Graph Convolutional Representation Learning for Sybil Detection in Online Social Networks

Heta Dasondi
Research Scholar, FCA
Ganpat University,
Gujarat, India

Meghna B. Patel, PhD
Associate Professor, A. M. Patel Institute of Computer Studies,
Ganpat University,
Gujarat, India

Satyen M. Parikh
Professor, A. M. Patel Institute of Computer Studies, Ganpat University,
Gujarat, India

ABSTRACT
Online Social Networks (OSNs) are increasingly vulnerable to Sybil attacks, wherein adversaries create numerous fake identities to distort information, manipulate influence, and compromise user trust. Existing detection methods, while effective in constrained settings, often struggle to scale and generalize across the complex and dynamic topologies of modern social graphs. In this paper, it propose SD-GCN, a scalable Sybil detection framework based on Graph Convolutional Networks. The proposed method leverages a GCN architecture that integrates both local and global topological features through multi-hop message passing, enabling the extraction of expressive node embeddings that capture structural and behavioral distinctions between benign and Sybil nodes. To enhance performance, the model undergoes comprehensive hyper-parameter optimization, balancing detection accuracy with computational efficiency. The proposed approach is evaluated on a real-world Facebook follower-followee graph and achieves a high classification performance, significantly outperforming established baselines such as SybilGAT and SybilWalk. Notably, the model achieves an Area Under the Curve (AUC) of 96%, demonstrating its robustness and generalization capability for large-scale OSN environments.

Keywords
Graph Neural Network, GCN, Online Social Network

1. INTRODUCTION
The exponential growth of Online Social Networks [1] has fundamentally transformed how individuals communicate, share information, and build communities in the digital age. Platforms such as Facebook, Twitter, Instagram, and LinkedIn have become integral to modern social interaction [2], hosting billions of users who generate vast amounts of content daily. However, this unprecedented connectivity and openness have also created new vulnerabilities that malicious actors actively exploit to undermine the integrity and trustworthiness of these platforms.

Among the most pervasive and dangerous threats facing OSNs today are Sybil attacks [3], sophisticated deception schemes wherein adversaries subvert network services by creating large numbers of pseudonymous identities to gain disproportionate influence. Named after the protagonist of a 1973 case study of dissociative identity disorder, the term Sybil attack was formally introduced by [4] at Microsoft Research to describe this fundamental vulnerability in distributed systems. The core principle underlying these attacks involves a single malicious entity operating multiple fake accounts simultaneously, thereby manipulating reputation systems, distorting information flows, and compromising the democratic nature of social interactions.

The sophistication of modern Sybil attacks has evolved considerably beyond simple fake account creation. Contemporary attack strategies include elite Sybil attacks, where adversaries recruit highly-rated legitimate users who normally post genuine content but are incentivized to participate in coordinated manipulation campaigns. These elite attackers craft convincing reviews and content that closely mimics authentic user behavior, making detection extraordinarily challenging through traditional analytical methods [5]. The impact of Sybil attacks extends far beyond individual user deception. In social media networks, successful attacks can disseminate misinformation, create artificial consensus, and manipulate public opinion on critical social and political issues.

