R-Susceptibility: An IR-Centric Approach to Assessing Privacy Risks for Users in Online Communities

Joanna Asia Biega[1], Krishna P. Gummadi[2], Ida Mele[3],
Dragan Milchevski[1], Christos Tryfonopoulos[4], Gerhard Weikum[1]

[1]Max Planck Institute for Informatics, Germany [2]Max Planck Institute for Software Systems, Germany
[3]Università della Svizzera Italiana, Switzerland [4]University of Peloponnese, Greece

{jbiega, dmilchev, weikum}@mpi-inf.mpg.de,
gummadi@mpi-sws.org, ida.mele@usi.ch, trifon@uop.gr

ABSTRACT

Privacy of Internet users is at stake because they expose personal information in posts created in online communities, in search queries, and other activities. An adversary that monitors a community may identify the users with the most sensitive properties and utilize this knowledge against them (e.g., by adjusting the pricing of goods or targeting ads of sensitive nature). Existing privacy models for structured data are inadequate to capture privacy risks from user posts.

This paper presents a ranking-based approach to the assessment of privacy risks emerging from textual contents in online communities, focusing on sensitive topics, such as being depressed. We propose ranking as a means of modeling a rational adversary who targets the most afflicted users. To capture the adversary's background knowledge regarding vocabulary and correlations, we use latent topic models. We cast these considerations into the new model of R-Susceptibility, which can inform and alert users about their potential for being targeted, and devise measures for quantitative risk assessment. Experiments with real-world data show the feasibility of our approach.

Categories: [H.1.0] **Information Systems**

Keywords: Privacy Risk Assessment, Online Communities

1. INTRODUCTION

1.1 Motivation and Background

The goal of this paper is to provide privacy risk assessments for users in online communities. An online post may directly or indirectly disclose personal information, such as gender, age, political affiliation, or interests. An adversary can combine such observations with his background knowledge of correlations between different attributes to infer privacy-sensitive information and discriminate against users. We argue that existing privacy models for structured data, such as k-anonymity [33], l-diversity [24], t-closeness [22], membership privacy [23] and differential privacy [10], are inherently inappropriate to capture these situations. One reason is that user posts in social media are mostly of textual form, inducing a very-high-dimensional data space of word-level or phrase-level features. A second reason is that users might not want to be prevented from posting contents, but instead be selectively warned about emerging privacy risks. In our setting, certain assumptions also differ from the assumptions of prior work on privacy-preserving data publishing [13]: users do want to post information, but they should be aware of possible exposure and targeting risks. For these reasons, this paper pursues an IR-centric approach to privacy, making novel use of topic models and ranking.

Scenario: To understand why adversaries and user risks are different from the privacy concerns for structured databases, consider the following scenario. An unscrupulous drug company wishes to advertise its new anxiety-reducing drug to Facebook users. It decides to target ads at a million users that are most susceptible to be afflicted by depression within the 1 billion population of Facebook. The company plans to infer users' demographics by text mining their posts and combine it with the background knowledge correlating demographics and certain vocabulary usage with depression, obtained from text mining an archive of medical journals. In such a scenario, how can a Facebook user estimate her risk of being targeted? Similar issues arise also within specialized online communities such as `healthboards.com` or `patient.co.uk`. Although these have a much smaller scale, a smart adversary would still target only a subset of highly susceptible users to avoid the impression of mass spamming.

Targeted ads of sensitive nature is one kind of risk, but there are even more severe threats with real cases reported: scoring users for financial credit worthiness or insurance payments, factoring a user's social-media posts in assessing her job application, and more (e.g., [7, 12]). Despite these being big trends, most users do not need hard guarantees regarding privacy (e.g., preventing de-anonymization by all means), and perfect anonymity cannot be guaranteed without severely diminishing the utility of social media. For example, someone who always posts using an unlinkable anonymous identity cannot build up a reputation as a credible information source. Conversely, making all posts under a pseudonym is insufficient to prevent tracking-and-rating companies (e.g., `www.spokeo.com`) from linking user accounts across different social platforms.

SIGIR '16, July 17-21, 2016, Pisa, Italy

 ISBN 978-1-4503-4069-4/16/07…$15.00

DOI: `http://dx.doi.org/10.1145/2911451.2911533`

Therefore, we focus on the assessment of privacy *risks* and on alerting users to support their awareness, rather than pursuing the elusive goal of enforcing privacy.

Existing privacy models fail to capture theses issues, along the following dimensions:

- *Data model*: Privacy models like k-anonymity or differential privacy are primarily geared for structured data or content that can be cast into low-dimensional feature spaces. Capturing risks from textual contents in online communities faces the problem of high-dimensional feature vectors (e.g., word bigrams). Prior work that coped with text focused on specific settings, such as predicting sensitive posts [30], sanitizing the information from query logs (e.g., [6, 28]), or publishing high-dimensional datasets [9]. Our goal, on the other hand, is to be able to quantify privacy risks from text in a generic way.
- *Adversary's background knowledge:* Prior work on privacy assumes computationally powerful adversaries, but disregards or makes special assumptions about the background knowledge that an adversary may have beyond the dataset at hand. However, adversaries may easilly tap into many datasets including large text corpora, thus obtaining a model of the typical vocabulary used by potential targets as well as semantic dependencies or statistical correlations between topics.
- *Disclosure vs. discrimination risk:* Existing privacy models focus on limiting information disclosure, but they do not capture the exposure within a community with regard to sensitive properties. Standing out in a community this way may result in discriminatory treatment, such as being rejected for loans or job applications, or receiving ads of sensitive nature.

