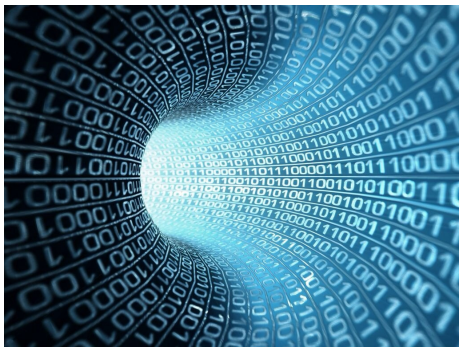
Estratégias big data para análise de tráfego hostil

Liane Tarouco, Cesar Loureiro ,
Afonso Comba de Araujo Neto, Leandro Bertholdo
POP/RS e UFRGS
GTS 24 , São Paulo 28 de novembro de 2014

Resumo

- Estratégias de big data passíveis de uso na análise de registros de acessos hostis com vistas a auxiliar na identificação de padrões ocultos.
- Etapas de análise preditiva usadas para construir um modelo dos dados capaz de auxiliar e dar suporte às decisões de aprimoramento das regras de controle de acesso, com vistas à sua otimização.
- O estudo de caso: registros de log de acessos a um honeypot e de acessos rechaçados por um firewall.

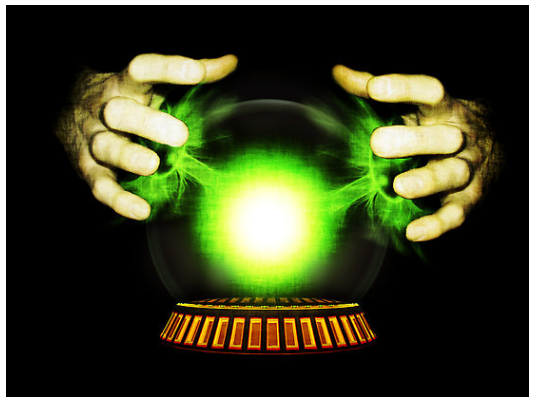


Big data

- Quantidade de informação no mundo dobra a cada 20 meses
- Pesquisa automatizada ou aumentada por computador
- Mineração de dados – elucidar parâmetros, identificar padrões → prever futuras ocorrências
- Mineração de dados é definido como o processo de descobrir padrões nos dados.
- O processo deve ser automático ou semi-automático
- Os padrões descobertos devem ser significativos e levar a alguma vantagem
- Os dados usados são usualmente volumosos

Big data

- Padrão dos dados → previsões
- Padrão detectado
  - Caixa preta
  - Caixa transparente



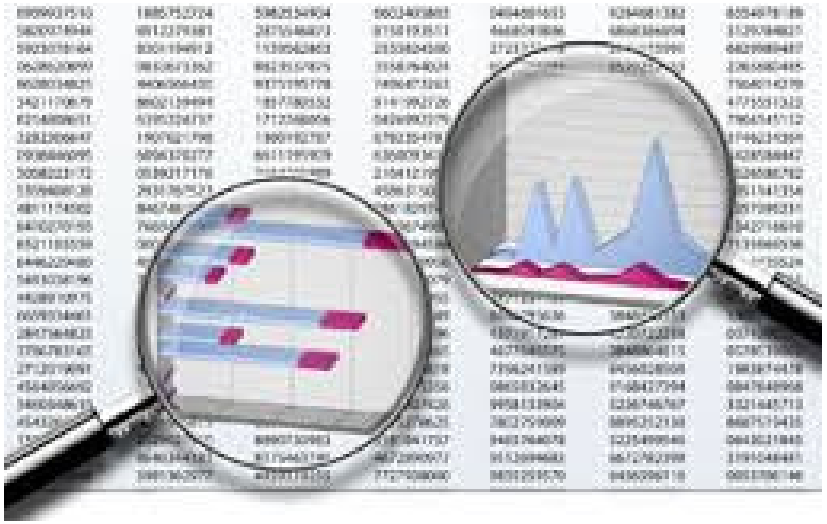
- Padrão percebido pode ser representado em termos de uma estrutura que pode ser examinada, pensada e usada para apoiar decisões futuras
- Técnicas para descrever os padrões estruturais dos dados

Algoritmos usados em big data

- Aprendizagem automatizada
  - machine learning
- Conjunto de exemplos usado para treinar o algoritmo
- Aprendizagem deve levar a um desempenho melhor no futuro
  - Decisões
    - Aceitar
    - Rejeitar
    - Zona cinza

Exemplos

- Previsão de carga → planejamento de capacidade
- Diagnóstico → manutenção preventiva
- Marketing e vendas



Performance

- Fidedignidade
  - Falsos positivos e negativos
- Aspectos éticos e institucionais
  - Barrar preventivamente ?

Estilos de aprendizagem em mineração de dados

- Classificação – forma de atribuir classificação aos futuros dados
- Associação – relacionamentos entre características
- Agrupamento – reunião em grupos para tratamento diferenciado
- Previsão numérica – prever valor derivado dos dados

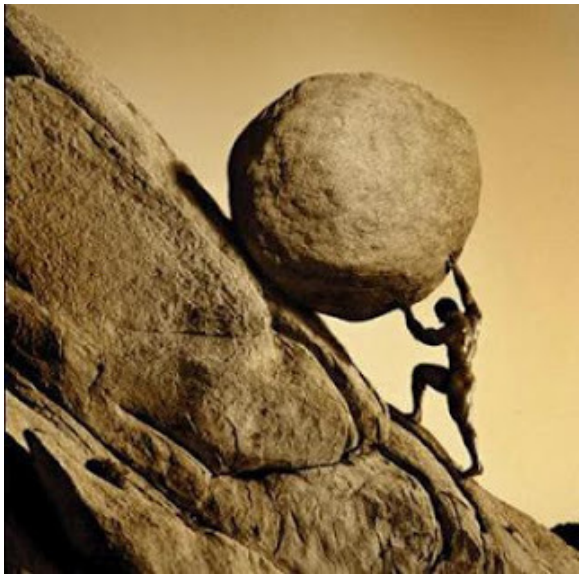
Tecnologia

- Software de **Big Data** recolhe todos os dados que uma organização gera e permite que os administradores e analistas encontrem novas maneiras de usá-los para fins de análise preditiva

