MEGI

MESTRADO
Estatística e Gestão de Informação

Sistema para deteção de fraude na indústria seguradora:

A aplicação de redes ao ramo da saúde

Margarida Martins Pinheiro Amado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Estatística e Gestão de Informação

UNIVERSIDADE NOVA ISEGI

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

SISTEMA PARA DETEÇÃO DE FRAUDE NA INDÚSTRIA SEGURADORA: A APLICAÇÃO DE REDES AO RAMO DA SAÚDE

por

Margarida Amado

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Estatística e Gestão de Informação, Especialização em Análise e Gestão de Risco.

Orientador: Doutor Rui Gonçalves

Novembro 2014

AGRADECIMENTOS

Quero agradecer em primeiro lugar ao meu irmão que sem o saber foi a força e a motivação em momentos que só me apeteceu desistir, porque acredito que o exemplo vem de cima e a vontade de mostrar que com determinação e empenho tudo é possível foi fundamental para a conclusão desta dissertação.

Quero também agradecer à minha irmã e aos meus pais por me terem incutido a necessidade e o desejo de alcançar sempre mais.

Ao Pedro pelos dias e noites de ajuda, pela motivação, paciência e apoio cruciais para a conclusão desta dissertação.

Um agradecimento especial ao Prof. Rui Gonçalves por todo o apoio, orientação e disponibilidade determinantes para a realização desta dissertação.

Por último, e porque não podia deixar de ser, à Carolina, Rita e Candice porque começamos este desafio juntas e por terem sido as minhas companheiras neste período.

RESUMO

A fraude nos seguros de saúde é uma realidade que causa perdas bastante significativas às empresas do setor. A presente dissertação tem como objetivo fornecer uma perspectiva das técnicas já utilizadas para detetar fraude e apresentar as potenciais vantagens da utilização de redes para a deteção e consequente prevenção deste tipo de comportamentos. A metodologia de investigação baseia-se na revisão da bibliografia bem como no estudo e aplicação prática de um algoritmo de redes aos dados de uma seguradora da área da saúde. A escolha da área da saúde teve por base o facto desta ser uma área onde a fraude tem um impacto bastante elevado e onde os esquemas fraudulentos se desenvolveram muito nos últimos anos. Com base nos requisitos identificados foi possível avaliar as vantagens da utilização de redes para deteção de fraude bem como descrever o processo de como as redes podem responder às necessidades atuais e futuras das empresas seguradoras no que respeita ao combate à fraude.

PALAVRAS-CHAVE

Palavras - Chave: Seguros de Saúde; Análise de Redes; Fraude;
Códigos JEL: I13; D85; G20;

ABSTRACT

Health insurance fraud has been causing significant losses to all insurance companies acting on this area. This dissertation has the objective to provide a perspective of fraud detection techniques already in use and present the potential advantages of networks on this behaviors detection and consequent prevention. The investigation methodology is based on the bibliography review along with the study and practical application of a network algorithm to an insurance company data. The option for this specific area was based on the high impact of fraud on health insurance and also on the high development of fraudulent schemes on the latest years. Based on the identified requirements it was possible to evaluate the advantages of network usage to detect fraud and also describe the complete process so it will be possible to cope with insurance companies present and future needs regarding fraud combat.

KEYWORDS

Keywords: Health Insurance; Network Analysis; Fraud;
JEL Codes: I13; D85; G20;

ÍNDICE

ÍNDICE DE FIGURAS

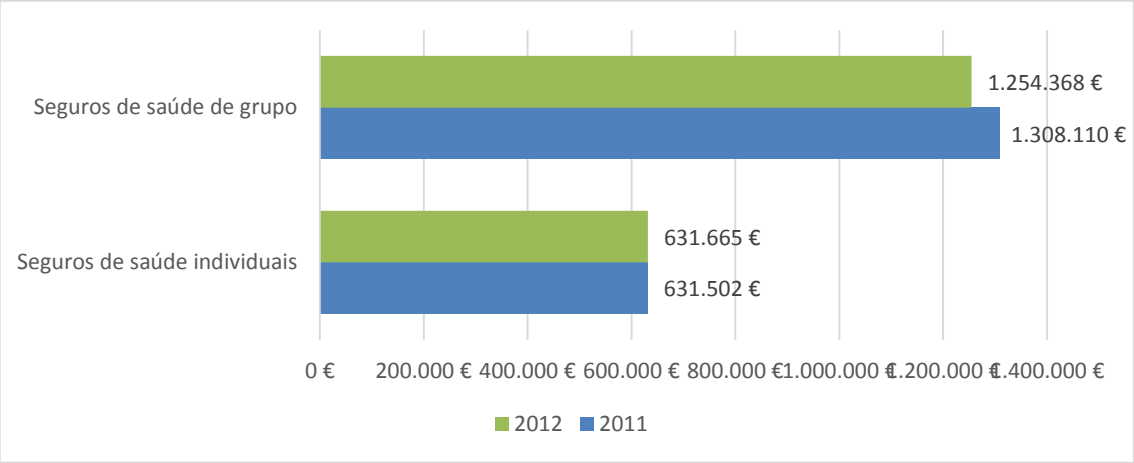
1. OBJETIVOS E RELEVÂNCIA

1.1. INTRODUÇÃO

O sistema de saúde português assenta no Serviço Nacional de Saúde (SNS), responsável por garantir proteção universal a todos os cidadãos de forma gratuita (Guiomar, 2010), no entanto, é também composto "pelos vários subsistemas de saúde públicos e privados, pelo setor segurador, e pelo setor privado "puro", financiado por pagamentos diretos dos indivíduos" (Silva [APS], 2009). De acordo com dados da Associação Portuguesa de Seguradores (APS) a percentagem de cidadãos com seguros de saúde tem vindo a aumentar, tendo atingido os 63,7% em 2012, o que revela o perfil de crescimento que este tipo de seguros tem evidenciado em Portugal. Seja devido à importância que este tipo de seguro tem no âmbito de proteção social individual, seja porque muitas empresas oferecem como regalia aos seus funcionários, é importante sublinhar este crescimento, não podendo deixar de se referir que o seguro de saúde individual (riscos referentes a uma única pessoa) apresenta, em Portugal, valores de crescimento em oposição ao seguro de saúde de grupo (conjunto de pessoas ligadas ao tomador do seguro por uma relação distinta do seguro) que apresenta um comportamento decrescente.

De acordo com a análise da figura 1, é possível perceber um decréscimo (- 4,11%) no que respeita aos montantes gastos em seguros de saúde de grupo e um ligeiro aumento (0,03%) nos montantes gastos nos seguros de saúde individuais. O aumento dos seguros individuais em detrimento dos seguros de grupo pode justificar-se com a crise financeira que se viveu em Portugal que levou a muitas empresas a proceder a cortes significativos nos custos, estando os seguros de saúde abrangidos nesses cortes. Por outro lado, e conhecendo-se a realidade do sistema nacional de saúde, existiu a necessidade dos indivíduos anteriormente abrangidos por seguros de grupo optarem por soluções individuais.

Figura 1 - Montantes gastos em Portugal em seguros de saúde individuais e de grupo (Fonte: Elaboração própria, recurso a dados de 2011 e 2012 da APS)



A importância dos seguros de saúde em Portugal pode ser justificada com a percentagem que as despesas com a saúde[1] representam no PIB de 2012 (conforme dados retirados da Pordata). Através da análise dos dados divulgados é possível verificar que entre 2008 e 2012 houve um aumento de 0,7 p.p nesta percentagem, atingindo em 2012 os 9,2%. Apesar da aposta que se tem vindo a verificar na investigação da fraude neste setor, para ser possível reduzir os custos verificados com a saúde e contrariar o crescimento que se tem vindo a verificar, seria importante proceder à identificação de potenciais comportamentos fraudulentos.

A fraude pode ser encontrada em todos os tipos de seguros, incluindo o seguro de saúde (Kirlidog & Asuk, 2012). Conforme afirmou Rui Gil (2008), Presidente da Comissão Técnica de Fraude, ao longo dos anos a fraude nos seguros evoluiu e "apresenta uma crescente sofisticação de métodos e técnicas utilizadas pelos elementos fraudulentos, que isoladamente ou em grupo, pretendem retirar benefícios ilegítimos para si ou para outros".

Numa tentativa de combater a evolução da fraude nos seguros, a APS criou em 2006 uma comissão técnica composta por representantes de companhias de seguros que, em conjunto, refletem sobre os aspetos centrais deste fenómeno, partilham experiências, desenvolvem ações conjuntas, identificam práticas de fraude e tentam fornecer soluções (Gil, 2008). Por outro lado, obriga as seguradoras a definirem e

[1] Pordata, Despesa corrente em cuidados de saúde em % do PIB em Portugal . http://www.pordata.pt/Portugal/Despesa+corrente+em+cuidados+de+saude+em+percentagem+do+PIB-610, Acedido em 02-04-2014.

implementarem medidas para a prevenção, deteção e reporte de fraude (CEA, 2010). A publicação do artigo 131.º-F do Decreto-Lei N.º 2/2009, de 5 de Janeiro, veio reforçar o combate à fraude nos seguros, uma vez que obriga todas as empresas de seguros a definir uma política de prevenção, deteção e reporte de situações de fraude nos seguros. A definição destas políticas não tem sido linear devido à atual conjuntura económica , aos custos de averiguação elevados e aos prémios baixos (Correia, 2012).

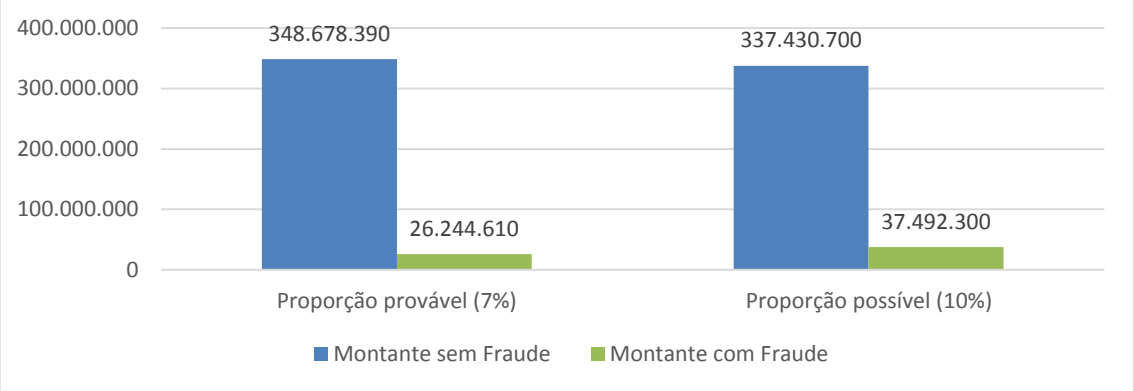
Em Julho de 2014 um comunicado da Comissão Europeia (CE) afirma que "os Estados-Membros devem reforçar os seus esforços para impedir, detetar e comunicar casos de fraude que afetem os fundos da EU (…)". No mesmo comunicado Algirdas Šemeta, Comissário responsável pela Fiscalidade e União Aduaneira, destaca a importância dos Estados-Membros desempenharem um papel eficaz na redução da fraude através do fortalecimento das ações em matéria de deteção e repressão dos comportamentos fraudulentos.

No que respeita à fraude no setor dos seguros de saúde, de acordo com Kirlidog e Asuk (2012), "é feita por engano intencional ou deturpação para a obtenção de algum benefício nos gastos com a saúde". São vários os tipos de fraude no setor dos seguros de saúde, conforme publicado pelo National White Collar Crime Center (2013), nomeadamente: (I) faturar medicamentos, procedimentos ou serviços que nunca foram realizados; (II) faturar acima do custo real; (III) divisão em diversas faturas de um só tratamento; (IV) alterações na natureza ou valor do serviço médico; (V) alteração nas datas dos serviços; (VI) alteração da condição médica do paciente à data do tratamento; (VII) alteração do diagnóstico ou da identidade do cliente; (VIII) execução propositada de serviços médicos desnecessários; (IX) pagamentos de compensações a fornecedores em troca de encaminhamento de doentes para serviços especializados; (X) recorrência a vários médicos para obter várias receitas de forma a conseguir prescrições de mais medicamentos do que os necessários.

De forma a ser possível quantificar o impacto da fraude neste setor de seguros, é importante analisar os resultados do estudo "The Financial Cost Of Healthcare Fraud 2014" (Gee & Button, 2014) que se baseou em dados mundiais. Este estudo divide as perdas em: perdas efetivas, perdas prováveis e perdas possíveis. De acordo com este estudo em qualquer organização da área dos seguros de saúde as perdas em fraude vão ser de pelo menos 3% (perdas efetivas), no entanto provavelmente serão mais de 7% (perdas prováveis) e possivelmente mais de 10% (perdas possíveis). Aplicando os resultados do estudo aos montantes pagos pelas seguradoras no ano de 2012 (segundo dados das Estatísticas do Seguro de Saúde da APS, em Portugal) é possível perceber que,

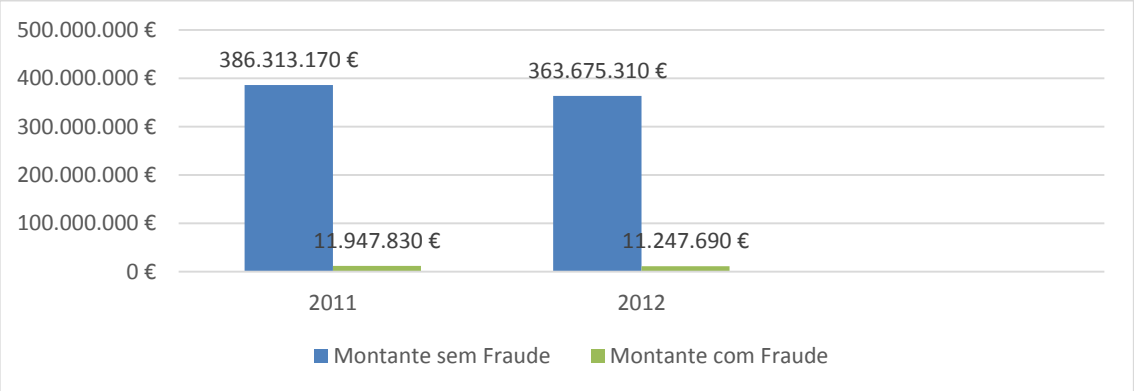
se as perdas forem as prováveis (ou seja, mais de 7%), serão de pelo menos 26.244.610€, no entanto, se atingirem o patamar mais elevado, ou seja, se forem as perdas possíveis (10%) serão de 37.492.300€ (figura 2).

Figura 2 - Montantes pagos pelas seguradoras da área da saúde em Portugal
(Fonte: Elaboração própria, recurso a dados do ano de 2012 da APS, aplicando as proporções definidas no estudo The Financial Cost Of Healthcare Fraud 2014)



Utilizando os mesmos dados e aplicando desta vez à proporção de fraude existente nos Estados Unidos da América (EUA) (3% de acordo com dados da National Health Care Anti-Fraud Association [NHCAA]) é possível perceber que corresponde a uma perda anual de 11.247.690€ (figura 3).

Figura 3 - Montantes pagos pelas seguradoras da área da saúde em Portugal em 2012 face ao ano de 2011
(Fonte: Elaboração própria, recurso a dados do ano de 2011 e 2012 da APS, aplicando a proporção dos 3% definida pela NHCAA)



Assim, os elevados montantes pagos pelas seguradoras agregados ao aumento de 2.32 p.p na percentagem de pessoas seguras que deram origem a custos para as seguradoras (figura 3) justificam o desenvolvimento de um sistema de deteção de fraude que permitirá às seguradoras prevenir e diminuir o impacto da fraude nos seguros de saúde e consequentemente reduzir a percentagem de incidência de utilizadores (percentagem de utilizadores que originam custos para a seguradora). A diminuição desta percentagem é uma meta significativa sobretudo se for tido em conta que apesar de se ter verificado um aumento de aproximadamente 2% entre 2011 e 2012 neste indicador (figura 4) o número médio de pessoas seguras decresceu entre 2011 e 2012. Estes números espelham uma relação inversa entre a incidência de utilizadores e o número de pessoas seguras, ou seja, através da análise da figura 4 é possível verificar que, embora existam menos pessoas detentoras de apólices de seguro existem mais custos para as seguradoras.

Figura 4 - Incidência de utilizadores e número de pessoas seguras na área dos seguros de saúde em Portugal

(Fonte: Elaboração própria, recurso a dados de 2011 e 2012 da APS)

Ano	Incidência de utilizadores	Nº de pessoas seguras (valor médio)
2011	61,3%	1940
2012	63,7%	1886

A presente dissertação pretende contribuir para uma evolução no que respeita à prevenção de fraude no setor dos seguros de saúde, considerada de alto risco devido ao número de casos reportados e investigados e aos sérios impactos que tem nos sistemas de saúde. Com o desenvolvimento de um sistema capaz de diminuir os impactos e número de fraudes nos seguros de saúde será possível reduzir os números apresentados e consequentemente tornar o processo dos seguros de saúde mais eficaz e menos vulnerável a este crime.

1.1.1. Relevância do tema

A deteção de situações fraudulentas cresceu e tornou-se numa prioridade para as seguradoras. (Viaene et al, 2002). Agregando a necessidade de definir políticas anti-fraude, com o facto da fraude no setor dos seguros de saúde ser cada vez mais importante devido aos custos crescentes desta área (Fisher, 2008), a criação de um

sistema capaz de identificar quais os clientes com maior grau de propensão a fazer fraude, tendo em conta as relações entre eles e os vários intervenientes do processo, permitirá às seguradoras classificar os mesmos de acordo com o seu perfil de risco e tomar medidas concretas e direcionadas para cada cliente.

O facto deste tipo de fraude ter sido a atividade criminosa que mais cresceu nos EUA na última década, devido sobretudo à passagem de foco para áreas mais seguras e lucrativas, como é o caso da área dos seguros de saúde em detrimento dos cartões de crédito ou outras áreas de maior risco (Allmon, 2005 as cited Fisher, 2008), reforça a importância de um modelo capaz de identificar quais os clientes mais propensos a fazer fraude.

Apesar das dificuldades relacionadas com a identificação de um comportamento fraudulento, devido não só ao facto de ser difícil definir a fronteira que separa o normal do fraudulento mas também devido à desacreditação das empresas nos processos legais (por receio de manchar a sua imagem ou por não acreditarem no impacto destes processos), a nível internacional é possível obter algum detalhe relativo à quantificação da fraude nomeadamente que 62% das empresas internacionais com mais de 5000 trabalhadores declararam ter sido alvo de fraude o que torna este tipo de empresas o alvo preferencial deste crime (Pimenta 2009).

Para Fisher (2008) o aumento do investimento nos seguros de saúde, torna o crime da fraude mais atrativo neste setor e consequentemente serão registadas mais evidências de fraude o que fará os processos evoluir e tornarem-se cada vez mais complexos e difíceis de prevenir ou detetar. Posto isto, apostar num modelo capaz de prever quais os intervenientes mais suscetíveis a cometer fraude neste ramo permitirá às seguradoras mitigar o risco e complementar os sistemas reativos existentes através da adoção de uma atitude rigorosa e proactiva em relação aos mesmos.

Os factos apresentados acima agregados à realidade da lei portuguesa que prevê a obrigatoriedade das seguradoras tomarem medidas de prevenção do risco levam à necessidade de desenvolvimento de um sistema de deteção de fraude no ramo dos seguros da saúde, capaz de detetar quais os intervenientes mais suscetíveis a fazer fraude.

Atualmente existem várias técnicas para combater as situações de sinistros fraudulentas, no entanto, muitas seguradoras optam pela utilização de sistemas transacionais de monitorização para detetar transações fraudulentas num determinado período temporal (semanas, meses, anos, etc.). Outras medidas utilizadas pelas seguradoras para combater a fraude no seguro de saúde são: encontrar pagamentos de

compensações em troca de direcionamento de pacientes, identificar *outliers* nos valores dos tratamentos, sumariar diferentes faturas para o mesmo tratamento, encontrar números de identificação falsos (por exemplo, segurança social), identificar faturação excessiva por um único prescritor e identificar pagamentos múltiplos para uma mesma conta bancária (ACL, 2010).

A identificação e exposição dos diferentes mecanismos de deteção de fraude permitirá compreender quais os algoritmos que melhor se adequam à realidade do mercado segurador, percebendo a sua evolução e especificidades. Adicionalmente a exploração de conceitos de *data mining* permitirá analisar de que forma estes algoritmos estatísticos poderão apoiar o processo de deteção de fraude através de uma maior perceção comportamental das entidades passiveis de cometer este crime.

A conjugação destas medidas com o facto da maioria dos processos de negócio estarem dependentes de tecnologia transforma essa mesma tecnologia num meio necessário para detetar a fraude permitindo assim o desenvolvimento de programas contínuos de deteção de fraude que garantirão a salvaguarda das empresas em relação ao risco e reduzirão o tempo necessário para operacionalização de ações relativas às fraudes detetadas (ACL, 2010).

