Características do Ruído de Fundo da Internet Brasileira a partir de dados do Consórcio Brasileiro de Honeypots

Eduardo G. Barros[1], Stephan Stephany[2], Antonio Montes[3]

[1]Programa de Doutorado em Computação Aplicada – CAP, INPE

[2]Laboratório Associado de Computação Aplicada – LAC, INPE

[3]Centro de Tecnologia da Informação Renato Archer (CTI)

edugdb@gmail.com, stephan@cea.inpe.br, antonio.montes@cti.gov.br

Abstract. *During the last years has been noticed the existence of a constant and not expected Internet traffic: the background noise or radiation. It consists of malicious traffic, like the one due to worms and bots, and/or benign traffic due to applications misconfiguration of. Generalized malicious activities are mixed with background noise. The Brazilian Honeypots Consortium (CBH) uses sensors with real IP addresses of the Brazilian portion of the Internet to capture malicious traffic. From this data is presented, in this work, an initial characterization of the background noise that allows administrators, researchers and stakeholders to perceive generalized malicious activity that are occurring in the Brazilian portion of the Internet.*

Keywords: *internet, honeypot, background radiation, characteristics, sensors*

Resumo. *Nos últimos anos tem-se percebido a existência de um tráfego constante e não esperado na internet: o ruído de fundo. Ele é composto por tráfego malicioso, como o originado por worms e bots, e/ou benigno devido a má configuração de aplicativos. Atividades maliciosas generalizadas se confundem com o ruído de fundo. O Consórcio Brasileiro de Honeypots (CBH) usa sensores com endereços IP reais da parcela brasileira da internet para capturar tráfego malicioso. A partir deste dados apresenta-se, neste trabalho, uma caracterização inicial do ruído de fundo que permite que administradores, pesquisadores e interessados perceber atividades maliciosas generalizadas ocorrendo da parcela brasileira da internet.*

Palavras-chave: *internet, honeypot, ruído de fundo, sensores*

1. Introdução

Todo tráfego que pode ser caracterizado como complexo, altamente automatizado, podendo ser malicioso ou não e que pode sofrer mutações num curto espaço de tempo constitui o que se chama de ruído ou de radiação de fundo. É possível, desde que haja conhecimento prévio do comportamento do ruído de fundo, utilizá-lo para detectar atividades maliciosas, comportamentos não esperados ou, até mesmo, atividades corriqueiras.

Segundo Savage (2006), descrever este comportamento implica em adquirir da-

dos em quantidade suficiente para garantir a representatividade do fenômeno observado.

Uma ferramenta que fornece a representatividade e a distribuição necessária é a rede telescópio – monitoramento do tráfego enviado para porções não alocadas do endereçamento IP. Estas redes são entidades totalmente passivas. Elas não respondem às solicitações dos invasores, não podem ser infectadas e não conseguem detectar varreduras que não sejam aleatórias.

Ainda segundo Savage (2006), a combinação da capacidade de detecção em larga escala da rede telescópio com a capacidade de resposta dos honeypots permite obter mais e melhores informações.

Para Baumann e Plattner (2002), honeypots são recursos computacionais cuja intenção é ser atacado e comprometido para obter mais informação sobre o atacante e as ferramentas sendo utilizadas.

O objetivo primário de um honeypot é coletar tanta informação quanto possível sobre um ataque sendo executado e características dos programas usados pelos atacantes.

Honeypots individuais não são as ferramentas apropriadas para monitoramento do ruído de fundo por não apresentarem a distribuição geográfica, isto é, a representatividade, necessária. Esta deficiência é superada com o uso de um grande número de honeypots.

O Consórcio Brasileiro de Honeypots (CBH)[1] – uma aliança de mais de 50 instituições distribuídas por todo o Brasil, coordenadas pelo CenPRA e pelo CERT.br, que usam honeypots de baixa interatividade dentro do espaço de endereçamento válido da internet brasileira – usa honeypots distribuídos para viabilizar a geração de avisos precoces e análise de tendências a partir dos dados coletados.

Como não há nenhum motivo legítimo para que se envie pacotes para os honeypots do CBH pode-se inferir que todo o tráfego capturado sugere a ocorrência de atividades não desejadas ou não esperadas, sejam elas maliciosas – negações de serviço, varreduras, monitoramento, … – ou não – pacotes gerados por má configuração de máquinas, ...

