Identifying, Collecting, and Monitoring Personally Identifiable Information: From the Dark Web to the Surface Web

Yizhi Liu
Management Information Systems
University of Arizona
Tucson, Arizona
yizhiliu@email.arizona.edu

Fang Yu Lin
Management Information Systems
University of Arizona
Tucson, Arizona
fylin@email.arizona.edu

Zara Ahmad-Post
Management Information Systems
University of Arizona
Tucson, Arizona
zahmadpost@email.arizona.edu

Mohammadreza Ebrahimi
Management Information Systems
University of Arizona
Tucson, Arizona
ebrahimi@email.arizona.edu

Ning Zhang
Management Information Systems
University of Arizona
Tucson, Arizona
zhangning@email.arizona.edu

James Lee Hu
Management Information Systems
University of Arizona
Tucson, Arizona
jameshu@email.arizona.edu

Jingyu Xin
Management Information Systems
University of Arizona
Tucson, Arizona
jingyuxin@email.arizona.edu

Weifeng Li
Management Information Systems
University of Georgia
Athens, Georgia
weifeng.li@uga.edu

Hsinchun Chen
Management Information Systems
University of Arizona
Tucson, Arizona
hchen@eller.arizona.edu

Abstract—**Personally identifiable information (PII) has become a major target of cyber-attacks, causing severe losses to data breach victims. To protect data breach victims, researchers focus on collecting exposed PII to assess privacy risk and identify at-risk individuals. However, existing studies mostly rely on exposed PII collected from either the dark web or the surface web. Due to the wide exposure of PII on both the dark web and surface web, collecting from only the dark web or the surface web could result in an underestimation of privacy risk. Despite its research and practical value, jointly collecting PII from both sources is a non-trivial task. In this paper, we summarize our effort to systematically identify, collect, and monitor a total of 1,212,004,819 exposed PII records across both the dark web and surface web. Our effort resulted in 5.8 million stolen SSNs, 845,000 stolen credit/debit cards, and 1.2 billion stolen account credentials. From the surface web, we identified and collected over 1.3 million PII records of the victims whose PII is exposed on the dark web. To the best of our knowledge, this is the largest academic collection of exposed PII, which, if properly anonymized, enables various privacy research inquiries, including assessing privacy risk and identifying at-risk populations.**

Keywords—PII, privacy, data breach, dark web, surface web, data collection

I. Introduction

With the rapid growth of online platforms, data privacy has become a major societal concern. On one hand, personally identifiable information (PII) of internet users has been one of the main targets of cyberattacks [1]. The stolen PII in data breach attacks is often disseminated on the dark web hacker communities for further exploitation (e.g., filing fraudulent loan applications, medical claims, and tax returns), leading to financial loss and reputation damage. On the other hand, internet users often unknowingly expose their PII in people search engines and social media platforms on the surface web. Such PII contains users' name, age, gender, address, contact, occupation and education, which can be exploited by hackers. In 2018, 87 million Facebook user profiles were harvested by Cambridge Analytica without users' consent [2]. At-risk populations, including the elderly and children, are particularly vulnerable to PII exposure as they lack the capabilities and resources to protect themselves [3].

Accordingly, research has been proposed to collect stolen PII from the dark web for privacy risk assessment [4]. Additionally, the surface web has also been collected to assess the extent of privacy exposure [5]. Nonetheless, little research has collected and analyzed exposed PII from both the dark web and the surface web. Cybercriminals often leverage the stolen PII they obtain from the dark web in conjunction with the surface web to obtain a comprehensive profile of data breach victims. Therefore, relying on partial PII might lead to underestimating the extent of PII exposure, thereby compromising the accuracy of privacy risk assessment [6].

However, identifying and collecting exposed PII across the dark web and the surface web is a non-trivial task for two reasons. First, the timeliness of PII exposures in the dark web necessitates constant monitoring of data breaches. Due to the covert nature of the dark web, the monitoring of exposed PII has mostly been a manual process, requiring experts to actively search for emerging data breaches. This challenge has prevented prior studies from building a timely collection of exposed PII. Second, collecting from various dark web and surface web platforms entails tailored strategies, as these platforms are often different in terms of the availability of APIs, anti-crawling measures, and response time. As such, developing a PII collection from both the dark web and the surface web has been rare in prior research.