Através do sistema de deteção de fraude será possível analisar não só a entidade mas também analisar todas as atividades e relações envolvidas na sua rede de influência. De acordo com Soares (2009), perceber uma entidade passa por "entender o seu comportamento através das suas atividades na rede". Assim, pela observação da rede é possível perceber a tipologia e características das entidades que cometem fraude. A utilização das redes permitirá ainda identificar quais as entidades com mais alertas/com alertas de maior severidade, permitindo assim conhecer as entidades com maior probabilidade de ter feito/vir a fazer fraude.

1.1.2. Objetivo da tese

Numa tentativa de responder à problemática relacionada com o papel dos algoritmos de redes sociais nos seguros de saúde, foi definido um objetivo geral e vários objetivos específicos. Respondendo aos objetivos específicos vai ser possível esclarecer a problemática definida para esta dissertação:

"Qual poderá ser o papel dos algoritmos de redes no desenvolvimento da deteção de fraude no setor dos seguros de saúde?"

O recurso aos algoritmos de redes sociais (também denominadas ao longo da dissertação como redes de influência) permitirá modelar as relações entre os detentores de apólices de seguros de saúde, os médicos e os serviços de saúde (centros de saúde, hospitais, clínicas, etc.) de forma a perceber as ligações entre os intervenientes da rede.

O desenho da rede de influências permitirá verificar se existem relações com padrões indicativos de potenciais casos de fraude. Esta identificação será uma mais-valia para a seguradora na medida em que permitirá: reduzir os falsos positivos, melhorar a eficiência das investigações, aumentar a prevenção e deteção de fraude e diminuir os casos reincidentes.

Com o intuito de conseguir responder à questão principal foram definidos os seguintes objetivos específicos:

- Qual a informação necessária e possível de utilizar no desenvolvimento de um sistema de deteção baseado em redes na área dos seguros de saúde?
- Quais os algoritmos de redes possíveis de usar na área dos seguros de saúde?
- Identificar o valor que as redes podem trazer na deteção e prevenção da fraude na área dos seguros de saúde?

2. REVISÃO BIBLIOGRÁFICA

O conceito de risco operacional, de fraude, de seguro de saúde e de métodos de deteção de fraude são conceitos chave para a utilização de modelos cooperativos para deteção de comportamentos fraudulentos. Adicionalmente a identificação e exposição de métodos e algoritmos de deteção de fraude (nomeadamente, *data mining* e redes) permitirá esclarecer e classificar os diversos métodos existentes.

2.1. RISCO OPERACIONAL

O conceito de risco operacional foi ignorado até à década de 90 uma vez que as seguradoras se preocupavam mais com os riscos tradicionais, nomeadamente, o risco de seguro e o risco financeiro (Gonçalves, 2011). No entanto, é importante começar por defini-lo, dado que este tipo de risco é o mais antigo que as instituições financeiras enfrentam (Geiger, 2000). O primeiro conceito deste tipo de risco data de 1993 quando o Grupo dos 30[2] o definiu como a "incerteza relacionada com perdas resultantes de sistemas ou controlos, inadequados, erros humanos ou gestão" (as cited Magalhães, 2012). Mais tarde, em 1998 Shepheard-Walwyn e Litterman declararam que o risco operacional era o termo aplicável a todos os fatores de risco que tenham impacto na volatilidade de custos da instituição (financeira), o que é o caso da fraude.

A crescente preocupação por parte dos reguladores, executivos e da comunicação socia levou a que as entidades começassem a investir em novas regras, nomeadamente o Acordo Basileia II e o Acordo Solvência II (Saidenberg & Schuermann, 2003). O Acordo Solvência II surgiu em 2001 e veio ajudar a criar um espaço próprio para o risco operacional, isto porque, foram identificadas várias limitações no projeto Solvência I, nomeadamente de riscos e de situações que poderiam por em causa a solvência de uma empresa de seguros (Borginho, ISP[3]). Desta forma, este acordo exige que as instituições avaliem os dados de várias fontes para quantificar o risco operacional, assim, funciona como um indicativo do reconhecimento por parte dos supervisores que este risco resulta de conjugações complexas entre riscos e processos de negócio (Grinsven &

[2] Grupo consultivo privado internacional formado em 1978, composto por elementos do setor público e privado e da Academia.

[3] A importância estratégica do Solvência II. http://www.isp.pt/NR/rdonlyres/8370611D-BA30-4885-8A42-244A45268747/0/F33_Art3.pdf, Acedido em 05-07-2014

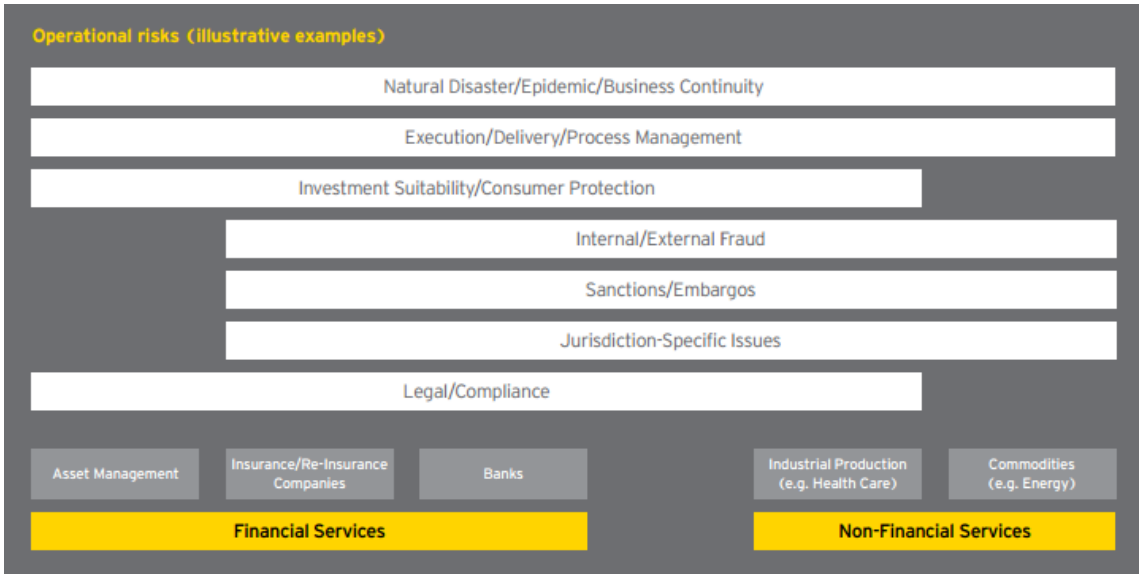
Bloemkolk, 2010). É neste contexto, e no âmbito do Solvência II, que Guiné (ISP[4]) define risco operacional como "o risco de perdas decorrentes de falhas ou inadequação do sistema de controlo interno, pessoas ou sistemas, bem como de eventos externos".

Apesar da difícil identificação e mensuração, a importância do risco operacional tem vindo a crescer devido a vários fatores (Geiger, 2000): (a) a perceção do aumento do risco operacional nos últimos anos; (b) a confirmação da insuficiência da utilização de abordagens para captar alguns riscos e o reconhecimento de que a gestão de risco operacional deve ser uma disciplina separada; (c) a inclusão do risco operacional nos métodos globais de risco; (d) o interesse renovado das autoridades de supervisão no risco operacional.

Para uma melhor compreensão do tipo de risco apresentado torna-se importante enumerar alguns exemplos de riscos operacionais, nomeadamente, violações de segurança, regulação, desastres naturais e a fraude. A análise da figura 5 permite verificar que a fraude, interna e externa, é um dos riscos operacionais a que as empresas devem estar alerta. Dado que o objeto de estudo da presente dissertação é a fraude este conceito vai ser exposto abaixo.

Figura 5 – Exemplos de riscos operacionais

(Fonte: Ernst & Young, 2011)



[4] Solvência II – Resultados do Exercício QIS5. http://www.isp.pt/NR/rdonlyres/052195EE-AF23-4DDD-B225-6167019A726D/0/F31_art1.pdf, Acedido em 05-07-2014

2.2. FRAUDE

A partir dos anos 80, do século passado, passou a ser consensual a existência de um conjunto de atividades económicas não controladas oficialmente ou sequer registadas, sendo também reconhecido que este fenómeno é global e não dependente da atual conjuntura (Pimenta, 2009). Segundo Pimenta e Afonso (2012), quando em 1939 Edwin Sutherland criou o conceito de "crime do colarinho branco" como sendo um crime cometido por uma pessoa respeitável, e de alta posição social, no exercício das suas ocupações, estava a incluir a fraude numa problemática social com necessidade de ser estruturada. Assim, e se já na altura a fraude era um problema social, atualmente, a fraude evoluiu passando a ter "variadas manifestações, crescentes repercussões e maior complexidade em resultado da mundialização" o que justifica que o combate e a prevenção da fraude económico-financeira se tenha tornado ainda mais relevante tendo em conta os acontecimentos recentes, nomeadamente a crise de 2008 e a intervenção da Troika em Portugal (Pimenta e Afonso, 2012).

Segundo Wells, 2007 (as cited Pimenta 2009), "no sentido mais lato a fraude pode incluir qualquer crime para obtenção de lucro, utilizando como principal *modus operandus* o logro". No entanto, segundo Pimenta (2009), o logro não implica diretamente uma fraude, isto porque, para que seja considerado fraude terão de existir danos, geralmente financeiros. Assim, segundo este autor, existem quatro factos sequenciais que necessitam ocorrer para que estejamos perante uma fraude: uma declaração material falsa, o conhecimento de que a declaração era falsa no momento em que foi proferida, a confiança na declaração falsa por parte da vítima e por último danos daí resultantes. Uma vez perante fraude esta pode ser classificada tendo em conta os mais variados critérios: o local, o lesado, quem praticou e a natureza.

O conceito de fraude tem evoluído ao longo do tempo tornando-se cada vez mais exaustivo. Com origem latina "fraus, fraudis" que significa "dano feito a alguém", à definição apresentada por Piccket em 2000 (p. 550, as cited Moura & Silva, 2004), que define a fraude como qualquer comportamento através do qual uma pessoa tem intenção de ganhar uma vantagem sobre outra pessoa, este conceito tem-se tornado cada vez mais abrangente, não sendo ainda no entanto, um conceito fechado (Moura & Silva, 2004), até porque se encontra nas mais diferentes formas e domínios (Šubelj et al, 2011) sendo até assumido como tópico de estudo nas mais diversas ciências (direito, criminologia, psicologia, auditoria, etc.) (Pimenta & Afonso, 2012).

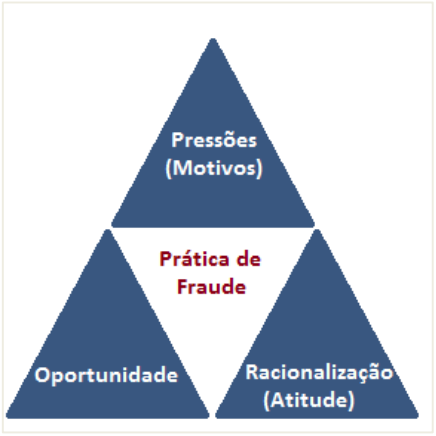
De acordo com o *Instute of Internal Auditors* (IIA), a fraude é "qualquer ato ilegal caracterizado por um engano intencional, ocultação ou violação da confiança. Estes atos

não dependem da utilização de ameaças de violência ou da força física. As fraudes são perpetradas por indivíduos e por organizações para obter dinheiro, bens ou serviços; para evitar pagamentos ou perda de serviços; ou para obter vantagens pessoais ou de negócio" (Pimenta, 2009). O grande diferenciador da fraude em relação a outros crimes (roubo e lavagem de dinheiro) é o facto de esta se basear num contrato em que uma das partes ou mesmo uma terceira atua de forma extrair vantagens sobre outra parte (Contador, 2011).

"Na base do comportamento fraudulento encontra-se um processo humano definido pela combinação de três fatores essenciais: pressões, oportunidade e atitudes. Cada um dos elementos que compõem este paradigma, vão-se ajustando mutuamente, fazendo oscilar a probabilidade de praticar fraude"(Soares, 2008).

Figura 6 - Triângulo de Fraude

(Fonte: Soares, 2008)



Assim, para uma melhor compreensão da figura 6 é importante analisar com maior detalhe cada um dos três fatores, de acordo com a perspetiva de Soares (2008). No que respeita à pressão, este fator conduz normalmente à prática de fraude de cariz financeiro tendo como motivos subjacentes, por exemplo, o desejo de melhor ou manter um padrão de vida elevado, vício do jogo, divórcio e desilusão com a vida profissional. As oportunidades são as perceções por parte de uma entidade da possibilidade de fazer fraude sem ser descoberta. Estas oportunidades surgem de fraquezas que as organizações possuem, por exemplo: inexistência de controlos internos, falta de comunicação anti-fraude, colaboração externa e inexistência de auditoria. Por último a racionalização que é a forma utilizada pela entidade fraudulenta para se justificar a si própria do comportamento fraudulento recorrendo a esquemas

morais, nomeadamente, “foi só desta vez”, “preciso mais disto que eles” e “é uma questão de justiça”.

A fraude tem impactos nas organizações em diferentes áreas, por exemplo a ”área financeira”, a “área operacional” e a “área emocional”. Apesar dos impactos a nível financeiro serem bastante significativos, a influência total da fraude na organização é bastante mais alargada, podendo evidenciar-se quer na reputação, quer nas relações com os clientes (ACL, 2010).

Em Portugal, de acordo com Pimenta (2009) a quantificação da fraude torna-se bastante complexa devido a um conjunto de fatores, institucionais, culturais e cognitivos. Apesar disso é possível obter, com baixa probabilidade de erro um conjunto de valores, como por exemplo, estimar que o total da fraude representa entre 1,5% e 2% do PIB. Uma das principais razões para esta complexidade prende-se com o facto da fronteira entre o normal e o fraudulento não ser suficientemente clara mesmo perante as práticas mais óbvias. O facto das grandes empresas seguradoras não recorrerem ao código penal, para evitar que a imagem da empresa seja lesada e por não acreditarem na rentabilidade dos processos penais, é outra razão que reforça esta dificuldade (Pimenta, 2009).

Em virtude do que foi mencionado é importante referir as diversas tipologias de fraude: fraude interna (de um empregado a uma organização onde pertence), fraude externa (de uma entidade exterior a uma organização), fraude interpessoal (de um ou mais indivíduos a outros) e fraude empresarial (de uma organização a outra organização/indivíduo) (Soares, 2008).

2.3. Fraude no Mercado Segurador

No que respeita ao mercado segurador, segundo Townsend (1979), a fraude é um fenómeno bem conhecido nesta área, que tem bastante impacto dada a importância dos seguros na sociedade atual: “os seguros são uma pedra basilar na vida moderna. Sem seguros, segmentos da nossa sociedade e economia não funcionariam” (Insurance Europe, 2012, p. 3). A fraude nos seguros pode ser consequência da violação de um contrato civil entre a seguradora e o detentor da apólice de seguro, em qualquer fase da relação contratual. A violação do contrato poderá ser através de ações ou omissões deliberadas por parte do consumidor do seguro numa tentativa de obter ganhos para si ou para terceiros (Maio, 2013). Os impactos desta atividade não se cingem apenas à

seguradora mas estendem-se de forma negativa a toda a indústria e às próprias estruturas económico-sociais (Viaene et al, 2005).

A fraude nos seguros é um fenómeno complexo, amplo e dinâmico (Maio, 2013). Segundo Bacher (1995), a fraude nos seguros pode ter diversas formas e baseia-se no ato do segurado enganar a seguradora de forma a obter benefícios que não teria direito. O Instituto de Seguros de Portugal (ISP), 2013, define a fraude contra as empresas seguradoras como a "prática de atos ou omissões intencionais, ainda que sob a forma tentada, com vista à obtenção de vantagem ilícita para si ou para terceiros, no âmbito da celebração ou da execução de contratos de seguros".

Assim, e conforme citado por Maio (2013), a fraude contra as seguradoras pode ocorrer em qualquer fase da relação contratual, podendo ocorrer de forma individual (ou seja, pelo detentor da apólice) ou com o apoio de terceiros (intervenientes ou intermediários dos seguros). Para cumprir esta relação contratual é necessário que, de acordo com Clarke (1989), Morley, Ball & Osmerod (2006) e Viaene & Dedene (2004) (as cited Maio, 2013), ambas as partes (tomador do seguro e seguradora) cumpram o princípio da boa-fé que obriga legalmente a fornecer toda e qualquer informação importante com impacto no contrato mesmo que não solicitada de forma explícita.

Aquando da prática de fraude nos seguros apesar das vítimas diretas serem as seguradoras que sofrem prejuízos na sua capacidade económica, indiretamente todos os cidadãos consumidores de seguros são prejudicados dado que "a fraude nos seguros enfraquece o sistema, uma vez que propostas e sinistros fraudulentos esgotam os recursos pagos pelos muitos clientes honestos para cobrir sinistros verdadeiros" (Insurance Europe, 2013, p. 5). Também Maio (2013) reforça esta ideia ao afirmar que os crimes de fraude resultam em "custos adicionais para os consumidores uma vez que as seguradoras incorporam esses prejuízos esperados no cálculo dos prémios de seguro" (Maio, 2013), assim, como consequência dos casos fraudulentos os detentores de apólice vão pagar prémios de seguro mais elevados e ver os seus sinistros tratados mais lentamente devido à sobrecarga das seguradoras com sinistros fraudulentos.

A fraude nos seguros foi classificada por vários autores, que avaliam os atos decorrentes das fraudes como oportunistas ou leves (maior incidência) e planeados ou graves (menor incidência). Os atos oportunistas têm pouco impacto a nível unitário e são aqueles que resultam de um acontecimento real que é aproveitado pelo detentor da apólice para fazer fraude. Os atos planeados têm um grande impacto a nível unitário e são aqueles em que um individuo ou um grupo organizado planeia um acontecimento

para cometer fraude (Brites, 2006; Ericson & Doyle, 2005; Niemi, 1995; Tennyson, 2008; Viaene & Dedene, 2004 as cited Maio, 2013).

Para além dos diferentes tipos de atos fraudulentos descritos acima é importante distinguir os diferentes tipos de ofensores (responsáveis por cometer fraude). Clarke (1989) cataloga os ofensores em três tipos: os oportunistas (indivíduos que numa situação real tiram vantagens), os amadores (indivíduos responsáveis por participações falsas e seguros múltiplos mas em número não significativo) e os criminosos profissionais (indivíduos que praticam fraudes organizadas que pertencem a grandes organizações ou redes internacionais criminosas). De acordo com Brites (2006) "conquanto existam cada vez mais casos de grupos organizados, a fraude nos seguros acontece no quotidiano e é cometida por pessoas que normalmente agem de acordo com a lei e que, em muitos casos, até não consideram estar a cometer uma ilegalidade". Face ao exposto, é possível concluir que a fraude no setor dos seguros é uma realidade cada vez mais presente e com impactos tanto para a empresa seguradora como para todos os cidadãos detentores de apólices pelo que se torna importante não só detetá-la mas, numa primeira instância, preveni-la.

Circunscrevendo-nos ao setor dos seguros de saúde, este é mais vulnerável à fraude do que a maioria das outras indústrias devido: à complexidade associada à área da medicina, ao facto dos regulamentos que regem esta área serem ambíguos e estarem em constante mudança e também ao fraco conhecimento em fraude associado às limitações existentes na área da saúde por parte dos investigadores de fraude (Fisher, 2008). Simultaneamente, segundo Sparrow et al (1996), a fraude no setor da saúde agrava-se porque as seguradoras são vistas como alvos socialmente aceitáveis para a fraude, por serem classificadas como ricas, grandes e anónimas. Segundo Evans (in The Financial Cost Of Healthcare Fraud 2014), apesar da fraude ter impactos em todos os setores, no que respeita ao setor dos seguros de saúde o impacto é direto na vida humana, uma vez que, aumenta os tempos de espera por tratamentos, faz com que as pessoas não estejam disponíveis para suportar o tratamento que precisam, impossibilita que todas as pessoas recebam a qualidade da assistência de tratamentos e prevenção que eram possíveis fornecer.