Este trabalho se propõe a apresentar características do ruído de fundo da parcela brasileira da internet usando dados coletados pelos sensores do CBH entre 01/01/2005 e 30/06/2006.

Para permitir a caracterização tem-se que, inicialmente, estabelecer um parâmetro que seja representativo do ruído de fundo. O CBH gera dois tipos de informação: a resumida e a completa. Neste trabalho usa-se a resumida.

A resumida condensa a informação coletada em um único arquivo com diferentes registros para cada tipo de protocolo:

- para o protocolo de transporte TCP, em três tipos de registro:
  - uma tentativa de conexão: este registro condensa a informação de que 1, ou 2 pacotes foram trocados. Representa toda troca de pacotes que não gerou um pacote de resposta da máquina alvo e as que tiveram pacote de resposta da máquina alvo mas não completaram o 3-way-handshake;
  - uma conexão estabelecida: ocorre toda vez que a troca de pacotes entre

1 http://www.honeypots-alliance.org.br/index-po.html

origem e destino completa o 3-way-handshake; e,
    - término de uma conexão estabelecida: ocorre quanto um flag FIN é enviado e, nesta situação é registrado o número de bytes trocados após o 3-way-handshake.
- para o protocolo de transporte UDP, em dois tipos de registro:
    - pacotes que chegam e não são respondidos; e,
    - pacotes que são respondidos. Embora o protocolo não suporte o conceito de sessão, é criada uma sessão virtual e só um registro final é informado. Nele é registrado o número de bytes.
- para o protocolo ICMP é apresentado um mesmo formato de registro que varia somente na informação do tipo.

O formato do arquivo não permite o uso do parâmetro pacote. Introduz-se, para a caracterização proposta um novo parâmetro, consistente com as informações disponíveis: fluxo. Para este trabalho fluxo é um conjunto de 1, 2, ..., n pacotes trocados por duas máquinas dentro do contexto de uma sessão, real ou virtual, de comunicação.

Para o protocolo de transporte TCP propõe-se uma subdivisão:
- fluxo com um ou dois pacotes, representado por “fluxo < 3”, representa, geralmente, um tráfego de varredura. Pacotes errados e oriundos de máquinas mal configuradas são assumidos como de baixa frequência e estão sendo considerados como parte do tráfego de varredura;
- fluxo com três ou mais pacotes, representado como “fluxo >=3”, indica uma conexão estabelecida.

2. Estado da arte sobre ruído de fundo

Sobre a coleta de dados na internet tem-se Vanderavero (2004) que propõe um sistema de honeypots, chamado de HoneyTank, para coletar grande quantidade de informação de tráfego malicioso simulando a presença de máquinas em endereços IP não usados de uma rede.

Sobre o uso de honeypots para detectar e/ou auxiliar sistemas de detecção de intrusão tem-se Yin, C. et all (2004) que apresentam uma aplicação e um projeto de honeypots capaz de ser usado em colaboração com sistemas de detecção de intrusão para gerar um sistema capaz de detectar varreduras de portas. Levine et all (2003) discutem formas de uso de honeynets com a finalidade de auxiliar o administrador de uma grande organização a identificar tráfego malicioso. Dagon et all (2004) apresentam um sistema local capaz de fornecer avisos precoces na detecção de worms, o HoneyStat, que usa honeypots modificados para gerar um fluxo de alertas preciso e com baixa taxa de falso-positivos. Grizzard et all (2005) faz uma comparação visual entre os dados capturados em uma honeynet e os dados coletados por usuários domésticos apresentando o relacionamento entre o endereço IP de origem e a porta de destino neste tráfego.

Pang et all. (2004) trata especificamente da caracterização do ruído de fundo através de uma rede telescópio. Ele usa os seguintes parâmetros para caracterizá-lo: a composição do tráfego dividido entre os protocolos TCP, UDP e ICMP; o tráfego TCP/SYN por porta de destino; a percentagem de endereços IP únicos que acessam as portas de destino; e, o tráfego TCP/SYN em portas de destino específicas.

3. Caracterização do Ruído de Fundo

Os dados coletados pelos sensores do CBH não representam uma amostra contínua ao longo do tempo, são dados discretos. Para permitir a análise de dados coletados em diferentes tempos adotou-se uma janela de observação.

