Black Sheep Wall: Towards Multiple Vantage Point-based Information Space Situational Awareness

Bernhards Blumbergs

This is the author's copy of the work. The paper has been and published in SECRYPT 2024 conference proceedings. Please use the following reference when citing this work:

B.Blumbergs (2024). Black Sheep Wall: Towards Multiple Vantage Point-based Information Space Situational Awareness. In proceedings of 21st International Conference on Security and Cryptography (SECRYPT 2024). SCITEPRESS. Dijon, France

Black Sheep Wall: Towards Multiple Vantage Point-based Information Space Situational Awareness

Bernhards Blumbergs[1 2]

[1]*Cyber Resilience Laboratory, Information Science Division, Nara Institute of Science and Technology, Nara, Japan*

[2]*CERT.LV, Institute of Mathematics and Computer Science, University of Latvia, Riga, Latvia*

{name.surname}.bb2[a]is.naist.jp, {name.surname}[a]cert.lv

Keywords: Incident Response, Information Space, Situational Awareness, Cyber Threat Intelligence, Distributed Website Data Mining

Abstract: CSIRTs rely on processing extensive amounts of incident and threat intelligence data. While the market is saturated with such solutions, they are limited to a narrow range of Internet positions for data collection, impeding the establishment of the security context and comprehensive awareness of the monitored Internet resources. To tackle this challenge, a novel approach is proposed for distributed content collection. Simultaneously employing multiple Internet positions and various content access techniques, a broader representation of the content may be obtained by combining data from all positions, followed by automated difference analysis and clustering. The solution enables fully automated large-scale deployments across globally distributed IP networks and seamless integration into existing toolsets. It enhances CSIRT capabilities in identifying content changes, access restrictions, contextual intelligence on cybercrime and threat actor campaigns, as well as detecting defacement and availability attacks, and misinformation attempts. Initial evaluation of the prototype demonstrated its effectiveness by detecting significant and distinct changes in website content, thereby providing expanded visibility and intelligence. Prototype code and validation datasets are released publicly for further use, research, and validation.

1 INTRODUCTION

In the current information age, collecting large volumes of data, performing near real-time processing, and deriving relevant and actionable information provide superiority over rivals and adversaries. Data aggregation, storage, and processing demand have increased, leading to competitive market saturation by commercial and community initiatives (G2.com Inc., 2024). This bolsters the development of unique approaches and solutions permitting each market player to focus on a bespoke aspect of data collection, analytics, and analysis of the information space. Incident response team (i.e., CSIRT/CERT) establishes their situational awareness based on the aggregated information via automated commercial, community-provided, or own-collected threat feeds and reports. In addition, targeted data collection from websites or other information sources may be performed to enrich the aggregated threat feeds. Threat intelligence collection and web resource content monitoring solutions may be used to accomplish this task. However, information resource access and data collection are conducted from a single or limited set of positions within the service provider-controlled IP address space. Simultaneous data collection from the same information source from a wide and geographically distributed set of positions is not being considered. From a broader cyber security perspective, such an approach may present limitations towards establishing comprehensive information space awareness, contextual information gathering, and conducting informed incident response activities.

This work proposes a hypothesis, that an information source (e.g., a website) may yield a different content based on how it is being accessed with the identified changes providing contextual intelligence information. Varying access methods may reveal the dynamic nature of the information source and deliver additional information to enrich the incident response process otherwise not provided by the current information space awareness solutions. In this work, the term of *vantage point* is understood as a combination of a connection method and the position within the Internet from which an information source is being accessed. By deploying a set of collectors, the content from the same information source may be collected simultaneously from multiple vantage points (Figure 1). For example, accessing an information source from a different geographically assigned IP address space, using private or well-known VPN or proxy services, routing through anonymization network exit nodes, or changing connection parameters (e.g., user-agent) may represent differences in the retrieved content. In this work, the *information space* is understood as a set of information sources either publicly available (e.g., clear-net) or ones with specific access requirements (e.g., deep-net or dark-net) relevant to the entity engaged in data collection. The information space may cover a broad range of resources, such as, websites, news articles, social network posts, online databases and repositories, malicious website resources, and leaked information dumps. Collecting and analysing retrieved multiple vantage-point-based data may reveal the information source's changing behaviour and provide additional visibility to the CSIRT team, expanding their capabilities. Differences in content representation based on the origin and position on the Internet is a commercially driven and well-accepted practice by benign services for performing geographic load-balancing, serving regionally relevant content, or search engine optimization. However, the broader contextual perspective is achieved by identifying particular vantage points for which distinct changes may be identified in the delivered content among all the employed vantage points.

This paper provides the following novel applied contributions (released publicly on https://github.com/lockout/b-swarm):

1. multiple vantage-point-based information space situational awareness concept description, its prototype design and implementation details, and complete source code (GPLv3 license);
2. multiple vantage-point-based benign and malicious website content snapshot collection, initial evaluation, and data set (MIT license).

This paper is structured as follows – Chapter 2 reviews the identified related work; Chapter 3 describes the proposed prototype design considerations and its implementation; Chapter 4 provides the overview of the prototype deployment, data set collection, and evaluation; and Chapter 5 concludes the paper and presents future development directions.