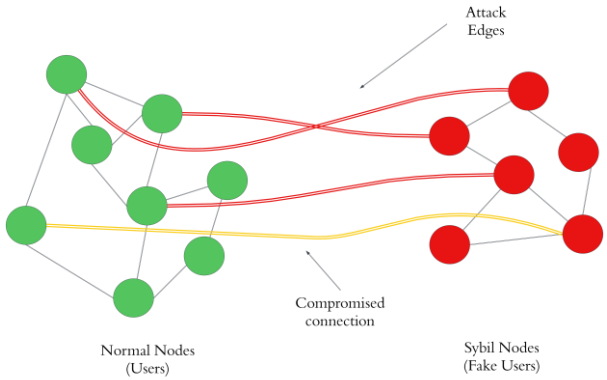

Fig 1. Basics of Sybil Attack Edges

The *Fig 1* illustrates the structure of a Sybil attack within an online social network, represented as a graph. On the left side are the normal nodes (legitimate users) connected to each other through typical social interactions. On the right side are the Sybil nodes, which represent fake or malicious accounts created by an attacker. These Sybil nodes attempt to infiltrate the network by forming connections—known as attack edges—with legitimate users. These connections are marked in red, showing the direct links used by the attacker to gain influence or trust within the network. Additionally, there is a yellow edge indicating a compromised connection, which suggests a legitimate user may have unknowingly accepted a connection from a Sybil node. This visualization highlights the threat posed by Sybil attacks and the challenge of distinguishing between genuine and malicious users in a social graph.

Rest of the paper is structured as follows. Section-2 reviews the extant literature. Section-3 describes the preliminaries. Section-4 explains the research methodology. Secttion-5 discusses the experiments and results. Section-6 summarises the paper.

2. LITERATURE REVIEW

The paper [6] shows that around 55% percent users between the age of 18-30 have shared their passwords online with family or friends which is not good thing. While the paper [7] discusses various levels of complex networks e.g. first type of network is multiple networks, the second type of network is temporal network and last is modal networks. The paper [8] observed various ways people use social media, how they spend their time on the social media, e.g. someone spend more time posting something, someone spend more time looking at others post etc. the research study conducted on various descriptions of user behaviour patterns.

Current Sybil detection methodologies, while demonstrating effectiveness in controlled environments, face significant scalability and generalization challenges when deployed across the complex, dynamic topologies characteristic of large-scale social networks. Traditional approaches typically fall into two primary categories: graph-connectivity-based methods and machine learning classifiers built on engineered features. Graph-connectivity approaches, such as those implemented in systems like SybilBelief [9], [10] rely on the assumption that social connections between legitimate users and Sybil accounts are relatively sparse. These methods leverage semi-supervised learning frameworks to propagate trust information from known benign nodes throughout the network topology. The paper [11] implements framework that is replied on GCN and it detects fake accounts in online social networks. This paper [12] uses hybrid graph based techniques to detect Sybil with aggregation of user behaviors, the SybilHunter is more accurate than SybilRank [13] and SybilWalk [14]. The paper [15] proposed a SATAR, it's a self-supervised social media bot detection system. This paper [16] combines Loopy Belief Propagation with an adaptive homophily estimator that dynamically predicts the assortativity for each node, enabling more accurate classification by accounting for the directional nature of social connections.

Machine learning-based detection systems typically extract handcrafted features related to user behavior patterns, account creation timestamps, posting frequencies, and social interaction characteristics. While these approaches can achieve reasonable performance on specific datasets, they suffer from several critical limitations. First, they require extensive domain expertise to engineer appropriate features for each platform and attack variant. Second, they often fail to generalize across different OSN platforms due to varying user behavior norms and platform-specific characteristics. Third, they struggle to adapt to evolving attack strategies as sophisticated adversaries continuously modify their techniques to evade detection.

Recent advances in deep learning [17], [18], particularly in the domain of Graph Neural Networks [19], offer promising new directions for addressing the limitations of traditional Sybil detection approaches. Graph Convolutional Networks [20] represent a significant breakthrough in learning from graph-structured data, enabling the automatic extraction of rich node representations that capture both local neighborhood characteristics and global network properties through iterative message passing mechanisms.

3. PRELIMINARIES

An **Online Social Network (OSN)** can be modeled as an undirected or directed graph $G = (V, E)$ where V is set of nodes (vertices), where each node $v \in V$ represent a user. $E \subseteq V \times V$ is the set of edges, where each edge $(u, v) \in E$ represents a social relationship between users u and v.

3.1 Sybil attacks and Sybil nodes

A **Sybil attack** involves an adversary creating multiple fake identities (**Sybil nodes**) that follow or are followed by a small number of real (honest) users to gain influence or manipulate the network.