Uma janela de observação, ou simplesmente janela, é um intervalo de tempo dentro do qual todas as observações, dos diversos sensores, são tratadas como sendo de um mesmo instante de tempo.

Os dados dos sensores do CBH são remetidos a uma central uma vez por dia. Isso sugere a janela dia. Esforços tem sido feitos para otimizar a transmissão de tal forma que, em breve, deverá se ter a possibilidade de se tratar com uma janela hora.

Alguns sensores do CBH podem apresentar comportamentos únicos em função do tipo de atividade da organização na qual estão inseridos. Para garantir que os dados analisados não sejam poluídos por características da organização e representem, na média, o comportamento do ruído de fundo na parcela brasileira da internet, foi estabelecida uma métrica que penaliza o conjunto de dados que mais se afasta dos seguintes critérios:

- número de dias, por ano, que cada sensor forneceu dados: quanto maior o número de dias mais relevante são os dados;
- relação entre número de fluxos TCP, UDP e ICMP: verifica a relação percentual dos fluxos TCP, UDP e ICMP em relação ao total de fluxos. A distribuição, empiricamente determinada, adotada para o trabalho considerou aproximadamente 90% de fluxos TCP, 7% de fluxos UDP e 3% de fluxos ICMP. Quanto mais os dados de um sensor se aproximam desta distribuição melhor a qualidade dos dados observados; e,
- relação entre número de fluxos < 3 e fluxos >= 3: verifica a relação percentual ente estes fluxos e o total de fluxos TCP. A distribuição, empiricamente determinada, adotada para o trabalho é de 75% de fluxos < 3 e 25% de fluxos >= 3. Quanto mais os dados de um sensor se aproximam desta distribuição melhor a qualidade dos dados observados.

O subconjunto dos sensores com menor número de penalizações é o que melhor representa o tráfego de dados na parcela brasileira da internet, isto é, eles representam melhor o que, neste trabalho, se chamou de “característica de normalidade” - o comportamento médio do ruído de fundo.

Aplicada a métrica foram selecionados 23 sensores. Eles representam 4.176 endereços IP válidos, únicos no espaço de endereçamento da internet brasileira. A partir dos dados deste conjunto de sensores foram realizadas diversas caracterizações.

A primeira caracterização foi o cálculo da média e do desvio padrão dos fluxos observados.

Foram agrupados, por dia, o somatório de todos os fluxos < 3 de todos os sensores. A amostra, neste caso é constituída de 546 observações, o total de fluxos em todos os sensores, por dia. Para esta amostra foi encontrada uma média de 330.811 fluxos/dia com um desvio padrão de 158.379. Isto é, a parcela brasileira da internet, tem um movimento médio de aproximadamente 330.811 ± 158.379 fluxos < 3, diariamente.

Outra amostra foi obtida agrupando-se os fluxos < 3 por dia, por sensor. A amos-

tra, neste caso, é constituída por 546 observações, um para cada dia, para cada um dos 23 sensores totalizando 12.558 observações. Devido a erros tais como erros de transmissão dos dados, paradas para manutenção, etc, obteve-se somente 9.511 observações. Para esta amostra obteve-se uma média de 14.383 fluxos <3 diários, por sensor, com um desvio padrão de 33.479.

Não foram determinadas outras médias, com outros protocolos, por eles constituírem uma amostra muito menor do que a de fluxos < 3.

A segunda caracterização foi a distribuição dos fluxos, por protocolo, na parcela brasileira da internet a partir dos dados consolidados do CBH (Figura 1).

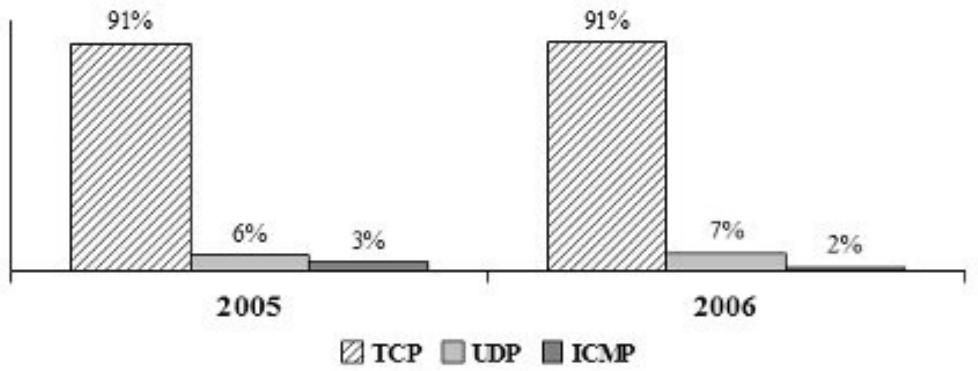

Figura 1. Distribuição de fluxos, por protocolo, na parcela brasileira da internet.