2 BACKGROUND AND RELATED WORK

Related work within open-source data gathering and information space awareness considers academic research, vendor whitepapers, solution descriptions, and technical blogs and write-ups. Searches were conducted using the Internet and academic database search engines (e.g., IEEE Xplore, ACM DL, Springer, Elsevier, ResearchGate, and Google Scholar). A primary time frame since 2017 was selected to represent the latest developments and advances in this evolving field. The search queries consist of keyword combinations – "request origin" OR "effect of origin", "impact of location", "website content change", "website content mining" AND "distributed website content mining", "open-source intelligence" OR "OSINT", and "cyber threat intelligence" OR "CTI".

A. Distributed Website Data Mining. News portal and social network content monitoring solutions for information space awareness are primarily aimed towards information space exposure assessment (Meltwater, 2024)(SK-CERT, 2024), emerging news alert identification (Google LLC, 2024), social media management (Hootsuite Inc., 2024), and marketing (Brand24 Global Inc., 2024). Open-source intelligence and cyber threat intelligence collection services aim at big data aggregation and analytics from clear-net (Gartner Inc., 2024) and dark-net resources (Flashpoint, 2024)(KELA, 2024)(Mandiant, 2024). There is an understandable lack of publicly available information related to the data collection specifics of commercial solutions. Based on the solution website descriptions and publicly available information, it has not been observed that these services provide information space visibility from multiple vantage points and offer identification of content differences among them. Instead, the end user accesses and queries a single stream of data related to specific monitoring, situational awareness, and data collection tasks. This may indicate that service providers do not consider the dynamic nature of the content based on the vantage point and will provide a limited single-faceted visibility of the content.

Wan et.al. (Wan et al., 2020) explore the impact of location against the IPv4 network layer 3 and 4 scans for TCP/80 HTTP, TCP/443 HTTPS, and TCP/22 SSH ports and identified that the origin of the scan impacts the results. The authors compared collected results to *censys.io* service results and recognized that visibility is lost due to the limited scan origins of the service. The paper proposes a multi-origin approach with three origins giving the optimal scan results. This paper does not consider application layer 7 data or approaches to trigger the content changes. Pham et.al. (Pham et al., 2016) explore web crawling strategies to bypass the *web robot* detection mechanisms used by websites. Crafted HTTP GET requests with six well-known crawler user agents were sent directly and

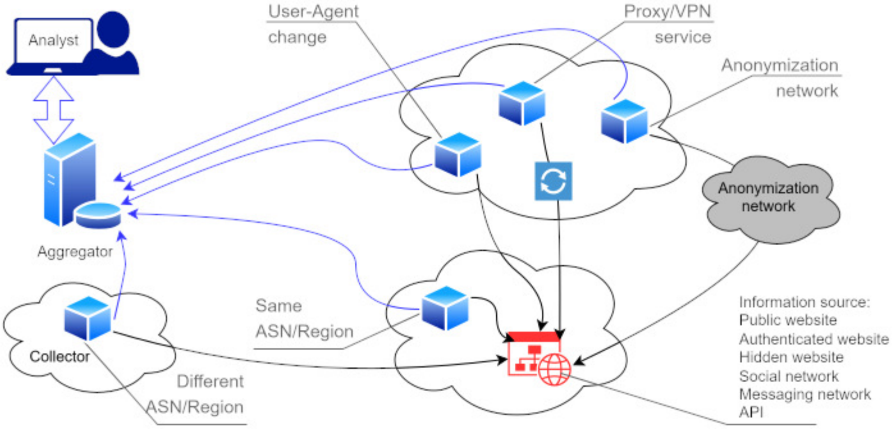

Figure 1: Information source interaction via multiple vantage points.

over the Tor network and observed the website's HTTP response header. The authors observed the differences in HTTP response status code and content length fields based on changed user agents and Tor network sources used for HTTP requests. This paper limits itself to HTTP response header analysis.

B. Website Data Mining and Analysis. The website's dynamic nature, adaptive content delivery mechanisms, and saturation with unrelated data (e.g., linked external resources and paid content) pose a challenge for identifying and extracting the main intended information of the webpage. This area has been well researched with multiple approaches being proposed for dynamic website content mining (Pujar and Mundada, 2021)(Russell and Klassen, 2018)(Neuendorf, 2017). More recognized approaches include webpage document model (DOM) parsing (Nadee and Prutsachainimmit, 2018)(Alarte et al., 2018), and natural language processing techniques (Nadee and Prutsachainimmit, 2018)(Thorleuchter and Van den Poel, 2012)(Gibson et al., 2007). Some researchers propose the use of document clustering or auto-regressive moving average algorithms for data processing and information extraction (Suganya and Vijayarani, 2020)(Calzarossa and Tessera, 2018), graph theory for pattern discovery (Leung et al., 2022), or analysis of website metadata (McGahagan et al., 2019a)(McGahagan et al., 2019b)(Song et al., 2020)(Kohli et al., 2012). Alternatively, the website content identification, labelling, and extraction may be crowdsourced to human operators (e.g., Amazon Mechanical Turk) (Kiesel et al., 2020).

