Cyber Shadows: Analyzing the Correlation Between Cyber Attacks and National Events

Lior Mishutin
314968306

Ben-Gurion University of the Negev, Israel
Department of Software and Information Systems Engineering
mishutin@post.bgu.ac.il

Abstract

This study explores cyber attack dynamics and their correlation with geopolitical events using a Kaggle synthetic dataset and the GDELT database. We map cyber battlegrounds through IP address conversion, apply the PageRank algorithm for network analysis, and develop a predictive model using link prediction techniques. Additionally, we assess national transparency in cyber incident reporting through GDELTDoc analysis. Our findings illuminate cyber adversaries tactics and national openness, offering a new framework for predicting cyber attacks and evaluating public disclosure.

1 Introduction

As digital threats escalate globally, analyzing cyber attack patterns becomes crucial for enhancing cyber security measures. The imperative to forecast and mitigate cyber threats is underscored by a recent study published in Nature (Almahmoud et al., 2023), which advocates for a proactive and predictive approach to cyber security, highlighting the shift towards anticipatory defenses against digital threats. This study leverages synthetic data from Kaggle (Team Incribo, Year of Publication) and real-world event records from the Global Database of Events, Language, and Tone (GDELT) (Leetaru and Schrodt, 2013) to dissect the dynamics of cyber attacks, their correlation with global events, and the transparency of affected nations in reporting these incidents.

Central to our analysis is the application of the PageRank algorithm, repurposed to evaluate the prominence of nations within the cyber conflict network. Furthermore, we develop a predictive model based on link prediction algorithms to forecast potential future cyber attacks, offering a proactive tool for cybersecurity preparedness. The methodologies, code, and data used in this study are accessible for review and replication in our project repository on GitHub (Mishutin, 2024).

Complementing our technical analysis, we examine global media coverage via GDELTDoc, assessing the transparency with which governments disclose cyber incidents to the public. Our research presents a holistic view of the cyber security landscape, providing insights into the tactics of cyber adversaries and the openness of nations in the digital age. By bridging technical analysis with socio-political considerations, we aim to contribute a comprehensive framework to the cybersecurity domain, enhancing understanding and facilitating the development of effective defense mechanisms against cyber threats.

2 Related Work

Our study intersects various innovative approaches in cyber attack prediction, network analysis, and the public disclosure of such attacks. This section organizes relevant literature into categories that mirror the facets of our research.

2.1 Cyber attack Graph Analysis Using Page Rank

The utilization of the Page Rank algorithm, initially developed for ranking web pages, has found its place in the cyber security domain, particularly in analyzing and prioritizing elements within cyber attack graphs. Studies have demonstrated the algorithm's efficacy in identifying critical vulnerabilities and attack paths within complex network structures, allowing for targeted defensive measures. For instance, research highlighted by Cornell University's Networks Course blog discusses leveraging Page Rank in cyber security to model potential security vulnerabilities within attack graphs, thus enabling system administrators to focus on the most critical threats(Mehta et al., 2006). Similarly, an investigation into the viability of Page Rank for attack graph analysis showcased its ability to effectively pinpoint vulnerable nodes, suggesting its potential as a valuable tool in cyber security

analysis(Dypbukt Källman, 2023).

2.2 Link Prediction in Graphs

Link prediction techniques aim to forecast connections within a graph based on its existing structure and features, focusing on the likelihood of a link's future appearance between nodes. While these methodologies are widely applied across social networks and biological data, direct applications to predicting cyber attacks between countries using graph-based link prediction remain unexplored. However, the foundational strategies (Liben-Nowell and Kleinberg, 2003) for link prediction showcase the potential adaptability of these methods for cyber security domains, hinting at an innovative frontier for research in the intersection of graph theory and cyber threat intelligence.

2.3 Public Disclosure of Cyber attacks

Exploring the timeline and transparency of cyber attack disclosures by affected nations adds a novel dimension to cyber security research. Our method of correlating cyber attack instances with subsequent news publications to gauge a country's openness provides valuable insights into the socio-political aspects of cyber security. While direct parallels in existing literature are scarce, this approach echoes broader interests in understanding the impacts of cyber incidents on national security policies and public awareness. The analysis of disclosure timelines can be related to studies focusing on the implications of public awareness and governmental transparency in the aftermath of cyber incidents.