Verifica-se que os fluxos do protocolo TCP são muito maiores do que a soma dos demais tendo se mantido estável ao longo dos dois anos de observações.

A terceira caracterização, a relação entre fluxos < 3 e fluxos >= 3 (Figura 2).

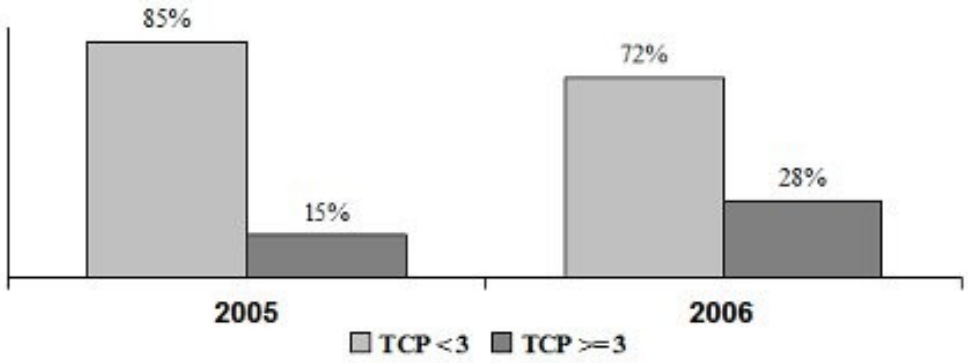

Figura 2. Distribuição de fluxos TCP: com menos de 3 e com 3 ou mais pacotes.

De posse destes dados considerou-se como “característica de normalidade” uma distribuição de fluxos TCP muito maior do que os demais (≈ 90%); fluxos UDP e ICMP pequenos com o UDP (≈ 7%) superior ao ICMP (≈ 3%); fluxo <3 (≈ 80%) muito maior do que o fluxo >= 3 (≈ 20%); e uma média diária, por sensor, de 14.383 ± 33.479 fluxos < 3 por sensor, por dia.

Outras informações podem ser agregadas à caracterização embora, a partir deste ponto, não sejam características constantes pois dependem das vulnerabilidades descobertas à época, da facilidade ou não de se usar um certo roteador, e assim por diante. Ainda assim são interessantes para que se tenha uma ideia do comportamento do ruído

de fundo.

Observou-se a quantidade de portas acessadas. Verificou-se que um grande número de portas são acessadas, porém, de forma tão esporádica que não possuem relevância estatística para o trabalho. Para evitar a poluição dos dados adotou-se critérios para seleção do fluxo, isto é, se ele será aceito na amostra a ser analisada. São os seguintes os critérios usados:

- se há fluxo continuado à porta por pelo menos 3 dias consecutivos (dia atual = D, D-1 e D-2) o fluxo é considerado;
- se há fluxo dirigido a uma porta de forma não continuada. Neste caso compara-se o total de fluxos destinados a esta porta ao longo de todo o tempo de observação com um limite[2].

Aplicado o critério e considerando somente as portas acessadas por fluxos < 3 verificou-se que 6.923 portas distintas, desde a porta 0 até a 65.506, foram acessadas. As portas mais acessadas foram listadas na Tabela 1.

Tabela 1. Distribuição de fluxo < 3, por porta.

Porta	***Fluxo Acumulado***	***Porta***	***Fluxo Acumulado***
445	40,7%	143	86,5%
139	64,3%	1.080	87,5%
135	70,7%	10.000	88,4%
80	76,6%	3.306	89,1%
1.433	81,7%	25	89,7%
4.899	85,4%	42	90,3%

Verificou-se que, embora 6.923 portas distintas tenham sido acessadas, 90% do fluxo < 3 foi direcionado para somente 12 portas. A porta 445 foi a mais varrida no período com quase 4,5x mais fluxo do que a segunda colocada, a porta 139 – deveu-se a uma descoberta de uma vulnerabilidade no protocolo para troca de mensagens SMB sobre TCP e UDP (microsoft-ds).

