A Study on Personal Identifiable Information Exposure on the Internet

Ningning Wu
Department of Information Science
University of Arkansas at Little Rock
Little Rock, AR, USA
nxwu@ualr.edu

Robinson Tamilselvan
Department of Information Science
University of Arkansas at Little Rock
Little Rock, AR, USA
rtamilselvan@ualr.edu

Talha Tayyab
Department of Information Science
University of Arkansas at Little Rock
Little Rock, AR, USA
ttayyab@ualr.edu

***Abstract*—Personal Identifiable Information (PII) is any information that permits the identity of an individual to be directly or indirectly inferred. It should be protected against random access. This paper studies the extent of PII exposure on the Internet. It is hoped that the results of this study can help raise the Internet users' awareness on privacy protection.**

Keywords*—*PII, data privacy, PII exposure, people search engines

I. Introduction

Cyberspace users often post and share information (texts, images, vlogs) that may contain private information. Moreover, there are more than abundant mobile apps and web applications that collect customer information (disclosed or undisclosed) through different channels. Some of that information is publicly accessible and searchable. On one hand, there is a need for effective regulation of those applications to collect and disseminate personal information. At the same time, there is a need for helping users: 1) to monitor what personal information has been collected, and 2) be provided with decision-making tools to help them identify information to be shared publicly; both in an effort to safeguard their privacy and prevent discrimination.

Private information can be easily spread and commonly shared as unstructured data, through news reports, web documents, images, and social media outlets. Unstructured data presents many unique challenges for determining if the content includes potentially sensitive information about an individual. This research studies the accessibility of Personal Identifiable Information (PII) on the Internet through either the people search engines or information retrieval from public web documents.

PII is any information that permits the identity of an individual to be directly or indirectly inferred, including any information that is linked or linkable to that individual. Figure 1 shows that our PII is used everywhere. The National Institute of Standards and Technology defines PII as follows:

"PII is any information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information."

Figure 1: PII Examples

Guarding PII is important to ensure the integrity of an individual's identity. With just a few bits of personal information, thieves can create false accounts in their name, start racking up debt, or even create a falsified passport and sell their identity to criminals. Protecting PII is essential for personal privacy. The leakage of PII can lead to privacy and safety issues like personal embarrassment, workplace discrimination, and identity theft. PII getting into the wrong hands can result in devastating consequences.

PII is often collected and sold by data companies. Users should be careful when releasing their personal information to such companies. It's important to read the 'terms and conditions' carefully to understand 1) how their information is used and shared, 2) what privacy laws the company is compliant with, and 3) if users have right to opt out the sales of their information to third parties.

Organizations should use the concept of PII to understand which data they store, process, and manage that identify people, so that they will practice due diligence to protect the data that are at rest, in transit, or in process. Depending on the natures of PII data organizations collect and use, they should ensure that their practices comply with appropriate privacy laws.

Ideally the owner of the information should have complete rights over their information. Unfortunately, with the fast advances in Web, wireless communication, cloud, and IoT technologies, the owner almost loses the control of their information which is collected anywhere and anytime with or

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

TALBE I. PII ATTRIBUTES DISCLOSED

websites	Name	Phone	Full address	Relatives	Age	Resume	Social media accounts
https://www.spokeo.com/	Y			Y	Y		
https://www.instantcheckmate.com/	Y			Y	Y		
https://www.intelius.com/	Y			Y	Y		
https://www.zabasearch.com/	Y		Y		Y		
https://radaris.com/	Y	Y	Y	Y	Y	Y	Y
https://www.yellowpages.com/	Y	Y	Y				
https://www.peoplefinders.com/	Y			Y	Y		
https://www.truthfinder.com/	Y			Y	Y		

without owner's knowledge. With the help of federal privacy laws and regulations, we hope organizations would strictly follow the laws when collecting, using and sharing the data.

According to a 2022 study by Javelin[7], over 14.4 million people per year are a victim of identity fraud. Identity thieves typically use a combination of data from different sources to get the information needed to open credit cards, take out loans, and make erroneous purchases in the victim's name. A similar report by IBM estimates that the average total cost of a data breach is $3.86M globally, with the most compromised and costliest type of record being customer PII at $150 per record.

