Arcabouço Multi-motor para Detecção de Vulnerabilidades na Internet Brasileira

Lucas M. Ponce,[1] **Igor Cunha,**[2] **Isabelle Matos,**[2] **Ítalo Cunha,**[1]
Elverton Fazzion,[2,1] **Cristine Hoepers,**[3] **Klaus Steding-Jessen,**[3]
Marcelo H. P. C. Chaves,[3] **Dorgival Guedes,**[1] **Wagner Meira Jr.**[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)

[2]Departamento de Ciência da Computação
Universidade Federal de São João del-Rei (UFSJ)

[3]CERT.br - Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança
NIC.br - Núcleo de Informação e Coordenaçãoo do Ponto BR

{lucasmsp,cunha,dorgival,meira}@dcc.ufmg.br
{igoraserpac,isabelle.mattos25}@aluno.ufsj.edu.br
fazzion@ufsj.edu.br {cristine,jessen,mhp}@cert.br

Abstract. *Device search engines play an important role in the vulnerability tracking process. However, there are few studies that analyze the capabilities of these systems. Our work compares two popular search systems, Censys and Shodan, in the context of Brazil. Due to the large volume of data generated by search engines, we implemented a unique data abstraction that simplifies complex queries and integrates them with external data. We propose a framework to evaluate both systems. Our results point to significant differences in the way the two systems operate, with Censys being the system with the largest device coverage in Brazil, while Shodan has a greater diversity of detected services and higher update rates. The combination of data from both engines increases the number of services detected and the scanning rate of up to 1.8 times, while obtaining more details about the services evaluated.*

Resumo. *Motores de busca de dispositivos desempenham um papel importante no processo de rastreamento de vulnerabilidades. Entretanto, existem poucos estudos que analisam as capacidades desses sistemas. Nosso trabalho compara dois sistemas de busca populares, o Censys e o Shodan, no contexto da internet brasileira. Devido ao grande volume de dados gerados pelos motores de busca, implementamos uma abstração de dados única que simplifica consultas complexas e que permite a integração de dados externos complementares. Propomos um arcabouço para avaliar ambos os sistemas. Nossos resultados apontam diferenças significativas no modo de operação dos dois sistemas, sendo o Censys o sistema com maior cobertura de dispositivos no Brasil, enquanto o Shodan possui uma maior diversidade de serviços detectados e atualizações mais frequentes. A combinação dos dois motores aumenta a quantidade de serviços detectados e a taxa de varredura em até 1,8 vezes, ao mesmo tempo que obtemos mais detalhes sobre os serviços avaliados.*

1. Introdução

O Brasil é um dos cinco países com maior número de dispositivos conectados à Internet no mundo [Statista 2023]. O investimento por parte de empresas na informatização de setores e o crescimento de dispositivos da Internet das Coisas, como smartphones, televisores e câmeras, contribuem para esse cenário positivo de modernização do país. Porém, ao mesmo tempo, novos desafios de segurança surgem, uma vez que o crescimento exponencial de dispositivos e o volume de tráfego gerado por eles desafiam a cobertura e eficácia de métodos tradicionais de controle e segurança de redes [Mousavi et al. 2020]. Por exemplo, em 2023, ataques de *ransomware* registraram um aumento de 13% no número total de ocorrências em relação ao ano anterior [IT Section 2024].

Segurança digital é uma das principais preocupações de todos os setores que utilizam a Internet, sejam empresas privadas, órgãos públicos ou usuários domésticos. Isso porque vazamentos de dados sensíveis, como informações de cartão de crédito, geram graves impactos sociais e custos financeiros. Exemplos disso foram a invasão de computadores de um hospital no Distrito Federal em 2022, quando hackers exigiram pagamento para não divulgar informações obtidas [Ortiz e Mendes 2023] e, mais recentemente em 2023, o ataque hacker na UFMS, em que uma série de dados pessoais foram vazados durante o ataque às informações digitais da instituição [Câmara 2023]. Nesse contexto, motores de busca como Censys, Shodan e Zoomeye desempenham um papel importante em segurança de sistemas, coletando informações sobre dispositivos conectados na rede para identificação de vulnerabilidades.[1]

Motores de busca desse tipo varrem (*scan*) a Internet em busca de dispositivos executando aplicações acessíveis pela rede, permitindo o reconhecimento em larga escala de dispositivos, aplicações e possíveis vulnerabilidades. Embora esses motores estejam se tornando cada vez mais populares, existem poucos estudos que os comparem em termos de aspectos de cobertura da rede e profundidade da varredura. Como as informações coletadas por esses motores são geralmente pagas, um estudo comparativo pode auxiliar o operador de rede na contratação desses serviços e permitir identificar como os serviços podem ser combinados de forma complementar. Embora a combinação seja pouco explorada, acreditamos ser uma frente promissora, por exemplo, no aumento da frequência de varredura, uma propriedade importante na detecção de vulnerabilidades recentes (*N-days*), ao mesmo tempo que obtemos uma diversidade maior de informações ao combinar as capacidades de cada motor.

Comparar motores de busca é desafiador, pois, embora tenham interseções de funcionalidades, eles possuem particularidades de projeto, implementação, funcionamento e operação que influenciam as observações coletadas. Por exemplo, a sequência e a frequência de sondagem de endereços IP são diferentes entre os motores, o que impacta na detecção e rastreabilidade de vulnerabilidades. Além disso, esses motores de busca capturam um enorme volume de dados, o que exige a adoção de uma estratégia eficiente para manipulação e análise de dados para realizar comparações.

Neste artigo, comparamos os dois sistemas de busca comerciais mais populares atualmente no contexto de análise de vulnerabilidades de rede no Brasil em larga escala: Censys e Shodan. Mais especificamente, buscamos responder questões sobre cobertura e profundidade das varreduras realizadas por esses motores de busca. Para isso, propomos

[1]Censys (https://censys.io); Shodan (https://shodan.io); Zoomeye (https://www.zoomeye.org)

um novo arcabouço que generaliza um trabalho anterior [Ponce et al. 2023], o qual era limitado ao Shodan, para também processar dados do Censys. Além disso, combinamos os dados de ambos os motores em uma interface única e amigável para realização de consultas e disponibilizamos o código da ferramenta publicamente. Nossas análises, utilizando o arcabouço desenvolvido, mostram diferenças significativas no modo de operação dos sistemas. Enquanto o Censys demonstra ter uma maior cobertura de IPs e portas, o Shodan demonstra realizar uma varredura mais frequente, coletando uma quantidade maior de informações em cada dispositivo sondado. Dessa forma, nossos resultados indicam que as informações das ferramentas são complementares e, se combinadas, oferecem uma visão mais ampla e aprofundada sobre dispositivos na Internet.

