Análise Estatística Multivariada de Incidentes de Segurança em Redes de Computadores

Érico M. H. do Amaral[1], Cláudio Bastos [1], Maria A. Figueiredo[1], Roben C. Lunardi[2], Adriano M. Souza[1], Raul Ceretta Nunes[1]

[1]Programa de Pós-Graduação em Engenharia de Produção
Universidade Federal de Santa Maria (UFSM)

[2]Programa de Pós-Graduação em Computação
Universidade Federal do Rio Grande do Sul (UFRGS)

ericohoffamaral@gmail.com,claudio68@ibest.com.br,
{mariaangelicafo, robenlunardi, amsouza.sm}@gmail.com,
ceretta@inf.ufsm.br

Resumo. *O gerenciamento de segurança em sistemas de informação exige muito estudo e observação de toda e qualquer forma de incidente. O reporte dessas informações é uma prática crescente, o que conseqüentemente leva ao incremento da base de dados que são relevantes para o tratamento de muitas ocorrências. Para que estas informações sejam tratadas corretamente e de forma pró-ativa, este artigo propõe a utilização de métodos estatísticos multivariados para definir um padrão comportamental de incidentes reportados pelo CERT-BR. O resultado mostra que é possível demonstrar de forma não subjetiva a correlação entre os incidentes de segurança.*

1. Introdução

O desenvolvimento da tecnologia da informação (TI) e de ambientes informatizados com alta conectividade tem provocado alterações no panorama organizacional das empresas [Tanembaum 2004]. O entendimento de que o bom funcionamento da infra-estrutura de TI é um ponto crítico tanto para a concretização de negócios como para a tomada de decisões é um exemplo destas alterações. Desta forma, observa-se um aumento substancial no volume das transações e serviços envolvendo informações confidenciais em ambientes de domínio público, como a Internet, as quais estão vulneráveis e continuamente expostas a ataques e atividades maliciosas, o que provoca um aumento no grau de risco. Como resultado, para proteger seus negócios as organizações precisam se preocupar cada vez mais com incidentes de segurança

O CERT.br [CERT.BR 2008] disponibiliza a todos usuários da Internet uma base de dados com informações quantitativas sobre reportes de incidentes de segurança mensais desde 1999. Segundo este repositório os incidentes mais comuns são a disseminação de vírus, *worms*, ataques a servidores web, negação de serviços, varreduras de portas, tentativas de invasão e fraude. Porém a interpretação destes reportes para possíveis ações de mitigação de vulnerabilidades é responsabilidade dos usuários da Internet, que quando administradores de redes ou gerentes de TI necessitam de ferramentas de apoio a decisão.

Este artigo explora o uso da análise estatística multivariada, especificamente a Análise de Agrupamento e Análise Fatorial, como ferramenta para avaliação de reportes

de incidentes. Para tal está organizado como segue. A seção 2 conceitua incidente e descreve a estrutura de quem os reporta no Brasil. A seção 3 descreve os principais métodos de análise multivariada. A seção 4 descreve os experimentos realizados e finalmente a seção 5 apresenta as considerações finais.

2. Incidentes e Reportes de Segurança no Brasil

Um incidente é o fato decorrente de uma atividade maliciosa que explora uma ou mais vulnerabilidades, levando à perda de princípios da segurança da informação [Sêmola 2001]. Um incidente pode gerar impactos no processo de negócio das organizações, devendo ser reportado afim de que seja analisado e avaliado em relação a seu grau de impacto. Neste contexto, são inúmeros os alertas recebidos pelos administradores de redes e/ou gerentes de TI informando sobre estas atividades em seus sistemas.

Os incidentes de segurança são gerenciados e reportados por alguns grupos conhecidos genericamente por CSIRT (*Computer Security Incident Response Team*). No Brasil, o grupo de resposta a incidentes de segurança na Internet é o CERT.br, mantido pelo Núcleo de Informação e Coordenação do Ponto br (NIC.br) e pelo Comitê Gestor da Internet no Brasil (CGI.br). O CERT.br notifica incidentes de segurança e prove a coordenação e o apoio necessário ao gerenciamento de resposta a incidentes, colocando as partes envolvidas em contato quando necessário [CERT.BR 2008].

