Linking Personally Identifiable Information from the Dark Web to the Surface Web: A Deep Entity Resolution Approach

Fangyu Lin
Artificial Intelligence Lab
Department of Management
Information Systems
University of Arizona
Tucson, Arizona, USA
fylin@email.arizona.edu

Yizhi Liu
Artificial Intelligence Lab
Department of Management
Information Systems
University of Arizona
Tucson, Arizona, USA
yizhilu@email.arizona.edu

Mohammadreza Ebrahimi
Artificial Intelligence Lab
Department of Management
Information Systems
University of Arizona
Tucson, Arizona, USA
ebrahimi@email.arizona.edu

Zara Ahmad-Post
Artificial Intelligence Lab
Department of Management
Information Systems
University of Arizona
Tucson, Arizona, USA
zahmadpost@email.arizona.edu

James Lee Hu
Artificial Intelligence Lab
Department of Management
Information Systems
University of Arizona
Tucson, Arizona, USA
jameshu@email.arizona.edu

Jingyu Xin
Artificial Intelligence Lab
Department of Management
Information Systems
University of Arizona
Tucson, Arizona, USA
jingyuxin@email.arizona.edu

Sagar Samtani
Department of Operations and Decision Technologies
Indiana University
Bloomington, Indiana, USA
ssamtani@iu.edu

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

Hsinchun Chen
Artificial Intelligence Lab
Department of Management
Information Systems
University of Arizona
Tucson, Arizona, USA
hchen@eller.arizona.edu

Abstract—The information privacy of the Internet users has become a major societal concern. The rapid growth of online services increases the risk of unauthorized access to Personally Identifiable Information (PII) of at-risk populations, who are unaware of their PII exposure. To proactively identify online at-risk populations and increase their privacy awareness, it is crucial to conduct a holistic privacy risk assessment across the internet. Current privacy risk assessment studies are limited to a single platform within either the surface web or the dark web. A comprehensive privacy risk assessment requires matching exposed PII on heterogeneous online platforms across the surface web and the dark web. However, due to the incompleteness and inaccuracy of PII records in each platform, linking the exposed PII to users is a non-trivial task. While Entity Resolution (ER) techniques can be used to facilitate this task, they often require ad-hoc, manual rule development and feature engineering. Recently, Deep Learning (DL)-based ER has outperformed manual entity matching rules by automatically extracting prominent features from incomplete or inaccurate records. In this study, we enhance the existing privacy risk assessment with a DL-based ER method, namely Multi-Context Attention (MCA), to comprehensively evaluate individuals’ PII exposure across the different online platforms in the dark web and surface web. Evaluation against benchmark ER models indicates the efficacy of MCA. Using MCA on a random sample of data breach victims in the dark web, we are able to identify 4.3% of the victims on the surface web platforms and calculate their privacy risk scores.

Keywords—Privacy, PII, Data breach, Dark web, Surface web, Data collection

I. INTRODUCTION

The proliferation of social media, e-commerce, and other online platforms has introduced new privacy concerns. Unauthorized access to Personally Identifiable Information (PII) of individuals is among the most important privacy concerns [1]. General Data Protection Regulation (GDPR) defines PII as “any information relating to an identified or identifiable natural person” [2]. Among the top five most compromised data types [3], emails, names, and physical addresses are listed as PII in GDPR. Internet users share a significant amount of PII across different online platforms. Additionally, government and local authorities also share PII in publicly accessible databases [4]. PII that is shared and stored electronically is vulnerable to data breach. In 2019, 5.182 data breaches were reported for a total of 7.9 billion records, resulting in an increase of 33% since 2018 [3]. The PII leaked in data breach incidents is a major societal concern as it can lead to significant financial loss, reputation damage, and safety threats [5].

To combat the issue, privacy researchers aim to identify at-risk populations and the factors that can encourage them to take protective actions. For example, the elderly and children are often more vulnerable due to their tendency to underestimate potential risks [6], [7]. Increasing the awareness of at-risk populations requires that they are proactively informed of the extent to which their PII is exposed across the online platforms [8]. Accordingly, there have been several established quantitative measures of PII exposure in the privacy risk assessment literature [9], [10]. However, the established measures are limited to analyzing data from a single source or multiple homogeneous sources.
within the surface web or within the dark web. This may lead to inaccurate assessment of the privacy risk. The effective assessment of PII exposure requires accessing a comprehensive set of exposed PII records on the internet. Analyzing heterogeneous platforms across online spaces, including the dark web and surface web, can facilitate a comprehensive assessment of users’ data exposure.