2. Motores de busca de dispositivos

O aumento significativo de dispositivos da Internet das Coisas (IoT) e a preocupação com segurança impulsionaram o interesse por motores de busca de dispositivos. O Shodan, lançado em 2009, foi pioneiro, sendo capaz de identificar dispositivos e aplicativos em execução e acessíveis na Internet. As informações coletadas pelo Shodan incluem endereço IP, sistema operacional, softwares e portas abertas. Para coletar informações, o Shodan se conecta a um endereço IP e porta aleatórios a partir de uma lista de portas predefinidas [Matherly 2015]. Em 2015, o Censys, baseado no ZMap, de código aberto, foi lançado. Em sua primeira versão [Durumeric et al. 2015], o Censys realizava uma varredura TCP SYN, a partir do ZMap, para verificar portas abertas em dispositivos na Internet. Atualmente, por ser comercial, sua arquitetura não é mais divulgada. Detalhes de implementação do Shodan também não são divulgados pelo mesmo motivo.

Apesar de suas diferenças, motores de busca possuem um núcleo de funcionamento em comum: centenas de módulos implementam o protocolo de comunicação para coletar informações, na forma de *banners*, da aplicação em funcionamento na porta sondada. O resultado de uma sondagem pode possuir campos de diversos tipos e são dependentes da aplicação sondada. Exemplos de campos são: (i) campos textuais, como nome da organização responsável; (ii) campos numéricos, como a porta sondada; (iii) campos com lista de valores, como a lista de vulnerabilidades; e (iv) campos complexos (sob uma subestrutura JSON), como o campo MongoDB no Shodan, que contém informações sobre o *status* do serviço e a lista de bases de dados. Dados de motores também podem utilizar catálogos de informações para enriquecer os dados com campos derivados, como o nome e versão do produto ou o sistema operacional do dispositivo.

Em geral, motores de busca disponibilizam uma interface *web* para a busca de registros e também fornecem acessos aos dados a partir de *snapshots* ou APIs. No entanto, o uso da interface *web* impõe limitações, em particular quanto à filtragem por campos. Por exemplo, expressões regulares, integração com bases externas ou transformações de dados, como agrupamentos por dispositivos, não são atualmente disponibilizados. Dessa forma, análises mais complexas exigem um processamento local por parte do usuário. Porém, em cenários de grandes volumes de dados, cabe ao usuário lidar com as complexidades da infraestrutura e da escalabilidade do processamento. Além disso, em muitos casos, é necessário integrar esses dados com fontes externas, de forma a realizar análises de vulnerabilidades mais completas. Por exemplo, o operador de rede pode estar interessado em verificar vulnerabilidades a partir dos serviços e suas versões coletadas pelo motor de busca, sendo necessária, uma integração com um catálogo de vulnerabilidades.

Por fim, é importante destacar que soluções de código aberto são escassas e limi-

tadas. Até o limite do nosso conhecimento, a única ferramenta alternativa da qual temos conhecimento é o Ivre.[2] Para cenários que exigem uma busca mais simples, é possível utilizar aplicações baseadas em ferramentas como o ZMap como alternativa.[3] Apesar do Shodan e do Censys serem sistemas comerciais, ambos oferecem planos gratuitos, além de planos especiais para pesquisas acadêmicas. Para este artigo, utilizamos dados do Shodan disponibilizados pelo CERT.br e dados do Censys obtidos através de uma licença acadêmica, ambos não permitindo o compartilhamento dos dados.

3. Arcabouço para processamento massivo de dados multi-sistema

Para superar os desafios citados na seção 2, propomos um arcabouço geral que permite o processamento em larga escala de dados de múltiplos motores de busca. Nosso arcabouço reduz a complexidade de processamento de grandes volumes de dados para operadores de rede e fornece uma interface comum para análise de dados. A figura 1 mostra sua arquitetura geral. O arcabouço recebe e converte dados de motores como *Shodan* e *Censys* para um formato mais eficiente (seção 3.1), coleta informações de bases externas complementares, como a *Common Vulnerabilities and Exposures* (CVE) (seção 3.2), que podem ser integradas com os dados tratados dos motores de busca utilizando uma camada de abstração mais próxima do operador de rede (seção 3.3). Essa abstração permite ao operador realizar consultas, escondendo a complexidade do seu processamento. Disponibilizamos o arcabouço como contribuição para a comunidade científica.[4]

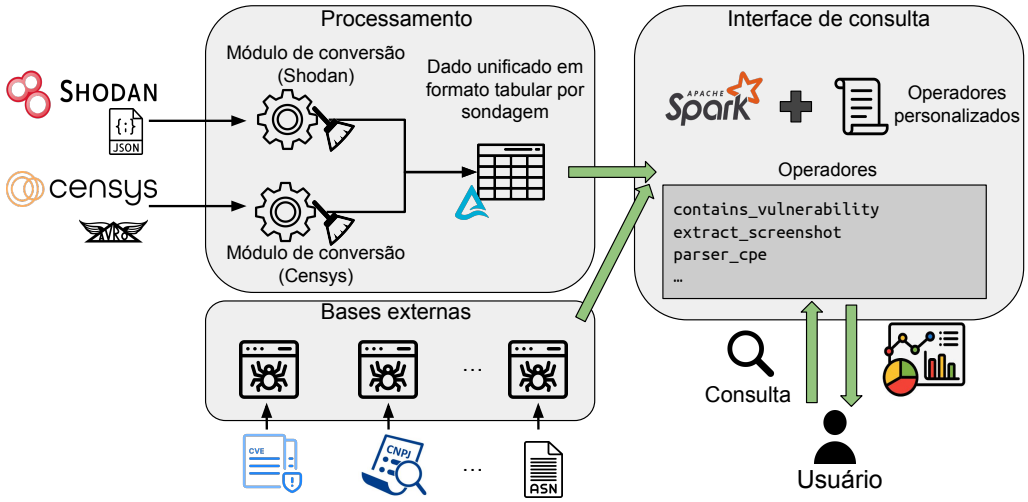

Figura 1. Arquitetura geral do arcabouço.