1.2 Approach and Challenges

This paper introduces *R-susceptibility*: a ranking-based privacy risk model for assessing users' privacy risks in online communities, accompanied by IR-style risk measures for capturing risks from textual contents. The model is very versatile: we demonstrate how it can capture user posts or search queries, but it can also be used with click streams, and other online activities. Semantic dependencies and statistical correlations among words and sensitive topics are represented using latent topic models, such as LDA [5] or Skip-grams [25]. This way, we anticipate adversaries with rich background knowledge. Adversaries are assumed to be rational: they target only a fraction of "promising" users. Therefore, we model a user's risk as the ranking position in the community when the users are ordered by relevance to sensitive topics, such as pregnancy, depression, financial debts, etc. This ranking-based model is meant to alert the users whenever critical situations arise. We posit that users might be then guided to selectively post anonymously.

Our model addresses several technical challenges:

- *Sensitive vs. general topics:* A trained latent topic model does not indicate which of the topics are privacy-sensitive. We carried out a crowdsourcing study to identify sensitive topics. Our study differs from the prior work of [30] as the latter relied on explicit categories.
- *Personal vs. professional interest:* A user who posts about a sensitive topic may merely have a professional or educational interest without being personally afflicted. To be able to rank such users lower, our model introduces the notion of topical *breadth* of interest, complementing the user's strength of interest in a sensitive topic.
- *Personal interest vs. curiosity:* A user may become interested in a topic out of curiosity, perhaps prompted by an external event (e.g., a celebrity scandal). To be able to rank such non-critical users lower, our model also considers the *temporal variation* of interest in a topic.

The paper's salient contributions are:

- a novel approach to privacy risks focusing on exposure in user rankings within online communities, and emphasizing risk awareness;
- a framework for quantifying privacy risks from textual contents in online communities, based on latent topic models and user rankings;
- measures for computing risk scores with regard to a given sensitive topic based on users' posts or search queries.

2. R-SUSCEPTIBILITY MODEL

2.1 Sensitive states and adversaries

We assess the risk of a user being perceived as afflicted by a *sensitive state*, such as being depressed, pregnant, or having financial debts. An adversary in our model attempts to find the most susceptible users, that is, the users who are most exposed with regard to a sensitive state. For instance, an adversarial insurance company might want to identify the users who are likely afflicted by certain diseases, an adversarial HR department of a company might want to screen for the users with likely drug or alcohol problems, while a seller of illegal anti-depressants might want to find the users most likely to be depressed, and thus prospective customers.

We therefore propose *ranking* as a means of modeling a rational adversary trying to identify the most susceptible users. To rank the users with respect to a given sensitive state, an adversary needs to choose a measure of quantitative risk assessment based on the contents of user profiles. We discuss several such measures in Section 3.

2.2 Sensitive topics

We associate sensitive states with a vocabulary distribution, i.e., distributional vectors of related words. For example, the topic *financial debts*, could be captured by related words and phrases like *loan, mortgage, money, problem, sorrows*, or *sleepless night*. Such salient phrases related to a sensitive state can be obtained by unsupervised or semi-supervised training of latent topic models over external datasets such as news archives, digital libraries or large crawls of social media. This way we capture the adversaries' background knowledge about the vocabulary for a topic and about semantic dependencies and correlations.

Sensitive states might manifest themselves in the online contents of users. User posts can also be characterized as distributional vectors of salient words. Then, the similarity between the distributional vectors of the user's posts and a sensitive topic can be used to assess the user's susceptibility to being exposed with regard to that topic.

2.3 Background knowledge

An adversary in our model is assumed to be interested in a sensitive state and aims to target a fraction of the most afflicted users. The adversary has *background knowledge*,

characterized by statistical language and topic models. This is a natural form of useful knowledge for a rational adversary who wants rank the users based on the textual contents, and to bound the cost of his targeting efforts.

In this paper, we consider three versions of adversary's background knowledge. The basic version is the knowledge of the most salient words for different topics, which is assumed in all the solutions we explore. The more advanced version assumes that the adversary is able to compute similarities between words, in the sense of semantic relatedness. Finally, in some of the solutions, we assume an adversary is able to assign latent topics to broader thematic domains, e.g., the topic of depression to the domain of psychiatry.

We believe that this model reflects a wide class of adversaries whose goal is to discriminate and target the most susceptible users in online communities.

2.4 R-Susceptibility

We propose R-Susceptibility (Rank-Susceptibility) as a measure of a user's privacy risk. To measure R-Susceptibilty with respect to a sensitive topic, we first rank all users within an online community based on their decreasing susceptibility of being exposed with regard to a sensitive topic (as described above) and then compute the position where the user is ranked.

Intuitively, the R-Susceptibility model could also have the following IR interpretation: we rank the users according to the relevance of their posts to a query being the words of a sensitive topic, and choose the top-ranked, who are most likely to be personally afflicted.

3. RISK ASSESSMENT MEASURES

Risk measures are plug-in components in the framework and orthogonal to the idea of R-Susceptibility. In this paper, we begin by investigating three kinds of risk scores, leaving an extended risk-measure study as future work.

The first two of the risk scores are baselines, inspired by standard measures in privacy research, namely, the entropy of attribute value distributions (as used in the t-closeness model) and the changes in the global probability distributions of attribute values incurred by the inclusion of an individual user's data (as used in the differential privacy model). The third measure is a novel IR-centric score based on topic models, capturing lexical correlations and three different characteristics of user interest in a topic: the strength of interest, the breadth of interest, and the temporal variation of interest.