The dark web is a valuable source of intelligence for breached PII data [11]. Dark web platforms such as Dark Net Marketplaces (DNMs) and carding shops are known as early indicators for PII data breaches [11]. A multitude of stolen credentials (e.g., credit card numbers or social security numbers) are advertised for sale on these platforms. The type of available PII varies on different platforms. Moreover, some platforms may provide only a specific part of a PII (e.g., the first 4 digits of SSN). Partial PII from the dark web can be linked to other publicly available PII from the surface web data sources [5], including conventional search engines and people search engines from the surface web. Linking the information from the dark web platforms and surface web can help provide a complete picture of an individual’s PII exposure. To this end, it is crucial to identify the real-world user whose PII has been exposed across the dark web and the surface web. However, due to the incompleteness and inaccuracy of PII records in each platform, linking the exposed PII to users is a non-trivial task [12]. Entity Resolution (ER) is a long-standing stream of research that can be suitable in linking exposed PII from the dark web or surface web to real-world users. Among the prevailing ER methods, Deep Learning (DL)-based models have emerged as a promising technique due to their capability to automatically extract features for linking records from multiple sources. Compared to traditional ER models that require manual feature engineering and entity matching rule development, DL-based models have shown to be more efficient and less labor intensive [13].

In this study, we propose a cross-platform privacy risk assessment framework to analyze PII exposure across the internet, including the dark web and surface web. At the center of our framework stands an advanced DL-based ER technique, known as Multi-Context Attention (MCA) model [14] to link heterogeneous PII records. We adopt MCA due to its robustness in processing incomplete, inaccurate, and redundant records. This is achieved by leveraging three attention mechanisms to focus on the most discriminative features from individual records, pairs of records, and the whole training dataset. We evaluate MCA against the state-of-the-art DL-based and traditional machine learning models. Our experiment result demonstrates that MCA outperformed the state-of-the-art machine learning and DL models for matching PII from the surface web to the dark web. We further present a case study based on the matched entities identified by MCA to illustrate the utility of the proposed framework in assessing users’ privacy risk.

The remainder of this paper is organized as follows. In Section II, we review the privacy paradox and at-risk populations, privacy risk assessment, and entity resolution literature. In Section III, we identify research gaps and propose research questions. Section IV details our proposed cross-platform privacy risk assessment framework. Section V presents benchmark evaluation results. Section VI summarizes key findings from a case study. Finally, Section VII concludes this study and presents future research directions.

II. LITERATURE REVIEW

Three relevant areas of the literature are reviewed. First, we examine the privacy paradox and at-risk population literature to understand at-risk populations’ perception of the privacy risk. Second, we summarize how information privacy risk was assessed in previous studies. Finally, we identify emerging entity resolution techniques to link exposed PII records across the dark web and surface web for assessing users’ overall privacy risk.

A. Privacy Paradox and At-risk Populations

The privacy paradox describes a phenomenon wherein users are significantly concerned about their information privacy. Meanwhile, they engage in online behaviors that disclose their personal information [1]. Users are subject to the privacy paradox due to biased perception, unawareness, and underestimation of the probability of data breach [6], [7]. The elderly, children, and other marginalized groups are among at-risk populations. The elderly’s increasing dependence on telemedicine brings convenience to their life. However, their information privacy is sacrificed by sharing health records via IoT devices and the Internet [6]. Children are also at risk of data exposure due to their parents’ lacking awareness of sensitivity of children’s personal information [7]. Since at-risk populations are particularly vulnerable to the negative impacts of PII exposures (e.g., safety threats), it is critical to proactively prevent them from excessive PII exposure.

Increasing users’ privacy awareness helps resolve the privacy paradox [8]. It has been shown that users implement protective measures if they recognize their privacy has been compromised [15]. A comprehensive assessment of exposed information across both the surface web and the dark web could help users recognize their compromised privacy. Therefore, there is a critical need to comprehensively assess the extent of individuals’ exposure and proactively inform them. Next, we review the privacy risk assessment literature to examine prevailing methods in evaluating users’ online information exposure.