3 Methodology

Our methodology is structured to facilitate a comprehensive analysis of cyber attack patterns, leveraging both synthetic and real-world data to construct a detailed landscape of global cyber conflicts. The methodology section is divided into several subsections, beginning with the creation of a graph that represents cyber attacks between countries.

3.1 Cyber Attacks Graph Creation

We developed a directed graph to depict cyber attacks, translating IP addresses into countries to illustrate both attack origins and targets. This graph, where nodes represent countries and weighted edges denote attack frequency, quantifies the intensity of cyber conflicts. Figure 1 showcases the graph's utility, highlighting attacks involving Israel as both aggressor and victim.

3.2 Application of the PageRank Algorithm

We applied the PageRank algorithm to our cyber attack graph, repurposing this web page ranking method to assess the roles of countries within the cyber conflict network. Nodes represent countries, and weighted edges show the attack frequency, allowing PageRank to score each country's prominence in the network. This analysis reveals key actors in cyber warfare, identifying nations central to the network as major targets or attackers. Figure 2 illustrates these rankings, with higher scores highlighting countries of significant influence or vulnerability in cyber conflicts.

3.3 Link Prediction Model for Cyber Attack Forecasting

We applied link prediction to forecast cyber attacks within our network, utilizing it as a strategic cybersecurity tool. This approach involves analyzing historical data and node relationships to predict future cyber interactions, thereby enriching our cyber threat understanding.

- **in_degree_v and out_degree_u**: The number of incoming and outgoing connections for nodes v and u, respectively, indicating the frequency of being attacked and initiating attacks.
- **total_countries**: The total number of unique countries involved in either incoming or outgoing attacks with nodes u or v, reflecting their engagement in the network.
- **common_countries**: The number of countries that have cyber interactions with both u and v, suggesting a shared sphere of cyber conflict.
- **jaccard_coefficient**: A similarity measure based on the common neighbors of u and v, indicating potential shared vulnerabilities or geopolitical contexts.
- **adamic_adar_index**: A measure that weights common neighbors by their degrees, emphasizing the importance of shared connections with less common countries.
- **page_rank_v and page_rank_u**: The PageRank scores of nodes v and u, providing insight into their centrality and influence in the network.
- **countries_measure**: A similarity measure based on the common neighbors of u and v.
- **avg_weight_out_u and avg_weight_in_v**: The average weight of outgoing and incoming

