

Fundamentando e Apresentando um Conjunto de Hipóteses para Embasar a Elaboração de dois Algoritmos de Mineração Fractal de Dados para a Detecção de Fraudes

Autoria: Reinaldo Cherubini Neto, Denis Borenstein

Resumo

Há fraudes em praticamente todos os ramos de atividades, nos esportes, nas artes, na ciência e, como não poderia deixar de ser, nos negócios. Todos os anos bilhões de dólares são perdidos no mundo devido as fraudes. Além disso, os fraudadores estão sempre desenvolvendo novos golpes e repaginando os antigos. Para isto, se utilizam de novas tecnologias, exemplo disso é o esquema Ponzi que vez ou outra reaparece, inclusive como corrente de e-mails. Porém, o desenvolvimento da ciência e da tecnologia também está do lado do combate às fraudes, criptografia de dados, senhas, reconhecimento de voz, cadastramento de usuários, certificados digitais, programas antivírus, programas anti-spam e sistemas de detecção de fraudes, por exemplo, compõem o arsenal básico para o combate a este tipo de crime. Embora as fraudes sejam quase tão antigas quanto às civilizações a detecção de fraudes é um assunto que começou a ser estudado recentemente, em 1962. A grande maioria dos artigos indexados no *Web of Science*, sob o termo “*fraud detection*” está relacionada ao tema “mineração de dados” (*data mining*). Este ensaio teórico tem como objetivo, considerando a fraude e o comportamento do consumidor como fenômenos complexos, apresentar e fundamentar um conjunto de hipóteses que permita a elaboração futura de dois algoritmos de mineração de dados para a detecção de fraudes; hipóteses estas baseadas em uma teoria que aborda a complexidade de forma simples. Contudo, apesar de existirem fraudes em todos os setores de atividade humana, o interesse deste trabalho é somente com as cometidas pelas pessoas físicas contra pessoas jurídicas nas transações que geram contas a pagar e que são registradas em banco de dados. As hipóteses apresentadas neste ensaio são obtidas por meio do método dedutivo da geometria fractal a partir de dois axiomas de detecção de fraudes, um para a detecção de fraudes cometidas por impostores e outro para a detecção de fraudes cometidas por vigaristas. Ambos os axiomas são deduzidos das definições de fraudes e da classificação dos fraudadores. A teoria do caos e a geometria fractal estão inseridas num campo da matemática, chamado de teoria dos sistemas dinâmicos, que foi desenvolvido para lidar com a complexidade envolvida em fenômenos estocásticos determinísticos, que são também chamados de caóticos. O estudo desses fenômenos não lineares possui duas abordagens distintas: uma qualitativa e outra quantitativa, sendo que a geometria fractal está inserida na segunda. Espera-se que os futuros algoritmos, desenvolvidos a partir das hipóteses H_1 e H_2 , sejam eficientes e eficazes o suficiente para atacar os problemas impostos à detecção de fraudes por meio da mineração de dados.

INTRODUÇÃO

As fraudes são quase tão antigas quanto às civilizações. Um exemplo disso são as falsas múmias de animais sagrados que eram vendidas por alguns cidadãos egípcios a ricos e nobres do antigo Egito, para serem utilizadas em suas cerimônias fúnebres, como a falsa múmia de Íbis, pássaro sagrado do antigo Egito, apresentada pela Figura 1, contendo somente gravetos em seu interior.

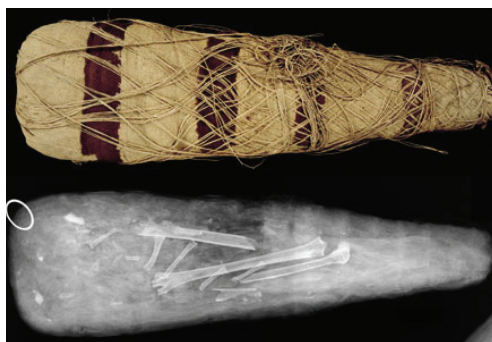


Figura 1. Falsa múmia de Íbis, foto externa e Raios-X do interior (gravetos e ossos de outros animais)

Fonte: Adaptado de PARODI, Lorenzo. **Introdução ao Mundo das Fraudes**. Monitor das Fraudes. Recuperado em 02 abril, 2011 de <http://www.fraudes.org/showpage1.asp?pg=2>.

Todos os anos bilhões de dólares são perdidos no mundo devido as fraudes. Estimativasⁱ apontam para perdas anuais de aproximadamente US\$ 40 bilhões com vendas fraudulentas de bens e serviços por telefone (telemarketing) nos EUA (Graycar & James, 2001). Ainda, de acordo com uma reportagem do jornal Zero Hora, de fevereiro de 2009, as fraudes virtuais são uma das modalidades de crime que mais crescem no Brasil. Segundo a Federação Brasileira de Bancos [FEBRABAN] (conforme citado em Zero Hora, 2009), este tipo de fraude causa um prejuízo anual de R\$ 500 milhões ao sistema financeiro nacional.

Há fraudes em praticamente todos os ramos de atividade humana, nos esportes, nas artes, na ciência e, como não poderia deixar de ser, nos negócios. Os fraudadores estão sempre desenvolvendo novos golpes e repaginando os antigos. Para isto, se utilizam de novas tecnologias, exemplo disso é o esquema Ponziⁱⁱ, que vez ou outra reaparece, inclusive como corrente de e-mails. Mas, o desenvolvimento da ciência e da tecnologia também está do lado do combate às fraudes, criptografia de dados, senhas, reconhecimento de voz, cadastramento de usuários, certificados digitais, programas antivírus, programas anti-spam e sistemas de detecção de fraudes, por exemplo, compõem o arsenal tecnológico básico para o combate a este tipo de crime.

A detecção de fraudes (DF) é um assunto que começou a ser estudado recentemente, embora as fraudes sejam quase tão antigas quanto às civilizações. O primeiro artigo indexado no *Web of Science* sobre este assunto é do ano de 1962, um segundo artigo só foi publicado vinte e dois anos depois, em 1984. Uma busca, no *Web of Science* (WS), com o termo “*fraud detection*”, retornou 148 documentos, o gráfico da Figura 2 apresenta a evolução do número de publicações de 1990 a 2009.

Os artigos sobre detecção de fraude, indexados na base de dados WS, abordam a aplicação de fraudes em diversos setores, tais como: telecomunicações, cartões de crédito e seguro de automóveis. Embora as fraudes possam ser cometidas pelas organizações, pelos consumidores e pelos funcionários contra suas organizações, a maioria desses estudos envolve as fraudes cometidas pelos consumidores contra empresas. Apenas quatro tratam de fraudes em produtos (ou adulteração de produto), três sobre vinho e um sobre café. A grande maioria das 148 publicações está relacionada ao tema “mineração de dados” (*data mining*) e, dos que

não estão, grande parte deles, estudam a auditoria contábil-financeira para identificar fraudes cometidas nas (ou pelas) empresas.

