Probabilistic Prediction of Privacy Risks in User Search Histories

Joanna Biega
Max Planck Institute for Informatics
Saarbrücken, Germany
jbiega@mpi-inf.mpg.de

Ida Mele
Max Planck Institute for Informatics
Saarbrücken, Germany
imele@mpi-inf.mpg.de

Gerhard Weikum
Max Planck Institute for Informatics
Saarbrücken, Germany
weikum@mpi-inf.mpg.de

ABSTRACT

This paper proposes a new model of user-centric, global, probabilistic privacy, geared for today's challenges of helping users to manage their privacy-sensitive information across a wide variety of social networks, online communities, QA forums, and search histories. Our approach anticipates an adversary that harnesses global background knowledge and rich statistics in order to make educated guesses, that is, probabilistic inferences at sensitive data. We aim for a tool that simulates such a powerful adversary, predicts privacy risks, and guides the user. In this paper, our framework is specialized for the case of Internet search histories. We present preliminary experiments that demonstrate how estimators of global correlations among sensitive and non-sensitive key-value items can be fed into a probabilistic graphical model in order to compute meaningful measures of privacy risk.

Categories and Subject Descriptors

K.4.1 [**Public Policy Issues**]: Privacy; G.3 [**Probability and Statistics**]: Multivariate statistics, Statistical computing; H.2.8 [**Database Applications**]: Data Mining

Keywords

Privacy Risk Prediction; Probabilistic Privacy; User-Centric Privacy; Query Logs

1. INTRODUCTION

1.1 Motivation

Privacy research has strongly focused on a *data-centric*, *local*, and *deterministic* perspective: given a dataset with publicly visible attributes and privacy-sensitive hidden attributes, ensure that an adversary cannot see any data or query results that allow him to infer any of the privacy-sensitive values. Measures to protect privacy include disallowing certain kinds of queries and/or coarsening or perturbing query results [18, 19, 21, 30]. Even the most successful of such models, differential privacy [9], is fairly limited in its scope and cannot address the modern situation with users having sensitive information spread across social networks, online communities, question-answer forums, and their search history.

Privacy risk has been considered in the field of web search, with the goal of masking real intents of search-engine users through the automatic generation of *dummy queries* [26, 27, 28]. However, these approaches are still prone to privacy attacks [32]. Moreover, none of this work considers background knowledge on the adversary side.

In the age of big data and social media, guarding a user's privacy in a single dataset is way insufficient. We advocate a paradigm shift towards a *user-centric*, *global*, and *probabilistic* approach: given a user and her data traces in the digital world, minimize or bound the chances that an adversary that observes as much as possible and has global background knowledge can successfully guess any privacy-sensitive information about the user. The notion of *guessing* is essential here: given a partial set of *observations* (i.e., attribute values or other information) about a user, can the adversary harness general *world knowledge* and *probabilistic inference* to make an *educated guess* at privacy-critical information?

For example, when a young woman anonymously posts questions in one or more online communities about "missed period," "morning sickness," etc., she is likely pregnant and presumably does not want everybody to know about it (yet). For the user, this is a privacy risk even if an adversary cannot be hundred percent certain about his conclusion. Consider tracking companies that sell user data to the advertisement industry or run analytics on behalf of large employers or health insurance companies. Even your online acquaintances may be viewed as (soft) adversaries in this situation.

There is little prior research that takes such a user-centric perspective. The recently proposed notion of stochastic privacy in [29] refers to a user-defined threshold for sharing data to be obeyed by the platform provider. This model does not consider any probabilistic adversaries nor does it provide guidance for users about their privacy risks. Another line of work focuses specifically on the linkability of user identities across communities such as Facebook and Twitter [23, 24, 33]. However, privacy risks are much wider and diverse than linkability and de-anonymization. A user would be irritated receiving ads about pregnancy tests even if her identity is not known to the originator of this target ad, especially if a friend of hers sees this on her screen. Moreover, none of this prior work considered the role of background knowledge

PSBD'14, November 7, 2014, Shanghai, China.

http://dx.doi.org/10.1145/2663715.2669609.

by the adversary, and this is a fundamental aspect that is poorly understood and largely under explored in the entire research on privacy.

1.2 Goals and Contribution

This paper proposes a departure from the established mainstream privacy research to a new avenue of user-centric, global, probabilistic privacy with explicit consideration of background knowledge. As perfect privacy is infeasible in an open setting with users posting information in a variety of online communities, we settle for the softer goal of raising awareness on the user side and helping users in assessing, understanding, and managing their potential privacy risks. We envision a privacy-advisor tool that alerts users when critical situations arise, explains the risk, and guides towards appropriate counter-measures, such as changing the privacy settings in an online community or using an anonymization tool to make certain posts or queries unlinkable to user's prior history.