For dark-net data mining content specifics, the highly dynamic flux nature of such websites, and access restrictions (e.g., rate limiting, DDoS protection, and CAPTCHA challenges) need to be taken into account. Dark-net information tracking has been rising in prominence and various initiatives relevant to CSIRT communities have been established (Shadowserver Foundation, 2024)(SISSDEN, 2024)(Lewis, 2024)(Expert Insights, 2024a)(Expert Insights, 2024b). The most common approaches include dark-net forums and marketplace crawling for the topic keyword extraction (Yang et al., 2020)(Takaaki and Atsuo, 2019) and content scraping and analysis (Lawrence et al., 2017)(Crowder and Lansiquot, 2021)(Pantelis et al., 2021)(Samtani et al., 2021). Measures to attempt to surpass dark-net crawling protection mechanisms have been evaluated, such as, a layer of SOCKS proxies, clearing stored cookies, changing *user-agent* (Pantelis et al., 2021), crawling speed limitations, and crowdsourcing or automating CAPTCHA solving (Samtani et al., 2021)(Lawrence et al., 2017).

3 PROTOTYPE DEVELOPMENT CONCEPTS

This section presents and describes the prototype's core operational concepts and design approaches, based on the published technical report (Blumbergs, 2023). In this work, a complete set of simultaneously collected information source content instances is referred to as a *snapshot* representing the dynamic nature of the same information source from all vantage points. Further automated snapshot analysis may identify any changes at a given time between the instances within the same snapshot. Furthermore, multiple snapshot creation over time against the same data source and its cross-correlation among the past snapshots may present the nature of changes in the data source over time. Detected changes are displayed to a human analyst in a structured and visual format for further analysis and evaluation. This work limits its scope to multivantage point-based data collection, basic snapshot analysis, and representation to the analyst. Although machine-assisted and guided, the in-depth analysis and deriving the meaning of the changes in most cases will be context-driven, dependent on the information source, and the reasons for engaging in the data collection.

The prototype aims at enhancing the following CSIRT operational capabilities:

1. identification of changes within the snapshot. Changes in the website content collected from geographically distributed vantage points may permit activities, such as, the evaluation of content changes due to geographical distribution, identification of access restrictions, assessment of phishing websites, and resources used for cybercrime and targeted attacks. For example, the detection of a website serving malicious content only to connections originating from specific IP address ranges or countries may provide contextual information on the potential targeted scope, victims, and likely adversarial intentions;
2. identification of dynamic changes between a sequence of snapshots over time. Such changes may permit activities, such as, observation of the content availability, tracking and identification of changes in the website content, which may lead to the disclosure and tracking of misinformation and disinformation campaigns, as well as (D)DoS and defacement attacks;
3. detection of keywords in the website content. May allow activities, such as, the identification of leaked data and the tracking of specific trigger words. For example, detection of information disclosure linked to organization-owned or affiliated domain names, employee email addresses, and user accounts.

The prototype is written in Python3 language, and its functionality and implementation consist of the following stages: 1) collector deployment and configuration, 2) collector initialization and process management, 3) information space interaction and data collection, 4) snapshot report assembly and storage, and 5) snapshot processing and analysis. These stages are represented in Figure 2 and are described further in this chapter. The complete prototype source code is released publicly under the GNU GPLv3 license on GitHub as listed in the contributions.

A. Collector Implementation. The collector engine uses Docker container technology to create a self-contained environment, enable flexible automation via Docker Compose, and permit scalable cloud-based deployments. The collector handles both the underlying connection establishment and data collection from the target source. Implemented data connection types include the implicit deployment IP range within the cloud services assigned region, Tor anonymization network, proxy services, and support for VPN connection services. To establish a vantage point for the collector (Figure 1), the following approaches are implemented: 1) geographically distributed IP address space deployments (e.g., cloud or virtual private server hosting services), 2) network proxy services (e.g., public or private HTTP(S) or SOCKS servers), 3) anonymization networks with Internet exit nodes (e.g., local Tor router as a SOCKS5 proxy), and 4) change of HTTP User-Agent (i.e., a list of mobile and desktop browser user agents). A Google Chromium headless browser is used and automated by using Selenium. A JSON-based configuration profile defines the required variables and settings for the collector instance, collection process, and interaction with the information source.

B. Snapshot Data Collection. The structure of the single interaction report holds the data related to the target information source with all the collected data and metadata representing the various features of that information source, its state, and its behaviour. The choices of the retrieved feature data are related to the unique identification of the target information source, its parameters, and served content. Such a single interaction report structure is chosen to be ready for unsupervised machine learning algorithms for cases, such as, collected data feature analysis, clustering, and pattern and outlier identification. A complete set of all interaction reports is assembled into a single snapshot in *Apache Parquet* columnar-based file format, which allows more efficient data storage, searching, and result processing.

