Censys: A Map of Internet Hosts and Services

Zakir Durumeric
Censys, Inc.
Ann Arbor, MI, USA

Hudson Clark
Censys, Inc.
Ann Arbor, MI, USA

Jeff Cody
Censys, Inc.
Ann Arbor, MI, USA

Elliot Cubit
Censys, Inc.
Ann Arbor, MI, USA

Matt Ellison
Censys, Inc.
Ann Arbor, MI, USA

Liz Izhikevich
Censys, Inc.
Ann Arbor, MI, USA

Ariana Mirian
Censys, Inc.
Ann Arbor, MI, USA

ABSTRACT

In 2015, we released Censys to lower the barrier to entry for researchers to study Internet devices by continually collecting and packaging Internet scan data. Since then, as we have learned more about how best to capture the complex behavior of Internet services and begun to serve commercial and government users, we have re-architected every aspect of how Censys operates. Motivated by requests from the community, we present Censys' evolution and current architecture, evaluate its visibility, and detail how Censys has been used by research, industry, and government. Finally, informed by our operational experiences, we discuss unsolved problems and the lessons we have learned. We hope that our work provides the transparency needed for researchers to soundly use Censys data and offers directions for future research.

CCS CONCEPTS

• **Networks** → **Public Internet**; • **Security and privacy** → **Network security**; • **General and reference** → **Measurement**;

KEYWORDS

Internet Scanning, Transition to Practice, Web, DNS, PKI

ACM Reference Format:
Zakir Durumeric, Hudson Clark, Jeff Cody, Elliot Cubit, Matt Ellison, Liz Izhikevich, and Ariana Mirian. 2025. Censys: A Map of Internet Hosts and Services. In *ACM SIGCOMM '25, September 8–11, 2025, Coimbra, Portugal.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3718958.3754344

1 INTRODUCTION

Enterprises, governments, and researchers use Internet scan data to remediate security vulnerabilities, track supply chain dependencies, uncover attacker-controlled infrastructure, and study Internet behavior [35]. However, while high-performance tools like ZMap and Masscan have reduced the time needed to conduct Internet scans [39], maintaining a comprehensive, up-to-date map of Internet infrastructure is a much more challenging problem that requires identifying, interrogating, and statefully tracking billions of continuously changing Internet services in close to real time.

In 2015, we launched Censys [32] to provide reliable Internet scan data to the research community. Since then, as we have learned more about the complex real-world behavior of Internet services, we have continually updated Censys behind the scenes to more comprehensively and accurately scan the Internet. Censys has also grown from an academic project into a standalone company. As we began to support commercial and government users, operational use cases pushed us to uncover security-critical services, collect deeper context, and build more intuitive data abstractions. Operators, for whom a single inaccurate data point can impact security, also led us to rethink our approach to data quality.

There is little overlap between how Censys operates today and how it was architected in 2015. Yet, despite the changes in how we collect and present data, we have not documented Censys' evolution in the research literature. Motivated by direct requests from the networking community [23], misconceptions about Censys' behavior [4, 16, 63, 117, 121, 122], and the importance of transparency for a platform used by hundreds of researchers [43], in this paper, we present how Censys has changed and operates today, as well as what we have learned in the process.

Censys' foremost goal has evolved from collecting raw Internet scan data to maintaining a comprehensive and cohesive map of Internet entities like hosts and websites. At our foundation, we have re-architected our scan engine to continuously scan 200+ protocols on 65K ports from multiple geographic vantage points using a combination of comprehensive scans and probabilistic models that predict likely service locations. We statefully track Internet entities and aggressively prune out stale data to ensure correctness. Our approach to presenting data has also evolved. While Internet scanning remains central to how we collect data, we no longer directly organize data by how it was collected. Instead, our pipeline extracts non-ephemeral data from scans, derives higher-level context like device manufacturers and software versions, and then assembles cohesive records about every Internet entity.

Today, Censys finds 188% more IPv4 services than in 2015 (794M vs. 275M IPv4-based services). In addition, we have added support for IPv6 services and name-based web properties. To estimate our coverage, we evaluate our visibility against both subsampled 65K port scans that approximate ground truth and other popular scan engines like Shodan and ZoomEye. We estimate that, at any given time, Censys sees 98% of IPv4-based services on the top 10 ports, 97% of the top 100 ports, and 62% of services across all 65K ports. Compared to other popular scan engines, we find that

SIGCOMM '25, September 8–11, 2025, Coimbra, Portugal

ACM ISBN 979-8-4007-1524-2/2025/09…$15.00
https://doi.org/10.1145/3718958.3754344

Censys surfaces 33–170% (180M–460M) more live IPv4 services and achieves an estimated 35–820% higher data accuracy.

After describing Censys' architecture and evaluating its coverage, we conclude with an overview of how Censys is used in research and industry, discuss our ethical considerations, and propose future research directions. We hope that our work provides the community with the visibility needed to soundly conduct research using Censys.

2 BACKGROUND AND RELATED WORK

Researchers have long used network scanners like Nmap to study Internet behavior [8, 37, 54, 55, 92]. In 2013, two new tools, ZMap [39] and Masscan [50], further popularized scanning as an Internet measurement technique by dramatically reducing the time required to survey the full IPv4 address space. Today, fast Internet-wide scanning is a commonly used research methodology that has been used in hundreds of studies [35]. In this section, we describe Internet service search engines that use Internet scanning, the complexities of real-world service deployment that affect our architecture, and relevant related work. We refer readers to Durumeric et al. [35] and Griffioen et al. [52] for a more in-depth discussion of how Internet-wide scanning has evolved over the past ten years.

2.1 Internet Scanning Search Engines

While tools like ZMap lowered the effort needed to conduct Internet scans, a single scan provides coverage for only a subset of the Internet (e.g., several ports and protocols) at a single point in time from a single vantage point. Internet scans are also bandwidth and resource intensive to operate, generate significant amounts of unstructured data (e.g., terabytes of HTML), and are prohibited by many cloud providers due to reputational concerns [32, 36]. To make scan data more accessible, several organizations built "Internet search engines" that regularly conduct scans and index comprehensive, longitudinal datasets of Internet services and devices.

The first and most notable Internet search engine is Shodan [78], which launched in 2009 and surfaced the Internet exposure of IoT and ICS devices. Shodan remains popular today. Frustrated by data quality issues (e.g., poor coverage and stale data), we launched Censys in 2015 to provide more dependable data to academic researchers. Censys originally ran weekly ZMap and ZGrab scans of 11 popular protocols, which were aggregated by IP address and served through Elasticsearch and Google BigQuery [32]. Censys initially scanned fewer ports and protocols than Shodan, but, for the protocols scanned, provided more accurate and complete data [32].

Since Censys' debut in 2015, several other engines have emerged, including ZoomEye, BinaryEdge, Netlas, and ONYPHE. While it is difficult to externally measure the popularity of these products, Li et al. that the most popular are Shodan, Censys, BinaryEdge, ZoomEye, and Fofa [16]. Since the study, BinaryEdge was acquired by Coalition (a cyberinsurance company) and shut down in March 2025 [25]. Relatively little is known about how these engines operate; we evaluate them along with Censys in Section 6.

2.2 Internet Scanning Complexities

Internet-wide scanning was initially thought to be straightforward using tools like ZMap [32, 39]. However, since then, several studies have shown that real-world service behavior is more complex than initially assumed [35]. We briefly summarize key research results below and their implications for surveying the Internet:

Service Diffusion. While early Internet scanning tools relied on scanning a handful of popular ports, Bano et al. [15] and Izhikevich et al. [60] showed that the vast majority of Internet services live on non-standard ports. Indeed, under 3% of HTTP services reside on TCP/80 [60]. In response, several ML-based solutions have been proposed for predicting service locations [61, 64, 72, 103, 107, 108]. This service diffusion introduces three challenges for scan engines: (1) continuously finding services across all 65K ports, (2) identifying the L7 protocol running on each port, and (3) building a data representation that supports any protocol running on any port.

Short Service Lifespans. DHCP churn and auto-scaling cloud services lead to short-lived services [89, 90, 95, 98], which can quickly degrade the accuracy of collected datasets. The severe impact on data coherency was most clearly shown by Antonakakis et al. in their attempt to combine scans across sequential days [13]. Maintaining up-to-date data requires balancing bandwidth between finding new services and updating services as they change and disappear. It also forces a trade-off between coverage and accuracy depending on how quickly potentially stale data is pruned out.

Fractured Visibility. Wan et al. showed that scans from a single vantage point achieve worse coverage than first estimated and that aggressive scanners experience significant blocking [118]. McDonald et al. further showed that web-based geoblocking is increasingly common [80]. After the 2022 invasion of Ukraine, Ramesh described how Russia began to block foreign access [96]. This has several implications: (1) scan engines need to account for Internet balkanization and transient Internet outages to reliably find all Internet hosts; and (2) engines cannot simply scan more aggressively to find all Internet services or to update data, because increased scanning leads to increased blocking, reducing coverage.

Beyond IPv4. While our work primarily focuses on IPv4 scanning—where there have been considerable challenges to address—related work has made significant progress in identifying IPv6-addressed hosts [1, 17, 46, 47, 57, 58, 70, 85, 102, 109, 124] and name-addressed HTTP(S) services [19, 29, 100, 104, 115].

The challenges discussed here do not have easy solutions and many remain active areas of study by the research community. Without clear solutions, these complexities oftentimes pose design tradeoffs (e.g., port coverage vs. potential blocking) that we have to balance as we maintain and evolve Censys.

3 DESIGN GOALS

Censys' foremost goal is to maintain a comprehensive and up-to-date map of Internet devices, services, web properties (i.e., websites), and certificates. We detail our design goals below:

User-Centric Data Model. While we originally built Censys to provide access to Internet scan data, we quickly learned that tying data presentation directly to data collection is unnatural for most people. Our users care primarily about the entities on the Internet (e.g., servers, embedded devices, and websites) and their configurations, not how we collect data about those entities. Today,

our goal is to build data abstractions that match users' conceptualizations of Internet entities and allow them to answer their questions, a UX driven shift that informs what data we collect (the opposite of where we started). Building an intuitive data model requires not only finding services, but also deriving higher-level attributes like service configuration, software version, device manufacturer and model, vulnerability to attack, geolocation, and organizational ownership. Storing ephemeral scan results is explicitly not a goal. While this precludes answering certain research questions (e.g., [34, 110, 113, 120]), storing trillions of handshakes poses significant cost and benefits only a small fraction of our user base.

Accuracy and Coverage. When scanning, Censys broadly optimizes for: (1) finding all Internet services, and (2) ensuring that returned data reflects the current state of the Internet and is accurate. While we measure coverage in terms of all Internet services, we emphasize that not all services are equal. Notably, uncovering many of the most security-critical services (e.g., industrial control systems) has negligible impact on aggregate coverage due to their infrequency but are critical for operational use cases. We have also found that people naturally expect different levels of comprehensiveness across different services (e.g., users often expect complete coverage of services on port 80 but not across all 65K ports).

Broadly, we have observed that most academic research considers longitudinal trends where individual data points are unimportant. Indeed, ZMap originally advertised its speed at the cost of losing an expected 3% of responsive services [39]. In stark contrast, most operational use cases have little interest in longitudinal data, but care deeply about specific exposures that could allow an attacker initial access or could identify threat actor infrastructure. Yet, despite the importance of coverage, accuracy (whether the data returned to the user is correct at query time) is more important than coverage for most of our users. False positives waste time that could otherwise have gone towards fixing real security problems and erodes users' trust. We prioritize accuracy over coverage when the two conflict (e.g., when considering how aggressively to prune out a service that just went offline and may return).

While scanning IPv6 is academically interesting, it is less important for many of our users today. Most IPv6 services are dual-stack (i.e., accessible over both IPv4 and IPv6) and patching IPv6-only services that would never have been practically uncovered by an attacker is a low priority for many security teams drowning in other work. Many of our users are however interested in global visibility into other types of named Internet entities (e.g., websites and S3 buckets) that are readily discoverable through other means like open source subdomain enumeration tools and public CT logs.

Continuous, Real Time, and Global. While many predictive scanning systems are designed to operate at one point in time [39, 61, 103], Censys must continuously maintain a real-time view of the global Internet, capturing the state of services as they change. Users expect to be able to accurately query Censys at any time and understand how Internet entities have changed over time. Continually learning about service deployment over time while updating an existing dataset is a fundamentally different problem than predicting as many services as possible at one point in time.

Beyond coverage and accuracy—which are metrics computed at one point in time—we optimize for minimizing the time to find new services that come online, to update them as quickly as possible after they change, and to remove them as they go offline. We note that because excessive scanning leads to blocking, we must maximize finding new services while minimizing bandwidth usage.

