Medical Fraud Detection: Network Graphs on Shared Patients

CS4991 Capstone Report, 2024

Byron Xu
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
bx7ugx@virginia.edu

ABSTRACT
Medical insurance providers in America deal with tens of millions of medical claims every year, in which tens of billions of dollars are lost due to fraud. On such a scale, it is simply not possible to manually review every case to determine which are fraudulent. To streamline the medical fraud detection process, I utilized Python frameworks such as Pyspark and NetworkX in order to plot the relationships between medical providers based on their shared patients. One common fraudulent scheme is for providers to bounce patients between physicians or clinics and submitting fraudulent insurance claims. The algorithm tracked the number of specific shared patients between a suspicious provider and other non-suspicious providers over a year in order to probe for any fraudulent networks. The proof of concept focused on a small network of providers totaling roughly $100,000 in claims. On a larger scale, the project could vastly increase the efficiency of fraud review by centering efforts on the most suspicious providers, allowing currently flagged providers in order to uncover other fraudulent providers that the company is unaware of.

1. INTRODUCTION

In 2022, 92.1 percent of Americans, or 304 million had health insurance at some point in the year. The U.S. individual health insurance sector was valued to be USD 1.60 trillion in 2022, growing by billions every year (Grand View Research, 2023). In such a financially large sector, even small percentages of fraudulent activity can equate to billions of dollars, affecting millions of Americans' lives.

The pervasiveness of medical fraud takes form in its various schemes, such as upcoding, unbundling, or medical identity theft. Each fraudulent scheme requires a different review process on the medical provider. Both the enormous volume of medical claims every year and the complexity of fraud schemes make manual review only possible on a highly limited number of providers. On the other hand, algorithms will all produce some type of classification error, marking either non-fraudulent providers vs. fraudulent, or the other way around. Blacklisting a non-fraudulent medical provider can result in serious consequences, affecting countless patients. Thus, multiple machine learning models and algorithms are used to reduce the error probability and funnel a select number of highly likely to be fraudulent providers to manual review teams, who carefully review each marked provider.

2. RELATED WORKS

Fraud can never be fully stopped, only hindered. As fraud detection technology improves, fraud schemes also grow more extensive and refined. The majority of fraud detection algorithms are based on machine learning models such as SVM, hierarchical

clustering, or k-means (Baesens, et al., 2015), which focus on detecting outliers from samples.

With recent advances in network analysis, data scientists are now better equipped to detect fraud rings through graph visualizers like Network X. The same types of machine learning models for medical fraud detection are continuously being used and combined into hybrid models (Li, et al., 2008). However, better data visualization software allows these models to be more easily understood by manual review teams rather than simply predicting a label for each provider.

3. PROJECT DESIGN

The goal of this project was to highlight the connections between fraudulent providers and their hidden associates by leveraging graph analysis. By examining the ratio of shared patients that suspicious providers had with other related providers, we were able to locate previously hidden medical fraud rings.

3.1 Data Utilization

The project utilizes two core datasets, in which one contains all providers based on their risk profile and the other contains all prejudicated claims (which we will label *prejudicated_claims*). All NPIs must be labeled to be either "unknown" (providers without current suspicion), or "flagged" (indicative of previous suspicious activity).

The *prejudicated_claims* enable us know the claim volumes and amounts tied to each provider. Having claim information allows us to both normalize different-sized providers and to pursue larger fraud schemes. Specifically, only providers whose financial transactions surpassed the threshold of $10,000 in the year 2021 were chosen. The financial threshold was chosen in order to focus efforts on non-trivial fraudulent networks, and the timeframe had to be set to a single year—2021—because of the large discrepancy in provider activity and the quantity of patients before, during and after the COVID-19 pandemic of 2020.

3.2 Methodology

The analysis is a graph-based model that displays the relationships between medical providers through a network of nodes and edges using the python framework NetworkX. Each node in this network represents a single provider, differentiated by color coding: red nodes signify providers with active tips that hint at potential fraudulent behavior, while blue nodes represent providers devoid of current suspicion but with substantial claim interactions.

What differentiates anti-fraud graph-based models in healthcare from one another is the how the edges are defined. In this POC, the edge symbolized the quantity of shared patients between a flagged provider and an unknown provider. The existence of an edge requires at least one shared patient between the two providers, an indicator of potential collusion or a patient being "bounced" between the two providers. To normalize the edge, the quantity of shared patients was converted into a *"% Ratio of Shared Patients"* as a metric to gauge the intensity of the relationship between two providers, allowing for a normalized comparison across providers of varying scales, from large viral testing laboratories to specialized surgery clinics.

4. RESULTS

A real-world example of the methodology entailed a network cluster involving 24 providers, a mix of flagged and unknown entities, centered around a patient with extensive interactions with flagged providers in 2021. The thickness of the edge was based on the % of shared patients, where bolder lines indicated higher shared patient ratios.

Along each edge, only one singular patient is tracked in order to create an individual small network of providers, as larger medical fraud rings are more difficult to detect.

In this specific example, certain NPIs shared up to an abnormally high 71% of their patients with one another. These NPIs tended to be smaller facilities or clinics, primarily focusing on acupuncture, physical therapy, and other related disciplines.

5. CONCLUSION

The project shows significant potential as an alternative model for medical fraud detection. Traditional fraud detection algorithms tend to focus on individual providers committing fraud, rather than the relationships between providers. Graph visualization on the connections between providers that share patients enables manual review teams to easily see whether two providers are supposed to be in constant contact with one another or not, showcasing abnormalities in the patient referral process between providers.

This approach not only reduces the number of providers to be reviewed, but also focuses on detecting fraudulent activity that may be less evident, as fraud rings are more difficult to detect. Throughout this project, I have not only honed my skills in data science and network analysis but have also gained deeper insights into the depth and complexity of medical fraud.

6. FUTURE WORK

While the current project was successful in detecting medical fraud, the overall scope of the project was rather limited. Ideally, this project should be scaled over a longer timeframe to depict larger networks of medical providers. Furthermore, the addition of other metrics such as physical proximity or specialization similarity to the graph edges would improve the precision of the mode.

Additionally, collaboration with medical providers in the industry could refine the model to better detect various types of fraud schemes. Outside of the medical sector, similar models could be used to detect other illegal, collaborative activities, such as embezzlement or drug rings.

REFERENCES

Baesens, B., Van Vlasselaer, V., & Verbeke, W. (2015). *Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection*. John Wiley & Sons.

Grand View Research. (2023). U.S. Individual Health Insurance Market Size Report, 2030. https://www.grandviewresearch.com/industry-analysis/us-individual-health-insurance-market-report

Li, J., Huang, KY., Jin, J (2008). A survey on statistical methods for health care fraud detection. *Health Care Manage Sci* 11, 275–287 (2008). https://doi.org/10.1007/s10729-007-9045-4