments and means is essential to modern people, helping those in financial difficulty can improve their lives. The personal credit score is a key factor in accessing the financial system. However, according to World Bank statistics, an estimated 1.7 billion people in the world are unbanked, while over 44% of the current banked population is not eligible for credit which means 72% of the world population is defined as not credible and it is impossible to use the traditional financial system (Fiocco, 2019). The global population's demand for credit usage increases every year, but most parts of the world have no access to credit, especially the underdeveloped regions. Obviously, the traditional methods of evaluating personal credit scores show some disadvantages, as most of them are usually centralized, limited access, inadequate assessment,

DOI: 10.4018/IJSI.315737

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

are expensive and have no transparency. Therefore, it is crucial to develop a solution that overcomes the drawbacks of the conventional financial system, has more efficiency in credit scoring, and makes finance accessible for everyone.

Recently, with the dramatic growth of blockchain technology, it has become a promising solution for a secure, decentralized, and transparent financial system (Xue, 2021). With the help of blockchain, many informal financial organizations can allow especially unbanked people to get access to micro-credit opportunities by facilitating personal accounts created by a digital identity system on the decentralized network. This protocol helps to overcome the limitations of the traditional credit scoring systems. Unbanked people can get a credit ranking and be able to borrow a small amount that banks do not consider. The distributed scoring system makes credit scores more transparent and reliable. Besides, users do not have to build a scoring system from scratch when moving to a new country with new unknown metrics, which is the downside of centralized credit scoring. Colendi (2019) looks at user data as a two-directed graph connecting lenders and borrowers, then applies machine learning algorithms to calculate credit ranking scores (Hassija et al., 2020).

Decentralized finance (Defi) uses cryptocurrency and blockchain technology to manage financial transactions. Defi aims to democratize finance by replacing legacy, centralized institutions with peer-to-peer relationships that can provide a full spectrum of financial services including banking, loans, mortgages, etc. Defi is an emerging phenomenon that is making its way into a wide variety of simple and complex financial transactions. It's powered by decentralized applications (dapps) and protocols that handle transactions in the two main cryptocurrencies, Bitcoin (BTC) and Ethereum (ETH). While Bitcoin is the more popular cryptocurrency, Ethereum is much more adaptable to a wider variety of uses, much of the app and protocol landscape uses Ethereum-based smartcontracts. The rapid development of Defi has led to the need to identify and rank entities on the blockchain network for customer protection, performing unsecured lending, and ensuring more trust in the transaction system (Zhou et al., 2021).

Continuing the first attempt at mathematical ranking research on Blockchain transactions (Thuat Do et al., 2019), in this study, we focus on how PageRank can be applied in the public blockchain space to build reliable, and verifiable social credit and reputation systems. The mathematics of PageRank are entirely general and apply to any graph or network in any domain. Thus, PageRank is now regularly used in social and information network analysis, link prediction, and recommendation. Inspired by the mechanism of Colendi protocols (Colendi, 2019), we consider entities on the Ethereum network (i.e. wallet addresses) as network nodes and transactions as edges. The PageRank algorithm is applied to calculate the ranking score. Thus, entities on the network are ranked based on the strength of social connectivity and network activities.

Besides, the article also examines the effectiveness of HodgeRank in ranking entities on the Ethereum trading network. Hodge theory is well known as an important tool in computational electromagnetism and fluid dynamics and plays a crucial role in the mathematical theory of geometry and topology. But the research on Hodge decomposition is still limited to the statistical ranking problems. HodgeRank, introduced in (Jiang et al., 2011), proposes a new method for pairwise rank analysis represented as edge flow on graphs using combinatorial Hodge theory. The advantage of HodgeRank is that it provides a way to analyze local and global inconsistent or conflicting components of rating data, resulting in a global ranking. The study also applies HodgeRank to the web ranking problem and demonstrates that it provides a new way to estimate PageRank and allows us to study inconsistencies in PageRank models. Inspired by this research, we implement HodgeRank algorithms on Ethereum transaction data to contribute to the ranking and credit score estimation. This method opens a new horizon for analyzing more characteristics in transactions, considering these features in the problem of assessing the importance of entities on the Ethereum network.

The main contribution of this study can be summarized as follows:

- We present a new approach to interpreting and measuring real social interaction and reputation ranking on a public blockchain network.
- For the first time, ranking algorithms (PageRank and HodgeRank) are investigated and implemented in the blockchain space to build up a reliable and verifiable social credit and reputation system.
- Finally, We design an efficient data pipeline and compute process to analyze public blockchain transaction data for future studies.

The rest of this paper is organized as follows. The conceptual background and related works are described in detail in Section II. Section III presents the methodology and experimental results. Finally, Section IV gives the conclusions and the future study.