Balanced Access. While we initially designed Censys to expose all of the data we collect to all of our users, recent results indicate that scan engines like Shodan and Censys have been used by attackers to identify vulnerable services [22, 62]. As we increasingly scan for security-critical systems, it becomes increasingly important to provide operators with visibility without arming attackers. Today, our goal is not to provide all users with the same global Internet visibility, but to provide tailored access driven by users' needs to minimize potential abuse and ultimately improve Internet security.

4 DATA COLLECTION

Censys primarily collects data about the IP-addressed hosts and name-addressed web properties that compose the Internet. While we originally ran weekly timed ZMap and ZGrab IPv4 scans of 11 popular protocols, Censys has since evolved to continuously discover services across all ports and protocols, as well as capture web properties. In this section, we describe how we discover and collect data about these Internet entities. We note that while some systems are used to scan both internal networks and the broader Internet, we have found that the many nuances of continuously collecting global Internet data require a tailored approach that does not translate well to internal networks. Building a portable scanner is not a goal.

4.1 Service Discovery

Censys collects data about IP-based services through continuous "two-phase" scanning [56]. In the first phase ("Service Discovery"), Censys identifies potential service locations through a combination of comprehensive IPv4 scans and a predictive engine that learns deployment patterns and recommends probable service locations to probe. Since L4 responsiveness does not reliably indicate the presence of an actual service [60], we do not directly publish L4 scan data (e.g., responses to TCP SYN scans). Rather, we treat L4 responsive (or predicted) services as candidates for L7 investigation.

Comprehensive IPv4 Scans. Censys continuously conducts stateless L4 discovery scans targeting all IPv4 addresses and ports at varying rates per port and network. Our scan engine operates similarly to ZMap [7, 39] by iterating over sets of cyclic groups that cover targeted IPs and ports, statelessly sending TCP SYN packets that mimic those sent by modern Linux systems and protocol-specific UDP packets using a simple userspace network stack. Responsive services are queued for stateful L7 application-layer interrogation. Censys runs three sets of discovery scans:

Common Ports and Protocols. To ensure comprehensive coverage of commonly used ports and protocols, Censys scans approximately 100 of the most responsive ports and approximately 100 ports with IANA-assigned protocols of interest daily. When a new CVE is discovered, Censys will also scan relevant ports more frequently for several weeks to support reporting [31, 33].

Dense, High-Churn Networks. To maintain up-to-date data about cloud environments where elastic services have shorter lifespans,

Censys scans known cloud networks (e.g., Amazon EC2 [14]) on 300 ports associated with cloud infrastructure at least once a day in addition to the global scans.

Background 65K Ports. To collect data for our predictive engine and to uncover long-lived services on non-standard ports, Censys continuously scans all 65K TCP ports at approximately 8 Gbps. This results in scanning every IPv4 address on 100 random ports daily and completing a full scan of all 65K ports across the IPv4 address space every nine months. This replaced a weekly scan targeting the top 5,000 ports that ran from 2000 to 2003 (Appendix B).

Scan Engine. Across all discovery scans, Censys sends around 26.5M probes/second (17 Gbps), identifying on average 11K potential services/second (950M services/day). While Censys originally used ZMap for L4 scanning and a fixed schedule for running scans, we found this approach rigid. We transitioned to a proprietary engine in 2018 that: (1) runs continuously, (2) supports scanning multiple ports with multiple probes (e.g., sending a TCP SYN probe on port 80 and UDP DNS probe on port 53), and (3) is written in Go (to eliminate memory-unsafe code; ZMap uses C [35]).

We scan continuously rather than on a fixed schedule, distributing traffic evenly across source IP addresses and time. A continuous approach minimizes temporal biases (e.g., weekdays vs. weekends), prevents cascading failures if a system goes offline, enables consistent bandwidth usage, reduces the likelihood of blocking by spreading scans across a larger pool of source IPs, simplifies hardware scaling, and allows for finer-grained bandwidth allocation—all problems that we experienced with Censys' original fixed schedule.

Predictive Scanning. Since many services reside on non-standard ports [60] and scanning all ports requires months at a reasonable rate, Censys attempts to predict the locations of responsive services across the full 65K port space. Censys implements several dozen probabilistic models that rely on transport and application layer features along with network and geolocation data in an approach inspired by Izhikevich et al [61]. Utilizing predictive approaches introduces a trade-off: while Censys finds otherwise unknown services and increases our overall coverage, the methodology also introduces biases since only some service locations can be predicted. In addition, it prevents us from concretely explaining why certain services appear in the dataset. Censys optimizes for maximizing coverage since most of our users prefer coverage over explainability and perfect sampling. However, some researchers may want to conduct subsampled scans of the IP–Port space if they want a more representative sample of services.

IPv6 Scanning. While we do not conduct comprehensive IPv6 scans, we track and scan IPv6 addresses that we find through DNS queries of known names (e.g., found through CT logs, passive DNS data, and HTTP redirects) and have recently begun to roll out more intelligent IPv6 scanning in a method similar to 6sense [58]. We do not evaluate IPv6 given the early stages of our deployment.

4.2 Service Interrogation

In the second phase of our scanning ("Service Interrogation"), Censys collects application-layer data about previously identified services. We then use this data to build a structured record that describes each Internet service. Censys scanners specifically: (1) fetch batches of scan candidates found during Phase 1 Service Discovery, (2) attempt to detect the underlying L7 protocol, (3) complete any associated L7 handshakes, (4) create a structured data record describing the service, and (5) serialize and send the record for downstream processing. We have implemented approximately 200 protocol scanners, ranging from IETF-ratified protocols like HTTP and LDAP to security-critical ICS protocols like General Electric SRTP and Red Lion Crimson.

Protocol Detection. Since most services do not run on their IANA-assigned port, Censys attempts to fingerprint each services' L7 application-layer protocol using a detection algorithm inspired by LZR [60]: Censys listens for server-initiated communication, attempts the IANA-assigned protocol for the port (if one is assigned), and then tries additional common handshakes (e.g., HTTP GET request) to elicit a response that Censys can fingerprint. For example, if Censys receives an SMTP error in response to an HTTP request, it identifies the service as running SMTP. Censys attempts the same identification process within a TLS session if one can be established. In cases where application layer data is sent but cannot be fingerprinted, we capture the raw server response.

Data Collection. Censys attempts to complete a protocol handshake with each service that exposes an identifiable L7 protocol using custom high-performance protocol implementations in Go, similar to ZGrab [30]. In some cases, Censys will perform additional follow-up handshakes to build a complete picture of a service's configuration (e.g., fetch a favicon or compute JA4S fingerprint [9]). We do not attempt to infer higher-level attributes like device manufacturer or software vulnerabilities at this stage, since this is more accurately computed with data from across multiple services.

Coalescing data collected during application-layer handshakes, we build a highly-structured record about each service that captures *non-ephemeral* data. Parsed scan results (including failed scans) are enqueued in Google Pub/Sub as serialized Protobuf [49] objects for further processing within our cloud environment.

4.3 Web Properties

Thus far we have discussed how we scan IP-addressed services. However, the majority of HTTP(S) services are only accessible when addressed by name via TLS Server Name Indication (SNI) and/or HTTP `Host` header. These name-addressed web properties are equally interesting to our users for a few reasons: (1) they are frequently exploited (e.g., CL0P ransomware attacks against MOVEit managed file transfer sites), (2) they can be entry points for credential stuffing, (3) they can serve phishing and impersonation sites, and (4) they are common locations for attacker infrastructure, such as command-and-control (C2) servers and web shells. To capture these name-addressed services, Censys performs HTTP(S) scans against known names, which it collects from public CT logs, HTTP redirects, and third-party passive DNS subscriptions.

We initially began to perform name-based scanning as an extension to IP-based scanning by creating a new asset type—*Virtual Host*—inspired by virtual hosting in web servers like Apache HTTP Server and keyed by (IP, Port, Name). Unfortunately, we found that this was a poor abstraction as: (1) users often failed to understand the data model, and (2) cloud and CDN-based services caused endless growth in the IP records associated with each website. In

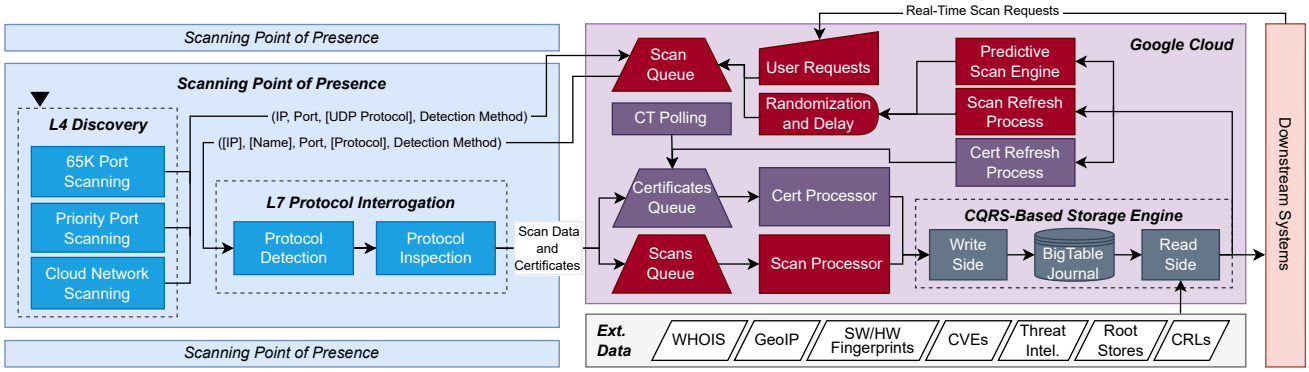

Figure 1: Overview of Censys Data Collection and Processing

2024, we migrated to a web-focused object type. We refer to these name-addressed HTTP(S)-served entities as *Web Properties*, because user testing revealed that users do not interpret the term *Website* to include services like back-office applications like Prometheus or REST APIs. We fetch the root (/) page of each web property and fetch additional endpoints based on the identified application.

4.4 X.509 Certificates

Censys collects X.509 certificates presented during TLS scans and by regularly polling public certificate transparency logs [66]. Certificates are valuable for locating web properties to scan and for identifying hijacked domains. In addition, we allow users to search across certificates since certificate transparency logs provide limited value without third-party indexing. Upon observing a new certificate, Censys parses out X.509 data, validates it against browser root stores, checks revocation via CRL, and lints it [65]. While we once used OCSP to track certificate revocation, we stopped checking OCSP in 2024 after CRL usage was mandated in CABF BR v2.0.1, and Let's Encrypt [5] began to publish CRLs.

4.5 Points of Presence (PoPs)

Wan et al. showed that geographically and topologically diverse vantage points surface different Internet services due to transient outages, routing anomalies, and, to a limited extent, geofencing [118]. To account for inconsistent visibility and maximize global coverage, Censys scans from multiple physical Points of Presence (PoPs) deployed at popular Internet Exchanges (IXPs) in North America (Chicago), Europe (Frankfurt), and Asia (Hong Kong).

Each PoP operates independently and houses Censys server and routing equipment for scanning. In each location, we route traffic through regionally dominant Tier-1 providers. For example, Censys purchases transit from NTT and PCCW in Hong Kong and from Orange S.A. and Arelion (previously Telia Carrier) in Frankfurt. We typically peer directly with Tier-1 ISPs (and other well-connected providers like Hurricane Electric) because this helps to diversify paths and minimizes the number of network hops where packets might be dropped (stateless scanning cannot detect packet loss).

Across the eight major Internet providers that Censys uses, we observe no correlation between bandwidth cost and scan coverage. Hurricane Electric, one of our least expensive providers, consistently sustains the highest hit rates, perhaps due to its large number of direct customer peerings. Between this observation and prior work demonstrating that most scanning packet drop is due to transient network outages [118], Censys optimizes for route diversity across PoPs and providers unless specific visibility weaknesses are observed in a geographic region or network.

4.6 Service Refresh and Eviction

To ensure data accuracy, Censys refreshes IP-based data at least daily and name-based HTTP(S) services at least monthly. In addition, Censys recomputes certificate validation and revocation status daily. To refresh scan data, Censys re-performs Service Interrogation as if the service had been found through an L4 discovery scan. If a service appears unresponsive from one PoP, we attempt to scan it from the other PoPs over the following 24 hours.

There is no clear answer as to how quickly to evict stale services after they appear to have gone offline. On the one hand, leaving stale data in the dataset leads to false positives during investigations and artificially inflates the time it appears to take security teams to resolve issues—a coveted metric tracked by many enterprises. On the other hand, removing data too quickly leads to churn where services are removed and then immediately re-added due to transient service and network outages. This is particularly problematic if the service belongs to a customer, because this can re-trigger a workflow that creates a ticket for remediation.