3.1. Processamento e padronização dos dados

Motores de busca como o Shodan e o Censys são capazes de coletar informações de mais de 500 milhões de dispositivos ao redor do mundo várias vezes por mês. Identificar vulnerabilidades nesse volume de dados pode ser desafiador, ainda mais quando o tempo é um fator determinante em análises de vulnerabilidades: quanto mais rápido uma vulnerabilidade for detectada, mais cedo contramedidas podem ser implantadas. Para tornar o processamento de dados desses motores de busca mais eficiente, tanto em aspectos de

[2]Ivre: https://ivre.rocks/
[3]ZMap: https://zmap.io/
[4]Disponível em: https://github.com/lucasmsp/tlhop-library

desempenho quanto em usabilidade, organizamos e comprimimos os dados no formato Delta Lake para processamento no Apache Spark.[5] Maiores explicações sobre os benefícios desse formato são discutidos em [Ponce et al. 2023].

Motores de busca possuem seu próprio formato de dados. Por exemplo, o Shodan disponibiliza dados em formato *JSON*, onde cada registro representa uma sondagem de um dispositivo (*banner*), enquanto o Censys disponibiliza dados no formato Avro e agrupa os *banners* por endereços IP. Para lidar com essa heterogeneidade, configuramos *módulos de conversão* para tratar as especificidades dos dados de cada motor de busca, conforme pode ser observado na figura 1. Cada módulo organiza os dados de entrada de acordo com um padrão pré-definido, aplica regras de limpeza de dados e disponibiliza os dados como uma tabela, separadas por motor, em formato Delta Lake. A organização padrão adotada por nosso arcabouço é por *banner*, uma vez que motores de busca como o Censys e Shodan não implementam lógicas para validação de dispositivos. Dessa forma, em redes que utilizam endereçamento dinâmico, diferentes *banners* associados a um mesmo endereço IP podem não representar necessariamente o mesmo dispositivo físico. Além disso, como o interesse é avaliar serviços vulneráveis, manter os registros organizados por *banner* torna mais simples a codificação e aumenta o desempenho de consultas. Embora os dados de cada motor sejam armazenados separadamente, nosso arcabouço permite ler ambas as tabelas como uma única fonte de dados.

3.2. Integração de bases externas

Em razão da diversidade de análises que podem ser feitas sobre dados de motores de busca, o enriquecimento feito por esses serviços geralmente é insuficiente e precisa ser complementado, por exemplo, para a classificação de uma vulnerabilidade de acordo com o *Common Vulnerabilities and Exposures* (CVE) ou para fornecer informações complementares sobre o sistema autônomo (AS) do endereço IP no *banner*. Informações desse tipo podem ser úteis para um operador de rede para tomar ações protetivas. Dessa forma, nosso arcabouço disponibiliza coletores para oito fontes de dados, como informações sobre os sistemas autônomos, organizações brasileiras, bases de vulnerabilidades e de softwares. Exemplos de bases de dados disponibilizadas são a já mencionada CVE, *CAIDA AS Rank* e o Cadastro Nacional de Pessoas Jurídicas. Utilizando nosso arcabouço, dados de fontes externas, como as mencionadas ou outras que se mostrem úteis, podem ser integrados a consultas em tempo de execução para o enriquecimento de análises.

3.3. Interface de consulta

O formato tabular unificado descrito na seção 3.1 é eficiente do ponto de vista do tempo para a realização de consultas. A construção das consultas, porém, pode ser complexa e exigir uma série de operações de transformação sobre os dados. Desenvolvemos 21 operadores sobre diversos tipos de dados que permitem aos usuários expressar consultas complexas em um nível de abstração elevado, delegando a construção e execução dos comandos Spark para nosso arcabouço. Por exemplo, o operador contains_vulnerability verifica se uma vulnerabilidade foi detectada em um *banner*, permitindo a filtragem de *banners* que possuem os CVEs passados como parâmetros. Informações sobre os demais operadores podem ser encontradas no repositório do projeto.

[5] Apache Spark (https://spark.apache.org/); Delta Lake (https://delta.io/)

Tabela 1. Número de endereços IPs, *banners* e tempos de sondagem dos motores.

MOTOR	# IPS ÚNICOS	# BANNERS	TEMPO POR SONDAGEM (S)
CENSYS	14.158.826	77.457.135	4,69 ± 58,81
SHODAN	8.370.733	39.591.808	5,97 ± 78,75
TOTAL	14.981.494	117.048.943	-

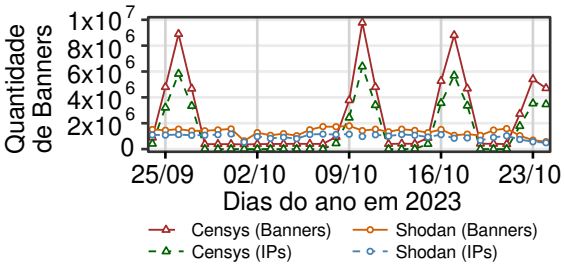

Figura 2. Registros coletados por dia.

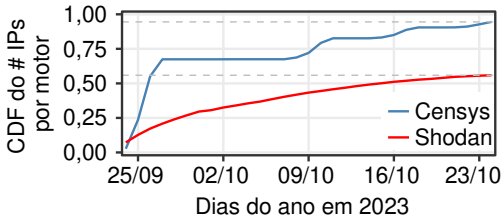

Figura 3. CDF da cobertura dos IPs por dia em relação ao total de IPs sondados.

4. Estudo comparativo entre o Shodan e o Censys

Nesta seção, utilizando nosso arcabouço, realizamos a análise comparativa entre os motores de busca Shodan e o Censys. Inicialmente, apresentamos uma caracterização dos dados utilizados para a comparação (seções 4.1 e 4.2). As seções 4.3 e 4.4 discutem a cobertura dos motores de busca na Internet, enquanto as seções 4.5 e 4.6 discutem a profundidade e riqueza dos dados coletados pelos motores. Por fim, a seção 4.7 realiza um estudo de caso para ilustrar como as informações dos motores se complementam.