As a first step in this ambitious direction, we address the sub-problem of predicting privacy risks for a given user. Here, risk refers to a probabilistic adversary that makes educated guesses over partial observations and background knowledge. We specifically consider search histories as observations and aim to predict if and when an adversary can infer a sensitive user property, such as being pregnant or depressed, from user's previous queries.

The technical contributions in this paper are the following:

- a model for user information suitable to capture posts in online communities and search histories,
- a model for the background world knowledge that an adversary may have, including statistical correlations,
- a probabilistic model, based on a Markov Random Field, that captures the transitive "cross-talk" between non-critical and privacy-critical attributes, and
- preliminary experiments with an Internet query log as a first proof of concept.

2. PRIVACY RISKS IN THE AGE OF BIG DATA

There is a huge body of literature on privacy research. However, this paper argues that the big data era brings new challenges that are not met by prior models and methods, and call for a paradigm shift towards user-centric, global, and probabilistic notions of privacy. This section discusses this point informally. The following section develops a first-cut model for capturing some of the issues.

Variety: Prior work focused on sanitizing (e.g., anonymizing or perturbing) single files or databases with privacy-sensitive contents. Typical examples are a table of patient records or the click log of an Internet provider. Today, this data-centric perspective is neither sufficient nor appropriate anymore. Users leave digital traces in many online services and associated datasets: social networks, discussion forums, knowledge sharing communities (e.g., Wikipedia editors), product review sites, commercial web sites (e.g., music subscriptions), query logs, and more. Each dataset may be individually uncritical, but when you put all of them together – which is feasible in the age of big data – they can reveal a scarily intimate picture of users' habits, hobbies, health, and other sensitive information.

Veracity and probabilistic risks: Putting these pieces together, an adversary may still not be able to draw perfect conclusions, like breaking pseudonyms and inferring hard facts about users. We argue, though, that severe damage is often done already when an adversary can make educated guesses about privacy-critical data. The reason is that high-probability cues could trigger additional actions by the adversary such as making explicit inquiries (e.g., financial credit history, previous jobs, international travel) or hacking accounts (e.g., in online communities). For some classes of "soft attacks" (e.g., spam or harassment in social networks), doubts on the veracity of a user's data will certainly not stop an adversary.

Longevity: This situation is further aggravated as user information is collected by service providers and other parties over extended periods of time. Issues that seem harmless today, say when the user is a teenager, might become sensitive years later, for instance when the user applies for a particular job. It is in the human nature to forget what has been said and posted, and hardly anybody keeps detailed records on their personal history on the Internet. However, big service providers can effortlessly keep user data over years if not decades, and advances in information extraction, entity matching, and machine learning will make it easier to compute comprehensive profiles of individual users. Even if a provider uses such data only in an aggregated way over a large population of users (e.g., to train the machine-learning component of a recommendation service), the individual data is still there and could be leaked or used in an unanticipated way.

Adversaries: The traditional picture of an adversary on user privacy is an individual hacker or perhaps a "big-brother-like" surveillance agency. However, the world today is much more complex than this simple model. There are Internet user tracking companies which collect and sell data to third parties, and similar threats exist even in social neighborhoods, for example, "boyfriend tracker" apps on smartphones.

Big Internet companies do not easily share their data for both commercial and privacy reasons; users may thus feel confident that their data is safely guarded. However, in the highly agile modern business world, nobody can really be sure how personal data is combined with other pieces of information and where the integrated data eventually ends. Sensitive information about your health, hobbies, or traveling, may become available to your insurance company, your employer, or a company where you apply for a job. Moreover, government agencies may force commercial companies to provide them with sensitive data, or data may be leaked by sabotage or negligence. Finally, service providers may change their data privacy policies over time, so that previously closely controlled data becomes more widely visible. For example, when companies are acquired or merge with other companies, it is totally unclear what happens with the user data collected so far.

Real-life examples: History shows that even official releases of anonymized datasets for research purposes bear high privacy risks. A prominent example is the scandal that involved the *American On Line* (AOL) search engine [2, 5]. In 2006, AOL released a sample of more than 20 million queries submitted by more than 650,000 users over 3 months. To protect user privacy, the usernames were replaced by anonymous identifiers, but still it turned out that this was insufficient to prevent the de-anonymization of some users. Each anonymous id connected all queries of the same user in the search history, and people could be identified by their entire search profiles. Typical examples of search behavior incurring privacy risk are: ego-surfing (typing one's own name to see the results), searching for one's social security number or phone number, and location directions. Although AOL publicly released its query log with the laudable purpose of helping research, it was criticized and sued. Moreover, two AOL employees were fired after the scandal.