Big data : preparação dos dados

- Conjunto exemplo
- Atributos
  - Nominal – enumerados, discretos
  - Numéricos

- Preparação e limpeza dos dados consome considerável tempo e esforço
  - Valores faltando
  - Registros duplicados
  - Pontos anormais
  - Valores derivados
  - Normalização



Big Data Security Analytics

Barbarians at the gate

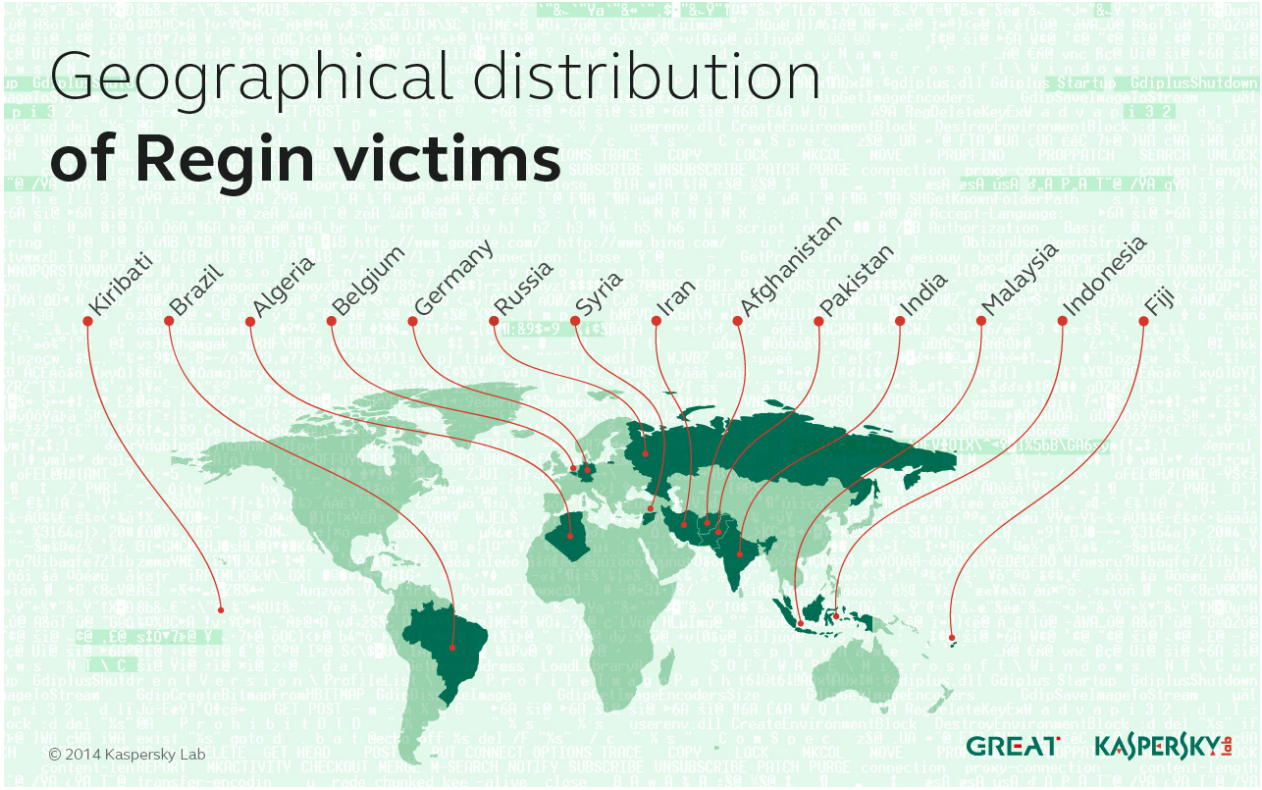
- Ajudar a reduzir falsos positivos
- Lidar com ameaças avançadas persistentes

Big data e segurança

- Ataques tradicionais
  - Estratégias mais conhecidas
  - Detecção – padrão e intensidade
- Ataques avançados persistentes
  - Lentos
  - Baixo tráfego
- Detectar uso abusivo de conta
- Identificar exfiltração de dados
- Alertar sobre a execução de novo programa

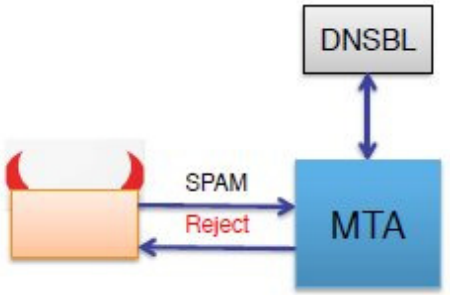
Regin

- Desde 2008
- Tem atacado nos últimos 6 anos
- Ataca principalmente empresas de telecomunicações e internet e órgãos políticos multinacionais e instituições de investigação financeira
- Ele é o mais discreto possível, dando acesso a invasores para que possam monitorar, não destruir



Cibercrime

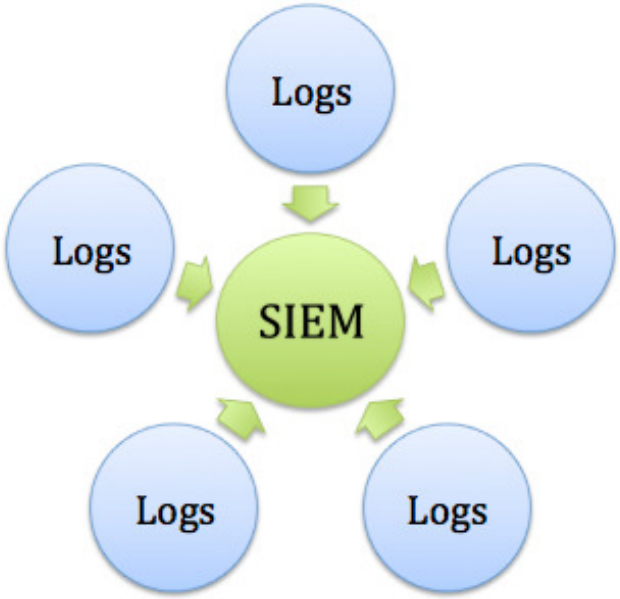
- Cibercrime é a maior preocupação das empresas
  - FBI – 50% das grandes empresas norte-americanas foram atacadas pelos chineses; 50% das grandes empresas norte-americanas foram atacadas pelos chineses e não sabem
- Uso de botnets se alastrando
- Blacklists
  - Manutenção
  - IPv6 – dificuldades (Verisign)



SIEM

- SIEM - Security Information and Event Management
- O conceito de **SIEM** é relativamente novo.
- Surgiu em 1999 e evolui gradativamente com novas funções.