4.1. Caracterização das amostras

Para a análise comparativa, utilizamos dados de ambos os motores no período entre 24/09/2023 a 25/10/2023 (31 dias) para dispositivos presentes no espaço de endereçamento brasileiro. Os dados do Shodan foram compartilhados pelo Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança no Brasil (CERT.br) em uma parceria para mapear vulnerabilidades no espaço de endereçamento IPv4 do Brasil. Enquanto os dados do Censys foram cedidos pelo motor de busca a partir de uma licença de pesquisa acadêmica. Após a conversão dos dados, o *dataset* do Shodan totalizou 78,6 GB, enquanto o Censys alcançou 56 GB. A tabela 1 mostra uma visão geral sobre o número de endereços IPs únicos coletados, o total de *banners* e o tempo entre sondagens.

A segunda coluna da tabela 1 mostra o número de endereços IP sondados pelo Shodan e Censys no período analisado. No total, 14,9 milhões de IPs brasileiros foram coletados; dos quais 5,5% foram sondados apenas pelo Shodan enquanto outros 44,1% foram sondados apenas pelo Censys. No total, o Censys sondou 1,7× mais dispositivos que o Shodan. Comparando os valores medidos com informações do LACNIC, os dois sistemas, quando somados, coletaram informações sobre 17,2% do espaço de endereçamento brasileiro em 31 dias.

4.2. Caracterização dos processos de sondagem

Motores de busca geralmente fornecem uma identificação do coletor interno responsável pelas sondagens. Utilizamos essa informação para estimar o tempo médio gasto entre cada sondagem de um coletor. Essa informação é representada na última coluna da tabela 1 em segundos. Embora o Censys possua uma coleta mais rápida, acreditamos que a diferença

de cobertura entre os sistemas deve-se às abordagens de sondagem. Por exemplo, nossos experimentos validaram a aleatoriedade do Shodan, com uma distribuição similar à uniforme. Já a documentação do Censys sugere que o sistema implementa múltiplas formas de escalonamento das sondagens. Por exemplo, enquanto portas menos populares são sondadas periodicamente com menor frequência (a cada 10 dias), endereços IPs em provedores de nuvem possuem maior prioridade e são sondados mais frequentemente (diariamente).[6] Esse comportamento pode ser observado no gráfico da figura 2. Enquanto o Shodan apresenta valores estáveis de sondagens diariamente, o Censys possui picos de sondagens periódicos a cada 7–13 dias.

Dessa forma, o Censys apresenta uma maior periodicidade na cobertura quando avaliamos a sondagem de endereços IP brasileiros. A figura 3 apresenta a distribuição acumulada da cobertura dos endereços IP sondados, em relação ao total (14,9M), por cada motor de busca em função do tempo. É possível observar que o Censys sonda 71% do total de endereços IP que ele sondou no período (14,1M) em quatro dias, coincidindo com o primeiro pico de sondagem. Por outro lado, para o Shodan atingir a mesma marca, foram necessários 14 dias.

4.3. Cobertura do espaço de endereçamento

Nesta seção, avaliamos com detalhes a cobertura do espaço de endereçamento brasileiro pelos motores sob aspectos de frequência de sondagem de endereços IP e diferenças geográficas. Pelo fato do Shodan escolher destinos aleatoriamente, pelo menos um endereço IP é sondado para a maior parte das sub-redes brasileiras durante o período analisado, conforme pode ser observado na figura 4. Por outro lado, o sistema de escalonamento de sondagens do Censys parece dividir a rede brasileira em dois conjuntos: um conjunto de sub-redes que são sondadas a cada 10–18 dias e um conjunto de sub-redes que são sondadas uma vez durante os 31 dias do período analisado. Inferimos, de acordo com a documentação do Censys, que as sub-redes frequentemente sondadas são provedores de computação em nuvem. Dessa forma, o Censys possui uma abordagem de sondagem hierárquica com frequência variável, enquanto o Shodan realiza uma sondagem estável.

O gráfico da figura 5 mostra que a maior parte dos endereços IP cobertos são pouco sondados. Como cada sondagem é vinculada a uma porta, apenas um serviço é avaliado por sondagem. Isso resulta em poucas sondagens para serviços em execução no período analisado ou em poucos serviços monitorados por cada endereço IP. Essa diferença é ainda maior para o Shodan: a maioria dos endereços IP possui poucas sondagens. O Shodan também apresenta casos extremos, com alguns endereços IP que foram sondados cerca de 17 mil vezes em diversas portas.

Considerando todos os endereços IP sondados pelas duas ferramentas (14,9 milhões), os mapas de calor da figura 6 mostram a distribuição de sondagem realizada pelos motores no espaço brasileiro. É possível observar que o Censys distribui melhor a sondagem de endereços IPs por regiões do Brasil (tons similares na figura 6a) com alta cobertura dos endereços (cor vermelha). O Censys sonda, em média, 93,3% do total de endereços IPs sondados pelas duas ferramentas em cada estado. Por outro lado, existem regiões brasileiras onde o Shodan sonda muito poucos endereços IP (cores mais próximas do azul na figura 6b). No geral, o Shodan sonda, em média, 50,1% do total de endereços IPs sondados pelas duas ferramentas em cada estado. Além disso, a contribuição do Sho-

[6]Censys Internet Scanning Intro: https://t.ly/48kM_

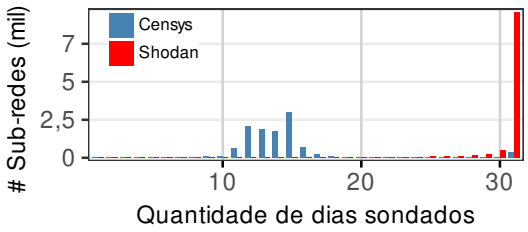

Figura 4. Frequência de sondagem das sub-redes brasileiras.

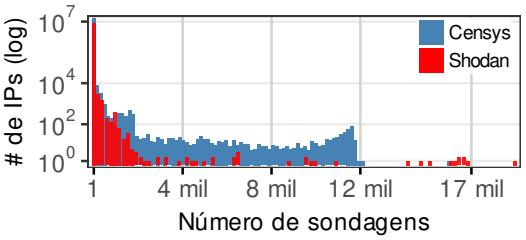

Figura 5. Relação do número de sondagens por IP.