In this paper, we summarize our work in developing a timely, comprehensive collection of exposed PII across the dark web and surface web. To the best of our knowledge, we have developed the largest academic collection of exposed PII. Our collection offers various prescriptive research opportunities, including the identification of at-risk populations in data breaches and comprehensive privacy risk assessment for data breach victims. Additionally, a multitude of privacy-related research inquiries can be advanced by our collection. For example, researchers may study password security and the privacy risk of using e-mail addresses as usernames in login credentials. With proper anonymization, our PII collection can be further shared among privacy research communities to foster privacy analytics.

The rest of the paper is organized as follows. Section II reviews the exposed PII on the dark web and surface web and discuss their value to privacy analytics research. Section III presents our methodology for data breach monitoring and cross-web data collection. Section IV summarizes our results and promising research opportunities. Section V concludes the study and discusses our future directions.

II. Research Background

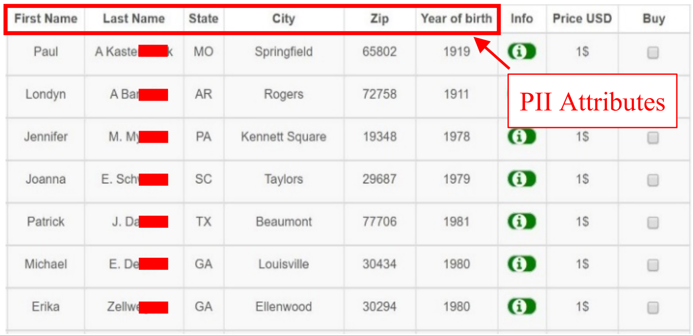
A. The Dark Web

The dark web [7] consists of a collection of illegal and covert platforms that facilitate communication and transactions among cybercriminals [8]. Accessing these platforms often require specific browsers (e.g., The Onion Router (TOR) browser) and specific configurations. The dark web is a valuable source of exposed PII because stolen PII from data breach attacks is mainly sold and shared on the dark web [9]. There is a huge demand for stolen PII, which can be used for making profits through identity theft (e.g., filing fraudulent loan applications, medical claims). Besides, hackers may share stolen PII to earn reputation and exchange for other hacking resources [7].In general, three major types of stolen PII are sold and shared on the dark web: Social Security Numbers (SSNs), credit/debit cards, and online account credentials [9]. Table I describes each type of stolen PII and its associated PII attributes. Such stolen PII can be found primarily on three major dark web platforms: Dark Net Marketplaces (DNMs) [8], carding shops, and hacker forums.

TABLE I. Summary of Stolen PII

Data Type	PII Attributes
SSNs	Full name, year of birth (YOB), Country, State, City, ZIP Code
Credit/Debit Cards	Full name, Country, State, City, ZIP Code
Account Credentials	E-mail Address, Password

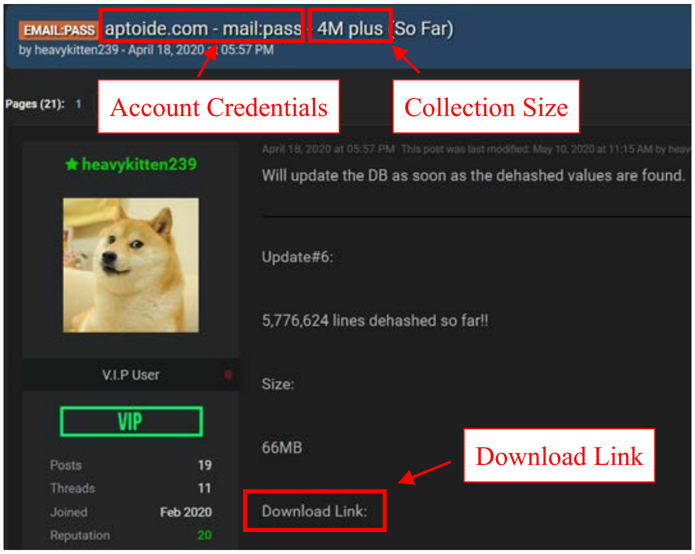
A DNM is a clandestine market on the dark web that hosts transactions of illicit products. Apart from physical products such as illicit drugs and weapons, SSNs and credit/debit cards stolen from data breaches are often sold on DNMs. This data contains various PII attributes of data breach victims, such as name, YOB, city, and ZIP code. Carding shops are another type of illegal markets dedicated to facilitating identity theft and carding fraud. Unlike DNMs, carding shops mainly sell digital goods stolen from data breaches, such as credit/debit card data, SSN, and account credentials. DNMs and carding shops can often serve as an early indicator of a data breach [10]. While SSN numbers and credit/debit card numbers on these platforms can be only obtained after the purchase, the sellers often provide certain PII attributes (e.g., name, city) of the victims to demonstrate the validity of the data for sale. Such information can be a valuable source for collecting exposed PII. Fig. 1 illustrates an example of a product listing page displaying names, location, and YOB of victims on a carding shop.