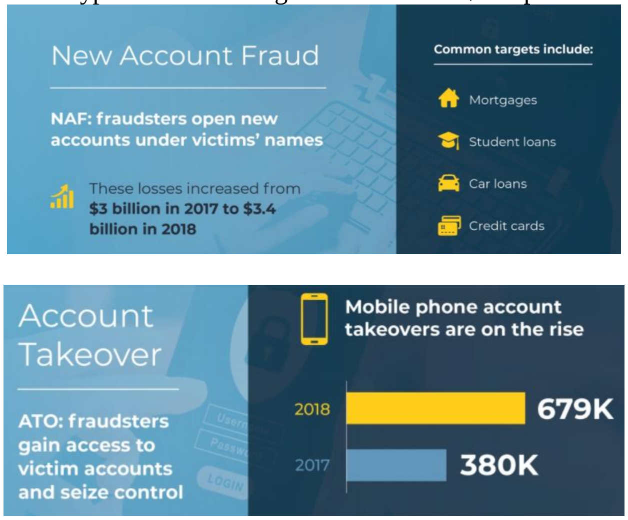

Figure 2: Identity Fraud Study by Javelin

This research explores how much PII is publicly accessible on the Internet. We hope the findings in this research can raise general public's awareness on protecting their private information. Social media become ubiquitous today. They provide convenient platforms for users to connect with family members and friends, make new friends, and share their life and experiences. On the other hand, user may accidentally disclose some sensitive information, like posting a funny picture of a messy desk with a driver's license on it and in clear sight. Indeed, social media posts published on social media platforms are a major and the most common source of PII exposure [8]. For a person lacking a strong sense of privacy protection, it would be easy for a determined criminal to compile sufficient information from his postings, and then steal his identity for fraud activities, which may cause a huge financial loss to the victim. Our preliminary analysis on a small sample of social media accounts found that around 12% of users have published their phone numbers, 3.4% have published their full address, and 57.6% have full birthday information. So general public should be more vigorous in protecting their personal information.

This research is a part of an ongoing project, which aims to help users proactively monitoring their information disclosed on Internet and assess their privacy risk scores.

II. PII CLASSIFICATION

Keeping PII private is important to ensure the integrity of an individual's identity. In general PII can be classified into two types: sensitive and non-sensitive PII.

- Sensitive PII is personal information, which if lost, compromised, or disclosed without authorization, could result in embarrassment, inconvenience, harm, or unfairness to an individual. PII can become more sensitive when combined with other information. Sensitive PII includes date of birth, passport number, fingerprints, mother's maiden name, driver's license number, credit or debit card number, Social Security number, etc.
- Non-sensitive PII refers to any information that is publicly available. Information such as business phone numbers, gender, business email, and job titles are typically considered non-sensitive PII.

III. PRIVACY LAWS

Stricter laws and regulations should be in place to restrict random dispersion and reckless handling of PII information. There are several federal laws in US that cover the privacy of different types of data, including the Health Insurance Portability and Accountability Act (HIPPA), the Fair Credit Report Act (FCRA), the Family Education Rights and Privacy Act (FERPA), the Gramm-Leach-Bliley Act (GLBA), the Electronic Communications Privacy Act (ECPA), the Children's Online Privacy Protection Act (COPPA), and the Video Privacy Protection Act (VPPA).

- HIPPA requires the creation of national standards to protect sensitive patient health information being disclosed without patient's knowledge.
- FCRA ensures the accuracy, fairness, and privacy of the information in consumer credit bureau file. It regulates the way credit reporting agency can collect, use, and share the data they collect.

- FERPA aims to afford parents the right to have access to their children's education records, and the right to have control over the disclosure of PII from education.
- GLBA requires financial institutions to explain their information sharing practices to their consumers and to safeguard sensitive information.
- ECPA protects individuals against unlawful interception of electronic communications by the federal government or individuals.
- COPPA imposes specific requirements on operators of websites and online services to protect the privacy of children under 13.
- VPPA regulates the disclosure of information about consumers' consumption of video content, imposing prescriptive requirements to obtain consumer's consent to such disclosure.

In addition, some US states have regulations also aims to protect data privacy including the California Consumer Privacy Acts(CCPA), California Consumer Privacy Rights Act(CPRA), Colorado Privacy Act, Connecticut Personal Privacy and Online Monitoring, Utah Consumer Privacy Act, etc.

IV. Our Research

This research is motivated by the fact that it is very easy to acquire a person's PII from the Internet either free or for a fee. We are interested in seeing how much personal PII can be publicly accessible on the Internet. We takes two approaches to mimicking the action of searching a person's information on the Internet. One is to use people search services from online data companies, and the other is to retrieve personal information from publicly accessible data on the Internet.