To allow users to decide on their own tolerance for churn vs. false positives, Censys' compromise is to mark services as pending eviction after the first scan fails, to include the last time Censys saw the service, and to remove the service after 72 hours. In addition, Censys' predictive engine will re-inject previously known services from the last 60 days into the scan queue after they have been pruned such that they will be quickly added back in case they were originally difficult to find but later return. As mentioned in Section 3, we optimize for data accuracy over coverage and we emphasize that because Censys prunes out stale services more aggressively than it finds new services across all ports, Censys' dataset underestimates the total services on the Internet.

5 PROCESSING AND PRESENTATION

Our data collection engine produces a significant continuous stream of inbound network handshakes, which have a fundamentally different structure than the abstractions that we want to present to

users. While we initially expected data collection to be the largest challenge, we have found that processing, storing, and contextualizing the network data is just as large as, if not a larger, challenge. In this section, we describe how Censys processes, stores, and presents a cohesive Internet Map as well as tracks its history.

5.1 Modeling Internet Entities

While Censys collects Internet data through scanning and web requests, we do not directly serve raw scan data, which many users find opaque and difficult to query. While academic researchers are often eager to explore raw datasets, industry users often are not. This is partly due to time constraints, but it also is due to differences in the types of questions each audience wants to ask. Researchers often have specific devices or protocols they are interrogating, whereas operators have use cases that relate more to a specific network, organization, or threat actor and can span hundreds of different devices and applications.

Instead of directly exposing scan data, Censys uses that raw data to construct data abstractions that model Internet entities like *Hosts*, *Web Properties*, and *Certificates*, which can be queried by users. We derive data like device type and manufacturer, but also surface lower-level network data (e.g., HTTP headers) for specific investigations. Similar to how our scanners extract structured, non-ephemeral data, Censys data abstractions are built to be stable and structured: data should not change if the configuration of the Internet entity does not. While this precludes certain academic research that directly studies network handshakes [101, 113], it dramatically reduces our storage and processing costs.

When possible, we use RFCs and other standards to guide our abstractions and naming conventions. Similarly, to enable interoperation with other data sources, we use MITRE CPE (Common Platform Enumeration), CWE (Common Weakness Enumeration), and CVE (Common Vulnerability Enumeration) schemas when deriving context about entities like device model or software version. However, because NIST's CPE Dictionary misses large swaths of devices, we do not restrict ourselves to official entries. Even with these standards, we have found building entity abstractions surprisingly difficult in practice. Even simple terms like "website" and "domain" carry different definitions across different communities.

5.2 Data Pipeline and Storage

To process inbound scan data and maintain a higher-level map of Internet entities, Censys utilizes a Command Query Responsibility Segregation (CQRS)-based architecture [123]. Most importantly, CQRS decouples the write-side and read-side data models and allows us to independently scale read-side and write-side processing [10]. This is critical for Censys because, unlike many systems that have a high read-to-write ratio, we are continually updating the state of Internet entities as scans complete.

Write Side (Command Side). Ingested scans are treated as *commands* in the CQRS model, which update the state of an entity (e.g., an IP-addressed *Host*, name-addressed *Web Property*, or SHA256-FP-addressed *X.509 Certificate*). To process an inbound scan, our processing pipeline: (1) retrieves the current state of the entity, (2) builds and applies an update *command* based on the inbound scan, (3) journals an *event* that updates entity's state (e.g., service found, changed, or removed); and (4) enqueues the event for later processing. In 2024, Censys processed around 5B events/day.

Backend Bigtable Event Journal. The journal of events that capture the entity's state is stored in Google Bigtable, keyed by *Entity ID* (e.g., IP address) and a monotonic *Sequence Number*. Journal events are delta encoded such that only differences to a service are stored to disk rather than the entire scan record since most services change very little across refresh scans. In addition, because journal events must be replayed to reconstruct the state of an Internet entity, Censys regularly snapshots entity state to minimize the maximum number of events that need to be retrieved for a query.

History. Our event journal is used not only for maintaining the current state of Internet entities, but also for storing their long-term history. Censys migrates journal events and historical snapshots prior to the latest snapshot from SSD-backed tables to HDD-backed tables. This guarantees that the current entity state can be quickly retrieved from SSD, but the bulk of our data is stored on less expensive rotational disk. Censys adds around 500 TB of data per year, post delta encoding and compression.

Asynchronous Event Processing. To support high throughput, Censys performs minimal processing during initial data ingestion. Instead, the write side processor enqueues any resulting update events for additional processing like updating our read-side data model, triggering follow-up interrogation, and sending data to downstream applications. For example, Censys asynchronously updates secondary tables that map from certificate fingerprint to IP address and triggers follow up JARM scans when a new TLS service is found or an existing TLS service changes.

Read Side (Query Side). Censys constructs the user representation of each Internet entity at read time by providing downstream applications with a set of APIs for querying entities by indexed IDs (e.g., IP address). Unless fetching the cached current state of an entity, lookup APIs operate by finding the latest snapshot prior to the desired timestamp, then fetching and applying any journal events after the snapshot, but prior to the queried timestamp. After an entity is reconstructed from the journal, the read-side processor fetches additional data (e.g., IP WHOIS, geolocation, origin ASN) and derives higher-level context like software, manufacturer and model, vulnerabilities, and ties to threat actors. We use a combination of first and third-party fingerprints (e.g., Rapid7 Recog). We implement static fingerprints through a combination of declarative filters (e.g., `html_title: "WAC6552D-S"`) and processors written in a Lisp-like DSL. In total, we check just over 10K static (i.e., non-ML-based) fingerprints, though we are increasingly moving to supervised models for fingerprinting. The journaled non-ephemeral scan data is combined with the additional derived context to return a cohesive entity record.

5.3 Data Access and Presentation

Many of our commercial users interact with our Internet Map through tailored downstream applications that solve specific operational needs. However, we also expose our global Internet Map for researchers and analysts to conduct investigations. Given the scale of the dataset, no single database engine practically supports all access patterns and we provide users with several interfaces:

Coverage	Censys	Shodan	Fofa	ZoomEye	Netlas
Top 10 Ports	96%	80%	63%	82%	63%
Top 100 Ports	92%	40%	62%	54%	27%
All 65K Ports	82%	10%	43%	26%	3%

Table 1: Coverage of Services in Engines—We show the coverage of each scan engine—broken down by non-overlapping port ranges—over the union of currently active services found in all scan engines.

	Censys	Shodan	Fofa	ZoomEye	Netlas
Self-Reported	794M	810M	3.1B	3.5B	877M
Est. % Accurate	92%	68%	20%	10%	49%
Est. % Unique	100%	100%	65%	99%	63%
Est. # Accurate	730M	550M	403M	346M	270M

Table 2: Coverage of Current IPv4 Services—Other scan engines self-report greater coverage of IPv4 services than Censys. However, after de-duplicating results and filtering out stale data, we find that Censys has fewer false positives in its data (Est. % Accurate) and higher coverage of currently active IPv4 services (Est. # Accurate).

Fast Lookup API. We surface our internal read-side storage API through a REST interface, allowing high-throughput lookups by entity ID and timestamp (e.g., "What did IP `A` look like at time `B`?" and "What IP addresses has certificate `X` been seen on?"). Since the API is backed directly by Bigtable, requests can be serviced in well under 100 ms and at significant scale.

Interactive Search and Exploration. We support interactive queries and full text searches against the current state of each entity through a web interface and REST API powered by Elasticsearch. Queries use a Lucene-like language, can reference any combination of fields, and take from 100 ms up to tens of seconds to execute, depending on their complexity.

Analytics Engine. Elasticsearch is limited in the complexity of queries it can execute. To support arbitrarily complex analyses and temporal queries, we snapshot and store our Internet Map daily to Google BigQuery, a serverless data analytics engine. After three months, we retain only one weekday snapshot per week to minimize costs, but allow longitudinal analysis.

Raw Data Downloads. To support on-premise analysis and model training, we publish daily snapshots in Apache Avro format.

Academic researchers and industry users differ in how they prefer to access data. All but the most sophisticated of our industry customers prefer not to download the full datasets and instead look for APIs and cloud access. In contrast, academic researchers typically prefer downloading the full datasets, even when their questions could be answered more efficiently through Google BigQuery or API.

6 EVALUATION

To evaluate Censys' coverage, accuracy, and timeliness in finding Internet services, we compare Censys against random sub-sampled scans of the IPv4 Internet across all 65K ports as well as other popular scan engines. While prior work has noted a significantly larger number of results from other tools like Shodan [68, 86], we find that self-reported statistics are often inflated with old data and/or duplicate records. After filtering out stale data (i.e., services that are no longer online at the time of analysis), we estimate that Censys serves 33% more currently active services (180M) than the next leading platform and that the data returned by Censys is 35–820% more accurate than other scan engines.

6.1 Methodology

We evaluate Censys' coverage against two datasets: (1) an approximated ground truth of Internet services, captured through random sub-sampled scans of all 65K ports, and (2) the results found in other popular Internet scan engines.

Ground Truth Approximation. Because no ground truth dataset of Internet services exists to compare to, we evaluate Censys against a random sub-sampled scan of 0.1% of the IPv4 service space across all 65K ports that we collect using ZMap (i.e., we independently identify a random subset of all Internet services across all 65K ports using ZMap and quantify what portion of those random services that engines have in their datasets). We identified a total of 4.1 million Internet services by scanning at 1 Gb/s for one week starting August 20, 2024. We filter out 0.2% of those hosts that respond on more than 20 ports with nearly identical "pseudo" services. These hosts distort results: despite accounting for a negligible fraction of total Internet hosts [24], they outnumber legitimate services in 65K port scans, since they respond on every port.

Alternative Scanning Engines. We evaluate Censys against four popular scanning-based search engines: Shodan [106], FofA [44], ZoomEye [125], and Netlas [3]. We cannot directly compare Censys' coverage against external scan engines because other engines do not provide access to raw data nor do they support extracting a random sample of data. As such, we compare coverage by: (1) generating 10K random IP addresses, (2) querying both Censys and external scan engines for the current state of those IPs, (3) filtering out services that are no longer present by conducting follow-up scans of returned services using ZGrab [30] with added support for additional protocols, and (4) comparing coverage between engines.

There are several potential limitations for this methodology. First, because only a small percentage of Internet hosts have public services, this approach provides a limited number of data points for comparison. However, as shown in Appendix C, results quickly converge. Second, while engines scan primarily standardized network protocols, it is possible that our implementations introduce bias compared to other engines' protocol implementations. We do not find evidence of this in spot-checks using other tools like Nmap, and later, we show that accuracy directly correlates with self-reported freshness of data, which would not be affected by ZGrab. Third, the network could affect scan results. To minimize this, our liveness checks are run from a different network than our production scanning to minimize network-related bias.

6.2 Coverage and Accuracy

We first measure the coverage (how much of the Internet each scan engine sees) and accuracy (correctness of data returned by the scan engine compared to a real-time follow up scan at query time) of Censys versus other scan engines. As seen in Table 2, all other scan

engines self-report better coverage than Censys. However, this is more indicative of each platform's storage/presentation strategy than their actual visibility. After filtering out stale and duplicate services, we find a dramatic change: Censys emerges as having the highest coverage of current Internet services.

This occurs for two reasons. First, some engines serve exceptionally stale data. In the most egregious case, only 10% of services returned by ZoomEye were responsive to a follow up scan at the time of our evaluation. Second, some engines double count scan entries for the same IP/port pair. For example, over a third of services returned by Fofa and Netlas are duplicates. Staleness leads to overestimation of *coverage*, but even more so directly affects the *accuracy* of data and whether operators can trust results. While other platforms have 10–68% data accuracy (Table 2), Censys's aggressive pruning of stale data results in 92% data accuracy.

Service Age. To understand data freshness and to ensure that our liveness check does not inject bias, we investigate the "last scanned date" for the services returned by each scan engine. As shown in Figure 2, there are dramatic differences: while 100% of services in Censys were scanned within the past 48 hours, some reported services in ZoomEye were more than three years old. Netlas publicly notes that a single scan over the Internet takes about a month; our data supports this claim [87]. There is perfect rank-order correlation between accuracy and data freshness of search engines.

Port Coverage. Finding services on the most densely populated ports requires a different strategy than finding services across all 65K ports. As seen in Table 1, while other platforms do relatively well at finding services on popular ports compared to Censys, this coverage quickly dissipates. For example, Censys sees only 20% more services on the top 10 ports compared to Shodan, but finds significantly more services on less popular ports.