CONCEPTUAL BACKGROUND AND RELATED WORKS

Credit Score and Reputation Systems

Credit score is a critical concept and an attractive object of research and application in the financial industry. The first model to score personal credit is FICO introduced in the U.S (1989) by Fair Isaac Corporation. It's been the most popular framework for decades, used by a vast group of American financial institutions (Shweta Arya et al., 2013).

FICO score gives lenders a hint of “how likely a person is to repay a loan”. Higher FICO score, better credibility, hence lower risk. We do not describe the FICO model in detail but summarize it as follows (fig. 1).

- *Payment history* (35%) shows how a person has paid his accounts over the length of his credit. Bankruptcy, judgments, charge-offs, and late payments will lower the FICO score.
- *Debt size* (30%) refers to the total debt but is precisely measured by different metrics including the debt to limit ratio, number of accounts with balances, the amount owed across different types of accounts, and the amount paid down on installment loans.
- *Length of credit history* (15%) evaluates the average age of the accounts on a report and the age of the oldest account.
- *Credit mix* (10%) appreciates the history of managing different types of credit, for example, installment, revolving, consumer finance, and mortgage.
- *New credit* (10%) will consider new credit accounts in a (short) period of time because this exposes greater risk.

Figure 1. Key factors that determine a FICO credit score

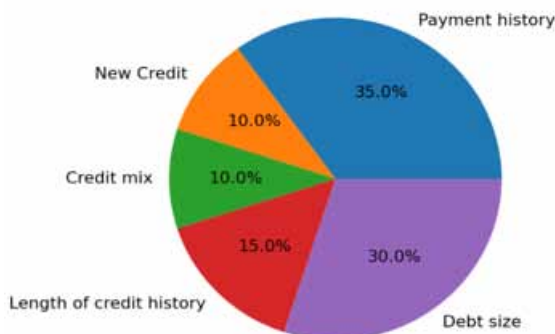


Table 1. FICO scores, rating, and meaning

Score ranges	Rating	Meaning
Below 580	Poor	Risky
580-669	Fair	Some lenders may approve loans.
670-739	Good	The score is near or slightly above the average of consumers and most lenders accept this as a good score.
740-799	Very Good	The score is above the average, indicating a dependable borrower.
800+	Exceptional	Exceptional borrowers.

FICO model (table 1) has many criticisms, although it has been used widely in the U.S for a long time. The credit score may imply the discrimination of the effect between favored and disfavored groups and broaden the moral and economical gaps between rich and poor people. Technically, the exact formulas to compute the scores are not a good indicator of risk. Of course, if a person hasn't used any credit, or doesn't have any payment history recorded in the banking systems, his credit score will be nearly zero. Although FICO and similar credit scoring systems have been popular in the U.S for many decades, they have several limitations.

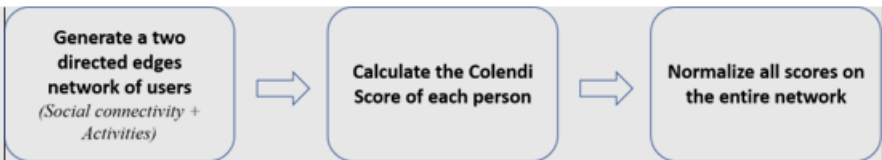
- *Data breach*: Equifax, one of the three biggest credit bureaus in the U.S, suffered a massive data breach, affecting nearly 148 million customers. Attackers could exploit the gathered credit card information to spend somewhere or for other bad purposes.
- *Limited access*: Credit scoring is not designed for unbanked people at all, even for bank account holders with no credit spending and payment history.
- *Limited data range*: Credit scoring is based on credit reports only, ignoring a lot of valuable information, for example, personal profiles, properties, or assets possessing.

The Chinese social credit system, since 2014, has had a much different approach (McWilliams et al., 2020). It intends to standardize the assessment of citizens' economic and social reputation (or social credit) as comprehensively as possible. The system gathers many types of behavioral data of individuals to analyze, then applies punishment upon fraudulent commits (e.g. government services, hospitality), to education and banking services (e.g. credit rating, lower/higher interest rate of loans). Experts have shown their concern about the Chinese social reputation system as it may harm citizens' privacy rights. Moreover, the centralized database and servers, by nature, have some cybersecurity issues and are targets for attacks.

Blockchain Application in Credit Industry

Colendi (2019) builds an application for the credit and micro-lending industry. Colendi Core is based on a distributed identity and credit scoring protocol. The firm acquires telco (mobile phone) data, social media, and trusted partners, then scores user credit ranking, offering Colendi Score on its platform connecting lenders and borrowers. The algorithm process is depicted in figure 2.

