Detection of Anomalous Activity in Diabetic Patients Using Graph-Based Approach

Ramesh Paudel, William Eberle, Doug Talbert
rpaudel42@students.tntech.edu, weberle@tntech.edu, dtalbert@tntech.edu
Tennessee Technological University
Cookeville, TN 38501

Abstract

Every year, billions of dollars are lost due to fraud in the U.S. health care system. Health care claims are complex as they involve multiple parties including service providers, insurance subscribers, and insurance carriers. Medicare is susceptible to fraud because of this complexity. To build a comprehensive fraud detection system, one must take into consideration all of the financial practices involved among the associated parties. This paper is focused on graph-based analysis of CMS provided Medicare claims data to look for anomalies in the relationships and transactions among patients, service providers, claims, physicians, diagnosis, and procedures. In our experiments, we create graphs from inpatient, outpatient, and carrier claims data of the beneficiary. We then demonstrate the potential effectiveness of applying graph-based anomaly detection to the problem of discovering anomalies and potential fraud scenarios.

Introduction

According to the Centers of Medicare and Medicaid Services (CMS), U.S. health care spending reached $3 trillion or $9,523 per person in 2014. The total health care spending in 2014 accounted for 17.5% of the nation's Gross Domestic Product and is expected to rise to 20.1% by 2025 (CMS 2016b). Unfortunately, roughly one-third of health care spending can be attributed to fraud, waste, and abuse (Kelley 2009). Because of this significant financial loss, there is a need to build better fraud detection mechanisms.

Health care claims are complex because they involve multiple parties including service providers (i.e., doctors, hospitals, ambulance companies and laboratories), insurance subscribers (i.e., patients and employers) and insurance carriers who receive regular premiums from their subscribers and pay health care costs on behalf of their subscribers, including governmental health departments and private insurance companies. To build a comprehensive fraud detection system, one must take into consideration all of the financial practices involved. There are basically two types of health-care fraud operations (Sparrow 1996):

Hit and Run: Fly-by-night operators who steal millions in a relatively short period, then vanish.

Steal a little all the time: Perpetrators who work to ensure fraud goes unnoticed and bill fraudulently over a long period of time. The provider may hide false claims within large batches of valid claims and, when caught, will claim it an error, repay the money, and continue the behavior.

There are various types of health-care fraud schemes, briefly described as follows (FBI 2009) (Li et al. 2008):

Identity Theft: Stealing identification information from providers or beneficiaries and using that information to submit fraudulent bills to Medicare.

Phantom Billing: Billing for services that are not actually performed.

Unbundling: Billing each stage of a procedure as if it were a separate treatment.

Upcoding: Billing costlier services than the performed.

Bill Padding: Providing medically excessive or unnecessary services to a patient.

Duplicate Billing: Submitting same claims multiple times

Kickbacks: A negotiated bribery in which a commission is paid to the bribe-taker (provider or patient) as a quid pro quo for services rendered (Albrecht et al. 2012).

Doctor shopping: Patient consults many physicians in order to obtain multiple prescriptions of drugs in excess of their own therapeutic need (He, Graco, and Yao 1998).

In this paper, we introduce an approach for discovering health care fraud using *graph-based data mining*. If we consider the *entities* involved in the process of medical claims as *nodes*, and the *relationships* and *transactions* between the entities involved as *edges*, we can represent the entire process as a *graph*. Using a known graph-based anomaly detection approach, we will show how anomalies that are potentially fraudulent can be discovered in data representing health care transactions. To empirically validate our proposed approach, we will apply the publicly available GBAD tool (Eberle and Holder 2007) on CMS provided Medicare data that has been made publicly available through the (CMS 2016a) web-site. Medicare data provides relationships among beneficiaries, their inpatient/outpatient care, carrier and drug claims, physicians and institutions they visit, procedures physicians perform on patients, and diagnoses they uncover. The GBAD system has been successfully applied to a wide variety of domains such as insider

threat detection, mobile telecommunications anomalies, and the discovery of illegal cargo shipments, but never to health care fraud. For this work, we will specifically target the treatment of diabetic patients in the state of Tennessee who were enrolled in Medicare in 2009 as a starting point for demonstrating the application of graph-based anomaly detection to the problem of health care fraud.

