Discovering Co-Occurrence Patterns Among Blockchain Address Categories Using the FP-Growth Association Mining Algorithm

Latasha Lenus[1,*]

[1]Computer Science and Design, Singapore University of Technology and Design, Singapore

ABSTRACT

This paper focuses on identifying recurring patterns among blockchain address categories using the FP-Growth algorithm, which is known for its efficiency in mining frequent itemsets within large datasets. The study provides insights into blockchain ecosystem dynamics by analyzing category associations across different blockchain networks like Ethereum and Bitcoin. Through this analysis, significant patterns were found, such as the frequent co-occurrence of categories related to smart contracts and exchanges, highlighting the central role of these categories in blockchain interactions. Additionally, the study delves into the influence of data sources on detected patterns, revealing that various data collection methods contribute to distinct biases, which affect category associations. The findings offer practical applications for blockchain analytics, such as improving classification models, anomaly detection, and enhancing regulatory compliance. This study contributes to blockchain research by showcasing how association rule mining can improve the categorization and understanding of blockchain address behaviors. The use of FP-Growth, as opposed to more traditional methods, enables faster and more comprehensive analysis, which is particularly valuable given the extensive nature of blockchain datasets. The research also points to potential directions for future work, such as integrating temporal data to observe changes over time and exploring additional blockchain networks to broaden the scope of insights. The study emphasizes the need for continuous advancements in blockchain address analysis to support security, transparency, and regulatory initiatives within this rapidly evolving digital ecosystem.

Keywords FP-Growth algorithm, blockchain address categorization, association rule mining, blockchain analytics, data source biases

Introduction

Blockchain technology represents a decentralized, distributed ledger system that facilitates secure and transparent transactions across a network of computers or nodes. Originally developed to support Bitcoin, blockchain has rapidly evolved to encompass a wide range of applications across various industries, including finance, healthcare, supply chain management, and governance. Its fundamental characteristics—decentralization, immutability, and transparency—enable blockchain to enhance trust and security in digital transactions [1]. This tamper-resistant record of transactions is achieved through a unique structure where each transaction is grouped into a block and linked cryptographically to the preceding one. This chain of blocks ensures that once data is recorded, it cannot be altered without the consensus of the network, thereby safeguarding against fraud and unauthorized access. Additionally, consensus mechanisms like proof of work and proof of stake

Submitted 27 December 2024
Accepted 2 February 2025
Published 8 March 2025

Corresponding author
Latasha Lenus,
13CSD0034@sutd.edu.sg

Additional Information and Declarations can be found on page 59

DOI: 10.47738/jcrb.v2i1.24

How to cite this article: L. Lenus, “Discovering Co-Occurrence Patterns Among Blockchain Address Categories Using the FP-Growth Association Mining Algorithm,” *J. Curr. Res. Blockchain*, vol. 2, no. 1, pp. 41-62, 2025.

bolster the security and reliability of blockchain systems by establishing a distributed verification process. This attribute is especially crucial in sectors such as healthcare, where the integrity of sensitive patient data must be maintained [2].

Beyond transaction security, blockchain technology offers innovative capabilities through smart contracts, which are self-executing contracts that enforce agreements automatically based on predefined conditions. These contracts significantly reduce the need for intermediaries, streamlining operations and lowering costs [3]. In supply chain management, for instance, blockchain can provide real-time tracking of goods, fostering transparency and accountability among stakeholders. The potential of blockchain to revolutionize traditional business models is substantial, as it promotes new forms of collaboration and value creation across industries [4]. Thus, as blockchain continues to gain traction across various sectors, its impact on enhancing efficiency and transparency in business operations is anticipated to grow, solidifying its role as a transformative force in data management and transaction processing [5], [6], [7].

Blockchain networks serve as essential platforms for implementing and executing a wide array of decentralized applications and services, with each network offering unique features that cater to specific use cases. Bitcoin, the first and most widely recognized cryptocurrency, operates on a proof-of-work (PoW) consensus mechanism, where miners solve complex mathematical problems to validate transactions. Its primary role as a decentralized digital currency is to enable peer-to-peer transactions without intermediaries, offering a secure and immutable platform upheld by a vast network of miners. Ethereum, launched in 2015, extends blockchain's functionality by introducing smart contracts, allowing developers to build decentralized applications (DApps) across domains such as decentralized finance (DeFi) and non-fungible tokens (NFTs). Ethereum’s proof-of-stake (PoS) mechanism further enhances its scalability and energy efficiency, making it a critical innovation hub within the blockchain ecosystem.

Other networks, such as BNB Chain, Polygon, and Avalanche C-Chain, contribute to the blockchain landscape by addressing specific challenges like transaction speed, cost, and interoperability. BNB Chain combines proof-of-stake and proof-of-authority mechanisms to support fast and low-cost transactions, with a particular emphasis on DeFi applications. Polygon functions as a Layer 2 scaling solution for Ethereum, utilizing sidechains and plasma chains to improve transaction throughput while maintaining security, thus enabling a variety of DApps and DeFi projects to flourish with enhanced scalability [8]. Lastly, Avalanche C-Chain, part of the broader Avalanche network, provides an Ethereum-compatible platform that supports rapid transaction processing through its Avalanche consensus mechanism, positioning it as a competitor in sectors like DeFi and NFTs. Collectively, these networks underscore the diverse and evolving nature of blockchain technology, highlighting the adaptability and versatility of blockchain as a foundation for innovative digital solutions across industries.

Cryptocurrency addresses are foundational to the blockchain ecosystem, acting as unique identifiers that facilitate transactions across decentralized networks. Each address, derived from a public key, enables users to send and receive digital assets while securely recording transactions on the blockchain. Beyond simple transaction facilitation, these addresses are integral to the structure and security of blockchain systems, allowing them to maintain a decentralized and transparent framework. For example, Bitcoin addresses are pseudonymous, meaning while transactions are transparent, the identities behind the addresses remain obscured. This pseudo-anonymity provides users with a degree of privacy; however, it also introduces challenges to regulatory compliance and the prevention of illicit activities, such as money laundering and fraud. Although Bitcoin transactions are often perceived as anonymous, analyses of transaction patterns can sometimes reveal the identities behind certain addresses, underscoring the need for deeper understanding and scrutiny of address behaviors within blockchain networks.

Cryptocurrency addresses also demonstrate varied activity patterns, with research showing a long-tail distribution, where a small number of addresses account for most transactions. This uneven activity distribution provides insights into user behavior and reflects the broader health and dynamics of the cryptocurrency economy. By studying these transaction patterns, researchers can identify prominent market participants and better understand the flow of funds within blockchain networks. Another key aspect of address analysis is address clustering, which links multiple addresses to a single entity based on observed transaction patterns and heuristics. Clustering facilitates a clearer picture of user behavior and the relationships between different entities, which is especially valuable for forensic investigations and regulatory compliance efforts [9]. These functionalities—enabling both transaction privacy and analytical insights—demonstrate the complex and essential role of cryptocurrency addresses in maintaining the security and integrity of blockchain ecosystems.

Categorizing cryptocurrency addresses is critical for enhancing security, supporting analytics, and facilitating regulatory compliance within blockchain networks. As cryptocurrencies continue to grow in popularity, the need for effective address classification systems has become increasingly important to manage the challenges posed by pseudonymity and the potential for misuse. In terms of security, the anonymous nature of blockchain transactions can attract malicious actors, and categorizing addresses can help security analysts track suspicious activities, such as those associated with scams or fraud. Address classification is essential for user protection, as it allows for the identification of various threat types and their connections within the ecosystem. De-anonymizing clusters of addresses also strengthen forensic investigations, providing key insights into illicit activities by revealing connections among entities involved in illegal operation.