B. Privacy Risk Assessment

Previous work has assessed privacy risk from three aspects: data de-anonymization, visibility, and sensitivity. Studies on data de-anonymization mimic an adversary who aims to identify the owners of the anonymized data and infer their sensitive information [16]. Organizations and security engineers can benefit from the results of these studies to improve their anonymization measures or restrict publishing sensitive information. Visibility studies examine the number of people with whom users share their sensitive information [17], [18]. This stream of research is mainly focused on preserving the shared content in social media and mobile apps via controlling privacy settings to limit access to the personal data by others. Providing a visibility score helps users better understand their privacy level and establish an appropriate privacy setting. Sensitivity studies measure privacy risk associated with the exposed PII attributes [10], [17]. Some attributes such as contact numbers are more damaging to the individuals’ privacy because they can be used to uniquely identify them or can directly interfere with their personal life. Sensitivity scores help users decide whether to share certain sensitive
information with others. Unlike sensitivity studies, data de-anonymization and visibility studies evaluate the privacy risk assuming the protection measures are already in place. Since such assumption does not hold for publicly available data on the dark web, we focus on sensitivity measures in this study.

Among sensitivity measures, Privacy Score is one of the first measures adopted for assessing privacy risk [9]. It was developed based on users’ perceptions of information disclosure in their social media profiles. The fewer the people willing to disclose a particular type of information, the higher the sensitivity of that information. Built upon Privacy Score, Srivastava et al. [10] considered PII attributes exposed in social media profiles, posts, and comments. They derived sensitivity scores for 11 common attributes (e.g., contact number, e-mail, etc.), based on the willingness of the users to share them. Srivastava et al.’s [10] enhanced Privacy Score has been extensively adopted by privacy risk assessment studies [17] to analyze the privacy risk on online platforms. However, previous studies focus on a single platform or multiple homogeneous platforms [9], [10], which may underestimate the extent of PII exposure. Overall, there is a dearth of work examining privacy risk across both the dark web and surface web. This can be due to the challenge of identifying corresponding individuals on heterogeneous platforms given the incomplete and inaccurate PII obtained from each platform. ER is necessary to address this challenge. Next, we review emerging ER techniques.

C. Entity Resolution (ER)

ER aims to identify different records from homogeneous or heterogeneous sources that refer to the same real-world entity [12]. These records are mainly divided into three types: structured data, textual data, and dirty data. Structured data contains pre-defined attribute sets with short and atomic values. Textual data includes raw text without pre-defined attribute sets. Dirty data contains pre-defined attribute sets with missing or incorrectly injected values.

The prevailing methods to process these three types of records can be categorized as schema-aware and schema-agnostic based on the presence of a pre-defined attribute set. Schema-aware techniques, including rule-based methods and conventional similarity metrics, achieve desirable performance in linking structured records that belong to the same entity [12]. However, they are not suitable for resolving textual and dirty data due to their lack of ability to automatically detect the corresponding attributes from two records. On the contrary, schema-agnostic methods consider each record as a bag of values that is not limited by pre-defined and fixed attribute sets. Given that breached PII records align with the characteristics of textual and dirty data, we focus on identifying emerging schema-agnostic techniques due to their flexibility in dealing with incomplete and inaccurate attributes.

Among the methods for schema-agnostic ER, emerging deep learning (DL)-based models are promising due to their ability to automatically extract important features from heterogeneous records. For example, DITTO [19], a DL-based model built on Google Bidirectional Encoder Representations from Transformers (BERT) [20]. BERT is a pre-trained deep bidirectional representation model that can be tuned to resolve a wide range of NLP tasks (e.g., classification). However, it requires to be pre-trained on large English Wikipedia corpora. Due to the scale of model parameters and the large size of training data, DITTO’s training process is time-consuming. DeepER is a Bi-LSTM binary classification that is able to capture the semantics from the local context for ER purposes [21]. DeepMatcher extends DeepER with an additional pair-attention layer [13]. The pair-attention layer analyzes two given records jointly and weights matching features between the records. MCA is built upon DeepMatcher and leverages a self-attention mechanism to disambiguate the meaning of the words within an exposed PII record. In addition, it adopts a global-attention mechanism to identify discriminative attribute values in the training corpus [14]. MCA has been shown to outperform the state-of-the-art ER models in structured, textual, and dirty datasets. Despite MCA’s potential in ER, it is unclear how it can be adopted for privacy risk assessment applications.

