Spring 4-11-2011

Towards A Complex Adaptive Link Prediction In Health Insurance Fraud

Szabolcs Feczak
The University of Sydney, Australia, szabolcs.feczak@sydney.edu.au

Liaquat Hossain
The University of Sydney, Australia, liaquat.hossain@sydney.edu.au

TOWARDS A COMPLEX ADAPTIVE LINK PREDICTION IN HEALTH INSURANCE FRAUD

Szabolcs Feczak and Liaquat Hossain
Project Management Graduate Programme, The University of Sydney, Australia
Email: {szabolcs.feczak|liaquat.hossain}@sydney.edu.au

Abstract

In this paper, we provide a comprehensive review of methodology for detecting anomalies based on the literature, industry expert and our findings. We elicit links between health service providers based on their shared customer base. Looking at the connections of providers, their geographical arrangements and statistical comparison of profile attributes, we highlight links which could be classified as suspicious or fraudulent behaviour.

Keywords: social networks, link prediction, graphs, anomaly detection, insurance, fraud

1.0 Introduction

The aim of fraud prevention is to maximise crime reduction without too much overhead on the core business activity (Cahill et al., 2004 & Sun, 2004). In Figure 1, we present a conceptual overview highlighting different boundaries of anti-fraud operation in insurance claims processing.

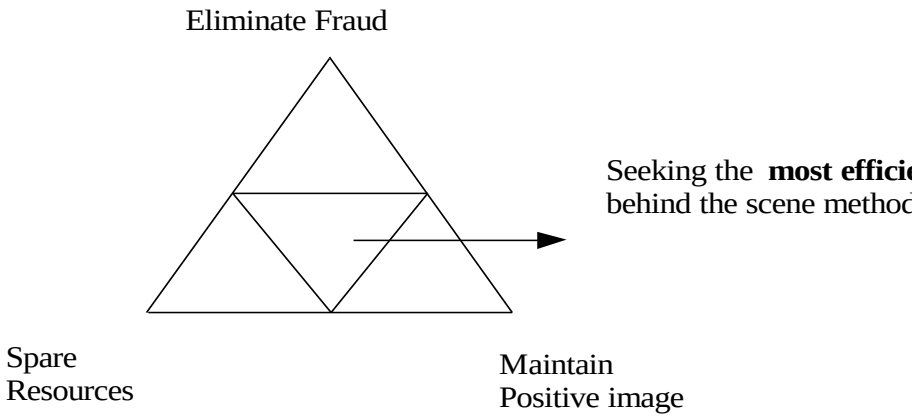

Figure 1: Boundaries of anti-fraud operation in an insurance company.

Towards sparing resources an automated screening facility is fundamental to improving efficiency of the process. It is possible to automate the process because decisions on the acceptance of a claim are governed by insurance policy and medical

rules. Bulk of these rules can be described with algorithms. Since human workforce is more expensive than computational processing power and system development in the long run, reducing the involvement of people in the decision making improves efficiency. Principally, these systems do not require interaction with the customer unless the claim is suspicious and put on hold for investigation. This helps to keep a better image of the company (Major & Riedinger, 2002).

The requirement for human intervention for most cases is dependent on the complexity of the claims as well as the rules associated with the claims. Anti-fraud operations can consume fair bit of human resources, because in most cases it is difficult to draw the distinction between fraud and error. Claims which were deliberately falsified or altered in order to gain financial benefit are unlawful and labelled as fraud, while genuine mistakes are considered to be errors (Sparrow, 1996). We therefore, suggest that an analogy between what constitutes intentional or non intentional error could be used as a metaphor for establishing the ground of fraud.

Additional clues are required to be gathered to make a reliable decision. This can be a lengthy process involving a legal team. To be time wise efficient paying claims, complex analytical processes are performed on historical data only not in real time. This however implies, that recovering money is uncertain from any fraudulent claim identified retrospectively, since the claims have already been paid, before the investigation was finished. Therefore, it is desired to capture problems upfront or minimise the possibility (Sparrow, 1996).

