Incognito: A Method for Obfuscating Web Data

Rahat Masood
Data61-CSIRO and UNSW
Sydney, NSW, Australia
rahat.masood@student.unsw.edu.au

Dinusha Vatsalan
Data61-CSIRO
Sydney, NSW, Australia
dinusha.vatsalan@data61.csiro.au

Muhammad Ikram
Data61-CSIRO and UNSW
Sydney, NSW, Australia
muhammad.ikram@data61.csiro.au

Mohamed Ali Kaafar
Data61-CSIRO, Macquarie University, and Optus
Macquarie University Cyber Security Hub
Sydney, NSW, Australia
dali.kaafar@mq.edu.au

ABSTRACT

Users leave a trail of their personal data, interests, and intents while surfing or sharing information on the Web. Web data could therefore reveal some private/sensitive information about users based on inference analysis. The possible identification of information corresponding to a single individual by an inference attack holds true even if the user identifiers are encoded or removed in the Web data. Several works have been done on improving privacy of Web data through obfuscation methods [7, 12, 18, 32]. However, these methods are neither comprehensive, generic to be applicable to any Web data, nor effective against adversarial attacks. To this end, we propose a privacy-aware obfuscation method for Web data addressing these identified drawbacks of existing methods. We use probabilistic methods to predict privacy risk of Web data that incorporates all key privacy aspects, which are uniqueness, uniformity, and linkability of Web data. The Web data with high predicted risk are then obfuscated by our method to minimize the privacy risk using semantically similar data. Our method is resistant against adversary who has knowledge about the datasets and model learned risk probabilities using differential privacy-based noise addition. Experimental study conducted on two real Web datasets validates the significance and efficacy of our method. Our results indicate that the average privacy risk reaches to 100% with a minimum of 10 sensitive Web entries, while at most 0% privacy risk could be attained with our obfuscation method at the cost of average utility loss of 64.3%.

CCS CONCEPTS

• **Security and privacy → Privacy-preserving protocols**; *Data anonymization and sanitization*; *Privacy protections*;

KEYWORDS

Web Data Privacy, Privacy Risk Evaluation, Data Obfuscation, Adversarial Machine Learning, Probabilistic Model, Semantic Similarity

WWW 2018, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5639-8/18/04.
https://doi.org/10.1145/3178876.3186093

ACM Reference Format:
Rahat Masood, Dinusha Vatsalan, Muhammad Ikram, and Mohamed Ali Kaafar. 2018. Incognito: A Method for Obfuscating Web Data. In *WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France,* Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3178876.3186093

1 INTRODUCTION

The wide-spread use of the Web to search or share information online introduces various privacy and confidentiality threats. One such most persistent threat is users' identification and tracking via their Web behavioral data [6, 16, 33]. Users unintentionally leave digital traces of their personal information, interests, and intents while using the online services, such as social networks, discussion forums, product reviews sites, and search engines, which could reveal sensitive information about them. The threat becomes more subtle when users are identified from anonymized datasets through inference analysis by an eavesdropper or a researcher who has access to the data. Few examples in the literature involving such threats are the re-identification of individuals in the anonymized AOL search histories of 650,000 users [16], Netflix training data of 500,000 subscribers [28], and Massachusetts hospital discharge data [34].

While there have been several works done on improving the privacy of users' Web data through obfuscation methods, these existing methods primarily lack in considering all key aspects/features of Web data privacy and they are not applicable to any Web data (e.g., search queries, posts, comments, reviews). Furthermore, these obfuscation methods are not resilient against adversarial attacks, where given the adversary's knowledge about the obfuscation mechanism and the users' Web behavior, they break the guarantees of protecting the privacy of users' Web data.

To this end, we provide answers to two key questions: *(1) What are the key features of Web data privacy; and how to quantify privacy risk by considering these features? (2) How to develop a resilient obfuscation mechanism to improve the privacy of Web data predicted with high risk, given the adversary has access to anonymized Web data and knowledge of obfuscation algorithm?* We propose an adversarial-resistant, quantitative method that predicts privacy risks of users' Web data and then obfuscates high risk data with the guarantee of protection against inference attacks by adversaries. The proposed

obfuscation method can be applicable to any type of Web applications, such as social networks, search engines, blogs, product review sites, and online forums.

*Definition 1.1 (**Privacy Risk in Web Data**).* We define privacy risk in (anonymized) Web data as a risk of identifying users and thereby learning their sensitive/private information through *(1)* **uniqueness** (distinguishability) of the sequences of a user's Web actions from other users' Web actions, *(2)* **uniformity** (non-diversity) of the user in his Web data, and *(3)* **linkability** of the user using his personal identifiable information (PII)[1] available in data.

A user's privacy is at a high risk when his Web data is distinguishable from other users, has non-diversity in own data or actions, and is linkable to an individual with high confidence based on the user's PII. For example, if a user searches or comments regarding a certain disease, drug, pregnancy, or terrorism, the Web history of the user could compromise privacy if the user's data is distinguishable, uniform for the user, and linkable to an individual based on PII available in previous search history.

We propose a privacy-aware obfuscation method for Web data that first quantifies privacy risk of users' data and then obfuscates high risk data entries with semantically similar lower risk data entries. The main contributions of our paper are as follows:

- We quantify users' privacy risk in Web data using probabilistic methods, the Hidden Markov Model (HMM) that calculates probabilities of uniqueness, uniformity, and linkability learned from training data. The model is generic (applicable) to any Web data, such as posts, shares, tweets, search queries, reviews, and clicks. Further, the model is dynamic in that the learned probabilities are updated with new data. To the best of our knowledge, no work has been done that allows such generic, comprehensive, and dynamic risk prediction in Web data.
- We propose a novel obfuscation method to obfuscate high risk (predicted) data using semantically similar low risk data retrieved from the trained HMM at the cost of some loss in utility. Using differentially-private noise addition, our proposed method is resilient against adversary who has knowledge about the method, HMM probabilities and the training dataset and therefore is able to estimate the privacy risk values and could differentiate between the original and the obfuscated data by getting all possible paths in the HMM that have higher risks.
- We conduct an extensive empirical study using two real Web datasets, the AOL dataset and our new app reviews dataset[2]. Our results indicate that privacy risk increases with sharing more data on the Web. For the AOL dataset, we found that an average privacy risk reaches 100% when a user enters 10 queries. For app reviews dataset, we found that average privacy risk associated with just 1 sensitive review is 80.5%, which increases to 87.5% with 7 reviews. We found that some obfuscated entries offer 0% privacy risk at the low cost of utility, however, there are some cases where obfuscated entries totally change the meaning of original entries. The addition of differentially private noise in the HMM model does not show significant difference in risk prediction, however, we see significant increase in utility loss for app reviews dataset i.e., 50% of the obfuscated entries has the utility loss of 64.3%, which increases to 90% for the perturbed entries by noise.

[1]Users often share or search for PII on the Web including names, contact details, address/location details of people, and ego-surfing [4]).

[2]We contribute a new large Web dataset in the domain of online app reviews by implementing a Google Play crawler that collects apps identifiers and apps meta-data.