From an analytics perspective, categorizing addresses enhances the understanding of user behavior and market trends. Analyzing transaction patterns and clustering related addresses allows analysts to identify significant players in the cryptocurrency economy, assess the market's overall health, and trace the movement of funds. By examining address categories and their transaction histories, researchers can perform comprehensive analyses of fund flows, which are crucial for market research, risk assessment, and economic modeling within the cryptocurrency space. Address categorization also supports regulatory compliance by helping authorities monitor for activities linked to money laundering, terrorism financing, and other illicit uses. As regulators implement stricter requirements for cryptocurrency exchanges and service providers, classifying addresses based on entity types and behaviors enables a more effective response to these mandates. Importance of regulatory frameworks that prioritize data privacy and security, both of which can be bolstered by robust address classification methods. In sum, categorizing cryptocurrency addresses is indispensable for maintaining the integrity of blockchain ecosystems, offering protection to users, and ensuring adherence to regulatory standards.

Address categorization in the blockchain environment presents significant challenges due to the complexity of multiple category assignments and the sheer volume of data involved. Many blockchain addresses belong to more than one category, reflecting a range of functions or activities. For instance, a single address may be associated with both exchange and gaming activities or may operate within DeFi (decentralized finance) while also engaging in smart contract operations. This multi-label characteristic complicates the process of categorizing addresses accurately, as each address can represent diverse and overlapping roles. Additionally, this study involves a dataset containing 10 million blockchain addresses, which demands efficient data processing techniques to manage and analyze the vast amount of information effectively. Handling large-scale data of this magnitude is resource-intensive and requires optimized methods to extract meaningful insights within a reasonable time frame.

To address these challenges, identifying meaningful patterns among categories becomes essential. Association mining methods, such as the FP-Growth algorithm, enable the discovery of frequent co-occurrences among categories, offering insights that enhance classification accuracy and address understanding. Traditional mining techniques often need help with multi-label data and may fail to capture complex relationships among categories effectively. Efficient association mining helps not only to uncover these hidden relationships but also to streamline the categorization process by focusing on category combinations that occur frequently. Recognizing patterns of co-occurrence among blockchain address categories can provide valuable information on ecosystem dynamics, revealing how different categories interact and identifying prevalent patterns that may indicate significant behaviors or trends within the blockchain landscape.

The exploration into discount strategies on consumer ratings and the comparative analysis of sentiment classification techniques underscore how nuanced approaches to consumer behavior can unveil critical insights in digital markets [10], [11]. Similarly, sentiment trends in Bitcoin-related tweets and predictive modeling of blockchain stability present novel opportunities to analyze decentralized networks using advanced clustering and machine learning methodologies [12], [13]. Further, a comprehensive analysis of Twitter conversations in the metaverse space highlights shifting public sentiment, while an empirical analysis of virtual property prices in Decentraland examines evolving digital asset markets, providing context to dynamic economic patterns [14], [15]. Combined, these works illustrate the diverse applications of data-driven methods in analyzing complex behaviors within blockchain ecosystems and beyond.

The primary goal of this research is to discover and analyze co-occurrence patterns among blockchain address categories, with the aim of enhancing the understanding of address behaviors within blockchain networks. By focusing on co-occurrence patterns, the study seeks to reveal how frequently different categories are associated with one another and how these patterns differ across various blockchain networks, such as Bitcoin and Ethereum. Understanding these patterns provides a foundation for developing more accurate classification models that capture the complexities of multi-label address categorization.

The research objectives are centered on utilizing the FP-Growth algorithm to efficiently mine frequent category combinations among blockchain addresses. Specifically, the study aims to compare association patterns across different blockchain networks to observe variations and commonalities in address behaviors. Additionally, the research assesses the influence of data sources on category associations, recognizing that variations in data collection methods and sources can affect the types of associations identified. The study provides insights into how address categories co-occur within the blockchain ecosystem, with the potential to inform both research and practical applications in blockchain analytics.

This study contributes to blockchain analytics by enhancing the understanding of address behaviors and the dynamics of the blockchain ecosystem. By identifying and analyzing co-occurrence patterns among blockchain address categories, the research provides insights into how different types of addresses interact within blockchain networks. This knowledge deepens the understanding of blockchain ecosystem dynamics, highlighting prevalent behaviors and identifying potential indicators of significant activities. Furthermore, the study’s findings can inform the development of more nuanced classification models that accurately reflect the multi-functional nature of blockchain addresses, improving the precision of address categorization in both academic research and industry applications.

From a practical perspective, the study’s findings have implications for enhancing security and regulatory measures within blockchain systems. Efficient and accurate address categorization aids in the detection of anomalies, such as unusual patterns that may indicate fraudulent activities or other security threats. Additionally, by improving the classification of blockchain addresses, the study supports the development of tools that assist regulatory bodies in monitoring blockchain activities, promoting transparency, and ensuring compliance with financial regulations. Overall, the study advances both the theoretical understanding and practical application of blockchain address categorization, contributing to the broader field of blockchain analytics and security.

Literature Review

Blockchain Address Classification

In the field of blockchain analytics, a range of classification techniques has been developed to improve the interpretation and utility of blockchain data. Various methods leverage machine learning, anomaly detection, and ensemble learning techniques to classify and analyze blockchain transactions effectively. For example, cascading machine learning models have been widely used to classify Bitcoin entities. Zola et al. demonstrated the effectiveness of this approach by utilizing input features from blockchain data to classify different Bitcoin entities, achieving impressive accuracy rates that highlight the robustness of ensemble learning techniques in this context [16]. Ensemble methods, which involve training multiple classifiers on diverse datasets, enhance the overall classification process by combining multiple perspectives and improving predictive performance [16]. Through this approach, researchers have been able to extract meaningful behavioral patterns, which contribute to a deeper understanding of blockchain transactions and potential anomalies within the network.

Another significant advancement in classification techniques is the use of decision trees within distributed networks, which facilitate secure and privacy-preserving data sharing. Blockchain-based ID3 decision tree framework, which leverages homomorphic encryption to maintain data privacy while enhancing classification accuracy. This method is particularly valuable in applications where data integrity and confidentiality are paramount, such as in healthcare-related blockchain analytics [17]. Anomaly detection has also become a critical component of blockchain analytics. Martin et al. explored the application of machine learning and network representation methods to detect unusual transaction patterns, demonstrating that supervised learning methods outperform unsupervised approaches in this context [18]. Additionally, dynamic attribute graph anomaly detection method uses graph attention mechanisms to capture temporal features, further improving the accuracy of blockchain transaction classification. Together, these approaches provide a diverse toolkit for enhancing security, understanding user behaviors, and detecting potential threats within blockchain systems.

Multi-label classification (MLC) presents distinct challenges in cases where each instance, such as a blockchain address, may belong to multiple categories simultaneously. Unlike traditional single-label classification, MLC requires models to account for multiple, non-exclusive labels, adding layers of complexity related to computational efficiency, prediction accuracy, and evaluation reliability. The computational demands of MLC can be significant, particularly as the number of label combinations grows exponentially. The classification process becomes increasingly challenging as the number of potential label combinations expands. To address this, researchers have introduced ensemble-based approaches that combine multiple classifiers, balancing improved performance with reduced computational overhead. Hemavati et al. developed a multi-layered stacked ensemble method that achieves dimensionality reduction while preserving classification accuracy, effectively streamlining the MLC process [19].

Prediction accuracy in MLC also faces obstacles due to label dependencies, where the relationships between labels can affect the accuracy of predicted label sets. Venkatesan et al. pointed out that MLC is inherently more complex than single-label classification because of the need to consider label correlations [20]. In response to these challenges, advanced models have been developed, such as Tang et al.'s improved Transformer model, which captures complex relationships between labels to improve prediction accuracy [21]. Another concern is the reliability of evaluation metrics in MLC. Traditional accuracy measures may not adequately reflect model performance due to label imbalances, making it essential to utilize metrics that account for the multi-label nature of the data. Charte et al. emphasized the importance of tailored evaluation metrics and resampling algorithms to ensure a reliable assessment of MLC models [22]. Through these innovations, researchers continue to address the unique challenges of MLC, leveraging advanced algorithms to enhance the accuracy and reliability of multi-label classification systems in blockchain analytics.