III. RESEARCH GAPS AND QUESTIONS

Past research evaluates users’ privacy risk from different aspects, aiming to increase the privacy awareness of individuals. However, the existing privacy analytics methods often assume that protection mechanisms (e.g., anonymization or privacy settings) already exist and they often focus on a single type of platform for risk assessment. There is limited work on leveraging ER to assess privacy risk across the dark web and surface web. In light of these limitations, we pose the following research questions:

- How can privacy risk assessment be improved by a DL-based ER system that functions across heterogeneous platforms in the dark web and surface web?
- To what extent are data breach victims susceptible to data exposure on different platforms across the dark web and the surface web?

In order to address these questions, we propose a novel cross-platform privacy risk assessment framework to link exposed PII from the dark web and surface web for privacy risk assessment.

IV. RESEARCH DESIGN

The proposed cross-platform privacy risk assessment framework consists of three major components: Data Collection, Entity Resolution, and Privacy Risk Assessment. Fig. 1 illustrates our proposed framework.

![Fig. 1. Our proposed cross-platform privacy risk assessment framework](image)

The Data Collection component aims to collect and pre-process the exposed PII records from heterogeneous online platforms across the dark web and surface web. For entity resolution, we leverage MCA, a promising DL-based ER model, to automatically link exposed PII records between the two collections. Lastly, to conduct a privacy risk...
assessment, we leverage the records linked by MCA and calculate individuals’ privacy risk score.

A. Data Collection

1) Dark Web Collection

As noted, DNMs and carding shops in the dark web are preeminent sources for breached PII data [11]. Using a well-known dark web directory, DeepDotWeb, we obtain a list of reputable dark web platforms for collection. We select platforms based on the amount of available exposed PII of data breach victims. To collect exposed PII from the selected platforms, we develop a Python crawler to crawl and parse the webpages and extract exposed PII records on each page into a SQL database.

2) Surface Web Collection

We extended our dark web collection to the surface web to estimate PII exposure of the data breach victims. Among surface web platforms, people search engines are specifically designed to collect personal information from social media, public records, proprietary databases, etc. While such platforms can be extensively used by law enforcement and intelligence services to acquire comprehensive details about targeted individuals, they are also accessible to the public and are therefore major sources of exposed PII [4]. Collecting data from people search engines using their provided search interface can be inefficient due to the long response time. Thus, we develop a crawler that utilizes Google search to acquire the search result pages from these platforms. Our crawler operates in three steps. First, the full name and city of a given victim and the name of the people search engine are used as a query for Google search. Second, the URLs from the retrieved results are crawled and extracted. Third, the crawler visits each URL separately and downloads the page.

3) Data Pre-processing

As PII usually does not contain symbols (with an exception of email addresses), we remove non-alphanumeric characters from the attributes of our collected data. Additionally, all characters are converted to lowercase, and all abbreviations are spelled out using a set of predefined rules. We further remove stop words to reduce the influence of these words on entity resolution. Lastly, to normalize the corpus, we leverage WordNet Lemmatizer [22] and acquire a unified form for words with different spellings.

B. Entity Resolution (ER): Multi-Context Attention Model (MCA)

The Entity Resolution component aims to automatically link the exposed PII records from the surface web to the dark web. In particular, we utilize a state-of-the-art DL-based Multi-Context Attention (MCA) model for matching records. MCA formulates ER as a binary classification problem, where the model takes a pair of records as input and predicts whether the records match (i.e., refer to the same entity). MCA considers record pairs in a bag-of-words fashion and leverages semantic contexts to extract salient features for matching different types and formats of records across heterogeneous sources. This enables MCA to achieve high performance in analyzing both textual and dirty data types. MCA is built upon four major components: Bi-GRU, self-attention, pair-attention, and global attention. Bi-GRU captures the semantic of each word based on its local context. Self-attention facilitates word disambiguation by capturing the relationship between a word and the record it belongs to. Pair-attention jointly learns from the record pair to automatically identify key words that are critical in determining a matched pair. Unlike the traditional schema-aware technique, key words can locate in any position in two records and under different attribute names. Global-attention emphasizes highly discriminative words in the entire training dataset. For example, if all the records in the training dataset contain different city attributes, city name can be considered as highly discriminative. Such discriminative words contribute to uniquely matching the input records. Leveraging these four components, MCA is robust to noise and thus, is useful to analyze textual and dirty datasets.