Figure 2. The algorithm process of Colendi credit scoring system



The protocol investigates W_{ij} , the “social connectivity strength” between two nodes i and j to define social relation strength and contribution to social credibility. It is a compound of relation scalar strength between two nodes, denoted by $W_{ij}^{relation}$ and dependent bilateral social relation activity, denoted by $W_{ij}^{activity}$. It reads (for a weighting factor β)

$$W_{ij} = \beta * W_{ij}^{activity} + (1 - \beta) * W_{ij}^{relation}$$

The Colendi Score of a node i is computed as

$$S_i = \alpha \left(\frac{1-d}{N} + \frac{d * \sum_j \frac{S_j^i * W_{ij}}{S_{max}}}{\sum_j W_{ij}} \right)$$

Where d the damping factor, N the total number of users (or nodes), α the normalizing coefficient, S_j^i the Colendi Score of the j -th neighbor of the node i , S_{max} the maximum Colendi Score of all nodes.

METHODOLOGY AND EXPERIMENTAL RESULTS

Ranking Algorithms

PageRank

The mathematics of PageRank, which originated from the work of L. Page et al., (1999), are entirely general and apply to any graph or network in any domain. Thus, PageRank is now regularly used in bibliometrics, social and information network analysis, and link prediction and recommendation. It's even used for system analysis of road networks, as well as biology chemistry, neuroscience, and physics. PageRank is the first ranking algorithm that used a mathematical method to order search results instead of Yahoo hierarchical search. PageRank implied the foundation of Google. Then it has become an active and extensive research topic. PageRank is well-known as a trillion algorithm. In fact, it has been an extensive and vibrant research area among scientists and technologists, with wide applications in search engines and other fields for decades.

PageRank aims to give universal ranking scores (as nonnegative real values) among sites in a Web graph. It is essentially based on the idea that the more referring linkages, the more important a website seems to be. Among millions of websites, the relative importance of a site is evaluated and ranked by hyperlinks referring to it. PageRank mathematically models web pages as vertices and hyperlinks as edges of a directed graph. Let A be an adjacency matrix among N pages in a Web graph. Then PageRank is a vector R that assigns real rating values over the pages based on the adjacency. Mathematically, $R = cAR$ is an eigenvector associated with the largest eigenvalue c . In practice, the PageRank model reads

$$H = dA + (1 - d)E, R = cHR.$$

Where d is the damping factor (usually $d = 0.85$), E is the teleportation matrix. The parameters, d means a probability of random surfer from a certain page to another, and E is to ensure that H is a Markov chain (i.e. stochastic matrix), hence the algorithm to estimate R certainly converges.

HodgeRank

HodgeRank was introduced by (Jiang et al., 2011) as a promising tool for the statistical analysis of ranking, especially for datasets with incomplete and imbalanced information. To apply this method, the required data is in the form of pairwise comparisons, meaning each voter would have rated items in pairs. Pairwise comparisons are natural unbiased methods due to the arbitrariness of the rating scale by adopting a relative measure (Dym et al., 2002). They have been popular in psychology, management science, social choice theory, and statistics. HodgeRank yields an orthogonal decomposition of the edge flows of the pairwise graph into three subspaces: a gradient subspace, a harmonic subspace, and a curly subspace. The combinatorial Hodge decomposition reads

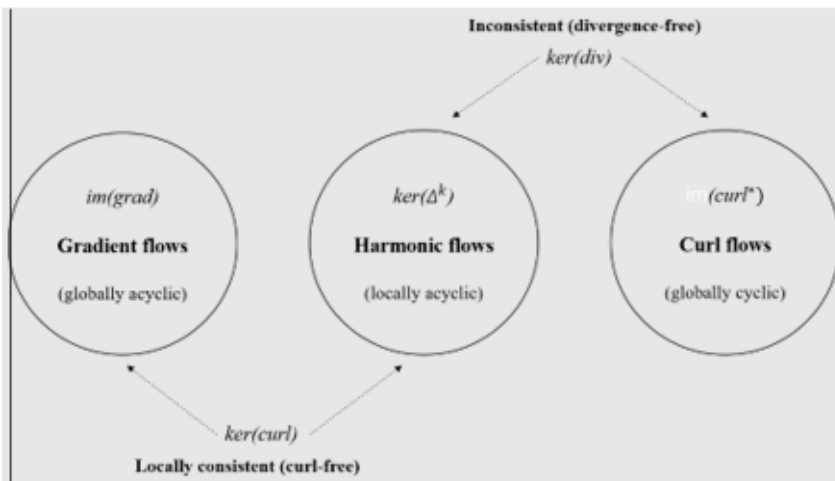
$$C_1(G) = im(grad) \oplus ker(\Delta^k) \oplus im(curl^*)$$