A similar incident happened to Netflix which offered an award to the research team able to improve its recommendation algorithm. Netflix also released a training dataset for the contest and, to protect its customers' privacy, user information was replaced with random ids. However, Narayanan and Shmatikov were able to partially de-anonymize the Netflix training database by linking Netflix anonymous data with the public IMDb database where users reviewed, rated, and discussed movies [25].

These two examples make clear that it is possible to compromise individual information even when data releases are carefully controlled. After these mishaps, some companies keep sharing their datasets but only under research licenses or adopting conservative policies. However, defining such policies is not easy. For example, a popular search engine has recently released a small sample of its query log. Such a log contains only queries, and each query has been posted by at least three different users. This way it is impossible to create sessions or link the queries to a particular user. On the other hand, the utility of this dataset is low as information about query reformulation, clicks after queries, etc. is missing. The query log has therefore no value for research on personalization or improved ranking.

All this needs to be taken into consideration in a modern view of privacy risks, in the presence of big data and powerful data-integration and analytics capabilities. This motivates our goal of helping users with a privacy advisor that, although not being able to perfectly guard a user's sensitive information, can alert and guide people about privacy risks.

3. PRIVACY RISK MODEL

3.1 Framework

User information is constituted by the entirety of their profiles in social networks, postings in online communities, replies, thanks and topical interests in question-answer forums, and keywords and phrases in search histories. For uniform representation, we model all this as a schema-free *data space* of *key-value (KV) pairs* (or equivalently, subject-predicate-object triples with the subject being the user). For example, a user may have data attributes like `hasGender=female, livesIn=USA, visitedCountry=India, visitedCountry=Brazil, tookMedicalDrug=Tylenol, searchedFor=anxiety, searchedFor=depression`, etc. Some of these may come from structured profiles, while others are assumed to be derived from textual contents or photo captions by information extraction (IE) methods. In cases when IE does not work or does not make sense (e.g., in search-engine query log), salient keywords or phrases can be represented as well.

We assume that a specific subset of keys or key-value pairs is considered privacy-sensitive by the user. For example, all values for `tookMedicalDrug` may be in this category, or only these where the value is in a prespecified set of critical drugs such as anti-depressants (as opposed to simple painkillers like Tylenol). The privacy-sensitive part of the dataspace could be explicitly specified by the user, or by an expert on behalf of user groups, or it could be automatically learned from interactions with users. In this paper, we do not deal with this issue but simply assume that the sensitive KV data is known upfront to the privacy risk model, while being inaccessible by the adversary.

The adversary aims to make an educated guess at one or more of the user's privacy-sensitive KV pairs. For example, the adversary could infer that the user with the aforementioned key-value pairs is likely to take also the drug Xanax or Prozac. The adversary's power comes from partial observations, $O = \{O_1, O_2, \dots\}$, of some of the user's KV data, and from general world knowledge, W, including statistical correlations among KV items. The latter could be mined by the adversary from large web crawls, big-data analyses of social-media contents or query logs, and other estimators over anonymized data at large scale. This is exactly what may lead the adversary to infer that `tookMedicalDrug=Xanax` may hold for the user at hand.

So the task of the adversary is to estimate the probability $P[S|O,W]$ that a critical KV property S holds for a user, given the observations O and the background knowledge W. In our example, this could be the test variable:

$$P[\mathit{tookMedDrug} = \mathit{Xanax} \mid \mathit{tookMedDrug} = \mathit{Tylenol}, \\ \mathit{searchedFor} = \mathit{depression}, \mathit{searchedFor} = \mathit{anxiety}].$$

In order to anticipate such potential inferences of statistical nature, the user needs to be aided by an equally powerful privacy-advisor tool, which has the same world knowledge and global statistics in addition to knowing which of the user's data is sensitive. This tool simulates the adversary and, this way, predicts privacy risks.

Without any knowledge about a specific user U, the adversary can only assume that the key-value distribution for this user is the same as for the world knowledge:

$$P_U[K_1, K_2, \dots] = P[K_1, K_2, \dots |W]$$

for keys (attributes) $K_1, K_2, \dots$ in the underlying data space. As the adversary observes certain attribute values $O_1, O_2, \dots$ about U, the multivariate distribution for U can be refined into:

$$P[K_1, K_2, \dots |W, O_1, O_2, \dots].$$

Thus, our goal in assessing the privacy risk for U is to estimate this joint distribution, or more specifically, the marginal probability for each sensitive key S, taking into account strongly positive or strongly negative correlations among keys.

