Health Care Fraud Detection with Community Detection Algorithms

Song Chen
Information Systems
University of Maryland Baltimore County
Email: song8@umbc.edu

Aryya Gangopadhyay
Information Systems
University of Maryland Baltimore County
Email: gangopad@umbc.edu

***Abstract*—Fraud detection is interesting research topic and it not only needs data mining techniques but also needs a lot of inputs from domain experts. In health care claims, relationships between physicians and patients form complex communities structures and these communities could lead to potential fraud discoveries. Traditionally, researchers have focused on clustering physicians and patients and tried to find the suspicious communities. In this paper, we studied and discussed different types of relationships and focus on small but exclusive relationships that are suspicious and may indicate potential health care frauds. We developed two algorithms to detect these small and exclusive communities. These algorithms can be applied to larger dataset and are highly scalable. We tested these algorithms with a set of synthesized datasets. These synthesized datasets were created to resemble the real health care claims datasets and used to test the fraud detection algorithms. The test results show the these algorithms are very efficient and can evaluate the communities structures of 50,000 providers in about 1 minute.**

I. Introduction

Frauds exist wherever when it involves money transactions. Health care is especially a tempting target for thieves. In the United States, total health spending in America is a massive $2.7 trillion, or 17% of GDP. No one knows for sure how much of that is embezzled, but in 2012 Donald Berwick, a former head of the Centres for Medicare and Medicaid Services (CMS), and Andrew Hackbarth of the RAND Corporation, estimated that fraud (and the extra rules and inspections required to fight it) added as much as $98 billion, or roughly 10%, to annual Medicare and Medicaid spending, and up to $272 billion across the entire health system [1].

There are different types of frauds in health care systems, such as drug abuses, counterfeit drugs, off-label marketing issues. In this paper, we will focus on the health insurance claims. When health services are provided, a set of claims is submitted to one or more insurers for reimbursements. Health insurance is like any other types of insurance that there is a claim processing system to adjudicate these claims to determine if a claim should be paid or by how much a claim should be paid.

To prevent the possible frauds, there are multiple levels of edits within the claim processing systems. Some edits are implemented to prevent the incorrect payments and are part of pre-payment system adjudication. Some edits are implemented after the payments have been made to the health providers and they are the post-payment edits.

This paper will discuss the claims data that are processed and paid to the providers. Using post-payment claims data, we can perform many types of data analytics and data mining techniques to discover potential frauds. There are many types of insurance frauds, the following is a list of 10 types of frauds in health insurance that are most commonly mentioned:

1) Billing for services not rendered.
2) Billing for a non-covered service as a covered service.
3) Misrepresenting dates of service.
4) Misrepresenting locations of service.
5) Misrepresenting providers of service.
6) Waiving of deductibles and/or co-payments.
7) Incorrect reporting of diagnoses or procedures (including unbundling).
8) Overutilization of services.
9) Corruption (kickbacks and briberies).
10) False or unnecessary issuance of prescription drugs.

In this paper, we developed algorithms that target at one type of frauds. That is the suspicious provider communities that either share patients between or refer patients to each other. These communities are usually small and have exclusive relationships within the communities and no outside connections. The relationships between these communities are suspicious, however we couldn't be 100% confident that these communities are conducting fraudulent activities. There are other factors we need to consider, for example, incomplete data etc. These communities can be put on a watch list for further investigations and review. The additional review will help prevent incorrect payments to go out to these groups of providers or patients.

II. Related Work

Community Detection is a well studied area. There are so many research that have been done in Community Detection algorithm development. In [2], it gives a real good comprehensive overview of community detection algorithms. One of the most popular algorithms is to use *Modularity* as the objective function to optimize the cluster assignments until it reaches to an optimal structure [3]. These algorithms are effective to parse the whole physician network into smaller communities based on their similarities. The selection of similarities is another research topic. Some algorithms use the distance between nodes as the similarity measurements. Some similarities are

978-1-5090-0898-8/16/$31.00 ©2016 IEEE

based on the existence of connections. In this paper, we select to use the connection-based similarities rather than the distance-based similarities. In another research we conducted, we developed an algorithm based on spectral analysis and it can parse a network into similar communities using *Fielder Vector* [4].

