Plenty of Phish in the Sea: Analyzing Potential Pre-Attack Surfaces

Tobias Urban[1 3][0000−0003−0908−0038], Matteo Große-Kampmann[1 2 3], Dennis Tatang[3], Thorsten Holz[3], and Norbert Pohlmann[1]

[1] Institute for Internet-Security, Westphalian University of Applied Sciences, Germany
[2] Aware7 GmbH, Germany
[3] Ruhr-Universität Bochum, Germany

Abstract. *Advanced Persistent Threats* (APTs) are one of the main challenges in modern computer security. They are planned and performed by well-funded, highly-trained and often state-based actors. The first step of such an attack is the reconnaissance of the target. In this phase, the adversary tries to gather as much intelligence on the victim as possible to prepare further actions. An essential part of this initial data collection phase is the identification of possible gateways to intrude the target.
In this paper, we aim to analyze the data that threat actors can use to plan their attacks. To do so, we analyze in a first step 93 APT reports and find that most (80 %) of them begin by sending phishing emails to their victims. Based on this analysis, we measure the extent of data openly available of 30 entities to understand if and how much data they leak that can potentially be used by an adversary to craft sophisticated spear phishing emails. We then use this data to quantify how many employees are potential targets for such attacks. We show that 83 % of the analyzed entities leak several attributes of uses, which can all be used to craft sophisticated phishing emails.

Keywords: advanced persistent threats · phishing · OSINT · reconnaissance · MITRE · cyber kill chain · measurement study.

1 Introduction

Today, *advanced persistent threats* (APTs) represent one of the most dangerous types of attacks, as a malicious actor focuses a tremendous amount of resources into an attack on a selected target. Often such attacks utilize social engineering methods—especially *spear phishing*—to initially infect the system in the target's network (e. g., via an email attachment) [19]. For an attacker, one of the first steps is to collect as much information as possible on the target to plan their further steps (e. g., used technologies or intelligence on employees to craft spear-phishing emails) [22]. This data collection mostly happens unnoticed since the adversaries often rely on *open-source intelligence* (OSINT) data, which can be accessed by anyone. The collection of such data cannot be measured, or at least the crawling cannot be distinguished from benign traffic.

In this paper, we aim to understand and measure which publicly available data malicious actors can potentially utilize to plan and conduct their attacks with a strong emphasis on data an adversary can use to design sophisticated phishing campaigns. To the best of our knowledge, all previous work exclusively aims to detect attacks while they happen, to investigate them after the adversaries performed the attack, or to compare different APT campaigns (e. g., [4, 10, 11, 16, 21, 25]). We aim to illuminate the data publicly available to adversaries during their *initial reconnaissance phase* by analyzing a diverse set of organizations ($n = 30$). In a first step, we analyze 93 APT reports with a strong focus on the different approaches how actors get access to a company's network and which techniques they use to do so. We show that an overwhelming majority of 80 % use targeted phishing emails to lure users to unknowingly infect their system (e. g., clicking on a malicious email attachment). Based on this finding, we crawled nearly 5 million websites, analyzed more than 250,000 documents, and over 18,000 social media profiles regarding data that can be used to create personalized phishing emails. We then quantify the magnitude of publicly available data companies (unknowingly) leak and show that 90 % of them leak data that adversaries can use for the desired task. Furthermore, we show that, on average, 71 % of the employees we identified leaked several attributes that can be used for phishing attacks as we found several work-related information on them that an adversary can use in a targeted phishing campaign (e. g., supervisors, the focus of work, or the used software).
In summary, we make the following key contributions:

1. We analyze real-world APT campaigns and identify the most common tactics adversaries use during an attack and map these tactics and techniques onto the *MITRE PRE-ATT&CK* framework.
2. We measure the magnitude of data that companies (unknowingly) expose that can be used by adversaries to craft spear phishing emails. To this end, we crawl several publicly available data sources (e. g., social networks and openly available information on data leaks) and the company's infrastructure.
3. We analyze how many employees of a company leak enough attributes to write highly sophisticated phishing mails. We find that over 83 % of all analyzed companies provide rich target for spear phishing attacks.

2 Background

Before describing our approach to determine the Internet-facing attack surface of a company, we provide background information necessary to follow our method.

2.1 Advanced Persistent Threats

Advanced Persistent Threats (APTs) are attacks executed by sophisticated and well-resourced, often state-sponsored, groups. In contrast to other adversaries, the actions of these groups are often politically motivated, but they also aim to

achieve an economic gain from their efforts. They target every business sector and design their attacks in a way that remains undetected for a very long time, in contrast to e. g., ransomware attacks. While common adversaries often choose their target by chance, APT threat actors typically target a specific company or business sector and invest a lot of time and energy until they eventually successfully obtain access. To enable such attacks, these groups utilize traditional attack vectors like social engineering (e. g., spear phishing), but also sometimes collect information by physically infiltrating the target companies (e. g., dumpster diving).

Spear Phishing In computer security, *phishing* describes the act when an adversary impersonates a trusted entity (e. g., a popular brand or bank) with the intent to trick users into exposing personal data (e. g., credit card numbers or credentials) or spreading malware via malicious attachments or links [27]. While these attempts commonly target tens of thousands of users, spear phishing targets a limited group of people (e. g., few people within a company or one research group at a university) or sometimes only a single person (e. g., the head of a department). As these phishing campaigns target specific individuals, adversaries can craft emails in a way that they perfectly suit the audience (e. g., by personal salutations in emails) and are often successful [3]. Adversaries persistently exploit phishing and spear-phishing because exploiting humans is often easier compared to bypassing technical security measures [8].

Cyber Kill Chain The term *cyber kill chain*, coined by *Lockheed Martin* [22], is referring to the military term "kill chain", and both terms describe the structure of an attack. However, the cyber kill chain is often used defensively in incidence response or digital forensics to model the attack performed by an adversary [32]. The chain maps each attack to seven phases that can be grouped into two sections, based on the stage of the attack. First, the attacker profiles the target (1: "Reconnaissance"), then she builds the malware used to infiltrate the target (2: "Weaponize"), which is then transferred to the target (3: "Delivery"). Afterwards, the attacker triggers the payload (4: "Exploitation"), and installs a backdoor and establishes a persistent bridgehead into the target's network (5: "Installation"). Finally, she builds a C&C infrastructure to communicate with the infected hosts (6: "Command and Control") and performs the malicious actions of desire (7: "Act on Objective"). Phases one, two, and sometimes three are referred to as the *pre-attack stage*, while the remaining stages are referred to as *attack stage*. In this work, we only focus on the pre-attack phase and specifically on the *reconnaissance* phase.