Desired properties. By considering the community and interpreting risk with respect to a user's rank in the community, our framework does not impose any restrictions on the absolute values or the value domains of valid risk measures. Intuitively, for the framework to function, we expect a good measure to correlate with human assessments on the sensitivity of user profiles: the more human observers agree that a user might be in a senitive state, the higher the value of the risk score should be.

3.1 Entropy baseline measure

The entropy baseline measure is inspired by comparing a global probability distribution (for an entire community) against a local distribution (for an individual user) using relative entropy (aka KL divergence). We apply this measure to textual data as follows.

Let X be a sensitive topic, and $\{x_1, ..., x_j\}$ be the salient words and phrases of X. The knowledge of this vocabulary for different topics is assumed to be a part of the adversarial background knowledge (e.g., derived from latent topic models). We treat $x_1, ..., x_j$ as database attributes and represent users as database records where the value of an attribute x_i equals to 1 if the word appears in the user's contents, and to 0 otherwise.

Let U_0 be the user for whom we wish to compute the risk score with respect to X, and $U = \{U_1, .., U_k\}$ be the set of other users in the community. Let further be $U^* = \{U_0\} \cup U$, and let P_U, P_{U^*} denote the distributions of attribute values for U and U^*, respectively.

We compute the risk score by averaging the relative entropy of the univariate distributions P_U, P_{U^*} for the individual attributes $\{x_1, ..., x_j\}$. Note that measuring the relative entropy over the multivariate joint distributions of attributes could be an alternative, but we do not pursue this here because of the data sparseness that we would face.

Definition 1 (Entropy baseline risk score of topic X for U_0)**.** *The entropy baseline risk score of the user U_0 with respect to a topic X is:*

$$risk_{\text{ENT}}(U_0, X) = \frac{1}{j} \sum_i \sum_{v=\{0,1\}} P_U[x_i = v] log(\frac{P_U[x_i = v]}{P_{U^*}[x_i = v]}).$$

The ranking method based on this definition is being referred to as ENT.

Measure properties. It holds that $risk_{\text{ENT}}(U_0, X) \geq 0$. The lowest value of 0 is reached when the user does not have any of the topic's salient attributes in her observable contents. Otherwise, the risk score is lowest when half of the community's users exhibit an attribute in their contents and highest when all or none of the users have the attribute.

3.2 Differential-privacy baseline measure

The differential-privacy-based measure is inspired by the definition of differential privacy, that is calculating the changes of attribute probabilities incurred by the inclusion of a user's data. Let X, $\{x_1, ..., x_j\}$, U_0, U, U^*, P_U, and P_{U^*} be defined as in the previous section. The differential privacy principle requires that

$P_U[x_i] \leq 2^\epsilon P_{U^*}[x_i]$ and $P_{U^*}[x_i] \leq 2^\epsilon P_U[x_i]$

for some small $\epsilon > 0$. To give an ϵ-differential-privacy guarantee, existing methods would perturb the data by Laplacian noise if the inequalities are not already satisfied. However, our "attributes" are words in user posts that the user intentionally chose and our goal is to quantify risk rather than perturb the data. We thus aim to determine the best possible value of ϵ for which the guarantee holds without perturbation. This is the minimum ϵ for each x_i, but the guarantee is only as strong as the weakest x_i, leading to the following formulation:

Definition 2 (Differential-privacy baseline risk score of topic X for U_0)**.** *The differential-privacy baseline risk score of the user U_0 with respect to a topic X is:*

$$risk_{\text{D-P}}(U_0, X) = \max_{x_i} \left(\max \left(log \left(\frac{P_U[x_i]}{P_{U^*}[x_i]} \right), log \left(\frac{P_{U^*}[x_i]}{P_U[x_i]} \right) \right) \right).$$

The ranking method based on this definition is being referred to as DIFF-PRIV.

Measure properties. It holds that $risk(U_0, X) \geq 0$. The risk value is lowest for a user who does not have any of the sensitive topic's salient attributes in her contents and highest for a user who has a critical attribute that is not present in the contents of any other user.

3.3 Topical risk measure

To this end, we construct a distributional representation of each of the sensitive topics X (e.g., financial debts), user contents U (e.g., from an online community such as `quora.com`), and each post P the user authors in the online community. We model X, P and U as vectors in a distributional vector space.

3.3.1 Distributional vectors for topics and users

Topic vectors. Topics are represented as vocabulary distributions found by collecting word statistics over suitably chosen corpora.

Definition 3 (Sensitive Topic Vector)**.** *For sensitive topic X, the topic vector is a distributional vector $\vec{X}$ constructed using words or bigrams weighted by topic relevance.*

For example, *hiv* and *positive* are salient for the topic of hiv infection. Such topics and their salient phrases can be automatically extracted by applying latent topic analysis to large, thematically broad text corpora.

User vectors. To be able to relate posts and users to topics, we map each user U and post P created by the user in an online community to a vector.

Definition 4 (User Post and User Vectors)**.** *The content of a post P of a user U is modeled as a distributional vector $\vec{P}$. User U in the context of a topic X is modeled as a distributional vector $\vec{U}$ defined as:*

$$\vec{U} = \max_{P \in U} cos(\vec{P}, \vec{X}).$$

Vector construction. The exact mapping of topics and posts to vectors depends on the vector space in which we are operating. We use three different configurations in our experiments: i) a bag-of-words model (BOW), ii) an LDA model (LDA), and iii) a Skip-gram model (W2V).

Note that the use of LDA here is to construct a lower-dimensional vector space; this is orthogonal to using LDA for obtaining topics with their salient phrases, which we discussed above.