One of the shortcomings of these algorithms is that they can only be used on smaller networks. Once the network grows to contain more nodes, for example, if it exceeds more than 1 million nodes, these algorithms will fail to procedure results. In this paper, we developed an algorithm that can be applied to larger networks to detect the suspicious communities.

Besides our original work in [4], we haven't seen much efforts to use the community detection algorithms in health care fraud detection. However, there are other similar efforts. For example, some clustering techniques have been occasionally applied in fraud detection. Those clustering techniques try to group similar providers based on the similarities of their services. They are not able to detect the communities of providers based on connections between providers. Another limit of traditional clustering algorithms is that it is hard to derive the real world meanings of the clusters. Mathematically, providers could be placed in the same or different groups in a cluster analysis, but from fraud detection perspective, it is hard to determine if any groups of providers have more fraud potentials than other groups.

There are other techniques being used in fraud detections, either supervised or unsupervised. Most of them are direct implementation of these techniques without tailoring them to the health care environment. It makes it hard to derive the meanings of these results.

The rest of this paper will discuss and evaluate an new algorithm that targets at discovering smaller communities with abnormal behaviors. These behaviors will indicate potential collusions between physicians or identity theft issues. At the end of this paper, we will discuss the sources of false positives and other applications of this algorithm.

III. Types of Communities

I am defining three types of community structures that exist within a health care dataset. As we discussed in [5] , it introduced one type of community connected by referral relationships between primary care physicians and specialty physicians. There are two other connections to establish a community structure between physicians or between physician and patients.

The two types of relationships between physicians are:

1) If two physicians treated the same patients.
2) If one physician refers a patient to a second physician.

Physicians have connections with each other through patients they served. When a physician provides any services to a patient, this physician establishes a connection this patient. This connection is reflected on the health care claim and submitted into claims database for payment processing and analytics. There are other types of relationships between physicians that we will not discuss in this paper. Examples of other relationships include similarities of services provided, geographic distances etc.

IV. Communities Between Physicians

We would like to focus on the patients are shared between physicians, or the number of patients that are treated by two or more physicians. In order to detect the relationships between physicians, we will need to first build a matrix that can reflect the relationships. The matrix values are the number of patients shared between any pairs of physicians.

In Figure 1, it shows a count relationship example between four physicians, *A,B,C and D*. *X* represents all other physicians. If we want to normalize this matrix, it can be converted to a percentage matrix as shown in Figure 2. This percentage matrix is similar to what we discussed in previous paper [5]. There are advantages and disadvantages of both metric. A count matrix represents the connections better where a larger number indicates a stronger connection. However, in count matrix, large numbers dominate the smaller numbers and it will skew the results from the algorithms. A normalized matrix is good at represent the real connections, however, the percentages won't show the actual number of connections. A 100% relationship could be normalized from either 1 connection or from 1,000 connections. This won't help us find the most important connections.

	A	B	C	D	X
A	100	20	30	30	20
B	20	200	30	50	100
C	100	50	300	150	0
D	150	60	100	400	90

Fig. 1. A Sample Relationships (Count) Between Four Physicians

	A	B	C	D	X
A	100%	20%	30%	30%	20%
B	10%	100%	15%	25%	50%
C	33%	17%	100%	50%	0%
D	38%	15%	25%	100%	23%

Fig. 2. A Sample Relationships (Percent) Between Four Physicians

In order to calculate this relationship matrix, we may use some existing software packages. However, some of these packages are not very efficient with calculating large datasets.

When we calculate the large datasets, the best way is to divide a large dataset into smaller segments, then calculate the results and aggregate the small segments into one matrix. The following Algorithm 1 gives further detailed steps of how this count and percent matrix are created. It first separates the large matrix into smaller matrices to calculate the relationships. The final results are aggregated into one matrix. These two matrices represent relationships and are the first step for our algorithms to detect suspicious communities.

Algorithm 1 The Algorithm To Calculate A Count and Percentage Matrix

Require: Two inputs

a. A *health care claims dataset*. It is a transaction database with each record is a transaction or a claim.

b. *Size of Each Batch (Size)*, This gives the size of each batch that needs to be evaluated.