Address categorization in blockchain analytics is a critical area that focuses on classifying blockchain addresses based on their usage patterns, behaviors, and associated entities. This process is fundamental for improving transaction insights, enhancing security, and enabling compliance with regulatory requirements. A notable study in this domain is Le's exploration of an Autonomous Coin Mixer (AMR), which addresses privacy concerns by clustering addresses to minimize the risk of deanonymization in blockchain transactions [23]. This work underscores the dual challenge of maintaining privacy for users while effectively categorizing addresses to improve blockchain transparency and security. Similarly, Sahoo et al. proposed a hierarchical model for categorizing nodes within blockchain networks, which can be applied to address classification by defining address roles and behaviors [24]. This model provides a structured framework for understanding the functions of various addresses in the blockchain ecosystem, aiding in regulatory compliance and transaction monitoring.

Another important contribution is Joshi et al.'s comprehensive survey on blockchain security and privacy issues, which discusses address categorization as a vital tool for enhancing security. Joshi and colleagues argue that categorizing blockchain addresses allows for better identification of security risks, improving the robustness of blockchain systems against illicit activities [25]. By optimizing address categorization processes, blockchain systems can potentially handle more transactions, thus contributing to scalability and overall system performance. These studies collectively highlight the importance of address categorization as a means to improve both the functionality and security of blockchain networks, addressing critical issues such as privacy, scalability, and compliance within the evolving blockchain landscape.

Association Rule Mining in Blockchain

Association rule mining (ARM) is a key data mining technique focused on uncovering significant relationships and patterns within large datasets. First introduced by Agrawal et al. in 1993, ARM has since evolved as an essential tool across multiple domains, from retail to healthcare and beyond [26], [27]. Its primary objective is to identify strong association rules that reflect co-occurrence relationships between items in transactional datasets, typically expressed in "if-then" statements, such as X → Y, where X and Y represent itemsets [28], [29]. The appeal of ARM lies in its ability to derive actionable insights from extensive data, enabling organizations to understand underlying patterns that can inform strategic decisions and enhance operational efficiency.

ARM has substantial relevance in practical applications. For instance, in market basket analysis, retailers use ARM to determine frequently co-purchased products, which can improve store layout, cross-selling opportunities, and inventory management strategies [30], [31]. Beyond retail, ARM is employed in healthcare to analyze treatment outcomes, aiding clinicians in predicting patient responses based on treatment combinations. It is also applied in finance for fraud detection, where ARM identifies irregular transaction patterns that may signal fraudulent activities [32], [33]. Typically, ARM involves two steps: first, identifying frequent itemsets that meet a minimum support threshold, and second, generating association rules that fulfill a specified confidence level [28], [29]. To accommodate growing data volumes, various algorithms, such as

Apriori and FP-Growth, have been developed to optimize ARM's efficiency by reducing computational demands and enhancing rule quality [34], [35]. Recent advances also include mining both positive and negative associations, offering a holistic view of relationships within datasets (Jiang et al., 2008; Mani, 2012). Thus, ARM remains an invaluable asset for uncovering hidden relationships within data, driving informed decision-making across industries.

The Apriori and FP-Growth algorithms are two of the most widely used methods in ARM, each with distinct advantages and limitations suited to various data mining tasks. The Apriori algorithm, pioneered by Agrawal et al. in 1994, employs an iterative approach to identify frequent itemsets by generating and pruning candidate itemsets based on a minimum support threshold, a process that requires multiple database scans [36]. Apriori's main strength lies in its simplicity and ease of implementation, making it a popular choice for introductory applications and smaller datasets. However, this algorithm becomes less efficient with larger datasets, as the necessity for multiple scans increases computational overhead and memory usage, often resulting in longer execution times and challenges in real-time applications [37].

In contrast, the FP-Growth algorithm, which was developed to overcome Apriori's limitations, operates by constructing a compact data structure known as the FP-tree. This structure enables FP-Growth to mine frequent item sets without generating candidate sets, reducing database scans to just two [38]. The primary advantages of FP-Growth include its ability to handle large datasets efficiently and to quickly generate frequent item sets, making it well-suited for applications with high transaction volumes [38]. Additionally, the tree structure used by FP-Growth improves memory management compared to Apriori's candidate generation approach [38]. However, FP-Growth can be more challenging to implement and may present a steeper learning curve, particularly for practitioners new to data mining [38] . Thus, while Apriori is user-friendly and effective for smaller datasets, FP-Growth provides a more scalable and efficient solution for large-scale data mining tasks, highlighting the importance of choosing an algorithm that aligns with the specific requirements of the data analysis.

FP-Growth Algorithm

The FP-Growth algorithm is a highly efficient method for mining frequent item sets within large datasets, particularly when compared to traditional approaches like the Apriori algorithm. At the core of FP-Growth is the Frequent Pattern Tree (FP-Tree) structure, which allows the algorithm to compress the transaction database into a compact form, eliminating the need for candidate generation and significantly reducing the computational load [39]. The algorithm operates in two primary phases: FP-Tree construction and frequent itemset mining. In the first phase, the algorithm performs a preliminary scan of the dataset to identify all items that meet a specified minimum support threshold, filtering out infrequent items to focus on those with significant co-occurrence. After determining the frequency of each item, the transactions are sorted in descending order of item frequency, which aids in constructing a compact and efficient tree structure.

In the second phase, the FP-Growth algorithm mines the FP-Tree for frequent itemsets by recursively building conditional FP-Trees. These conditional trees represent subsets of the original FP-Tree, each corresponding to a frequent item. By traversing these smaller trees, the algorithm can efficiently extract all frequent patterns associated with each item, employing a divide-and-conquer strategy that further enhances performance [40], [41]. This approach allows the algorithm to bypass the resource-intensive candidate generation step, a common bottleneck in other algorithms, thus providing a scalable solution for frequent itemset mining. Additionally, FP-Growth's recursive mining process is tailored for parallelization, which can be advantageous when handling large, complex datasets [42]. The FP-Growth algorithm's efficiency and scalability make it an invaluable tool in various data mining applications, enabling rapid extraction of meaningful patterns from vast amounts of transactional data [43], [44].

The FP-Tree, or Frequent Pattern Tree, is the foundation of the FP-Growth algorithm, facilitating efficient data compression and mining of frequent item sets without the need for iterative candidate generation. The FP-Tree construction begins with a first pass through the transaction database to count item frequencies, filtering items that do not meet the minimum support threshold. This step generates a sorted list of frequent items, which is then used to organize subsequent transactions in a way that maximizes the sharing of common prefixes [45], [46]. During the second pass, each transaction is transformed into a path in the tree, with items added in descending order of frequency. Suppose items in the transaction share a prefix with existing nodes in the FP-Tree. In that case, they are added as extensions of those nodes, effectively compressing the dataset by consolidating similar transactions [47].

As the FP-Tree is constructed, each node records an item and its frequency count and links to sibling nodes, creating a compact data structure that stores essential information for subsequent mining phases. Additionally, a header table is maintained to track each item's occurrences and provide quick access to nodes within the tree [43]. This structure enables the algorithm to mine frequent item sets by recursively constructing conditional FP-Trees for each frequent item, focusing only on relevant subsets of the data and thus enhancing mining efficiency. The ability to represent the transaction database in a tree format allows the FP-Growth algorithm to effectively compress large volumes of data, providing an optimized pathway for extracting frequent patterns [41], [48]. The FP-Tree's compact design not only improves memory usage but also supports rapid traversal and retrieval of frequent itemsets, reinforcing the FP-Growth algorithm's reputation as a scalable and powerful data mining tool [49].

Method

The research method for this study consists of several steps to ensure a comprehensive and accurate analysis. The flowchart in Figures 1 outlines the detailed steps of the research method.