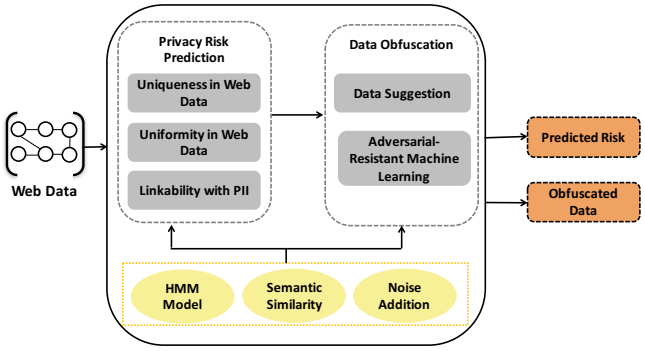


Figure 1: Overview of our privacy-aware obfuscation method for Web data

The rest of the paper is organized as follows. In Section 2, we present the methodology that we propose for obfuscating Web data. In Section 3, we first present our datasets (Section 3.1), then experimental results (Section 3.2), and finally discussion summary (Section 3.3). We provide the literature review on existing obfuscation methods and privacy risk quantification in Section 4. In Section 5, we conclude our work and discuss venues for future work.

2 THE METHODOLOGY FOR OBFUSCATING WEB DATA

In this Section, we describe how users' privacy risk in Web data can be predicted and measured using probabilistic methods, and obfuscated if the predicted risk is high. We begin with an overview, followed by the risk quantification, and then obfuscation method.

2.1 Overview

Our aim is to develop a method to predict privacy risk of Web data that comprehensively includes all key aspects of privacy and then obfuscate the high risk Web data using probabilistic methods. An overview of our proposed method is shown in Figure 1. The threat model we consider is the inference attack by a researcher or an eavesdropper who has access to anonymized (i.e., user identifiers are removed or encoded) Web data as well as knowledge about our probabilistic model. The proposed method is generic and can be applicable to various applications of Web data, such as online social networks, product reviews, forums, and professional networks.

The privacy risk (see Definition 1.1) of a user in the Web data is determined by three key aspects: *(1)* uniqueness of the data, *(2)* uniformity of the user's data, and the *(3)* linkability of data to the user based on personal identifiable information (PII) available in the Web data. The probability of uniqueness or distinguishability

of a certain data or a sequence of data is measured as the non-likelihood of it by a user being similar to Web data of other users such that it is unique or distinguished to reveal the user's identity. For example, if a user data contains 'Smith' it is less likely to be identifiable as it is a very common name in Australia, while data containing 'Dijith' (which is a less common name) is more likely to be identifiable (and therefore not anonymized). Similarly, if a user data contains a less common topic (e.g., a specific drug) it is more likely to be re-identified and the probability of distinguishability and linkability becomes even higher when the user's previous data contain personal information such as names and locations.

The probability of uniformness of a user based on the user's previous data (i.e., history) is measured as the likelihood the user has entered the data (and thereby interested in the data). The more the user has entered a certain data in previous history, the more confidence of the inference that the user is interested in this data. The joint probability of uniqueness and uniformity measures the probability of identifiability of the user in his Web actions (i.e., inverse of privacy gain for the user). The probability of linkability of a user's data to an individual is based on how much PII available in the user's data. PII could reveal personal identity of a user and therefore allows linking the corresponding data to the user. The overall privacy risk is measured as the joint probability of identifiability (uniqueness and uniformity) and linkability probabilities.

The probability of inference from a sequence of Web data is often conditional probability on previous data and therefore the risk of inference becomes higher along with the user's sequence of Web data (i.e., the probability of privacy preservation becomes lower with the sequence of user's data). The reason behind this intuition is that a user learns or reveals more with the sequence of Web actions and therefore the data become more refined or specified to a certain topic enabling the Web data sequence to being highly linkable (less anonymized) to an individual. Therefore, the inference probability becomes higher and the following Web data/action by the user might be at an even higher risk of disclosure.

2.2 Risk Prediction

The aim of our risk prediction module is to measure users' risk of their Web data being distinguishable, uniform and linkable as probabilities in a hidden Markov model (HMM). A user is represented by u_i and a data entered at a time t is represented by X_t. We train the HMM model using previous Web data in order to predict a user's privacy risk of his Web data entered at the current time being. HMM is a probabilistic model for representing probability distributions over sequences of observations. They are used in speech recognition systems, computational molecular biology applications, computer vision applications, and other applications of artificial intelligence and pattern recognition [17]. Assume a sequence of events (Web data entered by a user) over time t as $X_1, X_2, \cdots, X_T$. These events satisfy the (first-order) Markov property, i.e., the current event X_t is independent of all the events prior to X_{t-1}. Each of these events X_t outputs observations Y_t which also satisfy the Markov property, i.e., X_t and Y_t are independent of the events and observations at all other time indices. These Markov properties state that the joint distribution of a sequence of events and their observations can be factored as:

$$p(X_{1:T}, Y_{1:T}) = p(X_1)P(Y_1|X_1)\prod_{t=2}^{T} p(X_t|X_{t-1})p(Y_t|X_t). \quad (1)$$

A Web data entered by a user becomes a node and the probabilities of uniqueness, uniformity, and linkability are modelled in the HMM. The three probabilities modelled are:

(1) Uniqueness is modelled as transition probabilities in the HMM. Transition probabilities are conditional probabilities of a data by all users given previous data sequence by all users. This is required to calculate the indistinguishability or non-uniqueness of a user's data from other users' data. The risk of a data being distinguishable depends on the previous data. The reason is that the information gain from a data becomes higher if the previous data in the same topic are considered. Nodes in the HMM include data at a time (X_t) related to personal identifiable information topic, and/or a private/sensitive topic (such as cancer, drugs, and pregnancy). Edges contain the transition probabilities between nodes ($p(X_t|X_{t-1})$). These transition probabilities are weighted by their confidence in terms of how many transitions have occurred, which is $w_T = 1/count(X_t|X_{t-1})$. For calculating the privacy risk of a user with his Web data, the weighted transition probabilities are considered, i.e., $w_T \times p(X_t|X_{t-1})$.
(2) Uniformity is modelled as observation probabilities in the HMM. Observation probabilities are probabilities of the data found in previous Web data by different users (u_i) including the user whose risk is to be predicted (if available). Each node contains a set of observations with observation probabilities. We model these observation probabilities as different users' probabilities of the given data, X_t, found in previous data ($p(u_i|X_t)$). This is required to incorporate the non-uniformity aspect of a user as the frequency of the data entered by the user. The more a user has entered a specific data the more confidence (and therefore higher risk) in the inference that the user is interested in this data. Again these probabilities are weighted by $w_O = 1/count(u_i|X_t)$ and then inversed (as more uniform a user is higher the privacy risk is and therefore lower privacy probability), i.e., $(1 - w_O \times p(u_i|X_t))$.
(3) In addition to these two probabilities, we have prior probabilities of the user based on previous searches that include PII (names, locations, and organizations). In order for the Web data (related to sensitive/private topics other than PII topic) to be linkable to a user, the PII revealed by the user in his previous data needs to be taken into account. This prior probability of risk (of linkable using PII revealed) for a user u_i is calculated from the HMM of PII. The privacy risks of data related to PII topic are modelled in a separate HMM. For a given user u_i, the prior risk probability is calculated by getting the minimum privacy probability (maximum privacy risk) from all the paths in the PII HMM which include nodes X_t that contain an observation probability for the user, i.e., $p(u_i|X_t) > 0$. For users who do not have revealed any PII in previous search history the prior privacy probability becomes 1.0.