For example, suppose S is the key-value pair `tookMedDrug=Xanax`. The world knowledge of the adversary may suggest that:

$$\begin{aligned} P_U[S|W, O_1, O_2, O_3, O_4] = P[\mathit{tookMedDrug} &= \mathit{Xanax} \mid \\ \mathit{hasGender} &= \mathit{female}, \\ \mathit{livesIn} &= \mathit{USA}], \end{aligned}$$

is reasonably low, say 0.01, based on global statistics for pharmaceutical drug consumption and the correlations with gender and geo-location. However, after observing the search history, the refined probability becomes:

$$\begin{aligned} P_U[S|W,O_1,O_2,O_3,O_4] = P[tookMedDrug &= Xanax \mid \\ hasGender &= female, \\ livesIn &= USA, \\ searchedFor &= depression, \\ searchedFor &= anxiety], \end{aligned}$$

which would typically be much higher, say 0.3. This may be sufficient for the adversary to guess that the user takes anti-depressants, which in turn may trigger notifying the user's health insurance or employer.

3.2 Use-Case: User Search History

As a special case of our framework, we examine the risks hidden in an Internet search history, aggregated per user. Sensitive states of a user, be it pregnancy or depression, can manifest themselves through certain queries. Some are directly indicative of the state, but some can be linked to the state only with the help of background knowledge, and only in the right context. We transform the user search histories into a KV representation to be able to apply probabilistic reasoning for sensitive state prediction.

Assume the query set contains queries $q_1, q_2, ..., q_n$. For each of the queries q_i, we introduce a binary random variable $searchedForq_i$ whose value is 1 if the user's search history contains the query q_i, and 0 otherwise. A user profile is then represented by a set of key-value pairs of variables and their values.

For example, if the queries known to the model are $\{anxiety, xanax, therapy, fatigue\}$, and a user's query history contains the queries $\{anxiety, therapy\}$, then the user's profile is the set:

$$\{searchedForAnxiety = 1, searchedForXananx = 0,$$
$$searchedForTherapy = 1, searchedForFatigue = 0\}.$$

4. PROBABILISTIC PREDICTION

For reasoning over a user's KV data and predicting the privacy risk for certain sensitive items, we use probabilistic graphical models that capture correlations among random variables. Specifically, we adopt the Alchemy toolkit for Markov Logic [1, 8]. This tool supports probabilistically weighted first-order logic formulas that are automatically compiled into a Markov Random Field (MRF), with various kinds of inference.

For the case of search histories, we introduce two kinds of predicates. The *user-state* predicates are of the form $IsAlcoholic(X)$, $IsDepressed(X)$, or $IsPregnant(X)$, where X is a user variable. The inference task is to predict the values of those predicates for each user. The *search-history* predicates are like $searchedForAnxiety(X)$, $searchedForStress(X)$, $searchedForUltrasound(X)$, etc.

On top of those predicates, we define rules of the form $P(X) \wedge Q(X) \Rightarrow R(X)$ that capture the potentially relevant correlations, e.g.:

$$searchedForStress(X) \wedge searchedForPsychiatrist(X) \Rightarrow searchedForXanax(X).$$

The weights of these rules reflect the adversary's background knowledge and general statistics. One can obtain such weights by mining large datasets or text corpora for co-occurrences and correlations. For example, for rules of the form:

$$F_i = searchedForTerm1(X) \wedge searchedForTerm2(X) \Rightarrow searchedForTerm3(X),$$

they can be estimated as conditional probabilities of the form:

$$w_i = P[term3|term1, term2].$$

We also introduce rules correlating sensitive KV items and the states we aim to predict. The weights for those are manually set to a relatively high value, reflecting the model's assumption that the sensitive queries directly indicate sensitive states, e.g.:

$$searchedForXanax(X) \Leftrightarrow IsDepressed(X).$$

The entirety of weighted formulas constitutes the adversary's model.

Starting with this model, the adversary would now obtain a set of observations about a user U in the form of KV items, excluding the sensitive ones. The Markov Logic model is then enriched with setting the observed predicates for U to their proper truth values, e.g., $searchedForAnxiety(U) = true$. Finally, we run Alchemy's inference algorithm for marginal probabilities on the sensitive state KV items, e.g., $IsDepressed(U)$. A high output value denotes a privacy risk for U regarding the given sensitive attribute. To conclude that there is a notable risk, we can either use a threshold, or simply compare the output values for different variables or across different users. The next section, on initial experiments, refines this into more detail.