In the BOW vector space, we create topic vectors directly over the characteristic topic words with binary scoring; we also use these words as features with tf-scoring for user and post vectors.

In the LDA model, topic vectors are indicator vectors of for the latent dimensions. Users and posts are treated as documents that LDA maps into its low-dimensional latent space.

The third technique that we consider, `w2v`, is based on a neural network for word relatedness, which can be trained over large text corpora [25]. To create the topic vectors in this word-centric vector space, we compute a weighted sum of words from the previously computed sensitive topic distributions. Since there is no natural mapping of documents to vectors in this setting, the procedure for posts is similar. However, to discount the impact of words unrelated to the topics at hand, we introduce a topic-dependent weighting scheme for user vectors. Namely, for a topic X and a post containing the set of words $\{v_1, v_2, \dots\}$, the post vector is $\vec{P} = \sum_j cos(\vec{v_j}, \vec{X}) \cdot \vec{v_j}$.

Risk scoring. Given these vectors, we can now compare a user posting history against a sensitive topic by vector-based similarity measures, like the cosine similarity. An advantage of this risk measure is that, unlike the ENTROPY or DIFF-PRIV measures, it does not require any community-level data, as the risk score of a user is independent of other users' data. Thus, each user can compute her score locally and privately, and send the value to a server to obtain an R-Susceptibility rank.

In addition to quantifying the *strength* of user interest in a sensitive topic, we also capture the *breadth* and *temporal variation* of that interest. This is crucial to avoid erroneously ranking higher those users who have a professional interest in a topic without being personally afflicted, or are temporarily interested out of curiosity. In our previous preliminary work in this area, we identified these two components to be crucial for reducing classification error in a similar setup [4].

3.3.2 Strength of Interest

Having a vector representation of a user U, we can now compute the similarity between U and a topic vector X.

Definition 5 (Topic-aware risk score)**.** *The strength-of-interest risk score for a user U with respect to a topic X is*

$$risk(U, X) = \cos\left(\vec{U}, \vec{X}\right).$$

We further refer to methods based on this definition as BOW, LDA, and W2V.

Measure properties. It holds that $-1 \leq risk(U, X) \leq 1$. A high value of this measure means the user has at least one post with vocabulary related to the topic. Thus, the strength of interest is reflected by the presence of the topic's salient vocabulary in user posts.

3.3.3 Breadth of Interest

When ranking users, an adversary might want to distinguish between users who show a focused interest in a topic and users who show a broad interest in many topics within a domain, ranking the former higher than the latter. Applying this strategy could help, for instance, to capture users who are not personally afflicted but rather showing educational, hobbyist or professional interest in a topic. For example, for the topic of financial debts, a bank agent or finance hobbyist could offer advice in Q&A communities; similarly, a medical doctor or student could engage herself in health forums.

The posts of a user with a broad interest should exhibit a diversity of topics within their respective domain. We aim to capture this behavior, by means of distributional vectors, assigning each topic X to a broader domain, like finance, medicine, psychology, etc.

Definition 6 (Domain Vectors)**.** *A domain D is a set of topics $X_1, ..., X_{|D|}$ and its vector representation is a set of corresponding topic vectors $(\vec{X}_1, ..., \vec{X}_{|D|})$.*

To assess the risk taking into account whether a user U has a focused or a broad interest in a topic X, we compute:

1. how similar $\vec{U}$ is to $\vec{X}$ and

2. how dissimilar $\vec{U}$ is to the domain D by computing the distances between $\vec{U}$ and $\vec{X}_j$ for $j = 1..|D|$ and taking the $\lceil k * |D| \rceil$-th largest value, for some $0 < k \leq 1$.

If both of these measures are high, then we conclude that U is personally afflicted by topic X.

Definition 7 (Domain-aware risk score)**.** *The domain-aware risk score for a user U with respect to a topic X from the domain D is*

$$risk_{\mathrm{D}}(U, X) = \cos\left(\vec{U}, \vec{X}\right) - max_{\lceil k*|D| \rceil}\{\cos\left(\vec{U}, \vec{X}_j\right)|\}_{j=1..|D|}.$$

We further refer to methods based on this definition as BOW-D, LDA-D, and W2V-D.

Measure properties. It holds that $-2 \leq risk(U, X) \leq 2$. The value would be high for a user who has a post containing topic's salient vocabulary, but whose contents do not exhibit any vocabulary from other topics in the respective domain. A low value occurs in a situation where the user has not written any posts related to the topic at hand, but has contents related to other topics in the domain. Studying the relative importance of the two components in different online communities is an interesting topic of future work.

The intuition for parameter k is that a personally afflicted user would not have high posting activities in k-fraction of different topics within the same domain. The value of the parameter controls how large the domain coverage should be for the users to be considered broadly interested. In practice, setting this parameter requires the knowledge of the breadth of topics discussed in a particular community.

3.3.4 Temporal Variation of Interest

Being interested in users most likely afflicted by a given state, we would like to rank the users who exhibit recurring activity regarding a topic X higher than the non-afflicted (possibly curious or exploratory) users exhibiting a short-term interest in the topic. Such a bursty activity might be prompted by prominent news related to X, be it sex scandals in the press, or social campaigns about depression.

To capture this issue, rather than computing a user vector $\vec{U}$ over the entire user history, we divide the history into time buckets and compute a sequence of vectors $\vec{U}_i$ using the contents from each bucket i separately. In our model, bucketization may be realized at different granularity levels depending on the user observation period and the characteristics of the community.