The aim of this paper is to provide methodological framework based on social network methods and related theories for examining possible fraudulent behaviour in health insurance claims processing. We draw the relationship between network structure and criminal activity for flagging potential boundary cross over. We build our model on the basis of classical theories originating from error detection and social networks for presenting a framework to increase the reliability and effectiveness of the anti-fraud operations.

2.0 Background to the study

Deception for personal gain is considered to be unlawful and can be labelled as fraud. As long as the intention behind the act is not proven, it is considered to be a

mistake or error, therefore we look through error detection methods first.

Traditional methods of error detection use threshold, decision based on profiling and tracking. Profiling has become best practice over threshold . There has also been a shift from detection towards prediction and prevention (Cahill, Lambert, Pinheiro, & Sun, 2004). Adaptivity and in combination of advanced technologies such as neural networks, fuzzy logic and genetic algorithms are required to reliably predict infringements (Phua, Lee, Smith, & Gayler, 2005).

Adaptive systems are usually based on one of the two types of machine learning--unsupervised and supervised learning. Unsupervised methods do not perform optimally under the condition of uneven class sizes and uncertain class membership. This is due to that fact that the system can let fraudulent transaction through, even though a false alarm would be triggered on a legitimate one which is a similar issue to the traditional threshold method. However, supervised methods can only be used if prior sets with classification are available and reliably identified by other methods manually. Supervised learning algorithms can help to satisfy the adapting requirement. Common algorithms used are: bayes, foil, ripper, cart, c4 (Bolton & Hand, 2002).

The observations in detecting irregularities specifically in the health insurance sector focuses on: behavioural heuristics, the flow of dollars, medical logic, frequency and volume of treatments, geographic origin, time and sequence of activities, granularity, identification process (Fawcett & Provost, 1997; Major & Riedinger, 2002).

A case is marked fraudulent once the accused is convicted by court order. However in a broad sense even in science literature fraud is analogical with potential fraud, unusual and suspicious behaviour

Differential association theory (Chambliss, 1984) links fraud detection with the study of social networks. Social network in this paper refers to a structure of individuals or organisations in reality, rather than as commonly used today describing online databases and services with a web interface. Differential association theory claims that a person turns away from the lawful life, when the balance of explanations for violation of the law surpasses law-abiding and supported by social affiliation.

Therefore, criminal behaviour is observed to be learned through socialising with

closed group of criminals. Not only the behaviour but motives and techniques and legal loopholes are also learned through the aforementioned association. Relation to a criminal society can vary in priority, duration and intensity. Therefore, learning by association has been seen as an emerging theory for explaining the link the socialisation and learned behaviour

Inspired by the fact that "… fraudsters seldom work in isolation ..." (Bolton & Hand, 2002) we have a good reason to believe that social network analysis could support prevention and detection methods. There is clearly a need for support of the existing techniques since machine learning algorithms do not perform reliably on their own. A combination of different algorithms was proposed to improve the reliability. However as visible on Figure 2 last column pair, even combination of Probabilistic, Best match, Density selection and Negative selection algorithms only reach 50% reliability in fraud detection.