Fig. 2 shows the main components in the MCA model. As shown in Fig. 2, MCA operates in two stages: Representation Generation stage and Binary Classification stage. In the Representation Generation stage, two attention-based neural network encoders generate a representation for each record in a pair, separately. These representations capture discriminative features within the record attributes, between the record pairs, and in the entire training data. Next, in the Binary Classification stage, these two representations are jointly used as features to predict whether the records match or not. Each stage is detailed as follows.

Fig. 2. Multi-Context Attention Model (adopted with modification from [14])
Each encoder for representation generation consists of seven components. First, given two input record $R^A$ and $R^B$, each word in $R^A$ and $R^B$ is represented by word embeddings $w^A_i$ and $w^B_i$, respectively.

$$R^A = \{w^A_1, w^A_2, ..., w^A_m\} \quad (1)$$
$$R^B = \{w^B_1, w^B_2, ..., w^B_n\} \quad (2)$$

where $m$ and $n$ denote the number of words in $R^A$ and $R^B$, respectively. Second, a Bi-GRU extracts the forward and backward context information for each word. Each node in the Bi-GRU generates a directional hidden state $\tilde{h}_i$ (or $\tilde{h}^{-}_i$) based on word embedding $w^A_i$ and $w^B_i$ and the previous hidden state $h_{i-1}$ (or $\tilde{h}^{-}_{i-1}$). A comprehensive summary of the current hidden state $h_i$ is generated by concatenating $\tilde{h}_i$ and $\tilde{h}^{-}_i$. Third, a self-attention mechanism learns a global context that can disambiguate the meaning of each word in the context of the whole record. The self-attention component takes the Bi-GRU hidden states $H = (h_1, h_2, ..., h_n)$ as input and calculates self-attention embeddings $S^A$ as follows (Given that all formulations for $R^B$ are identical to those of $R^A$, we only provide the equations for $R^A$).

$$S^A = H^A \cdot \text{softmax}(H^A^T H^A) \quad (3)$$

Forth, a pair-attention mechanism assigns weights to each element in the self-attention embedding $S^B$ based on the similarity between word embeddings $w^A$ in record $R^A$ and word embeddings $w^B$ in record $R^B$ and generates a pair-attention embedding $P^A$.

$$P^A = S^B \cdot \text{softmax}\left((w^A)^T W_p w^B\right) \quad (4)$$

where $W_p$ is a trainable weight matrix. Fifth, a fusion module aggregates the self-attention and pair-attention embeddings. It takes $S$ and $P$ as input and output a fused embedding $f$.

$$f = \text{Highway}(\{S, P, |S-P|, S \odot P\}) \quad (5)$$

where $|S-P|$ is the difference between $S$ and $P$. Also, $S \odot P$ is an element-wise product, and $[\cdot]$ denotes vector concatenation. $\text{Highway}()$ is a feed-forward neural network with a transform gate, which regulates the amount of information that is passed to the output [23]. Sixth, a weight matrix $A^A$ is learned as the non-linear combination of the fused embedding $f^A$ and the word embedding $w^A$. Subsequently, a gate $g^A$ uses this weight matrix to control the extent to which the output is affected by $w^A$ and $f^A$.

$$A^A = \sigma(W_1 w^A + W_2 f^A + b) \quad (6)$$
$$g^A = A^A \cdot w^A + (1 - A^A) \cdot f^A \quad (7)$$

where $W_1$ and $W_2$ are trainable weight matrices. Finally, in the seventh step, a global-attention mechanism learns to emphasize highly discriminative words within the training dataset. This attention mechanism takes $g^A$ as input and generates a global-attention embedding $G^A$.

$$G^A = g^A \cdot \text{softmax}\left(g^A^T \cdot c_x\right) \quad (8)$$

where $c_x$ is a trainable global context vector.

In the Binary Classification stage, two generated representations $G^A$ and $G^B$ from the Representation Generation stage are fed into a fusion layer. The fusion component from the binary classification stage follows a similar operation to the one in the record encoder and generate a unified embedding for the input record pair. This embedding is the input for a binary classifier to predict whether $R^A$ and $R^B$ match.