The next section of this paper presents existing research on detecting health care fraud. We will then follow that with a description of the dataset we will use in our experiments. After which, we will briefly discusses the GBAD approach, followed by a discussion of how we generated the graphs from the data. We will then conclude with our experiments, results, conclusions, and future directions for this work.

Related Work

Most of the research in health care fraud is focused on statistical analysis and the use of machine learning algorithms like clustering, k-nearest neighbor, decision trees, neural networks, etc. But, compared to the extent of financial loss in the health care sector, the research to date has been minimal.

(Ortega, Figueroa, and Ruz 2006) propose a supervised fraud detection system for the Chilean health care system which uses a committee of multilayer perceptron neural networks (MLP) for each one of the entities involved in the fraud/abuse problem: medical claims, affiliates, medical professionals and employers. Their application decreases the time it takes to detect fraud by 76% (from an average of 8.6 months to 2 months) than without the system.

(Williams and Huang 1997) use decision trees for detecting insurance subscribers' fraud for the Health Insurance Commission (HIC) of Australia. First, a clustering algorithm is built to divide all insurance subscribers' profiles into groups. Second, a decision tree is made to build a set of rules. Finally, each rule is evaluated by establishing a mapping from the rule to a measure of its significance. In the end, extremes are marked for further investigation.

(Yang and Hwang 2006) apply process-mining techniques to gather clinical-instance data to construct a model that identifies service provider fraud for the NHI in Taiwan. This approach eliminates the need to manually analyze and encode behavior patterns, as well as the guesswork in selecting statistics measures. It identifies some fraudulent cases not detected by a manually constructed detection model. However, building detection models that can be easily adjustable according to site-specific cost policies is challenging.

(He, Graco, and Yao 1998) propose the use of a k-Nearest Neighbor (kNN) algorithm with an optimized non-Euclidean distance metric using a genetic algorithm. Their study concluded differences in either the decision rule or the number of nearest neighbors had little or no impact, while optimizing the distance metric improved the classification accuracy of the kNN algorithm. However, their approach is focused only on detecting two types of fraud schemes: inappropriate practice of service providers and doctor-shoppers.

(Thornton et al. 2013) look at the data beyond the transaction level and build upon (Sparrow 1996) fraud type classifications and the Medicaid environment, to develop a Medicaid multidimensional schema and provide a set of multidimensional data models and analysis techniques that help to predict the likelihood of fraud. These data views address the most prevalent known fraud types and prove useful in discovering the unknown unknowns. The model is evaluated by functionally testing against known fraud cases.

Most of the above approaches need expert knowledge to design a set of rules, and the anomaly is detected by observing the deviation from such rules. The performance of these approaches is limited by the availability of domain experts. Furthermore, these techniques are not targeting the *relational* aspect of the involved entities - something appropriate for a graph-based approach.

(Liu et al. 2015) propose a graph-analysis technique called Xerox Program Integrity Validator (XPIV) to find fraud in health care by using an ego-net approach to examine narcotics relationships and temporal-spatial characteristics of patients migrating between pharmacies and providers; as well as the global structure of the health care relationship network to look for communities sharing a common abnormal practice. In preliminary work, they are able to identify millions of dollars lost in fraud for potential recovery. Though they use a graph-based approach, their work is focused on detecting anomalies in narcotic relationships, while our proposed approach targets broader anomaly detection.

Health Care Data

The dataset used for this research is the CMS Linkable 2008–2010 Medicare Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF) dataset (CMS 2016a). The data contains synthesized data taken from a five percent random sample of Medicare beneficiaries and their claims from 2008 to 2010. The data are fully "synthetic," meaning no beneficiary in the DE-SynPUF is an actual Medicare beneficiary, but they are all meant to represent actual beneficiaries. Out of the 20 random sample files made available by the CMS, we will use sub sample 1 for the following experiments. It should be noted that there is nothing that limits us to this particular sample or the use of multiple samples. It was an arbitrary choice for validating our proposed approach, and we will be expanding our dataset in the future with more samples. The database schema, showing all the tables and available attributes, is shown in Figure 1.