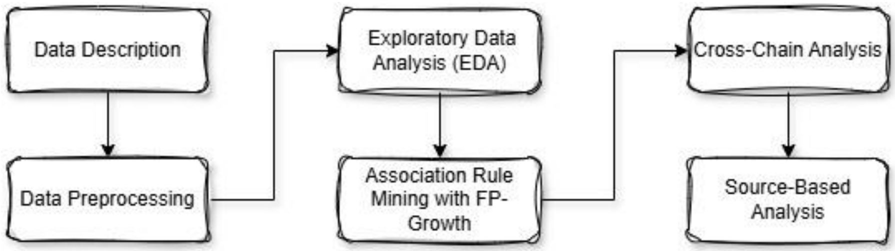

Figure 1 Research Method Flowchart

Data Description

The dataset used in this study, `dataset_10m_ads.csv`, comprises 10 million labeled cryptocurrency addresses, offering a robust foundation for analyzing blockchain address categories and their co-occurrence patterns. The dataset contains five columns: `chain`, `address`, `categories`, `entity`, and `source`. The `chain` column specifies the blockchain network associated with each address, such as Bitcoin mainnet, Ethereum mainnet, BNB Chain mainnet, Polygon mainnet, or Avalanche C-Chain. The `address` column represents the unique identifier for each cryptocurrency address, serving as the primary reference for transactions within the respective blockchain network. The `categories` column provides a multi-label classification for each address, reflecting its association with up to 62 possible categories, such as decentralized finance (DeFi), centralized exchanges (cex), smart contracts, and more. This column allows for multi-label assignments, capturing the diverse roles and activities that a single address may exhibit.

The `entity` column contains the associated entity or organization tied to the address, although it may be missing or null for some addresses. When present, it offers valuable context about the address's role in the broader blockchain ecosystem, such as whether it is linked to a specific exchange, wallet service, or other identifiable entity. Finally, the `source` column indicates the origin of the classification data, with values such as `ground_truth`, `heuristic`, or `external`. This field provides insight into the reliability of the data by specifying how the label assignments were determined. Together, these columns form a rich dataset for exploring and analyzing address categories and their relationships within the blockchain environment.

Data preprocessing was critical for ensuring that the dataset was suitable for association mining analysis. Given that the `categories` column allows for multi-label assignments, multi-hot encoding was employed to transform this field into binary columns, representing each unique category as a separate binary indicator. This transformation facilitated the identification of category co-occurrence patterns while enabling efficient data manipulation and analysis. The process involved extracting all unique categories, cleaning whitespace, and creating a set of dummy variables to represent the presence or absence of each category for each address.

Handling missing data was another important aspect of preprocessing. The `entity` field contained a significant number of missing values. To address this, missing `entity` values were filled with the placeholder "Unknown," ensuring that all records retained a consistent format. This approach preserved data integrity while acknowledging the limitations of the available information. Additionally, infrequent categories were filtered out to reduce noise and focus on the most relevant patterns. Categories appearing in fewer than a specified minimum number of addresses were excluded from further analysis, as their low frequency made meaningful pattern identification unlikely. This filtering step refined the dataset, enhancing the quality and interpretability of subsequent analyses by focusing on prominent and recurring category associations.

Exploratory Data Analysis (EDA)

The initial phase of EDA focused on understanding the distribution of blockchain address categories within the dataset. The dataset contained several multi-label categories, which required analyzing the frequency of each category across the entire dataset. The most frequently occurring categories included "smart_contract," "exchange," and "liquid_staking," with counts of 26,875, 18,345, and 14,956, respectively. These categories accounted for a substantial portion of the data, indicating their prevalence within the blockchain address ecosystem. Conversely, the least frequent categories, such as "business_or_services," "dapp," and "mixer," had significantly lower counts, highlighting the imbalance across different address types. This uneven distribution emphasized the need to focus on dominant categories while filtering out those with limited occurrences for more targeted analysis.

A more granular frequency analysis provided insights into the percentage representation of each category. For instance, "smart_contract" addresses comprised 26.88% of the total dataset, followed by "exchange" at 18.35% and "liquid_staking" at 14.96%. This breakdown underscored the concentration of specific address types in the blockchain ecosystem. Such detailed descriptive statistics helped identify key areas of interest and guided further analysis by shedding light on the dominant behaviors and interactions within the data.

The next step in the EDA involved analyzing category co-occurrences to identify relationships and patterns among different address types. Pairwise co-occurrence frequencies were calculated to determine how often specific categories appeared together within the same address. This analysis revealed that certain category pairs exhibited a high degree of co-occurrence, such as "cex" and "exchange," which appeared together in 3,044 instances. The co-occurrence of "nft" with "smart_contract" and "liquid_staking" with "smart_contract" were also noteworthy, with 1,497 and 341 occurrences, respectively. These pairwise associations highlighted potential functional linkages and shared behaviors within blockchain addresses, offering deeper insights into their interactions.

To provide a clearer picture of the co-occurrence dynamics, a co-occurrence matrix was constructed. This matrix captured the frequencies of category pairings, helping to quantify the strength of relationships among categories. The top co-occurring pairs underscored the prevalence of specific patterns within the dataset, providing a foundation for further exploration of complex associations and dependencies among blockchain addresses. This step proved crucial for uncovering potentially significant patterns and behaviors, informing subsequent stages of analysis.

Visualizations played a critical role in illustrating the findings from the EDA. Heatmaps were employed to display category co-occurrence matrices, offering a visual representation of the strength and frequency of category relationships. The heatmap for the top 20 categories revealed concentrated clusters of high co-occurrence, emphasizing strong interdependencies among certain address categories. Bar charts were also utilized to depict the most and least frequent categories, providing a clear and intuitive view of the distribution within the dataset. The top categories were shown to have a disproportionately large share of occurrences, while the bottom categories highlighted the presence of niche or less common address types.

To further explore the relational dynamics, network graphs were generated to visualize category relationships and their co-occurrence patterns. These graphs depicted categories as nodes and co-occurrences as edges, with edge thickness representing the strength of the association. The network visualization offered an intuitive way to understand the complexity and structure of interactions among address categories. By focusing on relationships with a co-occurrence threshold, the network graph provided a clear depiction of dominant associations while filtering out weaker, less significant connections. This comprehensive visual analysis underscored the interconnected nature of blockchain address categories and guided the identification of key patterns and trends within the dataset.

Association Rule Mining with FP-Growth

To apply the FP-Growth algorithm on the preprocessed blockchain address dataset, the first step involved loading the dataset into a DataFrame and ensuring that all category columns were converted into a Boolean format. This conversion was necessary for efficient frequent itemset mining, as the FP-Growth algorithm requires binary inputs for each category to indicate the presence or absence of a category for each address. The FP-Growth algorithm was then applied using a minimum support threshold of 0.5%, meaning that itemsets appearing in at least 0.5% of the transactions were considered frequent. This relatively low threshold was selected to capture meaningful but not overly rare patterns among the categories. The output of this process was a set of frequent itemsets, each representing combinations of categories that occurred together at a frequency above the specified support level.

Once frequent itemsets were identified, association rules were generated based on a minimum confidence threshold of 20%. Confidence measures the likelihood of the consequent category appearing given the presence of the antecedent category. Lowering the confidence threshold to 20% allowed for the generation of a broader range of rules while maintaining a focus on those with statistically meaningful support. This process provided a basis for exploring potential co-occurrence relationships among categories, ultimately leading to a deeper understanding of how various blockchain address categories interact.

The choice of minimum support and confidence thresholds was guided by both theoretical considerations and empirical observations. The minimum support threshold of 0.5% was selected to ensure that frequently co-occurring categories were captured without being overwhelmed by noise from rare occurrences. This balance was necessary for extracting meaningful patterns while avoiding an excessive number of insignificant itemsets that could hinder interpretability. Similarly, a minimum confidence threshold of 20% was chosen to focus on rules with a reasonable degree of predictive reliability. Confidence levels lower than this would have risked generating rules with little practical significance, whereas higher thresholds would have limited the scope of discovered rules (Han et al., 2000).

This combination of support and confidence thresholds enabled a focused yet comprehensive exploration of category associations. Parameter tuning was essential for tailoring the FP-Growth algorithm to the unique characteristics of the dataset, maximizing its ability to uncover co-occurrence patterns. Adjustments to these thresholds would have affected the number and nature of frequent itemsets and rules generated, highlighting the importance of informed parameter selection for effective association mining.