Considering coverage compared to our sub-sampled 65K port scan, we estimate that Censys sees 98% of IPv4 services on the top 10 ports, 97% of the top 100, and 62% of services across all 65K ports. The ability to identify services across all 65K ports is one of the foremost reasons for Censys' increased visibility over other scanners. While perhaps obvious, we emphasize that finding Internet services follows the "80:20 rule": finding the first 80% of services is the first 20% of the effort.

Protocol Coverage. While port coverage indicates greater Internet visibility, it is not immediately clear how that translates into coverage of important protocols. To evaluate protocol coverage, we analyze the 13 protocols that CISA lists as most critical to Internet exposure risk [26], filtered to three protocols with sufficient data to surface a ranking trend (Appendix C). As shown in Table 3, protocol coverage follows the same pattern as global coverage.

Country Coverage. Scanning platforms are run from different countries: Censys and Shodan are U.S.-based, ZoomEye and Fofa are Chinese, and Netlas is Armenian [68]. It is natural to wonder if engines demonstrate better visibility in certain regions, in particular for Fofa and ZoomEye in China given the Great Firewall and congestion at the national border. Using our sub-sampled 65K port scan, we investigate differences in country coverage. As shown in Table 3, while we see variation in coverage between countries, there is no clear pattern. Surprisingly, Shodan has better coverage in China and Germany than in the U.S. and Censys has better coverage in China than ZoomEye or Fofa.

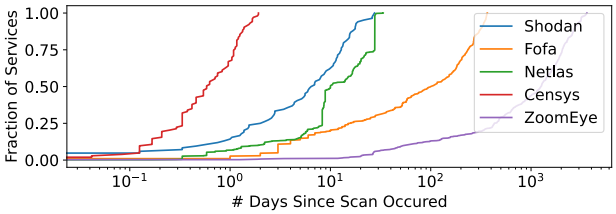

Figure 2: Service Data Freshness—Data freshness varies dramatically between scan engines, from under 48 hours for services in Censys to multiple years old for ZoomEye. Data freshness strongly influences the accuracy of data.

Category	Hosts	Censys	Shodan	ZoomEye	Fofa	Netlas
US	(625)	86%	44%	56%	58%	30%
CN	(100)	93%	52%	70%	45%	52%
DE	(59)	85%	56%	70%	72%	34%
HTTP	(761)	95%	45%	50%	60%	31%
HTTPS	(172)	92%	74%	76%	53%	51%
SSH	(73)	95%	42%	58%	56%	32%

Table 3: Country and Protocol Coverage—Censys' global visibility translates into the highest coverage of both countries and protocols. Notably, the country a scanner is headquartered in does not imply better coverage of the region. Results are based on our 65K port sub-sampled scan of the IPv4 address space.

Coverage Overlap. No scanning engine achieves complete coverage of all of the Internet services seen by other scan engines. This is likely due to differing scan methodologies and scan times. In Figure 3, we plot the coverage of confirmed active hosts that each scanning engine has (X-axis) of other engines (Y-axis). Censys has the greatest coverage compared to other engines. For example, Censys reports 96% of Shodan's accurate services. Censys has the least coverage of Fofa and ZoomEye, seeing only 90% of its responsive services. Censys is also the scanning engine that every other scanning engine has the least coverage of (39–57%), likely due to its coverage of all 65K ports.

6.3 ICS Exposure Case Study

While Censys has the greatest estimated coverage, the Internet service landscape is dominated by HTTP(S) services, many of which are not security critical. To concretely show how differences in Internet visibility impact a security use case, we investigate coverage of Internet-exposed Industrial Control Systems (ICS), an area of significant concern for governments after Russia- and Iran-affiliated attacks against critical infrastructure [27]. This second experiment also enables a slightly different methodology when comparing scan engines. In the prior section, we could not fully enumerate all Internet services from each engine due to API/pricing limitations. Thus, we compared a random set of IP addresses that we queried from engines. In contrast, because there are a small number of total control systems exposed on the Internet, we can nearly comprehensively

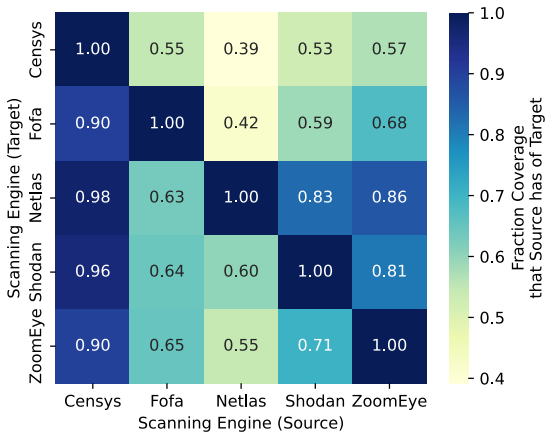

Figure 3: Scan Engine Coverage Overlap—Each engine finds a unique subset of Internet services, likely due to differing methodologies, vantage points, and scan times. To read the heatmap, (1) choose a scanning engine, *A*, on the x-axis, (2) the column will show *A*'s coverage of other engines. Censys has the greatest unique coverage of Internet services compared to alternatives.

Protocol	Censys		Shodan		ZoomEye		Fofa	
	Acc.	Rep.	Acc.	Rep.	Acc.	Rep.	Acc.	Rep.
ATG	**8.4K**	10K	2.9K	299K	–	–	4.4K	28K
BACNET	**13.1K**	14K	11.4K	46K	11.3K	76K	11.4K	20K
CIMON	**1K**	1.2K	–	–	–	–	–	–
CMORE	**2.3K**	2.4K	–	–	–	–	–	–
CODESYS	2.5K	2.7K	**2.6K**	216K	–	–	2.1K	4.8K
DIGI	**7.5K**	21K	–	–	–	–	–	–
DNP3	**1.2K**	1.4K	860	1.5K	506	2.4K	440	1.6K
EIP	**7.5K**	10.2K	4.7K	255K	–	-	-	–
FINS	**1.8K**	2.3K	891	1.4K	959	14.6K	-	–
FOX	**20K**	21.6K	5K	7.4K	–	–	17.7K	–
GE SRTP	**49**	54	45	56	39	8K	–	–
HART	**12**	12	8	9	1	75	1	3
IEC-60870	**6.9K**	8.2K	2.2K	2.9K	0	0	5.2K	10K
MODBUS	**42K**	42K	22.6K	36.5K	*4.1K	*30K	27.2K	91.5K
OPC UA	**2.4K**	3.1K	383	630	–	–	–	–
PCWORX	**228**	432	157	692	–	–	199	1.47K
ProConOs	**715**	809	235	295	556	1.4K	30	52
REDLION	**1K**	1.1K	1K	1.1K	931	16.1K	938	1.7K
S7	**6.5K**	8.2K	4.6K	7.2K	5K	32K	4.8K	12.2K
WDBRPC	**16K**	30K	2K	31.1K	2.7K	114K	2.5K	15.5K

Table 4: ICS Coverage—We show the *Self-Reported* and *Validated* number of services running common ICS protocols. Improved global visibility has direct implications for security. Netlas reports results for only S7: 5K services, of which 4K are accurate. * Results capped.

query the set of exposed services for each protocol from every scan engine. Experiments ran in August 2024; we provide queries for each engine in Appendix Table 8.

For all cases except for Modbus, we are able to iterate over the full set of results. However, because of stale data for ZoomEye, we stopped fetching results after 30K results for Modbus, the amount left over in our monthly download budget. We note that ZoomEye sorts services in descending order of update time. Our coverage allows us to download all Modbus services scanned in the past 22 days by ZoomEye. As seen in Table 4, Censys' increased visibility translates into better ICS coverage for all but one protocol, CODESYS. This is, in part, driven by many control systems using non-standard ports [60]. However, it is also likely affected by the fresher data from Censys since many control systems are connected through LTE or 5G networks that experience DHCP churn.

As seen in Table 4, some reported numbers dramatically differ from validated results. We find that alternative platforms often overreport ICS services due to poor labeling. For example, Shodan overreports coverage for ATG, CODESYS, EIP, and WDBRPC by multiple orders of magnitude because it does not explicitly verify the protocol handshake. For example, Shodan identifies CODESYS devices by searching for services on port 2455 that return the keywords "operating" and "system," criteria met by hundreds of thousands of HTTP services rather than services running CODESYS [105]. In contrast, Censys will only label a service as running a protocol if it is able to complete an L7 handshake for the protocol.

Port/Protocol	Censys (Hours)		Shodan (Hours)	
	Mean	Median	Mean	Median
80/HTTP	2.37	2.40	68.17	51.10
443/HTTPS	2.41	2.40	77.12	63.35
161/SNMP	6.27	4.67	65.99	49.27
3389/RDP	7.47	5.73	97.76	78.51
21/FTP	7.82	6.87	77.96	53.81
2082/HTTP	7.92	5.90	67.91	49.95
3306/MYSQL	7.96	6.62	87.63	76.68
2222/SSH	8.03	6.45	59.49	49.30
23/TELNET	8.20	7.27	88.18	75.28
5060/SIP	8.36	7.02	79.12	57.88
7547/HTTP	8.64	6.93	79.50	60.53
60000/HTTP	7.29	5.73	–	–
500/HTTP	75.54	65.92	–	–

Table 5: Time To Discovery—On average, Censys finds honeypots in 12.3 hours (5.7h median), while Shodan finds the same hosts in average 76.5 hours (60.9h median). Shodan did not find any services on 500/HTTP or 60000/HTTP.

6.4 Time To Service Discovery

Censys aims to not only uncover all Internet services, but to quickly find them as they come online. Here, we compare time to discovery for the two engines with clear scanner identification: Censys and Shodan. We note that after Censys, Shodan provided the freshest data (Figure 2), making it a meaningful comparison.

Methodology. To measure the time to discover services, we need to know when those services first came online. We deployed 100 honeypots between September 19–27, 2024 on Google Cloud Compute, staggering their creation every eight hours to account for variations in scan times. The honeypots run T-Pot's Python honeypot and listen for connections on 12 ports associated with common protocols (Table 5). We measure time until we see connections from the two identifiable engines: Censys and Shodan.

Results. As seen in Table 5, Censys identifies hosts in about 12.3 hours (5.7 hour median), while Shodan finds the same hosts in 76.5 hours on average (60.9 hour median). In addition to faster discovery, Censys achieves more comprehensive coverage: within 24 hours, Censys identifies 100% of services for six of the most popular protocols and achieves 100% coverage across all protocols within one week.

7 RESEARCH AND OPERATIONAL USAGE

While we originally built Censys for research [32], operators quickly outnumbered academic researchers on the platform. Industry interest, combined with the need for significant investment in the scan engine (e.g., implementing hundreds of protocols) and increasing operational costs, led us to spin out from the University of Michigan. Since 2018, Censys, Inc.'s goal has been to provide industry and government with the Internet visibility they need, while simultaneously using the commercial business to subsidize access for academic researchers. Here, we describe research that has used Censys data, our experiences running a research access program, and the operational use cases that Censys supports.

7.1 Research and Publication

Censys has been used in over 500 research papers as of January 2025. We remain committed to providing researchers with free access, but, as Censys has grown, our approach has also evolved.

Published Research. Researchers have used Censys to investigate topics ranging from mercenary spyware abuse [76] and hypergiant footprints [48] to recovering RSA private keys [113] and improving Cuckoo filters [119]. To understand research usage, one author downloaded and analyzed papers that cited the original Censys paper [32] or mentioned Censys using Google Scholar, identified the subset that used Censys data (as opposed to merely referencing its existence), and thematically coded [18] topics. We identified 509 papers: 358 journal/conference proceedings, 66 technical reports, and 85 theses.

Censys-based research has appeared at 160 venues, most commonly ACM Internet Measurement Conference (IMC, 35 papers), followed by top-tier security venues like USENIX Security (27), CCS (21), and NDSS (17). As can be seen in Appendix Table 6, papers frequently focus on security-relevant topics like the WebPKI (57 papers), IoT security (50), malware and attacker behavior (44), HTTPS and TLS (37), and critical infrastructure (37). While many papers cover expected topics like Internet exposure, we have been thrilled to see creative, unpredicted uses like measuring censorship [94] and reverse engineering NSO Group operations [12].

Beyond peer-reviewed research, close to 100 undergraduate and masters theses have used Censys as of July 2025, possibly hinting at how accessible data can be used as part of other educational activities. Several security courses use Censys data, including at Georgia Tech, NC State, UCLA, and Stanford. We are excited to partner with instructors to understand how Censys can be used to illustrate the Internet. We have also seen the data used to simply stress test new systems [6, 91, 119], highlighting a broader need for realistic large-scale datasets within the computer science community.

