Estudo da Representação do Tráfego de Rede como Série Temporal

Adriana C. Ferrari Santos[1], Lília de Sá Silva[1], José Demísio Simões da Silva[1]

[1]Instituto Nacional de Pesquisas Espaciais (INPE) – Caixa Postal 515 – 12.227-010 – São José dos Campos – SP – Brasil

aferrarisantos@gmail.com, lilia@dss.inpe.br, demisio@lac.inpe.br

Abstract. *This article presents the study in development that emphasizes the application of Computational Intelligence techniques and Time Series for representation and analysis of the computer network traffic in order to detect malicious events crossing the network traffic in an accurate mode and in good time. The theoretical bases addressed include characterization of network traffic flows as Time Series, and applications of Neural Networks in the analysis of network traffic. Finally, the steps of the study are presented: choice of attributes for use, storage and management of collected data, defining the hardware and software, the development phases and application of Computational Intelligence techniques for mapping network data.*

Resumo. *Este artigo apresenta o estudo que vem sendo desenvolvido com foco na aplicação de técnicas de Inteligência Computacional e Séries Temporais para análise e representação do tráfego de rede de computadores de modo que análises do tráfego possam resultar em descoberta de eventos anômalos de forma precisa e em tempo satisfatório. As bases teóricas abordadas incluem caracterização dos fluxos de tráfego de rede como Séries Temporais, bem como, aplicações de Redes Neurais na análise do tráfego de rede. Em seguida, apresentam-se as etapas do estudo: escolha dos atributos a serem utilizados, armazenamento e gerenciamento dos dados coletados, definição dos recursos de hardware e de software, as fases do desenvolvimento e a aplicação de técnicas de Inteligência Computacional para mapeamento dos dados da rede.*

1. Introdução

Diversas técnicas para o reconhecimento de eventos de intrusão em redes de computadores têm sido propostas ao longo dos últimos vinte anos. Entretanto, ainda hoje, observa-se a necessidade de uma técnica eficiente que proporcione a análise de grandes volumes de dados de rede em intervalos regulares de tempo, de modo que o comportamento normal do tráfego de redes monitoradas possa ser mapeado de forma precisa e em tempo satisfatório. Comparando-se dados correntes do tráfego de uma rede com a base histórica armazenada que representa o comportamento normal do tráfego desta rede, um sistema de detecção de intrusão (SDI) pode identificar as variações deste tráfego, ou seja, os eventos anômalos ou não usuais recentemente ocorridos na rede. Como diferentes serviços de rede são utilizados ao longo de dias, semanas e meses, estes geram dados de tráfego diversificados e em conjuntos volumosos, os quais

apresentam características similares e, ao mesmo tempo, o agrupamento destes dados pode prover informação do comportamento usual da rede.

Em continuidade a pesquisa sobre a aplicação de redes neurais para detecção de assinaturas em redes de computadores que vem sendo desenvolvida por um grupo de especialistas do Instituto Nacional de Pesquisas Espaciais – INPE em São José dos Campos desde 2004 [1],[2],[3],[4], o presente estudo visa aplicar técnicas de Inteligência Computacional e Sereis Temporais para armazenar os dados do tráfego normal (dados históricos), para a comparação futura com tráfego real (dados atuais) para detectar anomalias eventuais na rede de computador. Observando o tráfego e correlacionando-o a seus estados precedentes, pode ser possível predizer se o tráfego atual está se comportando de forma normal para aquele período de tempo. Esta aproximação é nomeada detecção de anomalia.

Neste artigo, as bases teóricas necessárias, incluindo conceitos de anomalias no tráfego de rede, atributos de rede, séries temporais e redes neurais serão descritas na segunda seção. As etapas para a representação do tráfego de rede, tais como a coleta de dados, seleção de atributos, armazenamento de dados, mapeamento de dados do tráfego de rede como série temporal utilizando redes neurais e recursos de *hardware* e *software* utilizados são abordados na seção 3. Finalizando, as conclusões deste trabalho são apresentadas na seção 4.