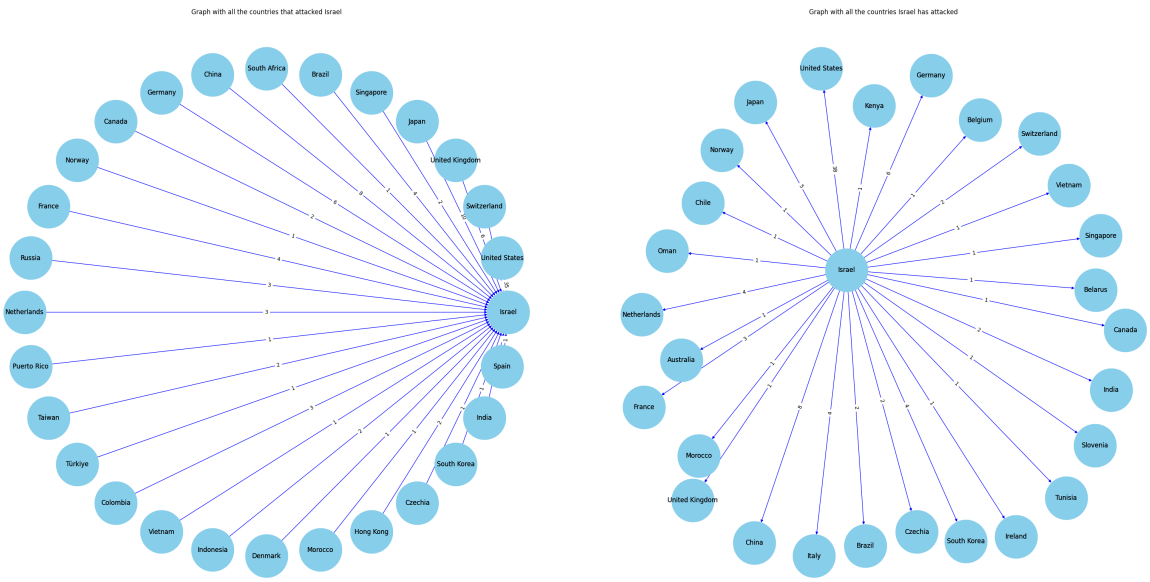

Figure 1: Left: Countries that have attacked Israel through cyber attacks. Right: Countries that Israel has attacked through cyber attacks.

edges for nodes u and v, offering a measure of the typical attack severity.

These features capture the nuanced dynamics of international cyber conflicts, aiding in the prediction of future attacks.

Negative Link Selection and Model Optimization: To refine our dataset, we selected negative samples using a second-degree neighbor approach, enhancing the model's realism. The selection process involves randomly choosing pairs of countries that are two steps away from each other in the network, ensuring that these pairs do not have a direct link. We trained our model with an XGBoost classifier, employing grid search for hyperparameter optimization. The optimized parameters — Colsample by tree of 0.7, learning rate of 0.1, max depth of 2, n estimators set to 100, and a sub sample rate of 1 — were chosen to balance model complexity and accuracy. This process enabled the effective forecasting of cyber attacks, as confirmed by high AUC and accuracy scores in our evaluations, illustrating the model's predictive strength and the potential of machine learning in cybersecurity.

3.4 Utilization of GDELT Database for Public Disclosure Analysis

A pivotal aspect of our study involves examining the transparency of nations in disclosing cyber attacks to the public. For this purpose, we leveraged the Global Database of Events, Language, and Tone (GDELT), which archives global news coverage across various themes, including cyber attacks. This database provides an invaluable resource for analyzing the temporal and thematic dimensions of how cyber attacks are reported and discussed worldwide.

To integrate GDELT's vast repository of news articles with our dataset, we developed a methodical approach for identifying and analyzing news coverage related to specific cyber attacks. This process began with the extraction of the first publication date and the total count of articles related to each cyber attack incident in our dataset, focusing on the two-week period following the attack. This timeframe was chosen to capture the immediate public and governmental response. **Our approach entails the following steps:**

1. Temporal Alignment: For each cyber attack instance, we calculate the period of interest, starting from the date of the attack and extending for two weeks. This period is crucial for capturing the most relevant and immediate media coverage following an attack.

2. Keyword and Country Filtering: Utilizing the GDELT API, we filter news articles based on keywords related to the cyber attack and the countries involved (both the attacker and the target). This step ensures that the articles retrieved are directly relevant to the incidents under scrutiny.