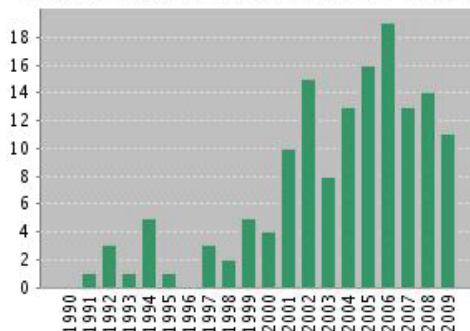


Figura 2. Evolução anual das publicações indexadas no *Web of Science* sobre detecção de fraudes

De modo geral, as fraudes podem ser cometidas por pessoas físicas contra outras pessoas físicas ou jurídicas ou ainda, contra o estado. Do mesmo modo, elas podem ser cometidas pelas pessoas jurídicas contra outras pessoas jurídicas, físicas ou contra o estado. Mas, o interesse deste trabalho é somente com as cometidas pelas pessoas físicas contra pessoas jurídicas nas transações que geram **contas a pagar** e que são **registradas em banco de dados**. Neste ensaio, objetiva-se, considerando a fraude e o comportamento do consumidor como fenômenos complexos, apresentar e fundamentar um conjunto de hipóteses, baseadas em uma teoria que aborda a complexidade de forma simples, que permita a elaboração futura de dois algoritmos de mineração de dados para a detecção de fraudes. A teoria do caos e a geometria fractal estão inseridas num campo da matemática, chamado de teoria dos sistemas dinâmicos, que foi desenvolvido para lidar com a complexidade envolvida em fenômenos estocásticos determinísticos, ou seja, caóticos. O estudo desses fenômenos não lineares possui duas abordagens distintas: uma qualitativa e outra quantitativa, sendo que a geometria fractal está inserida na segunda.

2 FRAUDES

Segundo Podgor (1999 as cited in Vasiu & Vasiu, 2004), fraude é um desses conceitos familiares que parecem ter um significado óbvio, até o momento em que tentamos defini-lo. É um conceito legal profundo que poucos realmente entendem ou, então, o empregam na definição comum (Ellingson, 1998 as cited in Vasiu & Vasiu, 2004). Fraude, em um sentido legal e amplo, significa um crime ou ato ilegal qualquer cometido por aquele que se utiliza de algum logro ou ilusão, como método principal, aplicado contra uma vítima para obtenção de lucro (Wikipédia, 2008). No dicionário de direito (Gilbert, 1997 as cited in Vasiu & Vasiu, 2004 p.3), fraude é definida como sendo –

Um ato com uso da enganação como a distorção intencional da verdade, de declarações falsas ou ocultação de um fato relevante para obter uma vantagem desleal em detrimento de outro, a fim de conseguir algo de valor ou privar outro do direito.

Para Podgor (1999), a legislação norte americana sobre fraudes se espelha na inglesa. A definição clássica do direito inglês foca na mentira intencional ou omissão, a legislação criminal federal americana foca no logro e a legislação sobre a fraude civil americana na ludibriação. Graycar and James (2001) comentam que a fraude na legislação australiana não é tratada como uma categoria de crime legalmente separada. Já, na legislação brasileira, a fraude é tratada juntamente com o estelionato no Capítulo VI do Título II do Código Penal Brasileiro nos artigos de 171ⁱⁱⁱ a 179.

Podgor (1999) comenta que não é fácil definir o termo exatamente, dependendo da legislação a definição pode variar. Jain (2008) também comenta sobre a diversidade de

definições para o termo. Para ele, fraude é um termo legal que pode ter vários significados dependendo do contexto em que é empregado. Mas, a despeito dessa dificuldade, Jain (2008) observa que a intenção de enganar ou lograr alguém é subjacente a todas as definições (o que pode ser verificado também na legislação inglesa e norte americana). Ele cita uma observação da Suprema Corte Indiana a qual diz que fraude é uma intenção de enganar, quer seja na expectativa de obter algum lucro para si, quer seja o objetivo, de prejudicar o outro, não material. Esse autor define fraude como sendo –

Um ato de deliberada enganação com o desejo de adquirir alguma coisa por meio da obtenção de lucro desleal de outro. É um logro a fim de obter ganhos por meio de perdas dos outros. É uma intenção fraudulenta para obter vantagens (Jain, 2008, p.7).

A definição de fraude dada por Devitt et al (1998 as cited in Podgor, 1999 p.9) também coloca a mentira, ludibriação ou logro como elemento essencial à fraude:

A fraude é uma deturpação intencional ou deliberada da verdade com a finalidade de induzir a outrem, com base nisso, entregar algo de valor ou ceder um direito legal. Fraude, então, é um logro que, se cometido por palavras, conduta, ou omissão, é destinado a causar a outra parte um prejuízo legal.

Ainda, Lopes de Sá e Hoog (2008 p.19) definem fraude como sendo “um ato doloso cometido de forma premeditada, planejado, com a finalidade de obter proveito com o prejuízo de terceiros”. O *Concise Oxford Dictionary* (as cited in Bolton & Hand, 2002) define fraude como uma “ludibriação criminal; o uso de representações falsas para obter uma vantagem injusta”.

Como pode ser visto nas definições apresentadas até aqui, a ludibriação (enganação ou mentira) é um elemento essencial à existência da fraude. Segundo Burgoon and Buller (1994 as cited in Jhonson & Santos, 2004 p.339), a ludibriação ou mentira (do inglês *deception*) é uma “ação deliberada cometida por um emissor para engendrar em um receptor crenças contrárias ao o que o emissor acredita ser verdade a fim de colocar o receptor em desvantagem”. Assim, a ludibriação (ou mentira) é a ferramenta do fraudador, que a emprega a fim de obter uma vantagem, geralmente de cunho econômico, ou prejudicar alguém. Portanto, com base nas definições até aqui apresentadas, pode-se afirmar que é necessária à ocorrência de uma fraude a existência conjunta dos dois requisitos abaixo:

- R₁- ludibriação (mentira, enganação) ou ocultação da verdade, ambas intencionais; e
- R₂- obtenção de um ganho (normalmente econômico) e/ou consequente prejuízo de outro.

Desse modo, a obtenção de uma simples vantagem (ou a realização de um prejuízo a um terceiro) não caracteriza por si só uma fraude, embora esse ato possa ser em muitos casos, um ato ilegal. Contudo, é imprescindível a existência deste requisito (R₂) à ocorrência da fraude, pois como a fraude não é um crime auto-revelável ela só será passível de detecção no momento em que o fraudador tentar ou conseguir realizar o seu objetivo final.

2.1 TIPOS DE FRAUDES E FRAUDADORES

Há diversos tipos de fraudes, em diversos setores da economia, fraudes em cartões de créditos, em telecomunicações, em seguros, na medicina, nas ciências, na web, etc. Afinal, em toda a atividade econômica há a possibilidade de ocorrência de fraudes. De acordo com a *Association of Certified Fraud Examiners* [ACFE] (2002) as principais categorias de fraudes são: deturpação de fatos materiais; ocultação de fatos; suborno; conflitos de interesse; roubo de dinheiro, bens, segredos comerciais ou propriedade intelectual; e violação de dever fiduciário.

Assim como as fraudes, os fraudadores também podem ser classificados. Eles podem ser divididos em duas categorias, como apresenta Bhargava, Zhong and Lu (2003): impostor (*impersonators*) e vigarista (*swindlers*). O impostor é um usuário ilegítimo que rouba

recursos de suas vítimas para “assumir” suas contas (ex.: clonar o telefone celular ou o cartão de crédito). O vigarista é um usuário legítimo que burla intencionalmente o sistema por meio da ludibriação. Mas, apesar de ser um usuário legítimo, ele não intenciona pagar as suas contas (ex.: cria um cadastro de usuário ou conta em banco com dados falsos). A fraude cometida por este tipo de fraudador pode ser chamada de fraude de subscrição.