The database consists of five tables: one for the beneficiary summary and one for each claim type, i.e., inpatient, outpatient, carrier, and prescription drug event. Each of the five tables has a primary key DESYNPUF_ID which uniquely identifies the beneficiary. In addition to DESYNPUF_ID, claim tables have CLM_ID to differentiate between different claims for the same beneficiary. The beneficiary summary table has beneficiary demographic information (like sex, race, birth date, state, county, etc.), medical information (like preexisting medical condition), and financial information (like coverage for each of the Medicare coverage types). Each of the claim types are linked to the beneficiary via DESYNPUF_ID. The inpatient, outpatient, and carrier claim tables have information about the institution visited, physician involved, disease diagnosed, procedure performed, cost and insurance coverage. The pre-

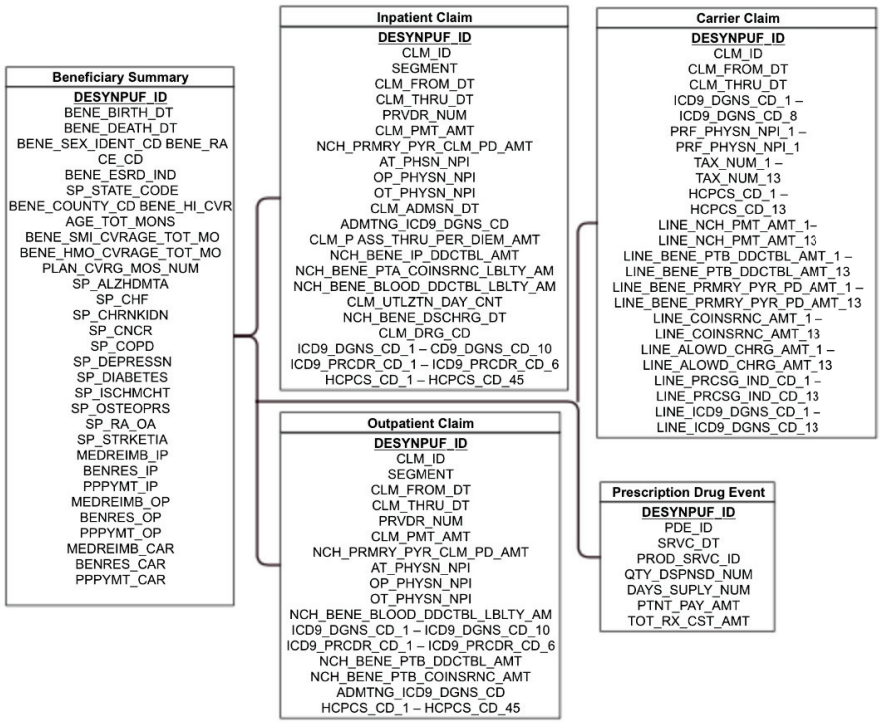

Figure 1: Database schema containing all the table and their attributes of DE-SynPUF subsample 1

scription drug event table has information about the amount of drug prescribed and the cost associated with it. Table 1 shows the number of data instances in each table. More detailed information about the data can be found at (CMS 2016a)

SN	Data Type	Number of rows
1	Beneficiaries	343,644
2	Carrier Claims	4,741,335
3	Inpatient Claims	66,773
4	Outpatient Claims	790,790
5	Prescription Drug Events	5,552,421

Table 1: Number of entries in each DE-SynPUF table.

Graph-Based Anomaly Detection

In order to lay the foundation for this effort, we hypothesize that a real-world, meaningful definition of a graph-based anomaly is an unexpected deviation to a normative pattern. The importance of this definition (which we more formally define below) lies in its relationship to any deceptive practices that are intended to illegally obtain or hide information (Eberle and Holder 2007).

Definition IV.A. A labeled graph $G = (V,E,F)$, where V is the set of vertices (or nodes), E is the set of edges (or links) between the vertices, and the function F assigns a label to each of the elements in V and E.

Definition IV.B. A subgraph SA is anomalous in graph G if $(0 < d(SA,S) < TD)$ and $(P(SA|S) < TP)$, where $P(SA|S)$ is the probability of an anomalous subgraph SA given the normative pattern S in G. TD bounds the maximum distance (d) an anomaly SA can be from the normative pattern S, and TP bounds the maximum probability of SA.

Definition IV.C. The score of an anomalous subgraph SA based on the normative subgraph S in graph G is $d(SA,S) * P(SA|S)$, where the smaller the score, the more anomalous the subgraph.

