Illinois Institute of Technology

CS-528 Data Privacy and Security

A Project Report on

Privacy Breach Detection and Security Analysis Using Differential Privacy

06/14/2024

By Poshan Pandey (A20519852)

1. Introduction

In the era of digital communication, social networks have become repositories of vast amounts of personal data. Ensuring the security and privacy of this data is crucial. This project aims to explore the enhancement of data privacy in social networks by detecting potential privacy breaches using advanced network analysis and machine learning techniques. Additionally, differential privacy is implemented to ensure the protection of individual data while maintaining the utility of the analysis.

2. Dataset

The dataset used in this project is the Twitter Social Network Dataset from the Stanford Large Network Dataset Collection (SNAP). This dataset contains information about Twitter users and their follower-followed relationships. It allows for the construction of a social network graph where nodes represent users and edges represent the follower-followed relationships. The dataset format includes:

Source: The user ID of the follower.

Target: The user ID of the followed user.

This structure is ideal for performing network analysis and applying centrality measures to understand the network's characteristics.

Dataset Size and Processing Constraints

The Twitter Social Network Dataset is extensive, containing millions of nodes and edges. Due to computational and processing power constraints, this project was conducted on a subset of 5000 nodes. This sample size was selected to ensure that the analysis could be performed efficiently within the available resources while still providing meaningful insights.

3. Algorithms and Implementation

3.1 Centrality Measures

Centrality measures are used to determine the importance of nodes within the network. The centrality measures calculated in this project include:

Degree Centrality: Indicates the number of direct connections a node has.

Clustering Coefficient: Measures the degree to which nodes in a graph tend to cluster together.

Betweenness Centrality: Indicates the number of times a node acts as a bridge along the shortest path between two other nodes.

Closeness Centrality: Measures how close a node is to all other nodes in the network.

Eigenvector Centrality: Measures the influence of a node in the network based on the influence of its neighbors.

3.2 Anomaly Detection Algorithms

Isolation Forest

Isolation Forest is an ensemble method based on decision trees. It isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. Anomalies are more likely to be isolated in fewer steps, making this method effective for anomaly detection in large datasets. This approach is particularly advantageous because of its efficiency and ability to handle large datasets intuitively.

Advantages:

- Efficiency in handling large datasets.
- Intuitive and easy to understand.

Disadvantages:

- Can be sensitive to the contamination parameter.
- May not perform well with high-dimensional data.

One-Class SVM

One-Class SVM is a type of Support Vector Machine used for anomaly detection. It attempts to find a decision boundary that best separates the data from the origin in a transformed feature space. This method is robust in high-dimensional spaces and can capture more complex patterns due to its flexibility with different kernel functions.

Advantages:

- Robust in high-dimensional spaces.
- Flexible with different kernel functions.

Disadvantages:

- Computationally intensive.
- Sensitive to parameter settings.

3.3 Differential Privacy

To ensure differential privacy, Laplace noise is added to the centrality measures. The Laplace mechanism helps to obscure the presence or absence of any single individual in the dataset by adding noise to the data. This method ensures that the privacy of individuals is protected while allowing for meaningful analysis of the dataset.

3.4 Community Detection

Community detection helps identify clusters of nodes that are more connected to each other than to the rest of the network. The algorithm used for community detection in this project is the Greedy Modularity Communities algorithm. This algorithm optimizes the modularity of the network by iteratively merging communities that result in the greatest increase in modularity. Modularity is a measure of the strength of division of a network into modules (or communities).

Advantages:

- Efficient in detecting large communities.
- Provides insight into the network structure.

Disadvantages:

- May not detect smaller communities effectively.
- Can be computationally intensive for very large networks.

3.5 Privacy Risk Scoring

Privacy risk scoring involves developing a system to score nodes based on their risk of privacy breaches. This scoring aggregates various centrality measures to compute the risk score. By considering multiple centrality measures, the scoring system provides a comprehensive view of the potential privacy risks associated with each node.

4. Implementation Steps

Data Preparation: Uploaded the social network dataset and created a graph using NetworkX.

Centrality Measures Calculation: Calculated various centrality measures for the nodes in the graph.

Anomaly Detection: Used Isolation Forest and One-Class SVM to detect anomalies based on centrality measures.

Differential Privacy: Added Laplace noise to centrality measures to ensure differential privacy and repeated the anomaly detection process.