First Name	Last Name	State	City	Zip	Year of birth	Info	Price USD	Buy
Paul	A Kaste[redacted]k	MO	Springfield	65802	1919	ⓘ	1$	□
Londyn	A Ba[redacted]	AR	Rogers	72758	1911			
Jennifer	M. M[redacted]	PA	Kennett Square	19348	1978	ⓘ	1$	□
Joanna	E. Sch[redacted]	SC	Taylors	29687	1979	ⓘ	1$	□
Patrick	J. Da[redacted]	TX	Beaumont	77706	1981	ⓘ	1$	□
Michael	E. De[redacted]	GA	Louisville	30434	1980	ⓘ	1$	□
Erika	Zellw[redacted]	GA	Ellenwood	30294	1980	ⓘ	1$	□

Fig. 1. An example of a product listing page in a carding shop. The first name attribute is anonimized for privacy concerns.

Hacker forums are online discussion platforms where hackers can post messages related to hacking tools, techniques, source code, and malicious assets [11]. Breached data, such as large collections of account credentials, are sometimes shared for free on hacker forums. The account credentials usually consist of e-mail addresses and passwords. Fig. 2 illustrates an example of a hacker sharing a download link for a collection of stolen account credentials in a hacker forum post.



EMAIL:PASS aptoide.com - mail:pass - 4M plus (So Far)

by heavykitten239 - April 18, 2020 at 05:57 PM

Pages (21): 1

★ heavykitten239

V.I.P User

VIP

Posts	19
Threads	11
Joined	Feb 2020
Reputation	20

April 18, 2020 at 05:57 PM This post was last modified: May 10, 2020 at 11:15 AM by heav

Will update the DB as soon as the dehashed values are found.

Update#6:

5,776,624 lines dehashed so far!!

Size:

66MB

Download Link:

Fig. 2. An example of a hacker sharing a download link for a collection of stolen account credentials in a hacker forum post

All these three types of platforms serve as the primary source for cybercriminals to obtain stolen PII in the dark web, posing significant privacy threats to data breach victims [12]. Collecting stolen PII from these platforms can help identify data breach victims and assess their privacy risk. However, collecting stolen PII from the dark web can be challenging. The advertised PII may be withdrawn from the platform shortly after the appearance. Also, dark web data collection requires circumventing anti-crawling measures installed by the platforms to prevent data collection. Four anti-crawling measures are commonly adopted by dark web platforms: User-agent checking, authentication, session timeout, and CAPTCHA [13]. User-agent checking is performed to ensure the request comes from a browser and not a crawler, Authentication enforces registration and logging in to the platform. Session timeout blocks excessive number of requests from the same user. CAPTCHA is utilized heavily to distinguish human users from crawlers.

B. The Surface Web

Cybercriminals not only leverage the dark web but also the surface web to develop comprehensive PII profiles of data breach victims. To estimate the full extent of PII exposure, it is therefore necessary to collect exposed PII on the surface web. The surface web is part of the internet that is accessible to the general public without requiring special software or configurations. With the proliferation of online services, internet users unknowingly expose an unprecedented amount of PII on the surface web [5]. For instance, many social media platforms encourage the users to share and update their PII, such as their name, age, and city. However, such PII can be further disseminated by other users in the friend list without permission. Also, the platforms and third-party applications can access the PII without users' further permission [2]. As a result, internet users are often unaware of the extent of their PII exposure [14]. The PII exposed on the surface web often contains attributes that are rarely available in stolen PII on the dark web. These attributes include gender, phone number, and occupation.