3 DETECÇÃO DE FRAUDES

As tecnologias de combate às fraudes podem ser divididas em duas grandes categorias: (a) prevenção de fraudes, que se destinam a impedir que as fraudes ocorram; e (b) detecção de fraudes, que envolvem a identificação das fraudes tão logo elas tenham sido cometidas (Bolton & Hand, 2002).

As tecnologias de DF entram em funcionamento principalmente quando as de prevenção falham, mas na prática elas devem ser continuamente utilizadas (Bolton & Hand, 2002). Segundo Provost (2002) a detecção de fraudes e a devida intervenção podem ser realizadas de duas maneiras: automaticamente e por iniciativa mista (humana/computacional). No entanto, ambas as maneiras fazem uso de algoritmos computacionais – quando possível. Uma das maneiras mais difundidas, tanto academicamente como comercialmente, para detectar fraudes automaticamente^{iv} é a utilização de algoritmos computacionais de mineração de dados (*data mining - DM*).

O termo *data mining* é usado normalmente como sinônimo do processo de extração de informações proveitosas de base de dados (BD). É uma etapa do processo de descoberta de conhecimento em base de dados (Fayyad & Stolorz, 1997). A descoberta de conhecimento em base de dados (*Knowledge Discovering in Databases - KDD*) é, segundo Fayyad and Stolorz (1997 p.102), o “processo não trivial de identificar padrões válidos, não conhecidos, potencialmente úteis e interpretáveis nos dados”. Especificamente, *data mining* é “um passo do processo de KDD que consiste em aplicar algoritmos de análise de dados e descoberta que, sob uma limitação aceitável de eficiência computacional, produzem uma enumeração particular de padrões (ou modelos) sobre os dados” (Fayyad, Piatetsky-Shapiro & Smyth, 1996 p.41).

As principais áreas de estudo de ferramentas para detecção de fraude, onde são empregados algoritmos computacionais, são segundo Bolton and Hand (2002): fraudes em cartões de crédito, lavagem de dinheiro, telecomunicações, invasão de computadores e fraudes em atendimentos médicos (planos de saúde).

Os casos de fraudes são notórios por sua complexidade, enquanto que as leis são muitas vezes relativamente simples se comparadas aos fatos e evidências que cercam as fraudes, mesmo porque, a complexidade é uma maneira usada pelos fraudadores para encobrirem seus atos (Kingston, Schafer and Vandenberghe, 2005). Além disso, os fraudadores estão sempre adaptando e aperfeiçoando seus métodos. Ao descobrirem a existência de um método de detecção em um local irão adaptar suas estratégias a outros. Há ainda os novos criminosos que constantemente estão entrando no “mercado”, muitos dos quais não estarão advertidos sobre os métodos de detecção que obtiveram sucesso no passado e, irão adotar estratégias que permitirão serem identificados. Assim, tanto os métodos mais novos como os antigos devem ser utilizados na tarefa de detecção de fraudes (Bolton & Hand, 2002).

Apesar dos inúmeros tipos de fraudes as ferramentas de detecção, que usam algoritmos computacionais para executar a sua tarefa, só podem ser utilizadas para detectar as fraudes cujos resultados das transações ficam registrados em bancos de dados, ou seja, principalmente as fraudes que são cometidas pelos impostores. Segundo Bhargava, Zhong & Lu (2003), a maioria dos esforços de pesquisas sobre detecção de fraudes se concentra nas cometidas pelos impostores (*impersonators*).

Considerando as transações que geram contas a pagar e levando em conta apenas o segundo requisito necessário à ocorrência de uma fraude – R_2 , pode-se prever os efeitos possíveis resultantes da ação de cada um dos dois tipos de fraudadores, a despeito da complexidade dos casos de fraudes. Sendo E_1 o efeito resultante da ação do impostor (*impostor*) e E_2 o efeito resultante da ação do vigarista (*swindlers*), tem-se que :

- E_1 = a conta de um terceiro sofrerá uma alteração de valor para maior devido à existência de uma ou mais transações não pertencentes a ela; e
- quando a fraude for cometida por um vigarista, três efeitos serão possíveis:
 - $E_{2.i}$ = não existirá um responsável real para a conta; se houver, então:
 - $E_{2.ii}$ = as garantias para o pagamento da conta não existirão ou
 - $E_{2.iii}$ = a conta terá um valor menor do que deveria.

Ainda considerando somente as transações que geram contas a pagar, a ação do impostor irá atingir diretamente no mínimo duas pessoas, uma física e uma jurídica, enquanto a do vigarista, normalmente, irá atingir diretamente uma pessoa jurídica.

3.1- QUESTÕES RELEVANTES SOBRE DETECÇÃO DE FRAUDES COM DM

Uma das questões mais importantes, senão a mais, envolvendo a DF por meio da mineração de dados é a necessidade de tratar a detecção de fraudes como um problema particular dentro do tema mineração de dados. Há, pelo menos, três razões para tal: i) o desbalanceamento das bases de dados mineradas; ii) a necessidade de uma eficiência realmente alta; e iii) a exigência de uma eficácia também muito alta.

Segundo Daskalaki, Kopanas, Goudara, and Avouris (2003), os casos de fraude são raros se comparados aos legítimos. Brause, Langsdorf and Heep (1999 as cited in Bolton & Hand, 2002) trabalharam com uma BD de transações de cartão de crédito cuja probabilidade de encontrar um fraudador era de 0,2% e, Hassibi (2000 as cited in BOLTON e HAND, 2002) com uma BD onde somente uma para cada 1.200 transações era fraudulenta. Como as bases são desbalanceadas encontrar os pesos adequados para as correções dos erros, como também para a avaliação dos modelos torna-se uma tarefa difícil e com certa carga de subjetividade. Além disso, a amostra para treinamento poderá ficar com poucas transações fraudulentas.

Contudo, este problema estaria resolvido com o uso de algoritmos de identificação de *outliers*, pois os métodos de detecção de fraudes baseados nos esquemas de identificação de *outliers* partem do pressuposto que, por se tratar de um dado bem diferente dos demais – a ponto de parecer pertencer a outra amostra – um *outlier* caracteriza uma transação fraudulenta. Assim, seria fácil para os identificadores de *outliers* detectarem as transações fraudulentas em BD desbalanceadas, elas seriam as poucas diferentes das demais. Porém, uma transação fraudulenta não precisa, necessariamente, ser discrepante como também, um *outlier* pode não ser resultado de uma transação fraudulenta. Para Tang, Chen, Fu and Cheung (2006), os *outliers* são identificados por suas propriedades distintas. Estas propriedades podem ser conceituais (uso não autorizado de um cartão de crédito) ou físicas (montante envolvido na transação), as conceituais estão na mente dos usuários e as físicas, por sua vez, estão descritas nas BD. Nem sempre as propriedades físicas dos *outliers* irão descrever suas propriedades conceituais, assim, por vezes, os *outliers* detectados nem sempre irão corresponder às expectativas dos usuários^v. “Falsos negativos” e “falsos positivos” podem ocorrer quando as propriedades conceituais dos *outliers* estão grosseiramente desconhecidas das suas propriedades físicas (Tang et.al, 2006).

