Social Network

Mohammed Yasser El.Sharkawey 2205149

Assignment 2 Report:

1. Introduction

The objective of this assignment is to develop a machine learning-based bot detection system using graph topology features and to evaluate its robustness against adversarial attacks. We utilized the **SNAP Facebook dataset**, injecting synthetic bots to create a supervised learning scenario. The study compares the model's performance under three conditions: Baseline (No Attack), Structural Evasion (Test-time attack), and Graph Poisoning (Training-time attack).

2. Methodology

Graph Construction:

- **Dataset:** SNAP Facebook Combined (4,039 nodes, 88,234 edges).
- **Feature Extraction:** We extracted three key topological features for every node:
    1. **Degree Centrality:** Number of connections.
    2. **Clustering Coefficient:** Measure of the degree to which nodes tend to cluster together.
    3. **PageRank:** Measure of node influence. *Note: Community detection features were excluded to test the robustness of local metrics.*

Bot Simulation:

- **Baseline Bots:** 201 synthetic bots were injected. They formed a "ring" structure (connected to each other) and established random connections with humans.

3. Attack Scenarios

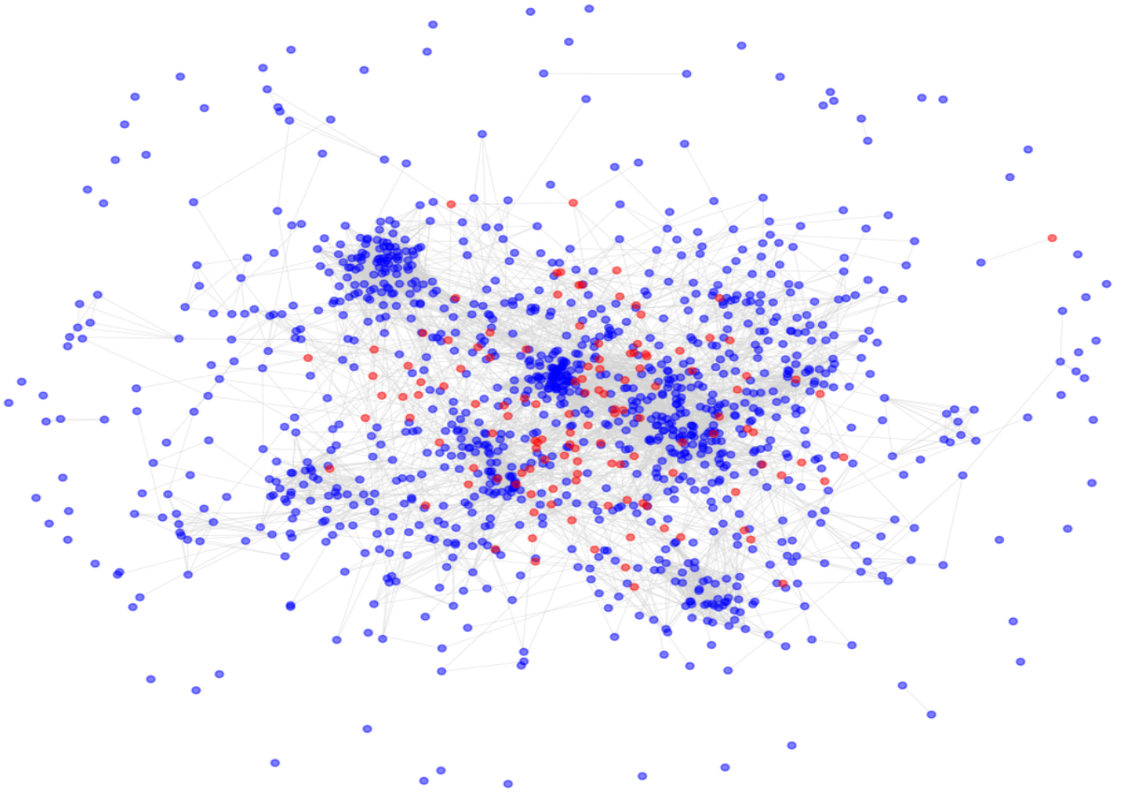
1. **Structural Evasion (Test-Time):** Bots actively altered their connections to mimic human behavior. They removed **100%** of their edges to other bots (breaking the botnet ring) and established connections with high-clustering human nodes to blend into social circles.
2. **Graph Poisoning (Training-Time):** 400 "Camouflage" nodes were injected into the training set. These nodes exhibited random behavior but were mislabeled as "Humans" (Class 0) to corrupt the model's decision boundary.