To implement the second approach, we developed an information retrieval framework that employs natural language processing and entity identification and resolution techniques in order to identify PII attributes in web documents. Given a name, the framework searches for the web documents containing the name, scrapes those documents, and then extracts PIIs.

To avoid privacy violation of random people, this study only uses publicly accessible data. With the people search services, the study only uses authors' names in the search so that information can be easily validated. With Internet search, a group of random names are used.

A. *People Search Engines*

People search engines provide online people search services. When conducting the people search by name, these search engines often give a profile preview on the person(s) of the searched name, such as name, age, relatives, etc. If users want more information, then they need to pay. For the privacy protection purpose, this study only focuses on the profile previews and examines which PII attributes are disclosed in the previews. The fee-based profiles provide rather comprehensive information about a person, including credit report, property records, criminal and traffic records, and so on.

We searched the authors' names on 8 popular data company websites. Table 1 shows the websites and PII attributes returned by the profile previews. Figure 3 shows the PII attributes covered in at least one website's preview and the percentage of the websites that contain those attributes. Besides name, most profile previews also return the age and relatives information. The information on relatives is much noisier and less accurate compared to other attributes.

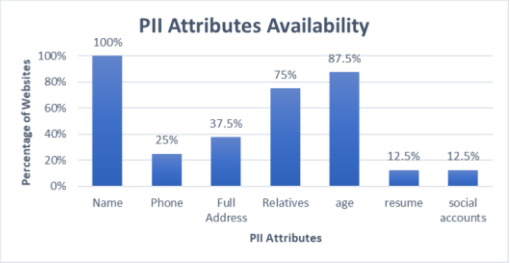

Figure 3: PII attributes disclosed in percentage of websites

Some distinct features of the websites:

- Both *Yellowpages.com* and zabasearch.com are powered by Intelius, Inc, a public records business that provides information services, including people and property search, background checks and reverse phone lookup. However, they all release different information in a profile preview.
- Several companies, including *instantcheckmate.com*, *intelius.com*, *peoplefinders.com*, and *truthfinders.com*, interact with the user during the search process. They all ask a sequence of questions about the person of the searched name to narrow down the search.
- Radaris.*com* searches data from public sources, and it does not possess or access to secure private financial information. It provides free profiles that combines public records with social media and other online mentions. Its profile preview contains more PII attributes including the resume and popular social media accounts such as youtube, flickr, facebook, googleplus, and classmates.
- All companies charge a fee for a person's complete profile. A complete profile from the most companies include information on current and previous address, phone numbers, relatives, age/birth month and year, property records, bankruptcies, judgement and liens, deceased indicator, misdemeanors, criminal check, and sex offender records.

TABLE II. INFORMAITON COVERED IN THE COMPLETE PROFILE BY EACH COMPANY

Data category	Radaris	Peoplefinders	truthfinder	Spokeo	instantcheckmate	intelius
Key personal information	Y	Y	Y	Y	Y	Y
Relatives	Y	Y	Y	Y	Y	Y
Property records	Y	Y	Y	Y	Y	Y
Business/Professional records	Y	Y	Y	Y		
bankruptcies		Y	Y	Y	Y	Y
Financial records			Y	Y	Y	Y
Traffic/criminal records	Y	Y	Y	Y	Y	Y
Government watch					Y	
Social media accounts	Y		Y	Y	Y	Y

TABLE III. PRIVACY POLICY COVERAGE

Privacy Policy Items	Radaris	Peoplefinders	truthfinder	Spokeo	instantcheckmate	intelius
Information collection practice	Y	Y	Y	Y	Y	Y
Information usage and sharing practice	Y	Y	Y	Y	Y	Y
Advertising	Y	Y	Y	Y	Y	Y
User rights	Y	Y	Y	Y	Y	Y
Children's privacy	Y	Y		Y		
CCPA compliance	Y	Y	Y	Y	Y	Y
HIPAA compliance			Y		Y	Y
GLBA compliance			Y		Y	Y
Driver's Privacy Protection Acts Compliance			Y		Y	Y

Data coverage of the websites

Table 2 shows the categories of information covered by a complete profile from each company. Since *zabasearch.com* and *yellowpages.com* are powered by Intelius and generate identical profiles as Intelius, they are not included in the table. Here are the summaries of Table 2:

- The key personal information category contains information on name, current address, phone, email, previous residences, age, and birth month and year. Besides the key personal information, relatives, property records, and traffic/criminal records are also included in every profile.
- *Truthfinder, spekeo, instantcheckmatch, and intelius* provide information in all categories in their profiles except the government watch.
- Only *Radaris* and *peopelfinders* don't provide information on financial records. In addition, *Peoplefinders* doesn't provide social media accounts information.
- Only *Spokeo* profile contains the government watch information.