Collected resource data consists of 1) collector metadata, 2) information source metadata, and 3) collected data and its metadata. Collector metadata includes the following features – vantage point unique identifier, specified target URL, visited URL which may differ from the initial one in case of HTTP redirection, used user-agent string, and interaction time measurement. Information source metadata includes the following features – HTTP header data (e.g., session cookies, identifiers, and parameters) and the SSL certificate fingerprint. This data may allow further collection of additional information (e.g., passive-SSL or passive-DNS lookups) helping identify and classify the information source. Collected data and its metadata include the following features – the retrieved HTML content and its SHA256 hash and *ppdeep* fuzzy-hash, retrieved data Shannon entropy calculation, and infor-

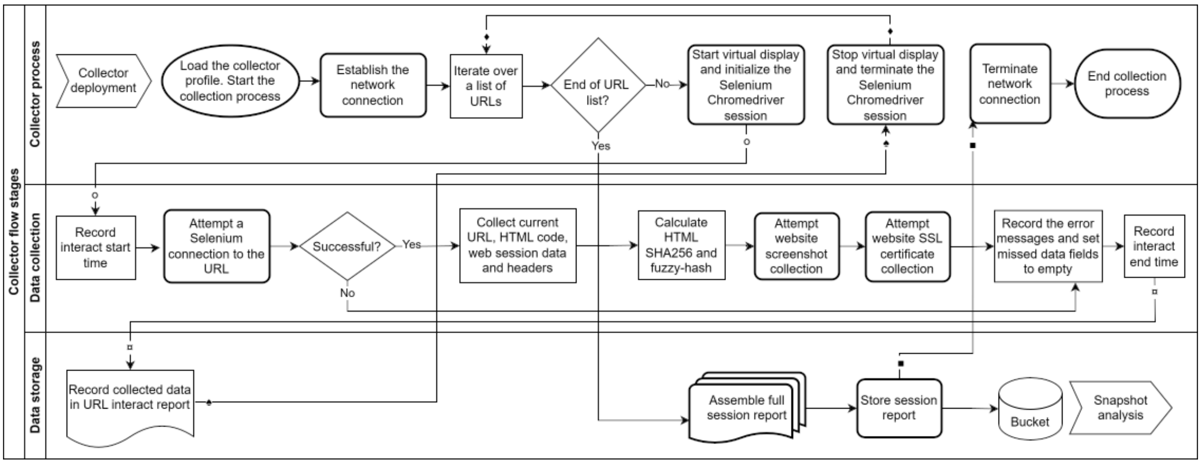

Figure 2: Collector process flow.

mation source graphical representation (i.e., screenshot). These retrieved content hashes are used for similarity analysis of the extracted content between other instances within the same snapshot. The graphical screenshot is collected via the Selenium Chromium web driver, which permits a visual comparison, change detection, and evaluation. To perform the similarity comparison between the captured screenshots, the visual difference mean squared error (MSE) is calculated. Calculating the derived content fuzzy-hashes and screenshot MSE values may permit the establishment of threshold criteria to identify and present relevant changes within and between snapshots when performing automated snapshot analysis.

C. Snapshot Data Parsing. Each collector instance produces a single file in Apache Parquet format. The collected snapshot data is loaded in *ClickHouse* high-performance column-oriented database management system to form a single table representing the collected data for the specified information space from the collector vantage points. For the initial snapshot data parsing and representation, a Jupyter Notebook with IPython Widgets is developed to give the analyst basic means for snapshot exploration and is released publicly on GitHub under the GPLv3 licence as a part of the prototype package. Notebook queries and loads snapshot table results for processing as Pandas DataFrames for each URL representing the collected data across all related vantage points. To permit further automatic snapshot analysis and change identification, additional values are calculated based on the collected SHA256, fuzzy-hash, MSE, and Shannon entropy feature values (i.e., key feature values). Within the URL-specific DataFrame, values from one vantage point are measured against all other vantage points to establish how one vantage point compares to the others within the same snapshot by calculating mean values for all key features. Lower mean values for the key feature indicate a lower content similarity against the rest of the elements in the set, except a lower mean value for the MSE feature indicates a higher visual similarity against other set elements. Based on the enriched data, an automated analysis is performed to identify clusters of vantage points by using DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm. Similar content is assumed if fuzzy-hash and MSE mean values are within the threshold range (91.0-100.0%) and all samples are assigned to the same cluster. The human analyst is represented with a smaller set of URLs and their clustering labels for which differences among vantage points have been identified.

The human analyst may use the analysis Notebook to visually interact with the snapshot data and focus attention on the URLs with changes in their content representation instead of manually analyzing the whole snapshot data set (Figure 3). This rudimentary way is sufficient for the initial evaluation of the collected snapshot data set and prototype validation. The applicability and evaluation of specific machine learning algorithms to perform a more efficient data analysis, feature detection, and clustering are out of the scope of the current research and will be explored in-depth in upcoming work.

4 PROTOTYPE ASSESSMENT

The validation aims to ensure, that the prototype code performs according to the designed functionality, all automation tasks are finished with no errors, completed deployment of the collector nodes across globally distributed network ranges, can process a set of target information space resources, and successfully performs the data collection and storage. Additionally, sample data sets are created to verify the collection process and assess the snapshot. For validation purposes only, the online resource selection is divided into two main categories – benign web resources (e.g., news portals, social networks, forums) and malicious web resources (e.g., phishing websites).