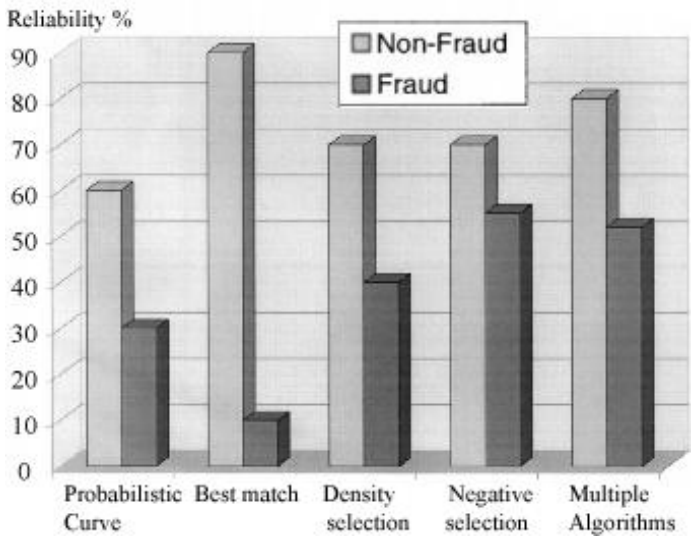

Figure 2: Reliability of different algorithms and combination of them.

We therefore, introduce social networks and its analytics in developing a better understanding of the detection of error and or fraud here.

2.1 Social Network Measures

Social network measures deal with quantifying network structure properties and provides models to understand information flow, the evolution of the network, the role of nodes and links with certain properties and the effect of the changes in the network. The networks are built through interactions of nodes, usually human beings but can be informational systems, which makes the research area to overlap with human

computer interaction. The principal measures are size, centrality, density and link weights.

In social networks, centrality denotes the structural power position of a node in a given network. Centrality has three basic measures Freeman degree centrality - number of adjacent nodes, closeness - reciprocal value of the total number of hops in the shortest possible way to every other node, and betweeness - number of times the node appears on the shortest path between other nodes. The higher the value is the more influence can a particular node has on the entire network. Competitive advantage for example can be explained by control on information flow. Entities bridging network segments which otherwise would not be connected just through the particular node are in advantageous position which they can benefit from by using the information flowing through them or controlling dissemination for their own good (Burt, 1995). Centrality is not only understood on nodes, a characteristic. Value can be calculated for the total network as well using any if the above measures.

Network density is the number of links divided by the number of all theoretically possible links (Freeman, 1979; Hanneman & Riddle, 2001).

Tie strength is another measure which is the weight of the links between the nodes and it can be used to explaining information flow. For example, too densely connected actors provide mostly redundant, already known information. This noise causes a delay which has a negative effect on cooperative work. It is also argued that new and innovative information is usually received through weak ties because strong ties share common information (Granovetter, 1973, 1983).

2.2 Health Insurance sector

There are unique contextual features which influence certain aspects of the case study. The room for inconsistency between interpretations is fairly large, because there are no common standards for codes and rules, therefore the compliance gap can be quite substantial.

To put error detection into the context of our case study, we look at claim processing. In case of misinterpretation of a rule we should label an unintentional act as 'error' and deliberate misinterpretation or deception leading to any kind of unfair advantage as 'fraud' by law (Sparrow, 1996).

The challenge here is to differentiate reliably between the two cases, since “it is

not possible to determine the presence of fraud from the data alone" (Bay, Kumaraswamy, Anderle, Kumar, & Steier, 2006).

Certain thresholds are common to use to filter distinct member and provider profiles against standards and averages of similar profile clusters. Profiles filtered with values above threshold can be used as a basis of feedback to the particular provider or member. If continuous abuse of the rules is recognised either from a fund member or a service provider some sort of higher level disciplinary action should be taken.

Health insurance funds are in a position to collect sensitive data, however the use of this data is regulated by governmental privacy laws. Depending on the country privacy laws can be substantially different in terms of strictness level. In Australia these laws are extremely protective, even in cases where it would be common to supply and use information it is prohibited.

3.0 Discussion

In this study, we ask: How does link analysis of customer/provider assist in providing clue related to fraud? We explore collective fraudulent activities, focusing on collusion between providers because it has a higher risk than collusion between health service providers and insurance fund members. Furthermore, discovering provider to provider agreements entails identifying the members associated with the questionable transactions.

In order to identify possible connections between providers we need to predict existing links. There are three types of approaches to do this (i) Probabilistic models supported by machine learning algorithms such as Relational Bayesian, Markov and Dependency Networks; (ii) Node wise and Topological similarity (iii) Maximum likelihood. Probabilistic and maximum likelihood algorithms are powerful, but very complex and computationally expensive (Lü, 2010). Configuring node wise prediction which is based on information distance and similarity disregards link information, they also required supervised methods and we did not have enough training data available for this.