5. PRELIMINARY EXPERIMENT

5.1 Setup

For a preliminary proof of concept, we experimented with the AOL query log, which comprises ca. 21 million queries by ca. 650,000 distinct users, collected between March 1 and May 31, 2006. The average number of queries per user is 31.5, and the 90th percentile of queries per user is 76.

For the purpose of predicting privacy risks in users' search histories, we identified three sensitive topics:

- `hasAddiction = Alcoholism`
- `hasDisease = Depression`
- `hasCondition = Pregnancy`

For each topic, we manually identified a number of keywords or phrases that are a) highly *indicative* of the topic, or b) *suggestive* of the topic, or c) *common* - including ambiguous ones with multiple meanings. By indicative we mean terminology that more or less paraphrases the topic, so that an adversary with strong background knowledge would guess the sensitive information with high certainty. Suggestive terms, on the other hand, are positively correlated with the sensitive topic, but by themselves are not sufficiently strong cues to reveal the sensitive information. Finally, common terms occur frequently in search histories, but are either harmless regarding privacy or have multiple meanings, only some of which are suggestive (e.g., words like "help," "delivery," etc.). Table 1 shows the terms for the three topics in our experiment.

	Alcoholism	**Depression**	**Pregnancy**
Indicative	alcoholism, alcohol abuse, alcoholics, alcoholic anonymous, alcohol dependence	clinical depression, depressive, antidepressant, xanax, prozac	pregnancy, childbirth, baby kicking, expecting child, prenatal diagnosis
Suggestive	drunk, alcohol, blackout, therapy, liver	anxiety, suicide, psychiatrist, psychotherapy, mood disorder	missed period, nausea, morning sickness, gynaecologist, obstetrical
Common	dependence, anonymous, drink, drive, problem, liquor, beer, medication, party, help	mood, single, solitude, alone, self, pain, stress, problem, help, hope	labor, delivery, trimester, birth, child, baby, mother, paternity, sex, craving

Table 1: Indicative, suggestive, and common terms for privacy-sensitive topics

For each of the three sensitive topics, we select 3×5 top-ranked users from the AOL query log based on the following ranking criteria: i) highest overlap of the terms in the user's search history with indicative or suggestive terms (as specified per topic), ii) highest overlap with common terms, iii) highest average of the two overlap measures. The rationale for this choice is that we wanted both users whose search history is likely privacy-critical and users whose queries are more general but contain some sensitive words including ambiguous ones.

For each of the selected "test users," we mask out all queries that contain at least one indicative term. The remaining search history is fed into our probabilistic model to compute a quantitative measure of privacy risk.

We compute the weights for the Markov Logic rules using a subset of the AOL query log, choosing the users who posted at least 20 queries, and who queried for at least one risky term, excluding the users from the test set. We are specifically interested in the co-occurrence of query terms in query logs. For each rule of the form:

$$searchedForTerm1(X) \wedge searchedForTerm2(X) \Rightarrow searchedForTerm3(X),$$

we retrieve the number of distinct users who queried for $term1$ and $term2$, and the number of distinct users who queried for $term1$, $term2$, and $term3$. The weight of the formula is then set to:

$$\frac{\#users(term1, term2, term3)}{\#users(term1, term2)}.$$

To assess the viability of our prediction model, we compare its output against ground-truth judgements obtained by showing the full history of each test user to knowledgeable humans to determine whether the history has strong evidence for one of the sensitive states (i.e., Alcoholism, Depression, Pregnancy). Our model computes a prior probability for a KV item, without any observations about the user, and a posterior marginal probability, given the observations. Whenever the difference between the posterior and prior probability is greater than a threshold of 0.05 [32], we interpret this as a privacy risk for the given user.

We compute *precision* and *recall* measures for our model by comparing these predictions against the ground-truth.

5.2 Results

The precision and recall results for the three test topics are shown in Table 2. Our preliminary results show that our method could indeed serve as a predictor of privacy risk. Employing correlations between different search terms, and logical rules that enhance the importance of variable co-occurrences, it is possible to predict privacy risks despite the most indicative terms being masked out. For two of three topics (Alcoholism and Pregnancy), we were able to detect users in sensitive states with high recall and good precision.

	Precision	**Recall**
Alcoholism	0.60	0.75
Depression	0.67	0.33
Pregnancy	0.50	0.83

Table 2: Experimental results for privacy risk prediction

5.3 Model Limitations

With respect to the precision of predictions, we observed mixed results due to different factors. First of all, some users post risky terms in non-sensitive contexts (professional, educational, and other). We acknowledge that a richer background knowledge model could help distinguish the truly private queries. Moreover, some examples in the dataset were borderline, so that even for the human assessors the sensitivity of the state was unclear. We illustrate the limitations of the current model with the following examples.