A. Prototype Deployment Process. The Google Cloud Platform (GCP) is chosen for the prototype validation due to its geographically distributed cloud network ranges, flexible command line automation, Python API support, native Docker engine support, and running Docker containers as one-time jobs. Docker specifications, configuration files, and automation scripts are released publicly on GitHub under the GPLv3 license as a part of the prototype package. For validation purposes, a subset of 18 out of all 37 publicly announced GCP regional networks is selected to ensure at least one vantage point from each regional network covering Asia, Australia, Europe, the Middle East, North America, South America, and the United States. This ensures that data collection is performed with a broad perspective from globally distributed vantage points. The job deployment and execution time heavily depend on the number of URLs to be accessed, their network distance from the collector, access method, and connection speed and latency. Additional checks are implemented to avoid timed-out sessions, partially loaded resources, or extended loading times. For information space awareness, near real-time data collection is not required and time constraints may be more flexible to complete collector deployment and snapshot data collection.

All automated container deployments and GCP Cloud Run Jobs were completed without errors and produced a prototype validation dataset. For informative purposes, Table 1 presents the individual data set collection metrics: 1) URL count in the specified information space and used vantage point count, 2) total Parquet format snapshot file size and an average value for each vantage point, 3) average GCP job execution time per vantage point, 4) average time spent by the collector instance to retrieve the data from a single URL, and 5) count and ratio of identified URLs with notable changes by the automated analysis process.

B. Data Set Collection and Evaluation. *Top benign domains.* To form the benign web resource data set, the information on top domain names is selected from two publicly available resources: 1) Cloudflare Radar domain rankings (top 500 domain CSV list) (Cloudflare, 2024), which represents domain rankings based on the Cloudflare service visibility, and 2) Similarweb top websites ranking (top 50 domain names) (Similarweb, 2024) claiming to represent analytics for the most visited websites for business brand and marketing competitive analysis. Both lists are merged with duplicates removed to represent a combined list of 516 unique URLs. The automated analysis identifies that most resources (75%) present the same or nearly exact content across all vantage points. This similarity is mostly related to the unified corporate identity and providing equal experience to end users globally by highly ranked entities (e.g., Microsoft, Amazon). It has to be noted that the list of top domains is produced by assessing the traffic volumes for the corresponding domain name and may include domains related to content delivery networks. These networks do not serve any content when accessed directly without request parameters and would yield empty collection results. The dynamic content, as anticipated, is primarily observed for user-generated content resources, such as, *youtube.com*, *twitch.com*, and *tiktok.com* to represent regionally-relevant material and trends.

As a first use case, it was observed that resource *yandex.net* (Internet services and search engine platform originating from the Russian Federation) displays the crawling bot detection prompt (Figure 3) only if the request originates from European IP networks. While such a broader perspective gives an additional assessment of the website's behaviour, it is not unusual to observe content providers placing restrictions or content filtering based on their policies or collected metrics. Within the current geopolitical situation (Antoniuk, 2023), such restrictions might be anticipated and may give additional perspective to the human analyst towards intelligence collection and maintaining awareness over the cyber domain.

Tor benign domains. To evaluate the implemented Tor connectivity and proxy usage functionality, the top 50 benign domains were accessed by enabling the Tor setting in the collector profile. It has to be noted, that Tor-based Internet-bound connection origin, regardless of container deployment location, are allocated from a larger pool of known and geolocated Tor exit nodes (BigDataCloud Pty Ltd, 2024). This makes such connections easily identifiable and may impose additional content access restrictions or modifications. Dynamic Tor exit node assignment implies a certain level of randomness, which may add an anticipated uncertainty to the collected results. As an example, it was observed that *google.com* displayed CAPTCHA challenge for all Tor-based connections (Figure 4), except for the collectors deployed in GCP europe-north1 and southamerica-west1. While this observation gives insights that Google may not treat all Tor-based con-

Data set name	URL count	Vantage count	Total snapshot size (average per vantage)	Average job execution time	Average single URL harvest time	Distinct URL count (ratio)
Top benign domains	516	18	2284MB (127MB)	6.17 hours	20 seconds	131 (25%)
Tor benign domains	50	18	420MB (23MB)	22 minutes	23 seconds	40 (80%)
Phishing URLs	499	18	2226MB (124MB)	1.65 hours	9 seconds	175 (35%)
Tor phishing URLs	50	18	118MB (6.5MB)	11 minutes	10 seconds	4 (8%)

Table 1: Snapshot data set overview.

Figure 3: Top benign domain data set: Yandex on *hxxps://yandex.net*.

nections equally, the exact Tor exit nodes and their geolocation are not known.

