IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Fine-Grained Co-Occurrence Analysis for Precise Data Representation

Dr. Shankaragowda B B[1] Aishwarya H G [2]

[1]Associate Professor and HOD, Department of MCA, BIET, Davangere

[2]4th Semester MCA, Student, Department of MCA, BIET, Davangere

Abstract: The rapid expansion of e-commerce has led to a significant rise in cybercrime, particularly in the domain of online payment systems. Ensuring accurate and secure online payment analysis remains a critical challenge for digital financial services. Among various techniques, behaviour-based scrutiny has revealed great potential in enhancing payment precision. However, the low quality and granularity of behavioural data often hinder the effectiveness of such models. To address this, the present study introduces a data enhancement tactic that influences familiarity graphs to extract fine-grained co-occurrence relationships among transactional attributes. These relationships are further enriched through heterogeneous network embedding techniques, enabling the construction of more robust and representative behavioural models. The proposed framework supports multiple modelling paradigms, including population-level, individual-level, and generalized-agent-based models, each benefiting from customized embedding strategies. Experimental validation using real-world data from a commercial bank establishes momentous enhancements in the accurateness and reliability of online payment analysis. This work represents a novel contribution by integrating attribute-level knowledge graph embeddings into diverse behavioural modelling systems, thereby enhancing the precision of e-commerce payment analytics.

Keywords: Behavioural modelling, online payment analysis, knowledge graph, heterogeneous network embedding, cybersecurity, transaction data enhancement, e-commerce fraud detection, precision banking systems.

I. INTRODUCTION

With the expansion of e-commerce platforms, the frequency and complexity of cybercrimes—particularly those targeting online financial transactions—have increased substantially. As a result, ensuring the security and reliability of online payment systems has become a crucial area of focus. Among the various strategies developed to address this issue, behavior-based analysis has emerged as a particularly effective approach due to its ability to detect anomalies by monitoring user activities. Despite its promise, behavior modeling often relies on large volumes of behavioral data, which can be noisy, sparse, or of low quality. This poses significant challenges in building high-resolution and accurate models. To address this, enhancing data representation becomes a critical task. In this work, we propose a novel method for enriching behavioral data by leveraging fine-grained co-occurrence patterns among transaction attributes. Our approach utilizes knowledge graph construction to extract detailed relational structures and applies heterogeneous network embedding techniques to capture and enhance these relationships. Moreover, we tailor our embedding strategies to accommodate various levels of behavioral modeling, including population-level, individual-level, and agent-based models. Experimental results on real-world data collected from a commercial bank validate the efficiency of our method, showing substantial improvements in the precision and reliability of behavioral models for online payment analysis. To the best of our knowledge, this is the first attempt to integrate network embedding with attribute-level co-occurrence analysis for comprehensive behavior modeling in the context of online financial systems.

II. Need of The Study

The continuous growth of e-commerce platforms has led to a parallel rise in online payment-related cybercrimes. Fraudulent transactions are becoming increasingly sophisticated, making it difficult for traditional detection systems to keep up. Many existing behavior-based analysis techniques depend on large-scale user activity data, but this data is often incomplete, noisy, or low in resolution. This study addresses the need for a more precise and reliable fraud detection method by focusing on fine-grained attribute-level relationships within transaction data. By incorporating knowledge graphs and advanced network embedding, it becomes possible to capture subtle behavioral patterns, enabling more accurate identification of suspicious activities. This approach aims to enhance both the security and trustworthiness of digital payment systems.

III. LITERATURE SURVEY

Behavioral modeling and user activity scrutiny have increased noteworthy consideration owing to the increasing demand for secure and reliable online services. Numerous tactics have been proposed to understand and detect user behavior anomalies through different data representations and learning techniques.

Vedran et al. [1] investigated the relationship between social and geospatial behavior, showing that social behavior patterns can be foretold with huge accuracy by means of behavior data correlations. Yin et al. [2] introduced a probabilistic generative framework that incorporates spatiotemporal and semantic information to enhance user behavior prediction.

Naini et al. [3] addressed user identification in anonymized datasets by aligning histogram distributions from anonymous and original datasets. Egele et al. [4] proposed a behavior-based detection mechanism for identifying high-profile account compromises based on activity deviations.