De acordo com o CERT.br os incidentes comumente reportados na Internet brasileira são: vírus, *worms*, ataques DOS, invasão, ataques web, *scan* e fraudes virtuais.

3. Análise Multivariada

A análise multivariada [Hair 2005] é uma ferramenta estatística indicada quando o número de variáveis envolvidas é grande e o pesquisador não percebe como as observações podem ser resumidas em uma ou mais características que condensem o volume de informações. Seu objetivo é processar informações de modo a simplificar a estrutura dos dados e a sintetizar informações das amostras, facilitando o entendimento do relacionamento existente entre as variáveis do processo. Há diversas técnicas para realizar análise multivariada, algumas estão descritas a seguir.

3.1. Análise de Agrupamentos

A análise de agrupamentos, ou de conglomerados, tem por objetivo agrupar as variáveis conforme sua proximidade ou suas características comuns, buscando mostrar a homogeneidade dentro do grupo e a heterogeneidade entre os grupos. Conforme Hair [Hair 2005], esta análise pode ser definida como “um grupo de técnicas multivariadas cuja finalidade primária é agregar objetos com base nas características que eles possuem”. Da mesma forma, Malhotra [Malhotra 2001] define a análise de conglomerados como uma técnica usada para classificar objetos ou casos em grupos relativamente homogêneos chamados conglomerados.

A formação dos diversos grupos homogêneos (agrupamentos) pode ter como objetivo um estudo exploratório que visa a formação de classes de objetos, uma simplificação das informações, ou ainda a identificação de relacionamentos entre as observações. Os grupos, que são obtidos através de uma ou mais técnicas de análise de agrupamentos, devem apresentar tanto uma grande homogeneidade interna (dentro de cada grupo), como uma grande heterogeneidade externa (entre grupos). Portanto, se a

classificação for bem sucedida, quando representados em um gráfico, os objetos dentro dos grupos estarão muito próximos e os grupos diferentes ficarão afastados.

A análise de agrupamentos é realizada em três passos: primeiro seleciona-se as variáveis ou características que serão determinantes na formação dos agrupamentos; em seguida é feito um tratamento das variáveis (transformações e padronização); para finalmente ser escolhida a medida de similaridade a ser adotada. Também podem ser empregadas várias técnicas aglomerativas para a análise, que são divididas em hierárquicas ou não hierárquicas.

De posse dos resultados das análises os agrupamentos podem ser caracterizados, interpretados e validados com base no perfil das suas observações.

3.2. Análise Fatorial

A Análise Fatorial (AF) é uma técnica estatística que envolve um processo composto de vários métodos estatísticos multivariados com o propósito de definir a estrutura subjacente em uma matriz de dados, ou seja, a AF é um nome genérico que denota uma classe de processos utilizados essencialmente para redução e sumarização de dados [Malhotra 2001].

Como uma técnica estatística multivariada, a AF realiza a redução e sumarização de dados explicando-as através de suas dimensões comuns, que são os fatores calculados. É possível encontrar tantos fatores quanto forem as variáveis, porém há uma perda de objetividade ao se utilizar um grande número de fatores. A redução do número de fatores se dá através da análise de correlação.

Para a execução da Análise Fatorial, alguns passos são necessários como: a formulação do problema, determinar a matriz de correlação e calcular os autovalores que fornecerão elementos para encontrar a variância total explicada por cada um dos fatores encontrados, através da utilização do software estatístico Statistica 7.0.

3.3. Análise de Componentes Principais

Análise de Componentes Principais é utilizada para obter a participação de cada variável na construção de um índice ou na busca pelos resultados. Segundo Souza [Souza 2000], a idéia matemática do método de Análise de Componentes Principais é conhecida há muito tempo, apesar do cálculo das matrizes dos autovalores e autovetores não ter sido possível até o advento dos computadores. Esta análise é utilizada na redução das variáveis e para identificar as variáveis que possuem maior influência.

4. Análise Estatística e Multivariada dos Incidentes de Segurança

A hipótese básica deste trabalho é que os inúmeros alertas (reportes) sobre incidentes de segurança recebidos pelos administradores de redes e/ou gerentes de TI confundem os administradores e gerentes, mas que a análise multivariada pode auxiliar na identificação de correlações entre os incidentes.