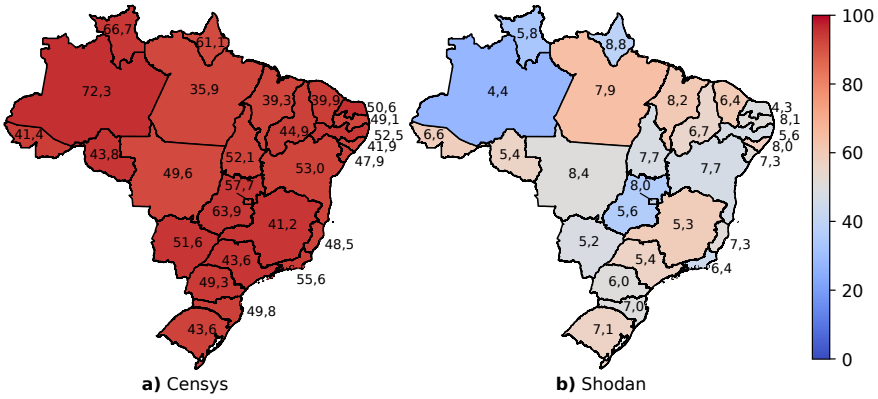

Figura 6. Mapa de calor da distribuição de endereços IP sondados entre os estados brasileiros. Quanto mais vermelho, maior é o percentual de endereços coletados por um motor de busca em relação ao total dos dois sistemas. Os números em cada estado representam o percentual de endereços IP no motor de busca analisado em relação ao total de endereços coletados.

dan em sondar IPs exclusivos é baixa, por exemplo, para o estado de São Paulo, 5,4% dos IPs atribuídos para o estado foram sondados apenas pelo Shodan, 43,6% apareceram apenas no Censys, e os demais 51% foram identificados pelos dois motores. Utilizar uma base com informações sobre as cidades brasileiras é útil não apenas para uma comparação entre estados, mas também para relacionar aspectos socioeconômicos.

4.4. Cobertura de portas

Nesta seção avaliamos o intervalo de portas sondadas pelos motores de busca. O Shodan opera sobre um conjunto pré-selecionado de portas, considerando que cada porta pode ter um conjunto de serviços em execução. Já o Censys opera em todas as 2^{16} portas. Para o período analisado, o Censys foi capaz de identificar serviços em 58.632 portas a mais que o Shodan. O Censys identificou ao menos um serviço a mais que o Shodan em 62,4% dos IPs, contra 17,4% para o Shodan em relação ao Censys. Na prática, essa diferença corresponde a aproximadamente 2,1 serviços a mais, mas incluindo um caso extremo de um endereço IP com 11.131 serviços identificados. Casos assim, embora não muito frequentes na amostra, podem ser atribuídos a ambientes de nuvem, por exemplo, a um servidor com Docker, onde cada uma dessas portas expõe um contêiner de uma aplicação.

A tabela 2 mostra uma caracterização de dez tipos de serviços sondados por ambos os sistemas. Por exemplo, para o MongoDB, dos 3.801 processos identificados pelo Censys, 22,3% deles estão em portas presentes exclusivamente no Censys. Essa fração é ainda maior para o Telnet, onde cerca de 75,6% dos processos estão em portas monitoradas apenas pelo Censys. Já a última coluna da tabela apresenta a porcentagem de serviços que foram sondados pelos dois motores de busca. Enquanto cerca de 40% dos

Tabela 2. Caracterização da quantidade de serviços sondados entre os motores.

Serviço	Censys		Shodan	Geral	
	Total	% Exclusivo	Total	Total	% Compartilhado
FTP	208.514	9,33	94.231	216.505	39,83
HTTP	13.798.638	18,28	6.785.327	15.608.469	31,88
Kubernetes	32.066	35,47	1.878	32.114	5,70
MongoDB	3.801	22,34	1.646	4.082	33,44
MQTT	8.437	15,51	2.538	9.256	18,57
MySQL	119.405	8,84	83.596	139.838	45,17
Redis	4.455	20,38	3.175	5.665	34,69
SSH	5.524.374	5,64	2.088.597	5.644.713	34,87
Telnet	836.765	75,61	76.693	845.980	7,98
UPNP	2.611	21,60	2.962	5.573	0,00

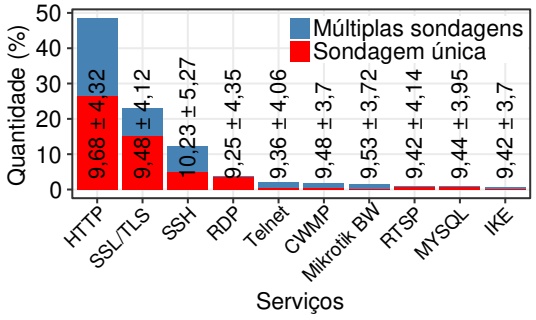

Figura 7. Top 10 serviços pelo Censys.

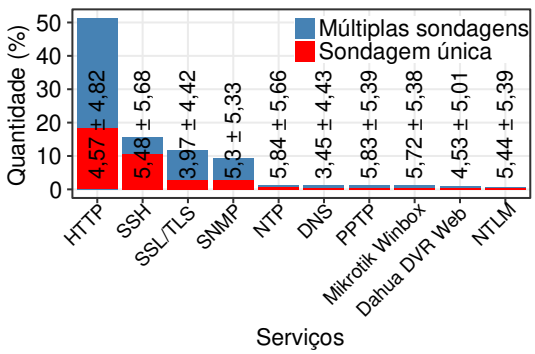

Figura 8. Top 10 serviços pelo Shodan.

serviços de FTP foram sondados pelos dois motores, serviços como o UPNP não apresentaram nenhum compartilhamento, ou seja, os dois motores sondaram conjuntos disjuntos de processos. Esses exemplos mostram o impacto que a diferença no intervalo de portas monitoradas pode causar na identificação de serviços.