$$ker(\Delta^k) = ker(curl) \cap ker(div)$$

Where the $C_1(G)$ is the space of edge flows on a graph G . The $im(grad)$ denotes the subspace of pairwise rankings that are the gradient flows of score functions. Gradient subspace comprises the globally consistent or acyclic pairwise rankings. The $ker(div)$ denotes the subspace of divergence-free pairwise rankings whose total in-flow equals total out-flow for each alternative. The $ker(curl)$ denotes the subspace of curl-free pairwise ranking with zero flow-sum along any triangle, which corresponds to locally consistent pairwise rankings. The $ker(\Delta^k)$ denotes the subspace of harmonic pairwise rankings. The $im(curl^*)$ denotes the subspace of locally cyclic pairwise rankings that have non-zero curls along triangles.

To apply HodgeRank in the Ethereum transaction network context, we describe the transaction graph (fig 3) in terms of pairwise comparisons. Each node on the network is a candidate for ranking data. The direction and amount of ETH of each transaction are the edge direction and the edge weights of the graph of the pairwise comparison. When performing actual computations, we generally store

Figure 3. Hodge/Helmholtz decomposition of pairwise rankings



these graphs as matrices, the graph is used mainly to illustrate the general principles. The HodgeRank then projects our data onto gradient subspace which contains no intransitive comparisons, resulting in a global ranking of the network. The harmonic and curly subspace contains the residual components which capture the inconsistencies and intransitive relations in the data, indicating a less reliable rating.

Dataset and Preprocessing

With the hype on blockchain technologies and particularly the Ethereum blockchain, the number of statements, actions, and transactions in the network are increasing quickly, and many BigData challenges arise. However, transactions are raw data and one cannot take advantage of them for further analysis. Fortunately, Google designed the BigQuery Ethereum public datasets that contain the Ethereum blockchain data (Allen et al., 2018). The datasets describe all Ethereum terms including blocks, transactions, contract messages, high-value data, derivatives-token transfers, smart contract methods, etc. The BigQuery structure allows users to access the Ethereum blockchain via SQL and find meaningful insights which can be fed for further analysis using graph databases, visualization tools, and machine learning frameworks.(fig 4)

The Ethereum datasets in the format of BigQuery structure contain all historical data of the Ethereum public blockchain network and are updated daily. For the sake of the experiment, we limited the sample dataset containing around 34153 addresses, including value transactions, timestamps, and spent gas fees for each transaction. The dataset is extracted during the mid of 2021 when Defi applications have a rapid development, addresses and activities on the Ethereum network are diversified and more meaningful in survey and evaluation.

Inspired by the Colendi protocol, the credit scoring of a node is based on its social connections and activities on the blockchain-based network. The Ethereum transaction graph consists of nodes

Figure 4. Ethereum ETL architecture diagram

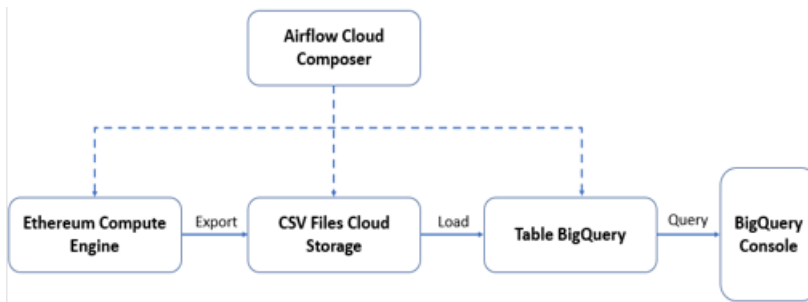
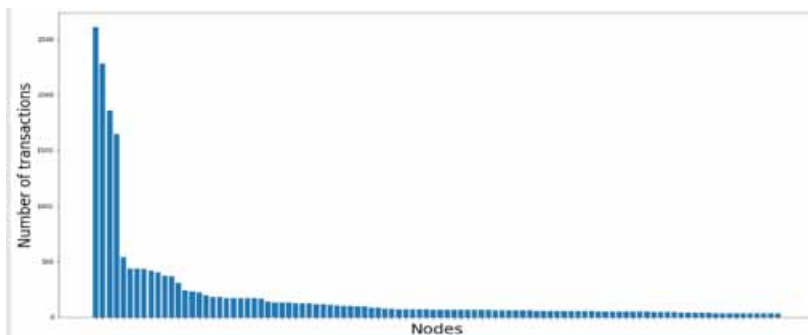


Figure 5. Distribution of the number of transactions of the top 100 nodes



(addresses), and the connections (edges) among them are transactions. A transaction, for example, a node sends ETH or any token to another, has a corresponding edge added to the graph with the direction of the transaction flow. We consider multiple edges between two nodes with the same direction as a single. Besides, all transactions from an address to itself (self-loops) are omitted.