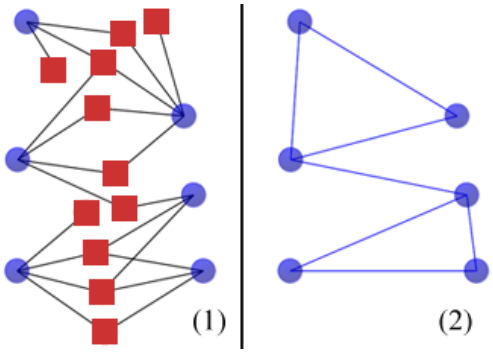

Figure 3: Demonstration of topological similarity link prediction based on common neighbours.

Topological similarity method was selected due to its relative simplicity and available implementation. There are several algorithms in the similarity framework, for example: common neighbours, Jaccard's coefficient, Adamic/Adar. Preferential attachment has the lowest computational requirement as it does not require information on neighbourhoods. (Lü, 2010) Multi dimensional networks are built of different network layers and known to be challenging link prediction problem scenario. (Lü, 2010) Network layers can be the same nodes linked up based on different, often conflicting criteria. Another type of multi dimensionality is having dissimilar class of nodes, like in our case providers and customers. On Figure 3 you see a demonstration network with two different node classes. Providers are blue dots and members members are red rectangular. Left hand side is the normal network information representing the connections based on the data available. Right hand side is the pure predicted network of providers.

Once we have identified the connections we would like to see how significant they are and give an indication of the extent they are suspicious. The link between providers and members is established based on claims. The insurance company reimburses the fund member with a fraction of the service fee defined as benefit based on the insurance policy in effect. These paid benefits are the subject of protection against fraudulent activities. Therefore summarising the benefits in all transactions between each provider, member pair gives us a dollar value significance in other words the link weight. Representing these weights do not indicate level of irregularity however they tell the possible impact factor.

It was demonstrated that link weights used for detection can be misleading. Using the unweighed version of the similarity index algorithms outperformed their weighted counterparts. Based on the weak ties theory: network connectivity is accounted

mainly to weak links, therefore they play an important role in link detection as well. (Lü, 2009) Therefore we apply the detection algorithm in its simplest form, establishing link on the slightest topological similarity and then apply weighting on these established links using domain related predictors.

These predictors will differentiate between coincidental and suspicious links. We propose to use a combined measure of (1) geographical distance (Phua, Lee, Smith, & Gayler, 2005) and (2) shared customer base. (1) Members visiting different providers in large geographical distances from each other is unusual practice, since most people visit practitioners in convenient distances, so all of those linked generally would be in the same area. Consequently providers linked based on shared customers in the same area specially if they provide different type of services considered to be normal. Another normal scenario is when a patient is referred to a specialist which can be far away, however if (2) the proportion of common neighbours against the entire client base is high that makes it look irregular. Combining one and two with a simple addition expected to result in a good predictor for irregularity.

We induced a threshold of minimum ten shared customers based on histogram of this value after performing link detection based on common neighbours in our sample. After imposing this limit we got the core network of thirteen providers as it can be seen on Figure 4.