Para comprovar a veracidade da hipótese, o experimento realizado considerou os incidentes de segurança reportados pelo CERT.br como ataques ocorridos na Internet brasileira desde o ano de 1999 até o final de 2007. Os dados representam 96 (noventa e seis) observações para cada categoria de ataque (*Worm*, DoS, Invasão, AW, *Scan* e Fraude) em um período de oito anos, com coletas mensais. Uma amostra destes dados e de sua organização está apresentada na Tabela 1, com a seguinte composição: a primeira coluna apresenta os meses, e as demais mostram os tipos de incidentes e o número total de reportes a cada mês.

Tabela 1: Dados de incidentes reportados ao CERT.br

Mês/Ataque	Worm	DOS	Invasão	AW	Scan	Fraude
jan/99	0	7	14	22	67	0
fev/99						
mar/99						
abr/99						
mai/99						
jun/99						
jul/99						
ago/99						
set/99						
out/99						
nov/99						
dez/99						

Mês/Ataque	Worm	DOS	Invasão	AW	Scan	Fraude
jan/00	0	11	3	57	217	0
fev/00						
mar/00						
abr/00						
mai/00						
jun/00						
jul/00						
ago/00						
set/00						
out/00						
nov/00						
dez/00						

Mês/Ataque	Worm	DOS	Invasão	AW	Scan	Fraude
jan/06	1.358	71	24	32	3.274	3.687
fev/06						
mar/06						
abr/06						
mai/06						
jun/06						
jul/06						
ago/06						
set/06						
out/06						
nov/06						
dez/06						

Mês/Ataque	Worm	DOS	Invasão	AW	Scan	Fraude
jan/07	18.109	0	8	21	3.132	2.911
fev/07	9.302	0	4	19	3.907	2.250
mar/07	9.124	1	3	125	4.850	2.530
abr/07	6.416	11	6	147	4.584	2.988
mai/07	5.776	179	5	139	3.107	4.417
jun/07	4.621	8	15	121	2.092	3.881
jul/07	4.349	2	18	88	1.779	4.622
ago/07	4.831	0	2	114	1.831	4.659
set/07	4.620	480	33	152	2.227	4.394

Para inferir o inter-relacionamento das variáveis em estudo (reportes de intrusão) o experimento realizado consistiu em uma análise estatística em 3 etapas: Estatística Descritiva, Análise de Agrupamentos e Análise Fatorial. O processamento dos dados, a geração dos resultados e os gráficos obtidos na análise em cada etapa foram efetuados com o auxílio do software de estatística STATISTICA 7.0 e da planilha Excel 2003.

4.1. Estatística Descritiva

A primeira parte do experimento consistiu na realização de uma estatística descritiva para identificar a incidência e o comportamento dos ataques ao longo do período (de 1999 a 2007). Com o objetivo de evidenciar a distribuição das ocorrências e suas amplitudes no período pesquisado, nesta etapa foi elaborado um gráfico *box-plot* (vide figura 1), que é uma representação dos valores amostrais. O gráfico mostra as incidências dos ataques no decorrer dos oito anos pesquisados, demonstrando a magnitude dos mesmos e sua variabilidade.

A partir da figura 1 observa-se que incidentes do tipo *worm* apresentaram variações de até 18000 ocorrências em um único mês, porém o maior número de reportes se manteve abaixo de 4000 incidentes mensais. Os ataques de DOS, invasão e AW apresentaram baixa incidência, ou seja, ficaram abaixo de 500 reportes por mês. *Scan* e fraude demonstraram picos em torno de 7000 ocorrências, contudo os ataques do tipo fraude apresentaram uma mediana muito baixa, próximo a 50 reportes.

De acordo com os dados fornecidos pelo CERT.br e com base na análise descritiva realizada, foi possível verificar que a incidência de ataques vem crescendo, evidenciando uma parcela de usuários dedicados a cometer delitos no mundo virtual, mas com conseqüências reais.