Como o Censys sonda um número maior de portas que o Shodan, o intervalo entre sondagens de um mesmo serviço no Censys tende a ser maior. Em geral, o Shodan possui um tempo médio entre sondagens de 5,44 $\pm$ 4,07 dias, enquanto que o Censys é de 7,60 $\pm$ 4,84 dias. O cálculo do intervalo de confiança de 95% nos diz que o Shodan possui um intervalo menor entre sondagens de (2,15–2,16) dias. Os gráficos da figura 7 (Censys) e 8 (Shodan) mostram a lista dos 10 serviços mais sondados por cada motor e o tempo médio (em dias) entre sondagens para um mesmo dispositivo. É possível observar que, enquanto um serviço HTTP (p.ex., página *web*) é avaliado a cada 4,57 dias no Shodan, no Censys esse intervalo é 9,68 dias. Além disso, cerca de 18% dos serviços HTTP identificados no Shodan foram sondados apenas uma vez (representado pela parte vermelha da barra). Para o Censys, esse número aumenta para 27%. Como um *banner* possui informações sobre apenas uma porta (e por consequência, apenas um serviço), o fato do Censys permitir sondagens em aproximadamente 9,5$\times$ mais portas que o Shodan aumenta sua área de cobertura. Esse fato pode estar relacionado ao intervalo maior para a atualização de uma sondagem. Em contrapartida, como o Shodan possui menos portas em seu conjunto, as sondagens de um mesmo serviço tendem a ser mais frequentes que as do Censys. Dessa forma, a taxa de atualização mais frequente do Shodan para serviços monitorados pode ser preferível caso os serviços de interesse do operador de rede sejam cobertos por ele.

4.5. Diversidade dos serviços monitorados

Nesta seção avaliamos a quantidade de tipos de serviços (*i.e.*, aplicações e protocolos) monitorados pelos motores de busca. O Censys identificou 116 serviços no período

Tabela 3. Relação dos tipos de serviços identificados em cada motor de busca.

Categoria	Shodan			Censys		
	#	%	Top 1 (%)	#	%	Top 1 (%)
Aplicações e Servidores	11	51,44	HTTP (51,37)	**15**	48,85	HTTP (48,57)
Casa Inteligente e protocolos para IoT	**21**	1,62	Dahua DVR Web (0,93)	6	0,07	MQTT (0,07)
Criptomoeadas	**5**	<0,01	Bitcoin (<0,01)	3	<0,01	Bitcoin (0,01)
Jogos e streaming	**9**	0,09	Plex (0,05)	6	0,11	SIP (0,01)
Protocolos e equipamentos de rede	**34**	15,48	SNMP (9,28)	29	6,12	CWMP (1,77)
Protocolos e equipamentos para controle e automação	10	<0,01	Siemens S7 (<0,01)	**14**	<0,01	Modbus (<0,01)
Repositórios & Compartilhamento de dados	**24**	1,61	FTP (0,71)	21	2,81	RTSP (1,04)
Segurança & VPN	8	12,84	SSL/TLS (11,74)	**11**	23,82	SSL/TLS (22,96)
Terminal e Acesso Remoto	6	16,91	SSH (15,81)	**11**	18,22	SSH (12,21)

analisado, sendo que para 50 serviços ele apresenta informações detalhadas; para os 66 serviços restantes, foi coletada apenas a resposta da requisição do Censys sobre o serviço. Já o Shodan permite detectar a execução de mais de mil tipos de serviços, que variam desde servidores de telefonia PBX a aplicações como o Apache ActiveMQ. Para 128 serviços, ele apresenta informações estruturadas com o enriquecimento de algumas informações a partir de metadados e de outras fontes, que são armazenados em campos específicos do serviço. Dessa forma, o Shodan apresenta uma maior diversidade de serviços identificados e uma qualidade maior de informações.

A tabela 3 mostra a distribuição dos serviços agrupados por tipo e a representatividade de cada grupo nos dados de cada motor de busca, bem como o serviço mais frequente de cada grupo. Da lista geral de serviços identificados, 37 deles são compartilhados entre ambos sistemas, sendo aplicações e protocolos de rede comumente utilizados como HTTP, SSH, SMTP, NTP e DNS (não mostrado na tabela). Pela tabela, podemos observar que o Shodan possui mais tipos de serviços identificados nas categorias relacionadas a IoT e equipamentos de rede, enquanto o Censys possui um pouco mais de destaque em serviços de segurança e serviço remoto. Dessa forma, apesar do Shodan monitorar mais serviços, a combinação dos dois motores fornece dados complementares e a escolha de um motor específico depende do tipo de serviço a ser avaliado pelo operador de rede.

4.6. Profundidade de coleta de informações

Nesta seção avaliamos a profundidade dos dados coletados por serviço pelos motores de busca. Classificamos o enriquecimento de informações em dois tipos: sobre os serviços e sobre os dispositivos. Exemplos de enriquecimento de informações sobre os serviços sondados são a forma de estruturação dos dados e inclusão de metadados sobre o serviço. Exemplos de enriquecimento do segundo tipo são localização e sistema operacional.

Ilustramos o enriquecimento de informações sobre o serviço HTTP para cada motor. Enquanto o Censys agrupa as informações sobre a coleta de um serviço HTTP em quatro colunas principais, o Shodan fornece uma divisão de até 18 campos que incluem informações tanto da página *web* e serviços dependentes quanto respostas das requisições. Esse comportamento se repete para 80% dos tipos de serviços compartilhados entre os dois motores, como CoAP (Constrained Application Protocol), com uma

mediana de 26,6× mais caracteres em seu conteúdo inferido; Redis (17,97×) e MQTT (9,83×). Em geral, grande parte dos serviços relacionados a repositórios de dados possuem mais tipos de informações no Shodan. Por exemplo, enquanto o Censys se limita a coletar informações de *build* do serviço, como versão e parâmetros de execução, o Shodan executa requisições adicionais e fornece informações específicas do serviço, como lista das tabelas em um banco de dados, se o serviço monitorado não possuir autenticação.

Em relação ao enriquecimento de informações sobre os dispositivos coletados, o Shodan é capaz de fornecer, em média, 16 informações complementares sobre a sondagem. Exemplos de complementação são o número do Sistema Autônomo (AS), Provedor de Serviços de Internet (ISP), localização, MAC, organização responsável pelo IP, *hostnames*, domínios, sistema operacional, tipo de dispositivos, tags sobre as inferências, códigos CPE[7] e CVEs dos produtos e suas vulnerabilidades inferidas. Já o Censys possui, em geral, apenas 5 informações complementares: informações sobre o AS, localização, tags, sistema operacional e código CPE do produto inferido. Dessa forma, o Shodan permite uma análise de vulnerabilidades mais aprofundada para estudos mais complexos. Por exemplo, a organização responsável pelo IP, fornecida pelo Shodan, pode permitir agrupar vulnerabilidades por área de atividade na tentativa de encontrar tendências de ataques a determinados tipos de empresas. A presença do CVE, também fornecido pelo Shodan, facilita a identificação de quais vulnerabilidades estão documentadas para um código CPE. Além disso, informações como endereço MAC e tipo de dispositivo podem ajudar pesquisadores a compreender melhor a utilização da rede. Dessa forma, do ponto de vista de enriquecimento de informações, o Shodan atualmente fornece um serviço melhor.