Aos algoritmos de detecção de fraudes é imprescindível uma alta eficiência^{vi} e escalabilidade, pois as bases de dados sobre as quais estes algoritmos trabalham são realmente de alta dimensionalidade. Por exemplo, a companhia de cartões de crédito Barclay card mantém aproximadamente 350 milhões de transações anuais, somente no Reino Unido (Hand et al., 2000 as cited in Bolton & Hand, 2002); o *Royal Bank of Scotland* mantém registros de

mais de um bilhão de transações de cartões de crédito por ano e a AT&T registra aproximadamente 275 milhões de chamadas por dia da semana (Cortes & Pregibon, 1998 apud Bolton & Hand, 2002). De acordo com Chen, Han and Yu (1996) algoritmos com complexidade exponencial, ou até mesmo de média ordem polinomial, não terão uso prático, para Fayyad and Stolorz (1997) os com complexidade superior $O(n^2)$ já são inviáveis.

Quanto a eficácia dos algoritmos de mineração de dados destinados a detectar transações fraudulentas em BD, não basta a eles uma alta acurácia. Bolton e Hand (2002) exemplificam: um método que classifique corretamente 99% dos registros (99% dos fraudadores como fraudadores e 99% dos não fraudadores como não fraudadores) seria considerado de alta eficácia. No entanto, se na BD existir somente 1 fraudador para cada 1.000 registros, então para cada 100 identificados como fraudadores, em média, apenas 9 realmente o seriam. A investigação detalhada dos 100 registros suspeitos em busca desses 9 terá, possivelmente, um custo considerável (Bolton & Hand, 2002).

Contudo, apesar dos métodos de DM terem a vantagem teórica sobre os métodos estatísticos de não exigirem suposições arbitrárias sobre os dados, os resultados reportados dos primeiros são levemente superiores ou ocasionalmente inferiores aos resultados dos segundos (Kirkos, Spathis & Manolopoulos, 2007). Viane et al. (2002) avaliaram o desempenho de seis algoritmos de DM no contexto de detecção fraudes em seguros de automóveis e obtiveram os seguintes resultados principais: (i) em todos os cenários avaliados, a diferença de desempenho entre os algoritmos foi pequena; (ii) técnicas relativamente simples e eficientes, como *Logistic Regression Classifiers* e *LinearLeast-Squares Support Vector Machine Classifiers* tiveram um bom desempenho global; e (iii) técnicas mais complexas e que demandam mais computação (*Bayesian Learning Multilayer Perceptron Neural Network Classifiers*, *RBF-Least-Squares Support Vector Machine Classifiers*, e *Tree-Augmented Naive Bayes Classifiers*) tendem a adicionar pouco ou nenhum poder preditivo.

Deve-se considerar que um algoritmo de detecção de fraude só pode realizar sua tarefa com base na obtenção do ganho (vantagem) por parte do fraudador (R_2), pois como a ludibriação (R_1) é um requisito conceitual ela não é registrada nas BDs, além disso, é a existência de R_2 que determina efetivamente a ocorrência da fraude. Por isso, esses algoritmos não poderão detectar, no sentido literal da palavra^{vii}, uma transação fraudulenta. O que poderá ser feito por um algoritmo computacional é indicar se há suspeita de fraude em uma transação. Mesmo porque, a determinação final do caráter fraudulento de uma transação vai além do escopo das ciências da computação^{viii}.

Com base no exposto até aqui é possível estabelecer duas condições essenciais para que a fraude possa ser detectada por um algoritmo computacional qualquer^{ix}:

1^a - o registro em uma BD das transações a serem investigadas (C_1); e

2^a - a presença de R_2 , ou seja, a ocorrência efetiva da fraude, E_1 e $E_{2,iii}$ ^x (C_2).

Respeitando C_1 e C_2 pode-se definir dois axiomas para a detecção de fraudes por meio da mineração de dados, um relacionado ao resultado da ação do impostor (A_1) e outro a do vigarista (A_2).

Sendo A o conjunto de todas as transações normais (legais e legítimas) registradas em uma BD; F o conjunto das transações fraudulentas efetuadas por impostores e registradas nesta mesma BD; e A'_i o subconjunto de A referente às transações do i -ésimo indivíduo. Para o indivíduo que sofreu um “ataque” de um impostor, tem-se que:

$$A_1) \quad A'_i \cap F \neq \emptyset.$$

Sendo a_1, a_2, \dots, a_n os valores corretos das transações de um vigarista e f_1, f_2, \dots, f_n os valores realmente registrados em uma BD, destas transações:

$$A_2) \quad \sum_{i=1}^n a_i > \sum_{j=1}^n f_j$$

De outra forma: A_1) quando os fraudadores forem impostores – haverá no conjunto de transações do representado no mínimo uma transação efetuada por um impostor, do contrário, por definição, não houve a fraude; e A_2) quando o fraudador for um vigarista – o valor total do conjunto de transações consideradas do vigarista será menor do que deveria ser, caso contrário, a fraude não terá ocorrido, poderá, então, ter ocorrido uma tentativa ilegal de ludibriação, mas fraude, por definição, não.

4 FRACTAIS E FRACTAL DATA MINING

Na natureza podem ser encontrados inúmeros fenômenos que possuem um comportamento aparentemente caótico, turbulência do ar, incêndio florestal, etc. No entanto, observando mais atentamente estes comportamentos é possível se encontrar auto-similaridade (Barbará & Chen, 2005).

De acordo com Stewart (1991 como citado em Savi, 2006), a ciência moderna tem uma tendência de nomear como caótico o comportamento de sistemas estocásticos determinísticos. Assim, os sistemas caóticos são determinísticos, ou seja, para uma entrada conhecida e determinada o sistema apresentará uma resposta aparentemente aleatória. Já, os fenômenos aleatórios não são determinísticos; para uma entrada aleatória o sistema apresentará uma resposta também aleatória (Savi, 2006).

Fractais e caos são estudados dentro de um tema mais abrangente chamado dinâmica (Strogatz, 1994). A teoria dos sistemas dinâmicos é uma parte da matemática que permite lidar com a complexidade envolvida nesses fenômenos não lineares (Capra, 2008/1996). O estudo de fenômenos não lineares possui duas abordagens distintas: uma qualitativa e outra quantitativa. A abordagem qualitativa tem como objetivo central entender o comportamento global de um sistema dinâmico qualquer. Já, na abordagem quantitativa o objetivo é analisar a evolução do sistema no tempo (Savi, 2006); o cálculo da dimensão fractal está inserida nesta abordagem. Capra (2008/1996) explica que quando se trata de não linearidade equações deterministas simples podem produzir uma grande variedade de comportamentos e, por outro lado, comportamentos complexos e aparentemente aleatórios podem dar origem a estruturas ordenadas e padrões sutis. Em sistemas não lineares pequenas mudanças podem ocasionar grandes efeitos, isto devido aos laços de realimentação presentes nesses sistemas. São estes laços, conhecidos matematicamente como iterações ($x \rightarrow kx$), os responsáveis pelas instabilidades e pelo surgimento súbito de novas formas de ordem. O processo de iteração conhecido como transformação do padeiro ou ferradura de Smale ($x \rightarrow kx(1-x)$) é o aspecto matemático que liga a teoria do caos à geometria fractal (Capra, 2008/1996).