Ensure:

1: Separate all providers into a list of batches according to *Size* and read them into a macro variable, $Batch\_List$
2: **for all** Batch(B) in $Batch\_List$ **do**
3: **for all** $Providers$ in B **do**
4: calculated pairwise count between providers in each batch
5: **end for**
6: Save the batch results
7: **end for**
8: Aggregate batch results into one table (Count Matrix)
9: **for all** Providers in $Variable\_List$ **do**
10: calculated percentages of each pair of providers
11: Save the batch results into one table (Percentage Matrix)
12: **end for**
13: **return** *A Count Matrix and A Percentage Matrix*

After we have established the relationships in a matrix. We would like to find the communities that are most suspicious in terms of potential fraudulent behaviors. As we have discussed earlier, the exclusive relationships within smaller communities are most suspicious. A simplest such community is a pair of two physicians who share a group of patients exclusively. These patients don't have any other connections with outside physicians and the two physicians don't see any other patients. As shown in Figure 3, this is a simplest two physician community. There could multiple potential frauds that relate to this community. Here just list a few:

1) Identify Theft Issue. When these groups of patients are fake patients and their IDs are for billing purpose only. These physicians are potentially fake as well.
2) Patient Recruiting Issue. These patients are real people, but they are paid to visit these physicians without actual medical needs.
3) Kickback Issue. These patients are referred by one physician to another when the referrers will financial incentives.

To expand the two-physician community, a slightly more complicated community is a group of three physicians when

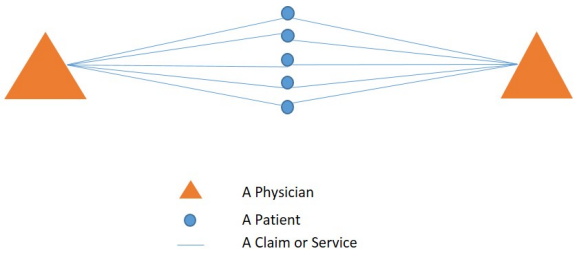

Fig. 3. Two physicians forming an exclusive relationship by sharing a group of patients who don't have claims with any other providers

they share patients exclusively. Similarly, these patients never go outside of this physicians group. In Figure 4, it displays one example of a three physician network. Some patients are shared between two physicians and one patient is shared between all three physicians.

there are a few methods to find the communities of three physicians. One method is to first find a two-physician community and then expands it to find the common physicians, if any, of the two-physician pair. If there is one common physician, then we find a group of three physicians with exclusive relationships.

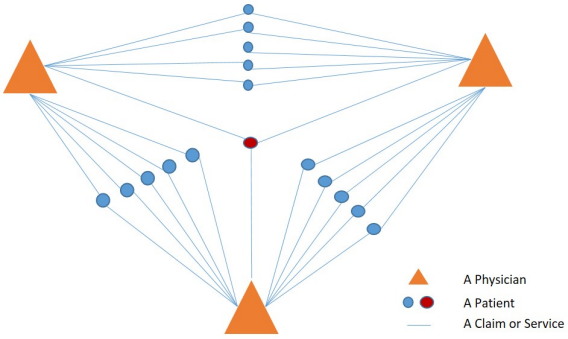

Fig. 4. Three physicians forming an exclusive relationship by sharing a group of patients who don't have claims with any other providers

However, how are we going to find more complicated communities when there are four or physicians in the community with exclusive relationships?

Next we will discuss an algorithm that can find a community of *N* physicians with probability of *P* of its network's connections are within the community.

The following Algorithm 2 illustrates the steps on how this algorithm is executed. This algorithm tries to first find physicians (List 2) that are connected to each physician on the list (List 1). It will evaluate if all the connected physicians (List 3) are all part of List 2. We want to make sure that the probability of List 3 are part of List 2 is greater or equal to *P*. This algorithm can be written as iterative operations, which will be real slow. It can also be implemented with some database operations to find communities of *N* physicians. The second method is very efficient to find the communities of more than three physicians in the health claim database. The results are discussed later in this paper.

To better explain the process of this Algorithm 2, a visual illustration is provided in Figure 5. The key idea behind this algorithm is to find the percentage of physicians on *Physician List 3 (C)* that are also in *Provider List 2 (B)* for every physician on *Physician list 1 (A)*. If there is one physician (A_n) that 100% of its connections(B)'s connections (C) are part of its connections (B), then we know these providers (A_n and B) form an exclusive community.

Since *Provider List 1 (A)* is the list of all physicians we will evaluate, the physicians on *Provider List 2 (B)* and *Provider List 3 (C)* are both part of *Provider List 1 (A)*. This is the reason that all the physicians use the same symbols on Figure 5.