3. Article Retrieval and Analysis: For every relevant article found, we record its publication

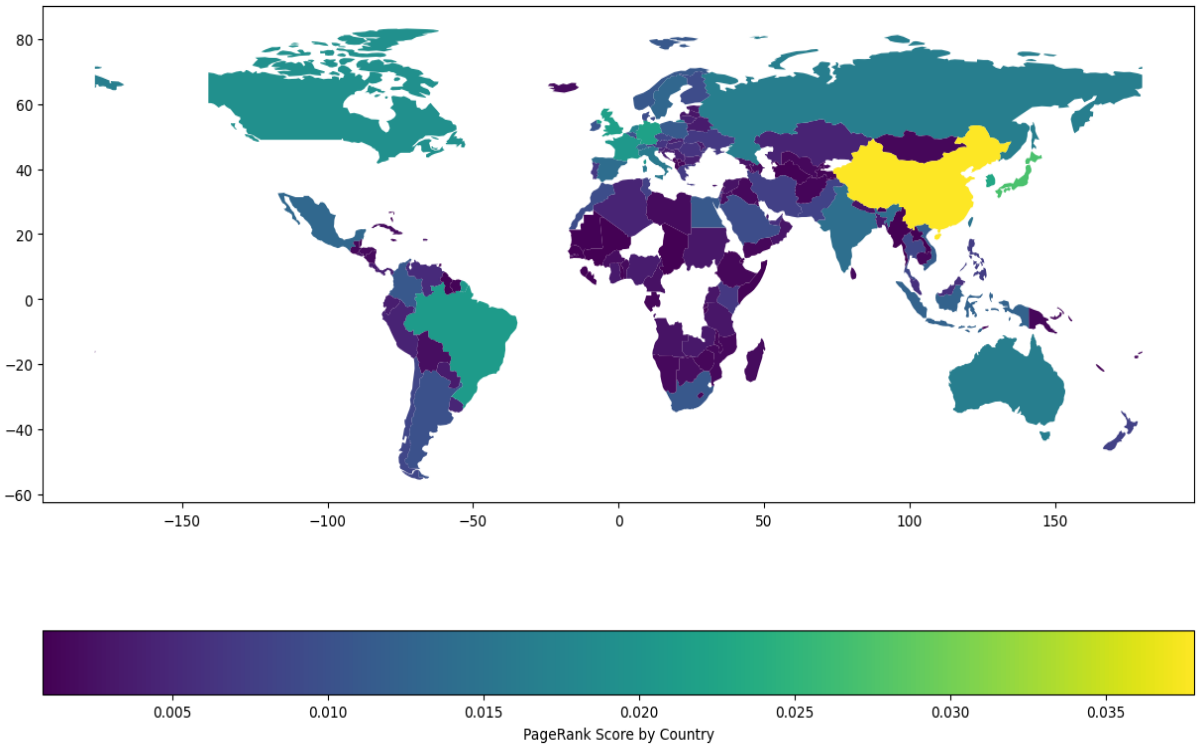

Figure 2: A heatmap of PageRank Score by Country.

date and increment the article count for the respective cyber attack incident. The earliest publication date among these articles is noted as the first public disclosure of the attack.

This process enables us to analyze not only the geopolitical and technical facets of cyber attacks but also the public and governmental openness in acknowledging and discussing these incidents. Through this analysis, we aim to shed light on the patterns of cyber attack disclosure, offering insights into how different nations address the public communication aspect of cybersecurity breaches.

4 Results

Our study's findings elucidate the dynamics of cyber attacks and the patterns of public disclosure associated with these incidents, leveraging a novel link prediction model and comprehensive analysis of global news coverage.

4.1 Performance of the Link Prediction Model

The efficacy of our link prediction model in forecasting potential cyber attacks was evaluated using two primary metrics: the area under the Receiver Operating Characteristic (ROC) curve (AUC) and accuracy. Figure 3 illustrates the ROC curve obtained from the model, showcasing an AUC of 1, indicating a high level of predictive accuracy. Additionally, the model achieved an accuracy of 99.9%, demonstrating its robustness in distinguishing between positive and negative instances of cyber attacks within the network.

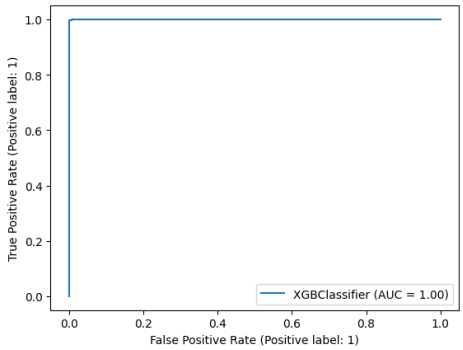

Figure 3: ROC Curve for the Link Prediction Model.

These results underscore the potential of machine learning algorithms, specifically ensemble methods like XGBoost, in enhancing cybersecurity defenses by providing anticipatory insights into likely cyber attack vectors.

