LinkedIn data breaches, 2021

Justin Chung, Aiguin Guseinova, Ashely Honn, and Kevin Salger

California State University, Long Beach

IS 656: Information Systems Security and Assurance

Joshua Morgan

7 May 2022

Introduction

As the world's companies have increasingly turned to computer-based systems for collecting, processing, and mining data to answer business-related questions over the last thirty years, the value of that data has also increased. Not only have the companies recognized the data's value, thieves and other nefarious actors have also recognized opportunities regarding that data.

Tens of thousands of attempted incidents occur annually on computer and network systems. The 2021 Verizon DBIR[1] Report authors, for example, examined 79,635 incidents, of which 29,207 met standards for inclusion in the study, and 5,258 of those were confirmed as data breaches. (Bassett et al., 2021) Verizon provides definitions for both Incident and Breach:

> **Incident:** A security event that compromises the integrity, confidentiality or availability of an information asset.
>
> **Breach:** An incident that results in the confirmed disclosure—not just potential exposure—of data to an unauthorized party. (Bassett et al., 2021)

With the volume of cyber incidents and breaches in the multiple thousands, even a modest discussion of the history of cyber security would fill a small book. Instead, this report will focus on one company in 2021.

LinkedIn

LinkedIn bills itself as the "world's largest professional network with 810 million members in more than 200 countries and territories worldwide". (LinkedIn, Inc., 2022) As a company (a service now operated by Microsoft since December 2016), and with hundreds of

[1] DBIR = Data Breach Investigations Report

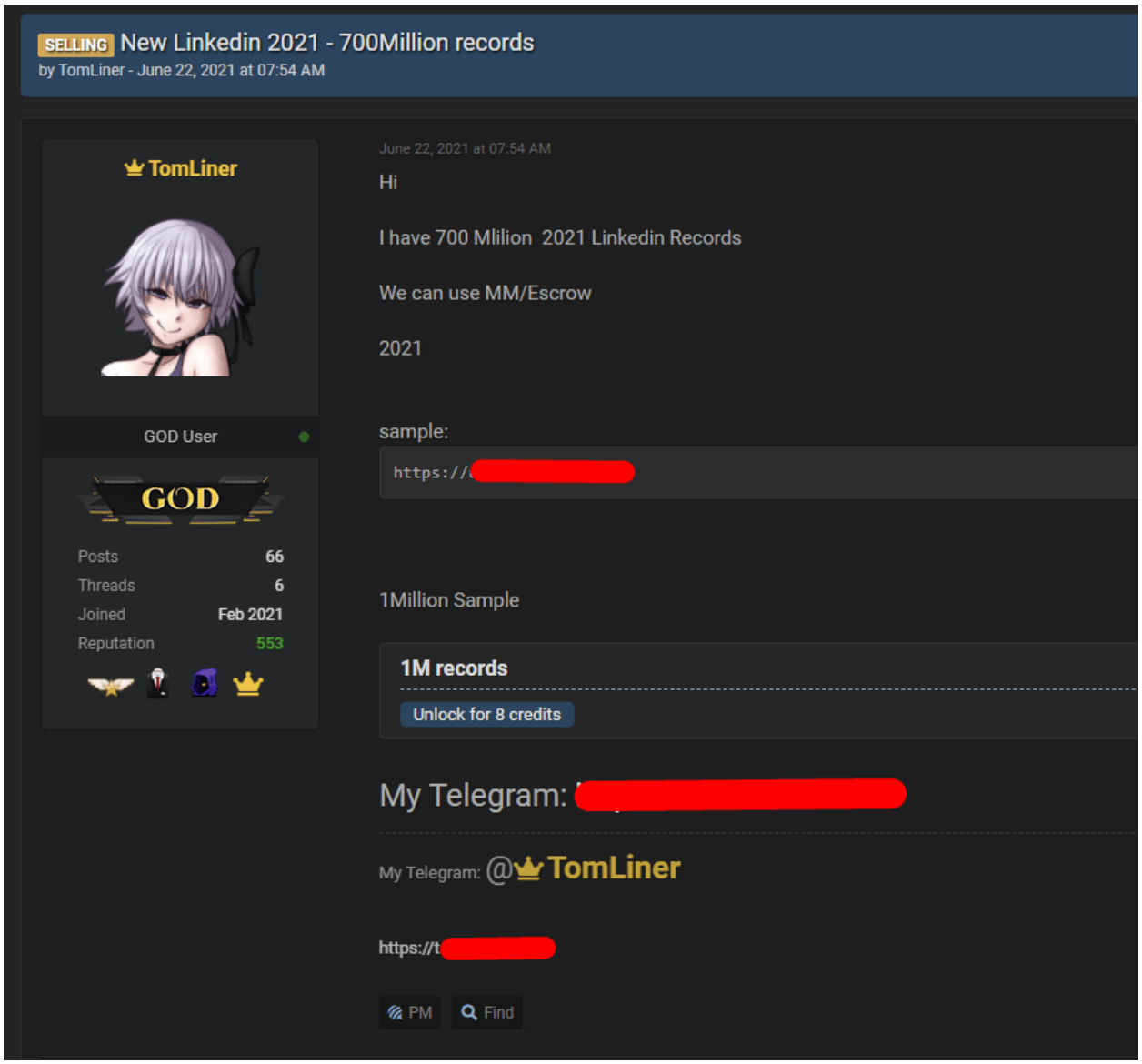
millions of worldwide users who use the site to share their educational and employment experiences, the treasure trove of user data is very tempting for bad actors.