In real health claims dataset, we might not always find the communities of 100% of its connections are within the network. We may set a threshold, for example 90% of its connections are within the network. This will give us more flexibility to find suspicious communities.

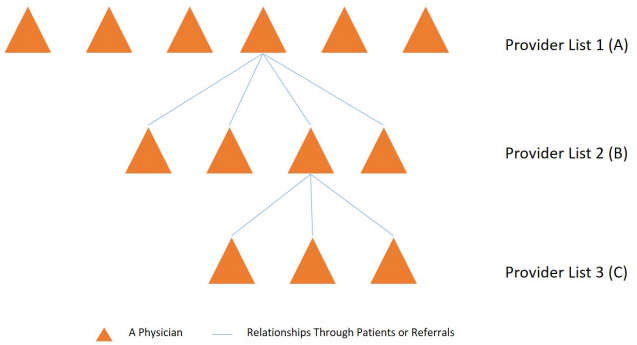

Fig. 5. Three physicians forming an exclusive relationship by sharing a group of patients who don't have claims with any other providers

One advantage of this algorithm is that it can assign a probability for each physician to measure the likelihood that its connections form an exclusive network. It is more practical to find communities that have shared a high percentages of connections within its own network. In the following section, we will talk about its implementation and test results of this algorithm.

V. Results

This section will discuss the results from this algorithm that was run on the synthesized datasets and how well this algorithm performs on finding the communities that we defined. We will compare this algorithm's performance on different sizes of the synthesized health claims data.

This algorithm could be implemented in different ways. We chose to implement this algorithm in database operations, rather than to execute loops to iterate through all physician pairs, which is probably the least efficient implementation. The algorithm is written in SAS programming language and tested on a Windows 10 64-bit operation system, with a Interl Core(TM)2 Duo CPU and 4.00 GB memory on the board.

Algorithm 2 The General Algorithm To Find Communities of N Physicians

Require: Three inputs

a. A *Count Matrix* this is a count matrix that is calculated in Algorithm 1.

b. *Count of Connections for Each Physician*, This list can be obtained by a summary on original data.

b. *A Threshold* (P), This is a threshold to set a probability that the connections of a physician are all within its network.

Ensure:

Transform the matrix into a physician pair list ($Pair1$) where it lists the number of patients that are shared between each pair. It could be other values between each pair for example, payments or claims etc.
Summarize the number of connected physicians each physician has. This creates a list of physician, $List\_1$.
for all Physicians in $List\_1$, A_n **do**
  find $List\_2$, a list of physicians who are connected with A_n
  for all Physicians in $List\_2$, B_n **do**
    find $List\_3$, a list of physicians who are connected with B_n
    calculate percentage of physicians on $List\_3$, C_n, that are also in $List\_2$
  end for
  find average percentage (Avg_A_n) of all physicians on $List\_2$
  if $Avg\_A_n \geq C$ **then**
    output A_n and Avg_A_n
  end if
end for
return *A list of providers that are part of communities with N physicians and with a probability (P) of its connections are all within network.*

A. Synthesized Data

In order to test and compare the performance of this Algorithm 2, we need to find the health claims datasets. However, the ideal health claims datasets are not easy to obtain due to privacy issues. Some public use files have limited information to test this algorithm.

We chose to create a set of synthesized health claims datasets and use them to test this algorithm. The following are a few features of this synthesized dataset.

1) it generates any *N* number of physicians.
2) its patients distribution follows a *Power Law*, which is also the distribution of most health claims datasets.
3) it contains more than 5,000 procedure codes that are used in professional health claims data.
4) its payments to each of procedure code are the average amounts according to the public data.
5) it creates referral relationships between physicians.
6) it creates the Date of Services within one year period of

time.

7) its claims frequency of each patient visit to one physician follows a *Power Law*.
8) it contains a few fraudulent features for further fraud algorithm testing. Some of the fraud features include impossible service days, work on holidays, exclusive referrals and impossible code pairs etc.

Six test datasets were created for the performance tests. These test datasets have 100, 1K, 5K, 10K,20K and 50K physicians. The largest dataset of 50K physicians has about half an million claims. The algorithm's performance is tested on the communities of size 2,3,5,10, 20 and 100 of physicians.