- A grande maioria dos sistemas e aplicações disponíveis em uma rede corporativa geram eventos que são armazenados em **logs**
- Dados adicionais podem ser obtidos ou comprados de fontes externas



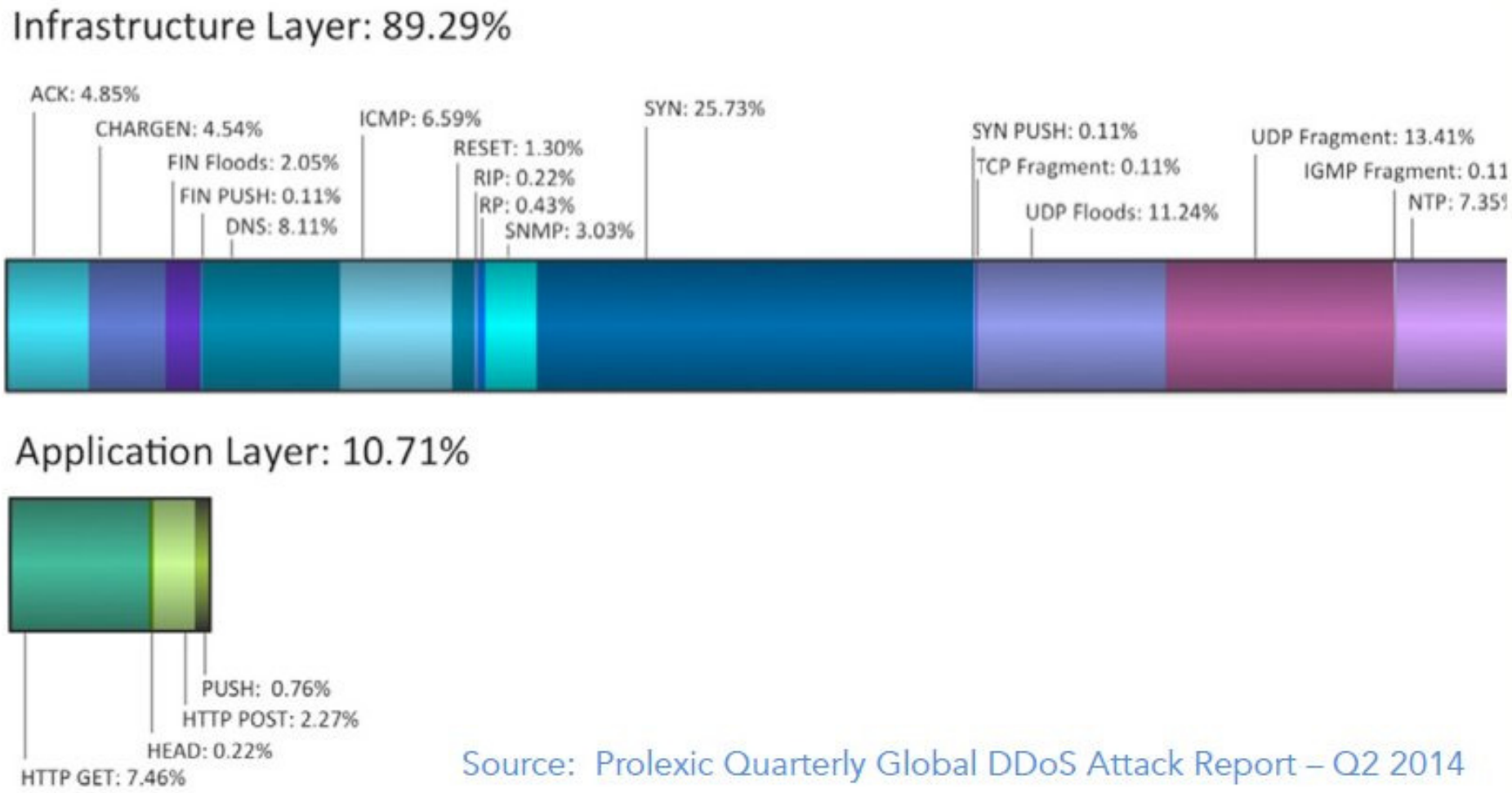
Tendências atuais

- 2013 → 2014
- Aumento da quantidade de ataques
- Aumento na banda
- Aumento dos ataques à infraestrutura
- Diminuição na duração dos ataques
  - Ataques mais curtos e mais intensos
- Aumento na banda de pico usada