2.2 MITRE Framework

The *MITRE* cooperation created and still maintains the *PRE-ATT&CK* [31] and *ATT&CK* ("Adversarial Tactics, Techniques, and Common Knowledge") [30] frameworks. The platform collects and systematizes techniques and tactics of real-world adversaries which were obtained from several attacks with the goal

that companies can learn from those attacks and improve their security concepts. All collected events are organized in different categories based on their appearance in the cyber kill chain [32]. The framework assigns a unique four-digit identifier to each category and technique so that it can be referenced easier (e. g., T1189 in TA0001).

The *PRE-ATT&CK* framework is designed to focus on the stages that usually occur before the attack is performed. For example, this includes choosing a victim, collecting data on the victim, or setting up the infrastructure needed to perform the attack (e. g., implementing the needed malware or setting up the C&C infrastructure). While the *ATT&CK* framework often contains very technical and specific information, the *PRE-ATT&CK* framework is often more general as it is by nature not as easy to determine which actions the actor performed. For example, if an adversary used a specific type of malware, one can analyze it and draw conclusions based on the sample. However, one cannot undoubtedly determine why a specific employee was phished based on technical data. Appendix A provides an overview of the pre-attack techniques and tactics of the framework that are relevant for our work.

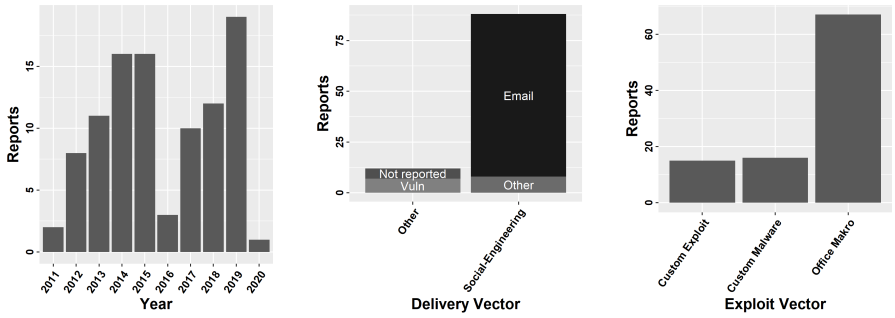
3 Advance Persistent Threat Analysis

In this section, we provide the results of an analysis of 93 real-world APT reports we studied. More specifically, we perform a technical mapping of these reports onto the *MITRE PRE-ATT&CK* framework.

3.1 APT Report Analysis

As noted above, the *cyber kill chain* describes the multiple stages of an attack. To the best of our knowledge, no systematic research went into the analysis of the early steps of this process in which adversaries collect data on their victims to plan and initiate their campaigns. To close this gap and to gain further insights into the methods adversaries use, we manually analyze 93 openly available reports and technical blogs on APT campaigns with a strong emphasis on these steps (i. e., the *reconnaissance* phase). We use these reports as security companies, in contrast to academic researchers, often have unique insights into these APTs, especially in terms of incident response. In total, 40 different companies provide the reports of the APTs (e. g., *Symantec*, *Kaspersky Lab* or *Palo Alto Networks*).

Overall, the analysis of the APT reports in our dataset attributed them to 66 different malicious actors. In 32 cases, the report does neither identify nor disclose the actor. We argue that this broad distribution of actors allows us to draw a more generalizable conclusion on the methods used by them. According to the analyzed reports, the attacks in our dataset happened between 2011 and 2020. Figure 1a provides a detailed overview of the number of analyzed APTs each year. Two reports do not report on the year in which the APT happened. Nearly all reports lack information about the reconnaissance phase (91 %). This



(a) Reported incidents by year.

(b) Reported delivery vectors.

(c) Reported exploits.

Fig. 1: Overview of the analyzed APT reports.

knowledge gap probably roots in the fact that this cannot easily be analyzed, especially from an incident response point of view. However, in 41 % of the cases, the target group (e. g., business sector or company type) could be identified. In rare cases where data on the reconnaissance phase is present, the actor used data publicly available ("Open-Source Intelligence" or OSINT) to identify promising targets for further steps. Reportedly, an overwhelming majority (88 %) of all APTs used social engineering techniques to deploy their attack tools (e. g., malware) in companies' infrastructure. Furthermore, email seems to be the most popular way to get in touch with the victim (80 %). Other means of communication with victims include social media (3 %), phishing websites (4 %), or SMS (1 %). In the cases where the malicious actor did not rely on social engineering, the attackers abused vulnerabilities collected from public data on the companies infrastructure (4 %), data collected from other services (3 %) or the reports only hold vague or inconclusive data on the delivery phase (e. g., "*banks in Russia*"). Figure 1b provides an overview of the reported delivery vectors. In the exploitation phase, the actors mostly used *Microsoft Office* documents that contained malicious macros (69 %). In the remaining cases, the adversaries either used case-specific malware or exploits they tailored for a product the company uses, as shown in Figure 1c.

In summary, there is a lack of knowledge of how attackers collect the data on their victims. However, in the early stages of an attack, social engineering is the most common attack vector. Most malicious actors use email (e. g., spear phishing) as a primary channel to get in touch with their targets. In these emails, they utilized office documents that contain malicious macros to infect the user's system. While the analysis of the APT campaigns yielded the most common ways of how adversaries try to infiltrate companies, it is unclear which data they used to perform these attacks, or how and where they acquired it.

3.2 MITRE PRE-ATT&CK Analysis

The *MITRE PRE-ATT&CK* framework consists of 15 groups that describe different stages of the pre-attack phase [31]. In this work, we focus on data that can be publicly accessed by an adversary that provides her insights on the target company, the used infrastructure, and employees of the company. As previously described, an adversary can use this data to perform sophisticated social engineering attacks, like spear phishing.