### C. Privacy Risk Assessment

As noted in our review of privacy risk assessment measures, the Sensitivity Score developed by Srivastava et al. [10] is suitable for privacy risk assessment in the dark web and surface web. Accordingly, we adopt Sensitivity Scores of 11 PII attributes developed in [10] and calculate the privacy risk score for each individual. These sensitivity scores are shown in Table I.

<table>
<thead>
<tr>
<th>PII Attribute</th>
<th>Sensitivity Score ($\beta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contact Number</td>
<td>0.6</td>
</tr>
<tr>
<td>Email</td>
<td>0.1833</td>
</tr>
<tr>
<td>Address</td>
<td>0.1166</td>
</tr>
<tr>
<td>Birthdate</td>
<td>0.1166</td>
</tr>
<tr>
<td>Hometown</td>
<td>0.1166</td>
</tr>
<tr>
<td>Current Town</td>
<td>0.1166</td>
</tr>
<tr>
<td>Job Details</td>
<td>0.2</td>
</tr>
<tr>
<td>Relationship Status</td>
<td>0.4166</td>
</tr>
<tr>
<td>Interests</td>
<td>0.3</td>
</tr>
<tr>
<td>Religious Views</td>
<td>0.5666</td>
</tr>
<tr>
<td>Political Views</td>
<td>0.6833</td>
</tr>
</tbody>
</table>

Each score ranges between 0 and 1 and reflects the extent to which users are willing to share the corresponding PII attribute with the public. A higher sensitivity score suggests that the privacy risk posed by information exposure is higher. For each individual, overall privacy risk score $PR$ is calculated by summing up the scores $\beta_i$ of exposed attributes across the dark web and the surface web. The resultant score can be leveraged for subsequent useful analytics. We conduct a case study to demonstrate the utility of the proposed framework. Our PII dataset was used in the case study to examine the privacy risk of selected users based on their overall privacy risk score $PR$.

### V. Evaluation

#### A. Ground Truth Dataset Development

We selected BuySSN platform as the source of the dark web collection since it provided a significant amount of PII of data breach victims. Table II summarizes our dark web collection.

<table>
<thead>
<tr>
<th>Data Sources</th>
<th>BuySSN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Product information of stolen SSNs</td>
</tr>
<tr>
<td>Number of Records</td>
<td>54,912</td>
</tr>
<tr>
<td>Date Ranges</td>
<td>5/2014 – 1/2018</td>
</tr>
</tbody>
</table>

We focused on the elderly as one of the major at-risk populations. Accordingly, consistent with [24], we randomly selected one thousand records from our dark web collection, for which the age attribute was greater than 65 and used them as seeds to search on the surface web.

Two people search engines (My Life and Been Verified) served as our primary surface web data source as they
provide a rich set of PII for a significant number of individuals [4], [5]. The full name and city attributes of records from the dark web collection were used to query the two mentioned people search engines on the surface web. Consequently, 66 result pages encompassing a total of 1,014 records were obtained from the people search engines. These records included sensitive PII such as physical address, email, phone number, etc. Each record in this set was a potential candidate for being the true individual whose stolen SSN was sold on the BuySSN dark web platform.

We pre-processed our dark web and surface web collections via the steps described in Section IV-A. We then developed a ground-truth PII dataset based on two collections for performing ER and evaluating benchmark ER models. To this end, we formed a panel of two privacy experts to manually match the records form our candidate set (1,014 records) with the information of the victims whose SSNs were breached on the dark web. The annotators considered the similarity of attributes such as name, city, state, and age. The agreement rate between the annotators was above 90%. As a result, out of our initial 1,014 candidate records, 52 records (more than 5%) were matched with the dark web data.

B. Entity Resolution Benchmark Evaluation

We evaluated MCA on benchmark structured, textual, and dirty datasets, and PII ground-truth datasets. Our constructed PII dataset is considered as a textual and dirty dataset. In addition to the PII dataset, we carefully selected three benchmark datasets: Abt-Buy (AB), Amazon-Google (AG), and DBLP-ACM (DA). Each has been extensively used in the past entity resolution literature [13], [14], [21]. The statistics of these four datasets are summarized in Table III. Well-established metrics for entity resolution included precision, recall, and F1-score [14], [21]. Hyper-parameters were tuned through 5-fold cross-validation [25]. MCA was compared against four state-of-the-art ER models, including a machine learning-based model and three prevailing deep learning-based models. TF-IDF combined with cosine similarity is a conventional frequency-based machine learning model, which predicts matched records based on the similarity of the attribute values between records. Three deep learning-based models were DITTO, DeepER and DeepMatcher, which were selected earlier. Experiments were conducted on a single Microsoft Windows 10 Pro Server with 128GB of RAM, an Nvidia GeForce GTX 1070 GPU, and an Intel CPU at 2.60 Gigahertz (GHz).