Two major types of platforms on the surface web expose a large amount of PII: people search engines and social media platforms. People search engines are publicly accessible search interfaces specifically geared for personal information [15]. These platforms gather PII from proprietary databases, public records, social media platforms, etc. Various attributes such as name and ZIP code, can be used as the search query. Retrieved results contain PII of individuals, including phone number, e-mail address and physical address, many of which are complementary to stolen PII from the dark web. Social media platforms are online services that connect users and are accessible with registered accounts and specialized interfaces [5]. Users exchange and update personal information on social media platforms to satisfy social needs. Such information is often publicly available and can be used to identify a user's real-world identity. As a result, social media platforms have become emerging sources for PII collection [5]. These platforms provide PII attributes complementary to the dark web and people search engines, such as the occupation and photos.

Collecting PII from people search engines and social media platforms are critical to estimate the full extent of PII exposure. To comprehensively collect the exposed PII of data breach victims, we can enrich the dark web data collection with exposed PII from the surface web. Anti-crawling measures are also employed by these platforms on the surface web. Specifically, most platforms still check the user agent to identify crawlers. Also, many social media platforms are only accessible with credentials. To avoid DDoS attacks, surface web platforms detect abnormal requests of IP addresses and block them. Lastly, people search engines usually have a long response time to prevent their data from being automatically crawled.

III. Collection Methodology

We propose a systematic approach to develop a timely, comprehensive collection of exposed PII across the dark web and the surface web. On a high level, the proposed approach consists of three steps. First, we automatically monitor data breach news to obtain the timely intelligence about data breaches. Second, we locate and collect the exposed PII on the dark web based on the data breach intelligence. Third, we search and collect exposed PII from the surface web to complement the dark web PII collection.

A. Automated Data Breach Monitoring

We design a system to frequently access data breach news sites via Really Simple Syndication (RSS) feeds and analyze the RSS feed to identify the intelligence related to data breaches. The intelligence is then disseminated to domain experts for further examination. Our automated data breach monitoring architecture is illustrated in Fig. 3.

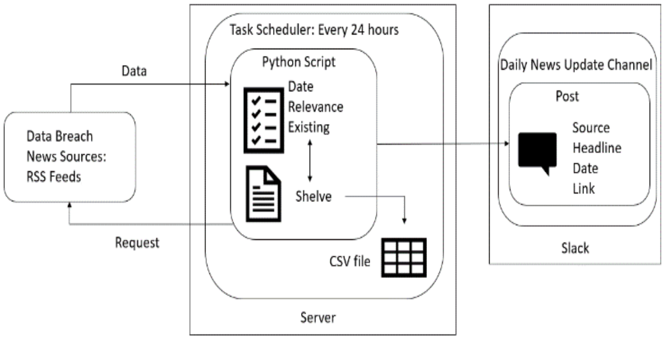


Fig. 3. Our Automated Data Breach Monitoring System

As shown in the figure, the system gathers the intelligence of emerging breaches from data breach news sites selected by domain experts. To gather data breach news automatically, we leverage a Python script to access these sites periodically (i.e., once every 24 hours) via their RSS feeds. The RSS feed is further validated in terms of their relevance and redundancy to identify recently published news related to data breaches. Specifically, we first identify the news whose titles contain keywords related to data breach (e.g., "hack," "breach"). Then, we examine whether the news has been already published within the past 24 hours. This is because the sites could repeat the same news over multiple days. The resultant news is archived and disseminated using Slack, a collaborative team messaging and file-sharing platform, which informs our team domain experts of emerging data breach news.

B. Dark Web Collection

Following the notification from the monitoring system, domain experts navigate dark web hacker communities to locate and collect the breached PII data. As noted, the breached data can often be found on hacker forums, DNMs, and carding shops. For hacker forums, hackers share breached PII data files for free with a link in the posts. For DNMs and carding shops, we develop crawlers to collect stolen PII. Specifically, our crawlers first set the TOR environment, which enables dynamic IP address assignment. Then, the crawlers load credentials for accessing the targeted platform. Within the platform, to crawl targeted pages incrementally, we build a URL list of pages based on the URL structure or leverage CSS selector to click the provided pagination button. For each page, we extract the body from the HTML content for further parsing. Random waits are used between crawling each page to avoid anti-crawling detection. The extracted web pages are further parsed using Regular Expression (RegEx), which recognizes and retrieves the PII attributes such as name, YOB, city, ZIP code, and state.