4.2 Analysis of Cyber Attack Event Disclosure

We analyzed cyber attack disclosures timing using the GDELT database, focusing on the delay to first news publication and attack severity across

countries (Figures 4 and 5). Our analysis aimed to discover correlations between attack severity and disclosure speed. However, no clear correlation was found, possibly due to the synthetic nature of our dataset, which may not reflect real-world disclosure behaviors accurately. This underscores the importance of further studies with real-world data to explore this potential relationship more deeply.

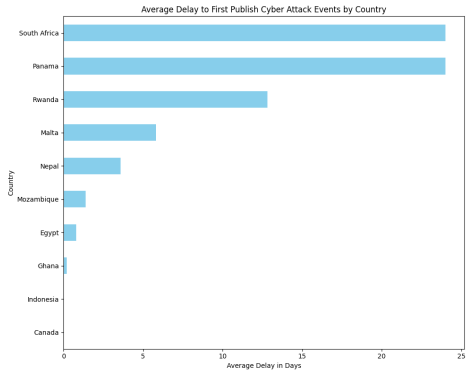

Figure 4: Average Delay to First Publish Cyber Attack Events by Country.

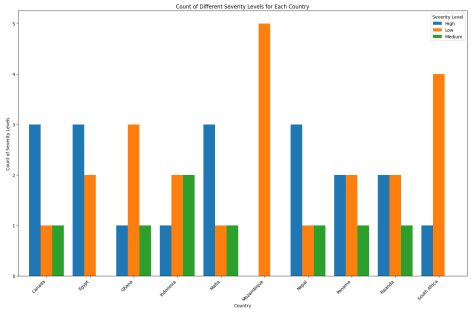

Figure 5: Count of Different Severity Levels for Each Country.

In summary, our results offer critical insights into the capabilities of link prediction models in cybersecurity and the complex dynamics of cyber attack disclosures. While the synthetic dataset presents certain limitations, our study lays the groundwork for future research aimed at enhancing the understanding and mitigation of cyber threats through advanced predictive analytics and public discourse analysis.

5 Conclusions and Future Directions

This study introduced a comprehensive framework for analyzing and predicting cyber attacks using a novel link prediction model and an in-depth examination of global news coverage on cyber incidents. Our findings demonstrate the effectiveness of machine learning algorithms in forecasting cyber attacks and reveal complex patterns of public disclosure across different nations. Despite the predictive success of our model, as evidenced by a high AUC and accuracy, the analysis of disclosure timeliness and attack severity did not yield a clear correlation, potentially due to the synthetic nature of the cyber attack dataset employed.

Conclusions: Our research underscores the potential of data science and network analysis techniques in enhancing cybersecurity measures and understanding the socio-political dynamics of cyber attack disclosures. The application of the PageRank algorithm and the XGBoost classifier within our study illustrates the valuable insights that can be gained from applying advanced analytical methods to cybersecurity data.

Future Directions: To build on the foundations laid by this study, future research should aim to incorporate real-world data to verify the patterns observed in synthetic datasets. Additionally, exploring other machine learning and deep learning models could offer further improvements in predictive accuracy and insights into cyber attack dynamics. Investigating the impact of different types of cyber attacks and their severity on the speed and nature of public disclosures would also be a valuable area for future study, providing a more nuanced understanding of global responses to cyber threats.

References

Zaid Almahmoud, Paul D Yoo, Omar Alhussein, Ilyas Farhat, and Ernesto Damiani. 2023. A holistic and proactive approach to forecasting cyber threats. *Scientific Reports*, 13(1):8049.

Marcus Dypbukt Källman. 2023. Exploring the viability of pagerank for attack graph analysis and defence prioritization.

Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pages 1–49. Citeseer.

David Liben-Nowell and Jon Kleinberg. 2003. The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559.

Vaibhav Mehta, Constantinos Bartzis, Haifeng Zhu, Edmund Clarke, and Jeannette Wing. 2006. Ranking attack graphs. In *International Workshop on Recent Advances in Intrusion Detection*, pages 127–144. Springer.

Lior Mishutin. 2024. Data mining project on cyber attack analysis. https://github.com/LiorMish/data_mining_project.

Team Incribo. Year of Publication. Cyber security attacks. https://www.kaggle.com/datasets/teamincribo/cyber-security-attacks. Accessed: date.