Research Access Program. When we first launched Censys, we provided open access to all of our data for non-commercial use. Unfortunately, access was abused by companies that incorporated the data into paid products and services. We also worried about malicious actors using the data to exploit vulnerable Internet systems. In response, we established an application-based research access program similar to VirusTotal and Farsight. This allowed us to restrict access to commercial users and enabled due diligence around research access. From 2018 to 2024, we processed 959 requests and provided access to 1,221 researchers. As of November 2024, 433 researchers from 239 organizations have continued access.

Equitably operating a research program is more challenging than we anticipated. While it is easy to verify the identity of well-established researchers with a Google Scholar profile or presentations at conferences like Blackhat or BSides, these constitute only a small fraction of requests. Most requests come from independent researchers and students who have no public reputation. To equitably provide access, we established evaluation criteria based on: (1) proposing a clear research plan, (2) intending to publicly disseminate results, and (3) confirming that work is conducted independently or as part of a non-profit or academic institution. These applications are then reviewed by our internal research team.

This strategy has enabled access for early-career researchers, but the system is imperfect. Many students lack coherent research plans and without significant back-and-forth, it is difficult to discern between poorly written requests, requests from first time researchers exploring, and fabricated plans. We struggle to process many international requests because of language barriers and mounting evidence that universities are being used to proxy offensive government operations in some countries [97], turning research access decisions political. Recently, we have also seen malicious actors use the research program to identify vulnerable systems. To minimize potential harm in these ambiguous situations, we established multiple access tiers that provide delayed access or access to a subset of data (e.g., excluding access to CVE or ICS data).

Much to our surprise, it is not uncommon for researchers to send vitriolic messages, accusations, and, in rare cases, threats. While we know these interactions are not representative of the community and we remain committed to providing data to researchers, these messages can quickly turn program administration into a thankless job, similar to the experiences expressed by open source maintainers [69, 82, 93]. Access requests are mediated by our research team and in situations when we deny access, this is due to us being unable to justify the potential for abuse over the potential for a research outcome for the presented case, not because of conflicting commercial interest. Not specific to Censys, we urge researchers to remember the humans behind research programs. Even at a company founded by researchers, there are always tradeoffs in investment and the ability to show positive impact is critical.

7.2 Industry and Government Usage

Many people familiar with Censys since it was an academic project equate the company with Internet scanning and our search interface. Some of our largest customers continue to directly ingest the Internet data that we collect. However, for many of our customers, data must be transformed into actionable insights to be practically useful. In these situations, the Internet Map is the foundation on which to build customer-facing products rather than the product

itself. Below, we describe the top use cases that drive commercial usage of Censys. While none of these use cases were planned when we first built Censys, they have subsequently shaped how we collect data. We also hope that by discussing how scanning is used operationally, we can inform translatable future approaches.

Attack Surface Management. Internet-facing infrastructure is the primary target for initial access [28, 116]: ransomware actors and other adversaries exploit Internet-facing software (e.g., MOVEit Managed File Transfer), compromise edge networking equipment (e.g., Fortinet firewalls), and log into Internet-facing VPNs using compromised credentials to gain an initial foothold into companies. Organizations use Censys to comprehensively discover, monitor, and remediate Internet vulnerabilities and misconfigurations on their external perimeter. While tracking Internet exposure might seem like a relatively easy task, we have found this to be shockingly difficult for large companies who have sprawling Internet footprints. It is not uncommon for customers to have dozens of cloud providers, thousands of cloud projects, and hundreds of thousands of Internet-exposed assets. For these companies, it can be difficult to know when new assets appear, and, when they do, to quickly understand who owns them and how to remediate any vulnerabilities.

Supply Chain Intelligence and Cyberinsurance. Similar to monitoring one's own attack surfaces, organizations additionally monitor the external attack surfaces of their supply chain dependencies. Insurance providers and re-insurers use scan data to quantify the risk associated with potential insurees and to price premiums for customers. We observe two use cases here: (1) scoring the risk profile of suppliers, and (2) using the data to actively help suppliers protect themselves as part of a company's own self-defense.

Critical Infrastructure Monitoring. Critical infrastructure has increasingly come under attack; Iran-affiliated and pro-Russia actors have compromised and in some cases manipulated industrial control systems (ICS) in the food and agriculture, healthcare, and water and wastewater sectors [88]. In response, governments are increasingly leveraging Internet scan data to monitor the attack surfaces of critical infrastructure sectors. This use case is similar to Attack Surface Management, but operates from a reverse perspective. Instead of mapping out the infrastructure that belongs to a company and then identifying security weaknesses, governments will map out classes of vulnerabilities and then identify the organizations that need help remediating. In one example that we are particularly proud of, in October 2024, Censys identified SCADA user interfaces (HMIs) for water distribution networks belonging to 268 U.S. towns and cities that allowed unauthenticated manipulation; the U.S. Environmental Protection Agency (EPA), with assistance from state water administrators, worked with utilities to remove over 97% of these HMIs from the Internet [20]. This project led to a partnership with the EPA to continuously monitor for these types of exposures.

Threat Hunting. Another top use case for Censys is threat hunting: investigating adversary-controlled Internet infrastructure (e.g., C2 servers). In practice, this includes identifying malicious servers through specific scanners (e.g., Cobalt Strike), mapping out relationships between servers (e.g., via SSH hostkey or JARM fingerprint), and finding patterns among compromised devices (e.g., SOHO routers) used for launching attacks. While there exists a major threat intelligence industry, much of the data provided by vendors is retroactive, collected after an incident occurs; in contrast, scanning can sometimes find malicious infrastructure prior to its operationalization or immediately after-the-fact during an incident response. We refer readers to the following reports for concrete threat hunting examples: [11, 40–42, 45, 71, 75–77]. Data about threat actors is utilized by most major threat intelligence vendors (e.g., Crowdstrike) as part of their reporting as well as directly by governments and sophisticated corporations (e.g., financial institutions) to block malicious infrastructure and respond to incidents.

Fraud and Impersonation. Companies use web and certificate data to identify and take down websites that impersonate their brands to prevent financial fraud, phishing, and credential harvesting (e.g., by searching for websites with similar domains or favicons, and/or page structure). Companies also use these data points to identify domain hijacking.

8 ETHICAL CONSIDERATIONS

The ethics of Internet scanning and publishing scan data are not universally agreed upon and continue to evolve [35, 43, 53, 79, 99]. We have also found that the academic and operational communities have different ethical norms. Our own mission and views have changed as we work more closely with industry and government partners to secure Internet infrastructure and interrupt adversary operations. For example, Wu et al. previously argued that Censys operates unethically by collecting data beyond service presence due to potential privacy concerns [122]. However, collecting only service presence precludes determining exposure and ownership, which is critical for remediating vulnerable critical infrastructure.

We broadly follow Durumeric et al.'s 2024 best practices for scanning [35]. We never attempt to exploit vulnerabilities, bypass authentication, or access devices behind NAT. Our data collection is compliant with U.S. law and GDPR regulation. The IPs we use have identifying rDNS, WHOIS, and an HTTP site that indicates ownership, intent, and contact information unless a specific reason prompts otherwise (e.g., tracking a threat actor as part of a takedown operation). When possible, scan probes identify Censys (e.g., HTTP User Agent). We emphasize the importance of testing new probes, slowly ramping up scans, and enabling operators to easily contact the research team. Our one notable incident that adversely affected a large number of devices was due to sending a protocol compliant SVR probe that inadvertently triggered a bug in some Juniper routers. In this situation, we were able to work with Juniper to halt scanning, test a new patch, allow them to patch devices, and slowly resume scanning in coordination with them.

Censys sends around 26.5M probes/second, which results in a public IP seeing a probe every 2.5 minutes. This is a non-negligible amount of traffic, but is relatively small compared to total scan traffic. Greynoise honeypots see on average 20 probes per minute [51]: Censys accounts for 1–2% of the total scan traffic a cloud host sees. We argue that that this is not out of balance compared to its widespread usage across the security industry. That said, we also argue that future work is needed to reduce the amount of scan traffic sent

to achieve real-time Internet visibility. Reducing bandwidth usage for service discovery is a goal.

We honor opt-out requests from operators who can verify network or domain ownership through public WHOIS data (Appendix D.1); we expire exclusion requests after one year. We note that the blocking and opt-out norms have evolved significantly from when ZMap was released. In 2014, Durumeric et al. noted that they had received 208 blocklist requests and excluded 0.15% of the public IPv4 address space [36]. In comparison, Censys has excluded IP ranges from 39 organizations (0.03% of the address space), despite scanning significantly more aggressively. Only four organizations have requested to exclude a /16 or larger prefix: CalTech, Carnegie Mellon University, Indiana University, and KU Leuven.

A less commonly discussed but increasingly critical ethical question concerns what Internet data we should expose and to whom. While the data we collect is publicly accessible to everyone, clearer visibility into Internet infrastructure is an innately double-edged sword. Visibility is crucial for practitioners and can draw attention to a growing problem, but a global index can also be used to identify targets. While there was little evidence Censys was used maliciously in its first few years, the attack landscape has since shifted and we have begun to scan for more security-sensitive protocols. Internet-facing infrastructure (e.g., RDP, VPNs, enterprise applications) is now the top initial attack vector and adversaries have begun to use search engines to identify vulnerable services [62]. In addition, attacks against Internet-facing devices are increasingly translating into physical harm, ranging from damaging critical infrastructure [83] and ransoming hospitals [81] to targeting missile strikes in Ukraine, Russia, and Israel [2, 59].

Today, we are more conservative in what data we expose to our users by default. We restrict access to data that is ripe for abuse, most notably control system, vulnerability, and adversarial infrastructure data (e.g., C2 servers) unless there is a specific need. We also limit specific types of searches against our data until we can verify user identity and goals. We appreciate that these restrictions create friction for researchers and may altogether exclude some researchers just starting in the field, but we lack better mechanisms for preventing increasing abuse of the platform.

9 DISCUSSION AND CONCLUSION

Censys has changed tremendously since the project launched in 2015. We hope that by documenting Censys' design goals, evolution, and current behavior in the literature, we can help researchers to accurately use Censys data. While much of our evolution discussed in this work has been driven by industry use cases over the past few years, we have been thrilled to see how researchers have applied Censys data in unanticipated but valuable ways. We also hope that any lessons we have learned can help future research transition to practice. Below, we present several areas of potential future research where we lack strong understanding or approaches within the networking community today:

Understanding Internet Dynamics. The more we have directly studied the real-world behavior of Internet services, the more we have learned about how to best conduct Internet scans. There remains much work to understand the ephemerality of Internet services, how and why Internet services appear in unexpected places, and how the Internet is balkanizing. There has also been little work investigating the behavior of UDP-based services on the Internet. A more nuanced understanding of the Internet also helps to ensure that otherwise seemingly design choices do not have unexpected consequences or introduce biases.

Intelligent Scanning. Comprehensively scanning the Internet requires efficiently finding services across all 65K ports. Several works have proposed predictive approaches [60, 61, 72, 103, 107, 108]. While these are promising first works in this space, intelligent 65K port scanning remains an unsolved problem. First, predictive approaches do not find most services. For example, GPS [61] found a largely non-overlapping set of services than Censys in 2021, exhibiting that new methods are not yet replacements for comprehensive scanning. Second, many proposed approaches do not scale across all 65K ports [107, 108] and require weeks of training time [103]. Third, predictive systems are designed to run once, requiring massive amounts of training data, rather than operating continuously over months, a fundamentally different problem [72].

Passive Fingerprinting. While internal vulnerability scanners and attackers run unsafe code to check for vulnerabilities, Internet scanners cannot. Safely checking for vulnerabilities and extracting software versions is critical for understanding Internet behavior, prioritizing responses, and conducting vulnerability notifications. However, building safe scanners is highly manual, sometimes impossible, and often under time pressure in response to a vulnerability notification. There are many research opportunities to understand how to best fingerprint software versions, build scanners to safely detect vulnerabilities and misconfigurations, and to automatically build safe scanners for software and malware.

Identifying Relationships. Threat Hunting remains a largely manual task where experienced analysts search for erroneous configurations that could be used to fingerprint infrastructure and try to identify patterns that could be used to link related assets. Industry standards like JARM and JA4+ are patented, static, and brittle. Similarly, uncovering the organizational owners of systems remains difficult to complete accurately and at scale. There are many opportunities for building better automated techniques for fingerprinting and uncovering relationships between Internet assets.

Device and Service Tracking. While we use identifiers like IP address as a unique handle for Internet entities, real-world entities often change IP addresses. We lack validated options for tracking devices and services as they change IP address or for grouping services together that belong to the same underlying application.