Terms of Use Policy

People search engines allow users to search a person's information for various reasons such as locating a long-lost relative, discovering details about someone, or simply finding a friend.

Profiles returned by the search engines contain plentiful sensitive and private information, and the uncontrolled dispersion of those information can lead to serious consequences such as reputation damage, privacy breach, and identity theft or fraud. The Terms of Use policy is needed to restrict undesired data usage. By going through profile generation process on each engine, we found that all companies have their Terms of Use policy that users must agree on in order to receive the data. Each company clearly states that users may not use their services for hiring someone, lending money, leasing property, or any other professionally related decisions that are restricted by The Fair Credit Report Act (FCRA).

In addition, when a user purchases data, all companies require the user's name, emails, and credit card information. So a user cannot buy data anonymously. Still it could be possible for cyber criminals to hide their true identity by using the information of stolen PIIs and credit card numbers to purchase the services.

Privacy policy

All companies have a privacy policy that covers many aspects, including data collection practices, data usage and sharing practices, user's rights, and compliances to federal and state privacy laws and regulations. Table 3 shows the key privacy policy items covered by each company. Note that *zabasearch.com* and *yellowpages.com* are not included in the table as they follow intelius' Terms of Use and Privacy Policy.

Among all policy items, children's privacy indicates a company does not knowingly collect personal information from individuals under 18 years of age; and user's right indicates that users have the right to opt out of the sale of their personal information to third parties. A user's request to opt-out will lead to the removal of their profile from a company's service.

Here are the summary of Table 3:

- *Radaris, peoplefinders* have a similar privacy policy which doesn't include compliance statements on HIPPA, GLBA, and Driver's Privacy Protection Acts. *Spokeo* doesn't have those compliance statements either, but it does have specific clauses for EU data subjects.
- *Truthfinders, instantcheckmate,* and *intelius* have a similar privacy policy that covers all listed privacy items except the children's privacy.

The study shows that data companies are becoming more restrictive in releasing personal information thanks to federal and state privacy laws and regulations. In the free profile previews, 5 out 8 companies release information only on name, age, and relatives; 3 out of 8 companies reveals address information, 2 out of 8 companies reveals phone information; only one company reveals social media accounts. No company allows users to buy data anonymously. In addition, all companies have a user rights policy that allows user to opt out of the sale of their information to third parties. At least half of companies claim to be compliant with HIPPA, GLBA, CCPA, and DPPA. We hope privacy laws and regulations will push more companies to improve their practices and privacy policies.

B. PII Information on Internet

The motivation of this approach is to study how much PII information of a person can be found from the public documents on Internet. We searched the information of a group of 19 names on the Internet. To avoid unnecessary noises, we use both name and location information such as city and state in the search. If a person has social media accounts, we also search those accounts for PII information. Currently only Facebook, Twitter, and Instagram accounts are used, and more will be included in the future study. Figure 4 shows PII attributes are discovered on the Internet, and the percentage of people having information on each attribute.

It's rather alarming to see how much PII is publicly accessible on the Internet. Some is stored in public documents or government records, and some is posted by the information owner in their personal websites or social media accounts. Out of 19 people, 26.3% have their address, date of birth, and birthplace information revealed; 47.4% have their cell information revealed, and 63.1% have their email addresses revealed. In addition, 52.6% have their FB accounts revealed, and 42.1% have Instagram accounts revealed. So individuals need be more cautious when disclosing their information and more active in protecting their information.

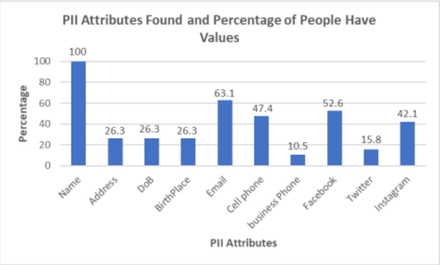

Figure 4: Percentage of people have values on each PII attribute

V. Related Work

This section only focuses on two relative research areas: studies on PII exposure on the Internet and studies on extracting PII attributes from unstructured text.