We analyzed the framework to test which of the listed tactics can be analyzed using publicly measurable data using only non-offensive collection methods, which we used as a basis to design our measurement. Adversaries probably also use offensive tools (e. g., vulnerability scanners or buying information leaks online) to collect information, but due to ethical considerations, we renounce to use such tactics. Three computer security experts with a strong background in online measurements or threat intelligence (i. e., the first three authors of this paper) analyzed the framework. The experts were instructed to analyze all techniques and tactics in the framework and assessed whether they are publicly measurable using only non-offensive collection methods. The final inter-rater agreement whether a technique is measurable in our setting or not shows substantial agreement (Fleiss' Kappa: $\kappa = 0.73$; agreement $> 90\,\%$). In the rare cases of discrepancies, the option that got the majority was selected to resolve such matters.

The results show that one cannot measure several techniques using data that is publicly available. As a result, we only consider four of the 15 groups (i. e., *Technical Information Gathering, People Information Gathering*, *Organizational Information Gathering*, and *Technical Weakness Identification*) in our analysis (see Table A in Appendix A). The remaining groups are not measurable without internal insights of the adversary. For example, an analyst could measure the *Target Selection* or *Adversary OPSEC* phase if she infiltrates the adversary's internal infrastructure and monitors all events. We consider this to be out-of-scope as (1) we want to identify protection mechanisms for companies, and only highly specialized experts can perform such infiltration and (2) such penetration is likely in a legal gray (if not black) area. The techniques that we exclude are often either (1) described too general in the framework (e. g., "Conduct active scanning"), (2) out-of-scope because we refrain from using offensive technologies (e. g., "Conduct social engineering "), or (3) can be done reliably in an automated fashion (e. g., "Identify supply chains").

Summary The analysis of the APT campaigns revealed that social engineering enables most of them, commonly conducted via spear-phishing emails. However, the reports could only rarely reconstruct which data attackers used to write the emails. The MITRE PRE-ATT&CK lists several techniques adversaries can use to collect such data. However, several of these described techniques are very broad, cannot be measured straightforwardly, and are sometimes not under the control of the company. Therefore, the question arises to what extend companies (unknowingly) expose such data.

4 Measuring Data Collection Opportunities

Based on the analysis of the framework presented in the previous section, we developed tools to collect the data types that a malicious actor can use to craft sophisticated spear-phishing emails as they are the most prevalent intrusion vector. We used two different crawling approaches to collect data for each company: (1) Analyzing sources directly maintained by the companies (e. g., websites) and (2) information present on third-party websites but that the company directly or indirectly provides (e. g., job postings or social media profiles).

4.1 Data Description

In our analysis, we perform an in-depth analysis of 30 entities (27 companies, two government agencies, and one non-profit organization). For the sake of simplicity, we use the term *company* for these 30 entities in the following. To choose these companies, we used a list of large, international companies and chose 27 from this list, with an emphasis on banking and e-payment companies. We focused on one sector as the described malicious actors tend to attack financial institutions or large organizations. However, the chosen companies are active in a variety of industry sectors and are of different sizes regarding revenue and number of employees. On average, the revenue of the analyzed companies is 60 billion USD (min: 27 million USD; max: 790 billion USD), and they employ 55,484 people (min: 49; max: 375,000). We took these numbers from the official figures the companies provided for 2018. Ten of the companies are active in the "banking" or "digital payment" sector (37 %), while the others are distributed over eleven sectors (e. g., "Food" or "Aeronautics").

For ethical reasons, we refrain from naming any of the companies and will use pseudonyms for all companies in the remainder of this work (i. e., *Comp. #X*). In our measurements, we used no legal or ethical questionable tools and only accessed data that is publicly available. More specifically, we use in this study three different types of data sources to measure the pre-attack surface of a company: (1) data the company (unknowingly) provides, (2) data publicly available through social media sites, and (3) data leaked in known data breaches. An extended ethical discussion is presented in Section 7.

4.2 Data Collection

As previously mentioned, we rely for our analysis on "Open Source Intelligence" (OSINT) data, i. e., data sources that are publicly available. In the following, we describe the used data sources in more detail.

Company Controlled Entities To crawl each companies' infrastructure, we built a crawler that we initialize with 1 to n domains owned by a company (*seed domains*). If possible, we read the TLS certificate present on these domains and

try to identify further domains that can be protected by this certificate (i. e., *Subject Alternative Name* (SAN) and multi-domain SSL certificates). Furthermore, we perform a DNS enumeration to discover further domains and infrastructure operated by the company. After identifying the "landing pages" of all domains associated with a company, we visit each page and recursively all first-party links occurring on each website to a certain depth ($n = 6$). Hence, we try to visit every single webpage publicly linked by a company. Using this approach, we miss resources that are only available if the user has a specific link to the resource.

Analyzing Metadata Most popular file types offer proprietary options to store additional information regarding the file ("metadata"). Such metadata, for example, includes authors of the document, the software used to create the document (e. g., `pdfTeX-1.40.17`), email addresses of the author, or its title. From an adversary's point of view, this information may provide specific insights into the victim. For example, the authors of a document, in combination with its title/content, can be used to craft specific phishing emails for a single or small group of users. With a given type of software, the adversary might also be able to attach a file that exploits a specific bug in that software to infect the user's system. We only used email addresses whose domain part's effective top-level domain (eTLD) +1 fit the eTLD+1 of the seed domain(s) we analyzed. For example, if our crawler scanned `foo.com` and extracted the email address `smith@bar.com` in one file, we dropped the file. Aside from metadata analysis, we identified emails by analyzing the content of websites and documents. For our study, we download all files that we find during the crawling process and extract the metadata. Overall, we analyze 36 different file types. These files includes `.pdf` files, office documents (e. g., `*.docx` or `*.odt`), and various image types (e. g., `*.png` or `*.jpeg`). If a document contains an author or other personally identifiable information (e. g., email addresses or names), we map them to other properties (e. g., used software). More specifically, we create relations between users, the software they use, and possible topics on which they work. For example, if we identified a *Microsoft Office v1.0* document written by two authors (*Alice* and *Bob*) with the title *World Peace—Status Quo and Outlook*, we can conclude that both worked on "World Peace" using Microsoft Office.

Company Infrastructure We mainly focus on the vulnerability of companies towards social engineering attacks, especially spear phishing. Thus, we describe our measurements regarding parts of the companies infrastructure that might be abused by an adversary for this specific kind of attack. Adversaries might use so-called *homoglyph domains* (e. g., changing 'l' to 1) to trick employees into visiting them with the belief to navigate on the secure infrastructure of the company (but an adversary, of course, controls this infrastructure). We perform a simple *cybersquatting* detection by creating a list—based on the seed domains—of URLs that "look" similar to humans by applying techniques like homoglyphs, simple permutations, or by using different eTLDs. Afterward, we test if any of these URLs exist and try to assess who registered them. We use `whois` requests and data from SSL certificates to identify the registering organization.