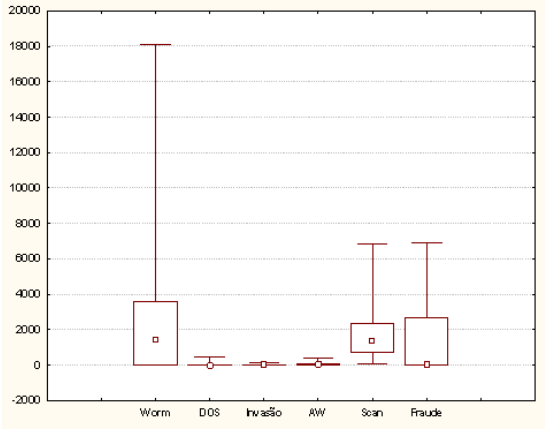

Figura 1: Gráfico de *box-plot*

4.2. Análise de Agrupamentos

Para realizar a análise de agrupamentos foi utilizado o método de Ward [Mingoti 2005], que tem como base principal os princípios de análise de variância e tende a produzir grupos com aproximadamente o mesmo número de elementos. Para o cálculo da medida de distância entre os respectivos vetores de dados (grupos) foi usado o coeficiente de "*pearson*", que mede o grau da correlação (e a direção dessa correlação - se positiva ou negativa) entre duas variáveis. No método de Ward, inicialmente cada elemento é considerado como um único agrupamento e, em cada passo do algoritmo de agrupamento, calcula-se a soma de quadrados das distâncias entre as variáveis dentro de cada agrupamento [Hair 2005].

A figura 2 mostra o dendograma formado a partir da matriz inicial dos dados mediante o método de Análise de Agrupamentos aplicado. A figura mostra a formação de três grupos distintos. O primeiro grupo é formado pelas variáveis *Scan* e *Worm*, indicando a ocorrência de ataques simultâneos e de forma crescente, acompanhando o crescimento da Internet. O segundo grupo é formado pelas variáveis representativas de ataques do tipo invasão e fraude, mostra invasões com crescimento não significativo, o que não ocorreu com as fraudes ao apresentar um crescimento maior, embora inferior ao aumento evidenciado dos ataques com *Scan* e *Worm*. O terceiro grupo é formado pelos Ataques à Web (AW) e *Denial of Service* (DOS), comprovando a semelhança de incidência desses dois tipos de ataques, não havendo ao longo do período estudado um aumento exponencial dos mesmos.

A Análise de Agrupamentos possibilitou visualizar como se processam os ataques às redes no Brasil, ficando clara a supremacia numérica dos *Worms*, seguida de grande número de *Scans*; exigindo dos administradores de TI uma maior preocupação com a vulnerabilidade da rede em relação aos mesmos. Foi possível inferir, nos *clusters,* que o segundo grupo de atividades maliciosas que mais ocorrem é formado por Invasões e Fraudes e, por último, em menor proporção ocorrem os Ataques a Web (AW) e *Denial of Service* (DOS).

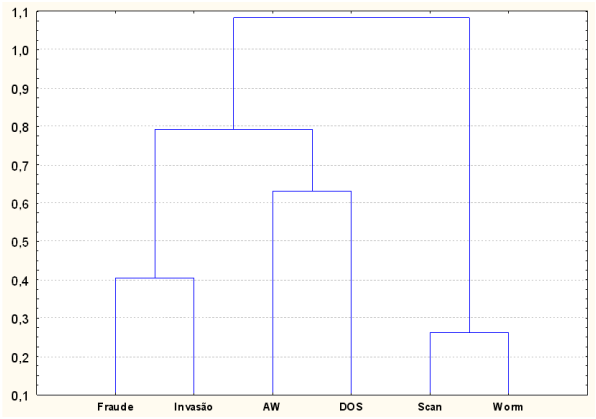

Figura 2: Dendograma com as variáveis, pelo método de Ward

4.3. Análise Fatorial e de Componentes Principais

Através da Análise Fatorial as variáveis são agrupadas em função de suas correlações e tem-se o propósito de explicá-las através de suas dimensões comuns, os fatores calculados.

Inicialmente, para realizar a AF foi calculada a matriz de correlação e a determinação dos autovalores e percentual de variância explicada por cada um (Tabela 2). Pela tabela observa-se que com três fatores pode-se explicar 82,49% da variância das variáveis sob análise. Por isto nessa análise, no lugar das seis variáveis iniciais, passam a ser utilizados apenas 3 fatores. Na tabela 2 estão demonstrados os fatores representados pelos autovalores resultantes da matriz de correlação, o percentual da variância explicado por cada um e percentual total de explicação da variância do conjunto de fatores.

