Análise das Séries Temporais formadas por dados do Consórcio Brasileiro de Honeypots

Eduardo G. Barros[1], Stephan Stephany[1], Antonio Montes[2]

[1]Laboratório Associado de Computação Aplicada – Instituto Nacional de Pesquisas Espaciais (INPE)
Av dos Astronautas, 1758 – Jd Granja – CEP 12227-010 – São José dos Campos – SP – Brasil

[2]Centro de Pesquisas Renato Archer (CENPRA)
Rodovia Dom Pedro I, km 143,6 – Amarais – CEP 13069-901 – Campinas – SP – Brasil

{edugdb, stephan}@cea.inpe.br, antonio.montes@cenpra.gov.br

Abstract. *The Internet network background noise includes both malicious and non-malicious traffic. Understand its behavior is important to allow early warnings of attacks. The current work employs TCP packets observed by the Brazilian Consortium of Honeypots (CBH) and represent them as a temporal series capable of characterize the background noise. This work concluded that the noise has a gaussian distribution and, therefore, several statistical methods can be employed to predict malicious activities.*

Resumo. *O ruído de fundo da Internet inclui tráfego malicioso e não malicioso. Entender seu comportamento é importante para possibilitar a realização de previsões precoces de ataques. Este trabalho usa pacotes TCP observados pelo Consórcio Brasileiro de Honeypots (CBH) e os representa como uma série temporal capaz de caracterizar o ruído de fundo. Conclui-se que esse ruído é gaussiano e, portanto, diversos métodos estatísticos podem ser empregados para prever atividades maliciosas.*

1. Introdução

O tráfego gerado por *malwares,* juntamente com todos os demais tráfegos que possam ser caracterizados como complexos, altamente automatizados, maliciosos e mutáveis em curto espaço de tempo, constitui o que se chama de ruído ou de radiação de fundo. Segundo Pang et all (2004) ruído de fundo é todo tráfego não produtivo seja ele malicioso (tentativas de DoS, varreduras, sondagens, ...) ou benigno (tráfego gerado a partir de uma máquina ou serviço mau configurado...).

Segundo Pang et all (2004), a maioria dos estudos que envolvem a coleta de dados maliciosos ou de ruído de fundo da Internet tem sido realizada em redes telescópio – redes que monitoram tráfego enviado para porções não alocadas do endereçamento IP.

Segundo Dagon et all (2004), as estratégias para detecção de worms têm medido as taxas com que pacotes de entrada chegam às redes telescópio. Entretanto, como detectam tanto ruído como ataques, os algoritmos usados para processamento são muito sofisticados e custosos em termos de tempo.

Segundo Savage (2006), a combinação da capacidade de detecção em larga

escala da rede telescópio com a de resposta dos honeypots permite que se obtenha mais informação.

Uma infecção em um honeypot pode ser detectada imediatamente. A probabilidade, porém, de identificá-la precocemente é pequena. Supera-se essa deficiência utilizando um grande número de honeypots.

A proposta do Consórcio Brasileiro de Honeypots (CBH), Aliança de aproximadamente 40 instituições coordenadas pelo Centro de Pesquisas Renato Archer (CenPRA) e pelo Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança no Brasil (CERT.br), é de usar honeypots de baixa interatividade em endereços válidos da parcela brasileira da internet como ferramentas para geração de avisos precoces e análise de tendências (http://www.honeypots-alliance.org.br).

Os sensores do CBH monitoram essencialmente o ruído de fundo da Internet, característica primordial na tentativa de realizar previsões.

O presente artigo analisa pacotes TCP obtidos pelo CBH usando técnicas de séries temporais. Esta análise permitiu a caracterização do ruído de fundo e, conseqüentemente, a predição de eventos futuros.

2. Desenvolvimento

O período de análise se situa entre 01 de janeiro de 2005 e 30 de junho de 2006. Nesse período, 44 sensores forneceram dados para análise, sendo escolhidos 24 após aplicação de critérios pré-estabelecidos.

Os sensores constituintes do CBH fornecem dois tipos de arquivo: um completo com todo o tráfego incluindo o *payload*; e, um menor, com um resumo do tráfego. No trabalho foi usado o arquivo resumido.

Este trabalho usa o conceito de **fluxo** como substituto ao de tráfego e ao de pacotes com o seguinte significado: *Fluxo é um conjunto de 1, 2, ..., n pacotes trocados por duas máquinas dentro do contexto de uma sessão de comunicação.*

Um fluxo com um ou dois pacotes representa, geralmente, um tráfego de varredura (pacotes errados e mal configurados são de baixa freqüência e são considerados como parte do tráfego de varredura). Um fluxo com três ou mais pacotes representa uma conexão.