We compared MCA to the state-of-the-art ER benchmark methods in terms of precision, recall, and F1-score. As shown in Table IV, the frequency-based method yielded the highest performance on the structured dataset (AB). This indicates that when the dataset is clean and attributes are well-defined, the conventional machine learning technique has a competitive performance. However, real-world datasets often contain noise or lack pre-defined attribute sets; and thus, require methods with the ability to process dirty and textual datasets.

In the textual dataset (AG), the sales information about software products contained a large portion of domain-specific proper nouns. However, DeepER, DeepMatcher, and MCA lacked enough training data to learn the semantics of these domain-specific proper nouns. On the other hand, DITTO that was pre-trained on English Wikipedia corpora, can better distinguish these proper nouns. The high performance of the frequency-based method is due to directly matching the proper nouns in the record pairs, which is effective in both structured and textual datasets. For the PII dataset, fewer domain-specific proper nouns were involved. Therefore, it is not necessary to learn from a large-scale domain-specific corpus.

MCA outperformed the state-of-the-art methods in both DA (dirty) and PII datasets in F1-Score (98.31% and 88.89%, respectively). Compared to DeepMatcher, which achieved the second highest performance in F1-score (98.09% and 85.93%), MCA leveraged additional self-attention mechanism to disambiguate words’ semantics and global-attention to emphasize discriminative words. As a result, MCA was more effective in processing different types and formats of PII data and was able to better leverage the context of the shared attributes between a pair of records.

### Table III. Statistics of Benchmark Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type of dataset</th>
<th># of Records</th>
<th># of Training Records</th>
<th># of Validation Records</th>
<th># of Testing Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abt-Buy (Structured)</td>
<td>Product</td>
<td>9,574</td>
<td>5,742</td>
<td>1,916</td>
<td>1,916</td>
</tr>
<tr>
<td>Amazon-Google (AG)</td>
<td>Software</td>
<td>11,461</td>
<td>6,874</td>
<td>2,294</td>
<td>2,293</td>
</tr>
<tr>
<td>DBLP-ACM (DA)</td>
<td>Citation</td>
<td>12,903</td>
<td>7,417</td>
<td>2,743</td>
<td>2,743</td>
</tr>
<tr>
<td>PII</td>
<td>PII</td>
<td>1,014</td>
<td>812</td>
<td>103</td>
<td>102</td>
</tr>
</tbody>
</table>