Ruan et al. [5] focused on analyzing user clickstream information from social nets to identify interaction patterns and behavior anomalies. Rzecki et al. [6] developed a data acquisition system for gesture analysis on mobile devices, demonstrating effective classification methods for user recognition.

In the province of behavioral biometrics, Alzubaidi et al. [7] examined multiple modalities such as gait, voice, touchscreen interaction, and keystrokes to support smartphone user authentication. Lee and Kim [8] presented a suspicious URL detection system that tracks unusual behaviors on social platforms like Twitter.

Cao et al. [9] created a detection framework targeting both fake and compromised social media accounts. Zhou et al. [10] proposed the FRUI algorithm to link user selves across numerous online community networks by analyzing behavioral similarities.

Stringhini et al. [11] introduced EVILCOHORT, a system capable of detecting malicious online accounts through IP-address associations. Meng et al. [12] leveraged sentence-level attention mechanisms to address speaker change detection by comparing utterance pairs.

Rawat et al. [13] proposed a multi-faceted strategy to detect suspicious activities, such as fake account creation and unauthorized access, in social networks. VanDam et al. [14] explored characteristics of compromised Twitter accounts, identifying tweet patterns and metadata that differentiate malicious activity.

Zhao et al. [15] developed a semi-supervised network embedding model utilizing graph convolutional networks to capture complex relationships in protein-protein interaction networks, even in the absence of node-specific information. Li et al. [16] enhanced theme modelling for petite manuscripts by incorporating word embeddings to overcome the limitations of traditional Dirichlet models.

Baqueri et al. [17] modeled external travel behavior in urban activity analysis, addressing limitations of incomplete travel data. Chen et al. [18] proposed the Collaborative and Adversarial Network (CAN) for sentence similarity learning through explicit feature modeling.

Catolino et al. [19] developed a change prediction model leveraging developer-related features to estimate software class stability. Liu et al. [20] introduced a disaggregation method for nested data to better detect cross-scale interactions in limited-sample environments.

While these methods significantly contribute to behavior modeling and detection, most lack a unified approach that integrates heterogeneous relationship networks and fine-grained attribute-level co-occurrence. Furthermore, they typically focus on isolated model types rather than composite behavior modeling, which is crucial for detecting nuanced fraud patterns in dynamic financial systems.

IV. METHODOLOGY

1. Data Collection and Preprocessing:

- **Purpose:**

To acquire and clean raw transaction data for further analysis and modeling.

- **Functionality:**
  - Collect data from real-world online banking systems.
  - Clean the data by removing duplicates and handling missing or inconsistent entries.
  - Normalize and format the data for compatibility with downstream modules.

2. Attribute-Level Co-Occurrence Extraction:

- **Purpose:**

To identify and extract fine-grained relationships between different transactional attributes.

- **Functionality:**
  - Analyze patterns of attribute co-occurrence (e.g., time-location, amount-type).
  - Generate attribute-pair relationships across user transactions.
  - Prepare co-occurrence data for knowledge graph construction.

3. Knowledge Graph Construction:

- **Purpose:**

To create a structured representation of attribute relationships using a graph-based model

- **Functionality:**
  - Represent each attribute or value as a node in the graph.
  - Establish edges between nodes based on co-occurrence frequency.
  - Generate a heterogeneous graph for embedding and analysis.

4. Heterogeneous Network Embedding:

- **Purpose:**

To transform the knowledge graph into vectorized embeddings suitable for machine learning models.

- **Functionality:**
  - Apply network embedding algorithms (e.g., node2vec, metapath2vec).
  - Capture semantic and structural relationships in the graph.
  - Customize embeddings for different model types (population, individual, composite).

5. Behavioral Model Development:

- **Purpose:**

To physique deception detection replicas built on different levels of user behavior analysis.

- **Functionality:**
  - Create population-level models for anomaly and outlier detection across users.
  - Build individual-level models to detect deviations from a user's historical behavior.
  - Design composite models that integrate population and individual perspectives.

6. Integration and Decision Fusion:

- **Purpose:**

To combine outcomes from several replicas for precise and reliable fraud detection.