Phishing URLs. To form a malicious web resource data set, publicly disclosed phishing website URL lists were used, which may represent both ad-hoc created or compromised resources. This data set is highly dynamic since large volumes of phishing URLs are reported, tracked, and taken down or blocked daily. The data set is comprised of the following publicly available resources: 1) OpenPhish (phishing feed list) (OpenPhish, 2024b), and 2) Phishing domain database (new phishing links today list) (Krog and Chababy, 2024). The final combined and deduplicated list includes 499 unique phishing URLs. The automated analysis identifies that 65% of phishing sites serve the same content across all vantage points. Public phishing URL lists present globally identified and reported broader mass phishing campaigns targeting brands globally, instead of having a more narrow target group (i.e., spear-phishing). As anticipated, it may be observed that such mass phishing campaigns are primarily financially motivated (e.g., gambling, cryptocurrency) and collection of sensitive data (e.g., email and office authentication, streaming service accounts, and social network and communication solution credentials). The specific nature of such campaigns primarily entails cloning and impersonating existing benign resources while providing limited dynamic content. According to the OpenPhish statistics (OpenPhish, 2024a), the most targeted brands and sectors are Facebook (Meta), Bet365, Telegram, Office365, cryptocurrency services, WhatsApp, AT&T, and Outlook and other webmail providers. The initial assessment of the phishing data set confirms, that the majority of collected resources fall within these categories.

As a second use case, a Microsoft scamming website (Figure 5) serves maliciously looking content (fake Microsoft Defender threat scanner (Meskauskas, 2024)) only for the connections originating from GCP Asian networks Taiwan, Japan, and Australia except India and Singapore. This finding provides threat intelligence from a global perspective indicating that the threat actor is selective about targeting users exclusively in Asia-Pacific specifically aiming at least at Taiwan, Japan, and Australia while excluding other countries in the same region. A dedicated targeted collection process was launched (snapshot data bundled with phishing URL data set) to observe the behaviour of this specific phishing website from all 37 global GCP IP networks instead of selected 18 glob-

Figure 4: Tor benign domains data set: Google on *hxxps://google.com*.

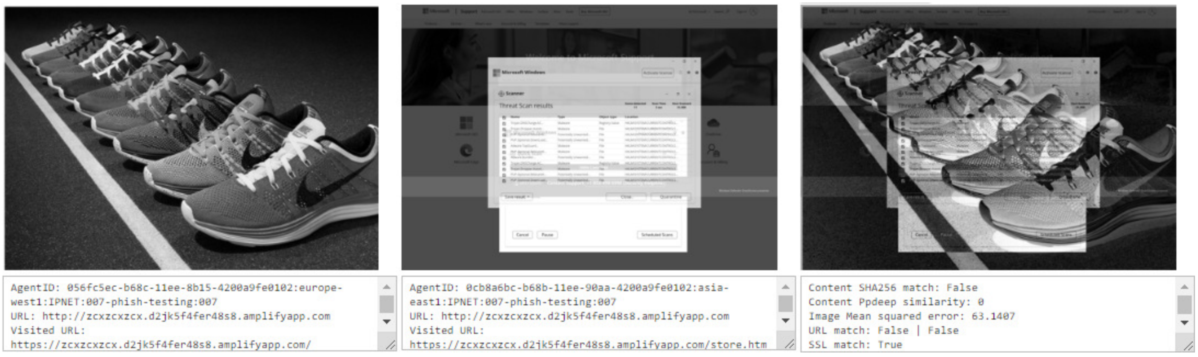

Figure 5: Phishing URL data set: Microsoft scamming on *hxxp://zcxzcxzcx.d2jk5f4fer48s8.amplifyapp.com*.

ally spread GCP networks. The collected data showed that the website displayed phishing content for connections originating from Japan, South Korea, Taiwan, and Australia, but excluded the following countries – Indonesia, Singapore, India, and Hong Kong. This intelligence may support the human analyst towards in-depth investigation and identification of threat actor *modus operandi*.

Tor phishing URLs. For data collection validation through the Tor network, 50 phishing sites from the phishing list were randomly selected. It was observed that phishing sites, hosted on major cloud providers or behind domain fronting services (e.g., Cloudflare) will impose connection checks and CAPTCHA challenges for Tor-based connections. Most of the collected data represent access restrictions across all vantage points related to the used service providers and not to the phishing resources. Such access control is primarily performed due to cases of Tor network abuse to perform malicious activities against online systems and resources. With most of the collected data containing access restriction messages or taken-down content, this data set does not present any significant findings beyond the already identified ones. However, it was observed that for the top 50 benign domains, the ratio of CAPTCHA-restricted service access was significantly smaller than for phishing websites for Tor-based connections. Although this may depend on many conditions, an assumption may be made that top benign websites have higher availability and security posture not requiring immediate potentially suspicious connection restrictions. Additionally, it may lead to the assessment, that cybercriminals are looking to maximise their impact, success ratio, and profit while investing the minimum required effort, being aware that phishing sites have a short life expectancy before being tracked and shut down.

5 CONCLUSIONS AND FUTURE WORK

This research paper delivers practical and applied contributions towards gaining expanded situational awareness of the information space (released publicly on https://github.com/lockout/b-swarm): 1) developed prototype and related code, and 2) collected snapshot data sets. It addresses the fundamental design limitations of current information space awareness solutions and proposes a novel approach towards expanding the information awareness capabilities of cyber security teams. A multiple vantage point-based approach offers a comprehensive view of the dynamic nature of the target information space. It provides contextual information, which may be ingested into the existing incident response tool set. Such augmented information space visibility may span beyond the CSIRT team requirements and apply to a wider range of sectors dealing with information space tracking, such as, law enforcement, combating financial fraud, intelligence officers, identifying disinformation, and strategic communications. The prototype implements a wide range of functionality, cloud deployment scalability and geographical distribution, and solves multiple design and operational complexities. Performed validation represents the capabilities and strengths of the developed approach by collecting, evaluating, and identifying the changes in the specified information sources. Initial data set collection and evaluation supports the hypothesis of this research and confirms the dynamic nature of content representation depending on the employed vantage point and identifying notable changes for increased context and intelligence collection.

