TgrApp: Anomaly Detection and Visualization of Large-Scale Call Graphs

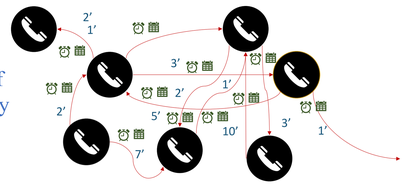
Mirela Cazzolato[1,2], Saranya Vijayakumar[1], Namyong Park[1], Meng-Chieh Lee[1],
Xinyi Zheng[1], Catalina Vajiac[1], Pedro Fidalgo[3,4], Bruno Lages[3],
Agma J. M. Traina[2], Christos Faloutsos[1]

[1]Carnegie Mellon University (CMU), [2]University of São Paulo (USP),
[3]Mobileum, [4]University Institute of Lisbon (ISCTE-IUL)

Introduction

Given: who-calls-whom and when network

Real life: millions of calls per day



Task: nodes with strange behavior

Given a dataset of semi-labelled call data, our goal is to:
1. Generate relevant features
2. Attention Route towards anomalous/suspicious nodes and patterns
3. Provide explanations for our predictions

Goals:

Effectiveness	Interactivity	Interpretability	Scalability
Separate fraudulent from non-fraudulent actors	Allow for Human-Computer Interaction	Well-defined features and intuitive plots	Handle millions of phone calls and subscribers

Challenges
1. Our training data is flawed – False Positives and False Negatives
2. High-dimensional data, millions of nodes → Latency
3. Explanations need to hold up in court of law

Methods: Our Tool

Figure 1: Feature extraction can point us to suspicious regions on the graph. We extract inter-arrival time statistics for time-evolving graphs and degree statistics to aid us in static analysis.

Figure 2: Pair plots help us summarize this information in n-dimensional space in a visual/interpretable way.

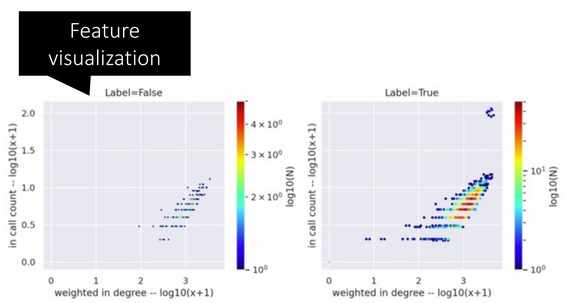


Figure 3: We also the pair plots of the nodes that are labelled as fraudulent versus non-fraudulent.

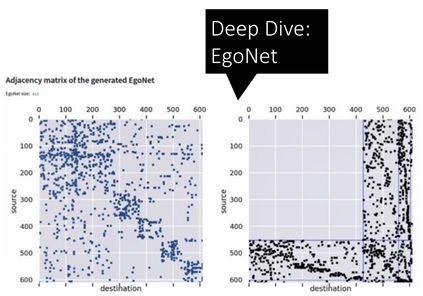


Figure 4: We visualize the adjacency matrix and a reordered matrix that shows cross-associations between the nodes of the EgoNet using an algorithm that groups similar nodes together. Therefore, if there are cliques, our visualization will highlight them.

Experimental Results

Case Study 1: international bypass. We found a group of phone numbers receiving one-second-long calls. People in country A call people in country B via a fraudulent telecommunications company that charges lower rates, in violation of regulations. Because of these practices, the calls end up dropping immediately.

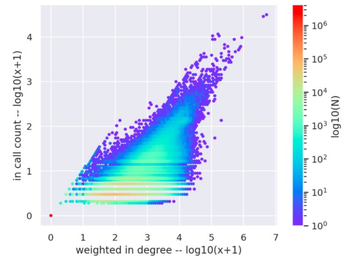


Figure 5: pair plot of weighted in-degree by in-call count (weight by duration)

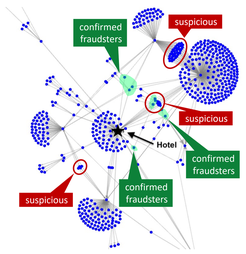


Figure 6: illustration of international bypass and pattern of fraudulent cliques

Case Study 2: camouflage. Here, callers with a lot of fraudulent international traffic evade filters that block numbers with a high fraction of international traffic by calling an equally high number of local numbers.

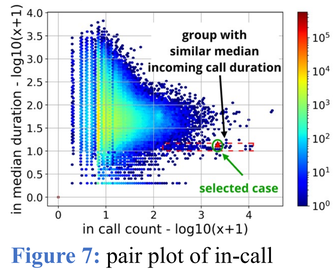


Figure 7: pair plot of in-call count vs in-median duration

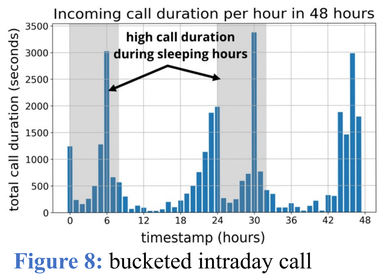
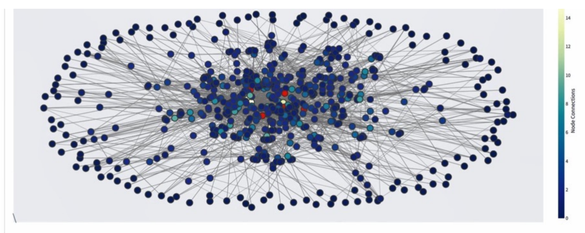


Figure 8: bucketed intraday call duration over time



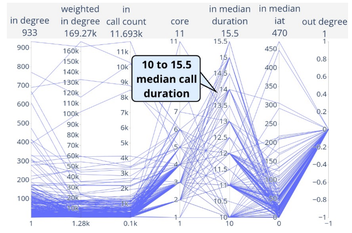


Figure 9a and 9b: Parallel coordinates assist the user to visually see the different feature values of nodes in the EgoNet, at the same time. Every orthogonal axis becomes a parallel axis, allowing us to visualize high-dimensional data.

Conclusions

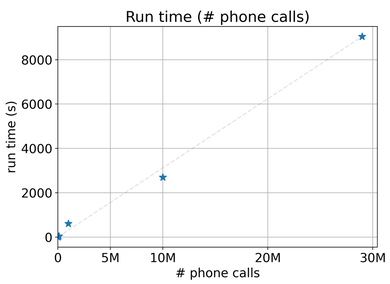


Figure 10: number of phone calls versus TgrApp runtime in seconds

- We need human-in-the-loop analysis, for fraud detection, interpreting results, and domain knowledge
- **TgrApp** provides these analysts with interpretable, clear visualizations and quick feature generation
- Runtime is linear (Figure 10)
- Visit github.com/mtcazzolato/tgrapp for open-source code for the tool

Fundação para a Ciência e a Tecnologia
EUROPEAN REGIONAL DEVELOPMENT FUND
EUROPEAN UNION
Conselho Nacional de Desenvolvimento Científico e Tecnológico