Aspect	Baseline	Structural Evasion Attack	Graph Poisoning Attack
Attack Type	No attack (clean scenario)	Test-time attack (bots modify their structure)	Training-time attack (poisoning the dataset)
Graph Manipulation	No manipulation, only injected bots with ring + random links	Bots remove all bot-bot edges and connect to high-clustering humans	400 fake nodes added, behave randomly but labeled as humans
Goal of Scenario	Train the model normally and measure natural detection ability	Make bots look like humans to evade detection	Corrupt the training labels to damage the model
Bot Behavior	Simple ring topology + random links	Mimic human clustering + no links between bots	Random noisy behavior but mislabeled as benign
Features Used	Degree, Clustering Coefficient, PageRank	Same features but bots modify them to overlap with humans	Same features extracted but training labels corrupted
Classifier	Random Forest	Same model (test-time evasion)	Random Forest retrained on poisoned data
Accuracy	**1.0000**	**0.9528**	**1.0000**
Bot Recall	**1.00** (perfect detection)	**0.00** (all bots misclassified as humans)	**1.00** (tested on clean set)
Impact on Model	Excellent detection	Completely breaks bot detection	No observed impact on performance
Reason for Result	Bots are structurally different from humans	Bots mimic human features, blending perfectly	Random Forest resists label noise and ignores poisoned nodes
Graph Visualization	Bots clustered together (easy to spot visually)	Bots dispersed and mixed with humans	Extra nodes appear but decision boundary unaffected

Aspect	Baseline	Structural Evasion Attack	Graph Poisoning Attack
Difficulty for Detection	Very easy	Extremely hard	Moderate–low
Key Insight	Local graph metrics detect naive bots	Graph-based models are vulnerable to feature-based evasion	Random Forest is robust to simple poisoning
Overall Risk Level	Low	**Very High**	Low