The overall privacy probability of a user u_i along a sequence of Web data $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_t$ is calculated as, following the Markov

probability in Equation (1):

$$p(X_1, \cdots, X_t|u_i) = min(HMM_{PII}|u_i) \times w_T \times p(X_1) \times (1 - w_O \times p(u_i|X_1)) \times \prod_{x=2}^{t} w_T \times p(X_x|X_{x-1}) \times (1 - w_O \times p(u_i|X_x)), \quad (2)$$

where $HMM_{PII}|u_i$ returns a list of privacy probabilities calculated from the PII HMM for all paths that include nodes where the user has an observation probability of > 0.0.

2.3 Obfuscation

Once a user data is identified as a privacy risk by our method based on the predicted privacy probability, the second step is to replace or modify the original high risk data with alternative data from different paths in the HMM to overcome the privacy risk with a loss in utility.

We quantify the utility loss (ul) in terms of semantic similarity between the original data X_x and the suggested data X_y.

$$ul(X_x, X_y) = 1.0 - sim(X_x, X_y), \quad (3)$$

where $sim(X_x, X_y)$ is a semantic similarity function [24] which returns the similarity value between the two data in the range 0 and 1. The larger the semantic similarity is the lower the utility loss is by using the alternative data.

The obfuscation module generates a list of alternative data suggestions (learned from the HMM model) along with their predicted privacy risk and calculated utility loss, from which one alternative data is chosen by the system to overcome privacy risk. It is important to note that the utility loss for the original data is 0.0 ($1.0 - sim(X_y, X_y) = 1.0 - 1.0 = 0.0$).

2.4 Adversarial Machine Learning

Given the training datasets and the HMM model learned probabilities can be accessed by an adversary, similar to all other existing obfuscation techniques our privacy-aware obfuscation technique can be susceptible to privacy attacks to learn the original data. The adversary is able to calculate or estimate the privacy risk values using the learned HMM probabilities and this could lead to privacy violation. For example, if a user's privacy risk increases with the data entered by the user and suddenly if the risk gets lower then the adversary might be able to guess that this could be a perturbed data by the system. In such a case, the adversary would be able to guess the actual data by getting all possible paths in the HMM that have higher risks.

In order to overcome this attack, we propose an adversarial machine learning technique by combining differential privacy-based noise addition with our HMM model. Noise is added in terms of counts/probabilities in the HMM model in order to perturb the original probability distribution. The magnitude of the noise depends on a privacy parameter ϵ and sensitivity S of query functions on the HMM model by an adversary.

*Definition 2.1 (**L1-sensitivity**).* Given two count dictionaries T_1 and T_2, such that $|T_1| = |T_2|$ and T_1 and T_2 differ in only one element/entry's count, the L1-sensitivity of q query functions on both dictionaries, is measured as:

$$S = max_{\forall T_1, T_2} \sum_{i=1}^{q} |Q_i(T_1) - Q_i(T_2)|, \quad (4)$$

where $Q(\cdot)$ is a query function on a dictionary and $|\cdot|$ denotes the cardinality of a dictionary.

Theorem 2.2 (Noise addition with differential privacy). *Let Q be a set of query functions and S be the L1-sensitivity of Q. Then, ϵ-differential privacy can be achieved by adding random noise r, i.e., $Q_i^T \leftarrow Q_I^T + r$, where r is a random, i.i.d. variable drawn from a Laplace distribution with magnitude $b \geq S/\epsilon$.*

A differentially private dictionary release (publishing) corresponds to issuing count queries by an adversary:

$$select\ count(*)\ from\ dictionary\ where\ count/probability \geq x \quad (5)$$

Given a set of query functions Q, differential privacy adds noise drawn from Laplace distribution with magnitude b to the true response value. As shown in Theorem 2.2, b is determined by two parameters: (1) a privacy parameter ϵ and (2) the sensitivity S of Q. In this context, it is known that a single update in the count/probability value of an element in a dictionary can change the result of at most two count queries by a magnitude of at most one. Therefore, we add Laplace noise to each element in the dictionaries with $b = 2/\epsilon$. Positive noise is incorporated by incrementing the count/probability values, while negative noise requires subtracting the count probability values.

3 EVALUATION

In this Section, we present and discuss our findings on adversarial machine learning based differentially private Web data obfuscation method. First, we present the datasets in use and then we discuss results of our experiments.

3.1 Datasets

To measure the privacy risks associated with online Web data and to evaluate the effectiveness of our obfuscation method, we use two datasets: *(1)* AOL users' search queries; and *(2)* reviews of Android applications on Google Play[3]. We summarize our datasets in Table 1.

AOL Search Queries: In 2006, AOL released an anonymized version of 20 million user search queries of more than 650,000 users over 3 months period. Usernames were replaced by anonymous identifiers with the aim to protect user privacy. However, it failed to prevent de-anonymization for some users who performed ego-surfing, or searched for personal details such as social security number, phone number, or location directions. Therefore, we use this dataset to quantify sensitivity of Web data and to evaluate the effectiveness of our obfuscation method. Each line in the in AOL search query data contains five fields: anonymous user ID, query string, query time, the rank of the item selected, and the domain of the selected item's URL path. Due to time limitation, we did not apply our method on the whole dataset, rather we set a criteria that selects only those users who have queries greater than 100. The statistics of our sampled dataset is given in Table 1.

[3] https://play.google.com

	AOL Search Queries	Android Apps Reviews
# of Entries (E)	36,389,567	16,335,480
# of Users (U)	657,429	11,196,960
# of Apps (A)	–	1,0186,560
Condition	E ≥ 100	5M Reviews where E ≥ 15 & E ≤ 20
Sampled dataset		
# of Entries (E)	23,927,203	13128
# of Users (U)	90,818	773
# of Apps (A)	–	6866

Table 1: Datasets in use.

Moreover, to highlight the consequences of searching privacy sensitive topics that could potentially reveal user information, we focus on three topics: Cancer, Pregnancy, and Alcohol. In order to extract queries in these topics, we need to identify some must words for each topic. For this purpose, we used Free Keyword Tool offered by Wordstream[4] that utilizes the latest Google keyword API. We then performed topic modeling on these keywords to get most accurate and relevant must words. We used `NLTK` [2] and `gensim` [1] to perform topic modeling and to extract relevant queries.

Android Apps Reviews: In order to collect users' reviews on Android apps from Google Play Store, we leveraged the crawlers developed in [20] and used the top 100 apps as a seed. Our crawler collects apps identifiers[5] and apps meta-data by following a breadth-first-search approach for the apps which are *"similar"* in description or published by the same developer at Google Play. In summary, we crawled 1,018,656 apps in a 4-week period of December 2016 and collected 16,335,480 reviews from 11,196,960 unique users. A given user review consists of anonymous ID of a user, review text, review time and date, app ID, and app category.

We selected four categories of apps i.e., Social, Lifestyle, Health, and Games and extracted 5 Million reviews from our crawled dataset and then applied a criterion to select only those users that provide reviews in a range of 15 to 20 on different apps. We found that most of the reviews have been given for games followed by Lifestyle and Health apps.

3.2 Experiments and Results

We analyze privacy risk prediction results from the three aspects of uniqueness, uniformity, and linkability, and also present overall risk prediction results combining all three. We then discuss our results on differentially private Web data obfuscation method using some validation cases. Finally, we present the efficiency results.