Entretanto, vale observar que grande parte dessa diferença de enriquecimento de informações sobre dispositivos é um tópico importante apenas quando o usuário utiliza apenas a interface *web* do motor de busca, onde suas análises são limitadas ao que é fornecido pelo motor. Para usuários que interagem via APIs ou por *dumps*, como o arcabouço proposto no presente trabalho, parte dessas informações podem ser obtidas de outra forma. Por exemplo, em nosso arcabouço, CVEs podem ser inferidos a partir dos códigos CPEs informados pelo motor (como demonstrado na seção 4.7); nomes das organizações podem ser obtidos de bases externas, como o Registro de Roteamento da Internet (do inglês, IRR), e até enriquecidos com informações socioeconômicas a partir do CNPJ; tags podem ser criadas a partir do cadastro de padrões, entre outros. Nesse contexto, nosso arcabouço provê uma maior compatibilidade entre os motores de busca, mas ainda haverá a diferença entre as sondagens de serviços e o tipo de informação inferida sobre eles.

4.7. Caso de uso sobre complementação de dados das ferramentas

Nesta seção apresentamos um caso de uso de como motores de busca podem ser utilizados de forma complementar para avaliar vulnerabilidades em serviços de repositórios de dados. Combinar os dados coletados por cada motor aumenta o número de serviços avaliados. Por exemplo, apenas o Censys traz informações detalhadas para o Postgres, enquanto o CouchDB é identificado apenas pelo Shodan. Essa combinação permite ainda aumentar a taxa de atualização desses serviços. A tabela 4 apresenta um conjunto de sete serviços de repositórios de dados populares presentes no período avaliado. O resultado combinado das sondagens aumenta o número de serviços coletados por dia em até 1,8 vezes, como o aumento de 5.517 para 9.793 sondagens diárias no caso do MySQL.

[7]Common Platform Enumeration (CPE) é um código estruturado para catalogar sistemas e *hardwares*.

Tabela 4. Serviços com repositórios expostos na Internet.

Serviço	Mesclado		Shodan		Censys	
	Total	# / dia	%	# / dia	%	# / dia
MySQL	142.638	9793,3	58,6	4585,5	83,7	5517,7
MSSQL	21.075	1712,2	13,1	189,7	86,9	1522,4
Postgres	37.581	3135,4	-	-	100,0	3135,4
Redis	5.665	140,7	56,0	79,7	78,6	65,2
MongoDB	4.082	406,9	40,3	126,7	93,1	291,6
Elasticsearch	244	40,8	63,5	22,3	97,1	20,8
CouchDB	158	10,9	100,0	10,9	-	-

Em muitas análises de vulnerabilidades, particularmente daquelas recém-publicadas (*N-days*), a capacidade de ter resultados recentes é um requisito importante. Por exemplo, em nossas análises encontramos 1.893 serviços Redis vulneráveis ao `CVE-2023-45145`, uma vulnerabilidade publicada no dia 18/10/2023, dentro do período avaliado. Embora o Shodan já identifique algumas vulnerabilidades, relacionando o seu código CVE em seu enriquecimento (seção 4.6), para o Censys essa informação ainda não está disponível. No entanto, utilizando os operadores que disponibilizamos em nossa abstração de dados (seção 3.3), podemos inferir as vulnerabilidades confrontando o CPE informado pelos motores com o intervalo de CPEs de cada vulnerabilidade presente no catálogo do NIST, fornecido pelo nosso arcabouço. Para este cenário, o uso de dados dos dois motores permite um ganho de 1,7$\times$ no número de serviços, aumentando o panorama de dispositivos vulneráveis logo no primeiro dia de publicação.

Outro exemplo de avaliação é partir de métricas como o Exploit Prediction Scoring System (EPSS), que estima a probabilidade de uma vulnerabilidade ser explorada nos próximos 30 dias e, por causa disso, tem seus pesos recalculados diariamente.[8] Métricas desse tipo tornam ainda mais necessário o processamento diário dos dados. Exemplificamos três tipos de *softwares* identificados pelos dois motores de busca, cujas versões foram vinculadas a vulnerabilidades com EPSS maiores que 80% durante o intervalo da amostra: O `CVE-2008-0226` foi identificado em 427 serviços MySQL, com valor de EPSS de 97,4%; o `CVE-2015-1427`, vulnerabilidade de 12 serviços Redis, com valor de 82,7%; e o `CVE-2015-8080`, vulnerabilidade com risco EPSS de 86,7% em 1.334 serviços ElasticSearch. Para esses três cenários, mesmo com número baixo de dispositivos vulneráveis, o valor do EPSS indica que as vulnerabilidades possuem um alto risco de serem exploradas. Serviços desse tipo, expostos na Internet, representam uma grande ameaça ao sigilo e integridade dos dados. Esses repositórios podem ser comprometidos por *hackers* que podem furtar dados sensíveis como identificação pessoal ou informações corporativas. Avaliações dinâmicas, como a análise de EPSS, podem auxiliar órgãos como o CERT.br a elaborar campanhas de conscientização mais efetivas e operadores de redes a avaliarem melhor a segurança dos dispositivos conectados em suas redes.

5. Trabalhos Relacionados

Grande parte dos estudos sobre motores de busca de dispositivos são limitados a análises utilizando a interface *web*, sem a possibilidade de pós-processamento nos dados, algoritmos mais complexos ou a integração com outras bases [Al-Alami et al. 2017, Raikar e Maralappanavar 2021]. Esse fato pode ser o motivo pelo qual a avaliação

[8]Nosso arcabouço também fornece um coletor para a obtenção do catálogo de EPSS diário.

de aspectos técnicos, como taxa de atualizações de sondagens, abrangência de IPs e comparações de informações, vêm sendo pouco discutidas. Além disso, como os motores de busca estão em constante evolução, conclusões baseadas em estudos antigos ou para outras regiões podem não refletir o estado atual para a Internet brasileira.