2. Bases Teóricas

2.1 Anomalias no Tráfego de Rede

Em uma rede TCP/IP, os sistemas se comunicam por meio de protocolos e as informações que trafegam pela rede são divididas em partes manipuláveis, denominadas pacotes de rede.

Os pacotes de rede TCP/IP contêm duas partes principais: cabeçalho (*header*) e dados de carga útil (*payload*). Os cabeçalhos dos pacotes são pequenos segmentos de informação, também chamados de atributos primitivos, que se localizam no início de um pacote para identificá-lo. Com base nos valores contidos em seus cabeçalhos, os pacotes são direcionados ou roteados.

Qualquer seqüência de pacotes de rede que caracterize a troca de informações entre dois *hosts* com diferentes endereços IP, durante um determinado intervalo de tempo, referente a um determinado serviço de rede, que apresente informação de inicio, meio e fim da comunicação, mesmo que toda esta esteja contida em um único pacote denomina-se, neste trabalho, uma sessão do tráfego de rede. E refere-se ao tráfego de rede como um conjunto de centenas de sessões de rede ocorridas em diferentes intervalos de tempo.

A reconstrução de cada sessão do tráfego corresponde à remontagem dos dados de pacotes coletados durante a comunicação entre dois *hosts* na rede. As análises de dados das sessões do tráfego de rede podem detectar alguns eventos anômalos não identificados por análises de medições de taxas de pacotes, como os ataques de negação de serviço [5].

Anomalias no tráfego de rede são ações diferentes observadas no comportamento normal do tráfego previamente observado, que podem ser indicativos de ataques, abuso

(ou mau uso) na rede, eventos de falha na rede, problemas de infra-estrutura na coleta de dados, entre outros. Portanto, nem toda anomalia na rede é um ataque, mas uma informação suspeita que deve ser analisada [5].

2.2 Séries Temporais

Uma série temporal é qualquer conjunto de dados ordenados cronologicamente (ao longo de dias, mês, trimestre, ano, etc.), os quais dão uma visão geral sobre o comportamento dos fenômenos estudados e permitem prever a evolução dos mesmos. [6]

Se o conjunto de dados é contínuo, a série temporal é chamada de contínua. Se o conjunto é discreto, a série temporal é chamada de discreta. Na prática, os dados do tráfego de rede observados para este trabalho são coletados em períodos de tempo discretos. Dessa forma, utiliza-se neste trabalho, séries temporais discretas ao invés de contínuas [7].

As componentes básicas de uma série temporal são classificadas em sistemáticas e não sistemáticas. As componentes sistemáticas apontam movimentos regulares, e são: a tendência ou secular, a sazão e o ciclo. O que difere sazão e ciclo é o período de avaliação (curto ou longo). Ambas definem oscilações relativamente regulares em torno da tendência. A tendência é a indicadora da direção global dos dados (ou movimento geral da variável) em um determinado intervalo de tempo. As componentes não sistemáticas apontam movimentos completamente irregulares e são denominadas: aleatórias, sendo uma mistura de perturbações repentinas, irregulares e esporádicas no movimento das séries que tipificam os fenômenos [6].

2.3 Redes Neurais

Redes Neurais Artificiais (RNA) são técnicas computacionais que apresentam um modelo inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento através da experiência. Equivale a uma estrutura de processamento de informação distribuída paralelamente na forma de um grafo direcionado, com algumas restrições e definições próprias. Os nós deste grafo são denominados elementos de processamento ou neurônios. Suas arestas são conexões que funcionam como caminhos de condução instantânea de sinais em uma única direção, de forma que seus neurônios podem receber qualquer número de conexões de entrada (estímulos do meio externo). Estas estruturas também podem possuir qualquer número de conexões nas camadas intermediárias (se houverem) e de saída, desde que os sinais nestas conexões sejam os mesmos[8].

Para a análise do tráfego de rede, muitos pesquisadores têm desenvolvido SDIs baseados em RNA supervisionadas ou não-supervisionadas, na tentativa de superar as limitações dos métodos baseados em regras e melhorar a qualidade da análise dos dados.