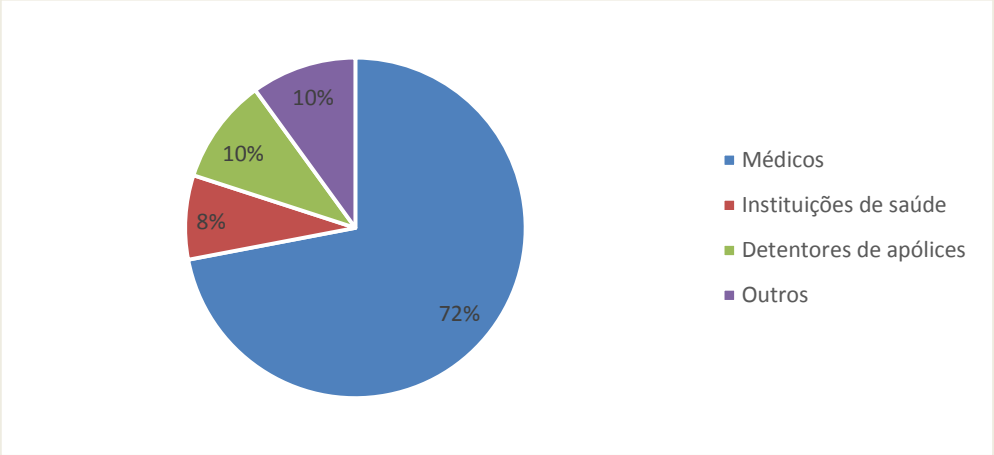
Atendendo ao exposto, a fraude neste setor é, de acordo com o National White Collar Crime Center (2013), uma declaração falsa que o individuo detentor da apólice faz sabendo que este ato poderá resultar em benefícios para o próprio ou para outro. De acordo com Fisher (2008), a fraude nos seguros de saúde "é um fator importante e visível associado com o aumento dos custos de saúde [...] Alguns dos outros fatores que

aumentam os custos, como a melhor tecnologia, têm implicações positivas, mas a fraude dos cuidados de saúde não".

São várias as práticas de fraude neste setor, nomeadamente: cobrar por serviços não prestados, várias prescrições ou excesso de prescrições, prestação de cuidados desnecessários, cobrar serviços não cobertos como cobertos, roubo de identidade e referenciação abusiva/ redes de fraude (Soares, 2008; National White Collar Crime Center, 2013). Assim, e embora a fraude entre médicos e pacientes seja o tipo de fraude mais comum, esta existe em várias áreas relacionadas com o setor dos seguros de saúde (Fisher, 2008).

De acordo com a figura 7, é possível verificar que os responsáveis por comportamentos fraudulentos na área dos seguros de saúde são, não só os médicos (72%) e os detentores das apólices (10%), mas também as instituições de saúde (8%) e outros intervenientes nos seguros de saúde (10%) (Datawatch, 2000 as cited Fisher, 2008).

Figura 7 - Quem comete fraude nos seguros de saúde?
(Fonte: Fisher, 2008)



Neste quadro, aliando a complexidade associada ao tema da fraude nos seguros de saúde ao impacto que tem, a redução da fraude nos seguros é uma prioridade para todas as seguradoras Europeias (CEA, 2014) que pretendem detetar os comportamentos fraudulentos de forma a mitigá-los e diminuir os impactos referidos.

Outro tipo de fraude é a fraude em rede, ou seja, no caso da saúde está relacionada com o facto de prestadores e/ou clientes se associarem entre si para obter algum tipo de proveito. Ao conseguir criar a redes que espelhem as relações existentes entre as entidades é possível retirar conclusões acerca do seu comportamento, desta

forma, quando uma nova entidade se junta à rede pode ser identificado como entidade com maior propensão à fraude se for identificado um comportamento similar ao de uma previamente detetada como fraudulenta (Soares, 2009). De acordo com Joshi (2008), "os atributos de um indivíduo na rede são menos importantes do que as suas relações com outros indivíduos na rede. Explorar a natureza e a força destas ligações pode ajudar a compreender a estrutura e dinâmica de redes sociais e explicar fenómenos reais desde a eficiência organizacional até à propagação de informação".

2.4. DETEÇÃO DE FRAUDE

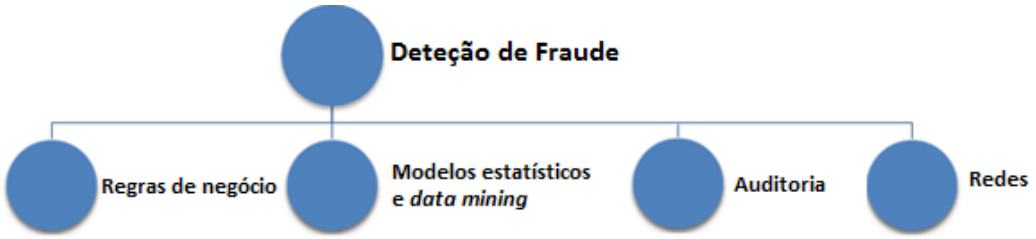
Sendo o objetivo de todas as seguradoras verificar a efetividade e sucesso das iniciativas atuais de combate à fraude e avaliar a necessidade de novas medidas, a forma como a fraude é abordada pelos países difere. Se para alguns países o importante é tratar e reduzir a fraude conhecida para outros é fundamental obter uma estimativa precisa das fraudes detetadas e das fraudes não detetadas (CEA, 2014). Assim, em muitos países, incluindo Portugal, as seguradoras trabalham em conjunto, trocando informações relevantes para a identificação de comportamentos fraudulentos. Mantendo presente o princípio da transparência as seguradoras partilham informações relevantes agindo sempre de acordo com os regulamentos de privacidade e proteção de dados. Para além da partilha interna de informação, em muitos países, nomeadamente em Portugal, as seguradoras têm vindo a aumentar a sua cooperação com os organismos responsáveis por aplicar a lei, ao mesmo tempo que apostam nas tecnologias de deteção de fraude e na formação dos seus colaboradores de forma a aumentar os conhecimentos sobre este tema com o propósito que a identificação seja efetuada numa fase mais precoce e de uma forma mais rápida (CEA, 2014).

A deteção de fraude, por se tratar de uma parte integrante do processo de controlo de fraude automatiza e ajuda na redução dos processos manuais de controlo (Phua et al, 2010). Segundo Bolton e Hand, 2002, antes da aplicação de métodos de deteção de fraude é importante a implementação de métodos de prevenção. Estes métodos têm como objetivo impedir que a fraude ocorra, no entanto, têm associado um custo elevado, como tal será necessário medir a efetividade do método em relação ao custo do mesmo. Assim, os métodos de deteção de fraude surgem apenas quando os métodos de prevenção falham, ou a opção recai sobre a não implementação dos mesmos. Ainda de acordo com os autores, os métodos de deteção têm como objetivo a identificação da fraude da forma mais célere possível, no entanto o desenvolvimento

destes métodos possuem uma dificuldade elevada relacionada com o facto da partilha de conhecimento sobre fraude estar limitada uma vez que a divulgação pública e detalhada destas técnicas poderá fornecer informações relevantes aos indivíduos que tenham como objetivo contornar os sistemas de deteção de fraude.

Apesar das dificuldades apresentadas, existem inúmeros métodos capazes de detetar a fraude sendo que a escolha destes métodos deve basear-se no conjunto de características que cada um apresenta face às necessidades e às particularidades do negócio. Assim, através da observação da figura 8 é possível verificar que alguns métodos utilizados para assertivamente detetar comportamentos fraudulentos são: técnicas estatísticas e *data mining*, definição de regras de negócio, auditorias e utilização de redes (redes de influências/ sociais). Cada um destes métodos vai ser analisado com detalhe neste capítulo.

Figura 8 - Métodos de deteção de fraude
(Fonte: Elaboração própria)



2.4.1. Regras de Negócio

As regras de negócio são um ativo bastante importante numa organização, assim, são muitas as empresas que fazem grandes investimentos na codificação das suas regras de negócio para que estas sejam contempladas nos seus sistemas de informação internos (Shao & Pound, 1999). Dado que as regras de negócio são diretrizes e limitações aos estados e processos de uma organização estas podem ser aplicadas para definir e cumprir objetivos, tratar dados, garantir a aplicação de regulamentações e leis, entre outros (Herbest, 1996).

A utilização de regras de negócio para deteção de fraude é bastante importante na medida em para além de servirem para representar os requisitos do utilizador e as condições dos processo internos, são essenciais para acompanhar o crescimento das empresas tornando assim o seu *software* útil e atualizado (Wan-Kadir & Loucopoulos, 2004).

Assim, no ambiente de alterações constantes da realidade empresarial a evolução e mudança do *software* é inevitável para que este seja capaz de espelhar a realidade do negócio. Face ao exposto as regras de negócio deverão ser atualizadas aquando de alterações políticas ou operacionais e deverão ser propagadas para o *software* (Wan-Kadir & Loucopoulos, 2004).

Abaixo serão referidos alguns exemplos significativos de regras de negócio utilizadas na área dos seguros de saúde que, pela sua essência, têm como principal objetivo a seleção de dados capazes de identificar quais os elementos potencialmente fraudulentos. As regras mencionadas, e de forma a melhor dirigirem o seu foco, estão divididas em três pilares: pacientes, profissionais de saúde e serviços hospitalares (hospitais, centros de saúde, etc.).

2.4.1.1. Regras aplicadas aos dados dos pacientes (detentores da apólice)

Algumas regras aplicadas pela seguradora em causa aos dados respeitantes aos detentores de apólice são:

- O paciente percorreu uma distância elevada num só dia;
- O paciente tem demasiadas horas de serviços hospitalares num só dia;
- O paciente é um *outlier* (em relação aos seus pares) no que respeita ao dinheiro permitido para sinistros para um determinado ano;
- O paciente é um *outlier* (em relação aos seus pares) no que respeita ao dinheiro total pago pela seguradora num determinado ano;
- O paciente é um *outlier* (em relação aos seus pares) no que respeita ao número de sinistros permitidos para um determinado ano;
- O paciente é um *outlier* (em relação aos seus pares) no que respeita ao usufruto do número de serviços de saúde autorizados pela seguradora num determinado ano;
- O paciente apresenta várias receitas com múltiplos prescritores para a mesma substância;
- O paciente tem um elevado número de sinistros num mês e apresenta mais diagnósticos raros que o normal;
- O paciente têm um elevado número de sinistros com uma idade que não corresponde com a idade normal de diagnóstico daquela doença;
- O paciente apresenta vários sinistros em hospitais com uma data de nascimento distinta;

2.4.1.2. Regras aplicadas aos dados dos profissionais de saúde

Algumas regras aplicadas aos dados dos profissionais de saúde de forma a selecionar e identificar dados são:

- O profissional de saúde é um *outlier* (em relação aos seus pares) no que respeita ao número de serviços fornecidos para um sinistro comparativamente aos permitidos pela seguradora para um determinado mês;
- O profissional de saúde informa a presença de um paciente no seu consultório no dia em que o paciente deveria estar no hospital;
- O médico pediatra trata pacientes mais velhos;
- O profissional de saúde reporta uma idade do paciente diferente da idade real;
- O profissional de saúde coloca um código de diagnóstico errado num elevado número de sinistros (quando comparado com o considerado normal) durante um determinado mês;
- O profissional de saúde faz demasiadas prescrições num curto período de tempo;
- O profissional de saúde reporta diagnósticos primários diferentes do previsto;
- O profissional de saúde faz/receita ao mesmo paciente o mesmo tratamento durante um curto período de tempo;
- Dados duplicados (pacientes, procedimentos, datas de serviço, etc, iguais) submetidos por dois profissionais de saúde diferentes;
- O profissional de saúde faz tratamentos em membros (pernas, braços, etc.) anteriormente removidos;

2.4.1.3. Regras aplicadas aos dados dos serviços de saúde (hospitais, centros de saúde, entre outros)

Algumas regras aplicadas pela seguradora em causa aos dados relacionados com os serviços de saúde (hospitais, centros de saúde, entre outros) são:

- Demasiados códigos de diagnósticos inválidos;

- O serviço de saúde é um *outlier* (em relação aos seus pares) no que respeita à variação (%) nos serviços cobrados aos pacientes num determinado mês em relação aos 6 meses anteriores;
- Muitos sinistros repetidos e ignorados por dois ou mais funcionários;
- O serviço de saúde é um *outlier* (em relação aos seus pares) no que respeita à quantia autorizada para pagamento de sinistros para um determinado mês;
- Referências a procedimentos não encontradas;
- Sinistros onde o género do detentor da apólice se alterou em relação ao sinistro anterior;
- Demasiadas datas de serviços incorretas;
- Serviços de saúde prestados para vários detentores de apólices da mesma família no mesmo local num curto período de tempo;
- O serviço de saúde é um *outlier* (em relação aos seus pares) no que respeita ao número de serviços autorizados para um determinado mês;
- Sinistros onde a data de nascimento do detentor da apólice se alterou em relação ao sinistro anterior;

2.4.2. Modelos estatísticos e *Data Mining*

Existem ferramentas estatísticas para deteção de fraude adaptáveis a diferentes tipos de dados que têm como denominador comum o facto de se basearem na comparação dos dados observados em relação aos dados esperados. Este tipo de análise estatística permite detetar uma observação fora do normal que provavelmente poderá ter maior propensão para se verificar como um registo fraudulento, conseguindo-se assim fazer um *ranking* de probabilidade de fraude onde um maior *score* indica um maior distanciamento do comportamento normal, ou seja, uma maior score identifica os registos que necessitam de ser investigados em detalhe (Bolton & Hand, 2002).

No que respeita à utilização de *data mining* para a deteção de fraude, é importante referir que este método obriga à existência de um problema bem definido baseado em modelos processuais e que não seja resolúvel com ferramentas de *query* e *reporting* (Lavrac et al, 2004). Esta técnica tem como objetivo encontrar factos desconhecidos com fundamento estatístico e despoletado a partir dos dados (Elkan, 2001), assim, a utilização do *data mining* justifica-se por ser financeiramente mais eficaz encontrando provas de fraude através da aplicação de algoritmos matemáticos nos

dados disponíveis (Phuaet al, 2010). Ao mesmo tempo, visto que muitas empresas não são financeiramente capazes de avaliar manualmente milhões de interações com entidades externas e internas a utilização do *data mining* serve para identificar as interações de forma a que apenas as interações suspeitas possam ser investigadas (Phua et al, 2010). De acordo com Cahill et all (2002, pp. 911-930) , todas as atividades da empresa deverão ser avaliadas e pontuadas através da comparação e identificação de similaridades com atividades fraudulentas.

Os métodos de *data mining* utilizados para a determinação das atividades suspeitas incluem regressões, *clusters*, algoritmos supervisionados individuais e múltiplos, algoritmos supervisionados e não supervisionados (Sherman, 2002 as cited Phua et al, 2010). As árvores de decisão, indução de regras, *case base reasoning* (Fan, 2004), redes neurais, sistemas de deteção de fraude de duas etapas baseados em regras (Rosset et al, 1999) e modelação estatística (Foster & Stine, 2004) são outros métodos que utilizam o *data mining* para a deteção de fraude.

Assim, o *data mining* é composto por inúmeras técnicas, sendo que cada uma apresenta as suas vantagens e desvantagens pelo que a escolha deve basear-se nas necessidades do negócio e nos dados disponíveis visto que nenhuma das tipologias consegue cobrir todas as situações. Desta forma poderão utilizar-se combinações das várias técnicas de *data mining* de forma a conseguir resultados mais eficazes (Phua et al, 2010).

Bolton e Hand, 2002, defendem que os métodos de deteção de fraude podem ser divididos em dois grupos: supervisionados e não-supervisionados. O primeiro grupo caracteriza-se por se basear em observações fraudulentas e não fraudulentas de forma a construir um modelo que disponibilizará *scores* de suspeita para novos casos. Ou seja, numa abordagem analítica utilizam-se os métodos supervisionados quando o alvo conhecido está disponível (fraude) e quando existe histórico da informação comportamental de comportamentos fraudulentos de forma a identificar comportamentos similares aos padrões de fraude anteriores. Este tipo de métodos inclui, entre outros, as redes neuronais, as árvores de decisão e os modelos preditivos (paramétricos e não paramétricos).

Relativamente aos métodos supervisionados destacam-se as redes neuronais que pretendem simular o cérebro humano, assim, consistem numa estrutura computacional baseada em unidades de processamento (neurónios) organizados em camadas que aprendem através da experiência (Dinn, 1998 as cited Lemos, 2003). Cada neurónio

processa *inputs* e *outputs* associados pesos. O cálculo do número de neurónios bem como os pesos de cada ligação são descobertos através de redes (Azevedo, 2011).

No que respeita à utilização de regressões pode definir-se como uma técnica de análise de informação onde se vai estabelecer uma relação entre as variáveis (Robinson & Officer, [Aspire][5]). Segundo Azevedo (2011) esta técnica pode ser dividida em regressões lineares e regressões não-lineares. Assim, define as regressões lineares como uma técnica com o objetivo de descobrir uma função linear capaz de retratar, o mais próximo possível da realidade, o comportamento de variáveis numéricas (variáveis dependentes) como combinações lineares das outras variáveis (variáveis independentes). No que respeita às regressões não-lineares esta autora define-as como bastante semelhantes às regressões lineares no entanto, ao invés de utilizarem a combinação linear das variáveis independentes para expressar o comportamento da variável dependente utilizam uma combinação não-linear das variáveis independentes para obter o comportamento da variável dependente.

Assim, a técnica de regressão é uma técnica de análise de informação onde se vai estabelecer uma relação entre as variáveis (Robinson & Officer, [Aspire][6]).

Em relação às árvores de decisão caracterizam-se por serem modelos sequenciais que combinam logicamente uma sequência de testes simples (Kotsiantis, 2011), ou seja, esta técnica de *data mining* pode ser definida como uma forma de representação de regras, tendo em conta a hierarquia que segue (Azevedo, 2011). Assim utilizam uma estratégia de dividir-para-conquistar, isto é, "um problema complexo é descomposto em problemas mais simples" (Gama, 2002).

O segundo grupo de métodos são os não-supervisionados, ou seja, são métodos que são utilizados quando não existe informação prévia sobre observações fraudulentas e não-fraudulentas. Neste segundo grupo de métodos procura-se encontrar os *outliers* tendo por base a normalidade de um universo de observações. Neste quadro, utilizam-se estes métodos quando não existe um alvo e como tal vai-se proceder à observação do comportamento atual para identificar comportamentos anormais e diferentes das transações expectáveis. Estes métodos incluem, entre outros, técnicas como análises uni e multivariadas, comparação de pares e análises de tendências.

[5] Data Mining: Predicting Laptop Retail Price Using Regression. http://www.spelman.edu/docs/aspire-research/joibritneypdf.pdf?sfvrsn=2, Acedido em 07-07-2014

[6] Data Mining: Predicting Laptop Retail Price Using Regression. http://www.spelman.edu/docs/aspire-research/joibritneypdf.pdf?sfvrsn=2. Acedido em 07-07-2014

Por último é ainda importante descrever a técnica de *clustering* que se define por ser uma técnica que faz a divisão dos dados em grupos de objetos semelhantes. Assim, caracteriza-se por ser uma técnica que vai organizar os dados em grupos de dados idênticos para uma maior simplificação dos dados (perdendo assim algum detalhe) (Berkin, 2006), ou seja, a análise de *cluster* vai dividir os dados em grupos úteis e significativos de forma a capturar o "normal" dentro daquele grupo de dados (Kumar, 2000).

Assim, o *data mining* através da utilização de algoritmos específicos identifica padrões e casos onde se suspeite da existência de fraude. Inicialmente é necessário codificar e implementar as regras de negócio nos algoritmos de modo a serem identificados potenciais casos de fraude, padrões, sequências de eventos e comportamentos que não seguem o padrão normal para prevenir possíveis perdas (Ferreira, 2009).

A escolha da técnica de *data mining* a utilizar está mais depende de temas de requisitos operacionais, restrições de recursos e estratégia relativa a redução do que propriamente dos dados disponíveis (Phua et al, 2010). Assim, o recurso a tecnologia de análise de dados como o *data mining*, permite a investigação aos dados de negócio de uma organização para que seja possível identificar comportamentos que indiciem atividades fraudulentas. Estas tecnologias possibilitam ainda tornar a organização mais proactiva no combate à fraude dado que fazem com que sejam detetados indicadores de atividade fraudulenta de forma bastante mais rápida permitindo que esta transação seja terminada antes de ser efetivamente realizada, evitando o impacto financeiro (ACL, 2010).

Face ao exposto as duas maiores críticas às pesquisas sobre deteção de fraude baseada em *data mining* são a escassez de dados reais públicos disponíveis para testes e a falta de publicações sobre métodos e técnicas bem fundamentadas. Para combater as críticas apresentadas muitos dos departamentos de fraude optam por colocar valor monetário nas previsões de forma a maximizar os lucros de acordo com as suas políticas (Phua et al, 2010), isto é, ao colocarem um valor nas previsões estas terão uma maior foco nas fraudes de maior impacto financeiro, visto estarem adaptadas às diretrizes internas, e consequentemente será possível maximizar lucro (ou minimizar perdas). Outras das medidas tomadas pelos departamentos de fraude passam pela definição de custos explícitos (Phua et al, 2004; Chan et al, 1999; Fawcett & Provost, 1997) ou modelos de benefícios (Fan, 2004; Wang et al, 2003). A utilização destas medidas é justificada pelo facto dos custos de classificações incorretas (falsos positivos e falsos

negativos) serem desiguais, incertos e poderem diferir ao longo do tempo (Phua et al, 2010).