[Lee et al. 2017] conduziram uma investigação para entender os padrões de acesso de cada motor com o objetivo de compreender como os administradores de rede podem bloquear as varreduras. Porém, devido ao avanço das ferramentas, as conclusões do artigo podem estar desatualizadas; por exemplo, em 2017 o Censys só explorava 35 portas. Outros trabalhos mais recentes [Li et al. 2020, Bennett et al. 2021, Zhao et al. 2022] buscam comparar os motores de busca em aspectos como o suporte a protocolos de Internet, a quantidade de dispositivos detectados, o nível de informações e a frequência de varredura, utilizando abordagens distintas e cenários mais limitados. Por exemplo, os experimentos realizados em [Bennett et al. 2021] baseiam-se em monitorar o acesso desses motores a um conjunto de máquinas preparadas pelos pesquisadores com quatro serviços em execução (HTTP, HTTPS, FTP e SSH), possibilitando análises do ponto de vista do dispositivo sondado. Como conclusão, os autores sugerem que o Censys é mais focado em HTTP/HTTPS, com coletores capazes de sondar diversas portas, enquanto o Shodan possui uma distribuição de portas sondadas mais balanceada. Similar aos nossos resultados, os autores também apontam que o tempo de sondagem de uma porta é ligeiramente mais rápida no Censys. No entanto, o Shodan é mais rápido na captura de novos *banners* e em torná-los acessíveis por meio de sua interface de pesquisa.

O trabalho de [Zhao et al. 2022] é o mais relacionado ao nosso. No artigo, os autores comparam quatro motores, incluindo o Censys e o Shodan, na descoberta de seis diferentes tipos de dispositivos de IoT ao redor do mundo, como roteadores, câmeras IP e impressoras, ao longo de um período de cinco meses em 2018. Encontramos alguns resultados similares ao trabalho de [Zhao et al. 2022]. Por exemplo, o Censys raramente repete a varredura de um mesmo serviço em um curto período de tempo e, dentre os motores avaliados, o Shodan foi o único capaz de sondar o espaço de endereços observados mais de uma vez para todos os tipos de serviços alvo das análises. No entanto, também observamos diferenças. Por exemplo, os autores argumentam que o Shodan é capaz de sondar todo o espaço de endereçamento observado em um intervalo médio de 17 dias, o que não observamos em nossos resultados.

6. Conclusão e Trabalhos Futuros

Este trabalho apresenta um arcabouço capaz de processar dados em larga escala de motores de busca de diversas fontes em uma interface única e amigável para operadores de rede e pesquisadores, reduzindo a complexidade de codificação para operadores de rede. Utilizando esse arcabouço, realizamos uma comparação experimental da capacidade de motores de busca como o Shodan e o Censys no rastreamento de dispositivos conectados à Internet e suas vulnerabilidades. Em nossos resultados, observamos diferenças significativas entre o modo de operação entre os sistemas. Por exemplo, o Censys mostrou ter uma maior cobertura de IPs e portas para o período analisado, enquanto que o Shodan mostrou um tempo de atualização de sondagens mais rápido e uma melhor capacidade de enriquecimento das informações coletadas. Também ilustramos como os dados de ambos os motores podem ser combinados para gerar análises mais precisas. Como trabalhos futuros, pretendemos incluir novos operadores, avaliar outros motores de busca e realizar uma caracterização temporal das diferenças entre os motores ao longo dos anos.

Agradecimentos

Este trabalho foi parcialmente financiado pelo NIC.br, RNP/CTIC (2955), FAPEMIG, CNPq, CAPES, MASWEB, INCT-Cyber e IAIA (INCT para IA).

Referências

Al-Alami, H., Hadi, A., e Al-Bahadili, H. (2017). Vulnerability scanning of IoT devices in Jordan using Shodan. In *Int. Conf. on the Applications of Information Technology in Developing Renewable Energy Processes Systems*, Amman, Jordânia. IEEE.

Bennett, C. et al. (2021). Empirical scanning analysis of Censys and Shodan. In *Workshop on Measurements, Attacks, and Defenses for the Web*, Online. The Internet Society.

Câmara, J. (2023). De CPF a fotos: UFMS confirma que dados pessoais de alunos foram acessados por hackers em vazamento. Disponível em: https://t.ly/TeTxE. Acessado em 12/01/2024.

Durumeric, Z. et al. (2015). A Search Engine Backed by Internet-Wide Scanning. In *Proc. of ACM SIGSAC Conf. on Computer and Comm. Security*, Denver, EUA. ACM.

IT Section (2024). Ransomwares aumentam 13% em 2023, atingindo quase 5 mil incidentes. Disponível em: https://t.ly/MrMPJ. Acessado em 12/01/2024.

Lee, S. et al. (2017). Abnormal Behavior-Based Detection of Shodan and Censys-Like Scanning. In *2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN)*, pages 1048–1052, Milão, Itália. IEEE.

Li, R. et al. (2020). A Survey on Cyberspace Search Engines. In *China Cyber Security Annual Conference*, pages 206–214, Beijing, China. Springer.

Matherly, J. (2015). Complete Guide to Shodan: Collect. Analyze. Visualize. Make Internet Intelligence Work For You. *Shodan, LLC (2016-02-25)*, 1.

Mousavi, S. H. et al. (2020). A fully scalable big data framework for Botnet detection based on network traffic analysis. *Information Sciences*, 512:629–640.

Ortiz, B. e Mendes, M. (2023). Polícia do DF prende hackers suspeitos de invadirem computadores de hospital em Taguatinga e exigirem resgate. Disponível em: https://t.ly/B-mSo. Acessado em 12/01/2024.

Ponce, L. et al. (2023). Um Arcabouço para Processamento Escalável de Vulnerabilidades e Caracterização de Riscos à Conformidade da LGPD. In *Anais do XXIII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais*, pages 15–28.

Raikar, M. e Maralappanavar, M. (2021). Vulnerability assessment of MQTT protocol in Internet of Things (IoT). In *Int. Conf. Cyber Secur.*, pages 535–540, Índia. IEEE.

Statista (2023). Countries with the largest digital populations in the world as of January 2023. Disponível em: https://bit.ly/3TQBqkb. Acessado em 12/01/2024.

Zhao, B. et al. (2022). A Large-Scale Empirical Study on the Vulnerability of Deployed IoT Devices. *IEEE Trans. Dependable Secure Comput.*, 19(3):1826–1840.