Effective Notifications. A significant body of work has investigated how to conduct vulnerability notifications [21, 38, 67, 73, 74, 84, 111, 112, 114]. These works have had statistically significant but minimal impact; our own direct notification attempts have been similar. However, in a recent partnership with the US Environmental Protection Agency (EPA), we were able to achieve near 100% remediation across hundreds of critical infrastructure providers. This may in part be because of the EPA's enforcement authority. However, in some cases, this required agencies to travel on-site to find and explain the problem to the appropriate personnel, highlighting the lack of dependable communication channels. As researchers and industry alike regularly surface information

that can be used to prevent attacks, we urge both the research community and trusted entities like the U.S. Cybersecurity and Infrastructure Security Agency (CISA) to establish better mechanisms for disseminating vulnerability notifications to organizations.

ACKNOWLEDGMENTS

We thank the many current and past staff at Censys, who have helped to make the project a success. We are particularly grateful to the operations team who ensures that Censys runs smoothly. The authors also thank Himaja Motheram, Emily Austin, Mark Ellzey, Aidan Holland, Ryan Kokoszka, David Adrian, Paul Parkanzky, J. Alex Halderman, Michael Bailey, Kimberly Ruth, Thea Rossman, and the broader networking and security community for their feedback and support.

REFERENCES

[1]
[2] Israeli officials say Iran exploiting security cameras to guide missile strikes, author=Antoniuk, Daryna. https://therecord.media/iran-espionage-israeli-security-cameras-missile-attacks.
[3] Netlas - Search the Internet Your Way. Accessed: 2024-11-25.
[4] A Survey on Network Attack Surface Mapping, author=Everson, Douglas and Cheng, Long. *Digital Threats: Research and Practice*, 2024.
[5] J. Aas, R. Barnes, B. Case, Z. Durumeric, P. Eckersley, A. Flores-López, J. A. Halderman, J. Hoffman-Andrews, J. Kasten, E. Rescorla, et al. Let's Encrypt: An Automated Certificate Authority to Encrypt the Entire Web. In *ACM SIGSAC Conference on Computer and Communications Security*, 2019.
[6] F. Abuzaid, P. Kraft, S. Suri, E. Gan, E. Xu, A. Shenoy, A. Ananthanarayan, J. Sheu, E. Meijer, X. Wu, et al. Diff: A Relational Interface for Large-Scale Data Explanation. *Proceedings of the VLDB Endowment*, 2018.
[7] D. Adrian, Z. Durumeric, G. Singh, and J. A. Halderman. Zippier ZMap: Internet-Wide Scanning at 10 Gbps. In *USENIX Workshop on Offensive Technologies*, 2014.
[8] M. Allman, V. Paxson, and J. Terrell. A Brief History of Scanning. In *ACM Internet Measurement Conference*, 2007.
[9] J. Althouse. JA4+ network fingerprinting. https://blog.foxio.io/ja4+-network-fingerprinting.
[10] Amazon AWS. CQRS pattern. https://docs.aws.amazon.com/prescriptive-guidance/latest/modernization-data-persistence/cqrs-pattern.html.
[11] Amnesty International. German-made FinSpy spyware found in Egypt, and Mac and Linux versions revealed. https://www.amnesty.org/en/latest/research/2020/09/german-made-finspy-spyware-found-in-egypt-and-mac-and-linux-versions-revealed/.
[12] Amnesty International. Morocco: Human Rights Defenders Targeted with NSO Group's Spyware. https://www.amnesty.org/en/latest/research/2019/10/morocco-human-rights-defenders-targeted-with-nso-groups-spyware/.
[13] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, et al. Understanding the Mirai Botnet. In *USENIX Security Symposium*, 2017.
[14] AWS. ip-ranges.json. https://docs.aws.amazon.com/vpc/latest/userguide/aws-ip-ranges.html, 2024.
[15] S. Bano, P. Richter, M. Javed, S. Sundaresan, Z. Durumeric, S. J. Murdoch, R. Mortier, and V. Paxson. Scanning the Internet for Liveness. *ACM SIGCOMM Computer Communication Review*, 2018.
[16] C. Bennett, A. Abdou, and P. C. van Oorschot. Empirical Scanning Analysis of Censys and Shodan. In *Workshop on Measurements, Attacks, and Defenses for the Web*, 2021.
[17] R. Beverly, R. Durairajan, D. Plonka, and J. P. Rohrer. In the ip of the beholder: Strategies for active IPv6 topology discovery. In *ACM Internet Measurement Conference*, 2018.
[18] V. Braun and V. Clarke. *Thematic Analysis*. American Psychological Association, 2012.
[19] J. Cable, D. Gregory, L. Izhikevich, and Z. Durumeric. Stratosphere: Finding Vulnerable Cloud Storage Buckets. In *Symposium on Research in Attacks, Intrusions and Defenses*, 2021.
[20] Censys. Turning Off the (Information) Flow: Working With the EPA to Secure Hundreds of Exposed Water HMIs. https://censys.com/blog/turning-off-the-information-flow-working-with-the-epa-to-secure-hundreds-of-exposed-water-hmis.
[21] O. Cetin, C. Ganan, M. Korczynski, and M. Van Eeten. Make Notifications Great Again: Learning How to Notify in the Age of Large-Scale Vulnerability Scanning. In *Workshop on the Economics of Information Security*, 2017.
[22] Y. Chen, X. Lian, D. Yu, S. Lv, S. Hao, and Y. Ma. Exploring Shodan from the Perspective of Industrial Control Systems. *IEEE Access*, 2020.
[23] K. C. Claffy and D. Clark. The 11th Workshop on Active Internet Measurements Workshop Report. *ACM SIGCOMM Computer Communication Review*, 2019.
[24] H. Clark. Beyond Noise: Filtering Pseudo-Services, 2024. Accessed: 2024-12-17.
[25] Coalition. BinaryEdge Transition FAQ. https://help.coalitioninc.com/hc/en-us/articles/34383910057371-BinaryEdge-Transition-FAQ.
[26] Cybersecurity and I. S. A. (CISA). BOD 23-02 Implementation Guidance: Mitigating the Risk of Internet-Exposed Management Interfaces, 2023.
[27] Cybersecurity and Infrastructure Security Agency (CISA). Industrial Control Systems (ICS). https://www.cisa.gov/topics/industrial-control-systems.
[28] Cyentia Institute. Information Risk Insights Study: Ransomware. https://www.cyentia.com/wp-content/uploads/2024/08/IRIS_Ransomware.pdf.
[29] A. Drichel, V. Drury, J. von Brandt, and U. Meyer. Finding Phish in a Haystack: A pipeline for Phishing Classification on Certificate Transparency Logs. In *International Conference on Availability, Reliability and Security*, 2021.
[30] Z. Durumeric and D. Adrian. Zgrab2, 2018. https://github.com/zmap/zgrab2.
[31] Z. Durumeric, D. Adrian, M. Bailey, and J. Halderman. Heartbleed bug health report. https://zmap.io/heartbleed.
[32] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, and J. A. Halderman. A Search Engine Backed by Internet-wide Scanning. In *ACM SIGSAC Conference on computer and communications security*, 2015.
[33] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, and J. A. Halderman. Tracking the FREAK attack, 2015. https://freakattack.com/.
[34] Z. Durumeric, D. Adrian, A. Mirian, J. Kasten, E. Bursztein, N. Lidzborski, K. Thomas, V. Eranti, M. Bailey, and J. A. Halderman. Neither Snow Nor Rain Nor MITM... An Empirical Analysis of Email Delivery Security. In *ACM Internet Measurement Conference*, 2015.
[35] Z. Durumeric, D. Adrian, P. Stephens, E. Wustrow, and J. A. Halderman. Ten Years of ZMap. In *ACM Internet Measurement Conference*, 2024.
[36] Z. Durumeric, M. Bailey, and J. A. Halderman. An Internet-Wide view of Internet-Wide scanning. In *23rd USENIX Security Symposium*, 2014.
[37] Z. Durumeric, J. Kasten, M. Bailey, and J. A. Halderman. Analysis of the HTTPS Certificate Ecosystem. In *ACM Internet Measurement Conference*, 2013.
[38] Z. Durumeric, F. Li, J. Kasten, J. Amann, J. Beekman, M. Payer, N. Weaver, D. Adrian, V. Paxson, M. Bailey, et al. The Matter of Heartbleed. In *ACM Internet measurement conference*, 2014.
[39] Z. Durumeric, E. Wustrow, and J. A. Halderman. ZMap: Fast Internet-wide Scanning and its Security Applications. In *22nd USENIX Security Symposium*, 2013.
[40] Embee Research. A Beginner's Guide to Hunting Malicious Open Directories. https://censys.com/a-beginners-guide-to-hunting-open-directories/.
[41] Embee Research. A Beginner's Guide to Tracking Malware Infrastructure. https://censys.com/a-beginners-guide-to-tracking-malware-infrastructure/.
[42] Embee Research. Combining Pivot Points to Identify Malware Infrastructure - Redline, Smokeloader and Cobalt Strike. https://www.embeeresearch.io/combining-pivot-points-to-identify-malware-infrastructure-redline-smokeloader-and-cobalt-strike/.
[43] T. Fiebig. Crisis, Ethics, Reliability & a measurement. network: Reflections on Active Network Measurements in Academia. In *Applied Networking Research Workshop*, 2023.
[44] FOFA. Fofa search engine. https://en.fofa.info/, 2024.
[45] fopwn. Hunting VIPER C2 Instances on Censys. https://community.censys.com/threat-hunting-38/hunting-viper-c2-instances-on-censys-140.
[46] P. Foremski, D. Plonka, and A. Berger. Entropy/IP: Uncovering Structure in IPv6 addresses. In *ACM Internet Measurement Conference*, 2016.
[47] O. Gasser, Q. Scheitle, P. Foremski, Q. Lone, M. Korczyński, S. D. Strowes, L. Hendriks, and G. Carle. Clusters in the Expanse: Understanding and Unbiasing IPv6 Hitlists. In *ACM Internet Measurement Conference*, 2018.
[48] P. Gigis, M. Calder, L. Manassakis, G. Nomikos, V. Kotronis, X. Dimitropoulos, E. Katz-Bassett, and G. Smaragdakis. Seven years in the Life of Hypergiants' Off-nets. In *ACM SIGCOMM Conference*, 2021.
[49] Google. Protocol buffers. https://protobuf.dev/.
[50] R. Graham. Masscan: Mass IP port scanner, 2014. https://github.com/robertdavidgraham/masscan.
[51] Greyniose. A week in the life of a GreyNoise Sensor: It's all about the tags. https://www.greynoise.io/blog/a-week-in-the-life-of-a-greynoise-sensor-its-all-about-the-tags, 2023.
[52] H. Griffioen, G. Koursiounis, G. Smaragdakis, and C. Doerr. Have you SYN me? Characterizing Ten Years of Internet Scanning. In *Proceedings of the 2024 ACM on Internet Measurement Conference*, pages 149–164, 2024.
[53] F. Hantke, S. Roth, R. Mrowczynski, C. Utz, and B. Stock. Where are the Red Lines? Towards Ethical Server-Side Scans in Security and Privacy Research. In *IEEE Symposium on Security and Privacy*, 2024.
[54] J. Heidemann, Y. Pradkin, R. Govindan, C. Papadopoulos, G. Bartlett, and J. Bannister. Census and Survey of the Visible Internet. In *ACM Internet Measurement Conference*, 2008.

[55] N. Heninger, Z. Durumeric, E. Wustrow, and J. A. Halderman. Mining your Ps and Qs: Detection of Widespread Weak Keys in Network Devices. In *USENIX Security Symposium*, 2012.

[56] R. Hiesgen, M. Nawrocki, A. King, A. Dainotti, T. C. Schmidt, and M. Wählisch. Spoki: Unveiling a New Wave of Scanners through a Reactive Network Telescope. In *USENIX Security Symposium*, 2022.

[57] B. Hou, Z. Cai, K. Wu, J. Su, and Y. Xiong. 6Hit: A Reinforcement Learning-Based Approach to Target Generation for Internet-wide IPv6 Scanning. In *IEEE INFOCOM*, 2021.

[58] G. W. M. E. A. Hsu, S. Bhat, and A. B. F. L. P. Pearce. 6sense: Internet-wide IPv6 scanning and its security applications. 2024.

[59] IPVM. Russian Military Used Hacked Cameras in Missile Strike on Capital, Alleges Ukraine. https://ipvm.com/reports/russia-hacked-hikvision.

[60] L. Izhikevich, R. Teixeira, and Z. Durumeric. LZR: Identifying Unexpected Internet Services. In *USENIX Security Symposium*, 2021.