C. Surface Web Collection

To enrich the dark web collection with exposed PII from the surface web, we use the stolen PII on the dark web as queries to search on the surface web platforms. This process allows us to collect additional exposed PII that is not available on the dark web (e.g., physical address). In particular, queries are developed based on the data type of the stolen PII. For stolen SSN and stolen credit/debit card collections, name and city are used as queries on people search engines and social media platforms. For stolen account credentials, queries are designed based on e-mail addresses, which are often used for registration on many surface web platforms (e.g., LinkedIn).

The collection process entails tailored strategies as surface web platforms are often different in terms of anti-crawling measures. Specifically, we leverage three collection strategies. First, when a platform provides APIs, these APIs are used to directly collect the data. Using APIs allows bypassing anti-crawling measures and the long response time. Second, most people search engines restrict data collection APIs to prevent the data from being shared. Hence, we collect search result pages by building surface web crawlers. To this end, we incorporate the user-agent information of a legitimate browser into the crawler. Furthermore, we use proxies and VPNs to avoid IP blocking. This strategy can be particularly useful for platforms that have weak anti-crawling measures and fast response time (e.g., That's Them). Third, most people search engines, such as BeenVerified and MyLife, have anti-crawling measures and slow response time. For these platforms, a Google-based crawler is implemented to bypass the restrictions. This crawler queries a combination of name, city, and platform name on Google (e.g., John Doe + New York + MyLife) and collect the URLs of retrieved results, which are then crawled separately by our surface web crawler.

IV. Data Collection Results

A. Dark Web Collection Results

In consultation with privacy experts, four major data breach news sites have been identified for automated data breach monitoring: Slashdot, Hackernews, HaveIBeenPwned, and DarkReading. Slashdot is a social news website that features news stories submitted and evaluated by site users and editors. Data breach news are often published under the security topics in Slashdot. Hackernews is a social news website focusing on computer security, where the stories are ranked by the users. HaveIBeenPwned is a website that allows internet users to check whether their PII has been compromised, and subscribe to notifications about future breaches. DarkReading is one of the largest cyber security news sites that contains intelligence about new cyber threats, vulnerabilities, and technology trends. We identified and collected three types of stolen PII from six sources, as summarized in Table II. BuySSN and WT1SHOP were the two largest DNMs we identified for selling SSN of U.S. victims. For stolen account credentials, we located and collected the breached data from the privacy subreddit on Reddit and RaidForums, both of which are active message boards for hackers. Tormarket and Yohohobay were the two carding shops that sold U.S. stolen credit/debit cards. Next, we detail our collection by data type and discuss promising research directions for each dataset.

TABLE II. Dark Web Collection Overview

Data Type	Source	Size
Stolen SSNs	BuySSN	54,912
	WT1SHOP	5,750,090
Stolen Credit/Debit Cards	Tormarket	831,949
	Yohohobay	13,324
Stolen Account Credentials	Reddit: Privacy Subreddit	1,199,527,942
	RaidForums	4,471,631
Total	**-**	**1,210,649,848**

1) Stolen SSNs

We collected 5,805,002 stolen SSNs between January 2018 and May 2020 from two DNM platforms (i.e., BuySSN and WT1SHOP). Table III summarizes our stolen SSN data collection.

TABLE III. Stolen SSN Collection Results

Source	Size	Attributes	# of Records
BuySSN	54,912	Full Name, YOB, State	54,912
		City	54,881
		ZIP Code	54,910
		Country	54,894
WT1SHOP	5,750,090	Full Name, State, City, ZIP Code	5,750,090
		YOB	3,933,674
Total	**5,805,002**	-	-

The collection includes six PII-related attributes: full name, YOB, state, city, ZIP code, and country. All these stolen SSNs are associated with U.S. victims. To the best of our knowledge, WT1SHOP is the largest DNM for stolen SSNs. Overall, 70% of the records from WT1SHOP include YOB, which can help identify the data breach victims and classify them by age groups. Hence, a promising research direction would be identifying at-risk populations (e.g., the elderly, children) in the stolen SSN victims. The name, city, and YOB attributes help identify them with an improved precision. In addition, cross-referencing the dark web and surface web collections could reveal useful patterns in the geographical location and education background of at-risk populations with exposed PII.

2) Stolen Credit/Debit Cards

We identified and collected 845,273 stolen cards issued in the U.S. from two carding shops between January 2018 and May 2020. As summarized in Table IV, our stolen credit/debit card collection provides PII-related attributes, including full name, country, state, city, and ZIP code. These PII attributes are sufficient to identify the cardholders.