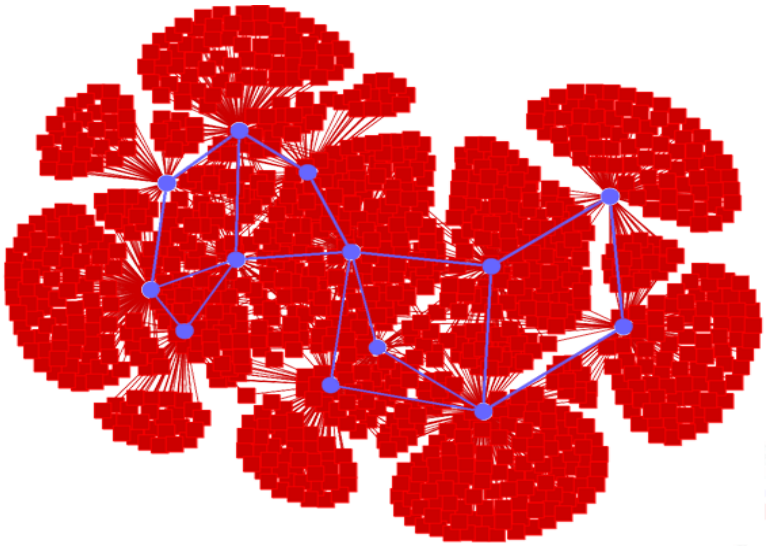

Figure 4: Predicted core network of providers extracted from real data with customers around them.

Further analysing this result graph customer nodes were hidden to focus on provider connection with the network. Visualisation parameters were set to colour nodes based on type of speciality and colour links according significance. The result is presented on Figure 5. As discussed earlier to indicate significance we use a summary of distance and shared customer base proportion as link weights. The more weight a links carries the more suspicious the association is. Further to this if a node is more central than others it means that it is more connected to other suspicious nodes. Centrality is visualised with the size of the node on the figure. Whenever we see nodes with same colour linked to each other that means same type of health service providers are sharing customers. All this information provides clues to spot irregularities. Looking at the figure we have to keep in mind that this is a limited set of the entire data with certain thresholds, so to get a comprehensive picture it would worth going back to the original full data set and extract all data related to the providers in question. We can then use the ego centric approach to keep the analysed data on a manageable level.

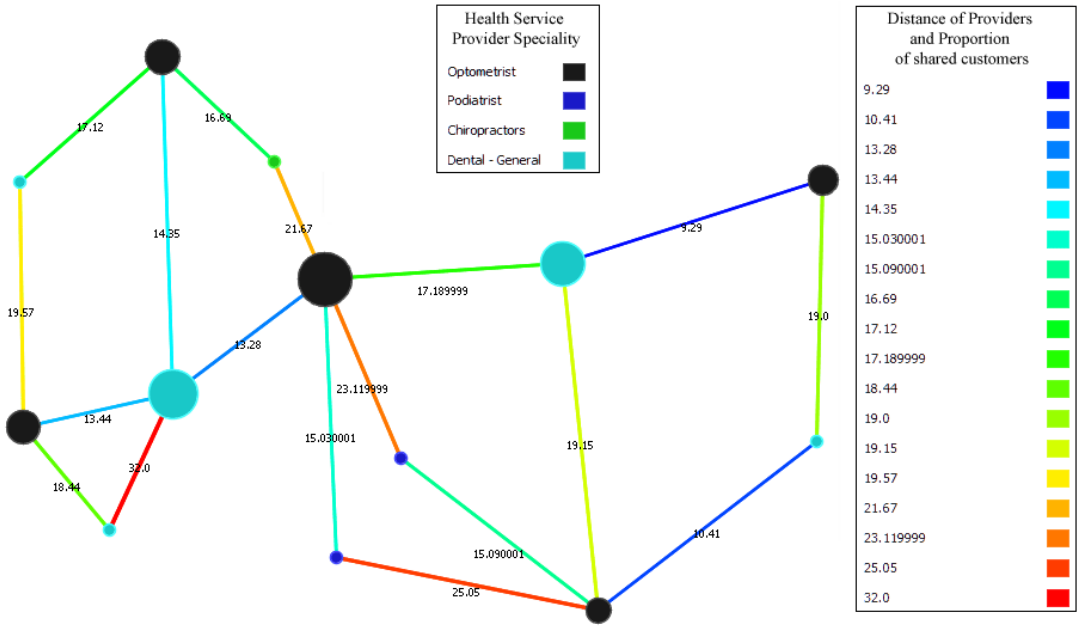

Figure 5: Predicted core network of providers in a filtered data set.