$V = V_H \cup V_S$, where V_H is set of honest nodes and V_S is set of Sybil nodes, such that $V_H \cap V_S = \emptyset$.

$E = E_H \cup E_S \cup E_B$, where $E_H \subseteq V_H \times V_H$ edges between honest users. $E_S \subseteq V_S \times V_S$ edges among sybil nodes. $E_B \subseteq (V_H \times V_S) \cup (V_S \times V_H)$

In a Sybil attack, the attacker generate large number of Sybil nodes V_S, and minimally connects them to honest nodes V_H through a small set of edges E_B to avoid detection.

$$|V_S| \gg |E_B| \quad (1)$$

The number of Sybil nodes is much greater than the number of bridge edges between the Sybil region and the honest region.

3.2 Overview Of Graph Convolutional Networks

Graph Convolutional Networks (GCNs) are a class of neural networks designed to operate directly on graph-structured data. In contrast to traditional neural networks that process grid-like data (e.g., images), GCNs can learn representations embeddings of nodes by aggregating information from their neighbors, capturing both feature and structural information.

$$G = (V, E), \text{ with } |V| = N \quad (2)$$

Where adjacency matrix of graph $A \in R^{N \times N}$ and $X \in R^{N \times F}$ feature matrix where each row x_i is a feature vector for node v_i.

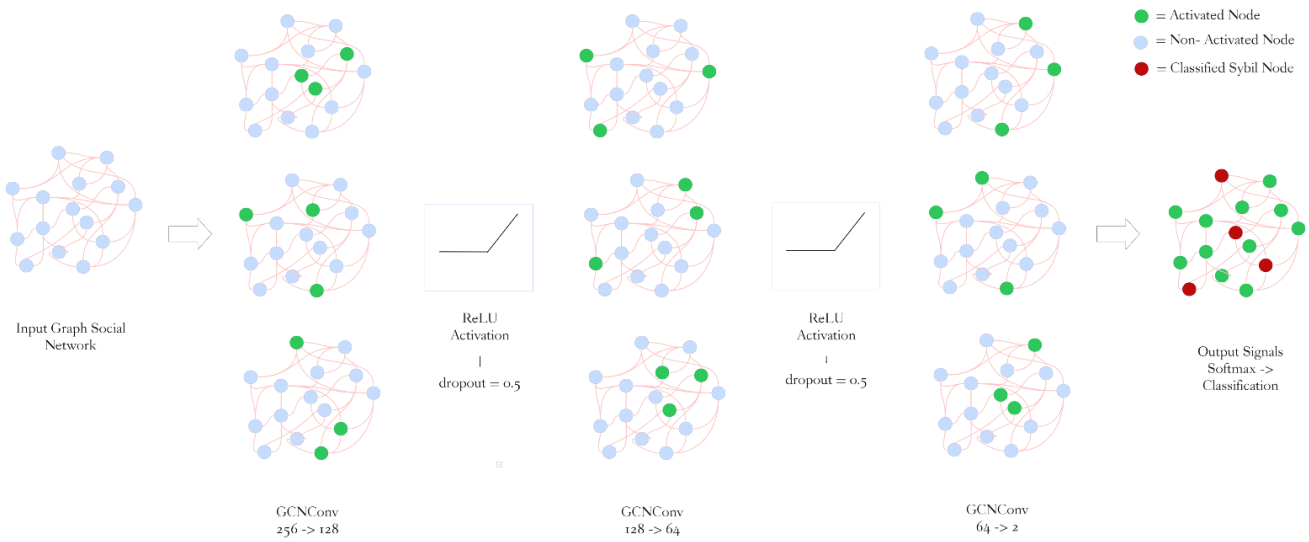

Fig 2. Graph Convolutional Networks (GCN) Architecture

GCNs operate based on the **message passing** paradigm, each node receives **messages from its neighbors** (i.e., the node gathers features from connected nodes) as shown in *Fig 3*. Messages are **aggregated** using a permutation-invariant function. The node's own representation is updated based on the aggregated message. Initially, the input graph consists of node features represented by $H^{(0)} = X \in R^{N \times 256}$, where N number of nodes and each node has a 256-dimensional feature vector.