An October 2021 article on Cognyte.com reported that a new trend of data theft is emerging. The actors are not using ransomware or outright theft (where the data is removed from its source). Instead, they are mining large quantities of publicly available data, copying it, categorizing and filtering it, then offering it for sale to other threat actors. (Cognyte CTI Research Group, 2021) . In a sense, the data miner or data scraper is providing a value-added service to threat actors, performing a pre-processing step previous bad actors had to perform themselves.

This scraping technique was levied against LinkedIn in April and June 2021. (Cybernews Team, 2021, Morris, 2021) The incident in April harvested the records of 500 million users while the incident in June reaped 700 million user records. With a user base of 756 million users, the June incident represented about 94% of the site's users. (Cimpanu, 2021) Basic mathematics reveals there would have to be a significant overlap of users between the two incidents, though a more thorough examination of the data is needed to know which records overlapped.

Several sources have commented that the data offered for sale in June by a "GOD User" named TomLiner was (a) aggregated by multiple websites, not just LinkedIn and (b) included "full names, email addresses, phone numbers, physical addresses, geolocation records, LinkedIn usernames and profile URLs, personal and professional experience/backgrounds, genders, and other social media account usernames." What was not taken were login credentials or financial information. The seller offered the collection to buyers for US$5000 and said the data was collected using LinkedIn's own API. (Malwarebytes Labs, 2021)

The following image is a screenshot of the RaidForums claim by TomLiner, redacted in parts, on the PrivacySharks website. PrivacyShark was the first to report the June incident to LinkedIn prior to publishing a news article about the incident.



(Hodson, 2021)

Despite the value-added approach, LinkedIn did not appear to be overly concerned following the incidents, at least in public. The communication in April from LinkedIn after the first scraping incident said:

> Members trust LinkedIn with their data, and we take action to protect that trust. We have investigated an alleged set of LinkedIn data that has been posted for sale and have determined that it is actually an aggregation of data from a number of websites and companies. It does include publicly viewable member profile data that appears to have been scraped from LinkedIn. This was not a LinkedIn data breach, and no private member account data from LinkedIn was included in what we've been able to review. (LinkedIn Corporate Communication, 2021)

A statement released after the June incident expressed a similar tone. It read, in part:

> Our teams have investigated a set of alleged LinkedIn data that has been posted for sale. We want to be clear that this is not a data breach and no private LinkedIn member data was exposed. Our initial investigation has found that this data was scraped from LinkedIn and other various websites and includes the same data reported earlier this year in our **April 2021 scraping update**. (LinkedIn Corporate Communication, 29 June 2021)

LinkedIn was not amused by the actions, though. Last updated 17 December 2018, the Prohibited Software and Extensions page of LinkedIn's Help page states that the use of any technology that "scrapes…or automates activity on LinkedIn's website. Such tools violate the User Agreement". Among the "Don't" listed are "Develop, support or use software, devices, scripts, robots, or any other means…to scrape the Services or otherwise copy profiles". LinkedIn also says don't "[c]opy, use, disclose or distribute any information obtained from the Services…without the consent of LinkedIn." (LinkedIn Help, 2018) TomLiner appears to have violated all of these directives.