Furthermore, we aim to identify isolated components in a company's infrastructure that is not connected to any other entity of the company's infrastructure. Examples for connections between the components are hyperlinks or shared IP addresses. Such isolated components could be legacy systems running without the direct knowledge of the responsible administrators, might be used as test systems, or in case of domains, might be run by an adversary in preparation of an attack. For all domains registered by a company (excluding homoglyph domains), we tested if the websites use trustworthy SSL certificates (e. g., not expired ones). If companies use certificates that are not trustworthy, adversaries might be able to intercept or eavesdrop the connection, which allows them to collect sensitive data. Finally, we check whether companies register domains with names similar to their original domain. Domain parking can be used to register domains up front before a service is run on the domain. Furthermore, a service provider can use this practice to avoid "domain drop catching". Domain drop catching is a (malicious) practice to registers a domain right after it expired and then to use it for different purposes [18, 26]. As users usually do not know when and if domains expire, they will still visit the domain and might be exposed to malicious content.

Social Media Employment-oriented social media platforms, like *LinkedIn*, are commonly used by millions of people [20]. As these platforms are supposed to maintain business relationships, they can also be abused by adversaries to collect intelligence on a company [1]. This data might provide several details about the internal workings of a company, and its employees and their careers, contacts, or supervisors. Furthermore, companies do not have real control over which data is shared and posted on such platforms, and adversaries might use these sites to get in touch with the employees, undetected by any security mechanism of the company.

In this work, we use data obtained from various sources (e. g., different APIs). Some of these APIs are deprecated as of July 2020 but were still available when we collected our dataset. One example was the *LinkedIn* API that allowed to crawl user data based on an email address (i. e., *https://www.linkedin.com/sales/gmail/profile/viewByEmail/mailaddress*. The malicious actor could use this endpoint to determine whether an identified email address had a corresponding profile. To mimic the potential workflow of an adversary, we utilized search engines to perform site-specific searches (e. g., *site:linkedin.com <COMPANYNAME>*). To further enrich our dataset, we utilized publicly available tools that automate the crawling process of social media sites (e. g., *CrossLinked* [24]).

Data Leaks Finally, adversaries may utilize data from previous data breaches to prepare their attack. In this work, we use the *Have I Been Pwned* API [15] to test if a company ever leaked data that can be used in another attack on that company. The API exposes data leaks from over 400 websites and over 110.000 "pastes". In this case, pastes are indications of data leaks in which the adversaries provides examples of the acquired data to prove that she actually got access to

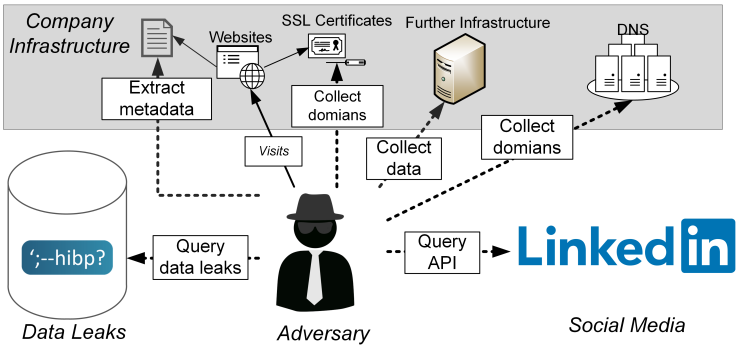


Fig. 2: Overview of our data collection approach.

sensitive information. The API does not directly provide any of the breached data but returns categories of data that the leak contained. For example, if one provides an email address, the API will return data types that were leaked along with the address in different data breaches (e. g., `foo@bar.com`) results in `dates of birth`, `employers`, and `job titles`). Figure 2 provides an overview of the three types of data sources we analyzed.

4.3 Identified Data

Data Crawled from Companies' Infrastructure In total, we scanned 30 entities and identified 492 domains (eTLD+1 and suffix) operated by them. Furthermore, we identified 18,873 employees, of which 8,994 appeared in data leaks, or they provided valuable data in public social media profiles.

Metadata Analysis During the measurement, we visited 4,912,938 distinct webpages and extracted metadata from 271,124 documents. Table 1a provides an overview of the identified data types identified based on the metadata of files we found on the crawled webpages. The `min`, `max`, and `mean` value describe how many instances were obtained for each company (e. g., we identified the names of 634 employees of a company). Ninety percent of the analyzed companies leak the names of their employees. Overall, we identified 22,361 email addresses, of which 6,335 were exclusively exposed via metadata (intentionally or unintentionally). As we extracted them from metadata, this might also provide insights to the adversary on which projects they work on (e. g., based on the file's content). Aside from names, the email address is essential as actors can use them to get in touch with potential victims. Almost three-quarters of all companies in our dataset leaked an employee's email address.

Furthermore, once the malicious actor understands the structure of a company's email addresses (e. g., `lastname@foo.com`), she can presumably make educated guesses on the local parts of further addresses if she knows the employees' names. In our dataset, the amount of identified email addresses would increase,

on average, by 52 % for each company. 81 % of the companies exposed third parties they work with (i. e., collaborating partners that created a document). The three named data types can be used to craft user/team specific spear-phishing campaigns. For example, an adversary could impersonate a partner the employee worked with. Aside from personal data, the metadata of a file might expose intelligence on the inner workings of a company. In our dataset, 90 % of the companies leaked the software they used to create a document, and almost two-thirds leaked data paths they use in the company to store documents (e. g., `Z:\Project_X\Results`). An attacker can use this information when preparing for the attack (e. g., zero-day exploits for the used software).

Company Domains Table 1b presents potential information on the infrastructure that an actor can collect and later use for an attack. Furthermore, it provides hints that adversaries already actively make use of homoglyph domains. The most troubling finding of this measurement is that for 18 (60 %) of the analyzed companies, an adversary actively abuse a homoglyph domain, at the time of our crawl. Note that we only counted domains for which we find a substantial string similarity of more than 95 %, and therefore, our results can be seen as a lower bound. For one company, we found twelve active domains of this type (avg: 4.5). The presence of such domains indicates that adversaries are likely already actively trying to misguide users or employees of such services (e. g., password phishing). However, we also observed that some companies are aware of this endangerment and acquire some of these domains and "park" them for brand protection purposes as a kind of proactive defense.