2.4.3. Auditoria

Em relação à deteção de fraude através da auditoria é importante clarificar a abordagem de Townsend, "The Costly Verification" (1979). Esta abordagem foca-se nas situações onde a empresa seguradora é capaz de verificar relatórios do agente de seguros através de auditorias e tem como princípio a completa honestidade dos agentes nos relatórios disponibilizados. A aplicação deste princípio torna-se complexo dado que existe um maior incentivo a atos fraudulentos por parte dos agentes de seguros. Assim, torna-se imperativo a execução de auditorias de forma a validar a veracidade de cada um dos relatórios disponibilizados à empresa seguradora. Uma vez que os processos de auditoria obrigam a um significativo esforço a nível financeiro para as seguradoras, torna-se indispensável a existência de uma triagem efetiva que indique quais as atividades mais propensas a fraude, para que as seguradoras possam ser mais assertivas na decisão entre realizar ou não o processo de auditoria. Um sistema de deteção de fraude só terá benefícios financeiros se a diminuição de fraude e de execução de auditorias compensar o custo deste sistema, ou seja, só fará sentido implementar num mercado de dimensão considerável (Schiller, 2006).

2.4.4. Redes

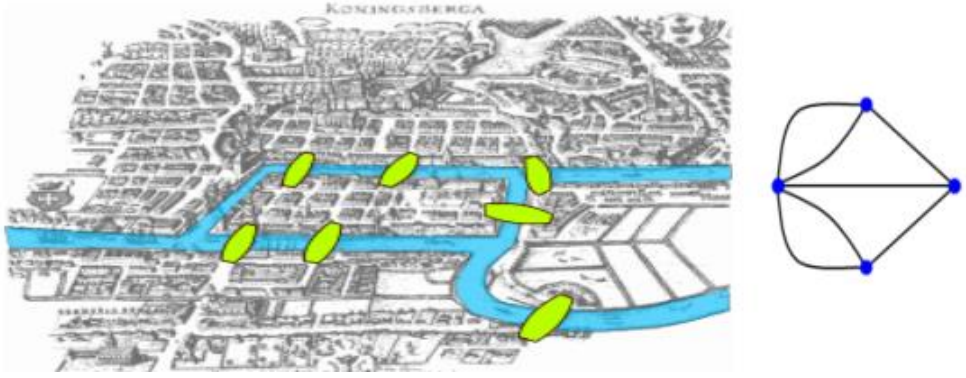
Por vivermos, conforme afirma Castells (1999), numa sociedade em rede este tema tem-se tornado cada vez mais relevante e atual. "As relações entre as partes de sistemas, os sistemas abertos, imprevisibilidade, não linearidade, auto organização, adaptabilidade, criatividade, instabilidade, emergência, incerteza, conectividade e fluxo são algumas das marcas da contemporaneidade, onde a informação e as redes emergem como elementos catalisadores da realidade" (Ferreira, 2012).

A adaptação desta abordagem aos problemas complexos prende-se sobretudo com duas características destes sistemas: a interdependência (o comportamento de cada elemento está dependente do comportamentos dos restantes elementos da rede) e a emergência (diferentes tipos de agregação desde o mais baixo – que se relaciona com os elementos da rede – ao mais alto – que assenta em conjuntos de elementos ou sub-redes da rede global) (Araújo, 2011, pp. 157-158).

Embora seja um tema bastante atual, as redes não são uma nova forma de organização social, isto porque, a evolução tecnológica forneceu novas capacidades a esta velha forma de organização social tornando-a na forma de organização mais flexível e adaptável (Castells & Cardoso, 2005). A evolução das redes passa sobretudo pelo facto de antigamente as redes serem interpretadas como objetos estáticos representativos de uma estrutura, e de atualmente as redes serem vistas como objetos dinâmicos em constante evolução e mudança de acordo com a passagem do tempo (Watts, 2003). Assim, podemos definir a rede como uma estrutura aberta que cresce através da adição ou remoção de nós consoante as mudanças necessárias, isto é, as redes são sistemas de nós interligados (Castells & Cardoso, 2005). É importante destacar que, "geralmente as redes organizam-se para cumprir finalidades. (…) Consequentemente pode pensar-se que é a função da rede que determina a sua forma." (Araújo, 2011, p. 192).

O primeiro conceito de rede remonta ao início do século XVIII, com Leornard Euler e o problema das Sete Pontes de Köningsberg (Araújo, 2006). Quando em 1735 Euler resolveu a problemática das Sete Pontes e determinou que não era possível começar e terminar um percurso pela cidade de Köningsberg sem repetir nenhuma das Sete Pontes, deu início à teoria dos Grafos. Euler afirmou que este problema não era possível de ser ultrapassada e provou-o através da eliminação de tudo o que era acessório, fazendo uma representação com recurso a pontos (que representavam as partes da ilha) e linhas (representavam as pontes que uniam as diferentes partes da ilha), ou seja, Euler representou este problema numa rede (Matos, 2013).

Figura 9 – Teoria dos Grafos
(Fonte: Araújo, 2006)

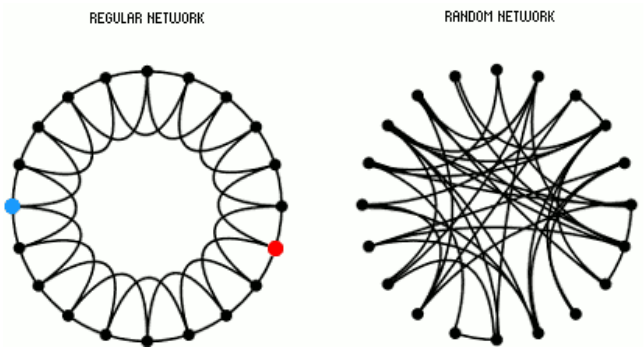


Através da representação que Euler apresentou (figura 9) foi possível perceber a importância da representação topológica, que ao contrário da representação geográfica

não que se preocupa não com a forma e posição exata dos objetos mas sim com as ligações estabelecidas entre os mesmos. A representação topológica, aplicada ao problema das sete pontes permite "restringir a consideração do problema à observação de uma rede com quadros nós e sete ligações" (Araújo, 2011, p. 160).

Os primeiros estudos da teoria dos grafos e das suas propriedades pertencem aos matemáticos Paul Erdös e Alfred Rényi destacando-se as teorias destes matemáticos acerca das redes aleatórias (Barabási, 2002). Definem-se assim duas classes de redes completamente distintas: as redes regulares e as redes aleatórias (figura 10).

Figura 10 - Redes regulares e redes aleatórias
(Fonte: http://cftc.cii.fc.ul.pt/PRISMA/capitulos/capitulo5/modulo2/topico3.php, Acedido em 27 Junho de 2014)



Pela análise figura 10 é possível perceber que as redes regulares são aquelas cujas ligações seguem um determinado padrão, pelo contrário as redes aleatórias não apresentam qualquer padrão nas suas ligações. Para além da sua construção estes dois tipos de rede distinguem-se pelo valor do coeficiente de agregação (*clustering)* e pelo caminho médio mais curto (*charecteristic path length*) entre cada dois elementos da rede. Através dos coeficientes é possível calcular a probabilidade de, para cada elemento da rede, os elementos a que está ligado estarem ligados entre si (*clustering*) e verificar o número médio de ligações necessárias para unir cada par de elementos da rede. (Araújo, 2006).

Ainda no contexto das redes aleatórias, são dois os algoritmos capazes de ensinar uma rede com ligações criadas de forma aleatória a desempenhar uma mesma função. O primeiro algoritmo é o RLM – algoritmo de aprendizagem por reforço nas ligações - que se caracteriza por um reforço ou enfraquecimento das ligações de todos os nós ativos (tendo o mesmo impacto em todos os nós). O segundo algoritmo é o LFM - algoritmo de aprendizagem através de erros – que tem como principal particularidade

um processo de enfraquecimento das ligações entre os nós que produzem um erro. (Araújo, 2011, pp. 197-198).

Surgiu ainda uma nova classe de redes quando se verificou que algumas redes apresentavam características determinadas como impossíveis de conciliar numa mesma estrutura. Esta nova estrutura é denominada *Small World* (SW) e caracteriza-se por apresentar um valor baixo no coeficiente do caminho médio. (Watts, 1999). Assim, as redes SW são redes ubíquas que apresentam, conjuntamente conectividade local (justificada pelo elevado *clustering*) e global (fundamentada pelo curto caminho médio) (Araújo, 2011, p. 169).

Em 1999, Barabási et al iniciaram a investigação noutra categoria de redes, as redes *Scale Free*, caracterizadas pela distribuição dos graus dos nós que seguem uma lei de escala com expoentes característicos identificados por através da experimentação. Esta categoria de redes é mais favorável à representação de processos evolutivos, isto porque, ao contrário das restantes categorias apresentadas, permite uma heterogeneidade do grau dos nós.

A utilização das redes para deteção de fraude pressupõe a utilização de um algoritmo apropriado para a situação que se pretende tratar, isto porque, o objetivo é encontrar algoritmos eficientes capazes de resolver o problema em questão, ou que demonstrem que não existe nenhum algoritmo capaz de solucionar o problema em causa (Even, 2011, p. 10).

É importante destacar alguns algoritmos utilizados em redes de forma a perceber melhor as suas características e capacidades. Para uma melhor compreensão dos algoritmos estes vão ser expostos de acordo com o problema computacional que resolvem. É ainda importante referir, que na descrição dos algoritmos as ligações entre os nós da rede vão ser definidas como arestas, ligações ou limites e os nós vão ser também chamados como vértices.

2.4.4.1. Análise de Redes

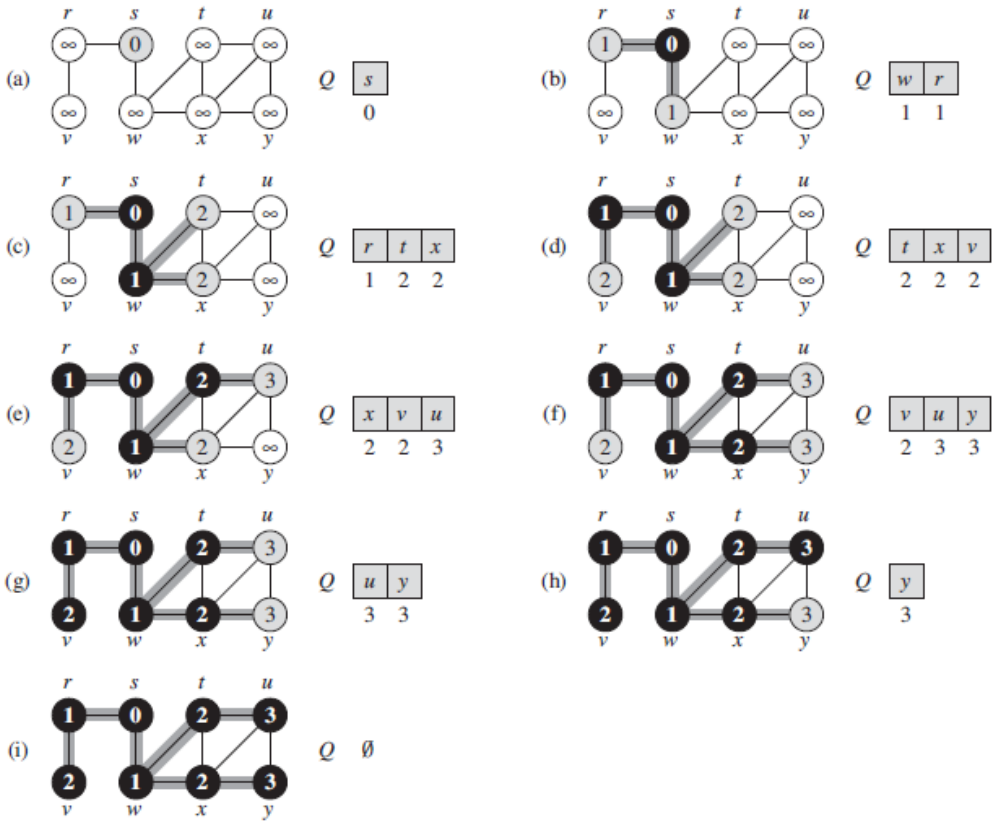
Para analisar uma rede é necessário seguir, sistemáticamente, as suas arestas, bem como prestar especial atenção aos seus nós, assim, a utilização de algoritmos de análise de redes permite conhecer melhor a estrutura da mesma (Cormen et al, 2009, p.589).

O algoritmo de *breadth-first search* é um dos algoritmos mais simples para análise de redes e consequentemente é a base de muitos algoritmos (Cormen et al, 2009, pp.

595-601; Kocay & Kreher, 2004). De acordo com os mesmos autores, a execução deste algoritmo passa pela exploração sistemática dos limites da rede para descobrir todos os nós acessíveis a partir do nó fonte, ou seja, o algoritmo vai atribuindo pesos aos nós consoante os limites que é necessário percorrer para que estes fiquem todos abrangidos (se um nó estiver diretamente ligado ao nó fonte [nó 0], então vai ser um nó com peso de 1, no entanto se for um nó que está não está diretamente ligado ao nó fonte e para o qual é necessário passar por um nó intermédio então vai ter um peso de 2). Pela observação da figura 11, é possível perceber a execução do algoritmo:

1. Encontrar os nós com ligações ao nó fonte e atribuir um peso de 1;
2. Encontrar os nós com ligações aos nós abrangidos no ponto 2 e atribuir um peso de 1;
3. Repetir os passos referidos tendo em atenção que cada nó só deverá ser abrangido uma única vez;

Figura 11 - Algoritmo de breadth-first search
(Fonte: Cormen et al, 2009, p. 596)



2.4.4.2. Problema da abrangência mínima

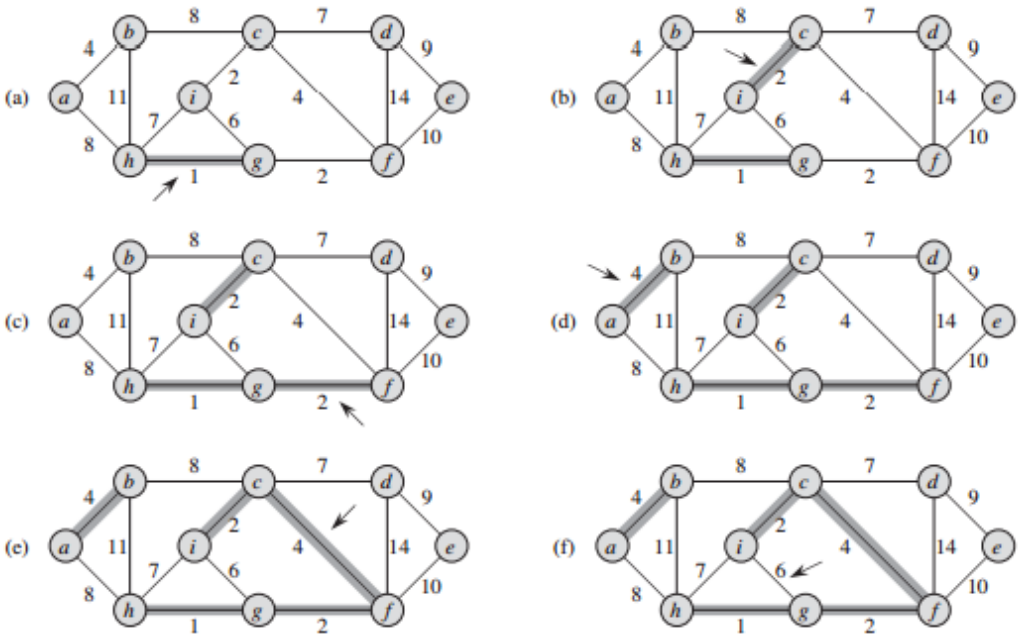
O problema da abrangência mínima (selecionar o caminho com menor peso de uma rede com pesos associados às ligações dos nós) é outro problema resolúvel por algoritmos, nomeadamente os algoritmos de Kruskal e Prim (Cormen et al, 2009, pp. 624-625).

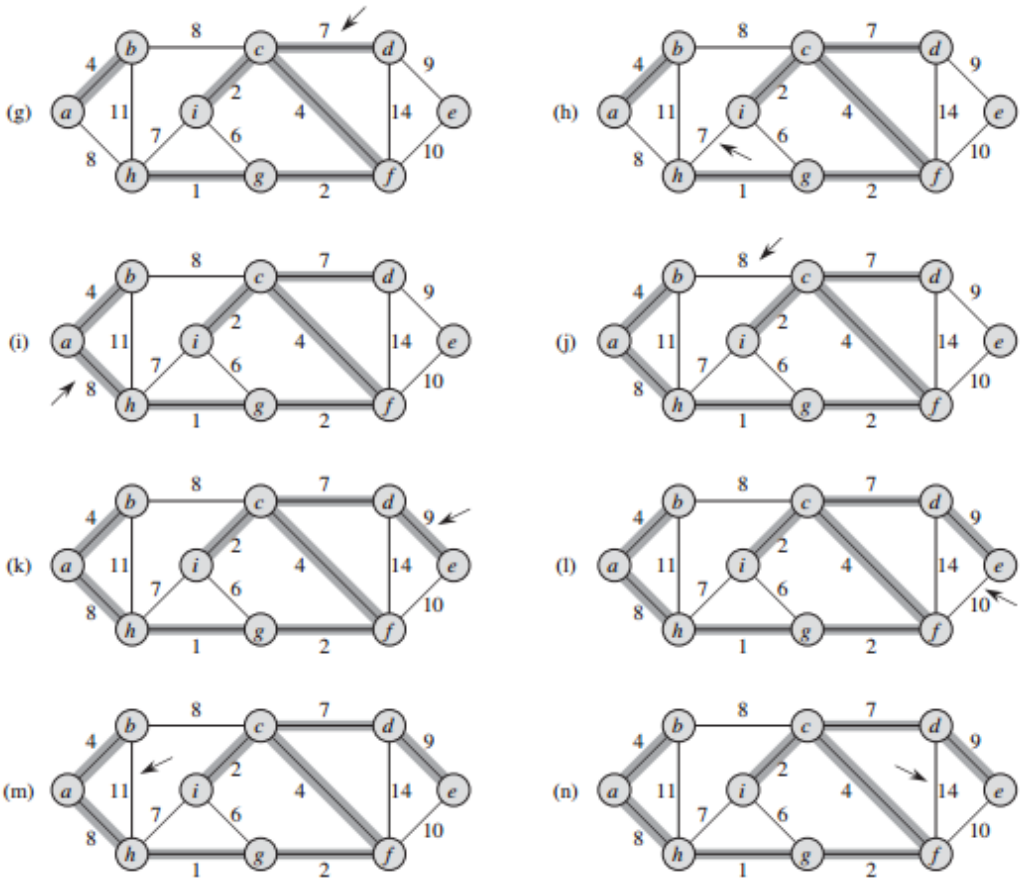
O algoritmo de Kruskal encontra qual a abrangência, partindo de qualquer nó da rede, tendo em conta que a primeira ligação a selecionar deverá ser a mais curta (ou seja com menor peso) (Cormen et al, 2009, pp. 631-633; Kamil, 2003). Assim, através da observação da figura 12, é possível verificar qual o comportamento deste algoritmo:

1. Seleção da aresta com menos peso (H-G);
2. Seleção da segunda aresta com menos peso (I-C);
3. Repetição dos passos anteriores até que todos os nós estejam abrangidos. Se a aresta mais curta subsequente pertencer a um dos nós já abrangidos então deverá passar-se para a próxima ligação com menor peso que contemple nós ainda não abrangidos (situação retratada na figura 12 na rede (f) e (g));

Figura 12 - Algoritmo de Kruskal

(Fonte: Cormen et al, 2009, pp. 632-633)



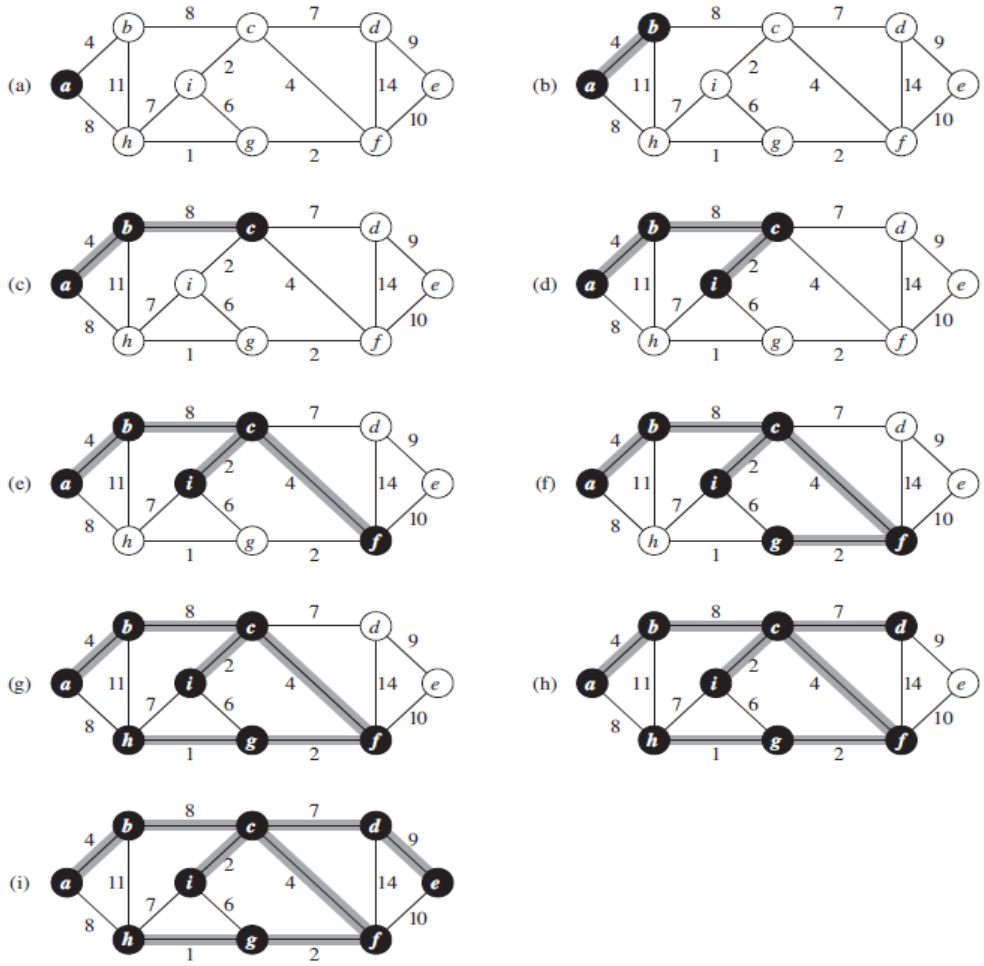


No que respeita ao algoritmo de Prim, tal como o algoritmo de Kruskal, também permite representar o problema da abrangência mínima, no entanto, este algoritmo parte de um nó aleatório (Cormen et al, 2009, pp. 634 – 646; Kamil, 2003; Grama et al, 2003). Na figura 13 é possível perceber o método de execução do algoritmo de Prim, nomeadamente:

1. Seleção aleatória de um nó de partida (A);
2. Seleção da aresta com menor peso do nó de partida;
3. Seleção da aresta com menos peso do nó abrangido no ponto 2;
4. Repetição dos passos anteriores tendo em conta que a rede não deverá nunca fechar-se, ou seja, se a o nó abrangido só possuir ligações a nós já abrangidos deverá voltar-se ao nó anterior e selecionar a segunda aresta com menos peso (situação representada nas redes (d) e (e) da figura 13);

Figura 13 - Algoritmo de Prim

(Fonte: Cormen et al, 2009, p. 635)



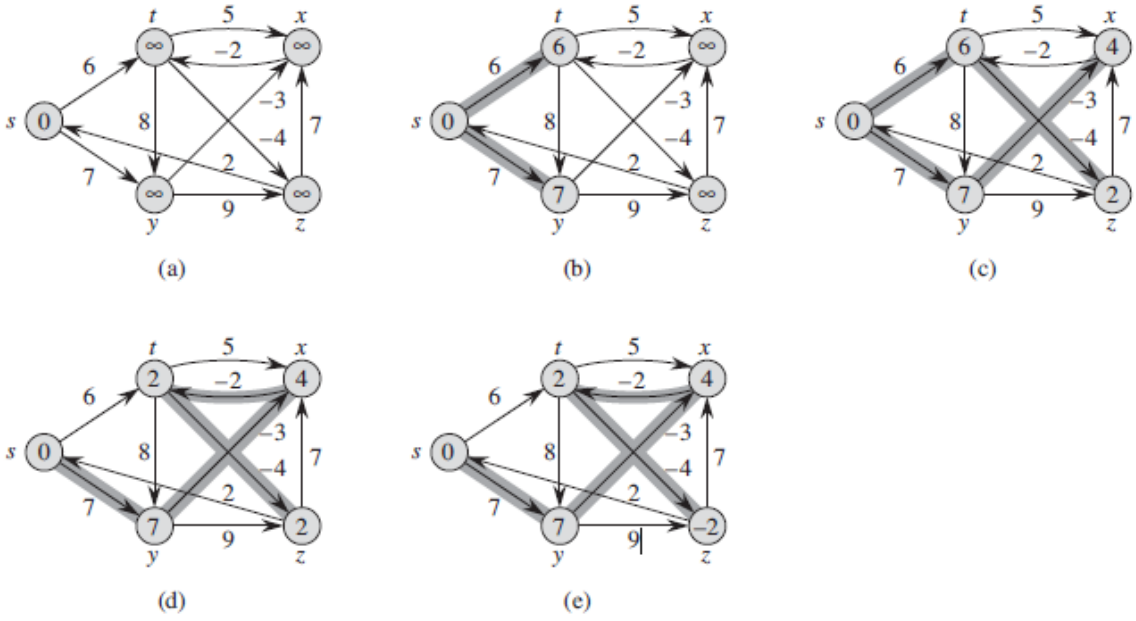
2.4.4.3. Problema do caminho mais curto partindo de um único nó fonte

O problema do caminho mais curto (caminho que inclua todos os nós da rede e cuja soma dos pesos das ligações seja o menor valor possível) é outro problema relacionado com redes capaz de ser solucionado através da utilização de algoritmos, nomeadamente, o algoritmo de Bellman-Ford e Dijkstra (Cormen et al, 2009, pp. 643-644).

O algoritmo de Bellman-Ford foi desenvolvido com o objetivo de resolver o problema do caminho mais curto (com uma única fonte) e a sua execução passa por numa rede com pesos indicar a existência de um ciclo com pesos negativos através de um valor booleano (se existir então o algoritmo retorna que não existe solução possível se não existir então o algoritmo vai indicar qual o caminho mais curto e os seus pesos) (Cormen et al, 2009, pp. 651-654). Assim, e pela observação da figura 14 é possível perceber melhor o método seguido por este algoritmo, a saber:

1. Partindo do nó fonte (s) verificar qual o caminho mais curto abrangendo todos os nós;
2. Verificar se existe de um ciclo com pesos negativos e retornar o valor booleano correspondente (*TRUE* se não existem um ciclo negativo e *FALSE* se existir). É importante destacar que os nós vão ficando com a soma dos pesos das arestas por onde se vai passando (exemplo: o nó fonte começa com 0 e os segundo nós a serem abrangido têm valor 6 e 7 porque são os pesos das arestas do nó fonte, no entanto, se ao longo do caminho este nó for atingido o seu valor altera para que fique com o valor atual da soma dos pesos [o nó que na rede (b) tinha valor 6, na rede (d) fica com valor 2 porque o valor da soma dos pesos que abrange aquele nó alterou [valor do peso S-T representado na rede (b) + valor do peso X-T representado na rede (c) ⇔ 6 -2 = 4]);

Figura 14 - Algoritmo de Bellman-Ford
(Fonte: Cormen et al, 2009, p. 652)



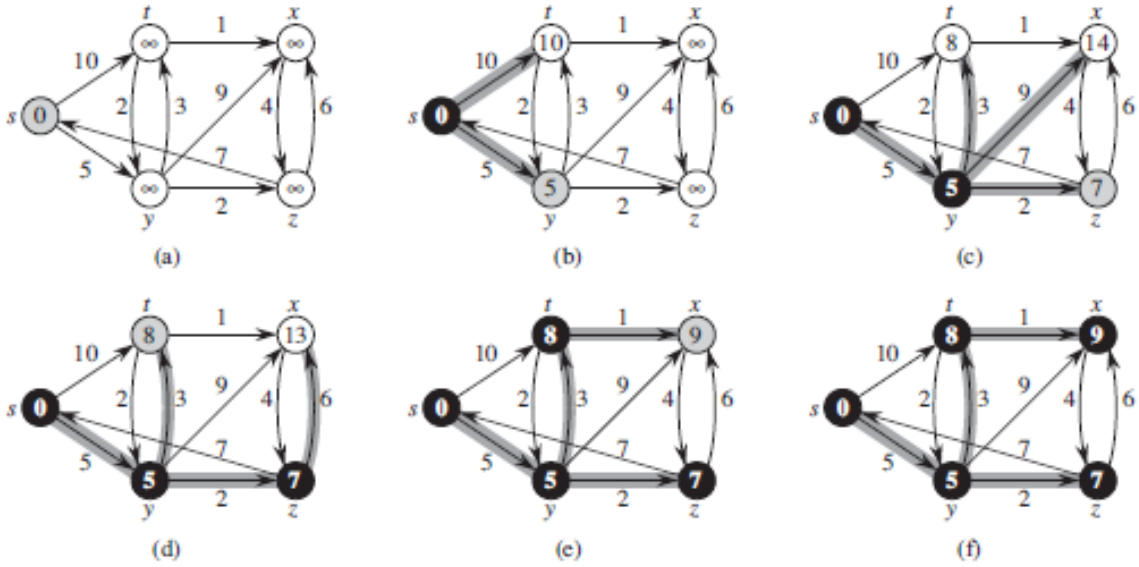
Outro algoritmo utilizado para a resolução do problema do caminho mais curto é o algoritmo de Dijkstra, o que distingue este algoritmo do algoritmo de Bellman-Ford é a particularidade de que para a execução do algoritmo de Dijkstra é necessário que todas as ligações dos nós sejam positivas, ou seja, não pode existir nenhuma aresta com valor negativo (Cormen et al, 2009, pp. 658-659; Kamil, 2003; Grama et al, 2003;

Johnson, 1973). Através da figura 15 é possível perceber com maior precisão a execução do algoritmo de Dijkstra:

1. Verificar qual das arestas do nó fonte tem menor valor (S-Y);
2. Verificar qual das arestas do nó incluído no passo 1 tem menor peso (Y-X);
3. Repetir os passos anteriores até que todos os nós estejam compreendidos na rede. É importante destacar que se a aresta mais curta a selecionar fechar o caminho e ainda não estiverem todos os nós abrangidos deve escolher-se a segunda aresta mais curta do nó anterior para que o caminho não fique fechado antes de contemplar todos os nós (situação verificada na rede (d));

Figura 15 - Algoritmo de Dijkstra

(Fonte: Cormen et al, 2009, p. 659)



2.4.4.4. Problema do caminho mais curto entre todos os pares de nós

O algoritmo de Floyd-Warshall e o algoritmo de Johnson foram desenvolvidos com o intuito de encontrar o caminho mais curto entre todos os pares de nós (Cormen et al, 2009, pp. 684-686).

O algoritmo de Floyd-Warshall vai encontrar através de várias instâncias (representadas em matrizes) os caminhos mais curtos entre todos os nós (quer sejam pesos negativos quer sejam pesos positivos mas assumindo sempre que não existem ciclos negativos) de uma rede colocando essas distâncias em matriz (representando a

distância com NIL se for a distância entre o próprio nó ou se não existir ligação direta entre os nós) (Cormen et al, 2009, pp. 693-696; Grama et al, 2003).

Através da figura 17 é possível compreender a execução deste algoritmo, nomeadamente:

1. Fazer a primeira matriz de distâncias;
2. Congelar a primeira linha e a primeira coluna da matriz de distâncias e verificar se algum dos valores que não está congelado (valores dentro do retângulo a azul) pode ser melhorado:
    a. Iniciando no primeiro valor que se pode melhorar (sublinhado a amarelo): Somar o primeiro valor da linha do valor que se está a avaliar com o primeiro valor da coluna do valor que se está a avaliar e verificar se é um valor mais baixo do que o atual, ou seja, NIL + 1 < NIL? Não. Como o valor não é mais baixo, então mantém-se o valor original;
    b. Segundo valor a avaliar (sublinhado a cor-de-rosa): O procedimento é exatamente igual ao descrito em a., ou seja, NIL + 1 < 3? Não. Como o valor não é mais baixo, então mantém-se o valor original;
    c. Terceiro valor a avaliar (sublinhado a cor-de-laranja): O procedimento é igual ao descrito em a., ou seja, 4+1 < NIL? Sim. Como o valor é menor, então vai ser substituído com o número da instância, neste caso 1;
    d. Repetição do procedimento descrito em a. para todos os valores que estão dentro do retângulo a azul;

Figura 16 - Algoritmo de Floyd-Warshall: Matrizes
(Fonte: Cormen et al, 2009, p. 696)

$$\Pi^{(0)} = \begin{pmatrix} \text{NIL} & 1 & 1 & \text{NIL} & 1 \\ \text{NIL} & \text{NIL} & \text{NIL} & 2 & 2 \\ \text{NIL} & 3 & \text{NIL} & \text{NIL} & \text{NIL} \\ 4 & \text{NIL} & 4 & \text{NIL} & \text{NIL} \\ \text{NIL} & \text{NIL} & \text{NIL} & 5 & \text{NIL} \end{pmatrix}$$

$$\Pi^{(1)} = \begin{pmatrix} \text{NIL} & 1 & 1 & \text{NIL} & 1 \\ \text{NIL} & \text{NIL} & \text{NIL} & 2 & 2 \\ \text{NIL} & 3 & \text{NIL} & \text{NIL} & \text{NIL} \\ 4 & 1 & 4 & \text{NIL} & 1 \\ \text{NIL} & \text{NIL} & \text{NIL} & 5 & \text{NIL} \end{pmatrix}$$

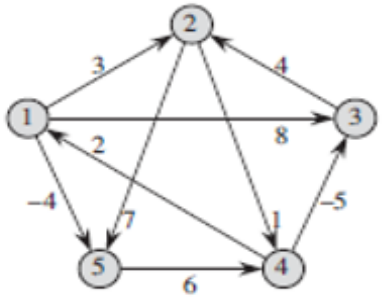
3. Congelar a segunda linha e a segunda coluna da matriz e verificar se algum dos valores que não está congelado pode ser melhorado (através do procedimento

descrito em 2. a.). Se o valor puder ser melhorado colocar o valor 2 (por se tratar da segunda instância);

4. Congelar a terceira linha e a terceira coluna da matriz e verificar se algum dos valores que não está congelado pode ser melhorado (através do procedimento descrito em 2. a.). Se o valor puder ser melhorado colocar o valor 3 (por se tratar da terceira instância);
5. Congelar a quarta linha e a quarta coluna da matriz e verificar se algum dos valores que não está congelado pode ser melhorado (através do procedimento descrito em 2. a.). Se o valor puder ser melhorado colocar o valor 4 (por se tratar da quarta instância). Assim, cada vez que o algoritmo Floyd-Warshall executa vai ter tantas instâncias quanto o número de nós que constituírem a rede, sendo que apenas na ultima instância vai ser possível verificar qual o caminho mais curto entre todos os pares de nós;

Figura 17 - Algoritmo de Floyd-Warshall

(Fonte: Cormen et al, 2009, p. 696)



$$\Pi^{(0)} = \begin{pmatrix} \text{NIL} & 1 & 1 & \text{NIL} & 1 \\ \text{NIL} & \text{NIL} & \text{NIL} & 2 & 2 \\ \text{NIL} & 3 & \text{NIL} & \text{NIL} & \text{NIL} \\ 4 & \text{NIL} & 4 & \text{NIL} & \text{NIL} \\ \text{NIL} & \text{NIL} & \text{NIL} & 5 & \text{NIL} \end{pmatrix} \quad \Pi^{(2)} = \begin{pmatrix} \text{NIL} & 1 & 1 & 2 & 1 \\ \text{NIL} & \text{NIL} & \text{NIL} & 2 & 2 \\ \text{NIL} & 3 & \text{NIL} & 2 & 2 \\ 4 & 1 & 4 & \text{NIL} & 1 \\ \text{NIL} & \text{NIL} & \text{NIL} & 5 & \text{NIL} \end{pmatrix}$$

$$\Pi^{(1)} = \begin{pmatrix} \text{NIL} & 1 & 1 & \text{NIL} & 1 \\ \text{NIL} & \text{NIL} & \text{NIL} & 2 & 2 \\ \text{NIL} & 3 & \text{NIL} & \text{NIL} & \text{NIL} \\ 4 & 1 & 4 & \text{NIL} & 1 \\ \text{NIL} & \text{NIL} & \text{NIL} & 5 & \text{NIL} \end{pmatrix} \quad \Pi^{(3)} = \begin{pmatrix} \text{NIL} & 1 & 1 & 2 & 1 \\ \text{NIL} & \text{NIL} & \text{NIL} & 2 & 2 \\ \text{NIL} & 3 & \text{NIL} & 2 & 2 \\ 4 & 3 & 4 & \text{NIL} & 1 \\ \text{NIL} & \text{NIL} & \text{NIL} & 5 & \text{NIL} \end{pmatrix}$$

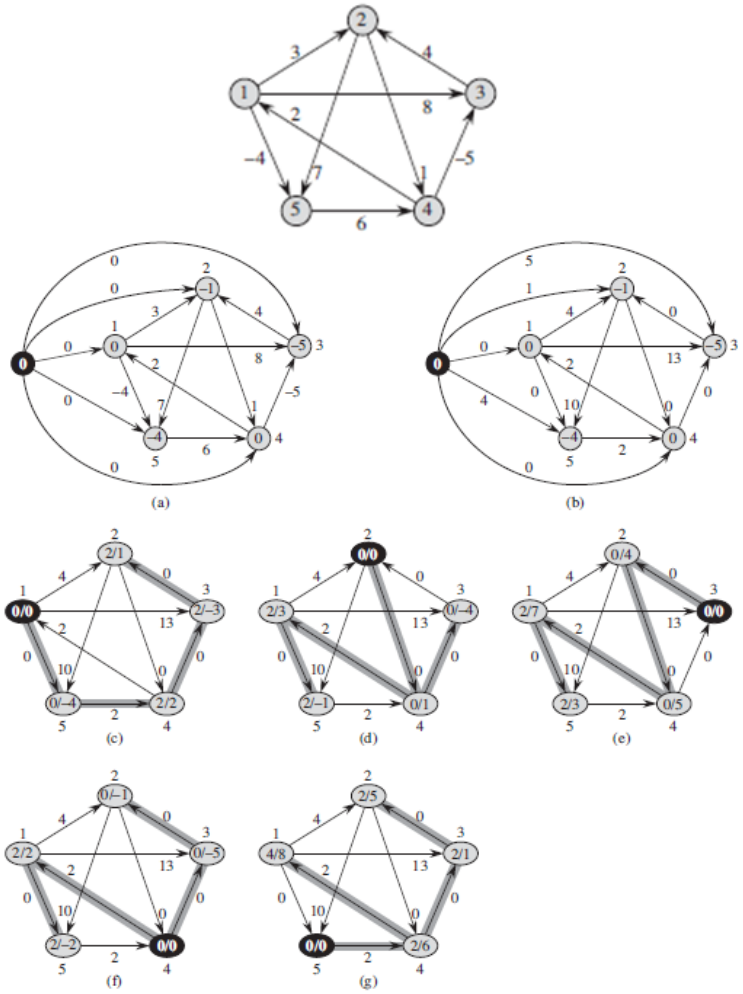
$$\Pi^{(4)} = \begin{pmatrix} \text{NIL} & 1 & 4 & 2 & 1 \\ 4 & \text{NIL} & 4 & 2 & 1 \\ 4 & 3 & \text{NIL} & 2 & 1 \\ 4 & 3 & 4 & \text{NIL} & 1 \\ 4 & 3 & 4 & 5 & \text{NIL} \end{pmatrix}$$

$$\Pi^{(5)} = \begin{pmatrix} \text{NIL} & 3 & 4 & 5 & 1 \\ 4 & \text{NIL} & 4 & 2 & 1 \\ 4 & 3 & \text{NIL} & 2 & 1 \\ 4 & 3 & 4 & \text{NIL} & 1 \\ 4 & 3 & 4 & 5 & \text{NIL} \end{pmatrix}$$

O algoritmo de Johnson foi desenvolvido com o mesmo objetivo que o algoritmo de Floyd-Warshal, no entanto, este tem uma execução bastante mais rápida para redes dispersas (Cormen et al, 2009, pp. 700-702; Grama et al, 2003). A figura 18 ajuda a clarificar a execução deste algoritmo:

1. Acrescentar um novo nó ligado a todos os outros nós através de arestas com peso 0;
2. Calcular o caminho mais curto entre cada um dos vértices utilizando ou o Bellman-Ford (se existirem arestas com pesos negativos) ou Dijkstra (se existirem apenas arestas com pesos positivos). Deverá executar-se o algoritmo selecionado uma vez a partir de cada vértice;

Figura 18 - Algoritmo de Johnson
(Fonte: Cormen et al, 2009, p. 703)



2.4.4.5. Problema das redes complexas

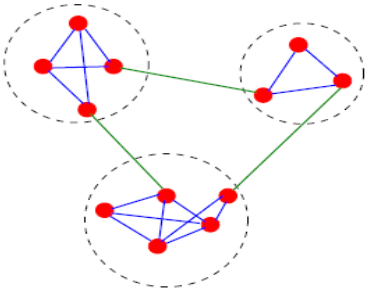
A representação de sistemas complexos, ou seja, sistemas que estão normalmente organizados em compartimentos que têm a sua própria função é feita com base em redes complexas que têm um conjunto de nós (que representam os compartimentos) com ligações internas bastante densas e relações externas (com outros compartimentos) com menor densidade (Lancichinett & Fortunato, 2009). De acordo com Tasgin et al (2008) uma comunidade é um conjunto de vértices com uma elevada densidade intra-grupo e uma baixa densidade inter-grupo.