Query context: There are users who submitted many queries containing sensitive keywords, yet for a human observer it is clear that the context of those queries is not sensitive. Consider for example a user who queried with many suggestive and common terms for alcoholism: *medication, blackout, help, beer, drinking, alcohol, drive, drink, drunk, party, therapy.* Upon closer inspection one can see that these terms appeared either without any context, like *medication*, or in a clearly non-sensitive context, like *blackout lyrics*, or *drinking too much water.* Sometimes the context was indeed sensitive but the target state was different than the one we tried to predict. Our model, for instance, captured the keyword *therapy* as sensitive for alcoholism in the query *therapy eating disorders.* Such a query of course does not indicate alcoholism, especially in the presence of other queries related to eating disorders, like *signs of bulimia.*

An example of a false negative was a user who posted relatively few sensitive keywords: *alcohol, liver, therapy*, yet those keywords were repeatedly used in the entire user history in queries of evidently sensitive context: *how liver detoxifies alcohol, xanax for alcohol withdrawal symptoms, inflammation of the liver* or *hep c false positive alcohol.*

User background: Another tricky setting for models based on bag-of-words representations is when users issue queries with sensitive terms, but the complete user history reveals to a human observer that the queries are related to a person's profession rather than a personal state. Such disambiguating queries include those with the intent of information gathering: *clinical psychology review, medical dictionary, google scholar*, those containing highly specific terms: *anorexia nervosa support groups*, or related institution names: *american counseling association*. Sometimes risky terms can be classified as posted by students, especially in the presence of other study-related queries: *scholarly articles on adolescent depression, papers on mozart, topics for music term paper, university master in anthropology.*

Temporal dimension: A full-fledged privacy monitoring model should take into account the temporal patterns of user behavior. Human evaluators would rather not judge that a person is depressed, alcoholic, or pregnant if they used risky terms only in one session over a longer period of time. Those are recurring inquires over sensitive topics that constitute stronger indicators of sensitive states.

6. RELATED WORK

Privacy-friendly data publishing: There is a large body of literature on sanitizing datasets to be published, in order to prevent adversaries from identifying users and inferring values of sensitive attributes for individual users. To this end, the notion of k-anonymity has been proposed by [30], where each database record should be indistinguishable from at least $k-1$ other records, regarding the values of non-sensitive attributes (e.g., age, gender, zipcode). This notion has been extended in [21] into the notion of l-diversity. In this variant, it is demanded that all the records within blocks defined by the values of non-sensitive attributes have at least l distinct values for each sensitive attribute. This line of work has been further continued in [18], with the introduction of the t-closeness principle, which says that the distribution of the sensitive values within a block of records with the same non-sensitive attributes should be similar to the distribution of this attribute in the entire table.

Privacy-friendly data analysis: As an alternative to publishing an entire dataset, many approaches assume that certain statistics will be computed over the data and only the results will be released. In differential privacy [9], the privacy restriction is applied to the algorithms that are run on a given database. To protect against inferring the presence or absence of particular individuals in the dataset, query answers are perturbed by adding controlled noise. Firm theoretical guarantees ensure that the presence or absence of individual records does not substantially change the output of the statistical queries. The PINQ platform has been developed by [22], to support data analysts in differentially private computations over statistical databases. In [14], it is argued that the differential privacy definition should be further generalized so that an adversary cannot infer whether an individual participated in the data generation process.

Privacy and linkability in social networks: Privacy has been examined also in the context of online social networks. In [33] the authors present methods for user de-anonymization in social networks, leveraging graph structures. In [10] it is shown how user accounts can be matched across different social networks. [23] proposes linguistic stylometry attacks to identify anonymous text authors. Some of this work can be seen as addressing a specialized version of the record linkage problem: matching records from different databases. So far, the emphasis has been on demonstrating attacks in such linkability settings. There is some work on preventing linkage by eliminating joinable attributes from the underlying datasets (e.g., [12, 16, 31]). However, as we deal with user-posted data in social media, this kind of control cannot be easily exercised in real-life settings. Moreover, it would not cover cases where users could possibly be linked across communities by a wider combination of different profile values, posted contents, and social behavior (replies, likes, thanks, etc.). In general, we view linkability as orthogonal to assessing privacy risks. Linking across datasets increases risks, but inferring values of privacy-critical attributes is yet another step.

Sanitizing query logs: Although the aim of this paper is not sanitizing query logs for public release, the prior work on this topic is remotely related. Search-engine logs capture queries and other user information. Such data is a valuable source of knowledge which can be used to improve the user experience on the web [3]. Some of the reasons why search engines retain query logs are: improvement of ranking algorithms and of language-based applications, query refinement and personalization, combating fraud and abuse, sharing data for academic research, marketing, and other commercial purposes [7]. The other side of the coin is that publishing query logs can reveal sensitive information about users.