Often websites or other services are connected by various mechanisms (e. g., hyperlinks or services that share the same IP address). In our dataset, half of the companies operate services that have no connection to others. These domains might pose a problem if the companies no longer maintain them and, therefore, could be less protected (e. g., legacy interfaces). On the other hand, these services might not pose a problem at all because the companies are fully aware of them. We found that eight entities (26 %) operated domains that use an invalid or outdated certificate. An adversary might abuse these by intercepting the TLS encryption to such domains to collect more data on users or employees, whoever primarily uses these services. All of these companies operated at least one isolated domain (avg: 5) that uses an expired or otherwise untrusted certificate, which reinforces the assumption that the companies no longer maintain them.

Data Available in Data Leaks and in Social Media Profiles In addition to the analysis of the companies' infrastructure and data they expose via metadata, we analyzed if the business accounts of employees (e. g., email addresses) occur in publicly known data breaches (see Section 4.2). For each company in our dataset, we found data on at least three of the identified employees (max: 1,102). In absolute numbers, 11 companies (46 %) leaked data of less than 30 employees, and only four (12 %) did not leak data on any employee that we identified. In relative numbers, two-thirds (20) of the analyzed companies leaked data of more

Table 1: Overview of the data extracted from the company's own infrastructure.

(a) Overview of the identified information. The `min` and `mean` values exclude companies that did not provide the type of data.

Data type	aff. comp.	min	max	mean
Names	90 %	7	634	227
Mail addresses	71 %	1	96	19
Third Parties	81 %	1	53	15
Software	90 %	5	205	71
Path	65 %	1	30	7

(b) Overview of the identified infrastructure information.

Type	min	max	mean
Homoglyph Domains	1	12	4.5
Parked Domains	5	379	41.6
Isolated Domains	7	89	27.3
Untrusted Certificates	1	20	5.9

than one-third of the identified employees (max: 88 %). We found no statistical significance between the amount of identified emails and the amount of leaked data (ANOVA-Test p-value ≈ 0.03). Hence, companies that expose more emails are not automatically likely to be present in more data leaks. As this might seem to be counter-intuitive, it hints that some companies have policies in place to reduce the potential of such data leaks (e. g., awareness campaigns). Overall, the analyzed data leaks include 65 different data categories. The categories range from personal data (e. g., *credit status information*, *government issued IDs*, or *device usage tracking data*) over data directly tied to the employee's professional live (e. g., *job titles*, *employers*, or *occupations*) to other data an adversary could use to plan an attack (e. g., *instant messenger identities* or *password hints*). The category of a data leak shows a statistical correlation with the number of instances that this data is leaked (ANOVA-Test p-value < 0.0001). Hence, some data types leak more often than others.

Figure 3 shows the type of leaked data for each company. The heatmap highlights the ratio of identified leaks with the email addresses that we could identify. The figure only lists the top 15 categories, which account for 89 % of all leaking instances. It shows that some companies leak excessively more data than others (ANOVA-Test p-value < 0.0001) but that there is no dominating data type that is leaked. In our dataset, the top leaked types are passwords (10 %), phone numbers(8 %), and geolocations (7 %), excluding the name and email addresses of the users that the adversary needs to identify an employee. The biggest challenge with data actors collect from data leaks is that companies have virtually no measure to delete it. Furthermore, in none of the cases, it was the company itself that leaked the data but other platforms on which the employees registered to use the service, using their business email address. Hence, one solution could be to raise awareness with employees only to use the work email if necessary and to provide as little information as possible when using the respective services.

Summary In this section, we demonstrated that companies excessively leak data that provides insights into their inner workings or on the employees of the com-

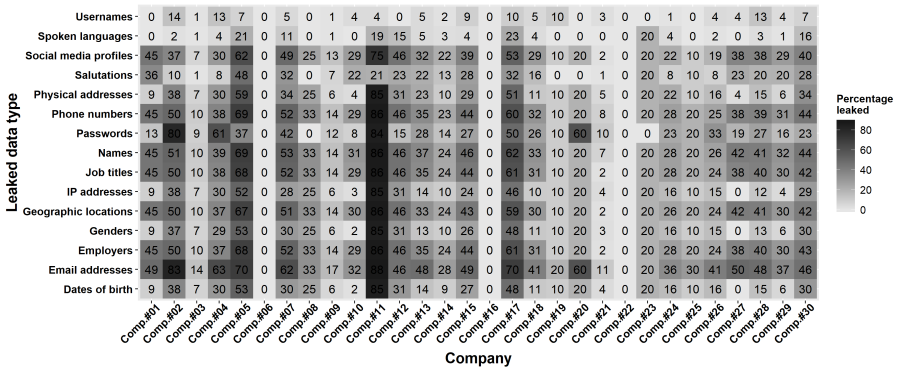


Fig. 3: Overview of the data extracted from other data sources.

panies. However, it is not clear whether an adversary can meaningfully combine this data to plan further steps link designing successful phishing campaigns.

5 Assessing Potential Phishing Targets

Based on the insights of our study, we now introduce a metric to asses the likeliness that an employee serves as a good spear phishing target. The presence of data that we identify in this work does not necessarily pose a security problem per se. Each data point on its own is properly not problematic if obtained by an adversary, but taken together, they reveal intelligence that can be used, for example, to craft personalized phishing emails. Therefore, it is important to analyze and interpret the collected data.

5.1 Identifying Potential Phishing Targets

We now numerically analyze whether users are promising targets for spear phishing attacks from an adversary's point of view. In this work, we only analyze technical aspects and not the personal experience of each person, which is out-of-scope of this work but an essential aspect if someone falls for a phishing attempt [17]. Previous work that analyzed the effectiveness of spear phishing found that sources that impersonate an individual from the victim's company (e. g., from the human resources department) are quite effective [3,12]. The work shows that 34 %–60 % of all participants clicked on a link in such email. Therefore, we assume that if we could identify other persons working in the company and especially if they are working together (e. g., co-worker, supervisor, or team member), a phishing attempt might be more effective. Furthermore, if the adversary knows the software used by the victim, she can craft and append an exploit specifically for the used software to the email, which increases the chances of a successful compromise. Thus, we also consider the used technology of each employee as an essential aspect.