2.1. Conjunto de Dados

O conjunto de dados de entrada é composto por todos os fluxos TCP dos sensores selecionados do CBH, agrupados por dia e descritos como uma série temporal.

Esses dados podem embutir discrepâncias que são minimizadas com o cálculo da média aritmética simples, por dia, de cada fluxo em relação ao número de sensores que forneceram dados.

Inicialmente verifica-se a existência de sazonalidade. Agrupa-se os dados em períodos de 7 e de 30 dias (semanal e mensal, respectivamente) e possíveis candidatos a períodos de sazonalidade. Calculam-se as médias e os desvios padrão para cada grupo e faz-se a representação gráfica da média contra o desvio padrão, como ilustrado na Figura 1.

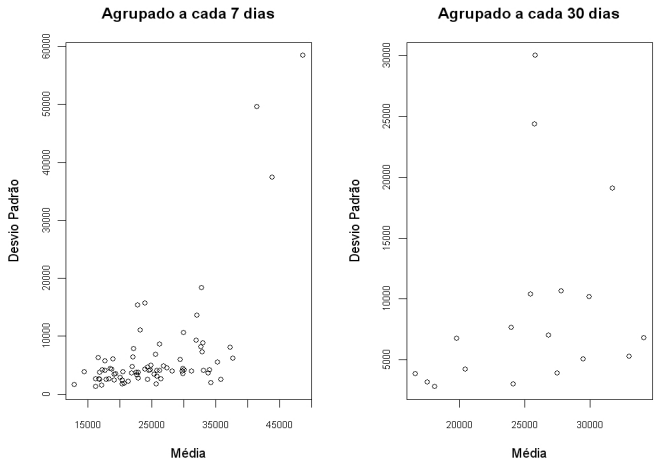

Figura 1. Média *versus* Desvio Padrão. Verificando a necessidade de transformação logarítmica.

A Figura 1, no agrupamento de 7 dias, parece mostrar uma variação linear entre a média e o desvio padrão. Este comportamento, o desvio padrão variando linearmente com a média, justifica a aplicação da transformação logarítmica para estabilizar as variâncias.

2.2. Caracterização do Fluxo de Dados

Caracterização é a representação gráfica de um conjunto de funções estatísticas que permitem a análise do sinal sendo estudado. Na visualização tem-se o sinal de entrada, o correlograma ou função de autocorrelação (ACF), o correlograma parcial ou coeficiente de autocorrelação parcial (PACF) e a distribuição de freqüências ou *spectrum*, como mostrado na Fiugra 2.

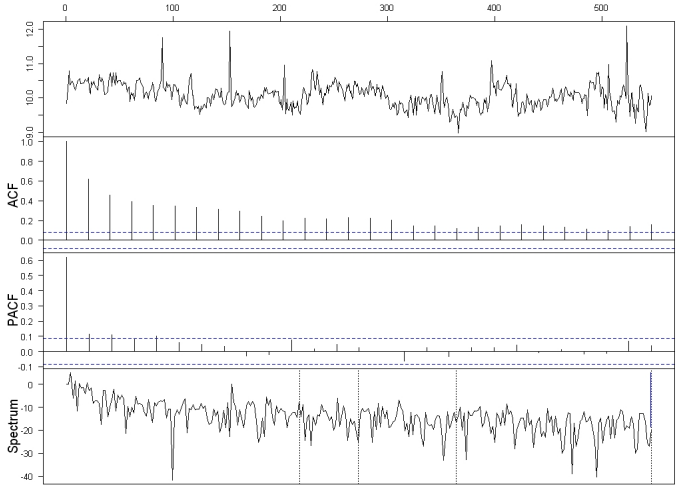

Figura 2. Caracterização do fluxo TCP diário.

O primeiro gráfico representa o sinal, isto é, o logaritmo da média aritmética dos fluxos TCP agrupados por dia.

A ACF mede a previsibilidade linear existente entre um de um instante t (x_t) a partir de outro instante s (x_s), isto é, captura a dinâmica linear dos dados.

A PACF é usada para determinar se a série é um processo auto-regressivo

genuíno (*AR(p)*). Se for, todos os valores representados na PACF são próximos de zero para todos os valores de $k > p$.

