Integrative Network Analysis: Unveiling Symptom-Disease Interactions and Enhancing Predictive Models

Andreoli C. • Ligari D. • Alberti A. • Scardovi M. [1]

[1] *Department of Computer Engineering, Data ScienSickness predictionce, University of Pavia, Italy*
Course of Financial Data Science

Github repository: https://github.com/DavideLigari01/financial-project

Abstract

This report presents the experimental results of a project focused on assessing the impact of a DNS reflection and amplification attack. The study explores different amplification factors based on the DNS request types used and analyzes their effects on the targeted system and the DNS server within a local network environment. The results indicate that higher amplification factors correspond to increased latency during the attack, with some query times exceeding a threshold of 100 ms. However, the attack primarily affects DNS requests rather than causing widespread disruption to the entire system, as demonstrated by the analysis using the ping command-line tool. Surprisingly, the attack has no significant impact on system resources, such as RAM and CPU utilization, suggesting efficient resource management by the targeted server. Additionally, an intensified attack configuration reveals notable changes in query times and increased CPU utilization on the DNS server, indicating its struggle to handle the intensified query traffic.

Keywords— *Graph theory • Features Engineering • Community detection • Null models • Random forest • MLP*

Contents

1 Introduction

In the dynamic landscape of healthcare, understanding the intricate interplay between symptoms and diseases is paramount for effective diagnosis and prediction. This report embarks on a comprehensive journey through the realms of network analysis, leveraging both theoretical foundations and empirical data to unravel the complexities of symptom-disease interactions. Our dual-fold objective is to provide a nuanced descriptive analysis of these interactions while identifying key features to bolster predictive models.

The foundation of this endeavor lies in an extensive review of existing literature, drawing insights from seminal works on network theory and disease prediction. By establishing a baseline through prior research, we pave the way for a deeper understanding of the subject matter and ensure the relevance of our findings in the broader context of scientific inquiry.

Guided by insights gleaned from the literature, our exploration extends to the realm of data, where we meticulously curate and analyze datasets of varying sizes. Through a systematic process of exploratory data analysis and cleaning, we prepare the groundwork for constructing meaningful networks that encapsulate the relationships between symptoms and diseases.

The heart of our analysis lies in the creation of intricate network structures, employing bipartite models and non-weighted links to distill meaningful patterns. We delve into a spectrum of network metrics, from fundamental measures like degree distribution and clustering coefficients to more nuanced assessments of node importance and betweenness centrality. Statistical significance is rigorously assessed through the lens of a null model, ensuring that our observations transcend mere chance.

Community detection algorithms further dissect the network, revealing hidden structures and relationships between diseases. This not only enriches our understanding but also lays the groundwork for subsequent analyses. As we traverse the terrain of network analysis, we introduce novel metrics inspired by the Hidalgo-Hausmann framework, stratifying symptoms and diseases based on their predictive importance. These metrics, coupled with traditional measures like betweenness centrality, contribute to the definition of features that fuel our predictive models.

With a robust foundation established, we transition to the realm of predictive modeling, where our feature-rich approach promises to enhance the performance of established models. Logistic regression, random forest, and multi-layer perceptron models are trained, tested, and validated, with a keen eye on feature importance and model improvement strategies.

This report unfolds as a holistic exploration, weaving together theoretical frameworks, empirical analyses, and predictive modeling into a cohesive narrative. As we traverse the intricate web of symptom-disease interactions, our aim is not only to elucidate the underlying dynamics but also to pave the way for more accurate and insightful predictive models in the realm of healthcare.

2 L1 and L2 metrics

- Meaning of Z-score, which is like p-value
- Meaning of L1 and L2 metrics, why L2 has no sense

References

Devi, G. U. 2015. "Detection of DDoS Attack using Optimized Hop Count Filtering Technique" [in en]. *Indian Journal of Science and Technology* 8, no. 1 (January): 1–6. ISSN: 09746846, 09745645. https://doi.org/10.17485/ijst/2015/v8i26/83981.

Fang, L., H. Wu, K. Qian, W. Wang, and L. Han. 2021. "A Comprehensive Analysis of DDoS attacks based on DNS." *Journal of Physics: Conference Series* 2024:012027. https://doi.org/10.1088/1742-6596/2024/1/012027.

Taylor, R. 2021. *Four major DNS attack types and how to mitigate them* [in en-US], August. Accessed May 10, 2023. https://bluecatnetworks.com/blog/four-major-dns-attack-types-and-how-to-mitigate-them/.