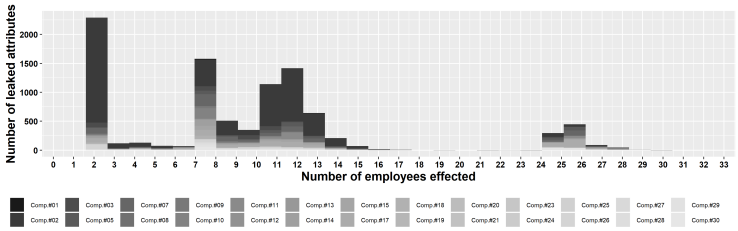


Fig. 4: Number of leaked attributes for each user in our dataset.

5.2 Spear Phishing Targets in the Wild

Figure 4 shows the number of attributes leaked for each employee in our dataset. Most employees only leak two categories of information (i. e., their name and email address), which we use to identify users of a company. For those individuals, it is not possible to craft targeted phishing mails (at least using our data). However, we identified 5,910 (62 %) employees that leak between seven and 15 attributes. Those employees are employed at 25 different companies (83 %). 878 (9 %) employees leak between 24 and 28 attributes and are employed at 18 companies (60 %). From both employee groups, an adversary can potentially pick several attributes and craft highly specific spear phishing mails. Only four (13 %) companies in our dataset do not leak any additional data on their employees, aside the name and email address). For all of these companies, we identified relatively few employees. One reason for this is that most of these companies are active in business fields with no (public) customer interaction (e. g., investment banking). The results show that almost all companies in our dataset provide a considerable pre-attack surface to motivated attackers. In absolute numbers, companies that leaked more data also provide more targets for an attacker (ANOVA-Test p-value ≈ 0.001). However, taking the relation between the amount of identified emails and leaked data into account (similar to Figure 3), we did not find a correlation. Hence, leaking more data does not necessarily mean that the adversary can identify more spear phishing targets. Our results show that the OSINT data sources that we utilized provide a rich data source from which threat actors can profit.

6 Related Work

Previously research on effective APT detection or prevention mostly focused on detecting them at and/or after the "Delivery" phase in the cyber kill chain [32]. APT detection is highly complicated as information from many sources (e. g., human behavior, intrusion detection systems, or system logs) have to be combined to make an informed decision. Machine learning approaches were studied to process this enormous amount of data to detect APTs [2, 10, 16, 21]. Furthermore, more heuristic solutions like correlating events [25], defining detection

rule sets [33], detecting misuse on application level [23], or annotating security events [11] have been proposed. Similar to our APT analysis, Lemay et al. [19] analyzed APT reports. In their work, they provide a summary of 40 analyzed APT reports. A large number of different works focus on the technical detection of spear-phishing emails, content analysis of phishing websites, and the detection of such websites based on their URLs (e. g., [13,14]). Furthermore, various studies analyzed human aspects to understand why spear-phishing attacks are successful (e. g., [3,12]). Finally, several papers systematize the extensive research that was conducted in this area (e. g., [5–7]). Our work differs from these approaches as we solely focus on the very first steps adversaries take when they plan their malicious actions, the *reconnaissance* phase. To the best of our knowledge, we are the first ones to show the variety and magnitude of information companies (unknowingly) provide that can be abused by adversaries to perform spear-phishing attacks. Furthermore, we do not aim to understand the effectiveness of specific phishing campaigns, but provide insights on how companies expose data.

7 Ethical Consideration

For this study, we gathered and analyzed sensitive information on companies and employees, thus individual persons. Our research institution does not require approval for this type of study, nor does it provide an Institutional Review Board (IRB). Nevertheless, we took strict ethical considerations into account. Additionally, we followed the research community's standard guidelines to protect those whose data was collected and the infrastructure of the services we use. A recent court ruling, according to the Electronic Frontier Foundation, found that "*automated scraping of publicly available data is unlikely to violate the Computer Fraud and Abuse Act (CFAA)*" [9]. As a general rule, the collection of personal information requires user consent; however, there are exceptions for cases where this is not practical. In our case, publicly available sources are the basis of the collection of information. By nature of our analysis, we cannot preempt to process personal data. We also want to highlight that none of our collections tools use any questionable tools to identify systems or persons. We do not perform any kind of penetration testing to collect data and send all requests at a courteous rate. The gathered data was collected for scientific purposes only, and we only disclosed it to the involved companies. To protect the collected personal data, we took additional safety measures: We encrypt the raw files for storage and delete unused samples and data.

8 Discussion and Conclusion

Our approach comes with limitations that need further clarification. The most decisive one, from a researchers' perspective, is that there exists no ground truth for our collected data. Hence, it implies that we do not know if adversaries profit from the data sources that we utilize to plan their actions or if all companies are unaware of the leakage of such data. However, the analyzed APT reports and

several other sources (e. g., [22,28,29]) indicate that adversaries make excessive use of OSINT data, and even if companies are aware of the leakage of data, it might still be used by the adversaries. There is very little raw data available on incidents especially how the attackers infiltrated their victims. Furthermore, to build a ground truth for our research, one would need to impersonate the malicious actor while she plans her attack, which is ethically not tenable. With a company's consent, we could perform an awareness phishing campaign using the identified data. However, previous work already performed similar studies and demonstrated that they are often successful (see Section 6). With our work, we do not aim to determine the exact data used by adversaries in each attack, which is probably impossible in an automated fashion, but we demonstrate the sheer scale of data leaked by companies. Our results highlight that all analyzed companies provide a large attack surface to adversaries that is not monitored or protected by state-of-the-art security solutions. Furthermore, this data leakage is not always under the control of the companies, nor is it always possible to revert the leakage. Therefore, there is no clear path how to circumvent this type of leakage or straightforward countermeasure. It is quite hard to successfully prevent attacks on third party providers or reduce attack surfaces and therefore to apply countermeasures. One way to decrease the potential damage by these data leaks is to raise awareness with employees that this kind of data is regularly abused by adversaries and that the principle of "data economy" should be followed. Actionable tools to counter misuse of our analyzed data sources can be to wipe the metadata from all uploaded files, to continually monitor data leaks if they include passwords or other personal data of employees, or to increase awareness in a way that empowers employees not to provide too much work-related information on social media platforms.