We then identify the top-m time buckets with the highest risk level, representing m different time periods (such as days or weeks). Let us denote these buckets of the user model as $U_1^*, \ldots, U_m^*$. A user whose interest in X is clearly above the level of a bursty interest (signifying occasional curiosity) would consistently have high risk scores in all of the top-m buckets. This leads us to our next definition of a user's privacy risk regarding topic X.

Definition 8 (Time-aware risk score)**.** *The time-aware risk score for a user U in time period i with respect to a topic X is*

$$risk_{\mathrm{T}}(U, X) = \mathrm{avg}_{i=1..m}\left\{\cos\left(\vec{U_i^*}, \vec{X}\right)\right\}.$$

We further refer to methods based on this definition as BOW-T, LDA-T, and W2V-T.

Measure properties. It holds that $-1 \leq risk(U, X) \leq 1$. The value would be high for a user whose posts contain relevant topic vocabulary in at least m observation buckets, and low for a user who does not exhibit topic's vocabulary in their contents.

The choice of a particular value of the m parameter depends on the available observation timeline and the characteristics of a given community. The parameter controls how often the activity regarding a topic should occur in order to not be considered occasional.

3.3.5 Combining Domain- and Time-Awareness

The final measure we introduce combines all the aforementioned dimensions of interest. Note that we use bucketized user contents for computing the temporal variation component, but the breadth-of-interest component is computed over the full contents.

Definition 9 (Domain- and time-aware risk score)**.** *The risk of user U in time period i for topic X in domain D is*

$$risk_{\mathrm{DT}}(U, X) = \mathrm{avg}_{i=1..m}\left\{\cos\left(\vec{U_i^*}, \vec{X}\right)\right\} - \cos\left(\vec{U}, \left(\vec{D} - \vec{X}\right)\right).$$

We further refer to methods based on this definition as BOW-DT, LDA-DT, and W2V-DT.

4. DISCOVERING SENSITIVE TOPICS

To complete our framework, we need to train a background knowledge model. The remaining question is then how to identify sensitive topics. Although our model is applicable to any topic irrespective of its sensitivity, in practice users would only be interested in their R-Susceptibility ranks for truly sensitive topics. There is indeed a systematic way of gathering such information in a reasonably inter-subjective manner: training a latent topic model on a background corpus and crowdsourcing sensitivity judgments for each topic. This section presents our results along these lines.

4.1 Experiments on topic sensitivity

Datasets. We trained 3 LDA models, using the Mallet topic modeling toolkit: i) with 500 topics, on 600K Quora posts we crawled ii) with 200 topics, on 3M posts from health Q&A online forums, and iii) with 500 topics, on a sample of 700K articles from the New York Times (NYT) news archive.

Crowdsourcing sensitivity and domain judgements. We collected human judgements regarding the sensitivity and the domains of topics using Amazon Mechanical Turk (AMT), employing only master workers from the USA, and collecting 7 judgements per topic. For each of the topics, the workers were shown the 20 most salient words computed by LDA, and asked whether they would consider a post in social media containing these words privacy-sensitive. We explained that by privacy-sensitive we mean that a person uses these words because he/she is in a privacy-sensitive situation (e.g, alcohol addicted), or that the usage of these words might lead to a privacy-sensitive situation (e.g., political extremism). The first condition can capture, for instance, words related to diseases, the second can capture words related to political or religious positions.

We computed Fleiss' Kappa to measure the inter-annotator agreement for this task, obtaining 0.241 for the Quora topics, 0.294 for the HF topics, and 0.157 for the NYT topics. These low values confirm that sensitivity is rather subjective. However, there is a considerable number of topics in all of these corpora, which were unanimously or almost unanimously rated as sensitive. These were mostly related to

Table 1: #topics with #judges agreeing on being sensitive topics.

#judges	#topics Quora	#topics NYT	#topics HF
7	29	8	38
6	43	27	32
5	48	60	30
4	56	84	21
3	68	73	22
2	99	90	28
1	106	111	23
0	51	47	6

Table 2: Examples: vocabulary of sensitive topics.

Topic	Vocabulary
clinical depression	depression depress suicide feel depressed suffer suicidal commit
drug addiction	drug addiction addict cocaine heroin substance meth addictive
pregnancy	baby birth pregnancy pregnant mother woman born child
hiv and viral diseases	hiv disease aids virus spread infection cure vaccine
financial debts	debt loan pay student interest payment money owe

health, private relationships, political and religious convictions, personal finance, legal problems and others. Table 1 shows the numbers of topics on which certain numbers of judges agree on their sensitivity.

The judges were also asked to assign a topic to one of seven high-level categories. Six of these, potentially containing some sensitive topics, were chosen based on the top-level Microsoft Academic Search categories. The annotators could also choose a generic category *other*.

Topics for evaluation in Section 5. For our further experiments, to make the much more laborious and costly evaluation of user profiles feasible, we leverage the above study to restrict the evaluation to 5 topics from the group of the most sensitive topics. The choice of particular topics is guided by the reported cases of social media screening by insurance companies, employers, and credit companies mentioned in Section 1. These are: *clinical depression*, *drug addiction*, *hiv*, *pregnancy*, and *financial debts*, assigned to the domains of *psychology*, *medicine*, and *finance&economy*. Table 2 shows the most prominent words for each of the chosen topics from the Quora topic model.

5. EXPERIMENTAL EVALUATION

5.1 Setup

5.1.1 Data sources

To test our methods in a variety of scenarios, we constructed three datasets using online communitites of different nature. As a first data source, we used the AOL query log collected between March and May 2006. The resulting data source amounts to around $107K$ users and more than $13M$ queries. The second data source consisted of over 5M posts spanning 13 years (2000-2013) from `healthboards` and `ehealthforum` Q&A health communities. We also collected data from the Quora Q&A community over a period of three months, between February and May 2015. The crawl focused on Quora users who were active in categories related to the considered sensitive topics and their domains, and comprised more than 200K users and 1.3M posts.