We applied some classic graph analysis algorithms to get meaningful insights into the datasets. Some indicators are counted and used as a basis to compare with PageRank scores such as the number of received (in) transactions (NoR), the number of transferred (out) transactions (NoT), the total of in-out transactions (node degree), the spent gas for receiving transactions, the spent gas for sending transactions, the total spent gas. Fig 5 shows the top 100 nodes based on the total transactions of nodes (node degree). The majority of transactions are concentrated in a small number of major nodes. These nodes can be predicted as exchanges, dapps, that play an important role in the network. Thus, they can achieve high credit scores. In fact, across all over 30000 nodes sampled, the majority of nodes only perform a few transactions. Fig 6.a depicts the distribution of the dataset according to the node degree. We divide the dataset into 4 folds with the mean transaction number being 193, 15, 4, 1, and the number of nodes 100, 200, 700, and 29000 respectively. In addition, the amount of ETH transferred is collected to contribute to the dataset for the HodgeRank investigation. As mentioned above, addresses on the Ethereum trading network are treated as vertices on the pairwise comparison graph, and the amount of ETH is the weight of the edges. The global ranking vector of the ETH transaction pairwise graph is found by solving the least squares problem of the graph Laplacian (Jiang et al., 2011). However, due to the computational overhead of the batch algorithm, the Hodge dataset is limited to 5000 transactions and 6315 addresses. Fig 6.b depicts the distribution of the dataset according to the amount of ETH transferred. We also split the dataset into 5 sets with an average ETH amount of 95.23, 2.0, 0.25, 0.01, 4.6×10^{-6} , the numbers of nodes are 100, 200, 700, 1000, and 4315 respectively. The distribution shows that a large amount of ETH transferred is concentrated in a small number of addresses, which possibly contributes to the high value and reputation of the network.

One key factor that affects the PageRank performance is the number of connected clusters on the graph and the number of nodes in these clusters. For isolated clusters with a smaller number of nodes, the nodes in these clusters should have smaller credit scores. However, these nodes have the ability to work as sinks and absorb the PageRank score, resulting in poor ranking results. The damping factor is included to prevent sinks from absorbing the PageRank of the nodes connected to the sinks. It is the probability of being on a random node after restarting. Reducing the tiny, isolated clusters containing few nodes will not significantly affect the entire transaction network, and selecting the appropriate damping factor can give a better ranking result.

Figure 6. Distribution of (a) group of the dataset according to the number of transactions, and (b) group of the dataset according to the amount of ETH transferred

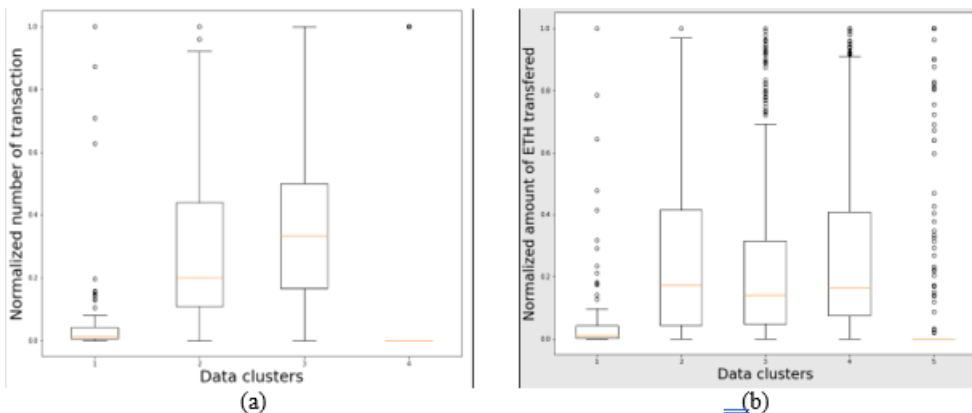


Fig 7 is the result of the clustering algorithm, describing the distribution of the number of nodes and isolated clusters on the entire dataset. As shown in the figure, the number of isolated clusters with more than 10 nodes accounts for only 5% of the total number of clusters but achieves 83% of the number of nodes in the entire network. Therefore, conducting experiments on clusters with a large number of nodes makes more sense for credit scoring. The number of nodes in a cluster can be considered as an indicator to investigate and evaluate the efficiency of PageRank.