TABLE IV. STOLEN CREDIT/DEBIT CARD COLLECTION RESULTS

Source	Size	Attributes	# of Records
Tormarket	831,949	Full name	709,380
		Country	831,949
		State	181,000
		City	225,891
		ZIP Code	193,300
Yohohobay	13,324	Full Name, Country, State, City, ZIP Code	13,324
Total	**845,273**	-	-

Given that stolen credit/debit cards are closely related to financial crimes, one potential research direction is to proactively identify the victims of these stolen cards and inform them of their potential PII exposure. This direction is meaningful to not only cardholders but also financial institutions and law enforcement agencies.

3) Stolen Account Credentials

We identified and collected over 1.2 billion stolen account credentials from data breaches between December 2017 and April 2020. All these account credentials contain the e-mail address of the victims. Table V details our stolen account credentials collection. One source of the collection was a trove of social media platform and e-mail accounts, collected and aggregated by an anonymous hacker from 256 data breaches. The dataset was briefly shared online, during which we found the link to the data on Reddit. About half of these records are U.S. e-mail addresses. At the time of collection, 79.3% of the passwords were still authentic [16].

TABLE V. STOLEN ACCOUNT CREDENTIAL COLLECTION RESULTS

Source	Size	Attributes	# of Records
Social Media and E-mail Accounts	1,199,527,942	E-mail (U.S.)	598,509,758
		Password (U.S.)	598,509,758
Aptoide	4,471,631	E-mail	4,471,631
		Password	4,470,937
Total	**1,203,999,573**	-	-

Another notable source was the data breach of Aptoide, an Android-based app hosting platform. Aptoide's user data was breached on April 18th, 2020, and was shared on RaidForums, a hacker forum. Our automated data breach monitoring system notified us about this dataset. 82.9% of e-mail addresses in the Aptoide collection are related to U.S. domains. As noted, e-mail accounts can often be used as the credentials for multiple online platforms. This can be exploited by adversaries to access these online platforms, increasing the privacy risks to the data breach victims. Therefore, a promising direction is to study the privacy risks of using e-mail addresses as usernames of login credentials. Another direction is the password security research. Specifically, researchers can analyze the commonalities of breached passwords and relate them to other attributes, such as age and occupation, to help increase individuals' awareness about password security.

B. Matching PII: From the Dark Web to the Surface Web

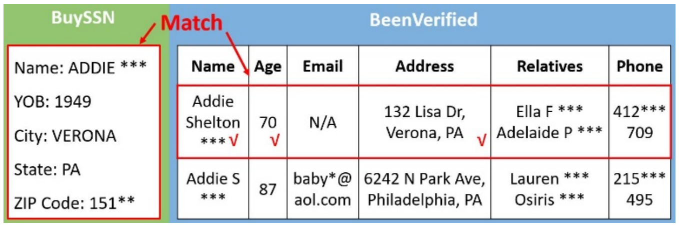
To show the potential of our surface web collection methods for matching PII, we selected four subsets from the dark web collection as queries to search on seven surface web platforms recommended by privacy research experts. Specifically, we randomly sampled 5,000 stolen SSNs, 5,000 stolen credit/debit cards, and 5,000 stolen account credentials. Additionally, to focus on at-risk populations, we selected 5,887 stolen SSNs belonging to senior citizens. Then, we used these subsets to search for PII on people search engines and social media platforms. Table VI summarizes our PII matching results.

TABLE VI. SAMPLE SURFACE WEB PII MATCHING RESULTS

Platform	# of Retrieved Records	Matched with Dark Web (%)	Additional PII Available on the Surface Web
Stolen SSNs – Senior Citizens (5,887 records)			
That's Them	3,255	7.81%	address, age, gender
MyLife	2,236	15.86%	address, alias, relative
BeenVerified	146,856	11.60%	age, e-mail, phone, relative
Spokeo	211,125	8.19%	age, alias, gender, relative
Twitter	11,270	4.18%	username, photo
Stolen SSNs – General (5,000 records)			
That's Them	2,441	7.14%	address, age, gender
MyLife	2,963	16.28%	address, alias, relative
BeenVerified	264,859	4.56%	age, e-mail, phone, relative
Spokeo	121,348	4.38%	age, alias, gender, relative
Twitter	8,268	3.04%	username, photo
Stolen Credit/Debit Cards (5,000 records)			
That's Them	1,936	14.32%	address, age, gender
MyLife	2,414	11.46%	address, alias, relative
BeenVerified	146,855	8.68%	age, e-mail, phone, relative
Spokeo	249,277	18.46%	age, alias, gender, relative
Twitter	9,132	58.16%	username, photo
Stolen Account Credentials (5,000 records)			
LinkedIn	999	19.92%	city, name, photo, occupation
Spokeo	132,785	42.4%	age, alias, gender, relative
That's Them	546	6.28%	address, age, gender
MyLife	1,965	12.26%	address, alias, relative
BeenVerified	34,110	0.42%	age, e-mail, phone, relative
Twitter	1,330	8.98%	username, photo

