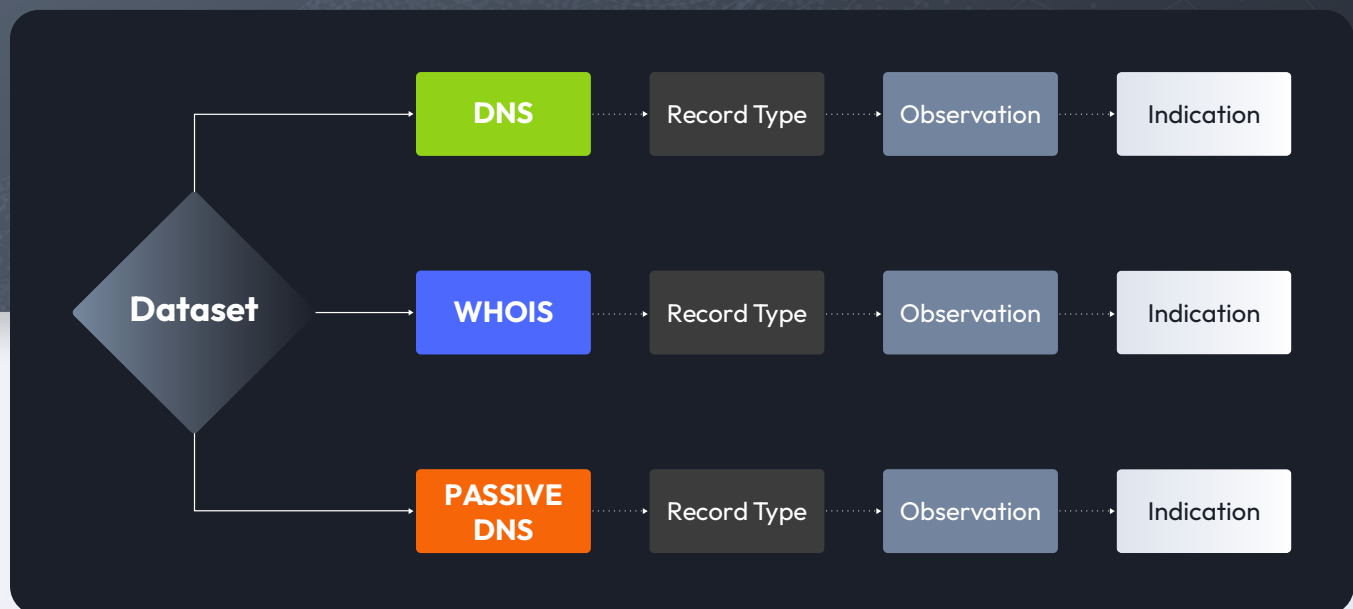



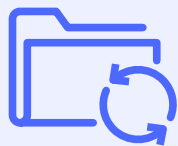
Valuable Datasets to **Analyze Network Infrastructure**



Dataset	Record Type	Observation	Potential Indication
 DNS	ISP Info / ASN	May be a bulletproof hosting company (or a “resurrected” out-of-business entity with associated netblock assets).	Bulletproof hosting is a type of service mostly used by malicious entities to combat potential takedown activities.
	IP Address	Frequently changing IP addresses associated with a domain name, often hosted on dynamic consumer broadband IPs.	Fast flux, a technique often used to make takedowns of malicious infrastructure more difficult.
	IP Address	Low number of domains hosted on a single IP address with consistent naming.	Can give high confidence that infrastructure is controlled by a single entity. Use DNSDB to validate.



Dataset	Record Type	Observation	Potential Indication
	IP Address	Where an IP address lives or “is announced” can tell you about its expected life.	As mismatch in expectations can mean a lot: Residential IPs should not host services, for example. Or a Russian geo-located IP providing services to Canadian customers.
	Nameserver	Cryptocurrency-themed hosters.	Suspicious infrastructure.
	Nameserver	Low number of domains with consistent naming pointing to a nameserver or nameserver IP.	Tight infrastructure correlations often imply shared ownership / control, so connected domains and infrastructure may act as valuable artifacts.
	Nameserver	Domain's NS record points to a nameserver owned by those operating a sinkhole.	Shadowserver and Microsoft do a lot of sinkholing, as do many independent security researchers.
	Nameserver	Single nameserver for a domain.	Suspicious infrastructure. A legitimate domain would likely have redundant authoritative nameservers.
	Nameserver	Cloudflare combination.	When several domains share the same combination of hostnames, there is a higher likelihood that they are related.
	SOA	Unique RNAME emails associated with known bad domain names.	The email address can cluster domains, which may represent malicious campaign infrastructure.
	SOA	Short TTL.	The average / Default TTL is 3600 seconds; very low TTLs for services that don't appear to require or need this may indicate a fast-flux network, especially if other red flag exist.
	MX	High entropy in the MX record name.	Malicious intent.
	MX	Typosquatting of a more well-known brand.	Malicious, likely phishing intent.
	TLD	Uncommon or inexpensive TLD (.top, .tk, .tv etc..).	Threat actor acquiring inexpensive domains in TLDs where realistic spoof names are sometimes more available.