Chris Beal – MCNC

Internet2 Technology Exchange

Tendências dos DDoS



Source: Prolexic Quarterly Global DDoS Attack Report – Q2 2014

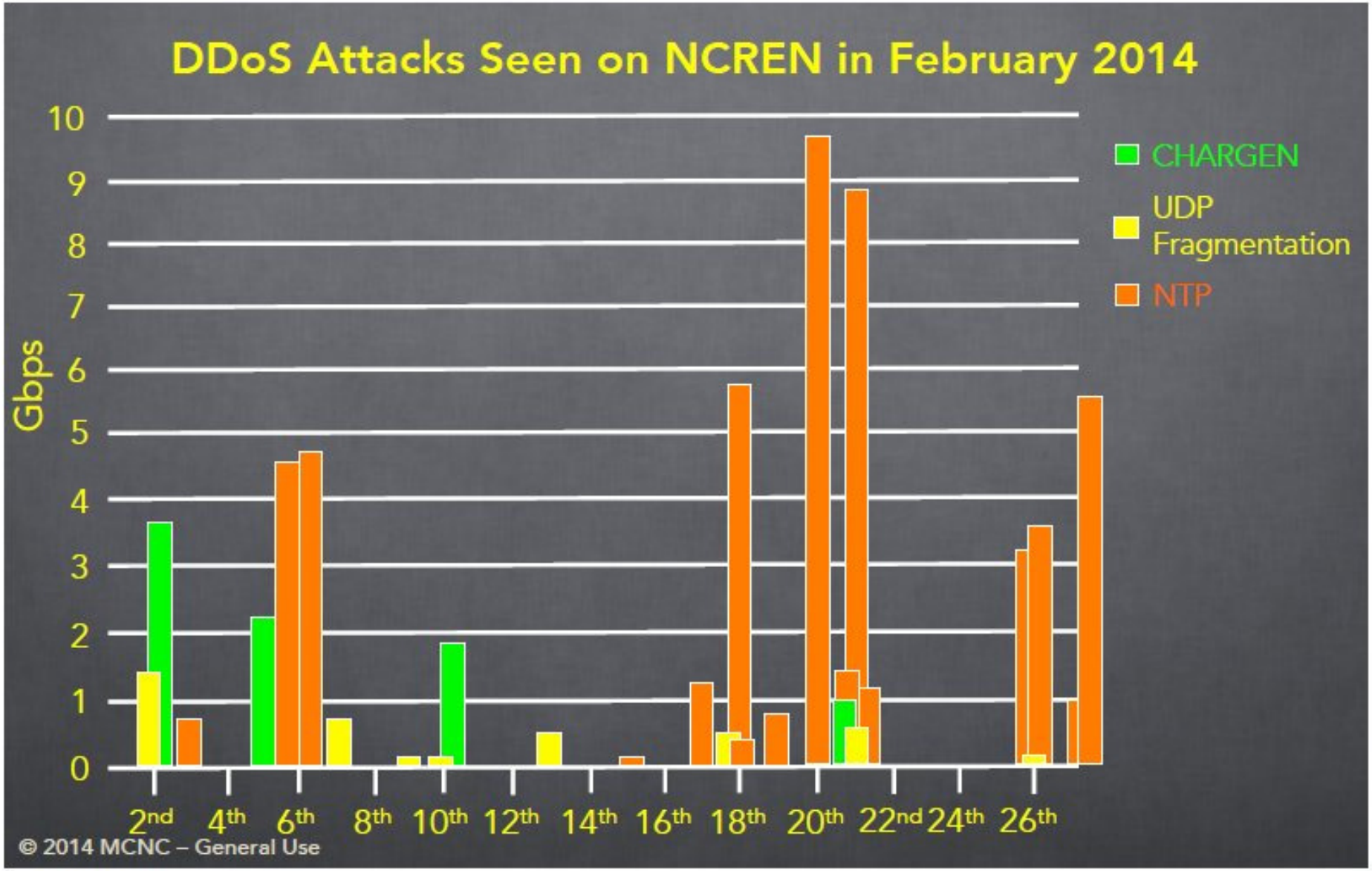
Ataques de amplificação

NetBIOS 3.8x
SNMP 6.3x

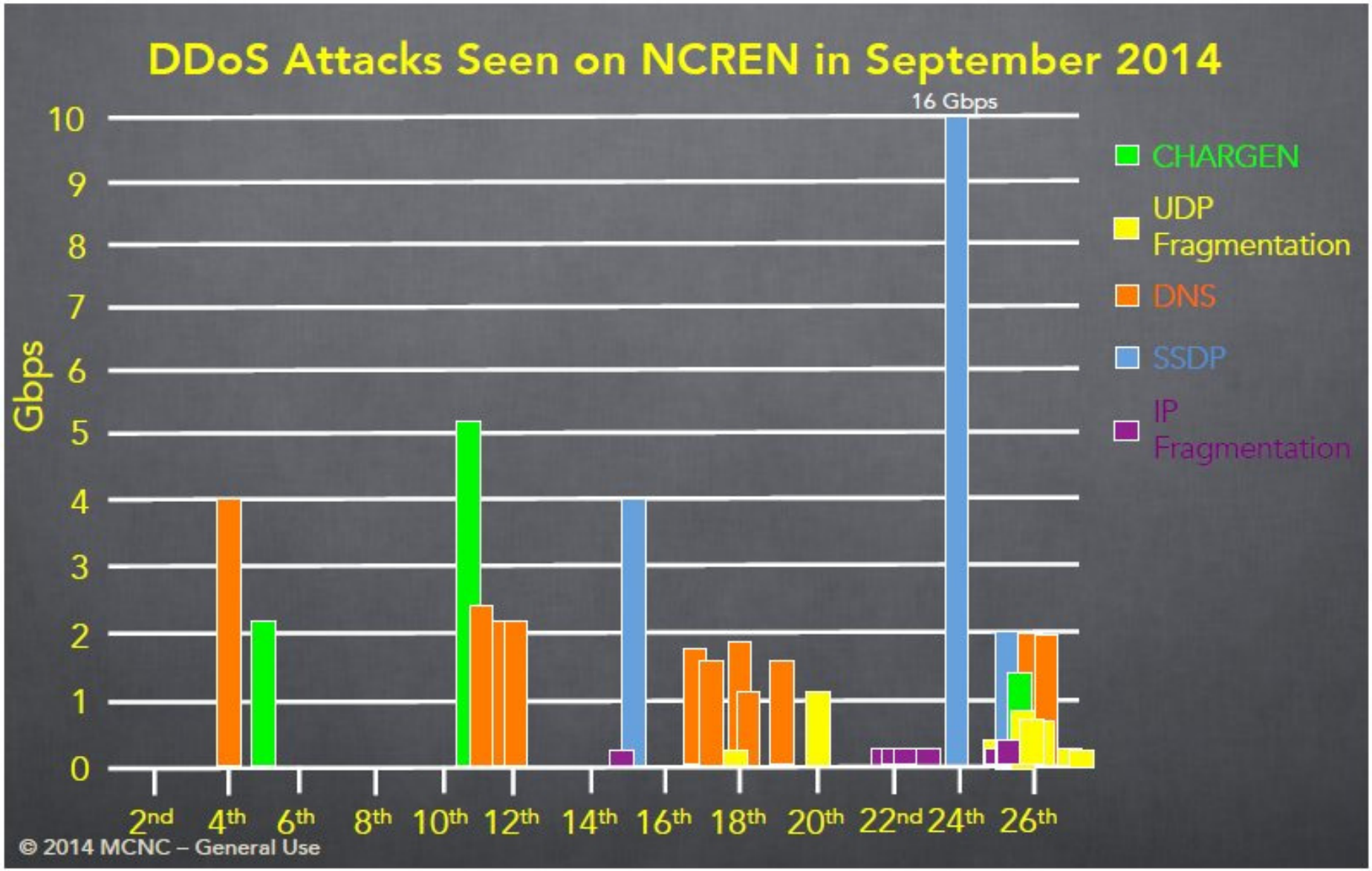
DNS 29x - 54x
QOTD 140x

CHARGEN 359x
NTP 557x

NCREN DDoS Attack Data



NCREN DDoS Attack Data



Observações sobre os ataques

- MCNC - tendências similares às da industria
  - Ataques mais curtos - NTP, + SNMP
- Maioria dos ataques visa clientes e não sistema da MCNC ou infraestrutura
  - Conteúdo Web
- Maioria dos ataques à infrastrurura
  - Fragmentação, CHARGEN, DNS, NTP, TCP SYN
- Escolas visadas mais frequentemente
  - Metade para cada segmento: K-12 & Higher-Ed
- Ataques acompanham as atividades escolares
  - Mais ataques quando as escolas estão funcionando
- Mistura de ataques direcionados e vítima “sem sorte”
  - Ataques se adaptam quando alguma mitigação é aplicada