4. Experimental Results

A. Baseline Performance

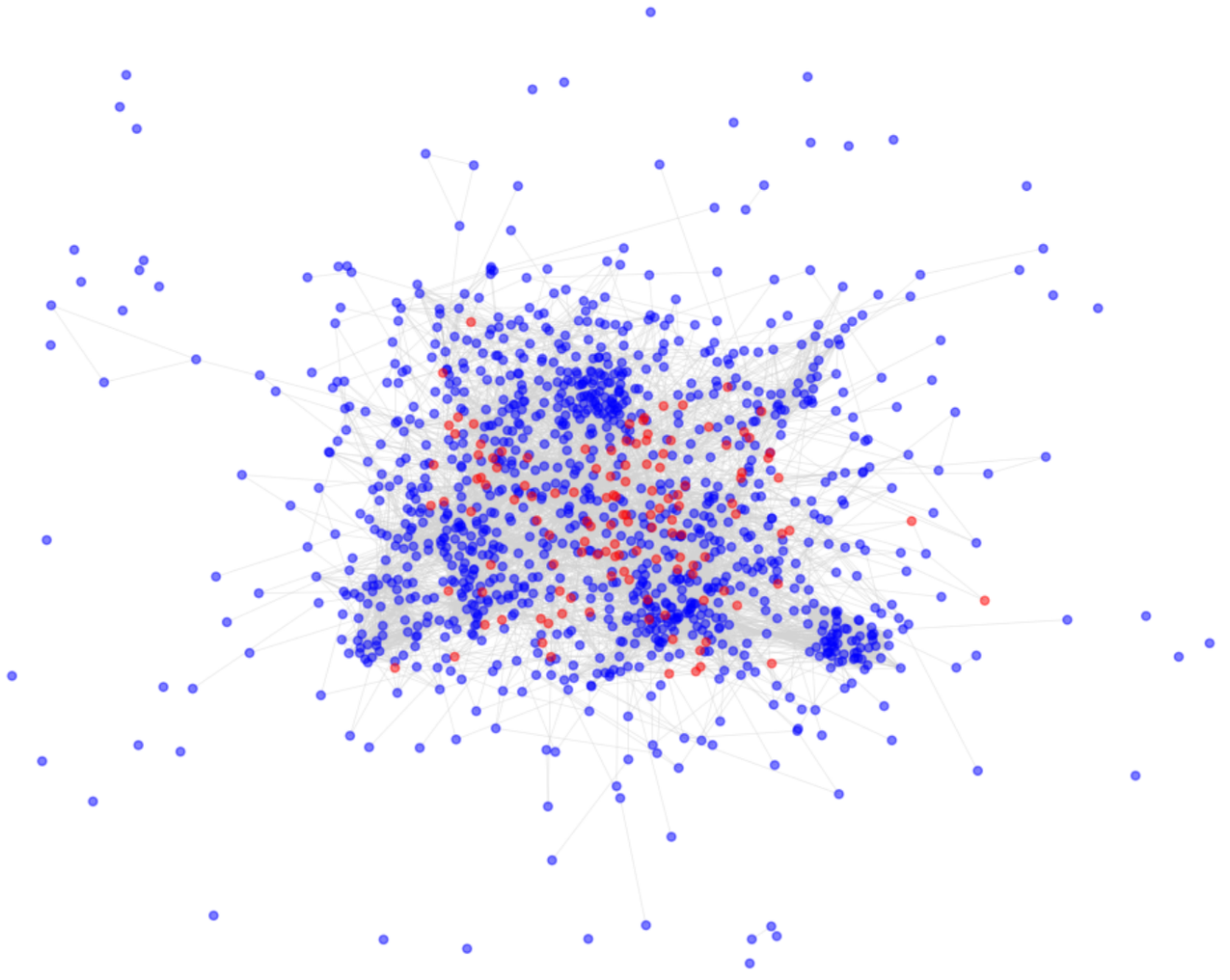


The Random Forest classifier achieved perfect detection on the initial dataset.

- **Accuracy:** 100%
- **Bot Recall:** 1.00

- **Observation:** Naive bots with random connections and ring structures are easily distinguishable from humans, who exhibit "small-world" properties (high clustering).