No que respeita às redes complexas, ou seja, redes utilizadas para modelar sistemas representando o seus componentes e as suas interações em nós e ligações e às suas propriedades os algoritmos de deteção de comunidades têm sido dos mais utilizados (Orma & Labatu, 2009). Segundo os mesmos autores estes algoritmos são baseado nos princípios do *clustering* mas adaptados aos grafos, assim, definem estes algoritmos como um subconjunto de nós com ligações bastante densas com os restantes intervenientes da rede. Cada comunidade deverá ser avaliada por si só e a rede não deve ser tratada como um todo, uma vez que possuem autonomia e características suficientemente distintas (Fortunato, 2009).

Os algoritmos de deteção de comunidades podem ter duas abordagens: a abordagem de aglomeração e a abordagem de divisão. A primeira abordagem inicia-se com tantas comunidades como nós, ou seja, cada nó é a sua própria comunidade e a cada interação vai fundindo os nós até ficar um única comunidade bastante grande. Pelo contrário, a abordagem de divisão começa apenas com uma comunidade que contém todos os nós e a cada iteração vai separando os nós de forma a que cada nó fique uma comunidade. As comunidades que vão ser fundidas ou separadas são escolhidas de acordo com a distância ou função que desempenham o que permite criar comunidades homogéneas (na abordagem de agregação) ou heterogéneas (na abordagem de divisão) (Orma & Labatu, 2009).

Figura 19 – Três Comunidades rodeadas por um círculo tracejado

(Fonte: Fortunato, 2009, p. 77)



2.5. REDES NO SETOR DA SAÚDE

A utilização de redes nas diferentes áreas de negócio oferece inúmeras vantagens, nomeadamente: obtenção informação, resolução de problemas, melhoramento da coordenação dos serviços e ainda criação de uma estrutura informal da organização (Friedman, 2011). Assim, sendo as redes baseadas em grafos, são a representação mais natural de muitos domínios e são fundamentais quando se quer aferir conclusões sobre as relações (representadas nas redes como arestas) entre as entidades (representadas na rede como vértices) (Šubelj, Furlan & Bajec, 2011). Se os vértices de uma rede representarem pessoas e as arestas relações entre as pessoas então a rede é denominada como rede social (Šubelj, Furlan & Bajec, 2011). A análise de redes sociais permite identificar nós influentes e dinâmicas de rede, ou seja, identifica nós influentes, estruturas locais e globais e ainda dinâmicas de rede, assim, as redes sociais são modelos matemáticos baseados em dados que se tornam nós e ligações que permitem analisar os dados e quando necessário executar simulações (Jang et al, 2012). Segundo os mesmos autores os conceitos base deste tipo de redes são o grau (que se refere ao número de ligações que um nó tem numa rede, permitindo descobrir os nós com mais ligações que são os nós centrais visto que podem mobilizar vários recursos) e a densidade (relação entre o número de relações existentes vs o número de relações possíveis).

A utilização das redes é importante na medida em que permite identificar comportamentos fraudulentos, isto é, um novo nó da rede (entidade) “pode ser identificado como fraudulento quando se verifica um comportamento similar ao de um previamente detectado como tal” (Ferreira, 2009). De acordo com a mesma autora, quando surgir uma nova entidade na rede, esta deverá ser comparada às restantes

entidades da rede para se procurar semelhanças e conseguir-se identificar a nova entidade como fraudulenta ou não fraudulenta.

No que respeita ao setor da saúde a utilização de redes sociais é bastante adequada quando se pretende "descrever, explorar e compreender estruturalmente e relacionalmente aspetos da saúde" (Luke & Harris, 2007). Assim, apesar de ter inicialmente sido ignorada deverá ser estudada uma vez que, quando uma intervenção médica é feita num paciente, essa intervenção poderá ter efeitos não só na própria pessoa mas também nas pessoas a quem este está ligado pelo que se torna importante medir o impacto na rede de relações do paciente e não apenas no próprio (Christakis, 2004). De acordo com mesmo autor, se um paciente estiver ligado a outras entidades, ou seja, se existir entre elas qualquer tipo de relação, então as intervenções médicas efetuadas a esse paciente podem ter efeitos não intencionais nas entidades ligadas a esse paciente.

O recurso a este tipo de redes tem vários impactos na área da saúde, nomeadamente: o aumento da inovação, desenvolvimento de comunidades de prática, colaborações de apoio, aumento do fluxo de informação e análise das alterações nas relações ao longo do tempo (Friedman, 2011).

3. METODOLOGIA

3.1. ESTRATÉGIA DE INVESTIGAÇÃO

A estratégia de investigação presente nesta dissertação é o estudo de caso, isto porque, esta é a metodologia que se deve utilizar para "compreender, explorar ou descrever acontecimentos e contextos complexos" (Araújo et al, 2008). Em 2003, Yin afirma que esta estratégia é utilizada em situações complexas, quando o investigador procura respostas para "como?" e "porquê", quando o investigador pretende relacionar fatores de uma determinada entidade, quando o investigador pretende analisar, descrever, ou conhecer a dinâmica do fenómeno, do programa ou do processo em estudo (as cited Araújo et al, 2008). Assim, a seleção desta metodologia pode justificar-se com os fundamentos de Yin (2003) "os casos de estudo são preferíveis quando as questões propostas são "como" ou "porquê", quando o investigador tem pouco controlo sobre os eventos e quando a questão se centra num fenómeno contemporâneo com contexto real".

A questão de investigação colocada "Qual poderá ser o papel dos algoritmos de redes no desenvolvimento da deteção de fraude no setor dos seguros de saúde?" é uma questão explicativa que pretende explorar o tema da deteção de fraude de forma a perceber o papel dos algoritmos de redes na deteção de fraude. Este objetivo pressupõe a recolha de dados de fases relativos a diferentes alturas do tempo e que estão dependentes entre si, no entanto, é ainda importante salientar que o investigador não tem qualquer influência sobre os dados procedendo apenas à sua recolha e análise.

É através da revisão da literatura que se pretende compreender e explorar os artigos já existentes bem como definir alguns conceitos que ajudarão a uma melhor compreensão da dissertação. A definição do conceito de seguros, a contextualização do mercado de seguros de saúde, a exploração do conceito de fraude e a apresentação dos modelos de deteção de fraude permitem uma contextualização mais abrangente face aos objetivos a que esta dissertação se propõe. Em paralelo a identificação dos algoritmos utilizados nos sistemas baseados em redes e a descrição dos mesmos permitirá a identificação do algoritmo capaz de responder ao desafio de apresentar um sistema baseado na aplicação de redes que permita a deteção de fraude no ramo da saúde da indústria seguradora.

3.2. DADOS

Os dados utilizados nesta dissertação foram cedidos por uma seguradora e são relativos a um período de dois anos e 8 meses (entre Janeiro de 2012 e Agosto de 2014). Os dados disponibilizados são dados da área da saúde dentária referentes a prestadores de serviço (clinicas, hospitais, centros de saúde), profissionais de saúde (médicos), serviços prestados (diagnósticos, procedimentos, tratamentos e pagamentos) e pacientes.

Os dados fornecidos, permitem fazer não só as análises necessárias (relações entre moradas, nomes e telefones) como acrescentar variáveis a estas análises de forma a apresentar análises mais completas e mais abrangentes. O acesso a informação dos médicos, dos tratamentos efetuados, dos procedimentos de diagnóstico permite construir o sistema de deteção de fraude com base na informação necessária em conjunto com a informação possível tornando-o mais rico e completo.

Figura 20 - Estrutura do sistema proposto
(Fonte: Elaboração própria)

É ainda importante realçar que como vão ser utilizados dados secundários, a qualidade da base-de-dados será analisada, uma vez que poderá comprometer a fiabilidade e robustez do sistema.

3.3. FERRAMENTA UTILIZADA

Para a criação do sistema para deteção de fraude na indústria dos seguros de saúde vai ser utilizada uma ferramenta SAS: SAS Fraud Framework – Network Analysis.

A utilização desta ferramenta permite a representação visual para que o analista seja capaz de descrever o comportamento da entidade dentro do sistema de saúde, para

que tenha capacidade de procurar por comportamentos considerados errados efetuados pelas entidades, capacidade de criar alertas ao nível da existência de fraude na rede e consequentemente calcular o impacto financeiro total das entidades envolvidas e capacidade de criar alertas de identificação de entidades que estão a agir em conjunto em comportamentos fraudulentos.

Esta ferramenta é utilizada para criar relações (as relações podem ser fortes [por exemplo: morada, nome, telefone e nome] ou fracas [por exemplo: localizações de ATM, pagamentos e tempo]) entre entidades com base em atributos. Assim, a rede é composta por nós (entidades) e ligações (relações entre os nós).

A rede que vai ser criada vai ser composta por quatro tipos de entidades, nomeadamente, responsável pelo diagnóstico (*resp_provider*), responsável pelo tratamento de saúde (*perf_provider*), responsável pela cobrança do tratamento de saúde (*bilg_provider*) e o paciente (*patient*). Esta rede vai basear-se, em dados do segmento dentário, isto é, informação relacionada com os tratamentos dentários. As relações entre as entidades da rede vão basear-se nos seguintes atributos:

- Número significativo de pacientes comuns dentro de um prazo específico (ligações inferidas);
- Número significativo de referências entre fornecedor;
- Locais onde a entidade efetuou o serviço;

Para que a rede seja mais consistente vão ainda ser criados alertas para situações em que aparecem numa rede mais entidades com *score* de fraude alto do que o expectável. Por outro lado, através de ligações de primeiro e segundo grau vão ser identificadas quais as entidades envolvidas em atividades suspeitas semelhantes.

3.4. ANÁLISE DETALHADA DAS ETAPAS SEGUIDAS PARA A CONSTRUÇÃO DO SISTEMA

Para a construção das redes dos resultados foram seguidas as seguintes etapas:

1. Construção de grupos homogéneos através da criação de *clusters* de entidades (as entidades são os *providers* - podem ser profissionais de saúde ou serviços de saúde [hospitais, centros de saúde, laboratórios, etc.]) . A metedologia utilizada para criar estes grupos foi a análise de *clusters* (organizar os dados em grupos com base em determinadas variavéis).

   Tendo em consideração que os dados disponibilizados pela seguradora pertenciam apenas ao segmento da saúde dentária, o primeiro passo foi separar os dados dos profissionais de saúde dos dados dos serviços de saúde

(hospitais, clinícas, consultórios, etc.) De seguida e com base no ficheiro composto por todos os sinistros foi-se verificar entidade a entidade quais as que colocavam implantes. A distinção das entidades que colocam implantes prende-se com o facto deste procedimento médico ser muito mais caro que os restantes procedimentos médicos e como tal foi necessário colocar estas entidades num grupo apenas com entidades que colocam implantes. Após esta distinção estar feita foram analisadas outras váriaveis de forma a criar-se *clusters* o mais homogéneos possiveis. Para isso foi necessário calcular para cada entididade a média de cada tipo de procedimentos médicos por ano, a média de pacientes por ano, a média de valor apresentado por ano, média de valor pago por ano e ainda o género e idade dos pacientes. Foi com recurso a estas métricas que foram criados os *clusters*.

Para que os *clusters* fossem significativos foi definido que deveriam ter pelo menos 30 entidades, assim, foram-se agrupando *clusters* com base no *nearest cluster* (*cluster* mais próximo). Apesar de se pretender que cada *cluster* tivesse pelo 30 entidades em alguns casos não foi possível garantir este número de entidades porque os *clusters* eram demasiado diferentes e não era possível agrupá-los. Após os *clusters* estarem criados foi necessário criar os alertas (os alertas representam o grau de propensão a cometer fraude). Assim, procedeu-se à execução de regras de negócio para criação de alertas e consequente deteção de fraude (estas regras foram recolhidas com base em projectos internacionais) . Alguns exemplos das regras de negócio que vão ser utilizadas são:

a. O serviço de saúde é um *outlier* (em relação aos seus pares) no que respeita à quantia autorizada para pagamento de sinistros para um determinado mês;
b. Serviços de saúde prestados para vários detentores de apólices da mesma família no mesmo local num curto período de tempo;
c. O serviço de saúde é um *outlier* (em relação aos seus pares) no que respeita ao número de serviços autorizados para um determinado mês;

É importante destacar que a comparação entre pares refere-se aos restantes constituintes do grupo em que a entidade em análise se encontra, ou seja, a comparação é feita intra-gupo, sendo os grupos os encontrados no ponto anterior.

2. Construção da rede com base nos resultados obtidos anteriormente. A construção da rede vai ser efetuada com base nas entidades (profissionais e serviços de saúde), sendo o objetivo verificar a existência de uma ligação entre as mesmas. O que vai definir se existe ou não relação entre as entidades é o número de pacientes partilhados, isto é, nas redes que vão ser construídas uma entidade vai estar ligada a outra se partilhar um número suficiente de pacientes redes.

Definidas as etapas principais é importante detalhar todo o processo efetuado para a criação das redes.

Após a seguradora ter disponibilizado os dados foi necessário definir qual a metodologia a utilizar para a construção das redes. A metodologia utilizada para a construção das redes foi a metedologia por entidades sendo as entidades profissionais de saúde e serviços de saúde. Conforme foi referido acima as redes foram construídas tendo em conta o número de pacientes que as entidades partilhavam. Para perceber quais as entidades que tinham ligação foi definida um percentagem mínima de pacientes partilhados para que se considera-se que as entidades tinham relação. Assim, para entidades com mais de 100 pacientes foi definido que a partir de 10% de pacientes partilhados é considerada a existência de uma ligação, no entanto, para entidades com menos de 100 pacientes apenas se metade dos pacientes (50%) é que se considera a existência de uma relação entre as entidades.

Antes de se iniciar a construção das redes foi efetuada uma limpeza aos dados, isto porque, muitas vezes as bases-de-dados possuem erros nos registos (por exemplo, a existência de uma entidade denominada Clínica Dentária Dentes Limpos e outra entidade com o nome Clínica Dentária Deentes Limpos) que levam à duplicação de registos que pertencem obviamente a apenas uma entidade. A importância da limpeza dos dados é explicada pelo facto de, se os dados não tivessem sido tratados antes da construção das redes iriam existir duas entidades iguais para a mesma clínica (no exemplo apresentado apenas pelo erro tipográfico iriam existir duas entidades distintas) . Para uma limpeza mais efetiva dos dados após se despistarem os erros tipográficos foram definidos componentes para fazer o despiste dos dados, ou seja, os dados com a mesma data de nascimento e o mesmo número de cliente mas com nomes diferentes são supostamente a mesma entidade. Assim, o grau certeza de que as redes são construídas com a informação correta aumenta. Para além da limpeza dos dados já descrita foi necessário verificar a existência de entidades muito grandes, nomeadamente, um serviço de saúde muito grande porque obviamente vão existir

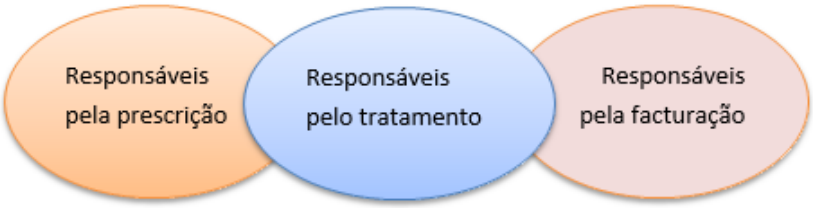
muitos médicos relacionados e como tal a rede gerada vai ser complexa e pouco relevante. Desta forma, se existirem este tipo de entidades estas devem ser excluídas uma vez que não vão acrescentar qualquer valor ao modelo.

De seguida foi necessário marcar as entidades, isto é, com base nos alertas gerados (que representam a probabilidade de fraude, isto é, quanto mais alertas/mais severos maior a probabilidade de fraude) as entidades foram marcadas com três cores: verde (menor probabilidade de fraude), amarelo e vermelho (maior probabilidade de fraude). A probabilidade de fazer fraude foi baseada nas regras de negócio, ou seja, cada vez que uma entidade não respeitava uma regra de negócio era gerado um alerta.

Após os dados estarem preparados o próximo passo foi a construção das relações. A base para a construção das relações foi um ficheiro ao nível do sinistro, ou seja, um ficheiro onde estavam colocados todos os procedimentos médicos ocorridos entre Janeiro de 2012 e Agosto de 2014, os códigos dos intervenientes (na área da saúde em geral, existem três tipos diferentes de intervenientes: responsáveis pela prescrição, responsáveis pelo tratamento e responsáveis pela faturação) e a data do serviço. Com base neste ficheiro foi possível analisar e identificar:

1. Quais os intervenientes que partilham pacientes num determinado espaço temporal, isto é, quais os intervenientes que partilham o mesmo paciente num espaço de 60 dias.
2. Quais os responsáveis pelo tratamento que partilham pacientes com os responsáveis pela facturação.
3. Quais os responsáveis pelo tratamento que partilham pacientes com os responsáveis pela prescrição.

Figura 21 – Intervenientes na área da saúde
(Fonte: Elaboração própria)



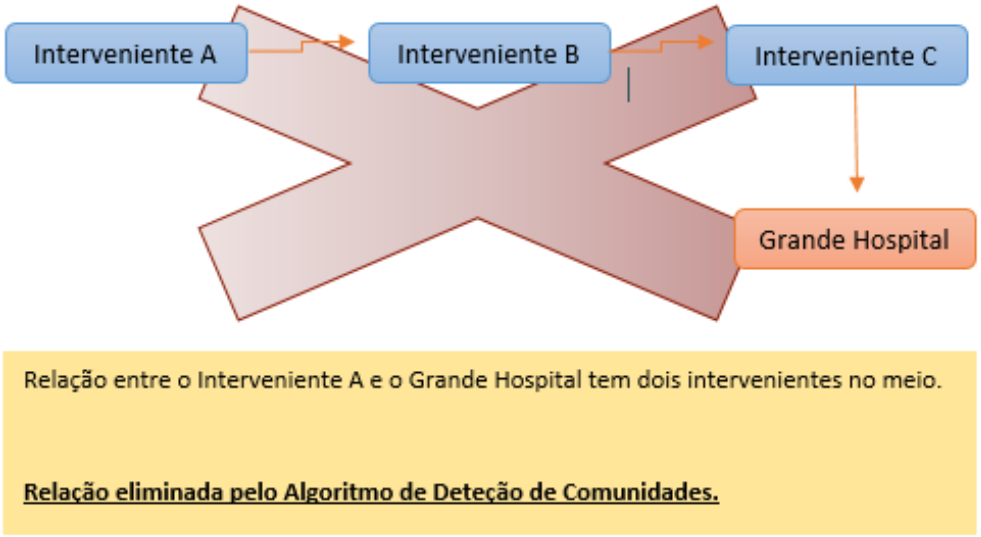
Como resultado desta análise surgiu um ficheiro com o *from_id* (código de um interveniente), *to_id* (código do interveniente que está relacionado), identificação do paciente partilhado, data de início do tratamento e data final do tratamento. Assim,

este ficheiro contém todas as relações, no entanto, nem todas as relações são importantes de colocar na rede (por exemplo, dois intervenientes que partilham um único paciente não é uma relação com peso suficiente para ser colocada na rede), para conseguir filtrar as relações e escolher aquelas que são relevantes foram colocados pesos nas ligações (número de pacientes partilhados) entre os intervenientes. Para colocar os pesos nas relações cada par de intervenientes foi agrupado para se contar o número de ligações e de seguida compará-lo com o número de pacientes que estes intervenientes normalmente têm. Esta comparação é muito importante, visto que é com base no número médio de pacientes que se vai aplicar a regra anteriormente descrita, ou seja, para intervenientes com 100 pacientes apenas foram mantidas as relações onde pelo menos 10% dos pacientes sejam partilhados, no caso de intervenientes com menos de 100 pacientes só se mantiveram as relações onde pelo menos metade dos pacientes eram partilhados.

Depois das relações estarem criadas o último passo foi "podá-las" para que as redes mostrassem a informação necessária. Para este passo foi utilizado o algoritmo de Deteção de Comunidades. Conforme já foi referido no ponto 2.4.5, este algoritmo vai observar todas as grandes redes e tentar encontrar quantos nós centrais (também conhecidos como centro da comunidade) existem. Partindo do centro da comunidade este algoritmo começa a expandir a rede e a definir os limites criando comunidades. Tendo as comunidades criadas é possível aplicar regras especificas a cada comunidade e não à rede toda o que é bastante útil. Com o recurso a este algoritmo é possível olhar para a rede e saber que toda a informação presente é útil, ou seja, se existir um alerta de três intervenientes fraudulentos sabe-se que o interveniente que estamos a investigar está relacionado de forma próxima com estes três indivíduos fraudulentos pelo que provavelmente poderá estar também a fazer fraude. Na prática o que este algoritmo faz é "podar" a rede com base nas comunidades (nós), assim, se o interveniente A estiver relacionado com o interveniente B que por sua vez está relacionado com o interveniente C que está relacionado com um grande hospital então esta relação não é mostrada na rede, isto porque, um grande hospital vai estar relacionado com vários intervenientes (devido ao seu tamanho), pelo que esta relação não vai acrescentar qualquer valor à rede do interveniente A. Tipicamente as relações com intervenientes muito grandes (neste caso por exemplo: grandes hospitais ou grandes clínicas) não acrescentam valor, sobretudo quando a distância entre o interveniente A e o grande hospital é elevada, ou seja, quando existem muitos intervenientes entre eles.