As shown in Table VI, for each subset, the first column lists the surface web platforms used for searching. The second column presents the number of retrieved candidate records from each surface web platform. We developed a rule-based matching program to automatically filter out duplicates and find matching records based on the similarity of attribute values between the records from the dark web collections and candidate records from the surface web platforms. The percentage of the records that matched with the dark web is presented in the third column. The fourth column lists the additional attributes from the surface web that are not available on the dark web.

MyLife had the highest match rate for stolen SSNs (15.86% for senior citizens and 16.28% in general), suggesting it has a higher coverage of SSN holders compared to the other platforms. For stolen credit/debit cards, the PII of 58.16% of cardholders was further exposed on Twitter. This suggests that cardholders and Twitter users have a significant overlap. For stolen account credentials, Spokeo has a 42.4% match rate, indicating the PII of the victims was highly exposed on people search engines. Besides, the privacy risk caused by PII exposure on the surface web varied by platform. For example, the address was often exposed on That's Them and MyLife, increasing the risks of location tracking. Contact information like phone number and e-mail were often exposed on BeenVerified and Spokeo, raising the risk of being spammed. Furthermore, the photos of victims exposed on Twitter and LinkedIn significantly increase the risks of identity threats. The matched records belong to victims whose PII was exposed in both dark web and surface web, suggesting their privacy was more likely to be compromised. Fig. 4 shows an example of matching exposed PII of a senior citizen. His/her name, YOB, city, state, and ZIP code were exposed on the dark web. On the surface web, the addresses, relatives, and phone were further exposed. As seen, the combination of the exposed PII increases the risks of identity threats, location tracking, and spamming.


BuySSN

Name: ADDIE ***
YOB: 1949
City: VERONA
State: PA
ZIP Code: 151**

Match

BeenVerified

Name	Age	Email	Address	Relatives	Phone
Addie Shelton *** √	70 √	N/A	132 Lisa Dr, Verona, PA √	Ella F *** Adelaide P ***	412*** 709
Addie S ***	87	baby*@ aol.com	6242 N Park Ave, Philadelphia, PA	Lauren *** Osiris ***	215*** 495

Fig. 4. An example of matching exposed PII of a senior citizen. Some sensitive attributes are anonimized for privacy concerns.

As shown, cybercriminals can leverage the stolen PII they obtain from the dark web in conjunction with the surface web to obtain a comprehensive profile of data breach victims. Thus, privacy risk assessment based on a single source can lead to an underestimation of the potential risk. Our initial analysis indicates a significant level of PII exposed when combining data from the dark web and surface web. Enabled by this holistic view of PII exposure, one promising research direction would be a comprehensive privacy risk assessment for the data breach victims, especially for at-risk populations and those whose PII is threatened on both the dark web and surface web. Also, there is a vital need for more advanced entity resolution techniques to facilitate matching records from the dark web and the surface web.