Motivação para ataques

- K-12
  - Muitos realizados pelos estudantes
  - Suspender aulas
  - Evitar exames
  - Vingança
- Higher-Ed
  - Motivações variadas
  - Explorar vulnerabilidade de sistemas
  - DDoS relacionado com jogos
  - Outros

Proteção contra DDoS

- Ferramentas para identificar tráfego DDoS
- Detecção apenas
- Algumas vezes o cliente chama
- Comunicação com clientes quando surge problema
- Implementar filtragem direcionada quando necessário
- ACLs, black hole routing
- Trabalhar com provedores quando as crises se apresentam

Ataques de amplificação - UDP

Protocol	Bandwidth Amplification Factor	Vulnerable Command
DNS	28 to 54	
NTP	556.9	
SNMPv2	26.3	GetBulk request
NetBIOS	3.8	Name resolution
SSDP	30.8	SEARCH request
CharGEN	358.8	Character generation request
QOTD	140.3	Quote request
BitTorrent	3.8	File search
Kad	16.3	Peer list exchange
Quake Network Protocol	63.9	Server info exchange
Steam Protocol	5.5	Server info exchange

Proteção almejada

- Detecção rápida
- Remediação automatizada
- Comunicação proativa

- Contrapartida almejada
- Impedir que tráfego forjado saia do backbone
  - BCP 38 / RFC2827 - Network Ingress Filtering
    - Restricting forged traffic

Análise preditiva

- Visão
  - Categorizar
  - Conhecer particularidades
  - Formular hipóteses
  - Novas categorias
- Decisão
  - Proporcionar base para tomada de decisão
  - Agilizar mediante automatização
- Precisão
  - Reduzir erro humano

Iniciando um projeto de análise preditiva

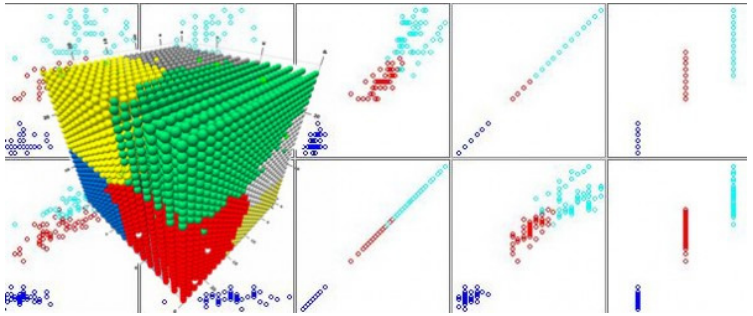
- Obter os fatos que estão ocorrendo atualmente
- Distinguir os fatos ocorrendo de outros que constituíram fatos isolados no passado
  - Tráfego incidental causado por ocorrências isoladas
- Derivar possíveis cenários futuros

Tecnologia envolvida

- Mineração de dados
- Estatística
- Algoritmos de aprendizagem automática (machine learning)
- Software para construir o modelo

Estudo por amostragem

- Um dia normal ...
  - Honeypot
    - >49000 entradas
    - > > 7 Mbytes
  - Firewall
    - >250000 entradas
    - > 1 Gbytes



Preparação dos Dados

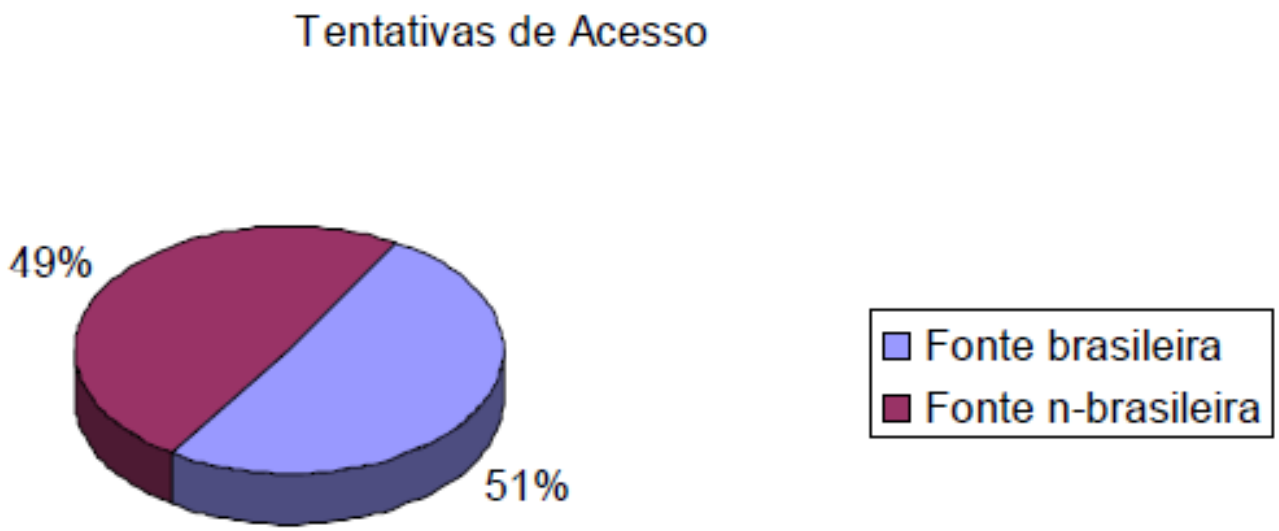
- Log do Honeypot do CERT/RS (AS 2716)
  - Formato único
- Log do Firewall da UFRGS (AS 19200)
  - 12 formatos de mensagens diferentes (muito cat, grep, cut, awk...)
- Conversão de formatos : IP2DEC