LinkedIn claims that neither of these incidents breached any important network systems. Paul Rockwell, writing on the LinkedIn blog, states an attacker has not breached "secure systems, subvert[ed] firewalls, or access[ed] protected network information." He continues to explain that the practice of scraping has existed since the beginning of the Internet. For instance, search engines are authorized to collect and index the Internet. The main difference is what the scrapers do with the data. Threat actors scrape without permission. (Rockwell, 2021)

Whether TomLiner was successful in selling the 700 million records or not in June, the data was offered for sale again in August. Whereas TomLiner offered about one million records as proof, a different threat actor in the first August incident leaked multiple millions of records "also filtered by country, on a Dark Web forum". A third threat actor filtered the data by profession and country, highlighting the IT personnel, HR professionals, and finance executives. It is not known exactly why these three careers were chosen above others, but the speculation is these groups respond in greater numbers to cyber attacks. (Cognyte CTI Research Group, 2021)

The Problem

If the scraped, cleaned, sorted, filtered, and presumably sold data was publicly available, what is the worry?

Assume for the moment that all the data sellers were successful finding buyers. Who are the buyers? What intentions do those people or groups have with the information? How can the data be used to their advantage? The data becomes a way to search and find high value targets. Some of those targets include chief executive officers, financial officers, sales team members, and the previously mentioned information technology and human resource personnel. What makes these people especially interesting?

Human resource (HR) team members are targeted because they often work with sensitive organizational data, including the personally identifiable information (PII) of the employees. HR can also be a good entry point for a company's network since the HR department personnel often open emails with attachments. Phishing and spamware attacks have a greater chance of success in those instances. Or malicious actors can impersonate HR personnel, gaining PII of employees

and potential employees. The LinkedIn data, public though it was, helps open doors into additional, more sensitive information on other systems. (Cognyte CTI Research Group, 2021)

Sales representatives frequently receive emails from unknown senders (potential buyers) but are also connected with payment and financial systems. Information technology members are often the recipient of targeted phishing emails, as the department has access to multiple company networks and business-critical applications. According to Barracuda Networks, Inc, in their report on spear phishing in July 2021, IT employees each are "targeted by 40 email attacks [per year], which is well above average". Administrative and executive assistants are also popular targets, with access to scheduling and sensitive accounts. (Barracuda Networks, Inc., 2021)

In short, the aggregation of publicly available data is not a danger in and of itself. But the data can be used to gain entry to other, more critical systems with greater value.

Mitigation

Among the public comments from LinkedIn was the following,

> "Members trust LinkedIn with their data, and any misuse of our members' data, such as scraping, violates LinkedIn terms of service. When anyone tries to take member data and use it for purposes LinkedIn and our members haven't agreed to, we work to stop them and hold them accountable." (LinkedIn Corporation, 2021)

Unfortunately, the statement "we work to stop them and hold them accountable" is vague. A detailed plan released to the public is akin to an army publishing their battle plans, strategies, and tactics to their opponent prior to the war. Certainly the techniques and tools are not unknown to both sides, but the question is how these are used.

LinkedIn's main purpose is to connect people in the working world. And they want customers to sign up and use their site. Completely disabling scraping is not a wise business choice (and may not be possible anyway). At least some scraping is necessary for the various search engines that are authorized to perform that task.

The scraping is a violation of LinkedIn's terms of service, allowing LinkedIn to suspend or cancel the subscriptions and users engaged in the nefarious activity. To do that, however, LinkedIn needs to have identifying information of the perpetrator. If the June seller known as TomLiner is a "GOD User" according to RaidForums and he/she/they have used LinkedIn's API, they are likely smart enough not to use their own PII on RaidForums. LinkedIn (and other forensic entities) will have to use other techniques to find the actual culprit.

Once found, is there a legal standing in which to prosecute the bad actor? If TomLiner or the other two sellers are outside the jurisdiction of the United States or its allies, prosecution may be a lengthy and costly affair.