The advantage of graph-based anomaly detection is that the relationships between entities can be analyzed for structural oddities in what could be a rich set of information, as opposed to just the entities' attributes. However, graph-based approaches have been prohibitive due to computational constraints, because graph-based approaches typically perform subgraph isomorphisms, a known NP-complete problem. Yet, in order to use graph-based anomaly detection techniques in a real-world environment, we need to take advantage of the structural/relational aspects found in dynamic, streaming data sets.

In order to test our approach, we will use the publicly-available GBAD test suite , as defined by (Eberle and Holder 2007). Using a greedy beam search and a minimum description length (MDL) heuristic, GBAD first discovers the "best" subgraph, or normative pattern, in an input graph. The MDL approach is used to determine the best subgraph(s) as

the one that minimizes the following:

$$M(S,G) = DL(G|S) + DL(S),$$

where G is the entire graph, S is the subgraph, DL($G|S$) is the description length of G after compressing it using S, and DL(S) is the description length of the subgraph. The complexity of finding the normative subgraph is constrained to be polynomial by employing a bounded search when comparing two graphs. Previous results have shown that a quadratic bound is sufficient to accurately compare graphs in a variety of domains (Eberle and Holder 2007).

For more details regarding the GBAD algorithms, the reader can refer to (Eberle and Holder 2007). In summary, the key to the GBAD approach is that anomalies are discovered based upon small deviations from the norm (e.g., insider threat, identity theft, etc.) – not outliers, which are based upon significant statistical deviations from the norm.

Dataset Generation

The DE-SynPUF dataset consists of Medicare data for 3 years, from 2008 to 2010. Since our view is limited to these three years, we want to make sure that the records we examine deal with patients at the same stage of their medical process. Thus, we will choose 2009 beneficiaries and their claims, as we can determine whether or not they were treated in 2008 and whether or not they were treated subsequently in 2010. In addition, while our future work will address the issue of big data as it relates to overall fraud detection in the health care industry, we will limit this initial work to only a subset of beneficiaries as a proof-of-concept. In this case, beneficiaries from Tennessee and their inpatient, outpatient, carrier and prescription drug claims, when they have an initial diagnosis of diabetes. The graph input file is built from the dataset to reflects the relationship between beneficiaries, their claims, physicians involved, service provider institute, procedure performed, etc. Each beneficiary might have multiple inpatient, outpatient, carrier or prescription drug claims. The edge between a patient and a claim indicates that the patient filed, or was related to, the corresponding claim. It should also be noted that if a beneficiary has more than one claim, prescription, physician, etc., then multiple claim, prescription, physician, etc., nodes are created for each unique value, resulting in potentially multiple edges between the patient and these entities.

Experimental Results and Analysis

Our experimental setup consists of parsing the required data from the DE-SynPUF dataset, constructing a single graph that contains the data for each beneficiary from all the claim tables, and processing the resulting graph with a graph-based anomaly detection tool. In order to create the graph input file, we will create a parser (written in the python programming language) that will read the CMS data and build the graph.