WHOIS

Dataset	Record Type	Observation	Potential Indication
	Domain Name	Typosquatting or non-typo spoofing (e.g. affixes/prefixes).	Suspicious infrastructure.
	Domain Name	High entropy strings or a combination of random words.	Potential use of DGA technique.
	Domain Name	Newly-registered, culturally-relevant themed domain names with close proximity to blocklisted infrastructure.	Suspicious infrastructure.
	Registrant Email	Unique emails associated with other malicious domains, SOA records, or SSL records.	Suspicious infrastructure.
	Registrant Email	Unique free email domains with a higher concentration of badness in combination with close proximity to known bad infrastructure.	Suspicious infrastructure.
	Registrant Address	Elements of a unique address shared between a small number of domain names (especially if the domain names share a collective theme).	Shared domain ownership.
	Registrant Address	Inconsistent or inaccurate address information that isn't associated with a legitimate entity.	Suspicious infrastructure.
	Registrant Phone Number	Inconsistent or inaccurate phone information that isn't associated with a legitimate entity, or country/area code that doesn't match the provided address.	Suspicious infrastructure.
	Registrar Name	Registrars operating out of countries who aren't likely to respond to legal actions by the US and EU.	Suspicious infrastructure.
	Registrar Name	Registrars accepting cryptocurrency as payment can provide a safe haven for cyber-criminals.	Suspicious infrastructure.
	Create Date	Domain age is less than 30 days.	Suspicious infrastructure.



PASSIVE DNS

Shown as separate category from DNS to differentiate between actively querying DNS servers and passively collecting DNS information from network traffic

Dataset	Record Type	Observation	Potential Indication
	Expiration Date	Recently expired domain with changed registration information from previous ownership.	Potential BEC or phishing infrastructure.
	Subdomain	Typosquatting or non-typo spoofing (e.g. affixes/prefixes).	Suspicious infrastructure.
	Hostname	Hostname with high entropy.	Potential DNS tunneling infrastructure.
	Hostname	Hostname (DNS RNAME) with 27+ unique characters.	High likelihood of DNS tunneling infrastructure.
	AAAA	Queries with odd A or AAAA responses.	Potential C2 infrastructure.
	IP	A single IP cycling quickly through queries or domains.	Fast flux.
	IP	(a) hosting on dynamic consumer broadband IPs (e.g., DHCP addresses) and (b) IPs from multiple ASNs, and (c) NOT something associated with a CDN.	Fast Flux
	Nameserver	Nameserver response associated with known badness associated with a provider.	Attacker is likely running their own infrastructure.
	CNAME	Typosquatting.	Potential phishing infrastructure.
	CNAME	CNAME connected to other known-bad infrastructure (e.g hostnames, domains, etc).	Suspicious infrastructure.
	TXT	TXT responses with SPF or DKIM records associated with typosquatting domains/ CNAMEs/Subdomains/ Nameservers.	Potential BEC or phishing infrastructure.

Other datasets to potentially consider (in alphabetical order)