Community Detection: Detected communities within the graph using the Greedy Modularity Communities algorithm.

Privacy Risk Scoring: Developed a privacy risk scoring system to score nodes based on their centrality measures.

Evaluation: Calculated accuracy and precision to evaluate the performance of the anomaly detection algorithms.

5. Results

5.1 Initial Accuracy (Without Differential Privacy)

Isolation Forest:

Accuracy: 0.94

Precision: 0.02

One-Class SVM:

Accuracy: 0.31

Precision: 0.05

5.2 Accuracy with Differential Privacy

Isolation Forest with DP:

Accuracy: 0.94

Precision: 0.04

One-Class SVM with DP:

Accuracy: 0.50

Precision: 0.05

Accuracy after Adding Laplace Noise

Isolation Forest with DP:

Accuracy: 0.94

Precision: 0.12

One-Class SVM with DP:

Accuracy: 0.50

Precision: 0.05

6. Analysis and Discussion

Isolation Forest

Without Differential Privacy: Isolation Forest demonstrated high accuracy but low precision, indicating that while it was effective at identifying a majority of normal data points, it struggled to accurately identify anomalies.

With Differential Privacy: The accuracy remained high, but precision improved, suggesting that the addition of differential privacy helped to better identify anomalies without compromising the overall detection rate.

With Laplace Noise: The addition of Laplace noise further improved precision, indicating a significant enhancement in anomaly detection capabilities while maintaining high accuracy.

One-Class SVM

Without Differential Privacy: One-Class SVM showed lower accuracy and precision, suggesting difficulties in correctly identifying anomalies in the initial setup.

With Differential Privacy: The accuracy improved significantly, though precision remained constant, indicating that differential privacy helped in better identifying the boundaries between normal and anomalous data.

With Laplace Noise: Accuracy improved further, but precision did not show a significant change, indicating that while the model became better at overall classification, its ability to precisely identify anomalies did not change much.

Community Detection

The Greedy Modularity Communities algorithm effectively identified clusters within the network, providing additional insights into the network structure. By grouping nodes into communities, we could analyze the interactions within and between these groups, aiding in the identification of patterns and potential areas of vulnerability. This method of community detection is advantageous as it helps in simplifying the complex network structure into more manageable subgroups, thereby facilitating targeted analysis and interventions.

Privacy Risk Scoring

The privacy risk scoring system was instrumental in highlighting nodes with higher risks of privacy breaches. By combining multiple centrality measures, we could assign a comprehensive risk score to each node, prioritizing nodes for further investigation and security measures. This scoring system helps in effectively allocating resources to safeguard the most vulnerable parts of the network.

Comparison of Anomaly Detection

Without Differential Privacy:

- Isolation Forest: High accuracy but low precision.
- One-Class SVM: Low accuracy and precision.

With Differential Privacy:

- Isolation Forest: Maintained high accuracy with improved precision.
- One-Class SVM: Improved accuracy but stable precision.

After Adding Laplace Noise:

- Isolation Forest: High accuracy and significantly improved precision.
- One-Class SVM: Improved accuracy with stable precision.

7. Challenges

The Twitter Social Network Dataset is extensive, containing millions of nodes and edges. Due to computational and processing power constraints, this project was conducted on a subset of 5000 nodes. Processing large datasets requires significant computational resources, which can be a limiting factor. Ensuring that the selected subset is representative of the larger dataset was crucial to maintain the validity of the analysis.

8. Conclusion

This project demonstrated that integrating differential privacy into anomaly detection processes in social networks can improve the precision of anomaly detection algorithms while maintaining high accuracy. The use of Laplace noise to achieve differential privacy proved effective in protecting individual data without compromising the utility of the data analysis. Community detection and privacy risk scoring further enhanced the analysis by providing deeper insights into the network structure and identifying high-risk nodes.

9. Future Work

Enhanced Anomaly Detection Techniques: Explore more advanced machine learning algorithms and hybrid models to improve anomaly detection.

Real-time Implementation: Implement real-time anomaly detection systems to monitor social networks continuously.

Scalability: Test the scalability of the approach on larger datasets and more complex social networks.

Parameter Optimization: Fine-tune the parameters for both differential privacy (epsilon values) and machine learning models to achieve optimal performance.