V. Conclusion and Future Directions

To identify at-risk individuals and assess their privacy risks, existing research largely focuses on collecting data from either the dark web or the surface web, which could result in an underestimation of privacy risk. Systematic PII collection from both the dark web and surface web can address this issue, whereas it is non-trivial due to the covert nature of the dark web and difficulty of data collection. In this paper, we summarize our effort to systematically identify and collect exposed PII across the dark web and the surface web. Enabled by our automated data breach monitoring system, we developed a collection comprising over 5.8 million stolen SSNs, 845,000 stolen credit/debit cards, and 1.2 billion stolen account credentials from the dark web. Using small subsets of our dark web PII collection as queries, we identified and collected 1.3 million PII records of data breach victims from the surface web. This large-scale data collection can facilitate various privacy research inquiries, such as providing the internet users with a holistic view of their privacy risks, increasing their privacy awareness, and helping at-risk populations in need. Future work can leverage advanced entity resolution approaches to facilitate the process of bridging the dark web and surface web collections. We also plan to integrate the collected data into a secure PII portal with a search interface. Additionally, we plan to anonymize our collection and make it accessible to the privacy research community. The portal, along with the anonymized collection can enable research inquiries in privacy analytics, proactive data breach notification, and privacy education.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) under Secure and Trustworthy Cyberspace (SaTC), (grant No. 1936370), Cybersecurity Innovation for Cyberinfrastructure (grant No. 1917117), and CyberCorps Scholarship-for-Service (grant No. 1921485).

References

[1] Risk Based Security, “Data Breach QuickView Report 2019 Q3 Trends,” 2019. https://pages.riskbasedsecurity.com/hubfs/Reports/2019/Data Breach QuickView Report 2019 Q3 Trends.pdf

[2] J. Isaak and M. J. Hanna, “User data privacy: Facebook, Cambridge Analytica, and privacy protection,” *Computer (Long. Beach. Calif).*, vol. 51, no. 8, pp. 56–59, 2018.

[3] B. Gupta and A. Chennamaneni, “Understanding Online Privacy Protection Behavior of the Older Adults: An Empirical Investigation.,” *J. Inf. Technol. Manag.*, vol. 29, no. 3, pp. 1–13, 2018.

[4] T. Floyd, M. Grieco, and E. F. Reid, “Mining hospital data breach records: Cyber threats to us hospitals,” in *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, 2016, pp. 43–48.

[5] G. Venkatadri *et al.*, “Privacy risks with Facebook’s PII-based targeting: Auditing a data broker’s advertising interface,” in *2018 IEEE Symposium on Security and Privacy (SP)*, 2018, pp. 89–107.

[6] D. A. B. Villalva, J. Onaolapo, G. Stringhini, and M. Musolesi, “Under and over the surface: a comparison of the use of leaked account credentials in the Dark and Surface Web,” *Crime Sci.*, vol. 7, no. 1, pp. 1–11, 2018.

[7] H. Chen, *Dark web: Exploring and data mining the dark side of the web*, vol. 30. Springer Science & Business Media, 2011.

[8] M. Ebrahimi, J. F. Nunamaker Jr, and C. Hsinchun, “Semi-Supervised Cyber Threat Identification in Dark Net Markets: A Transductive and Deep Learning Approach,” *JMIS*, vol. 37, no.3, 2020.

[9] P.-Y. Du *et al.*, “Identifying, collecting, and presenting hacker community data: Forums, IRC, carding shops, and DNMs,” in *2018 IEEE international conference on intelligence and security informatics (ISI)*, 2018, pp. 70–75.

[10] M. Ebrahimi, M. Surdeanu, S. Samtani, and H. Chen, “Detecting Cyber Threats in Non-English Dark Net Markets: A Cross-Lingual Transfer Learning Approach,” in *IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2018, pp. 85–90.

[11] V. Benjamin, W. Li, T. Holt, and H. Chen, “Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops,” in *2015 IEEE international conference on intelligence and security informatics (ISI)*, 2015, pp. 85–90.

[12] D. J. Solove and D. K. Citron, “Risk and anxiety: A theory of data-breach harms,” *Tex. L. Rev.*, vol. 96, p. 737, 2017.

[13] L. Von Ahn, M. Blum, N. J. Hopper, and J. Langford, “CAPTCHA: Using hard AI problems for security,” in *International conference on the theory and applications of cryptographic techniques*, 2003, pp. 294–311.

[14] A. Acquisti and J. Grossklags, “Privacy and rationality in individual decision making,” *IEEE Secur. Priv.*, vol. 3, no. 1, pp. 26–33, 2005.

[15] W. Weerkamp, R. Berendsen, B. Kovachev, E. Meij, K. Balog, and M. De Rijke, “People searching for people: Analysis of a people search engine log,” in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 2011, pp. 45–54.

[16] J. Casal, “1.4 Billion Clear Text Credentials Discovered in a Single Database,” 2017. https://medium.com/4iqdelvedeep/1-4-billion-clear-text-credentials-discovered-in-a-single-database-3131d0a1ae14