Lorenz (1963) ao estudar os sistemas hidrodinâmicos, que variam aparentemente de maneira irregular e aleatória e mesmo quando observados por longos períodos de tempo não repetem a história passada, mostrou que um simples conjunto de equações, utilizadas em um procedimento de integração numérica, pode gerar soluções não periódicas. O principal interesse de Lorenz (1963), neste estudo, foi com os fluxos determinísticos não periódicos, o que hoje seria chamado de fluxo com comportamento caótico. Segundo ele, a principal propriedade dos fluxos determinísticos não periódicos é a instabilidade diante de modificações de pequena amplitude. Isto significa que os resultados futuros dos sistemas não periódicos são altamente dependentes do seu estado inicial. Diferenças imperceptíveis entre dois estados iniciais podem, eventualmente, resultar em dois estados consideravelmente diferentes no futuro, o que, segundo Savi (2006), ficou conhecido como o *efeito borboleta*.

Segundo Strogatz (1994, p. 331) um comportamento caótico é “um comportamento aperiódico de longo prazo em um sistema determinístico que exibe uma dependência sensível das condições iniciais” e, de acordo com Oh and Thomas (2010 p.134), “a auto-similaridade é uma característica que descreve um comportamento caótico ou fractal”.

Os fractais são auto-similares. Isto significa que suas partes, as quais podem ser uma estrutura, um objeto ou um conjunto de dados, são exata ou estatisticamente similares ao fractal como um todo. Assim, eles apresentam, parcial ou integralmente, as mesmas características para diferentes variações na escala em que estão sendo analisados (Lee, 2005). De acordo com Peitgen et al. (1992 as cited in Jelinek & Fernandez, 1998), a semelhança observada entre os níveis de iteração/amplificação de um objeto geométrico sem um comprimento representativo é chamada de auto-similaridade.

A Figura 3 apresenta o conjunto de Mandelbrot, um dos mais famosos fractais, onde é possível visualizar a auto-similaridade. O conjunto de Mandelbrot, segundo Capra (2008/1996, p.125) “é a coleção de todos os pontos da constante c no plano complexo para os quais os conjuntos de Julia correspondentes são peças isoladas e conexas”. Ele é um catálogo de todos os possíveis conjuntos de Julia e este, por sua vez, é obtido com base no mapeamento simples. O conjunto de Mandelbrot é objeto matemático mais complexo já inventado (Capra 2008/1996). Isto revela o poder desta teoria para lidar, de modo simples, com a complexidade.

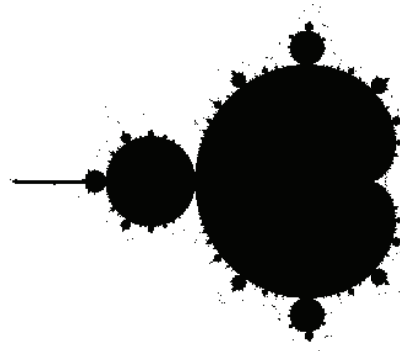


Figura 3: conjunto de Mandelbrot

Fonte: Capra, F. (2008). *A Teia da Vida: Uma nova compreensão científica dos sistemas vivos* (p.126). (11a Ed., N.R. Eicheberg). São Paulo: Cultrix. (Obra original publicada em 1996)

Capra (2008/1996) explica que é impossível calcular o comprimento ou área de uma forma fractal, mas pode-se definir, de maneira qualitativa, o seu grau de “denteamento”. A Figura 4 apresenta outro exemplo de fractal, o triângulo de Sierpinski que tem perímetro infinito e área nula. Uma vez que o número que representa o grau de denteamento de uma forma fractal tem propriedades semelhantes às de uma dimensão, Mandelbrot o denominou de dimensão fractal; uma linha dentada preenche mais espaço em um plano do que uma linha reta que tem dimensão igual a 1, porém menos espaço que um quadrado, cuja dimensão é 2 (Capra, 2008/1996). Jelinek and Fernandez (1998) explicam que a dimensão é denominada fractal porque é uma fração e não um número inteiro (dimensão fracionada) e é chamada de dimensão, pois fornece uma medida de quão completamente o objeto preenche o espaço.

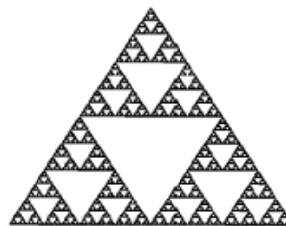


Figura 4: Triângulo de Sierpinski

Fonte: Lee, D. H. (2005). *Seleção de atributos importantes para a extração de conhecimento de base de dados* (p. 48). Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, Brasil. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-22022006-172219/pt-br.php>

Há muitas maneiras de medir a dimensão fractal de um objeto (Mandelbrot, 1985; Jelink & Fernandez, 1998; Barbará & Nazeri, 2000; Barbará & Chen, 2005) e, com base nisto, Mandelbrot (1985) diferencia auto-similaridade de auto-afinidade. Segundo ele, para os objetos ou conjuntos auto-similares todas as formas de medir a dimensão fractal levarão ao mesmo resultado, no entanto, para os que possuem auto-afinidade isto não acontecerá. Mandelbrot (1985) explica que para conjuntos auto-similares os valores produzidos por estas dimensões serão idênticos para todo o conjunto bem como para suas partes. Agora, ao mover-se pelas formas auto-afins, ver-se-á que os valores locais e globais deverão ser diferentes para cada dimensão e que os diferentes valores locais deixam de ser idênticos. De acordo com Lee (2005), teoricamente os fractais exatamente similares (ou auto-similares) são infinitos já, os conjuntos de dados reais, que apresentam um número finito de pontos, na prática, são estatisticamente auto-similares (ou são auto-afins), para um determinado intervalo de uma escala se obedecerem a uma regra bem definida nesse intervalo.

A dimensão fractal pode ser calculada, para fractais exatamente auto-similares (D), pela seguinte equação (Lee, 2005):

$$D = \frac{\log(R)}{\log\left(\frac{1}{e}\right)}$$

Onde:

- R = quantidade de réplicas da figura a cada iteração
- e = escala das réplicas geradas

Por exemplo, a dimensão fractal do Triângulo de Sierpinskyé igual a 1,58496, pois a cada iteração são gerados três réplicas com a metade do tamanho da anterior:

$$D = \frac{\log(3)}{\log\left(\frac{1}{2}\right)} = 1,58496$$

Mesmo considerando somente as dimensões de Hausdorff e de Minkowski e Bouligand há diversos algoritmos para calculá-las como: *calliper*, *box-counting*, *dilation*, *mass-radius*, e *cumulati veinter section methods*. Os três primeiros estão baseados na medição do tamanho e os dois últimos, da massa (Jelinek & Fernandez, 1998). Contudo, de acordo com Faloutsos and Kamel (1994) para o cálculo da dimensão fractal em uma BD o método mais utilizado é o *Box-Counting* ou *Boxcountplot*.

Incorporando-se o conjunto de dados em uma grade n -dimensional com células de lados com tamanho r , pode-se calcular a frequência com que os pontos de dados caem na célula p_i , e computar a dimensão fractal generalizada D_q , conforme a equação abaixo (Barbará & Nazeri, 2000).

$$D_q = \frac{1}{q-1} \frac{\partial \log \sum_i p_i^q}{\partial \log r} \quad (1)$$

Das dimensões descritas pela Equação 1 acima, a Dimensão Fractal de Hausdorff ($q=0$), a Dimensão Informacional ($\lim_{q \rightarrow 1} D_q$) e a dimensão de Correlação ($q=2$), são as mais utilizadas.