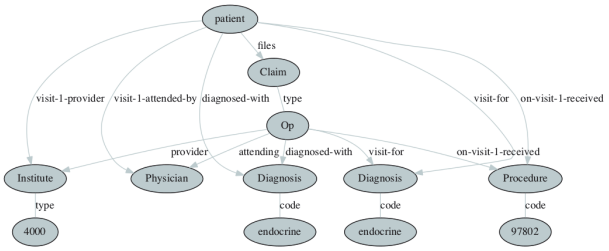

Figure 2: Visual Example of Outpatient Claim Graph

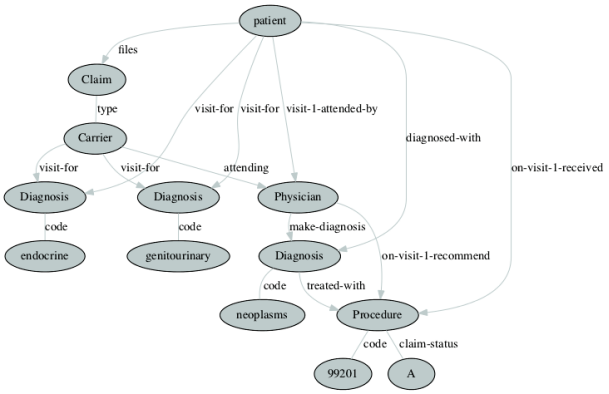

Figure 3: Visual Example of Carrier Claim Graph

Graph Input File

The structure of inpatient and outpatient claims are similar, and since the purpose of using a graph-based anomaly detection approach is to discover unusual structure, we combined the data from these two claim types into a single graph, which we will refer to as the ip-op claim graph. However, carrier claim data has a very different structure, so we will create a separate graph input file for that data, which we will refer to as the carrier claim graph. Figure 2 shows a visual representation of a portion of ip-op claim graph, and Figure 3 shows a portion of carrier claim graph. It should be noted that these are just visualizations, as the actual graph input files are just ASCII text files.

We limited our anomaly detection to only patients in Tennessee who have been diagnosed with diabetes. The choice of population and disease was arbitrary and was done to ensure that we are examining people with similar demographics and characteristics. In future work, we will expand to other populations and diseases.

Of the chosen population, we find that 62 beneficiaries diagnosed with diabetes have filed inpatient and outpatient claims, thus resulting in an ip-op claim graph with 62 examples, or subgraphs, for a total of 1,469 vertices and 2,139 edges. Similarly, there are 572 beneficiaries diagnosed with diabetes that have filed carrier claims, resulting in 572 examples, each representing a diabetic beneficiary, for a total of 21,082 vertices and 32,214 edges. From Figure 3 and Figure 2, one can see that each beneficiary is represented by a “patient” node, where a patient "files" a claim. Each of the claims is represented by a “claim” node, with an edge

linking to the one of the types of claims, i.e., “ip”, “op”, and “carrier”. Each claim can have an admitting diagnosis represented by a “visit-for” edge, and a final diagnosis represented by a “diagnosed-with” edge linking to that specific “Diagnosis”, which has a ICD-9 diagnosis code. In each claim, the patient is “attended-by” a physician and “receive”s a procedure. In a carrier claim, there is a clear relationship between the physician, what diagnosis was made, and what procedure was performed to treat that diagnosis. Thus, in the carrier claim graph, shown in Figure 3, this relationship is represented as "Physician" - "make-diagnosis" - "Diagnosis", and it is "treated-with" - "Procedure". Each procedure has an ICD-10 procedure code to uniquely identify the procedure.

Anomaly Detection

Running GBAD on the carrier claim graph, Figure 4(a), on the left, shows the discovered normative pattern. Similarly, for the ip-op claim graph, we get the normative pattern shown in Figure 6. GBAD does not discover any interesting anomalies in the ip-op claim graph, perhaps because the data set is small (only 62 instances). However, GBAD does discover anomalous substructures in the carrier claim graph which we will now discuss in detail below.

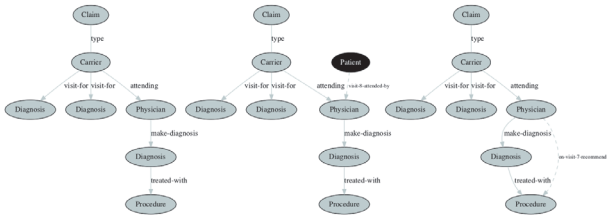

Figure 4: Carrier Claim Graph a) Normative Structure b) Anomalous Patient Visit c) Anomalous Procedure

It should also be noted that GBAD takes 459 seconds to analyze the carrier claim graph, and, because of its size, only 15 seconds to process the ip-op claim graph.

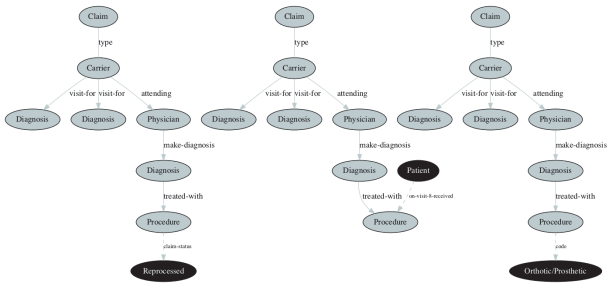

Figure 5: Carrier Claim Graph a) Anomalous Claim Status b) Anomalous Multiple Procedure c) Anomalous Procedure