Ethics. To adhere to ethical standards concerning incorporation of user data into research, we decided to only use data that is publicly available – either as online profiles (Quora, Health Q&A), or as datasets used in numerous other studies (AOL). We never attempted to identify the individuals whose profiles we analyzed.

5.1.2 User sampling

We created our datasets by sampling the users from the data sources described above. However, we encounter a technical challenge, as the number of sensitive users for a given topic is very small when compared to the size of the whole community. Sampling users uniformly would not constitute a good benchmark for risk scoring methods. For example, we could achieve high accuracy, in a misleading way, by the simplistic prediction that all users are non-sensitive.

What we want, though, is ranking evaluation – our goal is to see how sensitive the users are in different ranking regions. Therefore, our sampling method is non-uniform and proceeds as follows. We first rank all users for each of the datasets using our basic strength-of-interest method from Section 3.3.2, and then sample users from this ranking. To pick a user, the sampling procedure orders users by their score, then computes prefix sums Σ_i for all users up to user i, with Σ_n being the score sum for all users. Then we draw a random number between 0 and Σ_n. If the number falls between Σ_i and Σ_{i+1}, we choose user $i+1$ (with users numbered from 1 to n. However, given that risk scores are extremely skewed, this sampling does still not yield good coverage of all the ranking regions. Therefore, we transform the original risk score q into a^q, where constant a needs to be determined based on the score skew in a data source. The intuition is to give a higher probability of being sampled to users with higher scores, so that the final sample set has good coverage of users with both high and low scores. Figure 1 depicts the depression risk scores of our 100 samples from the AOL data vs. the scores of the original dataset of 170K users.

In our case, a value of $a = 10^2$ for the Health Q&A, and a value of $a = 10^3$ for the AOL were reasonable to compensate the skew. For each of these datasets, we sampled 100 users for each sensitive topic. We did not perform this kind of sampling for Quora, as our dataset was based on a focused crawl in the first place with focus on sensitive discussion threads. Since evaluating sizeable Quora profiles requires much more effort, for this data source we constructed smaller datasets with 40 users per topic. In total, our datasets comprised 1100 user profiles: personal histories of posts or queries.

5.1.3 User study for ground-truth labels

To assign sensitivity labels for user-topic pairs as ground truth, we used crowdsourcing and asked human judges to examine user profiles with chronologically ordered textual posts. Specifically, we asked whether based on the content

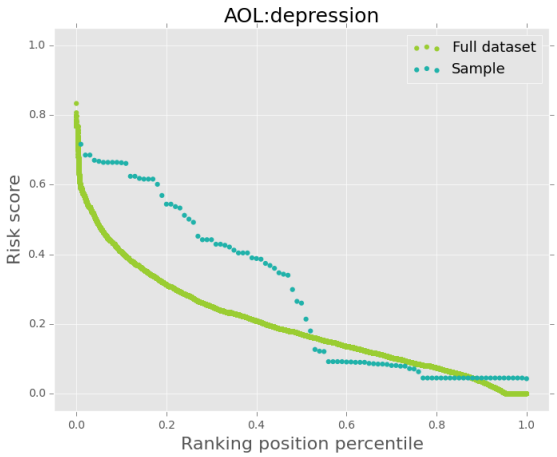

Figure 1: Example comparison of risk scores of sample vs. full data.

Table 3: Number of sensitive users according to judges' assessments (x/y means x sensitive users out of y in total).

	AOL	HF	Quora
depression	24	42	20
drugs	22	31	11
pregnancy	15	42	21
hiv	14	19	5
debts	24	n/a	11
TOTAL	99/500	134/400	68/200

of the profile, the judge suspects that the user (or a family member) is in a given sensitive state.

To evaluate the AOL and Health Q&A datasets, we employed AMT master workers from the USA and collected 5 judgements for each of the profiles. Since the majority of Quora profiles contain hundreds of posts, to ensure that proper care is given to evaluating them, we collected the judgements employing 19 students from our institution. We computed Fleiss' Kappa to quantify the global inter-annotator agreement across all the topics. The respective values for the AOL, Health Forums and Quora datasets were 0.442, 0.444, and 0.468 respectively, all corresponding to a moderate agreement. Table 3 shows the number of users who were marked by the human judges as sensitive by a majority vote.

5.1.4 *Configuration of methods*

To evaluate the topic-model-based method, we used three different distributional vector spaces: a bag-of-word vector space, as well as two 500-dimensional vector spaces trained with (i) the LDA implementation from the Mallet toolkit and (ii) `word2vec`[1] tool [25]. The latter two models were trained on NYT and Quora corpora described in Section 4.

In the breadth-of-interest model from Section 3.3.3, we set the parameter k to 0.3, i.e. we want a user of a broad interest in a domain to have at least a 30% coverage of topics from the domain.

[1] `https://code.google.com/p/word2vec/`

In the temporal-variance-of-interest models described in Section 3.3.4, we compute the results using weekly time buckets and set the number of buckets parameter m to 3.

We later analyze the robustness of the ranking methods with respect to these parameters.

5.1.5 *Effectiveness metrics*

We assess the effectiveness of risk ranking using the following metrics.

