Analyzing Potential Pre-Attack Surfaces

For an attacker, one of the first steps is to collect as much information as possible on the target to plan their further steps. 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.

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.

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.

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).

In computer security, phishing describes the act when an adversary impersonates a trusted entity with the intent to trick users into exposing personal data or spreading malware via malicious attachments or links.

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.

An overwhelming majority (88%) of all APTs used social engineering techniques to deploy their attack tools (e.g., mal- ware) in companies' infrastructure. Furthermore, email seems to be the most popular way to get in touch with the victim (80%).

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%)

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.

In our analysis, we perform an in-depth analysis of 30 entities (27 companies, two government agencies, and one non-profit organization). 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.

After identifying the "landing pages" of all domains associated with a company, we visit each page and recurse through 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.

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.

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.

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.

Platforms such as LinkedIn can be abused by adversaries to collect intelligence on a company. This data might provide several details about the internal workings of a company, and its employees and their careers, contacts, or supervisors.

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)

Finally, adversaries may utilize data from previous data breaches to prepare their attack. In this work, we use the Have I Been Pwned API to test if a company ever leaked data that can be used in another attack on that company. The API does not directly provide any of the breached data but returns categories of data that the leak contained.

In total, we scanned 30 entities and identified 492 domains 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.

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).

Almost three-quarters of all companies in our dataset leaked an employee's email address. 81% of the companies exposed third parties they work with (i.e., collaborating partners that created a document).

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. An attacker can use this information when preparing for the attack (e.g., zero-day exploits for the used software).

For 60% of the analyzed companies, an adversary actively abused a homoglyph domain, at the time of our crawl. The presence of such domains indicates that adversaries are likely already actively trying to misguide users or employees of such services. 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.

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.

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.

We identified 5,910 (62%) employees that leak between seven and 15 attributes. 878 (9%) employees leak between 24 and 28 attributes 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). The results show that almost all companies in our dataset provide a considerable pre-attack surface to motivated attackers.

Data leakage is not always under the control of the companies, nor is it always possible to revert the leakage.

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.

Actionable tools to counter misuse 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, and to increase awareness in a way that empowers employees not to provide too much work-related information on social media platforms.