5 HIPÓTESES PARA A DETECÇÃO DE FRAUDES COM FRACTAL DM

De acordo com Oh and Thomas (2010), dois fractais diferentes podem ter um mesmo valor para a dimensão fractal, mas dois fractais com valores diferentes para as suas dimensões fractais certamente serão diferentes. Ou seja, “o valor da dimensão fractal serve como uma assinatura de uma mudança de estado do sistema” (Oh & Thomas, 2010, p.133).

O exemplo do conjunto de George Cantor comparado ao “Conjunto Híbrido de Cantor”, utilizado por Barbará and Nazeri (2000) com o propósito de demonstrar o poder da

dimensão fractal para guiar algoritmos de clusterização; reforça a afirmação acima, de Oh e Thomas (2010). O Conjunto de Cantor é construído da seguinte forma: desenha-se um seguimento de reta com tamanho n , repete-se este mesmo seguimento abaixo dividindo-o em três seguimentos de tamanho $n/3$ e exclui-se o pedaço central; e assim sucessivamente, resultando no objeto da Figura 5.

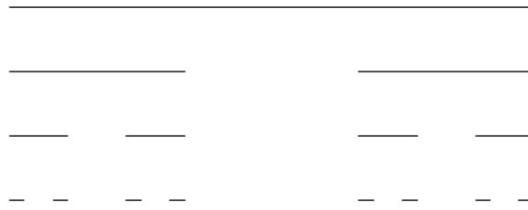


Figura 5: Conjunto de Cantor.

Já o Conjunto Híbrido de Cantor é construído da seguinte forma: inicialmente ele segue o mesmo modo de construção do Conjunto de Cantor – desenha-se um seguimento de reta com tamanho n , repete-se este mesmo seguimento abaixo dividindo-o em três seguimentos de tamanho $n/3$ e exclui-se o pedaço central – em seguida, no lado direito segue-se o mesmo procedimento e no esquerdo, divide-se o seguimento em nove seguimentos de reta de tamanho $n/9$ sendo que quatro seguimentos serão alternadamente excluídos, resultando no objeto da Figura 6.



Figura 6: Conjunto Híbrido de Cantor

Fonte: Barbará, D., & Nazeri, Z. (2000). *Fractal Mining of Association Rules Over Interval Data* (Technical Report, p. 5), Fairfax, Virginia, USA, George Mason University, ISE Dept. Retrieved December, 10, 2010, from <http://citeseerx.ist.psu.edu/viewdoc/versions?doi=10.1.1.24.3356>

A dimensão fractal do Conjunto de Cantor é 0,63:

$$D = \frac{\log(2)}{\log\left(\frac{1}{3}\right)} = 0,63$$

Para o Conjunto Híbrido de Cantor a dimensão fractal é de 0,73.

$$D = \frac{\log(5)}{\log\left(\frac{1}{9}\right)} = 0,73$$

A inclusão da parte esquerda no conjunto de cantor produz uma alteração na dimensão fractal da figura e, portanto, a parte mais a esquerda do conjunto híbrido de cantor é um conjunto anômalo em relação à parte mais a direita, ou vice e versa. Em se tratando de agrupamento, é fácil para o olho humano perceber o lado mais a esquerda e o lado mais a direita da Figura 6 como pertencentes a grupos diferentes e, de fato, um algoritmo que explora a dimensão fractal certamente irá agrupar estes dois conjuntos em dois grupos diferentes (Barbará & Chen, 2000).

Com base no demonstrado acima por Barbará and Chen (2000) e Barbará and Nazeri (2000) e considerando um conjunto de dados estatisticamente auto-similar (ou auto-afim) é possível estabelecer uma hipóteses sobre o comportamento da dimensão fractal em relação aos dados de um conjunto:

- H' – ao criar um novo registro, para uma instância de uma base de dados, com um valor gerado para um atributo específico, por meio de uma regra diferente dos demais, o valor da dimensão fractal, deste conjunto de dados, sofrerá uma variação significativa.

Considerando que uma linha reta horizontal tem dimensão euclidiana e dimensão fractal igual a um:

- H'' – para um atributo específico de uma determinada instância de uma base de dados, uma variação, em torno da média, próxima a zero produzirá um valor próximo a um para a dimensão fractal do conjunto de dados considerados.

Tomando-se como base as duas hipóteses acima, sobre o comportamento da dimensão fractal de um conjunto de dados estatisticamente auto-similar, H' e H'' , é possível, considerando os axiomas A_1 e A_2 , elaborar duas hipóteses sobre a detecção de fraudes com o uso de fractal *data mining*. Uma referente à E_1 (resultado do efeito da ação do impostor) – H_1 – e outra à $E_{2,iii}$ (terceiro resultado possível da ação de um vigarista) – H_2 .

Considerando o conjunto de dados que registram o valor das transações de um cliente qualquer, se este conjunto apresentar auto-similaridade estatística e se $A'_i \cap F = B$ e $B \neq \emptyset$, então: (H_1) o resultado da ação do impostor irá provocar uma alteração significativa no valor da dimensão fractal do conjunto de transações do representado.

Como já comentado anteriormente, para que uma fraude seja passível de detecção por mineração de dados as transações, tanto as fraudulentas como as não, devem estar registradas em alguma base de dados (C_1) e o fraudador deve ter alcançado o seu objetivo (C_2). Existindo estas condições, se o fraudador for um impostor ele terá atingido seu objetivo e gerado uma transação não pertencente ao titular (ou autorizados) na conta de um terceiro (A_1) e esta transação irá alterar significativamente o valor da dimensão fractal do conjunto de transações do representado. Entretanto, o contrário pode não ser verdadeiro, ou seja, uma alteração no valor da dimensão fractal pode não significar necessariamente uma fraude, mas indica a suspeita de fraude. A alteração pode ter ocorrido devido a uma mudança no comportamento do titular da conta (ou de seus autorizados).

Sendo o fraudador um vigarista, então deverá ter ocorrido $E_{2,iii}$. Ocorrendo $E_{2,iii}$ e respeitando as condições C_1 e C_2 , tem-se que: $\sum_{i=1}^n a_i > \sum_{j=1}^n f_j (A_2)$. Entretanto, não somente o

valor total do conjunto de transações consideradas do vigarista será menor do que deveria, mas também a variação destes valores será menor do que a de um consumidor normal, ou seja, o valor da conta do vigarista oscilará menos do que a de um não vigarista^{xi}. Com isso: (H_2) a dimensão fractal do conjunto de transações do vigarista terá um valor próximo de 1.

Também, neste caso, o inverso ($D_q \approx 1 \rightarrow E_{2,iii}$) pode não ser verdadeiro, ou seja, o fato da dimensão fractal de um conjunto de transações ser bem próximo de 1 pode não significar a ocorrência de uma fraude. O titular da conta pode ser uma pessoa muito controlada ou a conta pode estar inativa ou com pouco uso, nestes casos a variação seria pequena e D_q seria próximo a 1, mas não teria ocorrido uma fraude. Contudo, as transações pertencentes a um conjunto de transações com D_q próximo a 1 merecem investigação por indicar a suspeita de ação de vigaristas.

O quadro da Figura 7 apresenta um resumo das hipóteses a serem investigadas neste trabalho, bem como a base teórica de cada uma delas.