Perfil Ataques – GTS 2005

- Inicio Projeto Honeypot

POP-RS / CERT-RS

Origem das Tentativas de Acesso



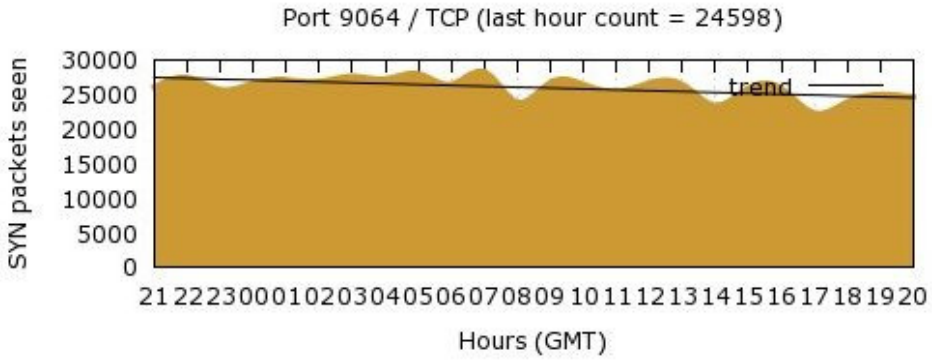
Acessos ao honeypot

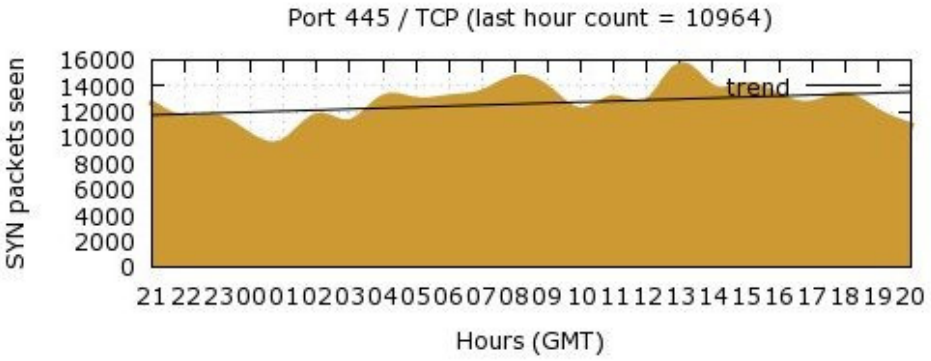
País	Frequência
USA	823
China	509
Russia	292
Venezuela	291
Taiwan	221
Brasil	202
França	171
India	169
Korea	162
Canada	127
Romania	115
Turquia	104

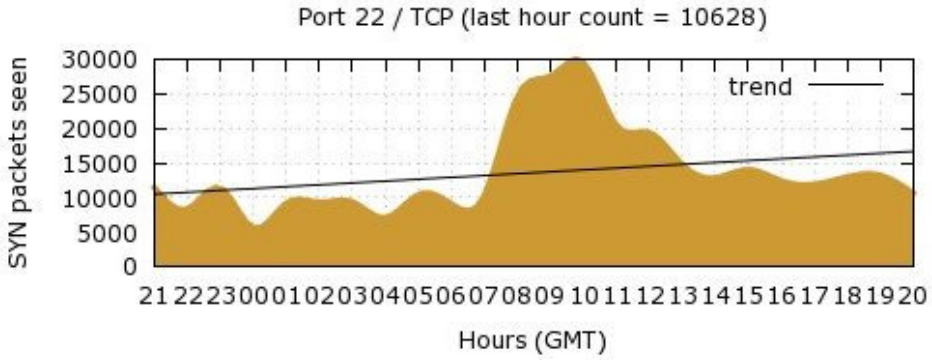
País	Frequência
Japão	91
Italia	87
Bulgaria	86
Alemanha	70
Indonesia	68
Mexico	68
Vietnam	64
Espanha	62
Holanda	59
Inglaterra	59
Hong Kong	56
Argentina	50

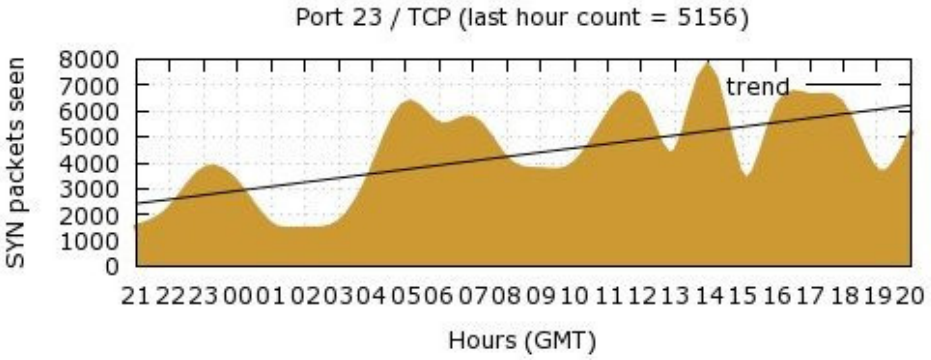
Amostragem de um 1 dia no honeypot local

Consórcio honeypot brasilieiro









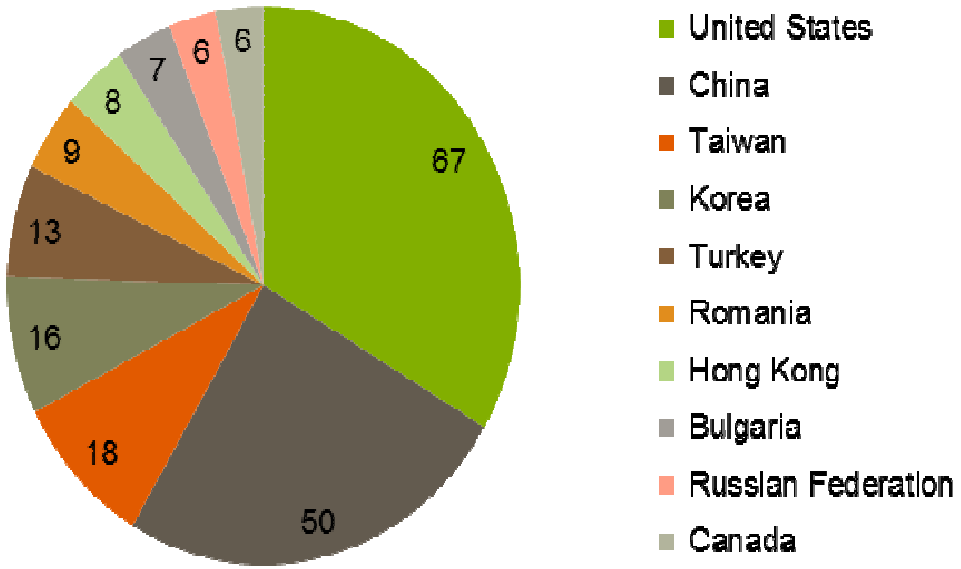
9064 – EMC2 (Legato) Networker or Sun Solcitice Backup (Official)
445 – SMB over TCP/IP
22 – SSH
23 - telnet