The process of extracting significant association rules from the identified frequent itemsets involved analyzing patterns that met or exceeded the specified support and confidence thresholds. Each rule, represented in the form "antecedent → consequent," illustrated a potential relationship between categories. For example, the rule "cex → exchange" indicated that addresses categorized as centralized exchanges frequently co-occurred with exchange-related activities. This particular rule had a support value of 0.03044 and a confidence level of 23.48%, suggesting a relatively strong association within the data. Similarly, rules such as "nft → smart_contract" highlighted important co-occurrences that offered insight into the interactions between non-fungible token addresses and smart contract functionalities.

The extracted rules were further analyzed based on lift, leverage, and conviction metrics to assess their strength and significance. Lift values greater than one indicated a positive association between the antecedent and consequent categories, while leverage and conviction provided additional context for understanding rule significance and potential causal relationships. This process ensured that only meaningful and actionable rules were retained for further analysis, providing a robust framework for understanding category dynamics within the blockchain ecosystem. The most significant rules were saved and visualized for further interpretation, offering a comprehensive view of category interactions and their implications within the dataset.

Cross-Chain Analysis

The cross-chain analysis focused on segmenting the data based on distinct blockchain networks, including Ethereum Mainnet, Bitcoin Mainnet, BNB Chain Mainnet, Avalanche C-Chain, and Polygon Mainnet. Each subset of data was analyzed separately to uncover unique patterns and frequent itemsets within each network. For example, the Ethereum Mainnet data, consisting of 63,462 records, revealed frequent itemsets such as "smart_contract," "exchange," and combinations like "cex, exchange," suggesting that smart contracts and centralized exchanges play prominent roles within the Ethereum ecosystem. The Bitcoin Mainnet data, with 18,086 records, highlighted different patterns dominated by categories like "sanctioned" and "cex," reflecting the distinct usage trends and regulatory concerns prevalent in Bitcoin transactions. Smaller chains, such as the BNB Chain, Avalanche C-Chain, and Polygon Mainnet, demonstrated varied itemsets, with liquid staking and NFT categories often appearing, suggesting unique user behavior compared to larger networks.

Following segmentation, the comparison of association rules across chains identified both unique and shared patterns. Ethereum Mainnet generated five unique rules with high confidence, including the frequent co-occurrence of "cex" and "exchange" categories, indicative of robust exchange activity. Bitcoin Mainnet revealed two association rules, with notable links between "ransom" and "scam," suggesting potential security risks. Conversely, no significant rules were generated for BNB Chain, Avalanche C-Chain, and Polygon Mainnet, possibly due to limited data or lower transaction volumes. Across all chains, a total of seven unique rules were identified, with no rules shared by all chains, underscoring the variability in user behavior and interactions across different blockchain networks. The identification of chain-specific rules provided insights into how user activity and category co-occurrences vary, driven by the underlying characteristics of each network.

Source-Based Analysis

The source-based analysis segmented data according to the origin of the data classification—specifically "ground_truth," "external," and "heuristic" sources. Each source was analyzed separately to assess its impact on category associations and frequent itemsets. For instance, the "ground_truth" segment, with 62,389 records, prominently featured categories like "smart_contract" and "liquid_staking." The "external" source segment, comprising 17,801 records, displayed frequent occurrences of "cex" and "wallet," reflecting the diversity of source-based classification criteria. Lastly, the "heuristic" segment revealed frequent itemsets dominated by "exchange" and "cex" categories, highlighting its focus on identifying transactional behaviors often associated with exchanges and centralized entities.

Analyzing the association rules generated across different sources highlighted potential biases introduced by source-specific classification criteria. For example, rules derived from "ground_truth" sources showed a strong emphasis on interactions involving "smart_contract," while "external" sources focused more on categories such as "cex" and "honeypot." Only one unique rule was identified for "heuristic" sources, indicating possible conservatism in rule generation or limited data diversity. No association rules were shared across all sources, demonstrating distinct classification perspectives. To detect potential biases, specific analyses were conducted to track how often certain categories appeared across different sources. For example, the "exchange" category showed no consistent association across sources, emphasizing variations in how each source captures and interprets data. Such findings highlighted the importance of understanding source-driven biases when analyzing and interpreting blockchain data patterns.

Result and Discussion

Frequent Category Patterns

The analysis revealed several frequent itemsets that offer insights into common blockchain address category co-occurrences. Figure 2 lists the top 10 frequent itemsets discovered using the FP-Growth algorithm, with the "smart_contract" category emerging as the most prevalent itemset, appearing in 26.88% of transactions. Other highly frequent categories included "exchange" with a support value of 18.35% and "liquid_staking" at 14.96%, reflecting their prominence within blockchain activities. Combinations such as "cex" and "exchange" also exhibited notable frequency, indicating a strong association between centralized exchanges and other transactional categories. This prevalence highlights the central role of smart contracts and exchanges in shaping blockchain interactions and user behaviors.

To further illustrate these patterns, Figure 2 presents a heatmap visualizing the co-occurrence of the top categories, emphasizing the most significant pairings, such as "cex" and "exchange."

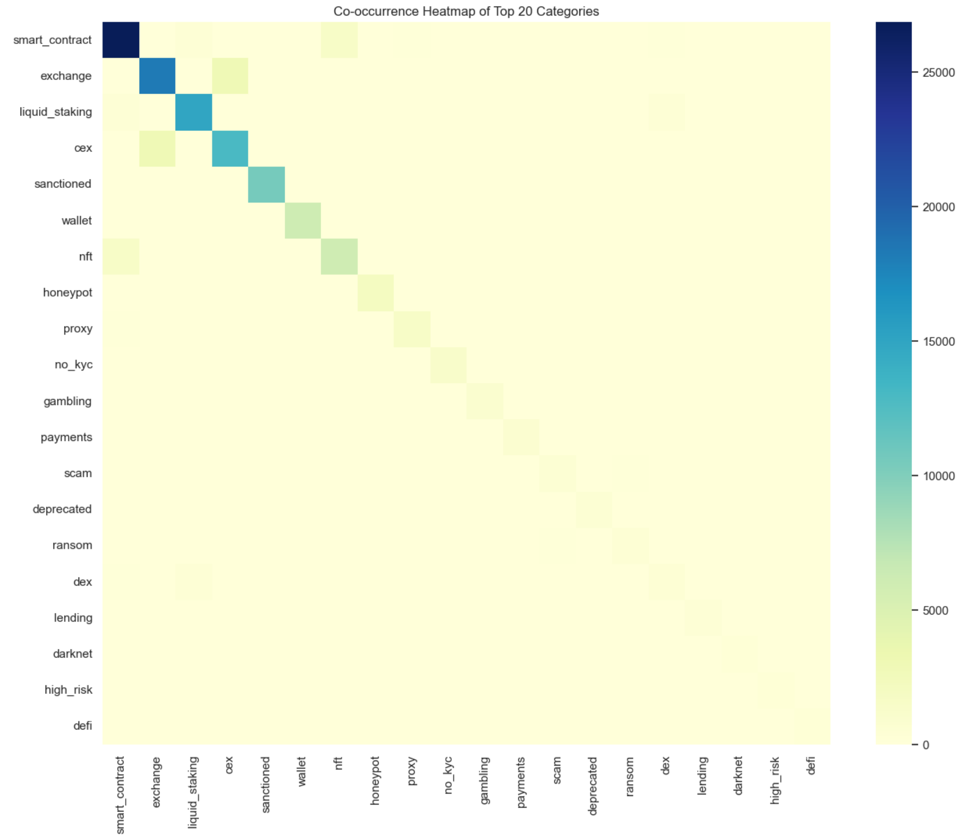

Figure 2 Co-occurrence Heatmap of Top 20 Categories

Figure 3 displays the top 15 most frequent categories among blockchain address classifications in the dataset. The "smart_contract" category appears as the most prominent, with a count exceeding 25,000, indicating that a significant portion of addresses in the dataset are associated with smart contract functions. This is followed by "exchange" and "liquid_staking" categories, each with high counts, suggesting substantial blockchain activity related to currency exchange platforms and liquid staking operations. The "cex" and "sanctioned" categories also feature prominently, highlighting a substantial presence of centralized exchanges and sanctioned addresses. Less frequent categories include "wallet," "nft," and "honeypot," each with moderate representation, likely reflecting diverse uses of blockchain for individual wallets, non-fungible tokens, and certain security traps (honeypots). Lower on the frequency scale are categories such as "proxy," "no_kyc," "gambling," and "payments," which exhibit specific but less prevalent blockchain activities. The least common categories in the top 15, such as "scam," "deprecated," and "ransom," represent more niche or illicit activities within the blockchain space. This distribution highlights the dominance of financial and smart contract-related applications on blockchain platforms, with various other categories representing a spectrum of additional functionalities and risks.