A maioria das pesquisas em detecção de ataques supervisionados tenta construir um modelo com base nos dados normais e então verificar se os novos dados se enquadram neste modelo. Algumas abordagens supervisionadas têm empregado as RNA para obter o modelo normal e detectar novos ataques. As RNA do tipo MultiLayer Perceptron (MLP), Radial Basis Function (RBF), Learning Vector Quantization (LVQ), bem como as Máquinas de Vetor de Suporte (SVM) têm sido utilizadas para treinamento supervisionado dos modelos de detecção de ataques em redes [5].

As pesquisas em detecção de ataques não-supervisionados tentam construir modelos que ajustam os parâmetros necessários de um detector que "por si só" é capaz de aprender sobre o comportamento normal do tráfego e identificar ataques. Algumas abordagens têm empregado algoritmos de *clustering*, tais como redes neurais SOM [5], [9]. É importante salientar que nenhuma das abordagens pesquisadas até a elaboração deste artigo, leva em conta o fator cronológico das variáveis observadas.

3. Estudo da Representação do Tráfego de Rede

Nesta etapa deverá ser definido o tipo de tráfego de rede a analisar, que depende da escolha da aplicação. Por exemplo, o tráfego http é proveniente de aplicação *Web*, enquanto o tráfego SMTP, é de aplicação de *e-mail*. Também deverão ser definidos os tipos de ataques que serão utilizados para gerar dados simulados de tráfego anômalo para testes.

Os dados analisados neste estudo, são dados reais e dados simulados na rede do Laboratório de Redes da Divisão DSS/INPE (LabRedes) e dados reais da rede do prédio Beta/INPE (Rede Beta). A seguir é apresentada a metodologia que vem sendo aplicadas no estudo da representação do tráfego de rede como série temporal.

3.1 Coleta de dados

É o primeiro passo da metodologia e consiste em capturar os dados da rede, por meio dos *softwares* tcpdump e Wireshark. Para a reconstrução de sessões TCP/IP e gravação de dados utiliza-se o sistema RECON – Sistema de Reconstrução de Sessões TCP/IP [10].

No RECON são introduzidos os dados de entrada em arquivos no formato tcpdump contendo o tráfego de rede e processados para remontagem de sessões e extração de atributos. Como saída, o RECON gera um conjunto de dados referentes às sessões, armazenados em uma estrutura de árvore binária balanceada que permite acesso facilitado, com bom desempenho. Esta aplicação está preparada para capturar dados do cabeçalho dos pacotes TCP, UDP e ICMP (em nível de Transporte), IP (em nível de Rede) e Ethernet (em nível de Enlace).

3.2 Seleção de atributos

Os atributos primitivos do cabeçalho dos pacotes, isoladamente, carregam informação semanticamente fraca. Alguns exemplos de atributos primitivos são extraídos do cabeçalho IP, tais como, endereço IP de origem e de destino, protocolo de transporte utilizado e outros exemplos são extraídos do cabeçalho TCP, como, portas de origem e de destino e *flags*. Observa-se que, para a representação do tráfego, é conveniente o uso de atributos derivados, construídos a partir do processamento de atributos primitivos, que modelam as sessões do tráfego de rede, apresentando informações de maior relevância para o mapeamento do comportamento da rede [5].

Os atributos de sessões selecionados para este estudo são derivados e correspondem aos nove atributos utilizados com êxito por [9], descritos a seguir: TMPC - tamanho médio dos pacotes de rede em *bytes* recebidos pelo cliente; TMPS - tamanho médio dos pacotes de rede em *bytes* recebidos pelo servidor; NPC - número de pacotes recebidos pelo cliente; NPS - número de pacotes recebidos pelo servidor; PPP - porcentagem de pacotes pequenos ou que tenham menos de 130 *bytes*; DIR - direção do

tráfego; TBC - total de dados (*bytes*) recebidos pelo cliente; TBS - total de dados (*bytes*) recebidos pelo servidor e DUR - duração da sessão em milissegundos.