Tabela 2: Autovalores e percentual de variância explicada

	Autovalores Extração dos componentes principais			
Fatores	Autovalores	% da variância explicada	Autovalores acumulados	% da variância acumulada
1	2,92	48,68	2,92	48,68
2	1,20	20,12	4,12	68,80
3	0,82	13,69	4,94	82,49

Analisando os fatores encontrados, também com o software Statistica 7.0, observa-se que o fator 1 é o mais importante para o estudo, pois é derivado do maior autovalor e possui uma variância explicada de 48,68%. As variáveis que mais contribuem neste fator são Fraude, *Scan* e *Worm*. O fator 2 é derivado do segundo autovalor e fornece uma explicação de variância de 20,12%, sendo representado pela variável *Denial of Service* (DOS). O fator 3, explica 13,69% da variância e é representado pela variável Invasão.

Observa-se na figura 3, a distribuição das variáveis no círculo de correlação, onde as variáveis mais próximas ao círculo de correlação são altamente representativas, notando-se que as Fraude, *Scan* e *Worm* estão bem próximas do círculo de correlação unitário, indicando a alta representatividade desses três elementos para o plano fatorial traçado. A relação de representatividade entre as variáveis e o fator é verificada através

da carga fatorial calculada e assinalada no plano, através de uma linha traçada formando um ângulo de 90° em relação ao eixo do fator, ligando o ponto que representa a variável no plano ao eixo fatorial. Quanto maior o valor deste ponto no eixo fatorial, maior é a representatividade da variável. Neste estudo, Fraude apresenta o maior valor em módulo, no eixo horizontal, em comparação com as demais variáveis; mostrando ser a de maior importância para este fator. Para esta análise deve-se considerar o valor em módulo, não importando se os números indicam valores positivos ou negativos. *Scan* e *Worm*, pela ordem, têm cargas fatoriais menores e, portanto não possuem a mesma representatividade; tendo *Worm* entre as três variáveis mais representativas, a menor influência para o fator. O fator 2, representado no eixo vertical, tem percentual de variância explicada de 20,12% e a variável com maior representatividade é *Denial of Service* (DOS); pois possui a maior carga fatorial neste eixo.

Uma das principais utilizações do círculo unitário de correlação é realizar a sua sobreposição sobre o plano fatorial; possibilitando com isso, identificar visualmente, as variáveis relacionadas com os casos em estudo.

Analisando-se as figuras 3a e 3b simultaneamente, verifica-se que no mês de setembro de 2007, houve a maior incidência de *Denial of Service* (DOS), no período estudado e, que no mês de janeiro do mesmo ano, não foram reportadas ocorrências do tipo. Os gráficos permitem inferir que no mês de outubro/07, março/06 e maio/07 ocorreu o maior número de ataques do tipo Fraude, evidenciados no mês de dezembro de 2007 como sendo o período de maior número de Ataques à Web (AW) reportados. Os demais meses, agrupados no primeiro quadrante da figura 3b, caracterizam-se pela pouca ou nenhuma incidência de ataques reportados.