Algumas observações

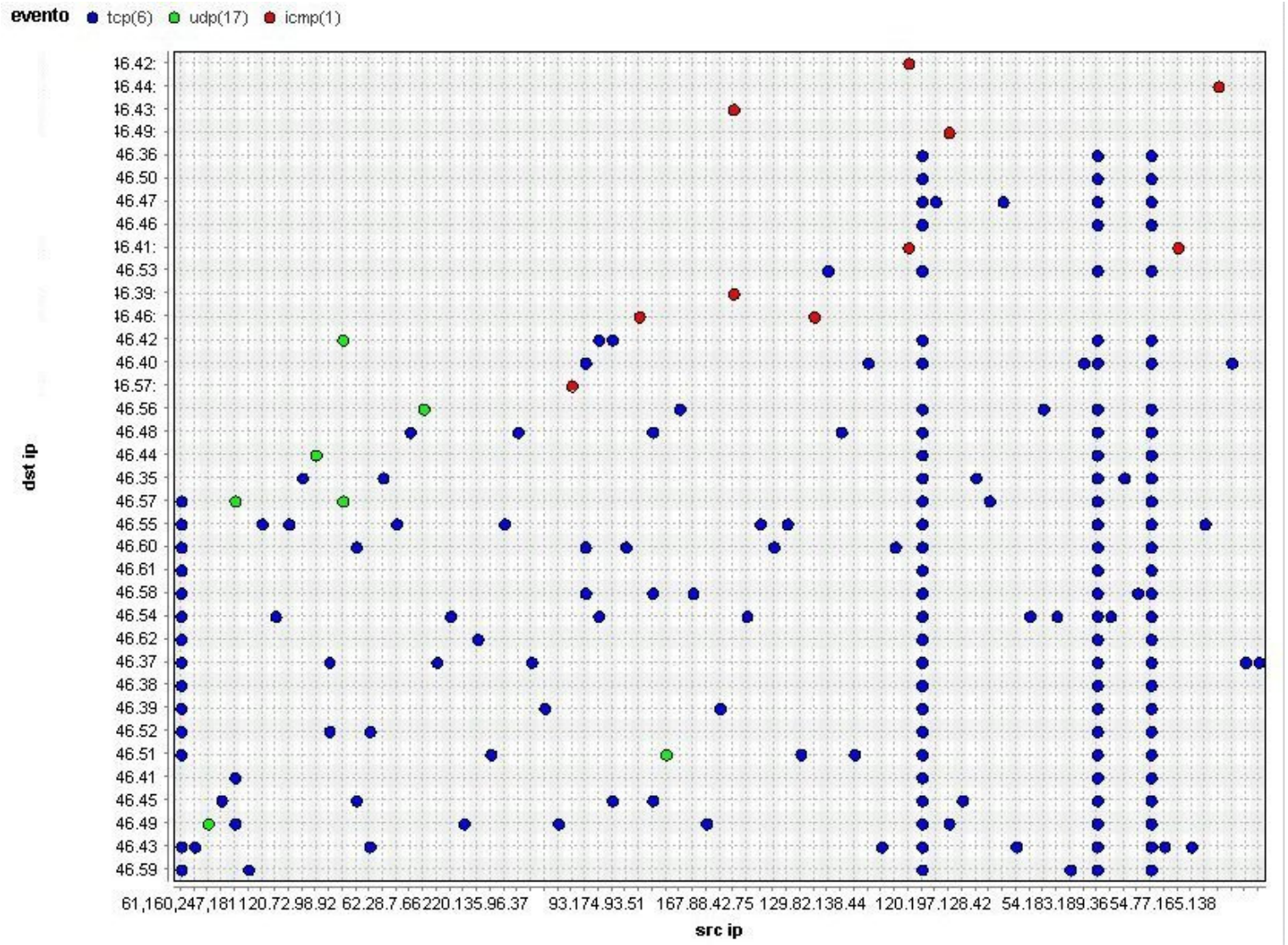
- Acessos honeypot
  - 97 % - tcp
  - 1848 (em 49036) também apareceram no log do firewall da UFRGS no mesmo dia (3,76 %)
  - Rajadas de port scan
  - Análise de valores atípicos