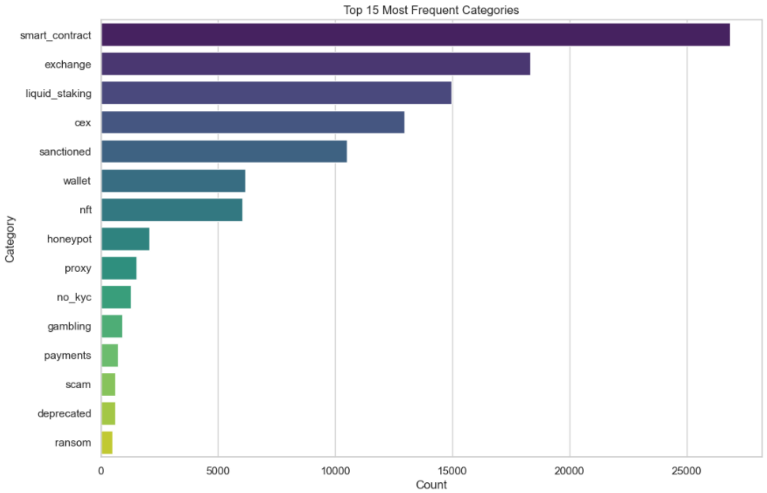

Figure 3 **Top 15 Most Frequent Categories**

Association Rules

The analysis of association rules derived from the frequent itemsets focused on identifying significant patterns based on metrics such as support, confidence, and lift. A total of two association rules met the minimum confidence threshold of 20%. Table 2 displays the top rules, including a notable rule where the antecedent "cex" leads to the consequent "exchange" with a support of 0.03044, a confidence of 23.48%, and a lift value of 1.28. This rule suggests a moderately strong relationship between centralized exchanges and transactional behaviors in the dataset, indicating that addresses categorized under "cex" frequently engage in exchange-related activities.

The second significant rule highlighted the association between "nft" and "smart_contract," reflecting a lower lift value of 0.92 but indicating the frequent use of smart contracts in NFT-related transactions. The limited number of highly significant rules underscores the complexity and specificity of blockchain address interactions. The relatively low number of association rules suggests that while some category patterns are prominent, there is still substantial diversity in how addresses operate across the blockchain networks.

The association between "cex" and "exchange" categories suggests that centralized exchanges play a critical role in facilitating liquidity and asset transfers within blockchain networks. This finding aligns with the well-known dominance of centralized exchanges in crypto market operations. The moderate lift value for this rule implies that while the association is notable, other transactional categories may also interact with exchanges, underscoring the interconnected nature of blockchain ecosystems.

The rule linking "nft" and "smart_contract" indicates the strong reliance of NFT transactions on smart contract functionality. This finding reflects the widespread use of smart contracts to govern the minting, transfer, and ownership of NFTs, providing automation and security in NFT markets. The relatively lower lift for this rule, however, highlights the prevalence of smart contracts beyond NFT transactions, suggesting their utility in a wide range of blockchain applications. Together, these insights reveal critical aspects of blockchain address behavior, emphasizing the interconnected roles of major categories and providing a foundation for further analysis and application development.

Cross-Chain Association Patterns

The cross-chain analysis focused on comparing frequent itemsets and association rules across various blockchain networks to uncover similarities and differences in address behaviors. In Ethereum Mainnet, the most common itemset, "smart_contract," appeared in 42.35% of transactions, demonstrating its dominant role within this chain. Similarly, "exchange" and "cex" were prominent with supports of 28.91% and 12.22%, respectively, often co-occurring, as seen in the "cex, exchange" pair with a support of 4.80%. Conversely, Bitcoin Mainnet displayed a markedly different pattern. "Sanctioned" was the leading category, found in 58.10% of the transactions, followed by "cex" at 28.79%, while categories like "ransom" and "scam" exhibited high co-occurrence, reflecting Bitcoin’s unique risk-associated profile. BNB Chain and Avalanche C-Chain revealed fewer frequent itemsets, with BNB Chain dominated by "liquid_staking" (83.20% support), suggesting concentrated activity in staking operations.

Figures 3 illustrate the observed patterns. Figure 3’s bar charts emphasize the varying prevalence of top categories across chains, while Figure 4’s network graphs highlight unique and common associations within each blockchain. Notably, Ethereum displayed a rich array of interconnected categories, whereas Bitcoin’s graph underscored risk-centric clusters like "ransom" and "scam." This variation in associations can be attributed to the distinct purposes, user bases, and regulatory landscapes governing each blockchain. For instance, Ethereum’s diverse decentralized application (DApp) ecosystem fosters complex associations, while Bitcoin’s focus on financial transfers often leads to specific risk-related interactions.

Source-Based Influence on Associations

To assess the influence of data sources on category associations, the dataset was segmented by ground_truth, external, and heuristic data sources. Ground_truth data exhibited high prevalence of "smart_contract" (42.99%) and "liquid_staking" (23.97%) categories, suggesting a focus on verified and decentralized applications. The most significant association rule derived here linked "dex" and "liquid_staking" with high confidence (85.89%), reflecting the natural pairing of decentralized finance (DeFi) activities. In contrast, external data sources prominently featured "cex" (37.03%) and "wallet" (23.11%), indicating a strong focus on custodial services and general user interactions, while rules centered around "ransom" and "scam" indicated potential threat monitoring in broader datasets. Heuristic data displayed a dominance of "exchange" (92.18%), emphasizing behavior-based inferences.

Figure 5 presents bar charts comparing category frequencies across data sources, showing how certain categories, such as "honeypot" and "scam," fluctuate depending on the source’s scope and methodology. The analysis revealed potential biases introduced by differing data sources. For example, ground_truth data provided a more comprehensive view of verified DApp activity, while heuristic data leaned heavily on exchange patterns due to behavioral clustering. These biases underscore the need to contextualize category prevalence and associations when developing security measures or user behavior models. Identifying these discrepancies is critical for accurate blockchain analytics and for ensuring that insights reflect true network dynamics rather than artifacts of data collection methodologies.

Implications of Findings

The patterns discovered through the FP-Growth association mining algorithm offer valuable practical applications in the realm of blockchain analytics. One key application is the enhancement of classification models. The frequent co-occurrence patterns among blockchain address categories, such as "cex" and "exchange," provide robust features that improve the accuracy of model predictions by capturing complex transactional relationships within the blockchain network. These patterns also play a crucial role in anomaly detection, where deviations from established co-occurrence norms can signal potentially fraudulent or suspicious activities. For instance, the frequent pairing of "ransom" and "scam" in Bitcoin transactions highlights areas that warrant closer scrutiny, aiding in the proactive identification of illicit behaviors. Moreover, security measures can be reinforced by leveraging these insights to develop dynamic monitoring systems that detect unusual category interactions indicative of emerging threats.

From a theoretical perspective, this study contributes to the broader understanding of blockchain address behaviors and ecosystem dynamics. The identification of recurring associations between categories like "smart_contract" and "nft" within Ethereum sheds light on dominant user activities and transactional flows. Such insights deepen our comprehension of how different categories interact, reflect market trends, and evolve in decentralized ecosystems. This enriched understanding allows researchers and practitioners to develop more sophisticated models of blockchain activity, ultimately fostering innovations in decentralized finance (DeFi), decentralized applications (DApps), and blockchain governance.