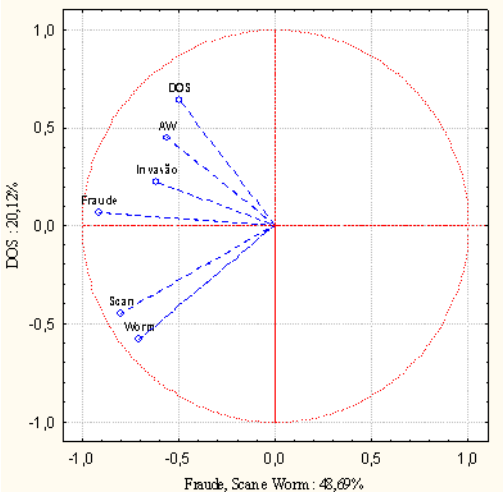

Figura 3a: Distribuição das variáveis no círculo de correlação

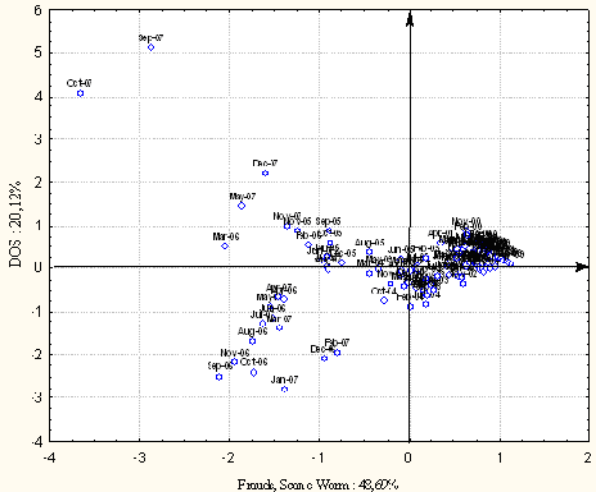

Figura 3b: Gráfico da distribuição da nuvem de pontos

Tais resultados mostram que a análise de incidentes via estatística multivariada possibilita gerar resultados claros e palpáveis sobre a relação entre os reportes, mostrando de maneira válida e não subjetiva o comportamento dos tipos de incidentes. Isto garante subsídios para elaboração de uma política de segurança de redes em que os administradores e gerentes poderão planejar e executar ações de forma a reduzir os danos e impactos provocados pelos ataques que ocorrem com maior incidência.

5. Considerações Finais

A gestão estratégica da informação tornou-se uma parte crítica e integrada a qualquer estrutura gerencial de sucesso. O desafio dos gestores de TI está na obtenção dos objetivos organizacionais específicos, os quais estão diretamente calcados sobre a informação devendo apresentar um grau satisfatório de integridade, disponibilidade e confidencialidade. Para que a informação possua essas características é importante à implementação de um conjunto de atividades de prevenção de incidentes de segurança e proteção dos sistemas de informação geridos por essas organizações. Com o foco na análise de incidentes, este trabalho aplicou um conjunto de ferramentas estatísticas com o intuito de apresentar uma abordagem prática de como podem ser abordados e gerenciados os incidentes de segurança reportados em uma empresa.

As técnicas multivariadas de Análise de Agrupamento e Análise Fatorial demonstraram ser capazes de representar as variáveis no estudo realizado, refletindo as características dos ataques e necessidades de implantação de um sistema de gestão de incidentes de segurança da informação.

O resultado da análise estatística multivariada demonstrou que para os dados analisados a ocorrência de ataques do tipo *Worm* precede uma seqüência de ataques na forma de *Scan* e Fraude (conforme demonstrado no dendograma da figura 2). Este resultado demonstra o poder a análise multivariada como ferramenta gerencial de apoio à decisão aos administradores de redes. Com ela o administrador pode disparar medidas preventivas para tentar reduzir ataques, salvaguardando as informações.

Referências

CERT.BR. (2008) Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança no Brasil, http://www.cert.br/

Tanenbaum, Andrew (2003). Redes de Computadores, 4 ª ed. – Rio de Janeiro: Campus.

Sêmola, Marcos. (2003) Gestão da Segurança da Informação - Uma visão estratégica, 9ª ed – Rio de Janeiro: Campus.

NBR/ISO/IEC 17799 (2002) Tecnologia da Informação: Código de prática para a gestão da segurança da informação. Associação Brasileira de Normas Técnicas ABNT.

Hair, J. et al. (2005) Análise Multivariada de Dados, 5ª ed. Porto Alegre: Bookman.

Malhotra, N.K. (2001) Pesquisa de Marketing. Uma orientação aplicada. Trad. Nivaldo M. Jr. e Alfredo A. de Farias. 3 ª ed. – Porto Alegre: Bookman.

Mingoti, S. Aparecida. (2005) Análise de dados através de métodos de estatística multivariada:uma abordagem aplicada. Belo Horizonte: Ed. UFMG.

Souza, Adriano Mendonça. (2000) Monitoração e ajuste de realinhamento em processos produtivos multivariados. Tese (Doutorado em Engenharia de Produção) – UFSC.

IT Infrastructure Library: ITIL Service Transition, v. 3. (2007) London: The Stationery Office, 270 p.

Peng, T., et al. (2007). Survey of Network-Based Defense Mechanisms Countering the DoS and DDoS Problems. In: ACM Computing Surveys.