- **R-Precision.** For a given sensitive topic, where r users were identified by the judges as sensitive, we compute the precision@r. When computing the average precision over all sensitive topics, we report both micro and macro average scores (summing over individual samples, and summing over topic precisions, respectively). To apply this measure, for each of the profiles we have to cast the five collected judgements to a binary score. We assume that an average of more than 0.5 classifies a user as sensitive. Note that r-precision imitates an adversary who, for instance, knowing that 1% of the population is depressed, ranks the users according to a depression-risk measure and chooses the top 1% of the users for further investigation.
- **Mean Average Precision (MAP).** For a given sensitive topic, we compute the average precision computed at the ranking positions of sensitive users.
- **Normalized Discounted Cumulative Gain (NDCG).** To asses the effectiveness of our methods using the actual non-binary judge assessments, we employ NDCG, which compares the rankings our methods yield with a perfect ranking obtained using the crowdsourced scores.

5.1.6 *Significance testing*

The number of topics in our experiments is too small to perform significance tests over macro-averaged metrics. We thus resort to performing a paired t-test over r-precision differences on individual test samples within a dataset, marking the significance in the micro-r-precision columns in the result tables. The $*$ symbols denote the case when the gain of a given ranking method over the strength-of-interest baseline is statistically significant with a p-value < 0.05. This lets us conclude that a good r-precision score of a ranking method does not likely depend on the particular choice of user profiles.

5.1.7 *Research questions*

The remainder of the experimental section seeks to answer the following research questions.

RQ 1: Do the proposed topical risk measures perform better than the ENTROPY and DIFF-PRIV methods in predicting human risk judgements? (Sec. 5.2.)

RQ 2: Does the topical risk scoring measure perform better when extended with the breadth and temporal dimensions of user interest? (Sec. 5.3.)

RQ 3: How robust is the proposed method against changes in the parameter configuration and the background knowledge of the adversary? (Sec. 5.4.)

5.2 Traditional vs. IR risk scoring

We begin the risk scoring methods analysis by comparing the effectiveness of traditional (ENTROPY, DIFF-PRIV) and

Table 4: Average metrics over all sensitive topics for different risk assessment measures

	R-precision		Prec@5	MAP	NDCG
	micro	macro			
AOL					
ENTROPY	0.455	0.428	**0.667**	0.465	0.766
DIFF-PRIV	0.445	0.418	0.433	0.441	0.743
LDA	**0.525**	**0.518**	0.600	**0.557**	**0.796**
Health Forums					
ENTROPY	0.560	0.537	**0.750**	0.613	0.870
DIFF-PRIV	0.560	0.559	0.500	0.542	0.794
LDA	**0.649***	**0.636**	**0.750**	**0.724**	**0.913**
Quora					
ENTROPY	0.244	0.216	0.267	0.317	0.630
DIFF-PRIV	0.256	0.246	0.233	0.311	0.614
LDA	**0.358***	**0.362**	**0.440**	**0.428**	**0.715**

the baseline strength-of-interest topical risk scoring methods. Here, we choose the baseline IR-based methods, while exteding the measures with dimensions of interest will be addressed in the sections to follow.

Table 4 shows that the LDA risk scoring outperforms the alternatives (similar observation holds for W2V), which confirms that these methods, designed to measure privacy in structured datasets, are not naturally applicable to textual data. The relatively good Precision@5 of these measures indicates that the most sensitive users tend to use highly salient words. However, operating on explicitly given salient attributes for each topic, the baseline measures do not capture any lexical correlations, an important prerequisite to capture users manifesting their sensitivity in a less direct way. This result validates the need to design new privacy risk measures tuned to textual contents and open web settings.

5.3 Risk scoring with dimensions of interest

We posited that extending the topic-model-based risk measures with breadth and time-variation dimensions of interest can help to predict sensitivity judgements better. Table 5 shows the evaluation results averaged over all topics, confirming that incorporating breadth and temporal variation into the risk score indeed improves the ranking performance.

We observe that breadth-of-interest is especially important for Quora, which is a Q&A community with a very wide variety of topics. Many Quora users seem to frequently post replies prompted by others rather than by their personal situation; hence the lower impact of the temporal component. Contrary, in AOL the temporal component takes over. With merely implicit cues in the form of queries, the temporal dimension is an important indicator of user sensitivity (also for the annotators). The breadth-of-interest component performs worse for AOL, possibly due to the short time span of the query log (3 months).

Note that in case of the proposed breadth-of-interest score, an underlying assumption is that an adversary is able to assign latent topics to broader thematic domains. Thus the best performing -DT methods imply a stronger background knowledge of an adversary.

Risk scoring for different topics.

Table 6 shows the values of r-precisions split by the topic, for different variants of LDA-based risk scoring. The trends observed in the results averaged over all topics can be seen here as well - there are consistent improvements across topics when incorporating the temporal and breadth dimensions. These results constitute anecdotal evidence that the proposed methods are general enough to be potentially applied to a variety of topics.

5.4 Robustness to configuration changes

Model changes. The BOW vector space models only an adversarial knowledge of salient words for different topics, whereas the latent vector spaces additionally enable an adversary to compute similarities between arbitrary words. The results show that this has a direct consequence in the risk ranking performance. The methods with the latent models as the background knowledge outperform the methods with the BOW background knowledge, while being comparable with each other. Thus the model seems resilient to rational background knowledge model changes, capturing a wide class of adversaries - the rational, cost-aware adversaries adopting latent models.

Training corpus changes. The results presented in the experimental section were obtained using the Quora topic model as the background knowledge model. We ran additional experiments using the NYT topic model described in section 4.1, and noticed that for topics which were captured in the other latent models, we observe similar trends and dependencies in the results. This would suggest that an adversary has the freedom to choose among the inputs where his topics of interest are well captured.