According to Bar-Ilan [4] it is possible to sanitize a query log to prevent disclosure of private information. Anyway, sanitizing query logs is all but an easy task. Jones et al. [13] proved that search logs containing session information may cause privacy risks, so scrubbing query content, which consists in removing identifying queries (e.g., names, addresses, phone numbers) is not sufficient to guarantee the protection of user privacy. In addition, the authors showed how to use simple classifiers to map a sequence of queries onto gender, age, and location of the user. These classifiers can be further combined to map a sequence of queries into a small set of candidate users. This would allow a real-life acquaintance of one user to use her background knowledge and identify the user in the log. Kumar et al. [17] proposed to tokenize the query log and apply a secure hash function upon each token. This anonymization technique seems strong, but it does not give full privacy guarantees. Indeed, the authors proved that using statistical techniques on a reference log can still allow to disclose private information. Adar [2] pointed out that traditional privacy preservation methods cannot be applied directly to query logs and that k-anonymity is too costly due to the rapid changes of data. He also suggested that deleting unfrequent queries may be an effective way to sanitize query logs. The drawback of this approach is that it is too conservative: Most of the user queries appear once or twice in query logs [6], so even an extremely low threshold may cause the elimination of a big fraction of non-identifying queries.

Recently, other researchers proposed algorithms to sanitize search logs. Korolova et al. [15] presented an algorithm which creates a private query-click graph containing only queries that are submitted by at least k users, where the value of k is fixed by the data publisher. This approach provides guarantees similar to differential privacy, and the query click graph can be used to improve query suggestion and spelling correction. [11] describes an algorithm to build a histogram of query counts that is modified by an additive noise and proved that search logs sanitized with this algorithm can be used for index caching and query substitution.

Finally, [7] is a survey on privacy-enhancing techniques including: log deletion, queries hashing, identifier deletion, identifier hashing, query content scrubbing, infrequent queries deletion, and sessions shortening. All these techniques are evaluated by considering 1) privacy protection, 2) utility preservation, which measures the usability of the anonymized data for statistical analysis, and 3) user control, which is defined as a mechanism that allows individual users to choose to have the technique applied to their own query logs.

Probabilistic models: Probabilistic inference is a central idea in our approach. Statistical models have been present in the field of data privacy in different forms. Some methods give probabilistic privacy guarantees, like differential privacy [9], membership privacy [19], and stochastic privacy [29]. In [20] privacy-preserving time-series analysis has been applied in the context of monitoring user browsing patterns. Last but not least, probabilistic graphical models, specifically latent topic models like LDA, have been used in [32] to enable obfuscating topical interests of users.

7. CONCLUSION

In this paper, we advocated a new model for privacy risk in a user-centric, global, and probabilistic manner, geared for adversaries with rich background knowledge and statistics for making educated guesses. We specialized our framework for the case of search-engine users where privacy risks come from observing a user's query history. Our preliminary experiments are a first proof-of-concept for our model and probabilistic inference methods.

Although the probabilistic predictions by our method gave informative results, we still face major limitations that are subject of future research. In particular, our model does not consider neither the search context, nor the structure of search sessions and temporal patterns in user behavior. For example, some users spend much time in long sessions about a sensitive topic, and others repeat certain kinds of queries almost every day. Our future work will include addressing such issues, and also exploring use-cases with social media beyond search histories.

8. REFERENCES

[1] Alchemy: Open Source AI, Alchemy Software Version 2.0, http://alchemy.cs.washington.edu/, last accessed on July 1, 2014

[2] E. Adar. User 4xxxxx9: anonymizing query logs. In Proceeding of the Workshop on Query Log Analysis at the 16th International World Wide Web Conference (WWW 2007), May 2007

[3] R. Baeza-Yates. Applications of web query mining. In Proceedings of the 27th European Conference on Information Retrieval Research (ECIR '05), pages 7–22, Berlin, Heidelberg, 2005. Springer-Verlag

[4] J. Bar-Ilan. Access to query logs – an academic researcher's point of view. In Query Log Analysis: Social And Technological Challenges Workshop at the 16th International World Wide Web Conference (WWW 2007), May 2007

[5] M. Barbaro and T. Zeller. A Face Is Exposed for AOL Searcher No. 4417749. The New York Times, August 2006

[6] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder. Hourly Analysis of a Very Large Topically Categorized Web Query Log. In the Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004), pp. 321–328, New York, NY, USA: ACM, 2004