### Table IV. Experiment Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Precision %</th>
<th>Recall %</th>
<th>F1-Score %</th>
<th>Dataset</th>
<th>Model</th>
<th>Precision %</th>
<th>Recall %</th>
<th>F1-Score %</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB (Structured)</td>
<td>TF-IDF</td>
<td>83.68</td>
<td>74.91</td>
<td>78.42</td>
<td>AG (Textual)</td>
<td>TF-IDF</td>
<td>73.27</td>
<td>72.72</td>
<td>72.99</td>
</tr>
<tr>
<td></td>
<td>DeepER</td>
<td>69.83</td>
<td>58.30</td>
<td>62.33</td>
<td></td>
<td>DeepER</td>
<td>65.72</td>
<td>58.31</td>
<td>61.79</td>
</tr>
<tr>
<td></td>
<td>DITTO</td>
<td>73.30</td>
<td>63.12</td>
<td>67.87</td>
<td></td>
<td>DITTO</td>
<td>79.39</td>
<td>71.24</td>
<td>74.46</td>
</tr>
<tr>
<td></td>
<td>DeepMatcher</td>
<td>64.39</td>
<td>64.08</td>
<td>64.23</td>
<td></td>
<td>DeepMatcher</td>
<td>55.38</td>
<td>76.92</td>
<td>64.40</td>
</tr>
<tr>
<td></td>
<td>MCA</td>
<td>71.88</td>
<td>66.99</td>
<td>69.35</td>
<td></td>
<td>MCA</td>
<td>67.56</td>
<td>73.45</td>
<td>70.38</td>
</tr>
<tr>
<td>DA (Dirty)</td>
<td>TF-IDF</td>
<td>97.43</td>
<td>94.42</td>
<td>95.63</td>
<td>PII (Ours)</td>
<td>TF-IDF</td>
<td>70.67</td>
<td>87.42</td>
<td>76.03</td>
</tr>
<tr>
<td></td>
<td>DeepER</td>
<td>95.38</td>
<td>87.69</td>
<td>90.95</td>
<td></td>
<td>DeepER</td>
<td>86.47</td>
<td>79.48</td>
<td>82.56</td>
</tr>
<tr>
<td></td>
<td>DITTO</td>
<td>95.35</td>
<td>90.23</td>
<td>92.33</td>
<td></td>
<td>DITTO</td>
<td>88.12</td>
<td>75.50</td>
<td>80.70</td>
</tr>
<tr>
<td></td>
<td>DeepMatcher</td>
<td>97.76</td>
<td>98.42</td>
<td>98.09</td>
<td></td>
<td>DeepMatcher</td>
<td>93.33</td>
<td>80.00</td>
<td>85.93</td>
</tr>
<tr>
<td></td>
<td>MCA</td>
<td>98.20</td>
<td>98.42</td>
<td>98.31</td>
<td></td>
<td>MCA</td>
<td>100.00</td>
<td>80.00</td>
<td>88.89</td>
</tr>
</tbody>
</table>
VI. CASE STUDY: PRIVACY RISK ASSESSMENT ON SURFACE WEB PLATFORMS

The strong performance of MCA in matching PII records allowed us to conduct privacy risk assessment across the dark web and the surface web. Using MCA on a random sample of one thousand data breach victims in the dark web collection, we were able to identify 43 of the victims on the surface web platforms and calculate their privacy risk scores. Overall, 4.3% of data breach victims had their PII exposed on two selected surface web platforms (i.e., My Life and Been Verified). Been Verified returned 36 matched records and My Life returned 14 matched records. We show the average privacy risk score for each platform in Fig. 3.

![Boxplot of Privacy Risk Score across Two Surface Web Platforms](image)

**Fig. 3.** Boxplot of Privacy Risk Score across Two Surface Web Platforms

In Fig. 3, the y-axis denotes the privacy risk score. The x-axis lists the platforms. The black bars show the minimum and maximum of privacy risk score for each platform. The green bars show the average privacy risk scores. The Data breach victims on Been Verified had an average of 0.72 privacy risk score with 0 standard deviation. This was because Been Verified returned contact numbers and the current town for every matched record. For My Life, the privacy risk score ranged from 0.58 to 2.25. My Life returned birthdate, hometown, relationship status, religious views, and political views. However, the available attributes varied for each individual. Although My Life returned fewer victims, the privacy risk of those victims was significantly higher than the risk for victims whose data was exposed on Been Verified. This finding suggests that an effective protective measure for data breach victims could be to consider requesting My Life service provider to remove their personal information in order to reduce their privacy risk.

VII. CONCLUSION AND FUTURE DIRECTIONS

With the rise of information privacy concerns, researchers have suggested increasing user's awareness in at-risk populations, who have their Personally Identifiable Information exposed on the internet. Identifying at-risk population requires a holistic privacy assessment across both the dark web and surface web. Current privacy risk assessment studies are limited to a single platform within either the surface web or the dark web. In this study, we enhance the existing privacy risk assessment with a DL-based ER method, MCA, to comprehensively evaluate individuals' PII exposure across different online platforms in the dark web and surface web. Using MCA, we were able to connect partial PII from carding shops and people search engines to thoroughly assess data breach victims’ privacy risk. Finally, we demonstrated the utility of the proposed framework via a case study. Users, governments, and security experts can leverage the proposed framework to identify at-risk populations and better protect PII across the internet.

There are several promising directions for future research. First, different types of additional platforms, such as social media, can be leveraged for holistic privacy risk assessment. Future work can implement a multi-view entity resolution model to handle heterogeneous data. Second, a privacy assistant tool built upon our proposed framework can be developed to proactively inform data breach victims of their data exposure.

ACKNOWLEDGEMENT

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

REFERENCES