Figura 22 – Funcionamento do algoritmo de detação de comunidades

(Fonte: Elaboração própria)



Assim, e retirando as relações que não fornecem grande informação, com recurso ao algoritmo apresentado, a rede está concluída e pronta para análise.

4. ANÁLISE DE RESULTADOS

4.1. CLUSTERS

De acordo com a metodologia detalhada no ponto anterior foi possível obter quatro grandes grupos compostos na totalidade por dez *clusters*.

Os grupos de *clusters* dividem-se em:

1. Profissionais de saúde que colocam implantes;
2. Serviços de saúde que colocam implantes;
3. Profissionais de saúde que não colocam implantes
4. Serviços de saúde que não colocam implantes.

A figura 23 apresenta uma análise detalhada de cada *cluster* e permite verificar que em média cada *cluster* é composto por 78 entidades com uma média de 1026 tratamentos efetuados numa média de 546 consultas efetuadas a uma média de 262 pacientes. É ainda possível verificar que nos *clusters* 2, 3, 7 e 10 não foi possível garantir o número mínimo de 30 entidades na medida em que eram bastante diferentes dos restantes *clusters* e como tal não era possível agrupá-los.

Figura 23 – Três quadros compostos pela análise detalhada dos *clusters*

(Fonte: Elaboração própria)

Cluster	Número de entidades	Média de tratamentos médicos (*)	Média de pacientes (*)	Média do valor apresentado (*)	Média do valor pago (*)
Profissionais de saúde que não colocam implantes					
1	177	177,47	45,72	11872,58	6347,69
2	1	7619,00	1998,33	447650,83	242703,16
3	14	1228,36	290,45	72543,10	38798,89
4	265	160,61	42,57	9489,57	5151,54
Serviços de saúde que não colocam implantes					
5	125	82,98	22,32	5251,18	2827,94
6	54	160,61	42,57	9489,57	5151,54
7	23	378.35	82.02	29492.58	15929.21
8	74	66,92	16,87	3915,89	2119,74
Profissionais de saúde que colocam implantes					
9	32	605,58	127,53	51698,05	31053,74
Serviços de saúde que colocam implantes					
10	13	155,21	36,87	17711,15	10913,33
(*) A média é anual					

Procedimentos					
Cluster	Consultas (*)	Endodontia (*)	Exames Complementares (*)	Oclusão (*)	Ortodontia (*)
Profissionais de saúde que não colocam implantes					
1	88,46	20,84	56,62	10,33	45,29
2	3946,33	1028,00	2268,00	189,33	1585,00
3	580,31	142,81	431,38	63,52	275,02
4	80,53	20,77	47,58	5,70	30,18
Serviços de saúde que não colocam implantes					
5	48,31	10,34	29,43	3,70	14,98
6	96,40	20,73	62,21	11,46	51,55
7	209,53	37,80	155,00	34,06	114,41
8	34,28	6,74	23,75	2,31	14,53
Profissionais de saúde que colocam implantes					
9	292,97	70,54	240,85	40,00	51,14
Serviços de saúde que colocam implantes					
10	84,67	18,87	43,13	21,87	15,49
(*) A média é anual					

Procedimentos					
Cluster	Paradontologia (*)	Pequena Cirurgia Oral (*)	Prótese (*)	Prótese Fixa (*)	Prótese Fixa (*)
Profissionais de saúde que não colocam implantes					
1	92,53	20,99	33,68	13,82	0,00
2	3926,67	1662,67	1299,33	517,33	0,00
3	696,67	175,62	230,02	53,62	0,00
4	85,18	20,22	29,58	10,70	0,00
Serviços de saúde que não colocam implantes					
5	46,85	10,12	13,66	7,32	0,00
6	99,66	18,04	36,14	18,97	0,00
7	264,12	45,94	74,20	43,71	0,00
8	32,58	6,70	12,77	4,36	0,00
Profissionais de saúde que colocam implantes					
9	140,59	345,49	68,23	93,51	52,23
Serviços de saúde que colocam implantes					
10	31,59	94,08	24,38	34,21	16,87
(*) A média é anual					

Através da análise dos quadros acima é possível perceber que o *cluster* 2 apresenta características bastante distintas dos restantes *clusters* e, apesar de ser composto apenas por uma entidade possui um número de consultas aproximadamente sete vezes superior ao segundo *cluster* com maior número de consultas. Visto que o *cluster* 2 representa apenas 1 entidade cujas características se apresentam completamente distintas face aos restantes *clusters* (nomeadamente no que respeita ao número de tratamentos médicos e de pacientes) este foi retirado da análise para que não enviesa-se os resultados.

É ainda importante destacar que no total estão representadas aproximadamente 778 entidades, divididas em 10 *clusters* compostos no total por aproximadamente 2705 pacientes que estiveram presentes num total de aproximadamente 5462 consultas. No que respeita aos valores apresentados e pagos pelas entidades, através da análise da figura 23, é possível verificar que em média os valores apresentados pelas entidades são de 65911,45€ e a média dos valores pagos é de 36099,68€.

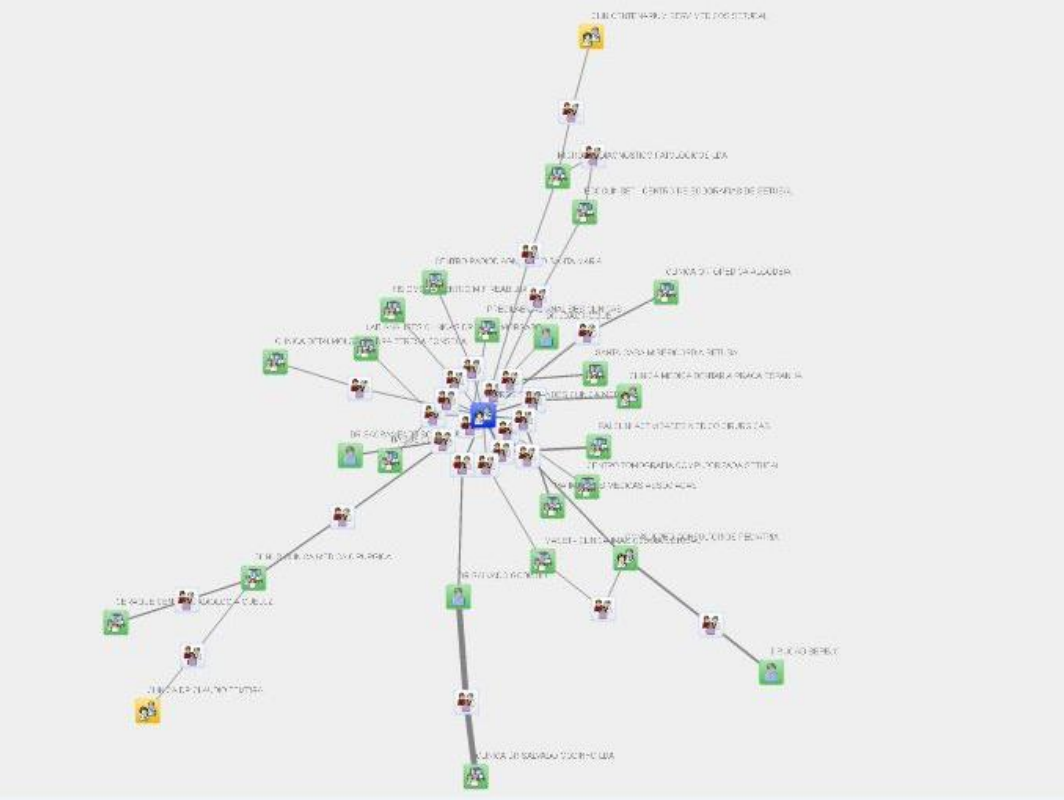
4.2. REDES

A figura 24 demonstra o aspeto de uma rede. Cada ponto representa uma entidade à excepção dos pontos brancos que representam grupos de pacientes partilhados. É ainda importante destacar que as entidades podem ter quatro cores: azul, verde, amarelo e vermelho. As cores atribuídas às entidades são representativas dos *socres* calculados com base nas regras de negócio referidas no ponto 2.4.1. O nó a azul representa o nó *source* da rede, isto é, a entidade selecionada para a visualização da rede. Os nós a verde representam entidades que possuem poucos alertas e/ou alertas pouco severos (entidades com *score* menor que 50), os nós a amarelo representam entidades que possuem alguns alertas e/ou alertas com alguma severidade (entidades com *score* entre 51 e 80), e por último os nós a vermelho representam entidades que possuem muitos alertas e/ou alertas com muita severidade (entidades com *score* maior que 80).

No que respeita aos alertas, conforme foi referido no ponto 3.4, foram calculados com base na comparação intra-grupo, isto é, os elementos do *cluster* são comparados e aqueles que apresentarem maiores desvios em relação à média do *cluster* são identificados e é criado um alerta.

Figura 24 – Exemplo de uma rede

(Fonte: Elaboração própria)



Através da análise da figura 24 é possível perceber que as relações de algumas entidades são representadas com traços mais grossos, o que significa que são relações mais densas, ou seja, relações onde a partilha de pacientes é bastante elevada.

É ainda importante destacar que as redes criadas são redes aleatórias, ou seja, são redes que não apresentam qualquer padrão nas suas ligações.

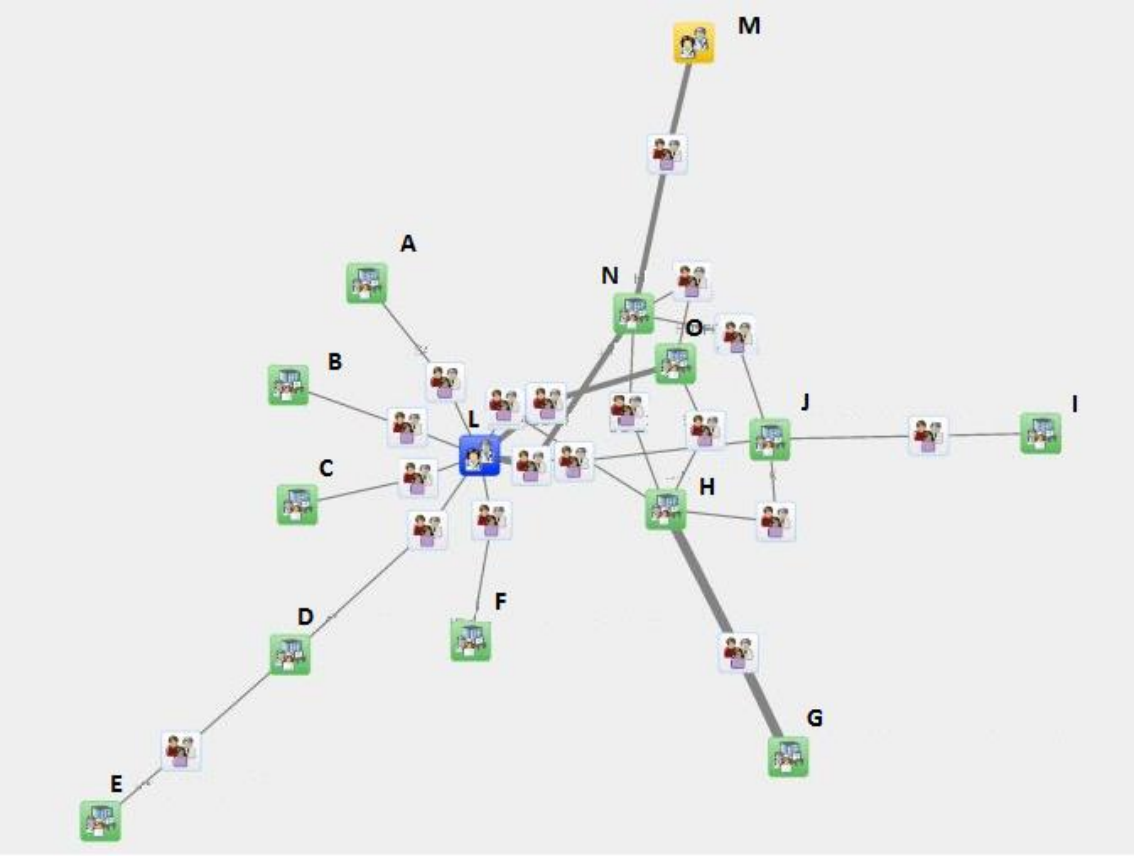
Antes de se iniciar a análise detalhada de algumas redes é importante destacar que os dados se encontram mascarados mas que a consistência foi mantida ao longo de todas as imagens, ou seja, a entidade representada com L é a mesma em entidade em todas as imagens.

Selecionando aleatóriamente uma entidade, de forma a ser possível uma melhor compreensão da importância das redes é possível verificar qual a rede detalhada dessa mesma entidade.

Através da observação da figura 25 conseguimos perceber qual a rede da entidade L.

Figura 25 – Rede da entidade L

(Fonte: Elaboração própria)



Pela análise da figura 25 é possível perceber que a entidade L partilha, de forma direta (ou seja, tem uma relação direta) pacientes com várias entidades, nomeadamente, entidade A, B, C, D, F, H, J, O e N. É importante destacar que todas as entidades com quem a entidade L tem relação direta possuem poucos alertas/alertas poucos severos. Assim, à partida esta não será uma entidade com propensão a cometer fraude, no entanto, é possível verificar que numa ligação não direta (ligação de segundo grau) existe uma entidade com alguns alertas/alertas com alguma severidade pelo que será importante analisar com mais detalhe a entidade em causa (entidade M).

Tendo em conta que na figura 25 foi possível compreender que a entidade M é uma entidade com alguma propensão à fraude é importante verificar e acompanhar as entidades com quem esta entidade partilha pacientes de forma a verificar a existência de mais entidades fraudulentas.

Pela análise da figura 26 é possível verificar que na rede da entidade M existe uma entidade com bastante propensão a fraude (representada a vermelho). De forma a

verificar qual as entidades que têm relação direta com a entidade P (com alerta vermelho) é importante observar a figura 27.

Figura 26 – Rede da entidade M
(Fonte: Elaboração própria)

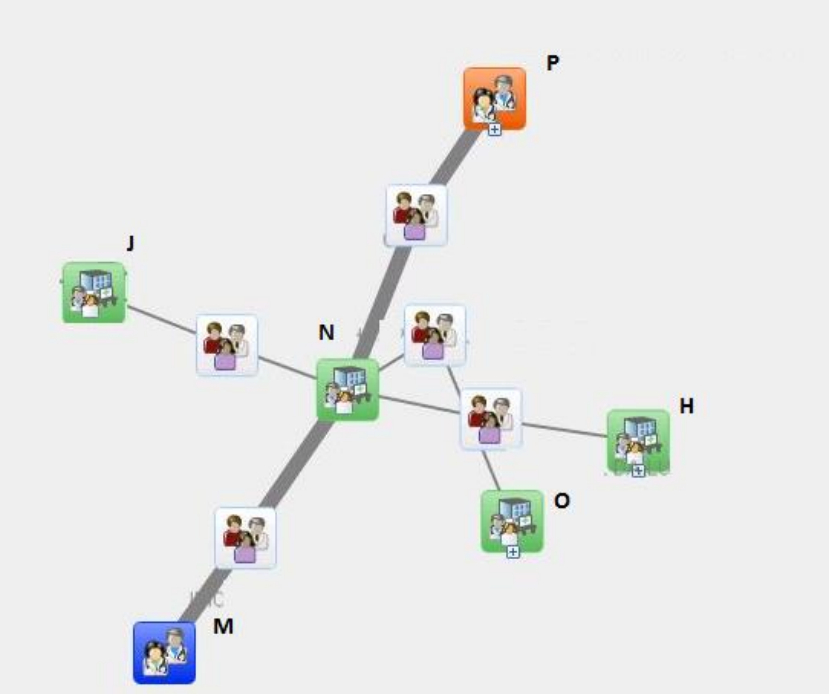
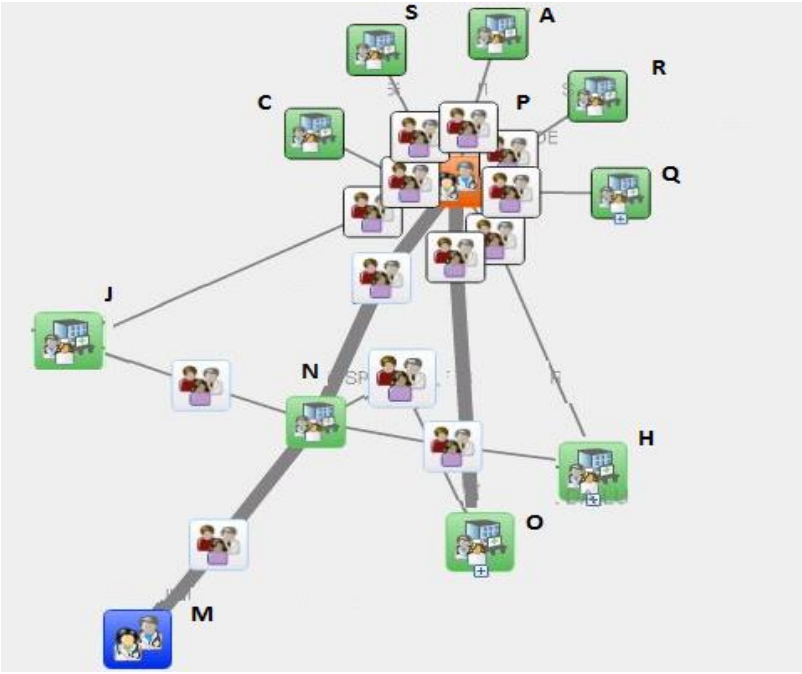


Figura 27 – Rede expandida da entidade M
(Fonte: Elaboração própria)



Tendo em conta que a entidade M é uma entidade com alguma propensão a cometer fraude e dado que possui uma relação de segundo grau com uma entidade que tem alertas bastante elevados (entidade P) é importante ir acompanhando o comportamento da entidade M. Ao mesmo tempo, torna-se útil acompanhar as entidades com ligação direta à entidade P (entidade C, S, A, R e Q) bem como a entidade N uma vez que tem ligação direta (e uma relação bastante densa) com duas entidades com comportamentos fraudulentos.

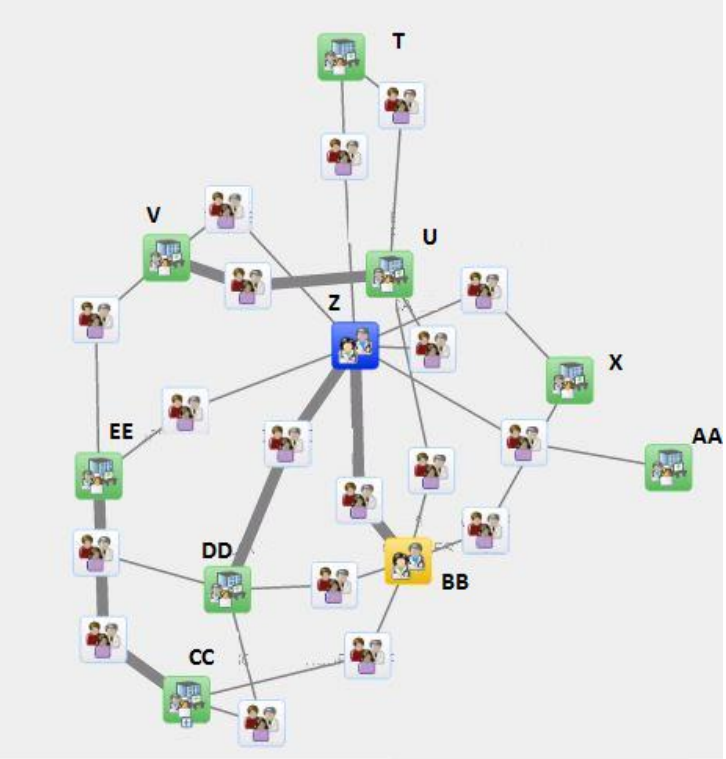
Através da rede acima é possível perceber que serviços de saúde/profissionais de saúde que se encaixem no *cluster* a que pertence a entidade P, bem como o *cluster* a que pertence a entidade M são serviços de saúde/profissionais de saúde com propensão a cometer fraude, pelo que deverão ser tomadas medidas específicas e apropriadas de forma a prevenir que estes cometam fraude.

Selecionando novamente uma entidade aleatória (figura 28), é possível verificar que na rede da entidade Z existe uma entidade com alerta moderado de fraude (entidade BB) que está diretamente relacionada (e com uma relação bastante densa, ou seja, com uma partilha de pacientes bastante elevada) com a entidade que selecionámos.

Assim, é importante perceber que esta entidade Z tem maior propensão a cometer fraude que a entidade L (figura 25) visto que a entidade Z tem uma ligação direta com uma entidade fraudulenta enquanto que a entidade L tem uma ligação de segundo grau. Ou seja, e aplicando o algoritmo do caminho mais curto entre cada par de entidades podemos fazer uma extrapolação e dizer que, quanto menor for o caminho entre uma entidade e uma entidade fraudulenta maior probabilidade dessa entidade fazer fraude.