3.2.1 Experimental Setting. Before applying our method, we first pre-processed the data by filtering the broken, invalid, or empty sentences, and then re-ordered them based on time sequence. We then split the data into 20-80 testing approach where 20% of the data were used for testing, while 80% were used for training the HMM. Furthermore, to reduce training time, we applied k-means clustering that partitions the training data into k clusters, and then used multi-processing technique to run each training cluster simultaneously [15]. k-means algorithm helps grouping similar Web data i.e., queries and reviews, based on the nearest mean (centroid). For our datasets, we selected 20 clusters using the elbow method [36]. Results from each multi-processed cluster are then combined to create one training model. For AOL dataset, we used semantic similarity algorithm for short sentences proposed by [25] to find similar queries, while term frequency-inverse document frequency (`TF-IDF`) was used to evaluate similarities of app reviews [15]. We used ϵ-differential privacy based noise addition for adversarial machine learning where the privacy budget parameter is set to $\epsilon = 0.3$.

3.2.2 Privacy Risk Prediction. Our results indicate that privacy risk increases with sharing more data on the Web. For the AOL dataset, we found that an average privacy risk reaches to 100% (1.0 privacy risk) when a user enters 10 queries. An exemplary user is shown in Table 2 (user ID 3058504), where the risk becomes 100% after entering 10 queries. Moreover, the average risk of predicting a user with just 1 sensitive query ranges between 78% and 83% (0.78 – 0.83). This is because our framework calculates risks based on three aspects, i.e., uniformity, uniqueness, and linkability. Even if a user does not have uniform data, he might be identified through the unique pattern of entering data, and vice versa. For instance, we can predict after 10 queries of the user shown in Table 2 with user ID '3058504' that either he or his family member is suffering from thyroid cancer. Similarly, we observe that another user (with user ID '3612363' as shown in Table 2) wants to know about Dr. Paul Mansfield, who worked at MD Anderson Cancer Center. Further queries would reveal that he is interested in prostate cancer at MD Anderson and its treatment. We also observe similar cases for pregnancy and alcohol topics, and found that users could be identified through their unique Web patterns. For instance, we discover that the user with ID '7894176' (shown in Table 2) is pregnant but has antiphospholipid and smoking problems. Likewise, the user with ID '4320454' wants to defy drug test by finding some ways.

For app reviews dataset, we found that average privacy risk associated with just 1 sensitive review is 80.5% (0.805), which increases to 87.5% (0.875) with 7 reviews. In Table 2, we observe that the user with ID '1559229' has some kind of association with Fibromyalgia disease and is using an app to improve his health issues. Similarly, we analyze that the user with ID '5995260' has the same writing pattern for all reviews and thus his privacy risk reaches to 99% (0.99) with only six reviews.

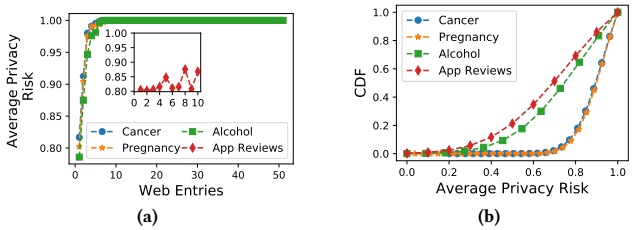
Considering our overall risk prediction results, we found that any data entry which contains words such as country name, person name, disease name, personal pronouns or uniformity has privacy risk of 75% (0.75) or above and is highly risky/sensitive. Therefore, we set our privacy risk threshold to 0.75, i.e., any entry which has a privacy risk above 75% is considered as highly risky which requires to be obfuscated with (semantically similar) entry.

Figure 2 shows the results of privacy risk prediction. It is clear in Figure 2a that our method is capable of re-identifying users even if the users' unique identities are not known. Our results indicate that an average risk reaches to 100% (1.0) if users have 10 or more data entries. The minimum average risk is 78% (0.78) for alcohol with 1 query. For app reviews, we achieve maximum of 87.5% (0.875) average risk with 7 reviews, and a minimum of 80.5% (0.805) with just 1 review. Figure 2b shows the CDF of users with their predicted privacy risks. For cancer and pregnancy, we found that more than 50% of users have risk higher than 0.85, while

[4] https://www.wordstream.com
[5] Each Android app has a unique identifier, termed as app ID in short.

Table 2: Few privacy risk evaluation cases

User Anonymized ID	Web Entries	Topic
3058504	'do you need surgery for underactive thyroid', 'why is physical therapy important after back surgery', 'why do you need physical therapy after back surgery', 'had back surgery but when i went to physical therapy my body hurt after', 'is it normal for my body to hurt after first visit to physical therapy', 'not being use to exercising can make physical therapy hurt', 'my husband did physical therapy one time and didnt go back due to pain', 'i dont like physical therapy because my body hurts after', 'physical therapy can be painful', 'is it normal for my body to hurt after first visit to physical therapy'	Cancer (10 Queries)
3612363	'md anderson cancer center and dr. paul mansfield'	Cancer (1 Query)
7894176	'getting pregnanct after being on birth control', 'getting pregnant with antiphospholipid disorder', 'having a healthy pregnancy with antiphosophlipid disorder', 'healthy pregnancy with antiphosophlipid disorder', 'chances of having a baby with antiphospholipid syndrome', 'costs of heparin during pregnancy', 'pregnancy and positive ana 1 640', ''pregnancy and positive ana 1 640', 'does lupus effect fertility', 'if i quit smoking in the middle of pregnancy will i miscarry', 'how much does smoking have an effect on fertility'	Pregnancy (10 Queries)
6143033	'pregnant no insurance denied by medicaid in florida'	Pregnancy (1 Query)
4320454	'cocaine drug testing', 'harms from herion addiction', 'opiate drug called suboxcine', 'national institute on drug abuse', 'how to clean out your urine for acocaine drug test .', 'how can we beat a cocaine urine drug test for employment', 'the longest time cocaine stays in our system for a drug test', 'how many days or hours for cocaine to leave the system to be clean for drug urine test for employment'	Alcohol (8 Queries)
3305139	'new jersey drug treatment rehab flynn house'	Alcohol (1 Query)
5995260	'Awesome app I am loving this app. Good work by the developers', 'Car wash for kids I am loving this app. Good work by the developers.', 'Awesome app I am loving this app. Good work by the developers.', 'Awesome app I am loving this app. Good work by the developers.', 'World hello Awesome game. I'm loving it. Good work by the developers', 'Car Racing Awesome game. Loved it'	Games (6 Reviews)
1559229	'Very useful tool This app is great for anyone going through health issues. Very easy to use, has many options for location of pain, what you were doing, and you can add different options. It's a great app if you have Fibromyalgia.'	Health (1 Review)



(a) (b)

Figure 2: (2a) Average privacy risk with the increasing number of Web entries and (2b) average privacy risk per user.

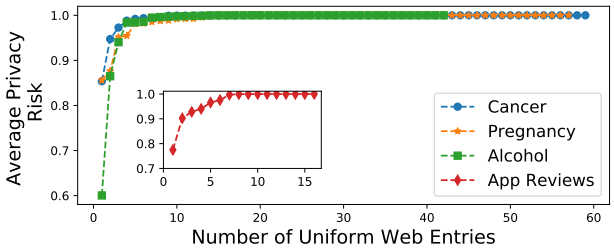


Figure 3: Risk prediction results of uniform Web entries.

alcohol has a prediction rate of 0.7 for more than 50% of users. We found similar results for reviews dataset, where more than 50% of users have privacy risk of 0.7 involved in their reviews.