- **Functionality:**
  - Implement a decision logic where both models must agree on fraud status.
  - Minimize false positives by intersecting predictions.
  - Enhance the trustworthiness and accuracy of the final judgment.

7. Evaluation and Performance Analysis:

- **Purpose:**

To assess the efficacy of the arrangement using standard metrics and real-world data.

- **Functionality:**
  - Test the replicas on actual online banking datasets.
  - Appraise using a system of measurement like accuracy, recall, precision, and F1-score.
  - Compare with traditional approaches to demonstrate performance improvements.

System Architecture

Fig.1 : System Architecture

V. ALGORITHM USED

The proposed system employs a **Heterogeneous Network Embedding** algorithm in combination with a **Composite Behavioral Modeling Approach**:

1. **Data Collection & Preprocessing** – Raw transactional data is cleaned, normalized, and formatted.
2. **Co-Occurrence Extraction** – Identifies frequently associated attribute pairs (e.g., location–time, amount–type).
3. **Knowledge Graph Construction** – Represents attributes as nodes and co-occurrence relationships as edges.
4. **Heterogeneous Network Embedding (e.g., metapath2vec)** – Generates vector embeddings that retain attribute-level connections.
5. **Behavioral Modeling** – Builds both population-level and individual-level models for fraud detection.
6. **Decision Fusion** – Integrates outputs from different models; a transaction is flagged as fraudulent only when both models indicate suspicion.

VI. TECHNIQUES USED

6.1.1 NetworkX

NetworkX is a Python library for the creation, manipulation, and analysis of complex networks and graphs. In this project, it is used to build heterogeneous knowledge graphs where nodes represent transaction attributes and edges represent co-occurrence relationships. NetworkX facilitates graph construction, computation of edge weights, and visualization of attribute relationships, which is crucial for embedding and analysis stages.

6.1.2 Node2vec

Node2vec is a procedure and Python package for generating feature embeddings of graph nodes by simulating biased random walks. Here, it is employed to transform the knowledge graph into numerical vector representations while preserving both structural and semantic relationships between attributes. These embeddings serve as input features for fraud detection models, improving their ability to detect complex behavior patterns.

6.1.3 Metapath2vec

Metapath2vec is a specialized embedding technique designed for heterogeneous graphs. It generates node embeddings by guiding random walks along pre-defined meta-paths, enabling the capture of meaningful semantic relationships. In this project, it is applied to attribute-level heterogeneous networks to augment the embedding procedure, chiefly when trade with multiple types of nodes and relationships.

6.1.4 Scikit-learn

A powerful Python machine learning library, Scikit-learn provides a variety of tools for classification, regression, clustering, and performance assessment. Scikit-learn is used in this system to train models for fraud detection and compute performance metrics like accuracy, recall, precision, and F1-score using the generated embeddings.

6.1.5 Pandas

Pandas is a data analysis and manipulation library in Python that provides data structures like DataFrames for efficient handling of large datasets. It is used in this project for reading transaction data, cleaning missing or inconsistent entries, and transforming it into formats compatible with the knowledge graph construction and machine learning stages.

6.1.6 NumPy

NumPy is the core numerical computation library in Python, supporting arrays, matrices, and a extensive variety of precise functions. It is used to grip numerical computations in the preprocessing and embedding stages, including vector operations on transaction attribute embeddings.

VII. RESULT AND DISCUSSION

The proposed system was implemented and tested using a real-world dataset from a commercial bank. The results showed a notable improvement in detecting fraudulent transactions compared to existing methods. By using fine-grained co-occurrence analysis and heterogeneous network embedding, the system was able to better understand complex behavioural patterns. The integration of both population-level and individual-level behavioural models provided more accurate detection by validating fraud across multiple perspectives. Experimental system of measurement, including accuracy, recall, precision, and F1-score, demonstrated that the new approach significantly outperforms