É interessante, também, notar que além das portas já consagradas em que se esperava encontrar tráfego (21, 22, 25, 80, ...) algumas portas com muito fluxo não são tão comuns ou mesmo não possuíam um serviço atribuído, segundo as portas atribuídas pela IANA.

Quando considerado o fluxo >= 3 verificou-se que o número total de portas foi bem menor, um total de 2.788, e, assim, não se aplicou nenhum tipo de critério. A distribuição encontrada (Figura 2) mostra que o fluxo dirigido somente a 13 portas foi suficiente para que se tivesse mais de 97,5% do total de fluxos >= 3.

Tabela 2. Distribuição de fluxo >= 3, por porta.

Porta	***Fluxo Acumulado***	***Porta***	***Fluxo Acumulado***
135	30,7%	9898	94,0%
1433	49,3%	4899	95,1%
1080	64,3%	4444	96,1%
139	77,2%	22	96,7%

2 O limite usado foi o número de janelas consideradas. Se o valor observado for menor do que o limite ele não é aproveitado, caso contrário o é.

Porta	*Fluxo Acumulado*	*Porta*	*Fluxo Acumulado*
445	87,7%	3306	97,2%
1025	90,3%	3380	97,6%
80	92,4%		

Observou-se, também, a quantidade de IP únicos que geraram fluxo para os sensores do CBH.

Para este trabalho é importante definir o que se entende por IP único uma vez que, na internet, não há como se garantir que um endereço remoto corresponda a uma única máquina. O uso de NAT (*Network Adrress Translator*), por exemplo, permite que várias máquinas usem um mesmo endereço. Logo, entende-se que endereço IP único é a quantidade de endereços mas não necessariamente a quantidade de máquinas.

Verificou-se que 3.168.461 endereços IP únicos acessaram os sensores do CBH. Este número é apresentado sem redundâncias, isto é, se um endereço IP acessou mais de um sensor ele foi considerado somente uma vez. Se comparado com o total de endereços disponíveis na Internet verifica-se que é um número pequeno, não mais do que 0,07% do espaço de endereçamento IP.

Observou-se, ainda, a quantidade de fluxos originados por endereço IP. Para esta análise agrupou-se os endereços IP a partir do primeiro byte de sua representação decimal. Como o endereçamento IP válido vai somente até o endereço 223.255.255.255 agrupou-se todos os endereços superiores a 224, inclusive, no endereço 224.

Verificou-se que para fluxos < 3 o endereço de origem 200 tem mais de 5,5x fluxos do que o segundo endereço mais usado, o 201. Que dos 224 endereços somente 40 tem algum significado estatístico embora tenham sido encontrados 78 endereços IP com mais de um fluxo.

Para fluxos >= 3 tem-se, novamente, que endereços iniciados por 200 são os que mais ocorrem e o fazem com mais de 3x o número de ocorrências do segundo endereço mais acessado, o 143. Do total de endereços representados 39 têm significado estatístico mas todo o espectro de endereços foram usados como origem.

4. Conclusão

Este trabalho empregou os dados do CBH para levantar características que auxiliam na caracterização do ruído de fundo na parcela brasileira da internet além de outras análises sazonais mas que permitem ver o tipo de tráfego que se pode esperar.

O conhecimento do ruído de fundo é um grande auxiliar na detecção de atividades não esperadas que estão ocorrendo na internet. Este tráfego irá chegar a praticamente todos os ativos diretamente conectados à internet.

Variações neste tráfego podem indicar tentativas de varredura ou infecções generalizadas, típicas da fase de propagação dos worms e dos bots.

Usou-se o conceito de fluxo em substituo ao de tráfego ou pacotes em razão do tipo de dado disponibilizado pelo CBH. Usou-se, ainda, uma janela de observação de um dia devido ao processo de transferência dos dados do CBH para o servidor central.

A “característica de normalidade” é constituída por:

- 90% de fluxos TCP, 7% UDP e 3% ICMP;

- 80% de fluxos < 3 e 20% de fluxos >= 3;
- cada sensor apresenta média de 14.383 ± 33.479 fluxos < 3.