O *spectrum* apresenta todas as freqüências de Fourier.

Pode ocorrer a exibição de um gráfico extra com os p-valores determinados pelo teste de Ljung-Box caso a série seja um ruído gaussiano.

A Figura 2 mostra que o fluxo TCP diário tem uma forte correlação linear já que os valores ACF são altos. O PACF mostra que esta série é auto-regressiva. A distribuição de freqüências mostra que há um componente sazonal. E não é um ruído gaussiano porque não aparece o gráfico extra.

Além das funções apresentadas na Figura 2, usa-se, também, o histograma e o gráfico de probabilidades – técnica usada para avaliar o quanto um conjunto de dados segue uma determinada distribuição. Verificou-se o ajustamento dos dados em relação à distribuição normal. Os dados reais são confrontados com os da distribuição normal como visto na Figura 3.

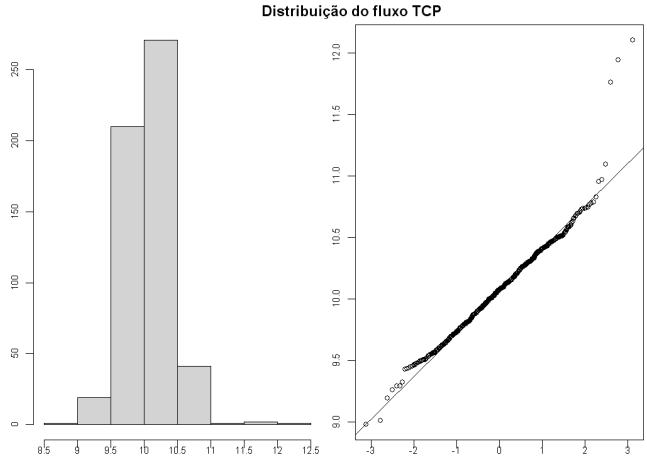

Figura 3. Histograma e distribuição de probabilidades do fluxo TCP diário.

Percebe-se, claramente, uma linha reta no gráfico de probabilidades. Conclui-se, portanto, que a distribuição é aderente à distribuição normal.

2.3. Análise do Fluxo de Dados

O objetivo da análise das séries temporais é tentar encontrar um modelo que transforme a série original em uma parcela determinística mais um ruído. Tenta-se extrair estruturas existentes na série temporal, os dados dependentes, e transformá-la em uma série de valores independentes. A extração é feita pela aplicação de modelos.

O ruído procurado não é um ruído qualquer! Deseja-se o ruído branco: *Ruído branco é uma série de variáveis aleatórias não correlacionadas, cuja expectativa é zero e possuem variância constante.*

A análise clássica tenta decompor a série temporal em uma soma, ou produto de, pelo menos, três termos: uma tendência (geralmente uma função afim), um componente sazonal (uma função periódica) e um ruído branco.

Várias técnicas podem ser aplicadas para realizar essa decomposição. Serão aplicadas, neste trabalho, a suavização exponencial tripla com o modelo aditivo e a

SARIMA (*Seasonal Auto Regressive Integrated Medium Avarage*) com o modelo multiplicativo.

2.4. Suavização Exponencial Tripla

A suavização exponencial tripla é usada quando os dados apresentam tendência e sazonalidade. Seu principal problema é a estimação de parâmetros.

A análise do ruído obtido mostrou que o mesmo ele não era ruído branco. Como visto anteriormente, o sinal original parece ser um processo auto-regressivo genuíno. Isto parece indicar que, no ruído gerado, ainda há estruturas estacionárias. Elas podem ser capturadas através de um modelo *ARMA(p,q) – Auto Regressive Medium Average.*

Há técnicas que permitem estimar os valores de p e de q a partir dos correlogramas. Porém, para garantir o melhor conjunto de parâmetros, analisou-se o AIC (*Akaike Information Criterion*) que compara os efeitos dos diferentes conjuntos de parâmetros. Neste caso, o melhor resultado apontou $p = 2$ e $q = 1$.

Na prática, o sinal foi modelado como um processo ARIMA, isto é, como processos ARMA integrados. Extraindo-se as estruturas estacionárias tem-se o ruído apresentado na Figura 4 que é um ruído branco uma vez que os p-valores apresentados no último gráfico são altos.