Uniformity: We now discuss our results on the uniformity of users' Web entries. As mentioned earlier, uniformity refers to the number of observations of data entry by a user on the Web. Our results compliment previous discussion, where users are exposed to higher risk with uniform data. We found that users who entered same entries two times have at least 85% (0.85) of privacy risk with all datasets. For instance, we observe that a user enters the query 'do i have liver disease if a small amount of billirubin is in urine' four times and thus gets the risk of 100% (1.0) being identifiable. Similarly, we found that a user enters the review 'NICE 1' 5 times, and has a privacy risk of 99.8% (0.998). Figure 3 shows the average risk for uniform queries. Overall, our results indicate that users involved in alcohol and pregnancy topics are 100% identifiable after entering 12 uniform queries, whereas users involved in cancer topics are 100% identifiable with 10 uniform queries. Similarly, we analyzed that users who entered 10 similar reviews are 100% identifiable.

Uniqueness: Uniqueness refers to the distinctive sequence of a user's data entries on the Web. Figure 4 shows the results. Our analysis shows that out of 700 unique data sequences of pregnancy, 680 sequences are 100% (1.0 risk) identifiable, and has the minimum average privacy risk of 98% (0.98). Likewise, cancer queries have 430 unique sequences out of which 410 are 100% identifiable and have the minimum average risk of 98.5% (0.985). For instance, in pregnancy topic, we found that a user is 98.5% identifiable after entering 7 unique queries in a sequence as shown below:

'how to increase fertility naturally', 'increasing fertility naturally', 'increasing the number of eggs released during ovulation naturally', 'increasing the number of eggs released during ovulation naturally', 'conceiving twins without fertility drugs', 'getting a baby girl', 'choosing babys sex with ovulation'

For alcohol queries, we realize that 40 out of 180 data sequences have 1.0 risk, and these queries have the minimum average risk of 0.71. App reviews dataset has the lowest number of unique sequences, i.e., 20. The minimum average risk is 0.79 and it shows 1.0 privacy risk for 2 unique sequences only.

Linkability: We now investigate the linkability of users' Web entries using their PII. We found few users who have PII information available in their Web entries. For instance, a user in pregnancy topic entered a query 'place son long island to have a baby shower', and another user in alcohol topic entered PII query 'drug cases that

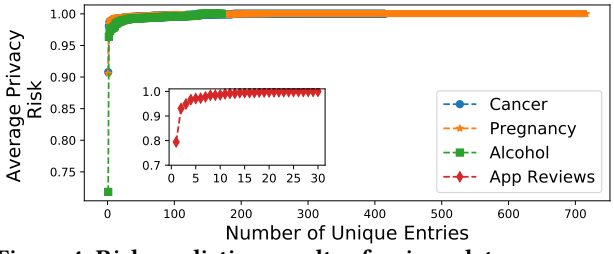


Figure 4: Risk prediction results of unique data sequences

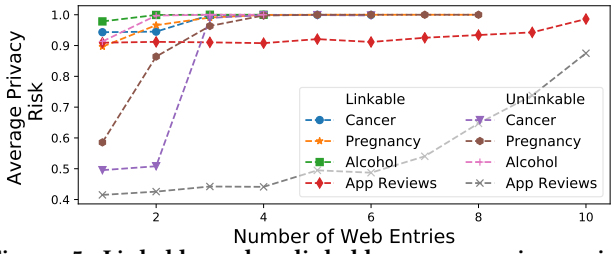


Figure 5: Linkable and unlinkable average privacy risk against Web entries having PII.

been through the US appellate court'. Similarly, for app reviews dataset, we found that a number of users have entered either email IDs or names in their reviews.

Figure 5 shows the average privacy risk for the queries having PII available. We also present results without linkability information i.e., we remove PII and evaluate the privacy risk for the same set of entries. Our results indicate linking data with PII has more privacy risk as compared to data with no PII. For instance, cancer has the minimum average risk of 95% (0.95) for linkability, which reduces to 50% (0.5) if we remove PII. Similarly, pregnancy has 89.5% (0.895) minimum privacy risk with PII and 59% (0.59) without PII. We observe less difference in alcohol queries, i.e., a minimum of 98.5% (0.985) risk for linkability and 90.5% (0.905) for unlinkability. For app reviews, the linkable reviews have 90.6% (0.906) of minimum average risk, but reduces to 40.5% (0.405) without PII. However, we found that entries with or without PII can reach to 100% identifiability (uniqueness and uniformity) except for app reviews, for which the maximum risk involved with and without PII are 99% (0.99) and 98.5% (0.985), respectively.

3.2.3 Obfuscation. In this Section, we discuss our results on the obfuscation of high risk Web entries. We first present results for adversarial resistant obfuscation method and then move to few validation cases where original Web entries are altered to low risk entries. As mentioned in Section 2, we obfuscated the data entries having higher privacy risk with lower risk entries that are semantically similar to original entries.

We compare original and obfuscated Web entries using two metrics i.e., privacy risk and utility loss. We found that some obfuscated entries offer 0% privacy risk at the low cost of utility, however, there are some exceptions where obfuscated entries totally change the meaning of original entries. Moreover, our results indicate that the addition of differentially private noise in HMM model does not show significant difference in the risk calculations of Web entries. However, we see a significant increase in utility loss for app reviews dataset, i.e., 50% of the altered entries has the utility loss of 64.3% (0.643), which increases to 90% (0.9) for the perturbed entries by noise. For AOL dataset, the utility loss remains between 58% to 63% (0.58 – 0.63) for 50% of both the perturbed entries with and without noise. The utility loss comparison between obfuscated data with and without noise for each topic is shown in Figure 6. Thus, these results indicate that the obfuscated entries come with the cost of loosing the original meaning of the data. We however are able to attain lower privacy risk, where the risk of all alternative entries suggested by our method are below 75% (0.75) and do not contain any name, location, specific writing pattern, uniformity in entries etc. On average, the privacy risk is reduced to almost 30% to 40% (0.3 – 0.4), however at the cost of utility.

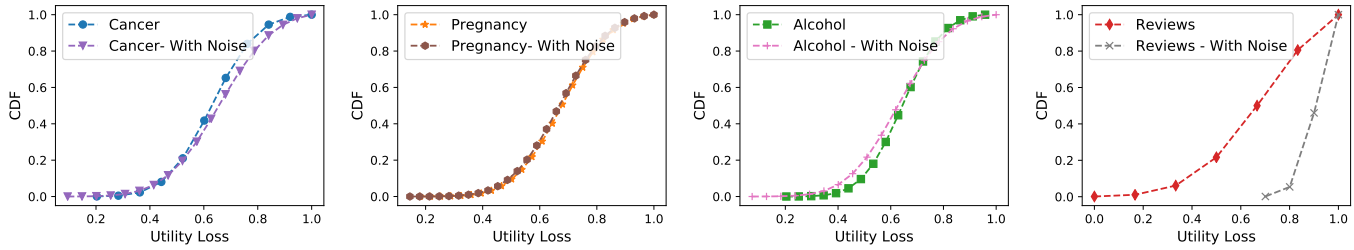
Table 3 shows validation cases where some Web entries are obfuscated to preserve privacy by our method at the cost of utility. We compare privacy risks of original and obfuscated Web entries along with the utility loss before and after the addition of differentially-private noise. We take three cases (best, average, worst) from AOL topics and two cases (average, worst) from app reviews dataset. These cases indicate that the addition of noise not only improves privacy but also helps in securing the obfuscation method against adversary attacks. Similarly, in Figure 7 we show that the addition of noise dodges adversary by changing the original risk to perturbed risk. Even if the adversary has access to datasets, HMM probabilities, and knowledge of framework, our addition of differential-private noise does not allow the adversary to guess the original risk as well as the difference between original and obfuscated Web data. Consider an example in Figure 7a where original risk reaches 83% (0.83) and suddenly falls down to 0% (0.0) risk by replacing original entry with obfuscated low risk entry. In this case, adversary is able to differentiate between original and obfuscated entries since sudden fall is an indication of obfuscated data. The inclusion of differential noise perturbs the risk such that it becomes difficult for an adversary to guess if it is an original or obfuscated entry. When a risk is above a certain threshold, the adversary model certainly replaces the original entry with low risk entries, however the addition of noise confuses the adversary to get to the original entry.