Embora estes números representem a “característica de normalidade” da amostra trabalhada ela contém algumas informações que, mesmo com o acréscimo de novos serviços na rede tem a tendência de se manterem constantes.

A relação entre os fluxos TCP e os demais não necessariamente é uma constante mas apresenta uma tendência verificada em todos os centros de coleta de dados sobre tráfego na internet. É lógico que este número poderá variar ao longo do tempo porém parece ser intuitivo que os novos serviços disponibilizados o tem sido no protocolo TCP. Porém quase toda solução de nome é feita usando-se UDP e as mensagens de erro são enviadas com ICMP. Daí, mesmo que estes números representem somente a informação da amostra ela pode ser estendida para a parcela brasileira da internet.

A relação entre varreduras e conexões também representa uma informação útil. Não se deseja afirmar que ao longo do tempo a relação será sempre de 4 x 1. Mas tão somente que o número de varreduras é muito superior ao de conexões numa ordem de grandeza quatro vezes maior.

Além disso, o número de fluxos recebidos por um sensor é um alerta ao administrador. Cada qual terá seu número próprio. Porém, todo gerente deve incluir na sua carga, na sua capacidade de tráfego, pelo menos o valor médio levantado. Este número pode e deve ser atualizado ao longo do tempo para que seja mais representativo da realidade daquele momento. Entretanto, serve como base para análise e permite que, quando comparado aos números levantados por outros métodos, verifique-se particularidades da parcela brasileira da internet (não é o foco deste trabalho).

Usando-se as análises complementares verificou-se que a grande maioria das varreduras concentram-se em um pequeno número de portas e que há tráfego continuado por mais de 3 dias para estas portas. As portas levantadas para fluxos >= 3 não devem ser extrapoladas do CBH para a internet porque nem todos os sensores possuem os mesmos recursos necessários para completar as conexões.

A grande maioria das varreduras e, consequentemente, do ruído de fundo, tem origem em endereços IP iniciados por 200, faixa de endereços distribuídos para o Brasil. No tráfego originado por varreduras nem todos os endereços são usados porém, nas tentativas de conexão, todos o são, sejam eles válidos ou não. Caracteriza-se, assim, tentativas de ataques usando IP forjados.

Para melhorar a caracterização sugere-se, como temas para pesquisas futuras:

- a busca por correlações ou tendências entre as portas acessadas nas varreduras e as acessadas nas conexões;
- a busca por correlações ou tendências entre os endereços IP de origem e as portas;
- a procura por características que permitam quantificar e/ou visualizar acelerações na velocidade das varreduras que caracterizem a presença de ataques generalizados na rede; e
- caso as caracterizações propostas possam ser estatisticamente validadas, verificar a possibilidade de empregar o CBH como gerador de alertas precoces.

5. Referências

Baumann, R.; Plattner, C. (2002). "White Paper: Honeypots", http://www.inf.ethz.ch/~plattner/pdf/whitepaper.pdf, July 2007.

Dagon, D. et all. (2004). "HoneyStat: Local Worm Detection Using Honeypots". In 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, pages 39 – 58.

Grizzard J. et all. (2005) "Flow Based Observations from NETI@home and Honeynet Data", In: Proceedings of the 2005 IEEE Workshop on Information Assurance and Security. United States Military Academy, West Point, NY, USA.

Levine J. et all. (2003). "The use of Honeynets to Detect Exploited Systems Across Large Enterprise Networks". In Proceedings of the 2003 IEEE Workshop on Information Assurance and Security. United States Military Academy, West Point, NY, USA.

Pang, R. et all. (2004) "Characteristics of Internet Background Radiation", In: Proceedings of the IMC'04, Taomina, Italy.

Savage, S. et all. (2006) "Center for Internet Epidemiology and Defenses", http://www.cs.ucsd.edu/~savage/papers/CIEDProposal.pdf, May 2010.

Vanderavero, N. Et all. (2004). "The HoneyTank : a scalable approach to collect malicious Internet traffic". In Proceedings of the International Infrastructure Survivability Workshop (IISW'04), Lisbon, Portugal.

Yin, C. et all (2004). "Honeypot and scan detection in intrusion detection system". In Canadian Conference on Electrical and Computer Engineering, Vol. 2, pages 1107-1110.