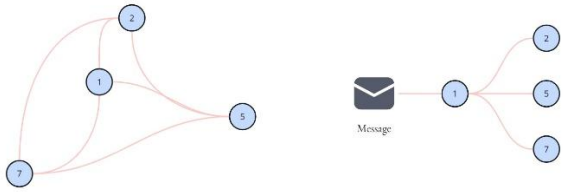

Fig 3. Message-passing architecture

The first *GCNConv* layer performs a neighborhood aggregation using the normalized adjacency matrix, transforming the features with a trainable weight matrix $W^{(0)} \in R^{256 \times 128}$, followed by a ReLU activation function. The second GCN layer continues this process, further compressing the features to 64 dimensions. In the final GCN layer, the features are reduced to 2-dimensional output, This output is passed through a softmax function. In the final classification result as shown in ***Error! Reference source not found.***, green nodes represent a ctivated nodes with meaningful embeddings, red nodes denote those classified as Sybil (having high probability for the Sybil class), and blue nodes are non-activated or uncertain, indicating low or no participation in the classification outcome.

3.3 Convolutional in CNNs & GCNs

Convolutional operations form the foundation of both Convolutional Neural Networks and Graph Convolutional Networks. Although both paradigms rely on the concept of local aggregation, the notion of *neighborhood* and the way convolution is performed differ fundamentally due to the underlying data structure regular grids for images and irregular graphs for networks.

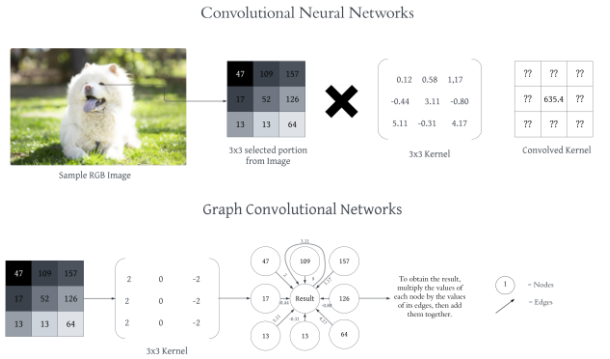

Fig 4. Difference between CNNs & GCNs

In CNNs, data are represented on a regular Euclidean grid, such as a 2D image. Convolution is performed by sliding a fixed-size kernel over the image and computing a weighted sum of pixel values within a local spatial neighborhood. Unlike images, graphs are non-Euclidean and irregular. Nodes do not lie on a fixed grid, and each node can have a different number of neighbors. Therefore, convolution in GCNs is defined as a neighborhood aggregation operation over graph connectivity. In *OSN*, users and their relationships naturally form a graph structure. Since Sybil accounts often exhibit abnormal connectivity patterns, GCNs are particularly well suited for capturing structural dependencies that CNNs cannot model.

4. METHODLOGY

In this study, a real-world Facebook follower–followee graph dataset is utilized to perform Sybil detection. The graph is modeled with nodes representing users and directed edges denoting social interactions. To enable effective learning, the graph is transformed into a structured representation by extracting relevant node-level features that capture both structural and behavioral properties of users. These features include metrics such as in-degree, out-degree, connectivity patterns, and neighborhood characteristics, which help distinguish Sybil nodes from benign users. The extracted features are used as input to the learning framework, where a Graph Convolutional Network (GCN) architecture is employed as described above. Through this process, high-quality node representations are learned, which are informative for distinguishing between Sybil and benign users. Node features

are propagated and transformed across multiple GCNConv layers, allowing each node to aggregate information from its neighbors and learn meaningful embeddings that capture both local graph structure and attribute information.