Prevention

The initial compromise could be understood as the point when the attacker successfully executes malicious code on one or more systems. Research shows a common method for gaining access is through some form of social engineering (most often spear phishing). Exploiting a vulnerability on an Internet-facing system or by any other means are used as well. As was acknowledged by the LinkedIn officials, the reason for the 2021 incidents was *web scraping***.** It implies the use of various methods (primarily the LinkedIn API, according to TomLiner) to collect information from across the internet. Generally, it is done with software that simulates human web surfing. The goal is pretty simple: a collection of data from different websites that

could be sold to other users to earn profit/financial gain. There are some general methods to detect and deter scrapers that could have been followed in time by LinkedIn to prevent web scraping:

- Logs should be checked regularly for many similar actions from the same IP address, for instance. In case of unusual activity indicative of automated access (scrapers), site access can blocked or limited.
- Users share some responsibility to regularly change their passwords to the site, actively use and manage the security settings offered by LinkedIn, and minimize the personal information available on the public side of the LinkedIn (or other sites) service.
- One of the simple ways to implement *rate-limiting* would be to temporarily block access for a certain period of time or until test is passed. Using a Captcha ("Completely Automated Test to Tell Computers and Humans Apart") could be a more effective alternative against stopping scrapers, even though it may irritate users.
- Use critical thinking when viewing emails from unknown senders. If something seems too good to be true, it probably is. The Nigerian prince is not going to send you any money. The Internal Revenue Service does not contact people through email. Check out this site for a primer on too-good-to-be-true: https://www.snopes.com/articles/396428/how-to-detect-avoid-online-scams/
- Keep in mind “TANSTAAFL”: There Ain’t No Such Thing As A Free Lunch. There is always a price to be paid for a good or service.
- Make sure APIs are not exposed, even unintentionally. For example, if you are using AJAX or network requests from within Adobe Flash or Java Applets to load your data, it is trivial to look at the network requests from the page and figure out where those

requests are going. A bit of reverse engineering can reveal the endpoints that can be used in a scraper program.

- Scrapers which process HTML directly scrape by extracting contents from specific, identifiable parts of an HTML page. It is advised to frequently change the HTML and page structures to prevent the scraping software from working correctly.
- Use up-to-date security software to regularly scan the system for malicious code.

Lessons

Twenty-twenty-one was a year that resulted in various companies having data breaches, but since these breaches occurred, information has come into light about what these hackers wanted. In regards to LinkedIn, Cognyte CTI Research Group reported that there may be a pattern to these attacks, as, "The fact that the threat actor divided his LinkedIn database specifically into human resources (HR), information technology (IT), and finance personnel may indicate that these employees are more likely to be targeted by cyber offenders." (Cognyte CTI Research Group, 2021) It seems that hackers find these departments within companies as targets to gain valuable information from the company. For HR, it seems to provide a good entry into the company's network which would make them more susceptible towards phishing or malspam attacks. HR is by far one of the most valuable since the department deals with a lot of sensitive information within the company or organization.

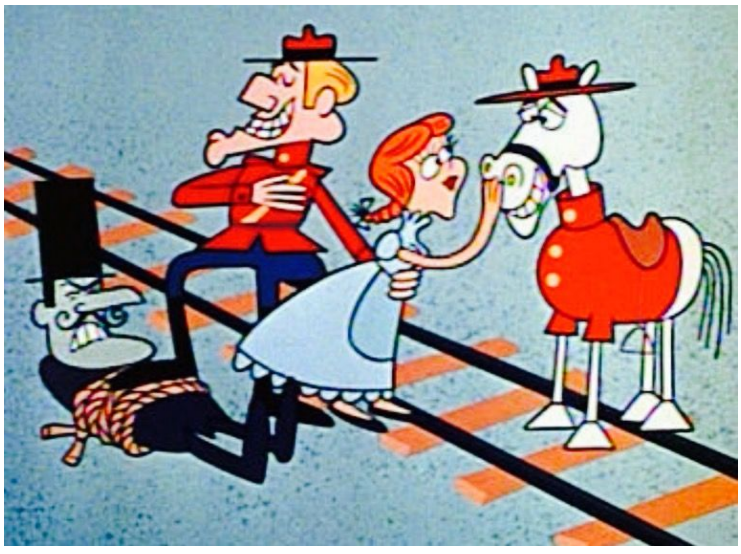
Since the three incidents that occurred during April, June, and August, LinkedIn has taken several steps to ensure these events do not occur again. First, LinkedIn has retrained their models to look for automated profile viewing, "Due to the adversarial nature of unauthorized scraping, our models are retrained and automatically deployed several times per day to quickly

adapt to new signals." (Rockwell, 2021) LinkedIn will also incorporate machine learning within these models in order for them to train and adapt accordingly to any scaping events.