Figura 28 – Rede da entidade Z

(Fonte: Elaboração própria)



De acordo com as figuras acima é possível concluir que uma entidade com relação direta com uma entidade fraudulenta tem maior possibilidade de vir a cometer fraude, ou seja, quanto menor for o caminho entre duas entidades (onde uma seja fraudulenta) maior a probabilidade da entidade em análise cometer fraude. Por outro lado é preciso ter em consideração que relações mais densas têm maior probabilidade de influência, ou seja, se duas entidades partilharem uma relação mais densa (com maior número de pacientes partilhados) existe a possibilidade do impacto da fraude ser mais alargado. Também é importante referir que entidades fraudulentas com muitas relações são mais preocupantes que entidades com poucas/fracas relações, uma vez que a sua rede de influência é muito maior, ou seja, têm a capacidade de manipular o comportamento e a informação de um número mais elevado de entidades.

Deste modo, a utilização de redes é importante na medida em que permite uma aferição mais rápida, gráfica e fácil de quais a entidades que cometem fraude, permitindo traçar um perfil de entidades fraudulentas que deverá ser tido em conta quando surgem novas entidades para que se possam tomar medidas capazes de prevenir a fraude. Assim, com o recurso às redes as empresas seguradoras conseguem uma seleção de quais as entidades com maior risco de fazer fraude e consequentemente

uma redução do desperdício de recursos em investigação de entidades que não são propensas a fazer fraude. Simultaneamente a análise das relações é outra vantagem da utilização de redes visto que permite verificar a probabilidade de influência entre as entidades, ou seja, por um lado permite a observação da densidade das ligações de forma a verificar onde é que o risco de influência é mais elevado, por outro lado, permite a análise da distância entre as entidades e a identificação das entidades onde o caminho que as separa é mais curto.

5. CONCLUSÕES

Embora a fraude nos seguros de saúde tenha várias consequências esta é muito difícil de prevenir e detetar, uma vez que os esquemas fraudulentos estão em constante desenvolvimento, e consequentemente é muito difícil de reduzir. Para o desenvolvimento de um sistema de deteção baseado em redes na área da saúde é necessário identificar as entidades e definir qual a variável que vai servir de base à criação de relações (para as redes apresentadas no capítulo 4 foram utilizados os pacientes partilhados).

Para criar uma rede mais rica os dados devem ser tratados com base nas regras de negócio (para as redes apresentadas no capítulo 4 foram utilizadas as regras de negócio fornecidas pela seguradora). Para além de tratar os dados as regras de negócio são importantes para criar alertas adequados ao negócio onde a rede foi implementada. Assim, ao utilizar as regras de negócio fornecidas pela seguradora foi possível verificar quais as entidades que não cumpriam as regras da seguradora e criar alertas sobre as mesmas.

De acordo com a investigação efetuada, e numa tentativa de responder ao segundo objetivo específico a que nos propusemos, os algoritmos (apresentados detalhadamente no ponto 2.4.4), são importantes em dois momentos: para tratar os dados, no caso desta dissertação foi utilizado o algoritmo de deteção de comunidades, (descrito no capítulo 4), mas também para analisar e ajudar a retirar conclusões das redes criadas, nesta dissertação foi utilizado o algoritmo do caminho mais próximo para proceder a esta análise (este algoritmo foi também apresentado no capítulo 4). Ao mesmo tempo é necessário avaliar e verificar o tipo de rede tendo em consideração as características que apresentam, com base nos resultados obtidos no capítulo 4 é possível perceber que as redes geradas são redes aleatórias.

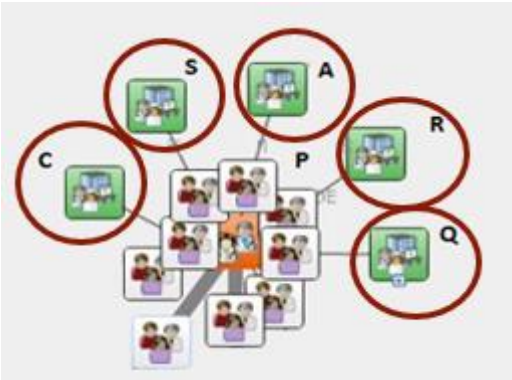
Assim, quanto mais rica for a rede, ou seja, quanto mais os dados estiverem trabalhados e filtrados, quanto mais informação útil acerca das entidades (telefones, moradas, gastos, etc.) estiver contemplada na rede, mais fidedignas e completas serão as conclusões que se poderão retirar. Considerando o primeiro objetivo específico a que nos propusemos, embora a informação necessária (no estudo de caso efetuado) seja apenas a que diz respeito aos identificadores das entidades (*perf_provider*, *resp_provider* e *bilg_provider*) e número de pacientes partilhados (variável responsável pela criação de relações entre as entidades), se a rede contiver mais informação

(telefones, moradas, volume de negócio, número de médicos, etc.) irá tornar-se, na opinião do autor, bastante mais rica e capaz de fornecer informação útil à seguradora.

No entanto, é importante ter em consideração que a rede não permite identificar e descrever as tipologias de fraude uma vez que permite apenas verificar quais as entidades que se relacionam com entidades fraudulentas.

Através da utilização das redes, conseguimos verificar que estas não são uma ferramenta suficientemente conclusiva para afirmar que uma determinada entidade vai fazer ou não fraude, isto porque, as redes vão permitir identificar e visualizar mais rapidamente as relações entre as entidades de forma a ser possível identificar quais as entidades mais próximas a entidades fraudulentas, no entanto, as relações entre as entidades não são suficiente para afirmar que uma entidade é fraudulenta. A existência de uma relação entre uma entidade fraudulenta e uma entidade não fraudulenta não é suficiente para afirmar que ambas as entidades vão fazer fraude, porque, não é por uma entidade estar relacionada com uma entidade considerada fraudulenta que virá obrigatoriamente a cometer fraude. Assim, podemos concluir que a utilização das redes pode gerar números elevados de falsos positivos, ou seja, identificar várias entidades como possíveis fraudulentas que na verdade não irão apresentar qualquer comportamento fraudulento (figura 29).

Figura 29 – Identificação das entidades mais propensas a fazer fraude
(Fonte: Elaboração própria)



De acordo com a figura 29 existem cinco entidades com propensão a fazer fraude pelo que deveriam ser as cinco investigadas, no entanto, no ponto de vista do autor esta questão pode não ser absoluta, ou seja, apesar da rede identificar estas entidades como mais propensas a fazer fraude (por estarem diretamente relacionadas com uma entidade fraudulenta), o facto da entidade P fazer fraude não implica que todas as

entidades que estão relacionadas com esta entidade façam, pelo que algumas destas entidades que pela observação da rede requeriam investigação/controlo poderiam ser falsos positivos.

Apesar de podermos verificar que as redes geram falsos positivos é ainda importante ter em consideração que a maioria da fraude nos seguros de saúde é efetuada em conjunto (paciente e médico, médico e laboratório, hospital e médico, dois hospitais, dois médicos, dois pacientes, etc.) pelo que, na opinião do autor, a observação da rede poderá ser pertinente e eficaz na medida em que quando uma entidade prepara um esquema fraudulento não vai dirigir-se às entidades com quem não tem relação, vai optar por contactar as entidades mais próximas.

A fraude nos seguros de saúde é um processo dinâmico que está em constante evolução, ou seja, a forma como a fraude é feita está sempre a desenvolver-se e vão surgindo novos esquemas fraudulentos pelo que se torna fulcral identificar e prevenir a fraude para diminuir as perdas das empresas seguradoras com este crime. No entanto, e em resposta ao terceiro objetivo específico apresentado, pudemos ainda aferir que embora a utilização de redes seja uma ferramenta com utilidade e com benefícios para a empresa seguradora deverá ser um complemento a outras metodologias de deteção e prevenção de fraude na medida em que a eficácia desta técnica pode ser contestada sobretudo por assumir, conforme foi referido acima, que uma entidade que esteja relacionada com uma entidade identificada como fraudulenta deverá ser investigada pois a probabilidade de fazer fraude é considerável.

6. LIMITAÇÕES E INVESTIGAÇÃO FUTURA

6.1. LIMITAÇÕES

A realização desta dissertação esteve sujeita a algumas limitações que a condicionaram.

A primeira limitação encontrada foi o facto dos dados que serviram de base ao estudo de caso presente nesta dissertação serem provenientes de apenas uma seguradora o que poderá ter limitado as conclusões. A utilização de dados de várias seguradoras permitiria conseguir redes mais abrangentes e permitiria verificar se as entidades consideradas fraudulentas, com base nos dados de uma seguradora, eram também consideradas fraudulentas pelos dados de outras seguradoras.

Por outro lado, o facto dos dados disponibilizados pertencerem apenas ao segmento da saúde dentária surge como outra limitação na medida em que, na opinião do autor, os seguros de saúde disponíveis no mercado apresentam baixo valor de cobertura na área da saúde dentária pelo que, e tendo em conta que esta é uma das áreas mais procuradas pelos clientes, pode fazer com que esta área se torne mais apetecível a fazer fraude e consequentemente a sua probabilidade seja maior que nas restantes áreas da saúde. Assim, o acesso a dados de diferentes setores permitiria não só comparar e verificar quais os setores da saúde onde existe mais fraude, como verificar se as entidades que atuam em vários setores (por exemplo hospitais ou clínicas) fazem fraude nos diferentes setores ou se esta se evidência em setores específicos.

Após terem sido criadas as redes foi encontrada outra limitação, isto porque, através das redes foram identificadas as entidades com maior propensão a fazer fraude, no entanto, os dados fornecidos pela seguradora não tinham esta identificação ou seja, não foi possível confirmar se de facto as entidades identificadas eram ou não fraudulentas. Esta particularidade dos dados pode também ser definida como uma limitação, visto que, se a seguradora tivesse identificado quais as entidades fraudulentas seria possível retirar conclusões mais fidedignas acerca da utilidade das redes uma vez que era possível comparar resultados e verificar se de facto este sistema de deteção e prevenção de fraude gera muitos falsos positivos.

6.2. INVESTIGAÇÃO FUTURA

Para enriquecer o sistema de deteção e prevenção de fraude e de forma a complementar as redes criadas seria importante aplicar os modelos preditivos criados às entidades identificadas pelas redes como possíveis fraudulentas para se verificar quais as entidades com maior probabilidade de fazer fraude reduzindo desta forma os falsos positivos gerados pelas redes. Assim, as entidades com maior probabilidade de fazer fraude deveriam ser acompanhadas e investigadas para se conseguir reduzir os custos da empresa seguradora associados à fraude.

Ao mesmo tempo seria interessante criar clusters sobre as entidades identificadas como fraudulentas de forma a conseguir uma análise mais profunda das suas características.

Após estarem criados os clusters a aplicação de modelos preditivos, por exemplo o recurso a árvores de decisão, permitiria identificar as características comuns às entidades que fazem fraude. Conhecendo melhor as entidades fraudulentas seria possível caracterizá-las e perceber de forma mais precisa e completa quais as características que efetivamente podem ser consideradas de risco.

Com os resultados dos modelos preditivos seria possível quando surgissem novas entidades classificá-las e compará-las com as entidades fraudulentas permitindo assim antecipar os possíveis comportamentos fraudulentos. Assim, se uma nova entidade pertence-se ao *cluster* de entidades com perfil fraudulento poderia afirmar-se que a probabilidade de vir a cometer fraude era bastante elevada.

7. BIBLIOGRAFIA

ACL. (2010). Fraud Detection Using Data Analytics in the Healthcare Industry.

Araújo, C., Lopes, E. M. F. P., Lopes, J., & Pinto, L. N. R. (2008). Estudo de Caso.

Araújo, H. De, & Ferreira, O. (2012). A Gestão do Risco Operacional em Eventos Desportivos.

Araújo, T. V. de. (2006). Redes em Economia : Criação de Estruturas e Auto-Organização em Sistemas Económicos Complexos.

Araújo, T. V. de. (2011). Introdução à Economia Computacional (pp. 157–189).

Azevedo, A. I. R. L. (2011). Data Mining Languages for Business Intelligence.

Bacher, J.-L. (1995). Insurance Fraud.

Bank for International Settlements. (2001). Basel Committee on Banking Supervision Consultative Document Operational Risk.

Barabási, A. L. (2002). Linked- How Everything Is Connected to Everything Else and What It means for Business, Science and Everyday Life.

Barabási, A.-L., Albert, R., & Hawoong, J. (1999). Mean-field Theory for Scale-free Random Betworks.

Berkhin, P. (2006). Survey of Clustering Data Mining Techniques (pp. 1–56).

Bolton, R. J., & Hand, D. J. (2002). Statistical Fraud Detection : A Review.

Borginho, H. (2014). A Importância Estratégica do Solvência II. Retrieved from http://www.isp.pt/NR/rdonlyres/8370611D-BA30-4885-8A42-244A45268747/0/F33_Art3.pdf, Acedido em 05-07-2014.

Brites, J. (2006). Fraude em Seguros. 2006.

Cahill, M. H., Lambert, D., Pinheiro, J. C., & Sun, D. X. (2002). Handbook oh Massive Datasets: Detecting Fraud in the Real World (pp. 911–930).

Castells, M. (1999). A Sociedade em Rede.

Castells, M., & Cardoso, G. (2005). A Sociedade em Rede.

CEA. (2010). The European Motor Insurance Market.

Chan, P. K., Fan, W., Prodromidis, A. L., & Stolfo, S. J. (1999). Distributed Data Mining in Credit Card Fraud Detection.

Christakis, N. A. (2004). Social Networks and Collateral Health Effects.

Clarke, M. (1989). Insurance Fraud.

Comissão Europeia. (2014). Comunicado de Imprensa: Luta Contra a Fraude.

Contador, C. (2011). A Fraude no Seguro: Aspectos Econômicos, 87–104.

Derrig, R. A., & Dedene, G. (2002). A Comparision of State-Of-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection.

Elkan, C. (2000). Magical Thinking in Data Mining : Lessons From CoIL Challenge 2000.

Ernst, Y. (2011). Operational Risk : Quantification Models.

Even, S. (2011). Graph Algorithms (p. 10).

Fan, W. (2004). Systematic Data Selection to Mine Concept-Drifting Data Streams.

Fawcett, T., & Provost, F. (1997). Adaptive Fraud Detection.

Ferreira, G. C. (2012). Redes Sociais de Informação em Organizações num Contexto da Sociedade Contemporânea .

Ferreira, M. I. da S. P. (2009). Detecção e Prevenção de Fraude.

Fisher, E. (2008). The Impact of Health Care Fraud on the United States Healthcare System.

Fortunato, S. (2010). Community Detection in Graphs. Physics Reports (Vol. 486). Elsevier B.V.

Foster, D. P., & Stine, R. A. (2004). Variable Selection in Data Mining : Building a Predictive Model for Bankruptcy.

Friedman, S. (2011). Measuring Social Networks : Understanding and Supporting Relationships to Transform Health Care.

Gama, J. (2002). Árvores de Decisão.

Gee, J., & Button, M. (2014). The Financial Cost of Healthcare Fraude 2014.

Geiger, H. (2000). Regulating and Supervising Operational Risk for Banks.

Gil, R. (2008). Lutando pela Transparência.

Gonçalves, R. (2011). Sistemas de Informação para Gestão de Risco Operacional em Instituições Financeiras.

Grama, A., Gupta, A., Karypis, G., & Kumar, V. (2003). Graph Algorithms.

Grinsven, J. Van, & Bloemkolk, R. (2010). Risk Management in Financial Institutions: Formulating Value Propositions (pp. 89–96).

Guimar, J. F. dos R. (2010). Os Seguros de Saúde Voluntários.

Guiné, C. (2014). Solvência II – Resultados do Exercício QIS5. Retrieved from http://www.isp.pt/NR/rdonlyres/052195EE-AF23-4DDD-B225-6167019A726D/0/F31_art1.pdf, Acedido em 05-07-2014.

Herbest, H. (1996). Business Rules in Systems Analysis: A Meta-Model and Repository System.

Instituto de Seguros de Portugal. (2012). Relatório do Setor Segurador e dos Fundos de Pensões.

Instituto Seguros de Portugal. (2013). Estatística do Seguro de Saúde.

Insurance Europe. (2012). Como Funciona o Seguro.

Insurance Europe. (2013). O Impacto da Fraude em Seguros.

Jang, H. L., Lee, Y. S., & An, J.-Y. (2012). Application of Social Network Analysis to Health Care Sectors.

Johnson, D. (1973). A Note on Dijkstra's Shortest Path Algorithm.

Joshi, A. (2008). Social Ties and their Relevance to Churn in Mobile Telecom Networks.

Kamil, A. (2003). Graph Algorithms.

Kirlidog, M., & Asuk, C. (2012). A Fraud Detection Approach with Data Mining in Health Insurance.

Kotsiantis, S. B. (2011). Decision Trees: A Recent Overview.

Kreher, W., & Kocay, D. L. (2004). 2004_Graphs, Algorithms, and Optimization.

Kumar. (2000). An Introduction to Cluster Analysis for Data Mining.

Lancichinetti, A., & Fortunato, S. (2009). Community Detection Algorithms: A Comparative Analysis (Physics and Society; Computational Physics).

Lavrac, N., Motoda, H., Fawcett, T., Holte, R., Langley, P., & Adriaans, P. (2004). Introduction: Lessons Learned from Data Mining.

Lemos, E. (2003). Análise de Crédito Bancário com o Uso de Data Mining: Redes Neurais e Árvores de Decisão.

Luke, D. A., & Harris, J. K. (2007). Network Analysis in Public Health: History, Methods, and Applications. Annual review of public health (Vol. 28, pp. 69–93).

Maio, L. S. C. G. da C. (2013). Fraude nos Seguros: A Tolerância à Fraude no Seguro Automóvel.

Matos, I. M. D. De. (2013). Teoria dos Grafos no Ensino Básico e Secundário.

Moura, H. da S., & Silva, A. C. R. da. (2004). Auditoria de Fraude: Instrumentos na Prevenção de Fraudes Contra as Empresas.

National White Collar Crime Center. (2013). Health Care Fraud.

Orman, G. K., & Labatut, V. (2009). A Comparison of Community Detection Algorithms on Artificial Networks.

Phua, C., Alahakoon, D., & Lee, V. (2004). Minority Report in Fraud Detection: Classification of Skewed Data.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A Comprehensive Survey of Data Mining-based Fraud Detection Research.

Pimenta, C. (2009). Esboço de Quantificação da Fraude em Portugal.

Pimenta, C., & Afonso, Ó. (2012). Notes on the Epistemology of Fraud.

Pordata. (2012). Despesa Corrente em Cuidados de Saúde em % do PIB em Portugal.

Robinson, B., & Officer, J. Data Mining: Predicting Laptop Retail Price Using Regression. Retrieved from http://www.spelman.edu/docs/aspire-research/joibritneypdf.pdf?sfvrsn=2, Acedido em 07-07-2014.

Rosset, S., Murad, U., Neumann, E., Idan, Y., & Pinkas, G. (1999). Discovery of Fraud Rules for Telecommunications Challenges and Solutions.

Saidenberg, M., & Schuermann, T. (2003). The New Basel Capital Accord and Questions for Research.

Schiller, J. (2006). The Impact of Insurance Fraud Detection System.

Shao, J., & Pound, C. J. (1999). Extracting Business Rules from Information Systems.

Shepheard-Walwyn, T., & Litterman, R. (1998). Building a Coherent Risk Measurement and Capital Optimisation Model for Financial Firms.

Silva, S. N. da. (2009). Os Seguros de Saúde Privados no Contexto do Sistema de Saúde Português.

Soares, M. (2008). Contributo do Data Mining na Detecção e Prevenção de Fraude.

Sparrow, M. K., A., M., & D., P. (1996). Health Care Fraud Control Understanding The Challenge.

Stein, T. H. C. C. E. L. R. L. R. C. (2009). Introduction to Algorithms.

Šubelj, L., Furlan, Š., & Bajec, M. (2011). An Expert System for Detecting Automobile Insurance Fraud Using Social Network Analysis. Expert Systems with Applications.

Tasgin, M., Herdagdelen, A., & Bingol, H. (2008). Community Detection in Complex Networks Using Genetic Algorithms.

Townsend, R. M. (1979). Optimal Contracts and Competitive with Costly State Verification.

Viaene, S., Dedene, G., & Derrig, R. A. (2005). Auto Claim Fraud Detection Using Bayesian Learning Neural Networks.

Wang, H., Fan, W., Yu, P. S., & Han, J. (2003). Mining Concept-Drifting Data Streams Using Ensemble Classifiers.

Wan-Kadir, W. M. N., & Loucopoulos, P. (2004). Relating Evolving Business Rules to Software Design.

Watts, D. J. (1999). Networks, Dynamics, and the Small World Phenomenon.

Watts, D. J. (2003). Six Degrees: The Science of a Connected Age. Choice Reviews Online (Vol. 40).

Yin, R. K. (2003). Case Study Research: Design and Methods.