- **BGP data** (particularly useful <https://bgp.he.net/> and/or Oregon Routeviews (<https://www.routeviews.org/routeviews/>) and/or Team Cymru IP to ASN mapping service (see <https://www.team-cymru.com/ip-asn-mapping>) can map ASNs to CIDR netblocks, or IPs to ASNs. This can be surprisingly useful as a “rough cut” when confronting large lists of IPs that need organization and analysis. Dataplane.org is a nice example of a site that leverages BGP data to organize the data it sees and shares, see for example <https://dataplane.org/signals/sshclient.txt> (scroll down to the list of hits sorted by ASN).
- **Blocklisting status** (example: check <https://multirbl.valli.org/> for listing status) – if a domain or IP is being used for abusive purposes, often it may already be tagged as a result of that activity. Note that not all blocklists are equally trustworthy!
- **CommonCrawl** (<https://commoncrawl.org/>): Everyone wants to build their own web crawler – why not just use the data that others have already collected? Find what pages exist, and what’s on them. Great source of FQDNs to actively resolve, too.
- **Corruption Index** (<https://www.transparency.org/en/cpi/2022>) – beware businesses located where corruption is endemic. Likely of fraud, failure to deliver with little recourse, etc., tends to be higher where corruption is rampant.
- **Cryptocurrency transaction log data**: Example of a company specializing in this: <https://ciphertrace.com/>.
- **Darknet (or “Network Telescope”) data**: Imagine blocks of network address space that are announced in network routing tables, but which have neither servers nor end users. Those networks SHOULD be quiescent, but normally see a steady barrage of inbound scans and other unsolicited traffic. Some of that traffic is associated with network measurement projects; other times the probes will be highly indicative of the exploit-de-jour or reconnaissance prior to a targeted attack. CAIDA has a particularly large network telescope (https://www.caida.org/projects/network_telescope/) ; see also <https://www.greynoise.io/>.
- **Geolocation** (<https://dev.maxmind.com/geoip/geolite2-free-geolocation-data?lang=en> for example) can provide valuable hints about where infrastructure is located. Some geolocations are unduly popular with bad actors (Seychelles, Cyprus, Panama, Iceland, etc.) due to strong general (or bank/corporate) privacy laws, easy corporate shelf registration availability, “Golden passport” availability (leading to concentration of organized criminals fleeing extradition), or other factors.
- **Log data (firewall logs, intrusion detection system logs, proxy server logs, web server logs, etc):** Tremendous amount of intelligence exists in log data, yet many times the data isn’t even centrally collected or reviewed. Has a given IP attempted to login as root 498,000 times? Perhaps it is attempting a password guessing attack. Is a site attempting to get at a non-existent web resource with a URL that looks hostile? (OWASP has great tips around suspiciously formatted URLs). This is Splunk’s bread and butter, or try the open source alternative, typically “Elastic Stack”.
- **Malware Hashes**: commonly shared as indicators of compromise. If an executable matches a known malware hash, this is a strong indicator that you’ve got a problem. “Mystery executables” will often be submitted to a publicly available sandbox to be checked against known malware hash databases, and/or for execution in an instrumented environment where suspicious behavior can be detected (“Hmm. This application claims to be a free flashlight, but it accesses my contact list and attempts to send email to each of them.” or perhaps “This application claims to optimize my system, but has been pegging the CPU and appears to be checking in with cryptocurrency mining sites.”).

- **Netflow ("IPFix") flow data:** flow data summarizes network traffic between a src (IP, port) and a dst (IP, port), including times and flow size in octets. This can be surprisingly useful as a "finger pointing" tool. For example, normally a consumer broadband IP will not be making outbound connections to large numbers of diverse MX servers (unless it is sending spam). Similarly, symmetric inbound and outbound flows can be highly indicative of a proxy server running on a system. Rarely shared in non-anonymized form with 3rd parties because it is so revealing to a trained analyst. Example talk: "A Look at the Unidentified Half of Netflow," <https://www.stsauver.com/joe/missing-half/missing-half.pdf>.
- **Open Proxy Server/ToR Exit Nodes:** Knowing that a site is an open proxy server or ToR exit node is an immediate strong indication that random people will be engaging in abuse through that IP. In the case of ToR: <https://check.torproject.org/torbulkexitlist>.
- **Volumetric Data:** often you won't even need full contents, just knowing the volume of traffic associated with something can be an indicator of badness, or that a site is being targeted or exploited to conduct an attack. Sometimes referred to as "metadata" or performing "traffic analysis." Example talk: "The Enduring Challenges of Traffic Analysis," <https://www.stsauver.com/joe/dublin-traffic-analysis/dublin-traffic-analysis.pdf>.
- **WHOIS for numeric resources:** While domain whois has largely been gutted due to GDPR and use of privacy/proxy services (and/or non-cooperative non-ICANN domains), IP WHOIS remains surprisingly useful if you shift from looking at domains to looking at IPs and ASNs. Particularly interesting: legacy netblocks, netblocks for defunct companies, netblocks used "out of region." Noteworthy expose: <https://chicago.suntimes.com/2021/10/4/22709624/afrinic-internet-protocol-addresses-ip-africa-lu-hen-g-china-broker>.

Another good way to identify cyber intelligence datasets is by looking at threat intelligence feeds such as:

- MISP Default Feeds <https://www.misp-project.org/feeds/>
(see also <https://www.misp-project.org/communities/>)
- OpenCTI Feeds
<https://filiгран.notion.site/3bb8723a91cc42488387ce6a9f06385f?v=f20da28ed8aa4adf879fd9ea6ee7faa0>
- Harpoon Feeds <https://github.com/Te-k/harpoon>
<https://www.anomali.com/marketplace/threat-intelligence-feeds>
<https://gosint.readthedocs.io/en/latest/configuration.html>