Acknowledgment

This work was partially supported by the Ministry of Culture and Science of North Rhine-Westphalia (MKW grant 005-1703-0021 "MEwM"), the federal Ministry of Research and Education (BMBF grant 16KIS1016 "AWARE7"), and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2092 CaSa – 390781972. We would like to thank Sweepatic NV—a cybersecurity company which maps, monitors and manages attack surfaces—for their support and access to their technology.

References

1. Balduzzi, M., Platzer, C., Holz, T., Kirda, E., Balzarotti, D., Kruegel, C.: Abusing Social Networks for Automated User Profiling. In: Symposium on Recent Advances in Intrusion Detection. (RAID'10) (2010)
2. Barre, M., Gehani, A., Yegneswaran, V.: Mining Data Provenance to Detect Advanced Persistent Threats. In: Proceedings of the 11th International Workshop on Theory and Practice of Provenance. TaPP' 19, USENIX Association, Berkeley , CA, USA (2019)

3. Caputo, D., Pfleeger, S., Freeman, J., Johnson, M.: Going Spear Phishing: Exploring Embedded Training and Awareness. IEEE Security & Privacy **12**(1), 28–38 (Jan 2014). https://doi.org/10.1109/MSP.2013.106
4. Chen, P., Desmet, L., Huygens, C.: A Study on Advanced Persistent Threats. In: Proceedings of the 15th International Conference on Communications and Multimedia Security. pp. 63–72. CMS '14, Springer International Publishing, Cham (2014). https://doi.org/10.1007/978-3-662-44885-4_5
5. Chiew, K., Yong, K., Tan, C.: A survey of phishing attacks: Their types, vectors and technical approaches. Expert Systems with Applications **106**, 1–20 (2018). https://doi.org/https://doi.org/10.1016/j.eswa.2018.03.050
6. Das, A., Baki, S., El Aassal, A., Verma, R., Dunbar, A.: SoK: A Comprehensive Reexamination of Phishing Research from the Security Perspective. IEEE Communications Surveys Tutorials (2019). https://doi.org/10.1109/COMST.2019.2957750
7. Dou, Z., Khalil, I., Khreishah, A., Al-Fuqaha, A., Guizani, M.: SoK: A Systematic Review of Software-Based Web Phishing Detection. IEEE Communications Surveys Tutorials **19**(4), 2797–2819 (2017). https://doi.org/10.1109/COMST.2017.2752087
8. Ferreira, A., Vieira-Marques, P.: Phishing Through Time: A Ten Year Story based on Abstracts. In: Proceedings of the 4th International Conference on Information Systems Security and Privacy. pp. 225–232. ICISSP '18, INSTICC, SciTePress, Setúbal, Portugal (2018). https://doi.org/10.5220/0006552602250232
9. Fischer, C., Crocker, A.: Victory! Ruling in hiQ v. Linkedin Protects Scraping of Public Data, https://www.eff.org/deeplinks/2019/09/victory-ruling-hiq-v-linkedin-protects-scraping-public-data
10. Ghafir, I., Hammoudeh, M., Prenosil, V., Han, L., Hegarty, R., Rabie, K., Aparicio-Navarro, F.J.: Detection of Advanced Persistent Threat Using Machine-Learning Correlation Analysis. Future Generation Computer Systems **89**, 349–359 (2018). https://doi.org/https://doi.org/10.1016/j.future.2018.06.055
11. Gianvecchio, S., Burkhalter, C., Lan, H., Sillers, A., Smith, K.: Closing the Gap with APTs Through Semantic Clusters and Automated Cybergames. In: Proceedings of the 15th International Conference on Security and Privacy in Communication Systems. pp. 235–254. SecureComm '19, Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-37228-6_12
12. Halevi, T., Memon, N., Nov, O.: Spear-Phishing in the Wild: A Real-World Study of Personality, Phishing Self-Efficacy and Vulnerability to Spear-Phishing Attacks. SSRN Electronic Journal (01 2015). https://doi.org/10.2139/ssrn.2544742
13. Han, Y., Shen, Y.: Accurate Spear Phishing Campaign Attribution and Early Detection. In: Proceedings of the 31st ACM Symposium on Applied Computing. pp. 2079—-2086. SAC '16, ACM Press, New York, NY, USA (2016). https://doi.org/10.1145/2851613.2851801
14. Ho, G., Sharma, A., Javed, M., Paxson, V., Wagner, D.: Detecting Credential Spearphishing in Enterprise Settings. In: Proceedings of the 26th USENIX Security Symposium. pp. 469–485. USENIX Sec '17, USENIX Association, Berkeley , CA, USA (2017)
15. Hunt, T.: Have I Been Pwned: API v3 (2020), https://haveibeenpwned.com/API/v3, accessed: 2020-04-15
16. Kumar, G.R., Mangathayaru, N., Narsimha, G., Cheruvu, A.: Feature Clustering for Anomaly Detection Using Improved Fuzzy Membership Function. In: Proceedings of the 4th International Conference on Engineering & MIS. ICEMIS '18, ACM Press, New York, NY, USA (2018). https://doi.org/10.1145/3234698.3234733

17. Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L.F., Hong, J., Nunge, E.: Protecting People from Phishing: The Design and Evaluation of an Embedded Training Email System. In: Proceedings of the 25thACM SIGCHI Conference on Human Factors in Computing Systems. pp. 905—914. CHI '07, ACM Press, New York, NY, USA (2007). https://doi.org/10.1145/1240624.1240760
18. Lauinger, T., Chaabane, A., Buyukkayhan, A.S., Onarlioglu, K., Robertson, W.: Game of Registrars: An Empirical Analysis of Post-Expiration Domain Name Takeovers. In: USENIX Security Symposium (2017)
19. Lemay, A., Calvet, J., Menet, F., Fernandez, J.M.: Survey of Publicly Available Reports on Advanced Persistent Threat Actors. Computers & Security **72**, 26–59 (2018). https://doi.org/10.1016/j.cose.2017.08.005
20. LinkedIn Corporation: Statistics (2020), https://news.linkedin.com/about-us#statistics, accessed: 2020-04-15
21. Liu, F., Wen, Y., Zhang, D., Jiang, X., Xing, X., Meng, D.: Log2vec: A Heterogeneous Graph Embedding Based Approach for Detecting Cyber Threats within Enterprise. In: Proceedings of the 26thACM Conference on Computer and Communications Security. pp. 1777—-1794. CCS '19, ACM Press, New York, NY, USA (2019). https://doi.org/10.1145/3319535.3363224
22. Lockheed Martin Corporation: Gaining the Advantage—Applying Cyber Kill Chain Methodology to Network Defense (2014), https://www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/Gaining_the_Advantage_Cyber_Kill_Chain.pdf, accessed: 2020-04-15
23. M. Milajerdi, S., Eshete, B., Gjomemo, R., Venkatakrishnan, V.N.: ProPatrol: Attack Investigation via Extracted High-Level Tasks. In: Proceedings of the 4th International Conference on Information Systems Security. pp. 107–126. ICISS '18, Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-05171-6_6
24. m8r0wn: CrossLinked (2020), https://github.com/m8r0wn/CrossLinked, accessed: 2020-04-20
25. Milajerdi, S., Gjomemo, R., Eshete, B., Sekar, R., Venkatakrishnan, V.: HOLMES: Real-Time APT Detection through Correlation of Suspicious Information Flows. In: Proceedings of the IEEE Symposium on Security and Privacy. pp. 1137–1152. S&P '19, IEEE Computer Society, Washington, DC, USA (2019). https://doi.org/10.1109/SP.2019.00026
26. Miramirkhani, N., Barron, T., Ferdman, M., Nikiforakis, N.: Panning for gold.com: Understanding the dynamics of domain dropcatching. In: International Conference on World Wide Web (2018)
27. Parsons, K., McCormac, A., Pattinson, M., Butavicius, M., Jerram, C.: The Design of Phishing Studies: The design of phishing studies: Challenges for researchers. Computers & Security **52**(C), 194—206 (2015). https://doi.org/10.1016/j.cose.2015.02.008
28. Paterson, A., Chappell, J.: The Impact of Open Source Intelligence on Cybersecurity, pp. 44–62. Palgrave Macmillan UK, London (2014). https://doi.org/10.1057/9781137353320_4
29. RSA Research: Reconnaissance—A Walkthrough of the "APT" Intelligence Gathering Process (2015), http://www.kerneronsec.com/2015/10/a-walkthrough-of-apt-intelligence.html, accessed: 2020-04-15
30. The MITRE Corporation: Mitre att&ck matrix for enterprise (2019), https://attack.mitre.org/matrices/enterprise/, accessed: 2020-04-15
31. The MITRE Corporation: MITRE PRE-ATT&CK Matrix (2019), https://attack.mitre.org/matrices/enterprise/, accessed: 2020-04-15

32. Yadav, T., Rao, A.M.: Technical Aspects of Cyber Kill Chain. In: Proceedings of the 2015 Security in Computing and Communications. pp. 438–452. SSCC' 2015, Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-22915-7_40
33. Yu, H., Li, A., Jiang, R.: Needle in a Haystack: Attack Detection from Large-Scale System Audit. In: Proceedings of the 19th International Conference on Communication Technology. pp. 1418–1426. ICCT '19 (2019). https://doi.org/10.1109/ICCT46805.2019.8947201

A Analyzed MITRE PRE-ATT&CK Techniques

Table A lists the groups analyzed in this work. For each group, the techniques and tactics are shown and we indicate whether we analyzed it (*"Meas."*), if we collected the needed information on third-party websites (*"3rd"* or from first-party resources (*"1st) "*, and how we collected them (*"How obtained"*). If we did not collect data on a technique, the column "How obtained" provides a brief explanation why.

Technique	Meas.	1st	3rd	How obtained
Technical Information Gathering (TA0015)				
Acquire OSINT data sets and information	✗	—	—	Too general
Conduct active scanning	✗	—	—	Too general
Conduct passive scanning	✗	—	—	out of scope
Conduct social engineering	✗	—	—	out of scope
Determine 3rd party infrastructure services	✓	✓	✓	Shodan and IP addresses
Determine domain and IP address space	✓	✓	✓	log addresses during crawls
Determine external network trust dependencies	✓	✓	✗	log 3rd party usage
Determine firmware version	✓	✓	✓	during crawl & metadata
Discover target logon/email address format	✓	✓	✗	extract from metadata
Enumerate client configurations	✓	✓	✗	during crawl & metadata
Enumerate externally facing entities	✓	✓	✗	during crawl & metadata
Identify job postings and needs/gaps	✓	✗	✗	No API present
Identify security defensive capabilities	✗	—	—	out of scope
Identify supply chains	✗	—	—	very hard automatically
Identify technology usage patterns	✓	✓	✗	logged during crawls

Continued on next column

Continued from previous column

Technique	Meas.	1st	3rd	How obtained
Identify web defensive services	✓	✓	✗	analyzing 3rd party usage
Map network topology	✓	✓	✓	based on identified data
Mine technical blogs/forums	✗	✗	✗	out of scope
Obtain domain/IP registration information	✓	✓	✓	`whois` queries
Spearphishing for Information	✗	—	—	out of scope
People Information Gathering (TA0016)				
Acquire OSINT data sets and information	✗	—	—	Too general
Aggregate individual's digital footprint	✗	—	—	very hard automatically
Conduct social engineering	✗	—	—	out of scope
Identify business relationships	✗	—	—	out of scope
Identify groups/roles	✓	✗	✓	Based on social media data
Identify job postings and needs/gaps	✓	✗	✓	No API present
Identify people of interest	✓	✗	✗	based on collected data
Identify personnel with an authority/privilege	✓	✗	✗	based on collected data
Identify sensitive personnel information	✗	—	—	out of scope
Identify supply chains	✗	—	—	out of scope
Mine social media	✓	✗	✓	APIs of platforms
Organizational Information Gathering (TA0017)				
Acquire OSINT data sets and information	✗	—	—	Too general
Conduct social engineering	✗	—	—	out of scope
Determine 3rd party infrastructure services	✓	✓	✗	extracted during crawl
Determine centralization of IT management	✗	—	—	very hard automatically
Determine physical locations	✓	✓	✗	extracted during crawl
Dumpster dive	✗	—	—	out of scope
Identify business processes/tempo	✗	—	—	out of scope
Identify business relationships	✗	—	—	social media data
Identify job postings and needs/gaps	✓	✗	✓	No API present
Identify supply chains	✗	—	—	out of scope
Obtain templates/branding materials	✓	✓	✗	extracted during crawl

Concluded