4.0 Data

4.1 Description and scope

We were given access to claims processing data of a major Australian health insurance fund. Four years of data contains over 65 thousand providers, 2 million customers and 28 million claim entries.

Clustering is inevitable to derive sensible results and to be able to process the data in reasonable amount of time. Based on the recommendations of the industry expert in health insurance fraud we have targeted a specific geographical area first, denoted by a single postcode. In order to avoid cutting off links of interest by the strict borders of a post code we extended our data set using a number of steps:

- selected all providers from the given post code
- took all customers related through claims to these providers.
- extracted all claims of these customers
- based on the resulted claim set collected all providers involved

4.2 Filtering

The result set was still extremely large after clustering still had over a hundred thousand claims and couple of thousand providers. At the particular fund we are working with, internal provider codes are used. Therefore we had to make sure we eliminate these providers first. Internal providers are considered to be trusted and much more claims are running through them than individual providers, so keeping them in would have biased our analysis. Our aim was to filter the sub-set of data to be able to focus on the top 100 health service providers of interest. The interest of the insurance fund is the amount of benefit paid to these providers, hence most of our data filtering is based on dollar values. The first steps in filtering included:

- calculate accumulated benefit over the years per provider and keep only the top 100

From network analysis perspective we need at least two links to different providers from a customer to be able to predict a connection between providers. Furthermore the links need to be significant

- remove customers who did not have claims through at least two providers
- remove transactions below average grouped by customer and provider pairs

4.3 Data preparation for analysis

We planned to use the Organizational Risk Analyzer (ORA) (Carley & Reminga,

2004) application from Carnegie Mellon University for analysis because it has the folding of networks based on shared links function implemented. First step was to format the relational data extract we have acquired from the insurance fund according to the needs of Social Network Analysis application. To do so we have taken the following steps in the first pass:

- selected the total dollar value of benefits grouped by customer, provider pairs and year month
- exported the result into a coma separated value file
- inserted two new columns manually: source node class, destination node class

The dollar value summaries served as link weights and the two new columns were filled with equal row values: customer, provider accordingly. The year month field was given in the database and we have used it as network identifier so once imported into ORA it further clustered the data based on time. Time information can also be used to analyse network evolution and dynamics.

4.4 GPS location of providers

We wanted to analyse link weights not only on the flow of dollars, but based on distances between providers as well. The following steps were prerequisite to efficiently do this. The address field of the provider table was used to geographically code the GPS coordinate of the providers. We have used the database of Google maps to assign each provider with the appropriate longitude, latitude pair. Additionally to our goals based on the coordinate set ORA automatically created regions finer than postcode, and enabled us grouping the providers more precisely.

4.5 Shortest geographical distance of providers

We needed to calculate the distances between providers where there is a link and assign the result as link weight. ORA was helpful to create network data from relational data, however the output of this operation was an adjacency matrix. For our purposes direction of the link is not important, therefore the matrix is symmetric, hence half of the output is redundant. Furthermore, to add distances we only need data on established links, however majority of the matrix was filled with zeros. Therefore, adjacency matrix is difficult to process with standard applications such as a spreadsheet calculator. NCOL format (Adai, Date, Wieland, & Marcotte, 2004) for social network data is much more condensed and suitable to further processing the data in relational database management systems as well. It has only three columns:

from node, to node, link weight and only established links are recorded. Therefore to overcome the above issue we wrote a script, which converts an adjacency matrix into NCOL format. (See in the appendix)

After the conversion we have loaded the result into a new relational database table and another table with provider identifiers, longitude and latitude values. We have matched the two tables based on the provider identifiers which resulted in coordinate to coordinate pairs. Finally using a great-circle distance algorithm custom spreadsheet macro we have calculated the distance between the linked providers.