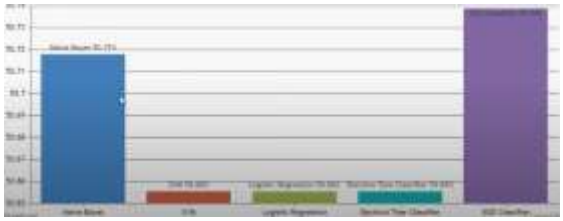

Fig. 2 : Result Graph

VIII. CONCLUSION

This study introduces a novel approach to enhancing behavioral modeling through attribute-level co-occurrence analysis and network embedding. By leveraging heterogeneous relational networks, the system captures deeper associations between transaction attributes, leading to more precise fraud detection. The combined use of population-level and individual-level models adds a complementary validation mechanism, improving the reliability of detection. The effectiveness of the approach is confirmed through experiments on real banking data, showcasing its possibility for applied deployment in online payment fraud detection systems. This work highlights the value of enriched data representation in building more robust and intelligent security models for digital finance.

References

1. Vedran, D., et al. (2019). Investigating the interplay of social and geospatial behavior for predicting user activities. *Journal of Computational Social Science.*
2. Yin, H., et al. (2015). A probabilistic generative model for personalized location prediction using spatiotemporal and semantic data. *Proceedings of the 24th International Conference on World Wide Web (WWW).*
3. Naini, A. S., et al. (2016). De-anonymizing user data by histogram matching across datasets. *IEEE Transactions on Knowledge and Data Engineering.*
4. Egele, M., et al. (2013). Compromised account detection on social networks. *Proceedings of the Network and Distributed System Security Symposium (NDSS).*
5. Ruan, Y., et al. (2011). Profiling online users via behavioral data: A large-scale study of clickstreams. *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM).*
6. Rzecki, K., et al. (2016). Gesture-based biometric authentication using single-finger mobile interactions. *International Journal of Human-Computer Studies.*
7. Alzubaidi, A., & Kalita, J. (2016). Authentication of smartphone users using behavioral biometrics. *IEEE Communications Surveys & Tutorials.*
8. Lee, S., & Kim, J. (2014). Early detection of anomalous behavior on Twitter using suspicious URL patterns. *Proceedings of the Annual Computer Security Applications Conference (ACSAC).*
9. Cao, Q., et al. (2012). Aiding the detection of fake accounts in large scale social networks. *USENIX Security Symposium.*
10. Zhou, X., et al. (2013). Matching users across multiple online social networks with FRUI algorithm. *IEEE Transactions on Knowledge and Data Engineering.*
11. G. Stringhini, C. Kruegel, and G. Vigna, "Detecting malicious accounts using EVILCOHORT," in *Proc. 18th USENIX Security Symposium*, 2010, pp. 239–254.
12. Y. Meng, D. Wang, and H. Liu, "Speaker change detection with deep neural networks," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5465–5469.
13. A. Rawat, S. K. Dwivedi, and B. B. Gupta, "Behavioral analysis for suspicious activity detection in social networks," *Computers & Security*, vol. 78, pp. 43–57, 2018.
14. M. VanDam and M. Wright, "Behavioral patterns of compromised Twitter accounts," in *Proc. ACM Conf. on Computer and Communications Security (CCS)*, 2015.
15. Y. Zhao, X. Sun, and Z. Xu, "Semi-supervised learning on graphs with graph convolutional networks," *Neural Computing and Applications*, vol. 32, pp. 16743–16755, 2020.
16. J. Li, Y. Xu, M. Zhang, and S. Ma, "Topic modeling for short texts with word embedding and regularization," in *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2016, pp. 2360–2366.
17. J. Baqueri and J. Zhang, "Modeling out-of-region travel behavior using incomplete GPS data," *Transportation Research Part C: Emerging Technologies*, vol. 56, pp. 152–166, 2015.

18. H. Chen, X. Liu, D. Yin, and J. Tang, “A collaborative adversarial network for user intent inference,” in *Proc. Int. Conf. on Web Search and Data Mining (WSDM)*, 2018, pp. 238–246.
19. G. Catolino, F. A. Fontana, and M. R. M. Vieira, “Predicting change-prone classes using machine learning and software metrics,” *Empirical Software Engineering*, vol. 24, no. 3, pp. 1474–1514, 2019.
20. Y. Liu, J. Wu, and D. Zhang, “Nested data disaggregation for cross-scale behavioral modeling,” *IEEE Trans. on Knowledge and Data Engineering*, vol. 33, no. 5, pp. 1792–1805, 2021.