Results

We applied PageRank to the post-processing Ethereum transaction dataset to examine the ranking performance. The ranking results based on PageRank are compared with some basic analysis indicators mentioned above such as NoT, NoR, node degree, and spent gas. We also examine the damping factor value and isolated clusters to find the parameter value for the best-ranking performance. Finally, based on PageRank score results, the high-ranking nodes with high credit scores are identified, proving the correctness of the ranking experiment.

Kendall's Tau Correlation

Kendall's tau correlation coefficient, developed by Maurice Kendall, is a nonparametric measure of the strength and direction of association that exists between two variables measured on at least an ordinal scale (Abdi, 2007). It is usually used as a measure of rank correlation. Intuitively, the Kendall correlation between two variables will be high when observations have a similar rank between the two variables and low when observations have a dissimilar rank between the two variables. The PageRank algorithm was run on the chosen Ethereum data sample with 34153 nodes (addresses), and 29742 edges (transactions). Damping factor is initially selected with a value of 0.85. Kendall's tau score between PageRank, NoT rank, NoR rank, Node degree rank, and spent gas rank are given in Table 2.

Figure 7. Distribution of groups of the normalized dataset

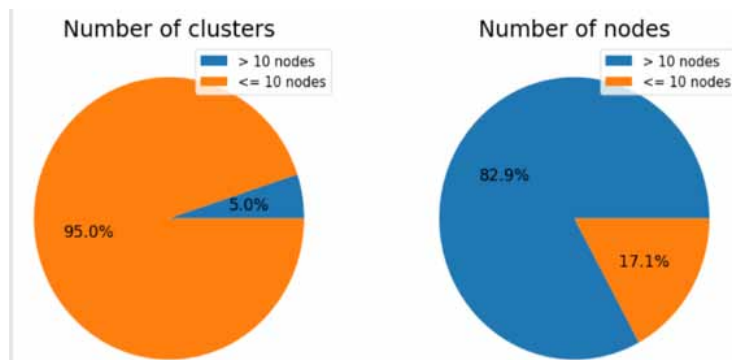


Table 2. Kendall's Tau correlation between PageRank and basis analysis indicators

Ranking vs Indicator	Kendall's tau correlation scores
PageRank-NoR rank	0.8975
PageRank-NoT rank	-0.082
PageRank- Node Degree Rank	0.416
PageRank – receiving spent gas rank	0.894
PageRank-Total spent gas rank	-0.349

Table 2 shows that PageRank has the highest correlation for the ranking of nodes with the NoR rank and the rank of spent gas for received transactions. That is, the node with the higher number of received transactions and the greater amount of spent gas for received transactions, the higher the score.

Kendall's Tau correlation results with HodgeRank are shown in Table 3. On the Ethereum transaction network, there are a lot of transactions has zero value in ETH transferred. This phenomenon happened because some smart contracts have been implemented without transferring any ETH. These transactions lead to zero weights on the edge of the graph. To overcome this problem, we fill the zero value with the mean value on that transaction. The result of the HodgeRank algorithm on the original dataset and the modified dataset is then called HodgeRank (1) and HodgeRank (2) respectively.

We also compare the ranking result with some basic indicators such as the amount of sending ETH, and the amount of receiving ETH. As shown in Table 3, the correlation between HodgeRank and some basic indicators is pretty low, so HodgeRank can be considered as an indicator to evaluate the important role of a node on the graph. However, because of the expensive cost of computation of the batch HodgeRank, only a small number of entities are examined. Besides, the identification of HodgeRank showed a poor result, so in this study, we focus more on optimizing the parameters of PageRank algorithms.

Damping Factor Investigation

The experiment with $d = 0.85$ on the entire dataset shows that some nodes achieving high ranking scores have low social connections and belong to some small clusters. A large number of tiny, isolated clusters can be considered sinks, hence they absorb ranking. With the large damping factor, the probability of being on a random node is low, resulting in the node in the sinks achieving a high-ranking score. To obtain a better PageRank result, we vary on a range of damping factor points from 0.1 to 0.9. Besides, the isolated tiny clusters are also eliminated. This means that the PageRank is only performed on large clusters, which are more meaningful for the network. PageRank results are compared with NoR rank through Kendall's tau correlation.