The current prototype uses HTTP(S) protocol for content collection, one of the most widespread access methods. However, more expanded support for diverse information sources (e.g., messaging platforms, and API interaction) should be implemented and validated. Despite having already full Tor network support, the Tor network resource interaction (i.e., *.onion* sites) has been outside the scope of this research paper due to their specific access requirements. Additionally, by using non-public CERT-provided URL lists and data feeds, it may be expected that the likelihood of identifying targeted attacks would increase due to a narrower perspective on specific entities. To advance the capabilities of the solution and address the identified limitations, the future work will focus on the following key directions: 1) automatic URL data acquisition and ingestion provided by a partner CSIRT (e.g., MISP API interface), 2) snapshot creation over time for change identification, 3) machine learning-based snapshot analysis for feature and pattern detection, and 4) improved visualization interface.

ACKNOWLEDGEMENTS

This research is conducted under the Japan Society for the Promotion of Science (JSPS) International Postdoctoral Research Fellowship, supported by the KAKENHI grant, and hosted by the NAIST Prof. Kadobayashi Cyber Resilience Laboratory.

REFERENCES

Alarte, J., Insa, D., Silva, J., and Tamarit, S. (2018). Main content extraction from heterogeneous webpages. In *Web Information Systems Engineering – WISE 2018*. Springer.

Antoniuk, D. (2023). Russia wants to isolate its internet, but experts warn it won't be easy. https://therecord.media/russia-internet-isolation-challenges. Accessed: 2024/03/05.

BigDataCloud Pty Ltd (2024). TOR Exit Nodes Geolocated. https://www.bigdatacloud.com/insights/tor-exit-nodes. Accessed: 2024/01/24.

Blumbergs, B. (2023). A multiple vantage point-based concept for open-source information space awareness. Technical report, The Institute of Electronics, Information and Communication Engineers, Hokkaido, Japan.

Brand24 Global Inc. (2024). Brand24. https://brand24.com/. Accessed: 2024/01/11.

Calzarossa, M. C. and Tessera, D. (2018). Analysis and forecasting of web content dynamics. In *2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, pages 12–17.

Cloudflare (2024). Cloudflare Radar. https://radar.cloudflare.com/domains. Accessed: 2024/01/09.

Crowder, E. and Lansiquot, J. (2021). Darknet data mining - a canadian cyber-crime perspective. *ArXiv (Draft publication)*, abs/2105.13957.

Expert Insights (2024a). The Top 10 Cyber Threat Intelligence Solutions. https://expertinsights.com/insights/the-top-cyber-threat-intelligence-solutions/. Accessed: 2024/01/23.

Expert Insights (2024b). The Top 11 Dark Web Monitoring Solutions. https://expertinsights.com/insights/the-top-dark-web-monitoring-solutions/. Accessed: 2024/01/23.

Flashpoint (2024). Flashpoint Ignite Platform. https://flashpoint.io/ignite/. Accessed: 2024/01/11.

G2.com Inc. (2024). Best threat intelligence services providers. https://www.g2.com/categories/threat-intelligence-services. Accessed: 2024/04/25.

Gartner Inc. (2024). Managed Detection and Response Services Reviews and Ratings. https://www.gartner.com/reviews/market/managed-detection-and-response-services. Accessed: 2024/01/11.

Gibson, J., Wellner, B., and Lubar, S. (2007). Adaptive web-page content identification. In *Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management*, WIDM '07, page 105–112, New York, NY, USA. Association for Computing Machinery.

Google LLC (2024). Google Alerts. https://www.google.com/alerts. Accessed: 2024/01/11.

Hootsuite Inc. (2024). Hootsuite. https://www.hootsuite.com/. Accessed: 2024/01/11.

KELA (2024). KELA Cyber Threat Intelligence Platform. https://www.kelacyber.com/. Accessed: 2024/01/11.

Kiesel, J., Kneist, F., Meyer, L., Komlossy, K., Stein, B., and Potthast, M. (2020). Web page segmentation revisited: Evaluation framework and dataset. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, CIKM '20, page 3047–3054, New York, NY, USA. Association for Computing Machinery.

Kohli, S., Kaur, S., and Singh, G. (2012). A website content analysis approach based on keyword similarity analysis. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence* *and Intelligent Agent Technology - Volume 01*, WI-IAT '12, page 254–257, USA. IEEE Computer Society.

Krog, M. and Chababy, N. (2024). Phishing Domain Database. https://github.com/mitchellkrogza/Phishing.Database. Accessed: 2024/09/09.

Lawrence, H., Hughes, A., Tonic, R., and Zou, C. (2017). D-miner: A framework for mining, searching, visualizing, and alerting on darknet events. In *2017 IEEE Conference on Communications and Network Security (CNS)*, pages 1–9.