3.2.4 Efficiency. Finally, we investigate the time efficiency of our method. We found that time increases with the increasing number of data entries. The average time to predict, add noise, and then alternate high risk Web entries is 0.0302, 0.0454, 0.0304, 0.0118 seconds per query of cancer, pregnancy, and alcohol, and app reviews, respectively. We found that the time to evaluate and obfuscate a query gets stable after entering certain number of queries. This is because either queries are repeating or we are training/updating our model continuously. However, when a new query comes in, which is not already seen by the training model, then it might take more time. Figure 8a shows the average time for each topic against the number of data entries. The maximum average time is 224 seconds for 47 cancer queries.

Figure 8b shows the distribution of users against average time. We found that 85% of AOL users are processed within 50 seconds, while 62.5% of app reviews users are processed in 0.002 seconds. The significant time difference between the two datasets is because of two different techniques used for semantic matching. The `TF-IDF` approach [15] is pretty faster than the semantic similarity function [25] used for AOL because of various functions involved (word order, sentence order, `NLTK` semantic dictionary etc.).

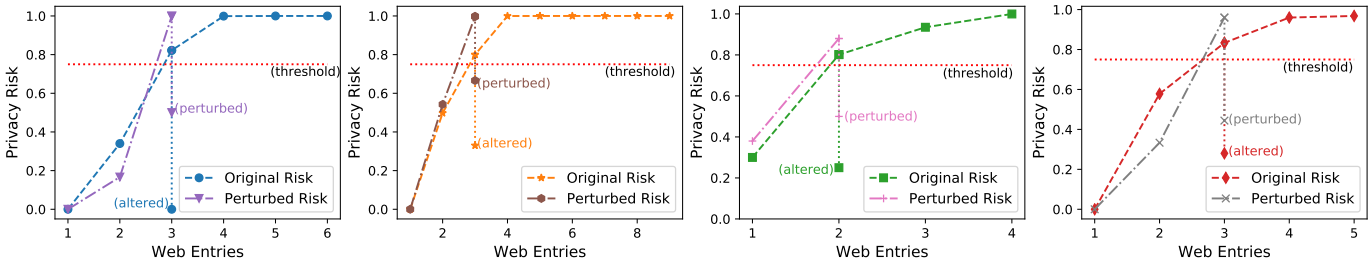
3.3 Discussion

The obfuscation method presented in this paper highlights three key aspects: *(1)* comprehensive privacy risk evaluation of Web data, *(2)* semantic similarity for obfuscated data, and *(3)* resilient against adversarial machine learning. We conducted experiments of our framework on two datasets, AOL search query and Android app reviews. We first measure the privacy risk associated with queries and apps reviews, and then obfuscate high risk entries. The results show



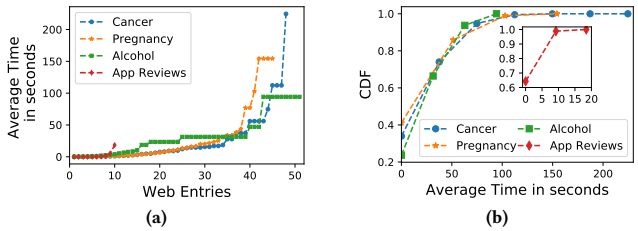
(a) Cancer (b) Pregnancy (c) Alcohol (d) Reviews

Figure 6: Comparison of utility loss between obfuscated data with and without noise for adversarial machine learning



(a) Cancer (b) Pregnancy (c) Alcohol (d) Reviews

Figure 7: Improving privacy and resistance against adversarial attack



(a) (b)

Figure 8: (8a) Average time in seconds in the increasing order of Web entries; and (8b) CDF average time per user.

that our privacy prediction technique is reliable enough to identify high risk Web entries via three aspects of uniqueness, uniformity, and linkability. In addition, our obfuscation method guarantees privacy against adversarial attacks with high effectiveness. Our results reveal some important findings which we enlist below.

(1) Privacy risk increases with sharing more data on the Web even if the users' unique identities are not known. Users who share their personal interest in a specific field are likely to be more vulnerable to privacy attacks. For instance, users who searched for information related to specific medical center in a specific area are more easily identifiable in terms of their location and disease. For app review dataset, we found many users have same writing pattern in their reviews, thus making them identifiable against other users.

(2) Privacy risk increases with sharing same data on the Web. Users who entered same queries or reviews multiple times are easily recognizable. The identification reaches to 100% with 10 uniform entries. Similarly, privacy risk increases with

Table 3: Validation cases - Comparison of original and perturbed Web entries

Entry	Privacy Risk		Altered Entries		Privacy Risk		Utility Loss	
	Without Noise	With Noise	Without Noise	With Noise	Without Noise	With Noise	Without Noise	With Noise
stomach cancer signs	97.8%	98.30%	testiclular cancer	testiclular cancer	0%	50%	0.406	0.406
best clinic for prostate cancer	98.9%	99.11%	best clinic for prostate cancer research	best prostate cancer treatment centers	50%	66%	0.094	0.289
Inoperable bladder cancer	98.41%	98.81%	bladder sonogram	failure to diagnose bladder cancer	0%	50%	0.65	0.482
i need help on getting pregnant	95.11%	96.25%	chances of getting pregnant	tips on getting pregnant	0%	75%	0.52	0.296
First response early detection pregnancy tests	99.30%	99.34%	Early sign pregnancy	Early stages of pregnancy	50%	50%	0.364	0.411
can you take tyelnoe while pregnant	98.1%	98.33%	Can you take medicine while pregnancy	Can pregnancy women take tyelon 3	0%	50%	0.364	0.45
drug addiction help and new jersey	92.3%	94.66%	alcohol and tobacco law new jersey	drug abuse counseling	0%	75%	0.547	0.588
How to deal with an alcoholic	83.33%	97.57%	Christianity and the alcoholic	Programs for drug abuse	50%	66%	0.45	0.63
low cost drug addiction program in ny state	90.9%	97.43%	drug rehabilitation through programs	drug rehab programs	50%	75%	0.71	0.69
Nice excellent work...good work...excellent graphics	93.75%	93.89%	Nice good excellent	Pretty good	66%	66%	0.49	0.68
My 3 year old loves i!	88.64%	90.9%	My daughter loves it	Fun game for kids	50%	25%	0.26	0.76

the distinct sequence of Web actions. This means that users who performed Web actions or shared data in a different way than others are likely to be identifiable among others. Moreover, we found that users who share PII on the Web are 100% identifiable in most cases.