B. Running Time Comparison

These tests were running using the Algorithm 2. Here is the comparison of the algorithm performance and their running time in seconds in Figure 6.

Real Time in Seconds		Community Size					
Dataset Names	Number of Claims	2	3	5	10	20	100
100	2,074	0.58	0.61	0.65	0.70	1.14	0.82
1K	9,649	0.64	0.67	0.68	0.62	0.59	0.89
5K	43,217	0.98	0.89	1.00	1.06	0.90	1.18
10K	86,364	1.74	1.78	2.38	1.86	2.73	1.52
20K	172,659	4.41	4.80	6.23	8.47	5.73	4.53
50K	427,538	24.39	23.08	28.84	1'59''	2'36''	1'10''

Fig. 6. Running Time (in seconds) of different claims datasets and community sizes to look for

From this results information, this algorithm is implemented really efficiently. It can evaluate 50,000 physicians in about 1 minute to find the communities of 100 physicians. It is even quicker if we were to find the smaller size of communities, which are probably more suspicious in terms of fraud potentials.

In general, the larger communities we were trying to evaluate, the longer it may take. However, we also noticed that the running time is reduced for some datasets when the community size increases. Theoretically, there should be more calculations and database operations when we want to find larger communities. The reason is due to the fact that when the community size we are looking for becomes bigger, there are less communities found in the claim datasets and thus reduced the running time in the following process. There might not any be any communities we can find. There are more communities with smaller number of physicians and less communities with larger number of physicians.

C. Community Sizes Comparison

In Figure 7, it summarizes number of physicians that are identified in each of these tests. We can see that the communities of 100 providers can only be found in the larger datasets of 20K and 50K physicians. it is reasonable that in the smaller dataset, it is harder to form a large and exclusive communities.

The probability of N have not been applied yet to filter to the set of physicians that are most interested to us. Once we have a desired probability, for example, we can filter these providers to exclusive communities or highly exclusive communities.

Number of Physicians		Community Size					
Dataset Names	Number of Claims	2	3	5	10	20	100
100	2,074	24	1	0	0	0	0
1K	9,649	197	107	27	2	0	0
5K	43,217	1,010	549	249	127	23	0
10K	86,364	1,519	1,219	619	192	132	0
20K	172,659	1,600	1,581	1,485	487	211	7
50K	427,538	1,579	1,565	1,697	1,648	734	92

Fig. 7. Number of Phyisicians idenfied in different claims datasets and community sizes to look for

VI. Discussions

As we discussed previously, there are two ways to connect two physicians in a claim dataset.

1) Physicians treating the same patients
2) physicians referring to other physicians

In this community detection algorithm, I only examined the physicians relationships when they treated the same patients. When two physicians treated the same patients, they are considered having a connection. For the referral relationship when one physician refers patients to another physician, we can apply this algorithm similarly by starting building the relationship matrices discussed in Algorithm 1.

Community Detection is an interesting topic in fraud detection. We tested this algorithm with health care data, but it can also be applied to other insurance data for fraud detections.

Our algorithm solved the big data issue by only looking for the suspicious communities, which are those communities with exclusive relationships and contain fewer physicians. By introducing a probability of P, this algorithms becomes more general and can detect more types of communities.

We did not test for accuracies of the test results, because we target at 100% detection rate in this algorithm. We will find all the communities that exhibit exclusive relationships.

When we don't know what the size of the fraudulent communities, we will need to run this algorithm for all possible sizes. Usually you may want to start with the small size of communities because most of the collusion frauds involve less physicians. For example, we can first examine the communities of size 1 through 10 to see if any of the small communities exist before checking larger ones.

References

[1] "The $272 billion swindle." [Online]. Available: http://www.economist.com/news/united-states/21603078-why-thieves-love-americas-health-care-system-272-billion-swindle

[2] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3, pp. 75–174, 2010.

[3] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[4] S. Chen and A. Gangopadhyay, "A novel approach to uncover health care frauds through spectral analysis," in *Healthcare Informatics (ICHI), 2013 IEEE International Conference on*. IEEE, 2013, pp. 499–504.

[5] A. Gangopadhyay, S. Chen, and Y. Yesha, "Detecting healthcare fraud through patient sharing schemes," in *Information Systems, Technology and Management*. Springer, 2012, pp. 421–426.