Limitations

Despite the valuable insights gained, this study faced several limitations, particularly with respect to data constraints. The analysis relied on data from multiple sources, including ground_truth, external, and heuristic data, each of which may introduce biases in category representation. For instance, heuristic data tends to emphasize behavior-based inferences, potentially skewing results towards frequently monitored behaviors, such as exchange activity. Similarly, ground_truth data often centers around verified applications, which may limit the generalizability of findings to broader, less regulated environments. In addition, the uneven representation of some categories across data sources could influence the prevalence and strength of discovered associations, leading to a potential underrepresentation of niche or less-monitored activities.

Methodological limitations inherent to the FP-Growth algorithm and association rule mining must also be acknowledged. While FP-Growth excels at identifying frequent patterns with minimal computational overhead, it can struggle with complex, highly interdependent datasets where nuanced, low-support interactions are significant. The reliance on fixed support and confidence thresholds may lead to the exclusion of potentially meaningful associations that do not meet these criteria. Furthermore, association rules often capture linear relationships, which may oversimplify the multifaceted behaviors present in blockchain transactions. Future research could explore hybrid models that integrate other data mining approaches, such as clustering and sequence analysis, to overcome these limitations and gain a more comprehensive view of blockchain interactions.

Conclusion

This research successfully uncovered key patterns and associations among blockchain address categories using the FP-Growth association mining algorithm. Prominent frequent itemsets, such as "smart_contract" and "exchange," demonstrated significant co-occurrence, indicating strong interactions within decentralized applications and financial services. The cross-chain analysis revealed notable variations across blockchain networks, with Ethereum and Bitcoin showcasing unique category distributions and associations. Ethereum's data was dominated by smart contracts and liquid staking, whereas Bitcoin highlighted categories such as sanctioned addresses and scam activities. Source-based analyses provided further insights, demonstrating how different data sources, including ground_truth, heuristic, and external datasets, influence category associations, sometimes introducing biases that shaped observed co-occurrence patterns.

The findings contribute substantially to the understanding of blockchain address categorization and association mining. This research bridges a critical gap by revealing how co-occurrence patterns differ across chains and data sources, enriching the collective knowledge of blockchain transaction behaviors. From a practical standpoint, the identified patterns offer avenues to improve classification models used in blockchain analytics, refine anomaly detection mechanisms, and bolster security measures against illicit activities. By analyzing category interactions, this study lays the groundwork for more targeted monitoring of blockchain ecosystems, thereby enhancing network security and operational efficiency within decentralized environments.

Future research can extend this study by incorporating temporal data to explore how category associations evolve over time, which could reveal dynamic shifts in blockchain usage and trends. Expanding the analysis to include additional blockchain networks, such as emerging chains with unique structures, may provide broader perspectives on category interactions. Integrating external datasets, such as regulatory databases or market trend data, could further enrich the analysis and provide a more holistic understanding of blockchain activities. Additionally, exploring alternative algorithms, such as hybrid approaches that combine association mining with clustering or graph-based methods, may enhance the detection of more complex and nuanced patterns within large-scale blockchain data.

Understanding category co-occurrences within blockchain data is crucial for developing robust insights into decentralized network behaviors and applications. This research highlights the potential of association mining techniques to uncover meaningful patterns that can inform decision-making, security protocols, and analytics frameworks. Continued exploration of these patterns is essential as blockchain ecosystems grow in complexity, providing a foundation for innovations that improve transparency, security, and trust in decentralized systems.

Declarations

Author Contributions

Conceptualization: L.L.; Methodology: L.L.; Software: L.L.; Validation: L.L.; Formal Analysis: L.L.; Investigation: L.L.; Resources: L.L.; Data Curation: L.L.; Writing Original Draft Preparation: L.L.; Writing Review and Editing: L.L.;

Visualization: L.L.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R. Beck, C. Müller-Bloch, and J. L. King, "Governance in the Blockchain Economy: A Framework and Research Agenda," J. Assoc. Inf. Syst., pp. 1020–1034, 2018, doi: 10.17705/1jais.00518.

[2] C. C. Agbo, Q. H. Mahmoud, and J. Eklund, "Blockchain Technology in Healthcare: A Systematic Review," Healthcare, vol. 7, no. 2, p. 56, 2019, doi: 10.3390/healthcare7020056.

[3] L. Jin, H. Liang, and C. Yang, "Accurate Underwater ATR in Forward-Looking Sonar Imagery Using Deep Convolutional Neural Networks," Ieee Access, 2019, doi: 10.1109/access.2019.2939005.

[4] T. Dirsehan, "Analysis of a Blockchain-Based Website Using the Technology Acceptance Model: The Case of Save Ideas," Int. J. Dipl. Econ., vol. 6, no. 1, p. 17, 2020, doi: 10.1504/ijdipe.2020.10031850.

[5] J. Yli-Huumo, D. Ko, S. Choi, S. Park, and K. Smolander, "Where Is Current Research on Blockchain Technology?—A Systematic Review," Plos One, vol. 11, no. 10, p. e0163477, 2016, doi: 10.1371/journal.pone.0163477.

[6] P. Song and Y. Liu, "An XGBoost Algorithm for Predicting Purchasing Behaviour on E-Commerce Platforms," Teh. Vjesn. - Tech. Gaz., vol. 27, no. 5, 2020, doi: 10.17559/tv-20200808113807.

[7] T. Z. Yi, N. A. M. Rom, N. Hassan, M. S. Samsurijan, and A. Ebekozien, "The Adoption of Robo-Advisory Among Millennials in the 21st Century: Trust, Usability and Knowledge Perception," Sustainability, vol. 15, no. 7, p. 6016, 2023, doi: 10.3390/su15076016.

[8] A. Singla, "Exploring the Potential of Dogecoin Promoted by Elon Musk," SJMBT, vol. 2, no. 1, pp. 35–43, 2024, doi: 10.36676/sjmbt.v2.i1.06.

[9] M. Jourdan, S. Blandin, L. Wynter, and P. Deshpande, "Characterizing Entities in the Bitcoin Blockchain," 2018, doi: 10.1109/icdmw.2018.00016.

[10] B. Berlilana, A. M. Wahid, D. Fortuna, A. N. A. Saputra, and G. Bagaskoro, "Exploring the Impact of Discount Strategies on Consumer Ratings: An Analytical Study of Amazon Product Reviews," J. Appl. Data Sci., vol. 5, no. 1, Art. no. 1, Jan. 2024, doi: 10.47738/jads.v5i1.163.

[11] Henderi and Q. Siddique, “Comparative Analysis of Sentiment Classification Techniques on Flipkart Product Reviews: A Study Using Logistic Regression, SVC, Random Forest, and Gradient Boosting,” J. Digit. Mark. Digit. Curr., vol. 1, no. 1, Art. no. 1, May 2024, doi: 10.47738/jdmdc.v1i1.4.

[12] T. Wahyuningsih and S. C. Chen, “Analyzing Sentiment Trends and Patterns in Bitcoin-Related Tweets Using TF-IDF Vectorization and K-Means Clustering,” J. Curr. Res. Blockchain, vol. 1, no. 1, Art. no. 1, Jun. 2024, doi: 10.47738/jcrb.v1i1.11.

[13] Hery and A. E. Widjaja, “Predictive Modeling of Blockchain Stability Using Machine Learning to Enhance Network Resilience,” J. Curr. Res. Blockchain, vol. 1, no. 2, Art. no. 2, Sep. 2024, doi: 10.47738/jcrb.v1i2.15.

[14] S. Yadav and A. R. Hananto, “Comprehensive Analysis of Twitter Conversations Provides Insights into Dynamic Metaverse Discourse Trends,” Int. J. Res. Metaverese, vol. 1, no. 1, Art. no. 1, Jun. 2024, doi: 10.47738/ijrm.v1i1.2.

[15] T. Wahyuningsih and S. C. Chen, “Determinants of Virtual Property Prices in Decentraland an Empirical Analysis of Market Dynamics and Cryptocurrency Influence,” Int. J. Res. Metaverese, vol. 1, no. 2, Art. no. 2, Sep. 2024, doi: 10.47738/ijrm.v1i2.12.