A. PII Exposure on Internet

Social media becomes one of the major platforms for PII exposure. Studies on user data privacy issues on social media[3] show that social media users often expose various PII information in their posts such as their names, birthdays, full address, telephone numbers, etc. Such information provides perfect opportunities for cyber criminals to exploit their information for identity theft. Most research on analyzing PII exposure on social media is done manually [1][4], still some efforts are made to use automated techniques such as LDA and supervised classification[2][5]. One research aims to systematically identify, collect, and monitor over 1 billion exposed PII records across both the dark web and surface web[14].

B. PII Extraction from Unstructured Documents

Deep learning and natural language processing are two popular techniques for automatic information extraction from unstructured textual data. To extract more fine-grained PII attributes, enhancing word representations with character-based representations are utilized[6][12][17]. In addition, recurrent neural networks and convolutional neural networks are widely used to extract character-level representations[9][11][15][16]. A most recent study proposed the Deep Transfer Learning for PII Extraction (DTL-PIIE) framework to extract users' exposed PII in social media automatically[13]. The framework can facilitate various applications to raise users' privacy awareness such as prediction of PII misuse and privacy risk assessment.

VI. Conclusion and Future Work

Guarding PII is important to ensure the integrity of an individual's identity, and it also prevents people from falling into a victim of identity theft. Strict privacy laws and regulations push data providers to be more diligent in avoiding data breaches and privacy violations in their practices. So individuals become to a weak link in privacy protection. We need to raise general public's privacy awareness so that they would step up in safeguarding their own information.

Identifying and classifying PII attributes in unstructured documents is challenging. When PII of multiple entities appear in the same document, entity disambiguation poses another big challenge. We will investigate advanced natural language processing and deep learning techniques for PII retrieval and entity disambiguation. We plan to develop a tool that helps individuals to monitor their PII dissemination on the internet so that they can proactively protect their privacy.

Acknowledgment

This work is sponsored by NSF EPSCoR grant No. OIa-1946391.

References

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

[2] DHS, "Handbook for Safeguarding Sensitive Personally Identifiable Information," The Privacy Office, US Department of Homeland Security, Washington DC.www.dhs.gov/privacy

[3] 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.

[4] 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.

[5] 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.

[6] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, "Deep learning with word embeddings improves biomedical named entity recognition," *Bioinformatics*, vol. 33, no. 14, pp. i37–i48, 2017.

[7] https://javelinstrategy.com/

[8] 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.

[9] Z. Jie and W. Lu, "Dependency-guided LSTM-CRF for named entity recognition," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3862-3872.

[10] B. Krishnamurthy and C. E. Wills, "On the leakage of personally identifiable information via online social networks," in *Proceedings of the 2nd ACM workshop on Online social networks*, 2009, pp. 7–12.

[11] J. Y. Lee, F. Dernoncourt, and P. Szolovits, "Transfer learning for named-entity recognition with neural networks," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[12] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," *arXiv Prepr. arXiv1603.01360*, 2016

[13] Y. Liu, Y. L. Fang, M. Ebrahimi, W. Li, and H. Chen. "Automated PII Extraction from Social Media for Raising Privacy Awareness: A Deep Transfer Learning Approach." In *2021 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 1-6. IEEE, 2021.

[14] Y. Liu *et al*., "Identifying, Collecting, and Monitoring Personally Identifiable Information: From the Dark Web to the Surface Web," *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2020, pp. 1-6, doi: 10.1109/ISI49825.2020.9280540.

[15] S. Moon, L. Neves, and V. Carvalho, "Multimodal named entity recognition for short social media posts," *arXiv Prepr. arXiv1802.07862*, 2018.

[16] Y. Nie, Y. Tian, Y. Song, X. Ao, and X. Wan, "Improving named entity recognition with attentive ensemble of syntactic information," *arXiv Prepr. arXiv2010.15466*, 2020.

[17] T. Schnabel, I. Labutov, D. Mimno, and T. Joachims, "Evaluation methods for unsupervised word embeddings," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 298–307.

[18] P. M. Schwartz and D. J. Solove, "The PII Problem: Privacy and a New Concept of Personal Identifiable Information," New York University Law Review, Vol. 86, 2011, p. 1814.