Second, LinkedIn has models that are being used to ensure that logged-in scaping doesn't occur. They are essentially using their models to look for bot activity by using, "Deep learning to classify sequences of user behavior as automated, and we can also use outlier detection to detect activity that appears to be non-human." (Rockwell, 2021) Thus, LinkedIn uses their models to detect abuse and to help companies by informing them if a malicious bot is detected.

Third, LinkedIn has employed "Additional defenses that detect and take down fake accounts engaged in scraping at multiple stages." (Rockwell, 2021)

However, even with these various defenses that LinkedIn now employs, these are not 100% foolproof and the company is very aware that scraping can occur again. Thus, one of the final lessons that LinkedIn has learned, and subsequently informed the users, is that they should be cautious about which information they want to be displayed to the public. The company has informed the users that the information they provide on their website is public information, likely meaning that scraping and data collection can occur again and that it is imperative the user displays only the information they want to display for public consumption.



References

Barracuda Networks, Inc. (2021, July). *Spear Phishing: Top Threats and Trends* (Issue Vol. 6). spear-phishing_report_vol6.pdf. Retrieved May 6, 2022, from https://assets.barracuda.com/assets/docs/dms/spear-phishing_report_vol6.pdf

Bassett, G., Hylender, C. D., Langlois, P., Pinto, A., & Widup, S. (2021, May 13). *2021 DBIR Master's Guide*. Verizon. Retrieved April 24, 2022, from https://www.verizon.com/business/en-au/resources/reports/dbir/2021/masters-guide/

Cimpanu, C. (2021, September 22). *Hackers leak LinkedIn 700 million data scrape*. The Record by Recorded Future. Retrieved May 2, 2022, from https://therecord.media/hackers-leak-linkedin-700-million-data-scrape/

Cognyte CTI Research Group. (2021, October 21). *2021 LinkedIn breach: cybercriminals are the new headhunters*. Cognyte. Retrieved April 24, 2022, from https://www.cognyte.com/blog/2021-linkedin-breach-cybercriminals-are-the-new-headhunters/

Cybernews Team. (2021, April 6). *LinkedIn Data Breach - 500M Records Leaked and Being Sold | CyberNews*. Cybernews. Retrieved April 24, 2022, from https://cybernews.com/news/stolen-data-of-500-million-linkedin-users-being-sold-online-2-million-leaked-as-proof-2/

Hodson, M. (2021, June 27). *Exclusive: 700 Million LinkedIn Records Leaked June 2021*. PrivacySharks. Retrieved May 5, 2022, from https://www.privacysharks.com/exclusive-700-million-linkedin-records-for-sale-on-hacker-forum-june-22nd-2021/

LinkedIn Corporate Communication. (2021, April 8). *An update from LinkedIn*. LinkedIn Pressroom. Retrieved April 24, 2022, from https://news.linkedin.com/2021/april/an-update-from-linkedin

LinkedIn Corporate Communication. (2021, June 29). *An update from LinkedIn*. LinkedIn Pressroom. Retrieved April 24, 2022, from https://news.linkedin.com/2021/june/an-update-from-linkedin

LinkedIn Corporation. (2021, June 29). *An update from LinkedIn*. LinkedIn Pressroom. Retrieved May 6, 2022, from https://news.linkedin.com/2021/june/an-update-from-linkedin

LinkedIn Help. (2018, December 17). *Prohibited Software and Extensions*. LinkedIn Help. Retrieved May 5, 2022, from https://www.linkedin.com/help/linkedin/answer/56347/prohibited-software-and-extensions?src=or-search&veh=search.yahoo.com%7Cor-search

LinkedIn, Inc. (2022, January 01). *About LinkedIn*. About LinkedIn. Retrieved April 24, 2022, from https://about.linkedin.com/

Malwarebytes Labs. (2021, June 30). *Second colossal LinkedIn "breach" in 3 months, almost all users affected*. Malwarebytes Labs. Retrieved May 5, 2022, from https://blog.malwarebytes.com/awareness/2021/06/second-colossal-linkedin-breach-in-3-months-almost-all-users-affected/

Morris, C. (2021, June 30). *LinkedIn data theft exposes personal information of 700 million people*. Fortune. Retrieved April 24, 2022, from https://fortune.com/2021/06/30/linkedin-data-theft-700-million-users-personal-information-cybersecurity/

Rockwell, P. (2021, July 15). *linkedin-safety-series-what-is-scraping*. Official LinkedIn Blog. Retrieved April 24, 2022, from https://blog.linkedin.com/2021/july/15/linkedin-safety-series-what-is-scraping