To ensure stable learning, preprocessing steps such as feature normalization are applied, and the dataset is divided into training and testing subsets. The model is trained in a supervised manner using labeled nodes, where each node is assigned a class label indicating whether it is a Sybil or a non-Sybil user. The final layer performs binary classification for each node using a softmax function, predicting the probability of a node belonging to either class. The performance of the model is evaluated using standard metrics such as accuracy, precision, recall, and F1-score to assess its effectiveness in identifying Sybil accounts.

5. EXPERIMENT AND RESULTS

The a real-world Facebook follower-followee graph dataset[1] used to evaluate Sybil detection approach. The graph consists of a total of 8,078 nodes and 353,136 directed edges, representing users and their social connections. The average node degree is 87.43, with degrees ranging from a minimum of 2 to a maximum of 2,090, and a median degree of 50.0. Each node is niquely identified in the range 0 to 8077.

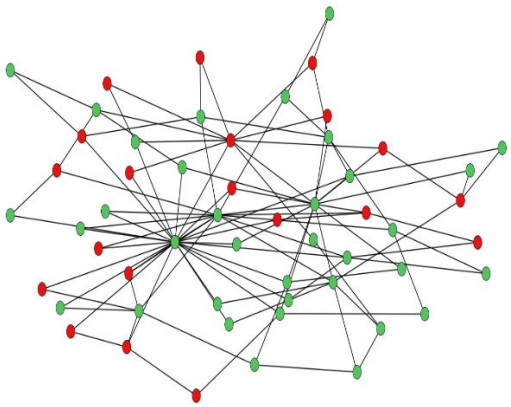

Fig 5. Network visualization

A small balanced subset of 200 nodes (100 benign and 100 Sybil) was used for training, while the remaining 7,878 nodes (3,939 benign and 3,939 Sybil) were reserved for testing.

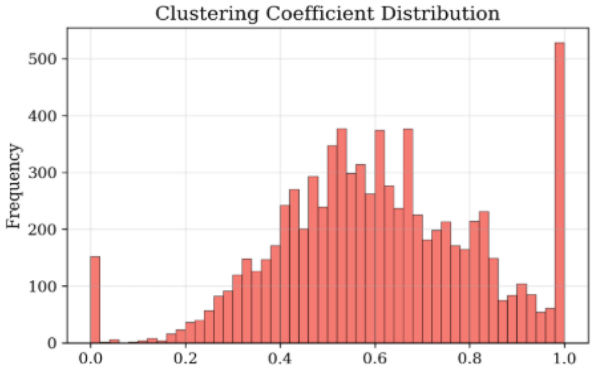

Fig 6. Clustering Distribution

5.1 Experimental Setup

Node features were generated randomly, resulting in a feature matrix of shape (8078, 256). These 256-dimensional vectors were used as input features in the Graph Convolutional Network. During pre-processing, isolated nodes were mapped to indices from 0 to 8077, and no additional attributes like degree or clustering coefficient were manually engineered. The graph was then converted into a format compatible with PyTorch Geometric[2], with node features $X \in R^{8078 \times 256}$, edge indices defining connectivity, and labels for supervised binary classification. All the experiments were conducted using the PyTorch framework with **PyTorch Geometric v2.6.1**, running on an environment configured with **CUDA 11.8** and an **NVIDIA RTX 3060 GPU**. The model was optimized using the Adam optimizer with a learning rate of 0.01, weight decay of 0.0005, and trained for 300 epochs. In the testing phase, the entire dataset of 8,078 nodes was relabelled, and fed through the trained model. The GCN processed these nodes and generated predictions for each, ultimately classifying ***4,044 nodes as benign*** and ***4,034 as Sybil.*** The model's adaptability and capacity to generalize learnt patterns throughout the whole social network graph are demonstrated by this full-graph inference.