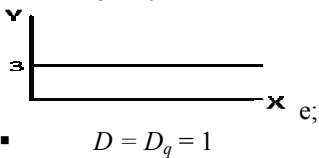
Base teórica	Hipótese
<p>Se:</p> <ul style="list-style-type: none"> os conjuntos de dados reais, que apresentam um número finito de pontos são estatisticamente auto-similares para um determinado intervalo de uma escala se obedecerem a uma regra bem definida nesse intervalo (Lee 2005); e o valor da dimensão fractal serve como uma assinatura de uma mudança de estado do sistema” (Oh & Thomas, 2010, p.133). <p>Então:</p>	<p>H’ – ao criar um novo registro, para uma instância de uma base de dados, com um valor gerado para um atributo específico, por meio de uma regra diferente dos demais, o valor da dimensão fractal, deste conjunto de dados, sofrerá uma variação significativa.</p>
<p>Se:</p> <ul style="list-style-type: none"> um conjunto de dados com uma variação igual a zero produz uma linha reta com $D_q = 1$; exemplo: dado o conjunto (3, 3, 3, 3), <ul style="list-style-type: none"> $\mu = 3$, $\sigma^2 = 0$  <p>Então:</p> <p>$D = D_q = 1$</p>	<p>H’’ – para um atributo específico de uma determinada instância de uma base de dados, uma variação, em torno da média, próxima a zero produzirá um valor próximo a 1 para a dimensão fractal do conjunto de dados considerados.</p>
<p>Se:</p> <ul style="list-style-type: none"> H’; e $A'_i \cap F = B$ e $B \neq \emptyset$ <p>Então:</p>	<p>H₁ – o resultado da ação do impostor irá provocar uma alteração significativa no valor da dimensão fractal do conjunto de transações do representado.</p>
<p>Se:</p> <ul style="list-style-type: none"> H’’; e $\sum_{i=1}^n a_i > \sum_{j=1}^n f_j$; e se o valor da conta do vigarista oscilar menos do que a de um não vigarista. <p>Então:</p>	<p>H₂ – a dimensão fractal do conjunto de transações do vigarista terá um valor próximo de um.</p>

Figura 7: Quadro resumo das hipóteses para futuras pesquisas

6. CONSIDERAÇÕES FINAIS

Neste ensaio apresentou-se um conjunto de hipóteses, bem como seus fundamentos teóricos, que possibilitam a construção de dois algoritmos de fractal *data mining* para a detecção de fraudes. Um para detectar fraudes cometidas por impostores e outro para as cometidas por vigaristas. De um modo genérico, as vantagens que os métodos propostos no ensaio teoricamente teriam sobre todas as outras técnicas de DM, de um modo geral, e todos os outros algoritmos específicos de detecção de fraude existentes, seriam: i) mais alta eficácia; com boa eficiência e escalabilidade; ii) necessidade de uma mínima preparação dos dados (podendo até ser automatizada); iii) baixa necessidade de retreinamentos; iv) capacidade de trabalhar com mudanças aleatórias, inclusive as ocasionadas por variações monetárias (inflação) e crises econômicas; v) resultados facilmente interpretáveis pelo usuário; e vi) baixo custo de implantação e operação.

Já as desvantagens principais, teoricamente seriam: (a) necessidade de um grande número de transações já efetuadas pelo mesmo indivíduo; e (b) mesmo trabalhando com a aleatoriedade, nem toda a mudança significativa no valor da dimensão fractal poderá implicar em uma fraude, poderá ter ocorrido apenas uma mudança de comportamento do indivíduo. Ou nem todo o comportamento constante será de um vigarista, poderá haver casos em que simplesmente o indivíduo é muito controlado ou uma conta com baixa atividade.

Contudo, essas vantagens e desvantagens ainda estão no campo das conjecturas, o método terá de ser posto a prova em um estudo empírico.

REFERÊNCIAS

Association of Certified Fraud Examiners. (2002, march-april) *FRAUD BASICS: The Many Faces of Fraud*. Retrieved May, 15, 2009, from <http://www.acfe.com/fraud/view.asp?ArticleID=190>.

Barbará, D., & Chen, P. (2000, August). Using the fractal dimension to cluster datasets. *Proceedings of ACM SIGKDD Conference on Knowledge Discovery in Data Mining*, Boston, MA, USA, 6. Retrieved April, 04, 2010, from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.22.2543&rep=rep1&type=pdf>

Barbará, D.; & Chen, P. (2005).. Fractal Mining. Self Similarity-based Clustering and its applications. In O. Maimon & L. Rokach (Eds.). *Data Mining and Knowledge Discovery Handbook* (chap. 28, pp. 628-647). Nova York: Springer.

Barbará, D., & Nazeri, Z. (2000). *Fractal Mining of Association Rules Over Interval Data* (Technical Report), Fairfax, Virginia, USA, George Mason University, ISE Dept. Retrieved December, 10, 2010, from <http://citeseerx.ist.psu.edu/viewdoc/versions?doi=10.1.1.24.3356>

Bhargava, B., Zhong, Y., & LU, Y. (2003, September). Fraud Formalization and Detection. *Proceedings of the Data Warehousing and Knowledge Discovery 5th International Conference*, DaWaK, Prague, Czech Republic, 5. Retrieved April, 04, 2010, from <http://www.springerlink.com/content/2nv0u647h8mm4u5p/>

Bolton, R.J., & Hand, D. J. (2002). Statistical Fraud Detection: A Review. *Statistic Science*, 17(3), 235-255. Retrieved June 30, 2008, from <http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.ss/1042727940>

Cammack, D. Mitigating Losses from Health Care Fraud and Abuse. *Healthcare Management Forum-ING-RE* 6(1), 1-5. Retrieved May 26, 2008, from http://www.ingreinsurance.com/pubs/group/man_care/index.html.

Capra, F. (2008). *A Teia da Vida: Uma nova compreensão científica dos sistemas vivos*. (11a Ed. , N.R. Eicheberg). São Paulo: Cultrix. (Obra original publicada em 1996)

Chen, M., Han, J., & Yu, P.S. (2008). Data Mining: An Overview from a Database Perspective. *IEEE Transactions on Knowledge and Data Engineering*, 6(8), 866-883. Retrieved October, 10, 2010 from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.12.7980&rep=rep1&type=pdf>

Daskalaki, S., Kopanas, I., Goudara, M., & Avouris. (2003). Data mining for decision support on customer insolvency in telecommunications business. *European Journal of Operational Research*, 145(2), 239-255. Retrieved June, 20, 2009, from <http://dblab.mgt.ncu.edu.tw/%E6%95%99%E6%9D%90/2004%20Data%20Mining/2004-45.pdf>

Faloutsos, C., Kamel, I. (1994 May). Beyond Uniformity and Independence: Analysis of R-trees Using of Fractal Dimension. *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, Minneapolis, Minnesota, EUA, 30. Retrieved December, 05, 2010, from <http://repository.cmu.edu/compsci/582/>

- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3), 37-54. Retrieved June, 20, 2009, from <http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf>.
- Fayyad, U., Stolorz, P. (1997). Data mining and KDD: Promise and challenges. *Future Generation Computer System*, 13(2), 99-115. Retrieved June, 20, 2009, from <http://wenku.baidu.com/view/93f6571b227916888486d737.html>.
- Fraude. (n.d.). Em *Wikipédia: a enciclopédia livre*. Recuperado em 12 maio, 2008, de <http://pt.wikipedia.org/wiki/Fraude>.
- Graycar, A., & James, M. (2004, june). Older People and Consumer Fraud. *Proceedings National Outlook Symposium on Crime in Australia*, Canberra, ACT, Australia, 4. Retrieved May, 26, 2008, from <http://www.aic.gov.au/conferences/outlook4/Graycar2.pdf>
- Jain, T. Fraud as a Vitiating Factor to Enforcement of Foreign Judgments: A Comparative Analysis of India, UK and US. *Social Science Research Network*. Retrieved May, 26, 2008, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1098293.
- Jelinek, H.F., & Fernandez, E. (1998). Neurons and fractals: how reliable and useful are calculations of fractal dimensions? *Journal of neuroscience methods*. 81(1-2), 9- 18. Retrieved October, 13, 2010, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1098293.
- Johnson, G. Jr., & Santos, E. Jr. (2004). Detecting Deception in Intelligent Systems I: Activation of Deception Detection Tactics. In *Advances in Artificial Intelligence* (Vol. 3060, pp 339-354) (Lecture Notes in Computer Science). Springer. Retrieved May, 26, 2008, from <http://www.springerlink.com/content/v0a30kthy8wa8fyr/>.
- Kingston, J.; Schafer, B.; & Vandenberghe, W. (2005). No Model Behaviour: Ontologies for Fraud Detection. In *Law and the Semantic Web* (Vol. 3369, pp.233-247) (Lecture Notes in Computer Science). Springer. Retrieved August, 15, 2008, from <http://www.springerlink.com/content/2j4ylumabgqe0850/>
- Kirkos, E., Spathis, C., Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert System with Applications*, 32(4), 995-1003. Retrieved August, 15, 2008, from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.81.8428&rep=rep1&type=pdf>
- Lee, D. H. (2005). *Seleção de atributos importantes para a extração de conhecimento de base de dados*. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, Brasil, Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-22022006-172219/pt-br.php>
- LopesDeSá, A., & HOOG, W. A. Z. (2008). *Corrupção, Fraude e Contabilidade*. (2a ed., Cap. 1, pp 1-55). Curitiba: Juruá.
- Lorenz, E. N. (1963). Deterministic Non periodic Flow. *Journal of the atmospheric sciences*, 20(2), 130-141. Retrieved January, 10, 2011, from <http://journals.ametsoc.org/doi/pdf/10.1175/1520-0469%281963%29020%3C0130%3ADNF%3E2.0.CO%3B2>.
- Mandelbrot, B.B. (1985). Self-Affine Fractals and Fractal Dimension. *Physica Scripta*, 32(4), 257-260. Retrieved February, 05, 2009, from <http://iopscience.iop.org/1402-4896/32/4/001>
- Oh, H.S. & Thomas, R.J. (2010). Nonlinear time series analysis on the offer behaviors observed in an electricity market. *Decision Support Systems*, 49(2), 132-137. Retrieved January, 10, 2011, from <http://portal.acm.org/citation.cfm?id=1775047>.

Parodi, Lorenzo. *Introdução ao Mundo das Fraudes*. Monitor das Fraudes. Recuperado em 02 abril, 2011 de <http://www.fraudes.org/showpage1.asp?pg=2>.

Podgor, E.S. (1999, abril). Criminal Fraud. *American University Law Review*, 48(4), 729-768. Retrieved May, 26, 2008, from <http://digitalcommons.wcl.american.edu/aulr/vol48/iss4/1/>

Provost, F. (2002) Comment. In: Bolton, R.J., & Hand, D. J. (2002). Statistical Fraud Detection: A Review. *Statistic Science*, 17(3), 235-255. Retrieved June 30, 2008, from <http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.ss/1042727940>

Savi, M. A. (2006). *Dinâmica não linear e caos* (pp. 15-25, 53-57). Rio de Janeiro: e-papers.

Strogatz, S. H. (1994). *Nonlinear Dynamics and Chaos* (pp. 1-11). Nova York: Perseus Books.

Tang, J., Chen, Z., Fu, A.W., & Cheung D.W. (2006). Capabilities of outlier detection schemes in large datasets, framework and methodologies. *Knowledge and Information System*, 11(1), 45-84. Retrieved June 30, 2008, from <http://www.cse.cuhk.edu.hk/~adafu/Pub/outlier-KAIS.pdf>

Vasiu, L, & Vasiu, I. (2004, January). Dissecting Computer Fraud: From Definitional Issues to a Taxonomy. *Proceedings of Hawaii International Conference on System Sciences*, Big Island, Hawaii, USA, 37. Retrieved September 20, 2009, from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.108.9517&rep=rep1&type=pdf>

Viane, S., Derrig, R. A., Baesens, B., & Dedene, G. (2002). A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. *The Journal of Risk and Insurance*, 69(3), 373-421. Retrieved September 18, 2009, from <http://onlinelibrary.wiley.com/doi/10.1111/1539-6975.00023/full>.

Witten, I.H., & Frank, E. *Data Mining, Pratical Machine Learning Tools and Techniques*, (2a ed.). San Francisco: Elsevier, 2005.

Zero Hora. (2009, fevereiro 15). Combate às fraudes virtuais. *Zero Hora*.ZH Classificados, Informática, p.1.

ⁱ Por definição, uma fraude não é um crime “autor-revelável”, como um assalto a banco, por exemplo, as perdas só podem ser contabilizadas após sua detecção, o que pode demorar anos, como no caso da múmia de Íbis (Figura 1), descoberta após três mil e quinhentos anos, aproximadamente. Assim, as perdas totais reais devido a fraudes não podem ser medidas, somente estimadas (Cammcak, 2004).

ⁱⁱ Esquema piramidal fraudulento de investimento que consiste em pagamentos de altos rendimentos aos primeiros investidores por meio do recebimento de aplicações dos investidores subseqüentes.

ⁱⁱⁱ O principal, e mais famoso, artigo que trata do tema no código penal brasileiro é o Art. 171: Obter, para si ou para outrem, vantagem ilícita, em prejuízo alheio, induzindo ou mantendo alguém em erro, mediante artifício, ardil, ou qualquer outro meio fraudulento.

^{iv} Entretanto, para Witten and Frank (2005) o mais frequente é que os processos de descoberta de padrões em dados sejam semiautomáticos.

^v Caso o usuário não tenha nenhuma propriedade conceitual em mente, ele poderá vir a considerar qualquer resultado como aceitável (TANG et al., 2006).

^{vi} Principalmente aos que visam detectar fraudes em tempo real, pois estes devem fornecer o resultado em segundos. Para a mineração de dados de um supermercado, por exemplo, não é necessário tamanha eficiência, podendo o algoritmo demorar minutos e, no caso de minerações acadêmicas é permitido até um pouco mais.

^{vii} Detectar, de acordo com o dicionário da língua portuguesa, é descobrir, revelar ou determinar a existência ou a presença de alguma coisa.

^{viii} É necessária a **prova** de que a vantagem foi obtida por meio da ludibriação, enganação ou mentira.

^{ix} Como o termo é amplamente utilizado na literatura sobre o assunto, neste estudo continuará sendo empregado detecção de fraude para a tarefa de identificar transações suspeitas de serem fraudulentas.

^xNo caso do vigarista, somente o terceiro efeito possível da sua ação é passível de ser investigado por algoritmos que varrem bases de dados que registram transações. Os outros dois são conceituais e envolvem a análise de documentação.

^{xi} Do contrário o vigarista não estará obtendo vantagem com a execução da fraude e considerando R_2 não terá havido fraude.