(3) It is possible for an adversary to differentiate between original and obfuscated Web entry given the dataset and obfuscation knowledge. The use of differential privacy in the method enables resistance to such attacks, however, at the cost of utility loss.

Limitations: We have only used the basic HMM model to measure privacy probabilities and corresponding privacy risk. We have not investigated different probabilistic models such as Gaussian distribution, Dirichlet distribution, and maximum entropy Markov model (MEMM) for comparison. Our method can be extended by replacing the HMM model with other probabilistic methods.

The AOL dataset (as used in most of the other related work) is outdated. Fresh Web datasets from search engines such as Google, and Yahoo as well as from social platforms such as Facebook may lead to high privacy risk rates. We also did not test our framework in an online environment, and thus another important aspect for future investigation is to develop a real-time privacy risk prediction and obfuscation system where Web entries are evaluated and obfuscated at run-time with or without user involvement. Perhaps a browser plug-in could be developed for our proposed method.

We used fixed privacy budget parameter for our differentially private obfuscation method. Similarly, we fixed our privacy risk threshold to 0.75. We need to further investigate different parameter settings. Moreover, the semantic similarity function is not efficient to calculate risk in milliseconds, which requires other efficient and effective similarity measure approaches to be studied for real-time applications.

4 RELATED WORK

Several works have been conducted on obfuscation techniques for Web search queries. TrackMeNot (TMN) [18] is proposed as a Firefox plugin to randomly issue dummy queries from predefined Rich Site Summary (RSS) feeds. GooPIR is a standalone application for noise addition to Google queries [12], which modifies the user queries by adding dummy keywords, and then the search results are re-ranked locally based on the original user queries. PRivacy model for the Web (PRAW) [32] is another technique, which continuously generates fake queries in different topics of interest of the user.

Few studies have been conducted on obfuscation methods for other Web data, such as social networks. Weinsberg et al. [37] studied the impact of obfuscation on the utility of recommendation systems with different classifiers. Salman et al. [31] and Li et al. [23] proposed methodologies to prevent inference attacks against published data by distorting data before making it publicly available while providing utility guarantee. A study by Chen et al. [7] investigated the effectiveness of different obfuscation strategies and policies for online social networks and proposed a novel obfuscation strategy based on the χ^2 feature selection metric without requiring knowledge about the classifier used by an adversary.

On the other hand, only limited works have considered quantifying privacy in Web data. Peddinti et al. [30] evaluated the privacy guarantees offered by the TMN based on machine learning classifiers. Gervais et al. [14] also evaluated the query obfuscation techniques such as TMN and fake query generation, by learning the linkability between users' original and fake queries via machine learning algorithms. Balsa et al. [3] performed qualitative analysis on six existing obfuscation techniques by investigating their privacy characteristics. The study provides insights into the deficiencies of existing solutions however, it did not analyze and compare the techniques quantitatively. Another study by Chow et al. [8] proposed two features that could be used to differentiate TMN dummy queries from real user queries.

A recent work by Biega et al. [4] studied quantifying privacy risk in Web data by manually developing rules for sensitive key-value pairs and performing probabilistic calculation of the rules based on user's search history. Rule-based approaches are time-consuming as well as non-reliable for real-time risk prediction. A ranking-based Information Retrieval-centric approach to privacy risk evaluation in online communities is proposed by Biega et al. [5]. This approach uses ranking as a means of modelling a rational adversary who targets the most afflicted users. In [26], a framework for computing privacy scores of users in online social networks was proposed based on the sensitivity and visibility of a set of profile items.

The threat of tracking users dates back to Sweeney, who showed for the first time that coarse-grained information such as birthday, gender, and ZIP code can uniquely identify a person [35]. This work was followed by several studies that provided measurement insights into Web tracking and device fingerprinting [9–11, 13, 21, 22, 27, 29, 38, 39].

However, none of these works allows risk prediction of Web data when the user actively participates in online Web activities. In addition, no work has addressed obfuscation based on risk prediction for online users who are about to exploit to inference attack. Adversarial machine learning has been an active area of research in the recent literature [19]. However, no work has so far considered adversarial machine learning for Web data obfuscation techniques. Our work is the first to address in this direction of privacy-aware obfuscation method for any Web data using a comprehensive risk evaluation method.

5 CONCLUSIONS

Web data privacy has received much attention in the recent times due to the wide spread use of the Web and the growing concerns of privacy and confidentiality. Several works on obfuscation methods to counter privacy risks of Web data have been conducted in the literature. However, these methods are not generic and applicable to any Web data and they do not consider obfuscation for high risk predicted data using semantically similar data. In addition, adversarial machine learning for Web data obfuscation has not been studied in the literature. In this paper, we propose a privacy-aware obfuscation method that addresses the shortcomings of existing methods. We conducted experiments using two real Web datasets and our experiment results show that our method is effective in predicting privacy risk in Web data and obfuscating data that are predicted with high risk. In the future, we plan to implement our obfuscation method as a user-centric application to be deployed as a browser plug-in.

REFERENCES

[1] 2018. gensim: Topic modelling for humans. https://radimrehurek.com/gensim/. (2018). Accessed on: 12-01-2018.

[2] 2018. Natural Language Toolkit. http://www.nltk.org. (2018). Accessed on: 12-01-2018.

[3] Ero Balsa, Carmela Troncoso, and Claudia Díaz. 2012. OB-PWS: Obfuscation-Based Private Web Search. In *IEEE Symposium on Security and Privacy, SP 2012, 21-23 May 2012, San Francisco, California, USA*. 491–505.

[4] Joanna Biega, Ida Mele, and Gerhard Weikum. 2014. Probabilistic Prediction of Privacy Risks in User Search Histories. In *Proceedings of the First International Workshop on Privacy and Secuirty of Big Data, PSBD@CIKM 2014, Shanghai, China, November 7, 2014*. 29–36.

[5] Joanna Asia Biega, Krishna P. Gummadi, Ida Mele, Dragan Milchevski, Christos Tryfonopoulos, and Gerhard Weikum. 2016. R-Susceptibility: An IR-Centric Approach to Assessing Privacy Risks for Users in Online Communities. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. ACM, New York, NY, USA, 365–374.

[6] Prima Chairunnanda, Nam Pham, and Urs Hengartner. 2011. Privacy: Gone with the Typing! Identifying Web Users by Their Typing Patterns. In *PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom), Boston, MA, USA, 9-11 Oct., 2011*. 974–980.

[7] Terence Chen, Roksana Boreli, Mohamed Ali Kâafar, and Arik Friedman. 2014. On the Effectiveness of Obfuscation Techniques in Online Social Networks. In *Privacy Enhancing Technologies - 14th International Symposium, PETS 2014, Amsterdam, The Netherlands, July 16-18, 2014. Proceedings*. 42–62.

[8] Richard Chow and Philippe Golle. 2009. Faking contextual data for fun, profit, and privacy. (2009), 105–108.

[9] Anupam Das, Nikita Borisov, and Matthew Caesar. 2014. Do You Hear What I Hear?: Fingerprinting Smart Devices Through Embedded Acoustic Components. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS), Scottsdale, AZ, USA, November 3-7, 2014*. 441–452.