Table 3. Kendall's Tau correlation between HodgeRank and basis analysis indicators

Ranking vs Indicator	Kendall's tau correlation scores
HodgeRank (1) – HodgeRank (2)	0.394
HodgeRank (1) – Sending ETH Rank	0.2
HodgeRank (1) – Receiving ETH Rank	-0.14

Figure 8. PageRank – NoR rank correlation for damping factor value

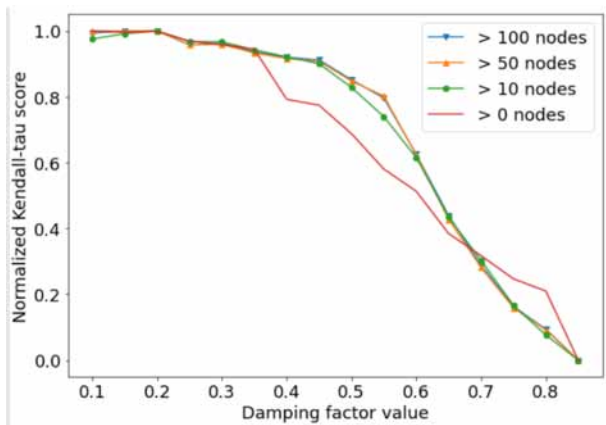


Fig 8 describes the change of Kendall's tau score between PageRank and NoR rank according to damping factor and filtering of the isolated tiny clusters. The results show that the smaller the damping factor leads the greater the correlation score. In fact, with a small value of the damping factor, the high-ranking nodes can also be better identified. Nodes with low connections and belonging to small clusters are gradually removed from the high-ranking group. In particular, with the clusters having higher 100 nodes and $d < 0.5$, the correlation score becomes more stable showing little change in the ranking results. Changing the damping factor and filtering out isolated tiny clusters helps to achieve better ranking results. Through this experiment, we choose $d = 0.5$ and filter out clusters with less than 10 nodes to perform PageRank to identify nodes with high credit scores.

High Ranking Nodes Identification

The higher the rank of the node, the higher the credit score of these nodes. Such nodes can be considered as entities that play a crucial role in the network, often having the ability to identify and trust to perform transactions such as deposits, loans, and other complex financial services. The nodes with low credit scores will have fewer benefits to use financial systems. After having PageRank results with reasonable parameters, we identify high-ranking nodes. Table 4 depicts the identification results of the top 20 nodes.

The results show that most of the high-ranking nodes are identifiable. These nodes are of different types like exchanges, dapps, NFT marketplaces, and protocol contracts. This proves that the algorithm is effective in calculating credit scores for entities on the network, being able to identify nodes with important and trusted roles.

Table 4. High-rank nodes identification

Node ID	PageRank Score	Wallet Address	Identification
49	0.03882566	0xa090e606e30bd747d4e6245a1517ebe430f0057e	Coinbase: Miscellaneous
10	0.033985347	0x7a250d5630b4cf539739df2c5dacb4c659f2488d	Uniswap V2: Router 2
7	0.031337637	0xdac17f958d2ee523a2206206994597c13d831ec7	Tether: USDT Stablecoin
207	0.010002127	0xeb2629a2734e272bcc07bda959863f316f4bd4cf	Coinbase6
22	0.007394333	0xa0b86991c6218b36c1d19d4a2e9eb0ce3606eb48	Centre: USD Coin
59	0.005696202	0xe592427a0aece92de3edee1f18e0157c05861564	Uniswap V3: Router
193	0.005527305	0x6c42627f2b140c1fc5c2cdb653d077e57abb95a8	Metamask: Swap Router
44	0.00320865	0x881d40237659c251811cec9c364ef91dc08d300c	
50	0.002677954	0x28c6c06298d514db089934071355e5743bf21d60	Binance14
594	0.002439676	0xa0c68c638235ee32657e8f720a23cec1bfc77c77	Polygon (Matic): Bridge
570	0.002359207	0x6b175474e89094c44da98b954eedeac495271d0f	Dai Stablecoin
128	0.001987958	0x7d1afa7b718fb893db30a3abc0cfc608aacfebb0	Polygon (Matic): Matic Token
453	0.001915217	0xda977e722ec56f9692a37d84a4e3212dca58114d	
338	0.001787715	0x7be8076f4ea4a4ad08075c2508e481d6c946d12b	OpenSea: Wyvern Exchange v1
524	0.00159456	0xfaff15c6cdaca61a4f87d329689293e07c98f578	Zapper.Fi: Zapper NFT
573	0.001581834	0xc02aaa39b223fe8d0a0e5c4f2ead9083c756cc2	Wrapped Ether
176	0.001502355	0x95ad61b0a150d79219dcf64e1e6cc01f0b64c4ce	Shiba Inu: SHIB Token
1410	0.00146114	0x1db92e2eebc8e0c075a02bea49a2935bcd2dfcf4	
371	0.001450488	0xd9e1ce1f2641f24ae83637ab66a2cca9c378b9f	SushiSwap: Router

CONCLUSION AND FUTURE STUDY

This study is modest compared to the big landscape of Blockchain space and the credit industry. We have tried to approach the problem and give some experimental results on ranking entities (addresses) based on transactions on the Ethereum blockchain. The research possibly has potential applications in the social credit and reputation systems in Blockchain and crypto space.