Leung, C. K., Madill, E. W., and Singh, S. P. (2022). A web intelligence solution to support recommendations from the web. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, WI-IAT '21, page 160–167, New York, NY, USA. Association for Computing Machinery.

Lewis, S. J. (2024). OnionScan. https://onionscan.org/. Accessed: 2024/01/12.

Mandiant (2024). Digital Threat Monitoring. https://www.mandiant.com/advantage/digital-threat-monitoring. Accessed: 2024/01/11.

McGahagan, J., Bhansali, D., Gratian, M., and Cukier, M. (2019a). A Comprehensive Evaluation of HTTP Header Features for Detecting Malicious Websites. In *2019 15th European Dependable Computing Conference (EDCC)*, pages 75–82.

McGahagan, J., Bhansali, D., Pinto-Coelho, C., and Cukier, M. (2019b). A Comprehensive Evaluation of Webpage Content Features for Detecting Malicious Websites. In *2019 9th Latin-American Symposium on Dependable Computing (LADC)*, pages 1–10.

Meltwater (2024). Meltwater Suite. https://www.meltwater.com/. Accessed: 2024/01/11.

Meskauskas, T. (2024). What is Windows Defender Security Center? https://www.pcrisk.com/removal-guides/12537-windows-defender-security-center-pop-up-scam. Accessed: 2024/03/05.

Nadee, W. and Prutsachainimmit, K. (2018). Towards data extraction of dynamic content from javascript web applications. In *2018 International Conference on Information Networking (ICOIN)*, pages 750–754.

Neuendorf, K. A. (2017). *The Content Analysis Guidebook*. SAGE Publications, Inc., California.

OpenPhish (2024a). Global Phishing Activity. https://openphish.com/phishing_activity.html. Accessed: 2024/01/09.

OpenPhish (2024b). Timely. Accurate. Relevant Phishing Intelligence. https://openphish.com. Accessed: 2024/01/09.

Pantelis, G., Petrou, P., Karagiorgou, S., and Alexandrou, D. (2021). On Strengthening SMEs and MEs Threat Intelligence and Awareness by Identifying Data Breaches, Stolen Credentials and Illegal Activities on the Dark Web. In *Proceedings of the 16th International Conference on Availability, Reliability and Security*, ARES 21, New York, NY, USA. Association for Computing Machinery.

Pham, K., Santos, A., and Freire, J. (2016). Understanding website behavior based on user agent. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 1053–1056, New York, NY, USA. Association for Computing Machinery.

Pujar, M. and Mundada, M. R. (2021). A systematic review web content mining tools and its applications. *International Journal of Advanced Computer Science and Applications*.

Russell, M. A. and Klassen, M. (2018). *Mining the Social Web, Third Edition*. O'Reilly Media, Inc., California.

Samtani, S., Li, W., Benjamin, V., and Chen, H. (2021). Informing cyber threat intelligence through dark web situational awareness: The azsecure hacker assets portal. *Digital Threats*, 2(4).

Shadowserver Foundation (2024). Darknet events report. https://www.shadowserver.org/what-we-do/network-reporting/honeypot-darknet-events-report/. Accessed: 2024/01/12.

Similarweb (2024). Similarweb Top Websites Ranking. https://www.similarweb.com/top-websites/. Accessed: 2024/01/09.

SISSDEN (2024). Secure Information Sharing Sensor Delivery Event Network Blog. https://sissden.eu/. Accessed: 2024/01/12.

SK-CERT (2024). Taranis NG. https://github.com/SK-CERT/Taranis-NG. Accessed: 2024/01/11.

Song, Y.-D., Gong, M., and Mahanti, A. (2020). Measurement and analysis of an adult video streaming service. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '19, page 489–492, New York, NY, USA. Association for Computing Machinery.

Suganya, E. and Vijayarani, S. (2020). Content based web search and information extraction from heterogeneous websites using ontology. *International Journal of Scientific and Technology Research*.

Takaaki, S. and Atsuo, I. (2019). Dark web content analysis and visualization. In *Proceedings of the ACM International Workshop on Security and Privacy Analytics*, IWSPA '19, page 53–59, New York, NY, USA. Association for Computing Machinery.

Thorleuchter, D. and Van den Poel, D. (2012). Using webcrawling of publicly available websites to assess e-commerce relationships. In *2012 Annual SRII Global Conference*, pages 402–410.

Wan, G., Izhikevich, L., Adrian, D., Yoshioka, K., Holz, R., Rossow, C., and Durumeric, Z. (2020). On the origin of scanning: The impact of location on internet-wide scans. In *Proceedings of the ACM Internet Measurement Conference*, IMC '20, page 662–679, New York, NY, USA. Association for Computing Machinery.

Yang, J., Ye, H., and Zou, F. (2020). pyDNetTopic: A Framework for Uncovering What Darknet Market Users Talking About. In Park, N., Sun, K., Foresti, S., Butler, K., and Saxena, N., editors, *Security and Privacy in Communication Networks*, pages 118–139, Cham. Springer International Publishing.