Unexpected Edges and Vertices. Using a probabilistic approach (one of several algorithms available in GBAD) with the amount of change (TD) set to 2 and probability (TP) set to 1, we discover various anomalies, as shown by the examples in Figure 4 and Figure 5. The anomalies in each of the figures are depicted using a black vertex to represent the anomalous existence of a vertex and a dashed line to represent the anomalous existence of an edge. Further inspection of the data confirms that the middle substructure in Figure 4 (b), is anomalous because it contains an unusually high number of visits. There are only two beneficiaries who have more visits than 8 out of 572 beneficiaries. The anomaly on the right in Figure 4 (c), is a case where multiple times the same procedure was used to treat the same diagnosis, resulting in multiple/duplicate billing. The anomaly on the left in Figure 5 (a), is a case where the claim status was reprocessed. Even though the filed claim was approved in this situation, which occurs in 3 examples out of 572, the claim is reprocessed. In the middle of Figure 5 (b), is another case where the patient receives the same procedure on multiple visits. Shown on the right of Figure 5 (c), is the case where a patient is treated with an "Orthotic/Prosthetic" procedure - something that only occurs once in all of the patients.

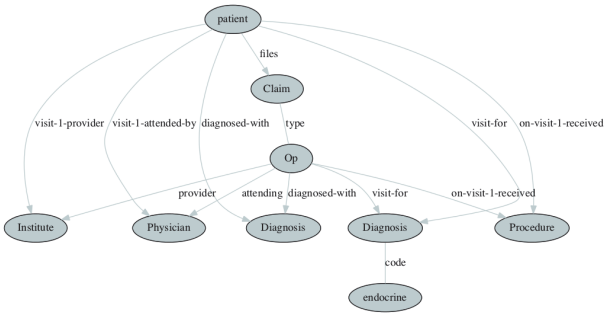

Figure 6: IP-OP Claim Graph Normative Structure

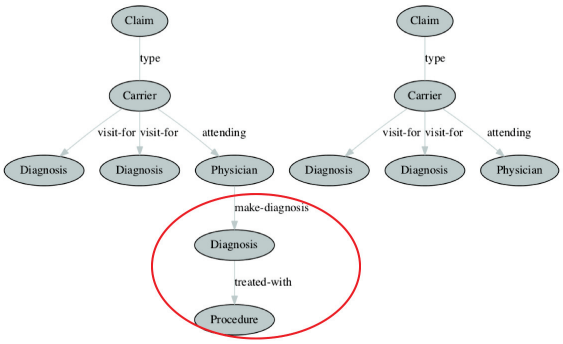

Figure 7: Carrier Claim Graph a) Normative Structure b)Anomalous Deletion of Diagnosis

Missing Edges and Vertices. Using a maximum partial substructure approach (another algorithm in GBAD) with TD set to 0.4 and TP set to 1, the anomaly is reported, as shown by the example on the carrier claim graph depicted in Figure 7. Using the normative pattern shown on the left in Figure 7 (a), we discover that a circled substructure is missing from Figure 7 (b), as shown on the right. Since the

anomaly is something that is missing from this patient, we further examine the source data to discover what is missing from this patient, i.e., what is encircled in the figure can be found in the substructure of other patients. In this example, we discover that this particular patient has visited the hospital multiple times. On the first and second visit, the anomalous patient was diagnosed with the same disease. However, on the anomalous patient's first visit, he/she is treated with one procedure and on the second visit, he/she is treated with another procedure. It might be the scenario where a physician is recommending multiple treatment so as to garner more revenue, or it could just be that the first treatment did not work, and they tried a different treatment the second time. Whether either scenario is true or not (or there is a different scenario), we cannot tell but we can say there is anomalous behavior.