3.3 Armazenamento dos Dados

Esta fase consiste em armazenar os atributos selecionados das sessões do tráfego para análise. O sistema de gerenciamento de bases de dados de domínio público MySQL é uma alternativa para armazenamento de dados e vem sendo utilizado neste estudo, permitindo rápida criação de bases de dados estruturadas, com recursos úteis de gerenciamento, incluindo *backups*, importação e exportação de dados.

Na construção das Tabelas do banco devem-se considerar os tipos de dados (inteiro, ponto flutuante, booleano, varchar, char) dos atributos de sessões a serem inseridos nas tabelas. Também se pode optar pela criação de um campo adicional usado para armazenar o valor de classificação (*decision*) para a sessão do tráfego: 0 (zero - para sessão normal) ou 1 (um - para sessão anômala).

3.4 Mapeamento de dados do Tráfego de Rede como Série Temporal utilizando RNA

Conforme pesquisa bibliográfica realizada, os modelos de RNAs mais aplicados para resolver problemas de previsão de séries temporais com bons resultados são as redes MultiLayer Perceptron (MLP) e Radial Basis Function (RBF). Portanto, um desses modelos de rede será implementado primeiramente no estudo.

De acordo com [7], a forma mais simples de fazer com que uma RNA realize processamento temporal é por meio de adição de memória à estrutura da rede. Estas memórias podem ser de longo prazo ou de curto prazo. A memória é inserida na rede através de atrasos de tempo que podem ser implementados no nível sináptico dentro da rede ou na camada de entrada da rede. Haykin [8], denomina esta forma de incorporar o tempo na operação da rede neural de representação implícita, a qual será adotada neste estudo, seguindo a metodologia proposta em [11]. Esta metodologia modela os dados para a entrada da rede neural de maneira a captar os componentes das séries temporal em três grupos de neurônios: o primeiro grupo recebe valores passados das séries; o segundo recebe os valores passados referentes ao mesmo mês em anos anteriores, visando captar tendências ou ciclos da série; e o terceiro grupo de neurônios recebe a sazonalidade da série, como, 100000000 para o mês de janeiro, 01000000 para o mês de fevereiro e assim sucessivamente até o mês desejado. A figura 1 apresenta de forma esquemática a utilização de tal abordagem:

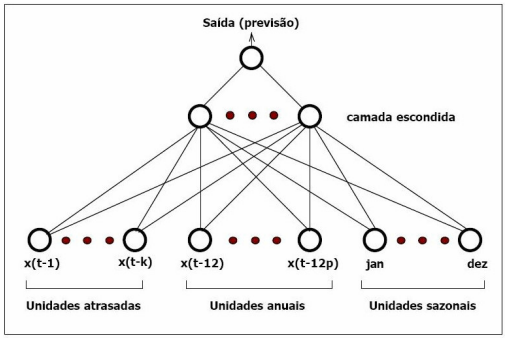

Figura 1: Modelagem dos dados do tráfego de rede como série temporal

3.5 Recursos de Desenvolvimento:

Os recursos de *hardware* e *software* que estão sendo utilizados para desenvolvimento deste estudo são: ***Software***: para a captura de pacotes de rede utiliza-se o Tcpdump (plataforma operacional Linux) e o Wireshark (plataforma Windows) e apoio na análise dos dados de pacotes em arquivos dump. Na reconstrução de sessões do tráfego da rede aplica-se o RECON, construído na linguagem C e plataforma Linux. O armazenamento dos atributos selecionados é feito por meio do MySQL em ambiente Linux). Para a análise das séries temporais (dados do tráfego de rede) pretende-se desenvolver RNA implementadas em Matlab. ***Hardware***: Para o desenvolvimento da aplicação de analise dos dados de rede, serão utilizadas duas estações de trabalho. Uma estação (CPU:xx, RAM:xx, HD:xxx) em plataforma Linux, contendo os arquivos de dados de pacotes de rede para análise, e a base de dados MySQL com os dados das sessões do tráfego e outra estação (CPU:xx, RAM:xx, HD:xxx) em plataforma Windows, com o software Matlab para desenvolvimento.