[7] A. Cooper. A survey of query log privacy-enhancing techniques from a policy perspective. ACM Transactions on the Web, vol. 2, no. 4, pp. 1 – 27, 2008

[8] P. Domingos and D. Lowd. Markov Logic: An Interface Layer for AI . Morgan & Claypool, 2009

[9] C. Dwork. Differential Privacy: A Survey of Results. TAMC 2008

[10] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira. Exploiting Innocuous Activity for Correlating Users Across Sites. In Proceedings of the 22nd International Conference on World Wide Web (WWW 2013), pp. 447–458, 2013

[11] M. Götz, A. Machanavajjhala, G. Wang, X. Xiao, and J. Gehrke. Publishing Search Logs: A Comparative Study of Privacy Guarantees. IEEE Trans. on Knowledge and Data Engineering, vol. 24, no. 3, pp. 1041– 4347, March 2012

[12] R. Hall and S. E. Fienberg. Privacy-Preserving Record Linkage. Privacy in Statistical Databases 2010: 269-283

[13] R. Jones, R. Kumar, B. Pang, and A. Tomkins. "I know what you did last summer" – Query logs and user privacy. In Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management (CIKM 2007), pp. 909 – 914, New York, NY, USA: ACM, 2007

[14] D. Kifer and A. Machanavajjhala. No Free Lunch in Data Privacy. ACM SIGMOD 2011

[15] A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas. Releasing search queries and clicks privately. In Proceedings of the 18th International Conference on World Wide Web (WWW 2009), pp. 171–180. New York, NY, USA: ACM, 2009

[16] H. C. Kum, A. Krishnamurthy, A. Machanavajjhala, M. K. Reiter, and S. Ahalt. Privacy preserving interactive record linkage (PPIRL). J Am Med Inform Assoc. 21(2):212-20, 2014

[17] R. Kumar, J. Novak, B. Pang, and A. Tomkins. On anonymizing query logs via token-based hashing. In Proceedings of the 16th International Conference on World Wide Web (WWW 2007). ACM Press, New York, pp. 629 – 638, 2007

[18] N. Li, T. Li, and S. Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. IEEE ICDE 2007

[19] N. Li, W.H. Qardaji, D. Su, Y. Wu, and W. Yang. Membership privacy: a unifying framework for privacy definitions. ACM CCS 2013

[20] F. Liyue, L. Bonomi, L. Xiong, and V. Sunderam. Monitoring Web Browsing Behavior with Differential Privacy. WWW 2014

[21] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. IEEE TKDD 1(1), 2007

[22] F. D. McSherry. Privacy Integrated Queries: An Extensible Platform for Privacy-preserving Data Analysis. ACM SIGMOD 2009

[23] A. Narayanan, H. Paskov, N. Z. Gong, J. Bethencourt, E. Stefanov, E.C.R. Shin, and D. Song. On the

Feasibility of Internet-Scale Author Identification. In Proceedings of the 2012 IEEE Symposium on Security and Privacy (SP 2012), pp. 300–314, IEEE Computer Society, Washington, DC, USA, 2012

[24] A. Narayanan, V. Shmatikov. De-anonymizing Social Networks. IEEE SP 2009

[25] A. Narayanan, V. Shmatikov. Robust De-anonymization of Large Sparse Datasets. In IEEE Symposium on Security and Privacy, pp. 111 – 125, May 2008

[26] H. H. Pang, X. Ding, and X. Xiao. Embellishing Text Search Queries to Protect User Privacy. PVLDB 3(1): 598-607 (2010)

[27] H. H. Pang, X. Xiao, and J. Shen. Obfuscating the Topical Intention in Enterprise Text Search. ICDE 2012: 1168-1179

[28] X. Shen, B. Tan, and C. Zhai. Privacy Protection in Personalized Search. SIGIR Forum 41(1): 4-17 (2007)

[29] A. Singla, E. Horvitz, E. Kamar, and R. White: Stochastic Privacy. AAAI 2014

[30] L. Sweeney. k-Anonymity: A Model for Protecting Privacy. Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5), 2002

[31] D. Vatsalan, P. Christen, and V. S. Verykios. A taxonomy of privacy-preserving record linkage techniques. Inf. Syst. 38(6): 946-969 (2013)

[32] P. Wang and C. V. Ravishankar. On Masking Topical Intent in Keyword Search. ICDE 2014

[33] A. Zhang, X. Xie, K. C. Chang, C. A. Gunter, J. Han, and X. Wang. Privacy Risk in Anonymized Heterogeneous Information Networks. In Proc. 17th International Conference on Extending Database Technology (EDBT 2014), Athens, Greece, March 24-28, 2014