This initial work shows that by representing Medicare data as a graph and using a graph-based anomaly detection approach, we can potentially detect various anomalous relationships. These anomalous instances can be of particular interest to fraud analysts, as focusing their efforts on these patients might lead to discovering health care fraud scenarios. For example, Figure 4 (b) is the case where the patient has an unusually high number of visits, which could be the *doctor shoppers* scenario discussed earlier. Also, it might be a case of *identity theft* where someone else uses their identity and files a claim on their behalf. Figure 4 (c) is the situation of recommending the same procedure multiple times, which could be a case of *duplicate billing* or *unbundling*, where each stage of a procedure is billed as if it were a separate procedure. Figure 5 (c) reports the case of an anomalous procedure recommended, which might be the scenario of *upcoding*, where the procedure was costlier than the usual procedure recommended by other physicians. In the case of Figure 7, a patient receives multiple procedures for treating the same diagnosis, a potential scenario of *phantom billing*, or perhaps even the scenario of a *kickback* where the physician and patient are involved in filing fake claims. Our initial experiments have indeed found some interesting anomalies. However, further work is needed to determine the basis of these anomalies in the realm of health care fraud.

Conclusion

In this paper, using a known graph-based anomaly detection approach, we showed how anomalies that are potentially fraudulent can be discovered in data representing health care transactions. We represented the Medicare claims data as a graph where the *entities* involved in the process of medical claims are *nodes*, and the *relationships* and *transactions* between the entities involved are *edges*. For this work, we specifically target the treatment of diabetic patients in the state of Tennessee who were enrolled in Medicare in 2009 to demonstrate the proof of concept of graph-based anomaly detection to the problem of discovering anomalies, particularly ones related to health care fraud.

In future, we will first extend this approach to the entire Medicare claim dataset. In order to discover other anomalies and address the scalability of this approach, we will investigate a graph-partitioning approach that will process multiple graphs in parallel. Then, we will involve medical practitioners who have offered their domain expertise. We also plan on including prescription drug claims, which will provide us with even more information as to potential fraudulent activities in the health care industry. Patients with certain disease may have certain phenotypic groups based on their comorbidity characteristics because they require totally different management and treatment paths. Our plan also include further investigating these phenotypic groups.

Acknowledgement

This material is based on work supported by the National Science Foundation (NSF) under Grant No. 1318957.

References

Albrecht, W.; Albrecht, C.; Albrecht, C.; and Zimbelman, M. 2012. Fraud examination south-western cengage learning. *Mason, OH*.

CMS. 2016a. Medicare claims synthetic public use files. [Online; accessed 11-August-2016].

CMS. 2016b. National health expenditure projections 2015-2025. [Online; accessed 8-November-2016].

Eberle, W., and Holder, L. 2007. Anomaly detection in data represented as graphs. *Intelligent Data Analysis* 11(6):663–689.

FBI. 2009. 2009 financial crimes report. [Online; Retrieved 21-February-2013].

He, H.; Graco, W.; and Yao, X. 1998. Application of genetic algorithm and k-nearest neighbour method in medical fraud detection. In *Asia-Pacific Conference on Simulated Evolution and Learning*, 74–81. Springer.

Kelley, R. 2009. Where can $700 billion in waste be cut annually from the us healthcare system. *Ann Arbor, MI: Thomson Reuters* 24.

Li, J.; Huang, K.-Y.; Jin, J.; and Shi, J. 2008. A survey on statistical methods for health care fraud detection. *Health care management science* 11(3):275–287.

Liu, J.; Bier, E.; Wilson, A.; Honda, T.; Sricharan, K.; Gilpin, L.; Guerra-Gomez, J.; and Davies, D. 2015. Graph analysis for detecting fraud, waste, and abuse in healthcare data. In *AAAI*, 3912–3919.

Ortega, P. A.; Figueroa, C. J.; and Ruz, G. A. 2006. A medical claim fraud/abuse detection system based on data mining: A case study in chile. *DMIN* 6:26–29.

Sparrow, M. K. 1996. *License to steal: Why fraud plagues America's health care system*. Westview Press Boulder, CO.

Thornton, D.; Mueller, R. M.; Schoutsen, P.; and van Hillegersberg, J. 2013. Predicting healthcare fraud in medicaid: a multidimensional data model and analysis techniques for fraud detection. *Procedia technology* 9:1252–1264.

Williams, G. J., and Huang, Z. 1997. Mining the knowledge mine. In *Australian Joint Conference on Artificial Intelligence*, 340–348. Springer.

Yang, W.-S., and Hwang, S.-Y. 2006. A process-mining framework for the detection of healthcare fraud and abuse. *Expert Systems with Applications* 31(1):56–68.