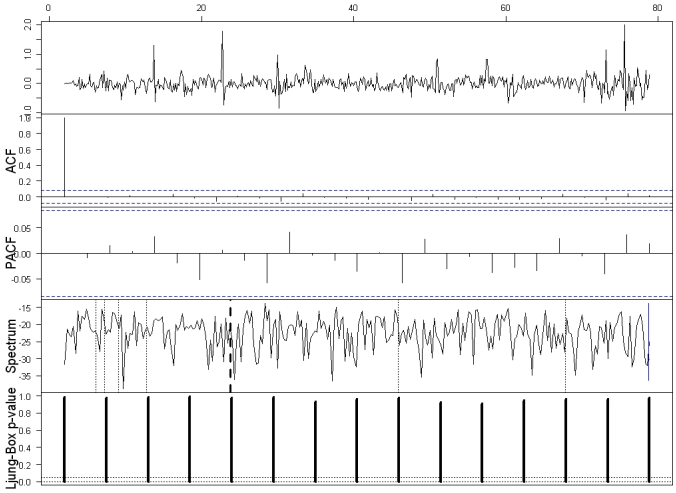

Figura 4. Ruído branco encontrado a partir da decomposição do fluxo TCP usando Suavização Exponencial Tripla e ARIMA(2, 1).

2.5. SARIMA

São processos ARIMA sazonais, isto é, tanto as partes MA ou AR podem ser sazonais. São representados geralmente por (p,d,q) x $(P,D,Q)_s$.

Neste trabalho testou-se todas as combinações de 0, 1 e 2 para todos os seis parâmetros e foram comparados os AIC. O melhor resultado indicou (2,0,1)x(1,0,1). O ruído resultante da aplicação dessa técnica tem sua caracterização na Figura 5.

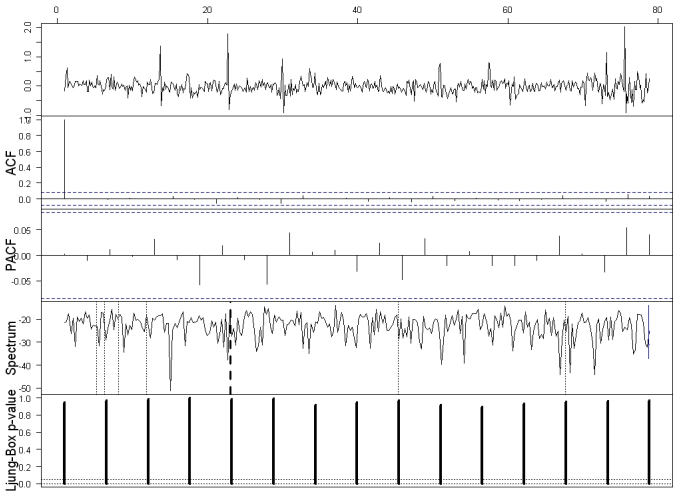

Figura 5. Ruído branco encontrado a partir da decomposição do fluxo TCP usando SARIMA.

3. Conclusão

O fluxo de dados obtido a partir do agrupamento das observações dos sensores do CBH, por dia, pode ser decomposto em séries estruturadas e em um ruído branco.

Técnicas estatísticas consagradas são empregadas para analisar os dados incluindo a previsão de eventos futuros.

As técnicas apresentadas mostraram que, independentemente do modelo sazonal adotado consegue-se realizar uma boa decomposição, permitindo previsões para eventos futuros e a prevenção de ataques. Logo, o CBH pode ser usado como um sistema de alertas precoces.

Uma limitação do presente trabalho foi o tamanho da janela de amostragem. Foi empregada a ordem de grandeza "dia". Ela foi escolhida para tornar viável a pesquisa, uma vez que há, atualmente, limitações técnicas e físicas para coleta dos dados em tempo menor.

Janelas mais reduzidas poderão aprimorar a geração de alertas precoces pelo CBH sendo uma sugestão para trabalhos futuros.

Referências

Dagon, D. et all. **HoneyStat: Local Worm Detection Using Honeypots**. Lecture Notes in Computer Science, Vol. 3224/2004. Proceedings of Recent Advances in Intrusion Detection: 7th International Symposium, RAID 2004, Sophia Antipolis, France. Ed Springer Berlin / Heidelberg . pp. 39 – 58. 2004.

Pang, R. et all. **Characteristics of Internet Background Radiation**. Proceedings of the IMC'04, Taomina, Itália, 2004.

Savage, S. et all. **Center for Internet Epidemiology and Defenses**. [online]. <www.cs.ucsd.edu/~savage/papers/CIEDProposal.pdf>. Out 2006.