The corresponding author published an investigation (Thuat Do et al., 2019) on blockchain ranking framework that applied PageRank, HodgeRank to compute ranking scores based on transactions. The results were proposed to build a reputation mechanism adding on top of Proof of Stake consensus. This paper is a continuation of the mentioned research but focuses deeper on the PageRank algorithm and transactions on Ethereum.

In the future, we shall continue and extend the research on a larger dataset, on multiple blockchains, and examine several mathematical ranking algorithms (e.g. online PageRank, online HodgeRank). As batch HodgeRank is expensive, new directions of HodgeRank such as online HodgeRank (Xu et al., 2013) can be the solution to increase the number of transaction data and improve the performance of HodgeRank in future studies.

A Public Github repo of the sample dataset and the code is available on (Do, 2022).

REFERENCES

- Abdi, H. (2007). The Kendall rank correlation coefficient. *Encyclopedia of measurement and statistics*, 2, 508–510.
- Allen, D., & Evgeny, M. (2018). *Ethereum in BigQuery: a Public Dataset for smart contract analytics*. Google. <https://cloud.google.com/blog/products/data-analytics/ethereum-bigquery-public-datas-et-smart-contract-analytics>.
- Arya, S., Eckel, C., & Wichman, C. (2013). Anatomy of the credit score. *Journal of Economic Behavior & Organization*, 95, 175–185. doi:10.1016/j.jebo.2011.05.005
- Colendi. (2019). *Blockchain and Decentralization*. Colendi. [https://www.colendi.com/blog?title=block chain-and-decentralization&id=12](https://www.colendi.com/blog?title=block+chain-and-decentralization&id=12).
- Do, H. D. (2022) *Blockchain based credit network*. Commit 6512a47. [https://github.com/ Octan-Finance/Blockchain_based_credit_network](https://github.com/Octan-Finance/Blockchain_based_credit_network).
- Do, T., Nguyen, T., & Pham, H. (2019). *Delegated Proof of Reputation: a Novel Blockchain Consensus*. In *International Electronics Communication Conference*, Okinawa, Japan.
- Dym, C. L., Wood, W. H., & Scott, M. J. (2002). Rank ordering engineering designs: Pairwise comparison charts and Borda counts. *Research in Engineering Design*, 13(4), 236–242.
- Fiocco, M. (2019). *Banking for the Unbanked: The Promises, Pitfalls and Potentials of Mobile Banking*. University of Ottawa.
- Hassija, V., Bansal, G., Chamola, V., Kumar, N., & Guizani, M. (2020). Secure lending: Blockchain and prospect theory-based decentralized credit scoring model. *IEEE Transactions on Network Science and Engineering*, 7(4), 2566–2575. doi:10.1109/TNSE.2020.2982488
- Jiang, X., Lim, L. H., Yao, Y., & Ye, Y. (2011). Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(1), 203–244.
- McWilliams, G. (2020). *A Study of the Social Credit Score System and the Implications it may have in Chinese Society*. University of Mississippi.
- Page, L. (1999). *The PageRank Citation Ranking: Bring Order to the Web* [EB OL]. Stanford University.
- Xu, Q., Xiong, J., Huang, Q., & Yao, Y. (2013). Online HodgeRank on random graphs for crowdsorceable QoE evaluation. *IEEE Transactions on Multimedia*, 16(2), 373–386.
- Xue, L. (2021, December). *The Application of Blockchain Technology in the Financial Field*. In *2021 International Conference on Forthcoming Networks and Sustainability in AIoT Era (FoNeS-AIoT)*, Nicosia, Turkey. doi:10.1109/FoNeS-AIoT54873.2021.00036
- Zhou, L., Qin, K., Torres, C. F., Le, D. V., & Gervais, A. (2021, May). High-frequency trading on decentralized on-chain exchanges. In *Symposium on Security and Privacy (SP)*. IEEE. doi:10.1109/SP40001.2021.00027

Huu-Dung Do received the B.E. degree in electronics and telecommunications engineering from the Hanoi University of Science Technology, Vietnam, in 2018, and the M.S. degree from the Department of Electronic Engineering, Pukyong National University, South Korea, in 2021. He is currently a Researcher at FPT Blockchain Lab, FPT Corporation. His research interests include blockchain, computer vision, and machine learning.

Thuat Do is a PhD candidate at Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong. His research focuses on Blockchain, cryptocurrency, DeFi and their applications. He has rich experience in developing Blockchain infra, multichain NFT marketplace and DeFi facilities, as well as enterprise blockchain applications. He is Head of Development, FPT Blockchain Labs, FPT Corporation. Thuat Do is the corresponding & leading author of this research article.