Análise do conjunto exemplo

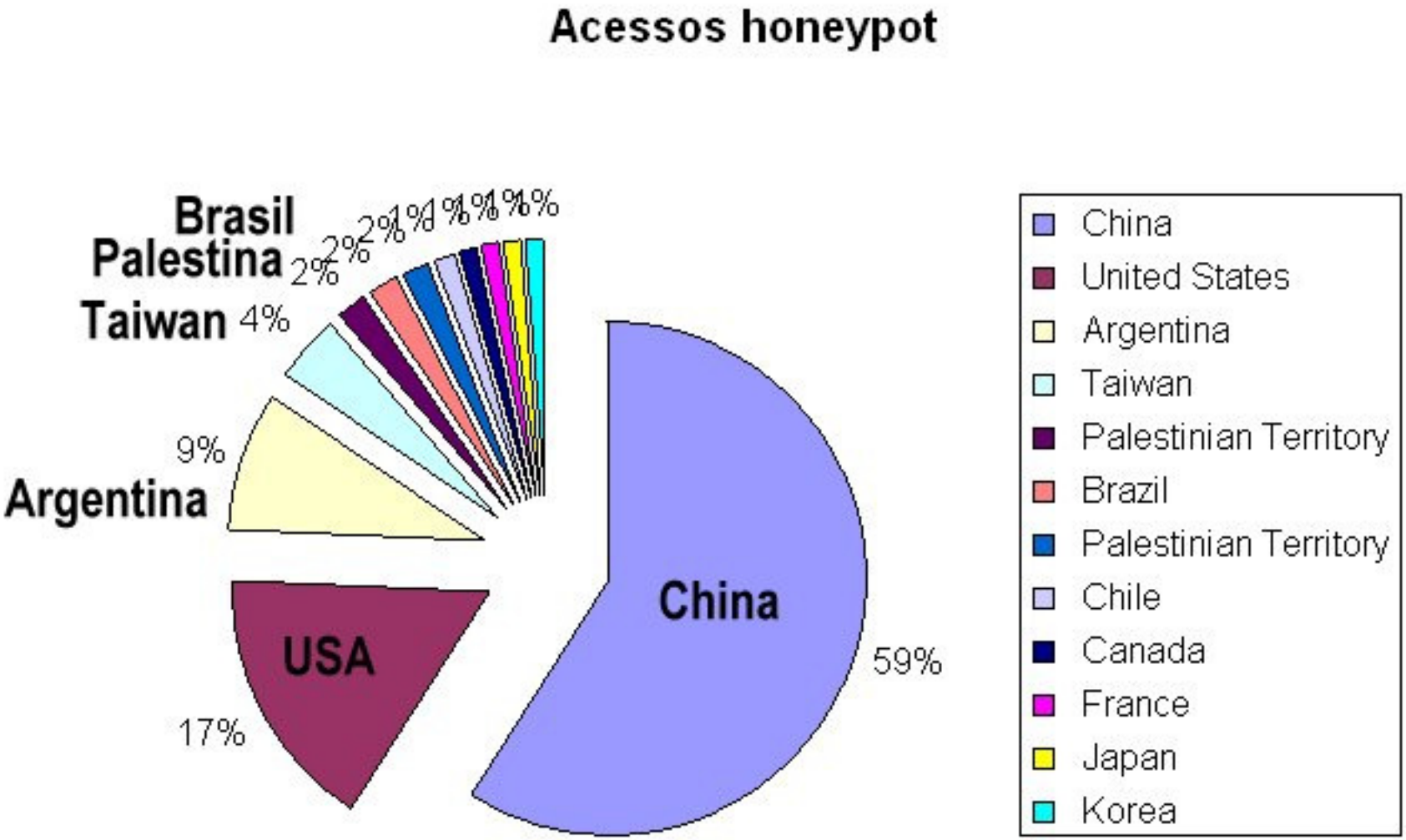
- IPs que acessaram o Honeypot e foram barrados no firewall da UFRGS – 1848 IPs
- Distribuídos em 50 países diferentes.
- Top 10
- Brasil: 4



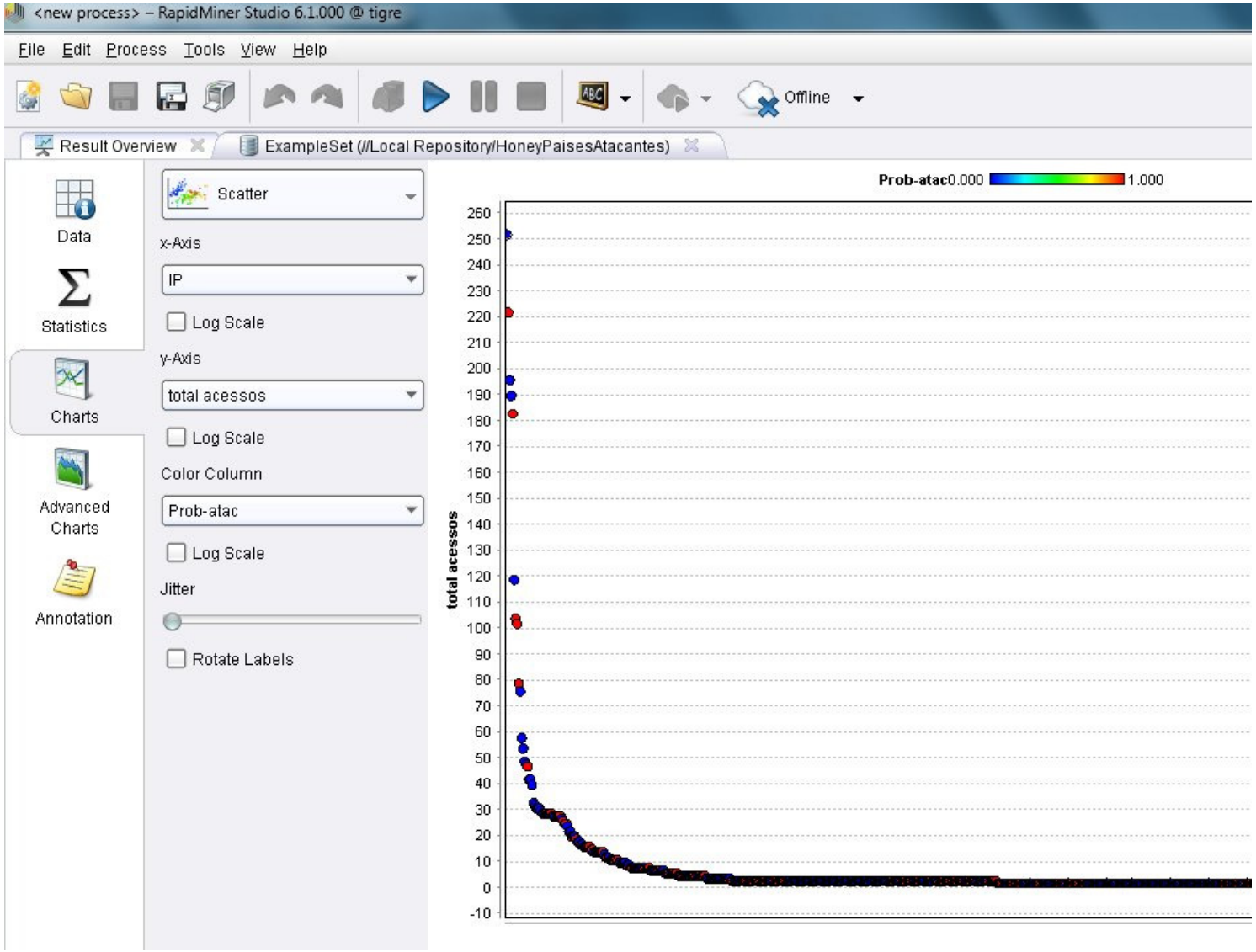
Distribuição dos IPs atacantes



Origem dos IPs atacantes



Honeyd & Firewall



Informações para contato

- CERT RS http://www.cert-rs.tche.br/
- TRI UFRGS http://www.ufrgs.br/tri/