[61] L. Izhikevich, R. Teixeira, and Z. Durumeric. Predicting IPv4 Services Across All Ports. In *ACM SIGCOMM*, 2022.

[62] L. Izhikevich, M. Tran, M. Kallitsis, A. Fass, and Z. Durumeric. Cloud watching: Understanding attacks against cloud-hosted services. In *ACM Internet Measurement Conference*, 2023.

[63] Y. Jiang, Y. Li, and Y. Sun. Networked Device Identification: A Survey. In *International Conference on Data Science in Cyberspace*, 2021.

[64] J. Klick, S. Lau, M. Wählisch, and V. Roth. Towards Better Internet Citizenship: Reducing the Footprint of Internet-wide Scans by Topology Aware Prefix Selection. In *ACM Internet Measurement Conference*, 2016.

[65] D. Kumar, Z. Wang, M. Hyder, J. Dickinson, G. Beck, D. Adrian, J. Mason, Z. Durumeric, J. A. Halderman, and M. Bailey. Tracking Certificate Misissuance in The Wild. In *IEEE Symposium on Security and Privacy*, 2018.

[66] B. Laurie. Certificate Transparency. *Communications of the ACM*, 2014.

[67] F. Li, Z. Durumeric, J. Czyz, M. Karami, M. Bailey, D. McCoy, S. Savage, and V. Paxson. You've Got Vulnerability: Exploring Effective Vulnerability Notifications. In *USENIX Security Symposium*, 2016.

[68] R. Li, M. Shen, H. Yu, C. Li, P. Duan, and L. Zhu. A Survey on Cyberspace Search Engines. In *Cyber Security: 17th China Annual Conference*, 2020.

[69] J. Linåker, G. Link, and K. Lumbard. Sustaining Maintenance Labor for Healthy Open Source Software Projects through Human Infrastructure: A Maintainer Perspective. In *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, 2024.

[70] Z. Liu, Y. Xiong, X. Liu, W. Xie, and P. Zhu. 6Tree: Efficient Dynamic Discovery of Active Addresses in the IPv6 Address Space. *Computer Networks*, 2019.

[71] Lumen Black Lotus Labs. Derailing the Raptor Train. https://assets.lumen.com/is/content/Lumen/raptor-train-handbook-copy.

[72] Y. Luo, C. Li, Z. Wang, and J. Yang. IPREDS: Efficient Prediction System for Internet-wide Port and Service Scanning. *CoNEXT*, 2024.

[73] M. Maaß, H. Pridöhl, D. Herrmann, and M. Hollick. Best Practices for Notification Studiesfor Security and Privacy Issues on the Internet. In *International Conference on Availability, Reliability and Security*, 2021.

[74] M. Maass, A. Stöver, H. Pridöhl, S. Bretthauer, D. Herrmann, M. Hollick, and I. Spiecker. Effective Notification Campaigns on the Web: A Matter of Trust, Framing, and Support. In *USENIX Security Symposium*, 2021.

[75] B. Marczak and J. Scott-Railton. The Million Dollar Dissident: NSO Group's iPhone Zero-Days Used Against a UAE Human Rights Defender. *Citizen Lab*, 2016.

[76] B. Marczak, J. Scott-Railton, K. Berdan, B. Abdul Razzak, and R. Deibert. Hooking Candiru: Another Mercenary Spyware Vendor Comes into Focus, 2021. https://citizenlab.ca/2021/07/hooking-candiru-another-mercenary-spyware-vendor-comes-into-focus/.

[77] B. Marczak, J. Scott-Railton, B. A. Razzak, N. Aljizawi, S. Anstis, K. Berdan, and R. Deibert. Pegasus vs. Predator Dissident's Doubly-Infected iPhone Reveals Cytrox Mercenary Spyware. https://citizenlab.ca/2021/12/pegasus-vs-predator-dissidents-doubly-infected-iphone-reveals-cytrox-mercenary-spyware/.

[78] J. Matherly. Complete guide to Shodan. 2015.

[79] J. Mazel, R. Fontugne, and K. Fukuda. Profiling Internet Scanners: Spatiotemporal Structures and Measurement Ethics. In *Network traffic measurement and analysis conference*. IEEE, 2017.

[80] A. McDonald, M. Bernhard, L. Valenta, B. VanderSloot, W. Scott, N. Sullivan, J. A. Halderman, and R. Ensafi. 403 Forbidden: A Global View of CDN Geoblocking. In *ACM Internet Measurement Conference*, 2018.

[81] C. C. McGlave, H. Neprash, and S. Nikpay. Hacked to Pieces? The Effects of Ransomware Attacks on Hospitals and Patients. *The Effects of Ransomware Attacks on Hospitals and Patients*, 2023.

[82] C. Miller, D. G. Widder, C. Kästner, and B. Vasilescu. Why Do People Give Up Flossing? A Study of Contributor Disengagement in Open Source. In *Open Source Systems: IFIP WG International Conference*, 2019.

[83] K. Miller. Rural texas towns report cyberattacks that caused one water system to overflow. *AP News*.

[84] A. Mirian, Z. Ma, D. Adrian, M. Tischer, T. Chuenchujit, T. Yardley, R. Berthier, J. Mason, Z. Durumeric, J. A. Halderman, et al. An Internet-Wide View of ICS Devices. In *IEEE Privacy, Security and Trust*, 2016.

[85] A. Murdock, F. Li, P. Bramsen, Z. Durumeric, and V. Paxson. Target Generation for Internet-Wide IPv6 Scanning. In *ACM Internet Measurement Conference*, 2017.

[86] T. Nasr, S. Torabi, E. Bou-Harb, C. Fachkha, and C. Assi. ChargePrint: A Framework for Internet-Scale Discovery and Security Analysis of EV Charging Management Systems. In *NDSS*, 2023.

[87] Netlas LLC. About netlas.io. https://app.netlas.io/about/.

[88] Office of the Director of National Intelligence. Recent Cyber Attacks on US Infrastructure Underscore Vulnerability of Critical US Systems. https://www.dni.gov/files/CTIIC/documents/products/Recent_Cyber_Attacks_on_US_Infrastructure_Underscore_Vulnerability_of_Critical_US_Systems-June2024.pdf.

[89] R. Padmanabhan, A. Dhamdhere, E. Aben, K. Claffy, and N. Spring. Reasons Dynamic Addresses Change. In *ACM Internet measurement conference*, 2016.

[90] R. Padmanabhan, J. P. Rula, P. Richter, S. D. Strowes, and A. Dainotti. DynamIPs: Analyzing Address Assignment Practices in IPv4 and IPv6. In *International Conference on Emerging Networking Experiments and Technologies*, 2020.

[91] S. Palkar, F. Abuzaid, P. Bailis, and M. Zaharia. Filter Before You Parse: Faster Analytics on Raw Data with Sparser. *Proceedings of the VLDB Endowment*, 2018.

[92] N. Provos and P. Honeyman. Scanssh: Scanning the internet for SSH servers. In *LISA*, 2001.

[93] N. Raman, M. Cao, Y. Tsvetkov, C. Kästner, and B. Vasilescu. Stress and Burnout in Open Source: Toward Finding, Understanding, and Mitigating Unhealthy Interactions. In *ACM/IEEE International Conference on Software Engineering: New Ideas and Emerging Results*, 2020.

[94] R. S. Raman, A. Stoll, J. Dalek, R. Ramesh, W. Scott, and R. Ensafi. Measuring the Deployment of Network Censorship Filters at Global Scale. In *NDSS*, 2020.

[95] S. Ramanathan, A. Hossain, J. Mirkovic, M. Yu, and S. Afroz. Quantifying the Impact of Blocklisting in the Age of Address Reuse. In *ACM Internet Measurement Conference*, 2020.

[96] R. Ramesh, R. S. Raman, A. Virkud, A. Dirksen, A. Huremagic, D. Fifield, D. Rodenburg, R. Hynes, D. Madory, and R. Ensafi. Network Responses to Russia's Invasion of Ukraine in 2022: A Cautionary Tale for Internet Freedom. In *USENIX Security Symposium*, 2023.

[97] Recorded Future Insikt Group. Chinese Cyberespionage Originating From Tsinghua University Infrastructure. https://go.recordedfuture.com/hubfs/reports/cta-2018-0816.pdf.

[98] P. Richter, G. Smaragdakis, D. Plonka, and A. Berger. Beyond Counting: New Perspectives on the Active IPv4 Address Space. In *ACM Internet Measurement Conference*, 2016.

[99] H. Rimlinger, K. Vermeulen, M. Gouel, O. Fourmaux, and T. Friedman. To Probe or Not to Probe? Reconciling High Speed Probing with Ethical Probing. In *CoNEXT Student Workshop*, 2023.

[100] R. Roberts and D. Levin. When certificate transparency is too transparent: Analyzing information leakage in https domain names. In *ACM Workshop on Privacy in the Electronic Society*, 2019.

[101] K. Ryan, K. He, G. A. Sullivan, and N. Heninger. Passive SSH Key Compromise via Lattices. In *ACM SIGSAC Conference on Computer and Communications Security*, 2023.

[102] E. Rye and D. Levin. IPv6 Hitlists at Scale: Be Careful What you Wish For. In *ACM SIGCOMM*, 2023.

[103] A. Sarabi, K. Jin, and M. Liu. Smart internet probing: Scanning using Adaptive Machine Learning. *Game Theory and Machine Learning for Cyber Security*, 2021.

[104] Q. Scheitle, O. Hohlfeld, J. Gamba, J. Jelten, T. Zimmermann, S. D. Strowes, and N. Vallina-Rodriguez. A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists. In *ACM Internet Measurement Conference*, 2018.

[105] Shodan. Industrial control systems, 2024. https://www.shodan.io/explore/category/industrial-control-systems.

[106] Shodan. Shodan search engine. https://www.shodan.io/, 2024.

[107] G. Song, L. He, T. Chen, J. Lin, L. Fan, K. Wen, Z. Wang, and J. Yang. PMap: Reinforcement Learning-Based Internet-Wide Port Scanning. *IEEE/ACM Transactions on Networking*, 2024.

[108] G. Song, L. He, T. Zhao, Y. Luo, Y. Wu, L. Fan, C. Li, Z. Wang, and J. Yang. Which Doors Are Open: Reinforcement Learning-based Internet-wide Port Scanning. In *IEEE/ACM Symposium on Quality of Service*, 2023.

[109] G. Song, J. Yang, L. He, Z. Wang, G. Li, C. Duan, Y. Liu, and Z. Sun. AddrMiner: A Comprehensive Global Active IPv6 Address Discovery System. In *USENIX Annual Technical Conference*, 2022.

[110] D. Springall, Z. Durumeric, and J. A. Halderman. Measuring the Security Harm of TLS Crypto Shortcuts. In *ACM Internet Measurement Conference*, 2016.

[111] B. Stock, G. Pellegrino, F. Li, M. Backes, and C. Rossow. Didn't You Hear Me?—Towards More Successful Web Vulnerability Notifications. 2018.

[112] B. Stock, G. Pellegrino, C. Rossow, M. Johns, and M. Backes. Hey, You Have a Problem: On the Feasibility of Large-Scale Web Vulnerability Notification. In *USENIX Security Symposium*, 2016.

[113] G. A. Sullivan, J. Sippe, N. Heninger, and E. Wustrow. Open to a Fault: On the Passive Compromise of TLS Keys via Transient Errors. In *USENIX Security Symposium*, 2022.

[114] C. Utz, M. Michels, M. Degeling, N. Marnau, and B. Stock. Comparing Large-Scale Privacy and Security Notifications. *Privacy Enhancing Technologies*, 2023.
[115] B. VanderSloot, J. Amann, M. Bernhard, Z. Durumeric, M. Bailey, and J. A. Halderman. Towards a Complete View of the Certificate Ecosystem. In *ACM Internet Measurement Conference*, 2016.
[116] Verizon. 2024 data breach investigations report (DBIR). https://www.verizon.com/business/resources/T32e/reports/2024-dbir-data-breach-investigations-report.pdf.
[117] M. Vermeer, J. West, A. Cuevas, S. Niu, N. Christin, M. Van Eeten, T. Fiebig, C. Ganán, and T. Moore. SoK: a Framework for Asset Discovery: Systematizing Advances in Network Measurements for Protecting Organizations. In *IEEE European Symposium on Security and Privacy*, 2021.
[118] G. Wan, L. Izhikevich, D. Adrian, K. Yoshioka, R. Holz, C. Rossow, and Z. Durumeric. On the Origin of Scanning: The Impact of Location on Internet-Wide Scans. In *ACM Internet Measurement Conference*, 2020.
[119] M. Wang and M. Zhou. Vacuum Filters: More Space-Efficient and Faster Replacement for Bloom and Cuckoo Filters. *VLDB*, 2019.
[120] F. Weimer. Factoring RSA Keys with TLS Perfect Forward Secrecy. *Red Hat Technical Report*, 2015.
[121] B. Wu, S. Zhang, Y. Liu, and Z. Yang. A Survey of Network Asset Detection Technology. In *International Conference on Network Simulation and Evaluation*, 2023.
[122] M. Wu, G. Hong, J. Chen, Q. Liu, S. Tang, Y. Li, B. Liu, H. Duan, and M. Yang. Revealing the Black Box of Device Search Engine: Scanning Assets, Strategies, and Ethical Consideration. *arXiv preprint arXiv:2412.15696*, 2024.
[123] G. Young. CQRS documents, 2010. https://cqrs.wordpress.com/.
[124] J. Zirngibl, L. Steger, P. Sattler, O. Gasser, and G. Carle. Rusty Clusters? Dusting an IPv6 Research Foundation. In *ACM Internet Measurement Conference*, 2022.
[125] ZoomEye. Zoomeye search engine. https://www.zoomeye.hk/, 2024.