4. Conclusão

As ferramentas para analisar o tráfego de rede de computador permitem tanto detectar anomalias no ambiente, incluindo ataques e eventos não usuais, quanto a execução rápida de ações para evitar que as ameaças detectadas possam propagar através da rede.

A aplicação de Séries Temporais e Inteligência Computacional para detectar anomalia em redes de computadores é inovadora, pois permite estabelecer o comportamento normal da rede levando em conta o fator tempo. Tendo como premissa que o tráfego de rede oscila conforme o período, alguns picos são muitas vezes considerados normais e estabelecer um único comportamento para a rede durante um determinado intervalo de tempo, pode produzir muitos falsos positivos. Utilizando Séries Temporais e Redes Neurais, os comportamentos normais da rede referentes a um determinado dia e hora podem ser estabelecidos. Qualquer pacote de rede analisado que ultrapasse o limiar estabelecido para um determinado período poderá ser um possível ataque. Desta forma, a utilização de Séries Temporais e Inteligência Computacional pretende minimizar o problema da alta taxa de alarmes falsos na detecção de anomalia, e, como qualquer experimento científico, pode não gerar os resultados desejados.

Referências Bibliográficas

[1] Silva, L.S, Santos, A.C.F, Silva, J.D.S, and Montes, A. “A Neural Network Application for Attack Detection in Computer Networks”, IJCNN’2004 International Joint Conference in Neural Networks, Budapest, Hungria, 2004.

[2] Silva, L.S, Santos, A.C.F, Silva, J.D.S, and Montes, A. “ANNIDA: Artificial Neural Network for Intrusion Detection Application – Aplicação da Hamming Net para Detecção por Assinatura”, CBRN’2005 VII Congresso Brasileiro de Redes Neurais, Natal, RN, Brasil, 2005.

[3] Silva, L.S, Santos, A.C.F, Silva, J.D.S, e Montes, A. “Estudo do uso da Hamming Net para Detecção de Intrusão”, SSI’2005 VII Simpósio de Segurança em Informática, Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos, SP, 2005.

[4] Silva, L.S, Santos, A.C.F, Silva, J.D.S, and Montes, A. “Hamming Net and LVQ Neural Networks for Classification of Computer Network Attacks: A Comparative Analysis”, SBRN’2006 IX Brazilian Neural Networks Symposium, Ribeirão Preto, SP, 2006.

[5] Silva, L.S. “Uma Metodologia para Detecção de Ataques no Tráfego de Redes baseada em Redes Neurais ”, Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelo Dr. Antonio Montes e Dr. José Demísio S. Silva, INPE, SJCampos, 2007.

[6] Milone, G. “Estatística Geral e Aplicada”. São Paulo: Pioneira Thomson Learning, 2004.

[7] Oliveira, M.A. “Aplicação de redes neurais artificiais na análise de séries temporais econômico-financeiras”. Tese de Doutorado da Universidade de São Paulo, 2007.

[8] Hayking, S. “Redes Neurais - Princípios e Práticas”, Porto Alegre: Bookman, 2001.

[9] Chaves, C.H.P.C and Montes, A. (2005), Detecção de Backdoors e Canais Dissimulados, V Workshop dos cursos de Computação Aplicada (Worcap), INPE, São José dos Campos, SP, 2005.

[10] Chaves, M. H. P. Análise de Estado de Tráfego de Redes TCP/IP para Aplicação em Detecção de Intrusão. Dissertação de Mestrado em Computação Aplicada - INPE, set 2002.

[11] Fernandes, L.G.L, Navaux, P. O. A, Portugal, M.S. Previsão de Séries de Tempo: Redes Neurais e Modelos Estruturais. Pesquisa e Planejamento Econômico, Rio de Janeiro, RJ, v. 26, n.2, p. 253-276, 1996.