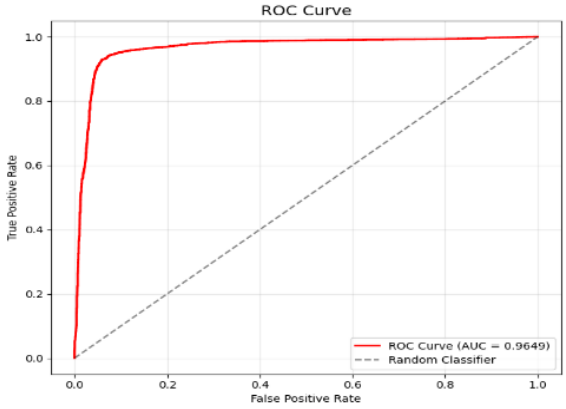

Fig 7. ROC Curve

5.2 Evaluation Metrics

To assess the effectiveness of the proposed GCN-based Sybil detection model several standard classification metrics. The GCN model achieved an accuracy of 93.23%, indicating strong overall performance. A precision shows that false positives were minimal, while a recall confirms that most Sybil nodes were correctly identified. The F1-score of 93.23% reflects a balanced trade-off between precision and recall. Additionally, an AUC score of 0.9649 demonstrates the model's excellent ability to distinguish between Sybil and benign nodes.

5.3 Results Comparison

To evaluate the effectiveness of proposed GCN-based Sybil detection framework *(SD-GCN)*, conducted extensive experiments on a real-world Facebook follower-followee graph. The model was trained on a small labeled subset and tested on the entire graph of 8,078 nodes, equally split between Sybil and benign classes. The model successfully classified 4,044 nodes as benign and 4,034 as Sybil, demonstrating its ability to generalize.

[1] https://snap.stanford.edu/data/ego-Facebook.html

[2] https://pytorch-geometric.readthedocs.io/en/latest/

Table 1. Comparison with existing methods

Algorithm	AUC
SybilRank [13]	0.82
SybilSCAR [21]	0.64
SybilWalk [14]	0.94
SybilSCAR-D [22]	0.95
SybilGAT-L2 [23]	0.76
SybilGAT-L4 [23]	0.60
SybilGAT-L8 [23]	0.46
SD-GCN (ours)	**0.96**

The model achieved a 93.23% accuracy, and AUC score of 0.9649 further confirms the model's excellent discriminative capability between Sybil and benign nodes. Proposed method achieves the highest AUC score among all listed approaches as shown in

Table *1*, outperforming even strong baselines like SybilWalk and SybilGAT.

6. CONCLUSION

In this study, we presented a Graph Convolutional Network (GCN)-based approach, **SD-GCN**, for detecting Sybil nodes in online social networks using the Facebook follower-followee graph as a real-world testbed. The system demonstrates that even with limited labeled data, GCNs can effectively leverage the structural properties of the social graph and learn meaningful node representations through message passing and aggregation. The proposed model achieved a high classification accuracy and significantly outperformed several existing methods in terms of AUC. By incorporating graph topology the GCN architecture captures both local and global patterns that are essential for distinguishing Sybil nodes from genuine users. Overall, this system provides robust solution for Sybil detection, which can be integrated into real-time social network monitoring tools to enhance trust and security.

In terms of future scope, several promising directions can be explored to extend this work. First, integrating attention mechanisms through Graph Attention Networks could allow the model to assign different importance weights to different neighbours during aggregation, improving detection of Sybil nodes with selective connectivity patterns. Second, extending the framework to temporal and dynamic graphs would enable detection of Sybil accounts that gradually infiltrate networks over time rather than appearing all at once. Finally, exploring federated learning approaches could allow SD-GCN to be trained across multiple decentralized social network platforms without sharing raw user data, addressing privacy concerns while maintaining detection effectiveness.

7. REFERENCES

[1] S. E. M. a. A. Hamzah, "Online Social Networking: A New Form of Social," International Journal of Social Science and Humanity, pp. 96-104, 2011.

[2] M. G. O. M. D. J. Thomas Aichner, "Twenty-Five Years of Social Media: A Review of Social Media Applications and Definitions from 1994 to 2019," CYBERPSYCHOLOGY, BEHAVIOR, AND SOCIAL NETWORKING, pp. 215-222, 2021.