[16] F. Zola, M. Eguimendia, J. L. Bruse, and R. Orduna-Urrutia, “Cascading Machine Learning to Attack Bitcoin Anonymity,” 2019, doi: 10.1109/blockchain.2019.00011.

[17] D. Elangovan et al., “The Use of Blockchain Technology in the Health Care Sector: Systematic Review,” Jmir Med. Inform., vol. 10, no. 1, p. e17278, 2022, doi: 10.2196/17278.

[18] K. Martin, M. Rahouti, M. Ayyash, and I. Alsmadi, “Anomaly Detection in Blockchain Using Network Representation and Machine Learning,” Secur. Priv., vol. 5, no. 2, 2021, doi: 10.1002/spy2.192.

[19] Hemavati, V. S. Devi, and R. Aparna, “Multi Layered Stacked Ensemble Method With Feature Reduction Technique for Multi-Label Classification,” J. Phys. Conf. Ser., vol. 2161, no. 1, p. 012074, 2022, doi: 10.1088/1742-6596/2161/1/012074.

[20] R. Venkatesan, M. J. Er, S. Wu, and M. Pratama, “A Novel Online Real-Time Classifier for Multi-Label Data Streams,” 2016, doi: 10.1109/ijcnn.2016.7727422.

[21] M. Tang, W. Yang, Y. Li, and Q. Zeng, “Research on Multi-Label Long Text Classification Algorithm Based on Transformer-Lda,” 2023, doi: 10.1117/12.2667798.

[22] F. Charte, A. J. Rivera, M. J. d. Jesus, and F. Herrera, “Addressing Imbalance in Multilabel Classification: Measures and Random Resampling Algorithms,” Neurocomputing, vol. 163, pp. 3–16, 2015, doi: 10.1016/j.neucom.2014.08.091.

[23] D. V. Le, “AMR:Autonomous Coin Mixer With Privacy Preserving Reward Distribution,” 2020, doi: 10.48550/arxiv.2010.01056.

[24] S. Sahoo, A. M. Fajge, R. Halder, and A. Cortesi, “A Hierarchical and Abstraction-Based Blockchain Model,” Appl. Sci., vol. 9, no. 11, p. 2343, 2019, doi: 10.3390/app9112343.

[25] R. Joshi, R. Gupte, and P. Saravanan, “A Random Forest Approach for Predicting Online Buying Behavior of Indian Customers,” Theor. Econ. Lett., 2018, doi: 10.4236/tel.2018.83032.

[26] Y. Wan, Y. Liang, and L. Ding, “Mining Multilevel Association Rules With Dynamic Concept Hierarchy,” 2008, doi: 10.1109/icmlc.2008.4620419.

[27] J. Kaur and N. Madan, “Association Rule Mining: A Survey,” Int. J. Hybrid Inf. Technol., vol. 8, no. 7, pp. 239–242, 2015, doi: 10.14257/ijhit.2015.8.7.22.

[28] S. Pramod and O. P. Vyas, “Performance Evaluation of Some Online Association Rule Mining Algorithms for Sorted and Unsorted Data Sets,” Int. J. Comput. Appl., vol. 2, no. 6, pp. 40–45, 2010, doi: 10.5120/670-941.

[29] Q. Wei, L. Ma, S. Ding, and H. Mi, “An Association Rule Algorithm Generated by Minimal Head-Item Set Based on Set-Enumeration Tree,” 2009, doi: 10.1109/iciecs.2009.5363139.

[30] B. Ramasubbareddy, A. Govardhan, and A. Ramamohanreddy, “Mining Positive

and Negative Association Rules," 2010, doi: 10.1109/iccse.2010.5593755.

[31] D. Liu, Y. Chen, Y. Fan, and G. Shen, "The Application of Association Rule Mining in Power System Restoration," 2005, doi: 10.1109/ipec.2005.207061.

[32] S. Mabu, T. Higuchi, and T. Kuremoto, "SemiSupervised Learning for Class Association Rule Mining Using Genetic Network Programming," Ieej Trans. Electr. Electron. Eng., vol. 15, no. 5, pp. 733–740, 2020, doi: 10.1002/tee.23109.

[33] S. Suresh, S. Uvaraj, and N. Raja, "To Allot Secrecy-Safe Association Rules Mining Schema Using FP Tree," Softw. Eng., vol. 1, no. 1, p. 1, 2013, doi: 10.11648/j.se.20130101.11.

[34] X. Jiao, L. Xu, and Q. Lin, "Association Rules Mining Algorithm Based on Rough Set," 2012, doi: 10.1109/itime.2012.6291318.

[35] H. Kong, "Itemsets of Interest for Negative Association Rules," 2018, doi: 10.48550/arxiv.1806.07084.

[36] P. Agarwal, M. L. Yadav, and N. Anand, "Study on Apriori Algorithm and Its Application in Grocery Store," Int. J. Comput. Appl., vol. 74, no. 14, pp. 1–8, 2013, doi: 10.5120/12950-9882.

[37] V. Vaithiyanathan, K. Rajeswari, and R. Phalnikar, "Improved Apriori Algorithm Based on Selection Criterion," 2012, doi: 10.1109/iccic.2012.6510229.

[38] D. Dwiputra, "Evaluating the Performance of Association Rules in Apriori and FP-Growth Algorithms: Market Basket Analysis to Discover Rules of Item Combinations," J. World Sci., vol. 2, no. 8, pp. 1229–1248, 2023, doi: 10.58344/jws.v2i8.403.

[39] S. Han, H. Kim, and Y.-S. Lee, "Double Random Forest," Mach. Learn., vol. 109, no. 8, pp. 1569–1586, 2020, doi: 10.1007/s10994-020-05889-1.

[40] L. Xiang, "An Improved Frequent Pattern-Growth Algorithm Based on Decomposition of the Transaction Database," 2015, doi: 10.22323/1.259.0023.

[41] S. Shan, X. Wang, and M. Sui, "Mining Association Rules: A Continuous Incremental Updating Technique," 2010, doi: 10.1109/wism.2010.39.

[42] K. Wang, T. Liu, J. Han, and J. Liu, "Top Down FP-Growth for Association Rule Mining," pp. 334–340, 2002, doi: 10.1007/3-540-47887-6_34.

[43] P. A. I. Lestari, "Application of the Association Rule Method Based on Book Borrowing Patterns in Bojonegoro Regional Libraries," J. Comput. Netw. Archit. High Perform. Comput., vol. 5, no. 2, pp. 751–759, 2023, doi: 10.47709/cnahpc.v5i2.2893.

[44] A. P. U. Siahaan, A. Ikhwan, and S. Aryza, "A Novelty of Data Mining for Promoting Education Based on FP-Growth Algorithm," 2018, doi: 10.31227/osf.io/jpsfa.

[45] M. Yin, W. Wang, Y. Liu, and D. Jiang, "An Improvement of FP-Growth Association Rule Mining Algorithm Based on Adjacency Table," Matec Web Conf., vol. 189, p. 10012, 2018, doi: 10.1051/matecconf/201818910012.

[46] Z. Abdullah, T. Herawan, N. Ahmad, and M. M. Deris, "A Scalable Algorithm for Constructing Frequent Pattern Tree," Int. J. Intell. Inf. Technol., vol. 10, no. 1, pp. 42–56, 2014, doi: 10.4018/ijiit.2014010103.

[47] S. K. Tanbeer, C. F. Ahmed, B. Jeong, and Y.-K. Lee, "Efficient Single-Pass Frequent Pattern Mining Using a Prefix-Tree," Inf. Sci., vol. 179, no. 5, pp. 559–583, 2009, doi: 10.1016/j.ins.2008.10.027.

[48] A. I. Idris et al., "Comparison of Apriori, Apriori-Tid and FP-Growth Algorithms in Market Basket Analysis at Grocery Stores," Ijics Int. J. Inform. Comput. Sci., vol. 6, no. 2, p. 107, 2022, doi: 10.30865/ijics.v6i2.4535.

[49] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns Without Candidate Generation," Acm Sigmod Rec., vol. 29, no. 2, pp. 1–12, 2000, doi: 10.1145/335191.335372.