4.6 Percentage of shared customer base

We have further explored the proportion of customers shared between the linked providers for achieving higher precision in our analysis. To assign this attribute value for each pair of linked providers, we have used a similar relational approach for calculating the distances. Our first step was to binarize the adjacency network data then fold the network based on shared customers. This resulted in a matrix where self loops are the total number of customers and each link weight is the number of shared customers. We divided these two numbers to get the proportion, which was calculated for both nodes in a pair and the higher value was used (Bolton & Hand, 2002).

5.0 Conclusion

We have presented a comprehensive method to analyse health insurance claims data combining social network methods with traditional profiling. A focus group interview with a group of industry expert confirmed that the highlighted situations are indeed suspicious. The industry expert panel further provides validation that existing methods of detection were not able to highlight these providers to be suspicious. We are waiting on the expert to verify our results by conducting further manual data analysis and in case of reasonable doubt on site investigations. We see great possibilities to continue and extend this work. Our next steps will be to analyse network dynamics over time and look for clues in the evolution of the network. Also to correlate the results with other moderating variables such as feedback, and to combine our proposed predictor variable with the risk index already used at the fund. Furthermore we have existing cases of investigated fraud which can help us to develop a set of typical social network measures for potentially fraudulent networks.

Appendix

Simple python converts an adjacency matrix into NCOL format:

```
#!/usr/bin/env python
import csv
import sys
import re
m = []
r = csv.reader(open(sys.argv[1],'rb'))
for row in r:
	m.append( [i for i in row] )
of = open(re.sub(".csv","-ncol",sys.argv[1]),'w')
for i in range(2,len(m)):
	for j in range(1,i):
		if (int(m[i][j])>0):
			of.write ("%s;%s;%s\n" % (m[i][0],m[0][j],int(m[i][j])))
of.close()
```

References

Adai, A., Date, S., Wieland, S., & Marcotte, E. (2004). LGL: creating a map of protein function with an algorithm for visualising very large biological networks. *Journal of Molecular Biology, 340*(1), 179-190.

Bay, S., Kumaraswamy, K., Anderle, M., Kumar, R., & Steier, D. (2006). Large scale detection of irregularities in accounting data.

Bolton, R., & Hand, D. (2002). Statistical fraud detection: A review. *Statistical Science, 17*(3), 235-255.

Burt, R. (1995). *Structural holes: The social structure of competition*: Harvard Univ Pr.

Cahill, M., Lambert, D., Pinheiro, J., & Sun, D. (2004). Detecting fraud in the real world. *Computing Reviews, 45*(7), 447.

Carley, K., & Reminga, J. (2004). *ORA: Organization Risk Analyzer.* Carnegie Mellon, Pittsburgh PA.

Chambliss, W. (1984). White Collar Crime and Criminology. *Contemporary Sociology: A Journal of Reviews*, 160-162.

Freeman, L. (1979). Centrality in social networks conceptual clarification. *Social networks, 1*(3), 215-239.

Granovetter, M. (1973). The strength of weak ties. *ajs, 78*(6), 1360.

Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological theory, 1*(1983).

Hanneman, R., & Riddle, M. (2001). Social Network Analysis: University of California, Riverside.

LÜ, L. & ZHOU, T. Year. Role of weak ties in link prediction of complex networks. *In*, 2009. ACM, 55-58.(Lü and Zhou, 2009)

LÜ, L. & ZHOU, T. 2010. Link prediction in complex networks: A survey. *Arxiv preprint arXiv:1010.0725*.

Major, J., & Riedinger, D. (2002). EFD: a hybrid knowledge/statistical-based system for the detection of fraud. *Journal of Risk and Insurance, 69*(3), 309-324.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2005). A comprehensive survey of data mining-based fraud detection research. *Artificial Intelligence Review*, 1-14.
Sparrow, M. (1996). Health care fraud control: understanding the challenge. *JOURNAL OF INSURANCE MEDICINE-NEW YORK-*, *28*, 86-96.