[3] M. B. P. S. M. P. Heta Dasondi, "A Proposed Blockchain-Based Model for Online Social Network to Detect Suspicious Accounts," Singapore, 2023.

[4] J. R. Douceur, "The Sybil Attack," in Lecture Notes in Computer Science, Berlin, Heidelberg, 2002.

[5] H. X. M. L. H. H. S. Z. H. L. X. &. R. K. Zheng, "Smoke screener or straight shooter: Detecting elite sybil attacks in user-review social networks.," arXiv preprint arXiv:1709.06916., 2017.

[6] S. R. S. J. K. Ankit Kumar Jain, "Online social networks security and privacy: comprehensive review," Complex & Intelligent Systems, p. 2157–2177, 2021.

[7] B. Hogan, "Online Social Networks: Concepts for Data Collection and analysis," The Sage Handbook of Online Research Methods, Second edition, pp. 241-258, 2016.

[8] D.-A. S.-T. a. I.-S. M. Daniel Mican, "User Behavior on Online Social Networks: Relationships among Social Activities and Satisfaction," Symmetry, pp. 1-16, 2020.

[9] N. Z. F. M. &. M. P. Gong, "Sybilbelief: A semi-supervised learning approach for structure-based sybil detection.," IEEE transactions on information forensics and security, pp. 976-987, 2014.

[10] N. Z. G. a. H. F. Binghui Wang, "GANG: Detecting Fraudulent Users in Online Social Networks via Guilt-by-Association on Directed Graphs," IEEE International Conference on Data Mining (ICDM), 2017.

[11] Z. Y. a. Y. D. Y. Sun, "TrustGCN: Enabling Graph Convolutional Network for Robust Sybil Detection in OSNs," in IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), The Hague, Netherlands, 2020.

[12] X. L. ,. L. L. Jian Mao, "SybilHunter: Hybrid graph-based sybil detection by aggregating user behaviors," Neurocomputing, pp. 295-306, 2022.

[13] M. S. Y. P. Qiang Cao, "Aiding the detection of fake accounts in large scale social online services," in Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, 2012.

[14] J. a. W. B. a. G. N. Z. Jia, "Random Walk Based Fake Account Detection in Online Social Networks," 2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), pp. 273-284, 2017.

[15] H. W. W. J. L. L. Shangbin Feng, "SATAR: A Self-supervised Approach to Twitter Account Representation Learning and its Application in Bot Detection," arXiv:2106.13089v4, 2021.

[16] D. G. L. L. L. Haoyu Lu, "SybilHP: Sybil Detection in Directed Social Networks with Adaptive Homophily Prediction," Applied Sciences., 2023.

[17] T. O. S. H. &. K. N. Talaei Khoei, "Deep learning: systematic review, models, challenges, and research directions," Neural Comput & Applic, p. 23103–23124, 2023.

[18] C. Y. B. P. K. R. Patel Devarshi, "A Comprehensive Deep Learning Model for Improved Person Re-identification

Using Multi-Camera Streaming Pipeline," Procedia Computer Science, pp. 455-466, 2025.

[19] B. P. S. K. K. e. a. Khemani, "A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions," J Big Data, 2024.

[20] S. T. H. X. J. e. a. Zhang, "Graph convolutional networks: a comprehensive review," Comput Soc Netw, 2019.

[21] L. Z. a. N. Z. G. B. Wang, "SybilSCAR: Sybil detection in online social networks via local rule based propagation," IEEE INFOCOM 2017 - IEEE Conference on Computer Communications, pp. 1-9, 2017.

[22] ,. J. J. Z. Z. G. Binghui Wang, "Structure-based Sybil Detection in Social Networks via Local Rule-based Propagation," IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, no. arXiv:1803.04321v2, 2020.

[23] S. P. A. &. W. R. Heeb, "Sybil Detection using Graph Neural Networks," arXiv preprint arXiv:2409.08631, 202.