Appendices are supporting material that have not been peer-reviewed.

A RESEARCH PAPER CATEGORIZATION

Topic	Papers	Perc.
BGP, Routing, and RPKI	9	4.31%
Blockchain and Cryptojacking	18	8.61%
Censorship and Balkanization	29	13.88%
Cryptography and Randomness	29	13.88%
Denial of Service and DDoS	11	5.2%
DNS, Naming, Domain Registration	22	10.53%
Email, Spam, and Phishing	14	6.70%
Fingerprinting and IoT Classification	24	11.48%
HTTPS and TLS	37	17.70%
Honeypots, Telescopes, Deception	22	10.53%
ICS Exposure and Critical Infra.	37	17.70%
Internet Scanning Methods	15	7.18%
Internet of Things (IoT)	50	23.92%
Malware, Infection, Attacker Behavior	44	21.05%
Other (Internet Measurement)	16	7.66%
Other (Security)	11	5.26%
Other (Systems Development)	5	2.39%
PKI, Certificates, Revocation, CT	57	27.27%
Privacy, Surveillance, VPNs	9	4.31%
Service Exposure and Vulnerability Tracking	31	14.83%
Web and CDNs	19	9.09%
Total Publications	509	

Table 6: Research Topics—Censys data has been used in over 500 research papers. Here, we show the breakdown of topics analyzed using Censys data.

Censys data has been used in over 500 research papers, a breakdown of which can be found in Table 6.

B DEPRECATED SCANS

As can be seen in Figure 4, we found that there is no clear inflection point for what defines a popular service and it was never clear how many and which non-standard ports that Censys should scan. After developing the predictive scan engine that we describe in the next section, we deprecated this scan and reallocated its bandwidth to the full 65K port scan. This had the added benefits of both finding additional long-lived services and providing richer training data for our predictive approach.

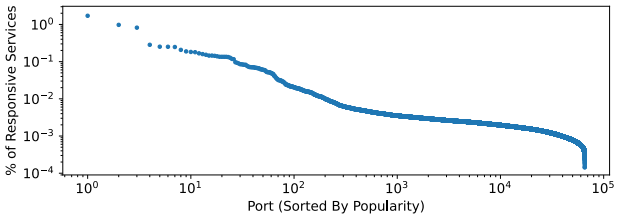

Figure 4: Service Population by Port—As seen from a sampled scan of all ports, port popularity follows a smoothly decaying distribution; no cut-off that divides "popular" from "unpopular" ports.

C SAMPLE SIZE

Measuring the coverage of alternative scanning engines is non-trivial, as described in Section 6. Consequently, we resort to randomly sub-sampling services across scanning engines to approximate metrics, such as the percentage of services returned that respond (i.e., "freshness") . In Figure 5, we show that only 50 services need to be found to accurately estimate the percentage of services expected to be responsive within a scanning search engine.

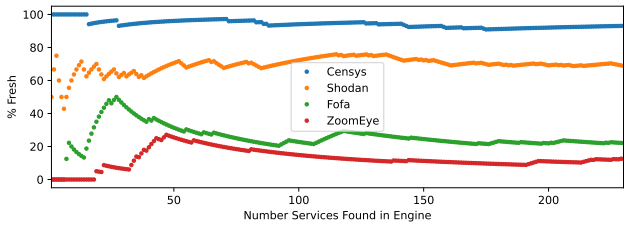

Figure 5: Sampling Services to Determine Scanning Engine Coverage—Sampling at least 50 services from random IPs is sufficient in reaching asymptotic behavior in determining the expected % of services that respond ("freshness").

D OPT-OUT PROCESS

Censys honors opt-out requests from operators who can verify network or domain ownership through public WHOIS data. On receipt of any initial reach out, Censys sends an informational email with details about Censys and stating our policy (Appendix D.1); Censys excludes prefixes if requested in follow up communication. Censys enacted this two phase policy after finding that (1) many operators rescinded their request for opt-out after understanding our scan intent and that Censys hosts were not compromised; and

(2) many initial emails are sent by automated processes and this dramatically reduces manual processing overhead. We show our response message below:

D.1 Abuse Autorespond Message

Hello {name},
These connections are part of a Censys network survey.

What is Censys?
Censys is an Internet security platform that helps organizations discover, monitor, and secure their devices on the Internet. We regularly probe every public IP address and website, curate and enrich the resulting data, and make it intelligible to defenders.

Enterprises use Censys to protect their networks. CERT groups use Censys data to alert system operators about critical Internet-facing vulnerabilities. Researchers use Censys to improve Internet protocols and understand attacker behavior. For example, Censys data has been used to inform the design of the TLS 1.3 protocol as well as to understand the

Censys Scanning and Data Collection
One way that Censys finds publicly-reachable services is by using Internet-wide scanning. We make small harmless connection attempts to every IPv4 address worldwide. When we discover that a server is configured to accept connections, we follow up by completing a protocol handshake (e.g., HTTP request) to learn more about the running services.

We never attempt to bypass any authentication or technical barriers, exploit security problems, or access internal, non-public-facing services. The only data we receive is publicly visible information to anyone on the Internet who connects to the service. If you would like to verify the benign nature of our scans, you can do so on the independent Greynoise platform.

Allowlisting or Blocklisting Scans
Networks and websites can request exclusion from our scans. To do so, please have a publicly-verifiable WHOIS contact associated with the IP block or domain respond to this email and request exclusion.

If you would like to have a host or website excluded from Censys, but do not control an IP allocation or namespace, you can blocklist the IP ranges that we use for scanning:

162.142.125.0/24
167.94.138.0/24
167.94.145.0/24
167.94.146.0/24
167.248.133.0/24
2602:80d:1000:b0cc:e::/80
2620:96:e000:b0cc:e::/80

Additionally, our HTTP-based scans use a Censys specific user-agent, which can be used to filter requests from our scanners.

```
Mozilla/5.0 (compatible; Censys Inspect/1.1;
+https://about.censys.io/)
```

Thank you,
Censys Abuse Team

E REPRODUCIBILITY OF EVALUATION

In Section 6, we compare Censys to alternative scanning engines. When querying for Censys-found services, we use Censys' Big Query SQL interface and query it the following way:

```
SELECT distinct host_identifier.ipv4 ip, s.port p
FROM ‘censys-io.universal_internet_dataset_v2.base‘,
UNNEST(services)s
WHERE s.extended_service_name = 'SERVICE'
and TIMESTAMP_TRUNC(snapshot_date, DAY) =
TIMESTAMP("YEAR-MONTH-DAY")
and s.pending_removal_since is null
```

Unlike Censys, we do not have direct access to raw data from other scan engines and we instead use the browser interface and API to query alternative engines. In Table 7, we list the syntax we use to query IPs in bulk (e.g., to randomly sub-sample). In Table 8, we list the syntax we use to query all results for specific industrial control system protocols. In Table 9, we list the syntax we use to query all results for specific protocols.

Engine	Bulk Query Syntax
Shodan	`ip:<address>,<address>`
FofA	`ip="<address>" || ip="<address>"`
ZoomEye	Upload a `.txt` file of IPs
Netlas	`ip:<address> OR ip:<address>`
BinaryEdge	`ip:<address> OR ip:<address>`

Table 7: Querying alternative scanning engines—Alternative scanning engines do not provide a SQL interface to directly query scanning data to researchers. We use the following syntax in every scanning engine web interface to bulk query scanning data for groups of IP addresses.

Protocol	Censys Query	Shodan Query	ZoomEye Query	Fofa Query
ATG	services.service_name="ATG"	shodan.module:"automated-tank-gauge"	-	protocol= "automated-tank-gauge"
BACNET	services.service_name="BACNET"	shodan.module:"bacnet"	+service:"bacnet"	protocol="bacnet"
CIMON_PLC	services.service_name="CIMON_PLC"	-	-	-
CMORE	services.service_name="CMORE"	-	-	-
CODESYS	services.service_name="CODESYS"	shodan.module:"codesys"	+service:"CoDeSys"	protocol="codesys"
DIGI	services.service_name="DIGI"	-	-	-
DNP3	services.service_name="DNP3"	port:20000 source address	+service:"dnp3"	protocol="dnp3"
EIP	services.service_name="EIP"	shodan.module:"ethernetip"	-	-
FINS	services.service_name="FINS"	port:9600 response code	+service:"fins"	protocol="fins"
FOX	services.service_name="FOX"	port:1911,4911 product:Niagara	+service:"fox"	protocol="fox"
GE_SRTP	services.service_name="GE_SRTP"	port:18245,18246 product:"general electric"	+service:"GE-SRTP"	-
HART	services.service_name="HART"	port:5094 hart-ip	+service:"hart"	protocol="hart"
IEC60870	services.service_name="IEC60870_5_104"	port:2404 asdu address	+service:"IEC 60870-5-104"	protocol="iec-104"
MODBUS	services.service_name="MODBUS"	shodan.module:"modbus"	+service:"modbus"	protocol="modbus"
OPC_UA	services.service_name="OPC_UA"	shodan.module:"opc-ua"	-	-
PCOM	services.service_name="PCOM"	-	-	-
PCWORX	services.service_name="PCWORX"	port:1962 PLC	-	protocol="pcworx"
PROCONOS	services.service_name="PRO_CON_OS"	port:20547 PLC	+service:"ProConOS"	protocol="proconos"
REDLION	services.service_name= "REDLION_CRIMSON"	port:789 product:"Red Lion Controls"	+service:"crimson-v3"	protocol="redlion-crimson3"
S7	services.service_name="S7"	shodan.module:"s7"	+service:"s7"	protocol="s7"
WDBRPC	services.service_name="WDBRPC"	shodan.module:"wdbrpc"	+service:"wdbrpc"	protocol="wdbrpc"

Table 8: Industrial Control System Protocols query syntax for different search engines—For Shodan, if a filter is not found or the recommended filter [105] only filters by port number, we query using the shodan.module filter.

Protocol	Censys Query	ZoomEye Query	Fofa Query	Netlas Query	Shodan Query
HTTP	services.service_name="HTTP"	+service:"http"	protocol="http"	protocol:http	shodan.module="http"
HTTPS	services.service_name="HTTPS"	+service:"https"	protocol="https"	-	shodan.module="https"
FTP	services.service_name="FTP"	+service:"ftp"	protocol="ftp"	protocol:ftp	shodan.module="ftp"
SNMP	services.service_name="SNMP"	+service:"snmp"	protocol="snmp"	-	shodan.module="snmp"
Telnet	services.service_name="Telnet"	+service:"telnet"	protocol="telnet"	protocol:telnet	shodan.module="telnet"
TFTP	services.service_name="TFTP"	+service:"tftp"	protocol="tftp"	-	-
RDP	services.service_name="RDP"	+service:"rdp"	protocol="rdp"	protocol:rdp	shodan.module="rdp"
rlogin	services.service_name="rlogin"	service:"rlogin"	protocol="rlogin"	-	shodan.module="rlogin"
RSH	-	-	-	-	-
SSH	services.service_name="SSH"	+service:"ssh"	protocol="ssh"	protocol:ssh	shodan.module="shodan"
SMB	services.service_name="SMB"	+service:"smb"	protocol="smb"	protocol:smb	shodan.module="smb"
VNC	services.service_name="VNC"	+service:"vnc"	protocol="vnc"	protocol:vnc	shodan.module="vnc"
X11	services.service_name="X11"	+service:"x11"	protocol="x11"	protocol:x11	shodan.module="x11"

Table 9: Query Syntax for Protocols across Platforms—All scanning engines support most protocols.