[10] Anupam Das, Nikita Borisov, and Matthew Caesar. 2016. Tracking Mobile Web Users Through Motion Sensors: Attacks and Defenses. In *23rd Annual Network and Distributed System Security Symposium (NDSS), San Diego, California, USA, February 21-24, 2016*. The Internet Society.

[11] Sanorita Dey, Nirupam Roy, Wenyuan Xu, Romit Roy Choudhury, and Srihari Nelakuditi. 2014. AccelPrint: Imperfections of Accelerometers Make Smartphones Trackable. In *21st Annual Network and Distributed System Security Symposium (NDSS), San Diego, California, USA, February 23-26, 2014*. The Internet Society.

[12] Josep Domingo-Ferrer, Agusti Solanas, and Jordi Castellà-Roca. 2009. h (k)-Private information retrieval from privacy-uncooperative queryable databases. *Online Information Review* 33, 4 (2009), 720–744.

[13] Peter Eckersley. 2010. How Unique Is Your Web Browser?. In *Privacy Enhancing Technologies, 10th International Symposium, PETS 2010, Berlin, Germany, July 21-23, 2010. Proceedings*. 1–18.

[14] Arthur Gervais, Reza Shokri, Adish Singla, Srdjan Capkun, and Vincent Lenders. 2014. Quantifying Web-Search Privacy. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14)*. ACM, New York, NY, USA, 966–977.

[15] Jiawei Han, Micheline Kamber, and Jian Pei. 2011. *Data Mining: Concepts and Techniques, 3rd edition*. Morgan Kaufmann.

[16] Saul Hansell. 2006. AOL Removes Search Data on Vast Group of Web Users. http://query.nytimes.com/gst/fullpage.html?res=9504e5d81e3ff93ba3575bc0a9609c8b63. *New York Times* (2006).

[17] Bunke Horst and Caelli Terry Michael. 2001. *Hidden Markov models: Applications In Computer Vision*. Vol. 45. World Scientific.

[18] Daniel C Howe and Helen Nissenbaum. 2009. TrackMeNot: Resisting surveillance in web search. *Lessons from the Identity Trail: Anonymity, Privacy, and Identity in a Networked Society* 23 (2009), 417–436.

[19] Ling Huang, Anthony D. Joseph, Blaine Nelson, Benjamin I. P. Rubinstein, and J. D. Tygar. 2011. Adversarial machine learning. In *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, AISec 2011, Chicago, IL, USA, October 21, 2011*. 43–58.

[20] Muhammad Ikram and Mohamed Ali Kâafar. 2017. A first look at mobile Ad-Blocking apps. In *16th IEEE International Symposium on Network Computing and Applications, NCA 2017, Cambridge, MA, USA, October 30 - November 1, 2017*. 343–350.

[21] Andreas Kurtz, Hugo Gascon, Tobias Becker, Konrad Rieck, and Felix C. Freiling. 2016. Fingerprinting Mobile Devices Using Personalized Configurations. *PoPETs* 2016, 1 (2016), 4–19.

[22] Pierre Laperdrix, Walter Rudametkin, and Benoit Baudry. 2016. Beauty and the Beast: Diverting Modern Web Browsers to Build Unique Browser Fingerprints. *Proceedings - IEEE Symposium on Security and Privacy, SP 2016* (2016), 878–894.

[23] Chen Li, Houtan Shirani-Mehr, and Xiaochun Yang. 2007. Protecting Individual Information Against Inference Attacks in Data Publishing. In *Proceedings of the 12th International Conference on Database Systems for Advanced Applications (DASFAA'07)*. Springer-Verlag, Berlin, Heidelberg, 422–433.

[24] Yuhua Li, David McLean, Zuhair A Bandar, James D O'shea, and Keeley Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE transactions on knowledge and data engineering* 18, 8 (2006), 1138–1150.

[25] Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockett. 2006. Sentence Similarity Based on Semantic Nets and Corpus Statistics. *IEEE Trans. on Knowl. and Data Eng.* 18, 8 (Aug. 2006), 1138–1150.

[26] Kun Liu and Evimaria Terzi. 2010. A Framework for Computing the Privacy Scores of Users in Online Social Networks. *ACM Trans. Knowl. Discov. Data* 5, 1, Article 6 (Dec. 2010), 30 pages.

[27] Rahat Masood, Benjamin Zi Hao Zhao, Hassan Jameel Asghar, and Moahmed Ali Kâafar. 2017. POSTER: TouchTrack: How Unique are your Touch Gestures?. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017*. 2555–2557.

[28] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy (SP '08)*. IEEE Computer Society, Washington, DC, USA, 111–125.

[29] Łukasz Olejnik, Claude Castelluccia, and Artur Janc. 2012. Why Johnny Can't Browse in Peace: On the Uniqueness of Web Browsing History Patterns. *5th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2012)*, 1–16.

[30] Sai Teja Peddinti and Nitesh Saxena. 2010. On the Privacy of Web Search Based on Query Obfuscation: A Case Study of TrackMeNot. In *Proceedings of the 10th International Conference on Privacy Enhancing Technologies (PETS'10)*. Springer-Verlag, Berlin, Heidelberg, 19–37.

[31] Salman Salamatian, Amy Zhang, Flávio du Pin Calmon, Sandilya Bhamidipati, Nadia Fawaz, Branislav Kveton, Pedro Oliveira, and Nina Taft. 2013. How to Hide the Elephant- or the Donkey- in the Room: Practical Privacy Against Statistical Inference for Large Data. In *IEEE Global Conference on Signal and Information Processing, GlobalSIP 2013, Austin, TX, USA, December 3-5, 2013*. 269–272.

[32] Bracha Shapira, Yuval Elovici, Adlay Meshiach, and Tsvi Kuflik. 2005. PRAW - A PRivAcy Model for the Web. *Journal of the American Society for Information Science and Technology (JASIST)* 56, 2 (2005), 159–172.

[33] Jessica Su, Ansh Shukla, Sharad Goel, and Arvind Narayanan. 2017. De-anonymizing Web Browsing Data with Social Networks. In *Proceedings of the 26th International Conference on World Wide Web, (WWW) 2017, Perth, Australia, April 3-7, 2017*. 1261–1269.

[34] Latanya Sweeney. 1997. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics* 25, 2-3 (1997), 98–110.

[35] Latanya Sweeney. 2000. Simple Demographics Often Identify People Uniquely. *Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000* (2000), 1–34. http://dataprivacylab.org/projects/identifiability/paper1.pdf

[36] Robert Tibshirani, Guenther Walther, and Trevor Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63, 2 (2001), 411–423.

[37] Udi Weinsberg, Smriti Bhagat, Stratis Ioannidis, and Nina Taft. 2012. BlurMe: Inferring and Obfuscating User Gender Based on Ratings. In *Proceedings of the Sixth ACM Conference on Recommender Systems (RecSys '12)*. ACM, New York, NY, USA, 195–202.

[38] Ting-Fang Yen, Yinglian Xie, Fang Yu, Roger Peng Yu, and Martín Abadi. 2012. Host Fingerprinting and Tracking on the Web: Privacy and Security Implications. In *19th Annual Network and Distributed System Security Symposium (NDSS), San Diego, California, USA, February 5-8, 2012*. The Internet Society.

[39] Zhe Zhou, Wenrui Diao, Xiangyu Liu, and Kehuan Zhang. 2014. Acoustic Fingerprinting Revisited: Generate Stable Device ID Stealthily with Inaudible Sound. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14)*. ACM, New York, NY, USA, 429–440.