B. Impact of Structural Evasion

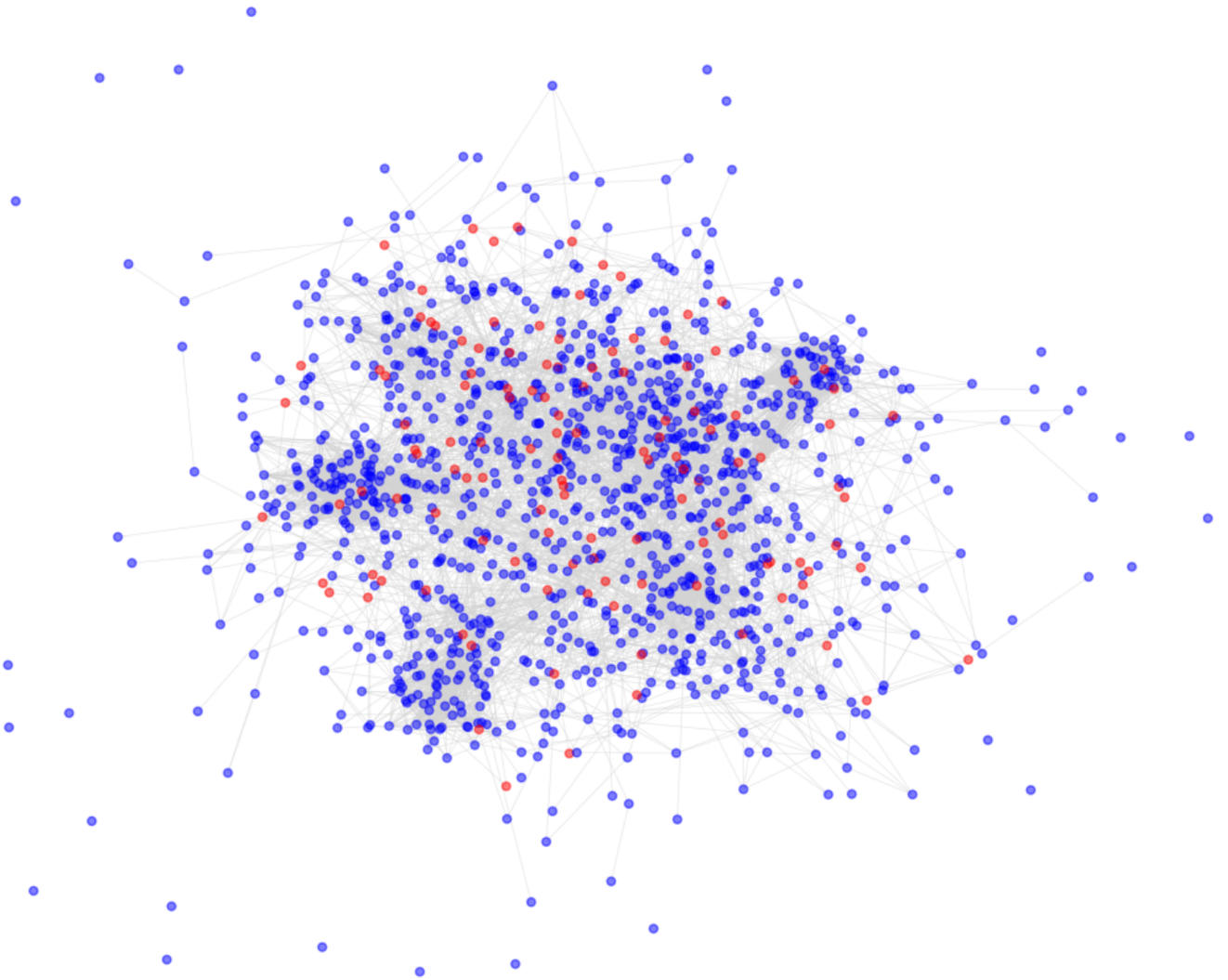


The evasion attack was **highly successful**, rendering the bots completely invisible to the classifier.

- **Accuracy:** Dropped to 95.28% (Only humans were correctly classified).
- **Bot Recall: 0.00** (Complete Failure of Detection).
- **Bot Precision:** 0.00

- **Analysis:** By removing links to other bots and connecting to high-clustering humans, the bots successfully manipulated their feature values (specifically Clustering Coefficient and PageRank) to overlap perfectly with the human distribution. The model classified 100% of the bots as humans.

C. Impact of Graph Poisoning



The poisoning attack had **no observed impact** on the classification of the original test set.

- **Accuracy:** Remained at 100%.

- **Analysis:** The Random Forest algorithm proved to be highly robust against label noise. Even with the injection of 400 mislabeled nodes, the ensemble nature of the model allowed it to filter out the noise and maintain the correct decision boundary for the distinct test set patterns.

5. Conclusion

This experiment demonstrates a critical vulnerability in graph-based security systems. While topological features are powerful for detecting naive attackers, they are extremely brittle against **Structural Evasion**.

- **Key Finding:** An attacker can evade detection with 100% success by simply mimicking the local connectivity patterns (Triadic Closure) of benign users.
- **Robustness:** Conversely, the model showed high resilience against **Poisoning**, suggesting that simple noise injection is insufficient to break ensemble classifiers like Random Forest.

6. Visualizations

(Attach the 3 generated images here: 1_baseline.png, 2_evasion.png, 3_poisoning.png)

- **Fig 1:** Shows bots clustered together (easy to spot).
- **Fig 2:** Shows bots dispersed among humans (impossible to spot visually or algorithmically).