Parameter changes in risk measures. The topical risk measures introduce two parameters: k for coverage of domain topics, and m for the number of (weekly) time buckets. Observing the values of r-precision and NDCG obtained when varying these parameters between $k = \{0.1, 0.2, ..., 1.0\}$, and $m = \{1, 2, ..., 12\}$, yields the following observations. First, when the parameters are set to values from the lower half of the ranges, we still observe improvements over the baseline strength-of-interest measure. Second, when the parameters are set to higher values, the results tend to deteriorate, possibly due to the incompleteness of user profiles in our datasets. Third, we observe higher sensitivity to parameters when a given dimension of interest is important for a given community (e.g. temporal for AOL, breadth for Quora). This result suggests that there is room for improvement within the framework of R-Susceptibility in that community-specific risk measures could be employed.

5.5 Discussion

The presented experimental results suggest that R-Susceptibility with appropriate risk measures is able to identify sensitive users with reasonable accuracy. The topical risk measures that quantify a user's exposure with respect to different topics work well, especially when the domain- and time-awareness components are included.

The R-Susceptibility framework allows the plugging of different risk measures, and in the future more advanced measures could be studied to address some of the limitations of this work. These could, for instance, model semi-experts, subtle vocabulary correlations, user contexts, or specific characteristics of a community.

Table 5: Results averaged over all sensitive topics.

	AOL				Health Forums				Quora			
	R-prec		MAP	NDCG	R-prec		MAP	NDCG	R-prec		MAP	NDCG
	micro	macro			micro	macro			micro	macro		
BOW	0.434	0.420	0.459	0.759	**0.642**	**0.625**	0.620	0.833	0.284	0.262	0.319	0.605
BOW-D	0.364	0.358	0.394	0.700	0.612	0.580	0.610	0.832	0.418*	0.398	0.400	0.672
BOW-T	**0.556***	**0.546**	**0.574**	**0.843**	0.582	0.574	0.619	0.873	0.284	0.285	0.317	0.605
BOW-DT	0.374	0.372	0.441	0.758	0.612	0.597	**0.667**	**0.894**	**0.463***	**0.444**	**0.440**	**0.688**
W2V	0.556	0.533	0.589	0.836	0.664	0.634	0.696	0.894	0.343	0.341	0.352	0.637
W2V-D	0.414	0.395	0.427	0.738	0.642	0.600	0.647	0.874	**0.493***	**0.465**	**0.532**	**0.776**
W2V-T	**0.586**	**0.580**	**0.645**	0.859	0.619	0.616	0.643	0.884	0.313	0.312	0.401	0.695
W2V-DT	0.545	0.530	0.601	**0.860**	**0.687**	**0.678**	**0.768**	**0.939**	0.463*	0.434	0.489	0.763
LDA	0.525	0.518	0.557	0.796	0.649	0.636	0.724	0.913	0.358	0.362	0.428	0.715
LDA-D	0.566	0.563	0.557	0.803	**0.716***	0.703	0.772	0.921	**0.493***	**0.485**	**0.489**	**0.752**
LDA-T	0.576	0.567	**0.655**	**0.879**	0.709	0.704	0.748	0.925	0.299	0.264	0.378	0.669
LDA-DT	**0.616***	**0.616**	0.649	0.859	**0.716***	**0.709**	**0.825**	**0.957**	0.418	0.389	0.481	0.751

Table 6: Comparison of r-precision of LDA, LDA-D, LDA-T and LDA-DT for different topics.

	AOL				Health Forums				Quora			
	LDA	LDA-D	LDA-T	LDA-DT	LDA	LDA-D	LDA-T	LDA-DT	LDA	LDA-D	LDA-T	LDA-DT
depression	0.542	0.542	0.625	**0.667**	0.762	0.762	**0.833**	0.762	**0.650**	0.550	**0.650**	0.600
drugs	0.545	**0.636**	0.591	**0.636**	0.710	**0.806**	0.677	0.774	0.364	**0.545**	0.273	0.364
pregnancy	0.533	0.667	0.533	**0.733**	0.571	**0.667**	0.619	**0.667**	0.095	**0.429**	0.095	0.381
hiv	0.429	0.429	**0.500**	**0.500**	0.526	0.579	**0.684**	**0.632**	**0.400**	**0.400**	0.200	**0.400**
debts	0.542	0.542	**0.583**	0.542	n/a	n/a	n/a	n/a	0.300	**0.500**	0.100	0.200

5.5.1 *User guidance*

The R-susceptibility model and risk measures can work on a user history in a streaming manner, considering all contents up to a given point and periodically or continuously repeating the risk assessment. These methods could also be embedded in a privacy advisor tool that would help users assess their privacy risk, raising an alert when they become too exposed with regard to a sensitive topic.

6. RELATED WORK

Data-centric privacy. Methods for privacy-preserving data publising [13] aim at preventing the disclosure of individuals' sensitive attribute values, while maintaining data utility, e.g., for data mining [3], using concepts like k-anonymity [33], l-diversity [24], t-closeness [22], and membership privacy [23]. All these models are geared for and limited to dealing with structured data, and this holds also for the most powerful and versatile privacy model, differential privacy [10]. In the field of Private Information Retrieval the goal of retrieving data from a database without revealing the query is mainly addressed by query encryption/obfuscation [36]. Generating dummy queries to obscure user activity is another intensively studied technique (e.g., [29]).

Sensitivity prediction. There is little research on characterizing what consitutes a sensitive topic. The recent work of [30] analyzed features of posts and user behavior in Quora, and developed a classifier that can predict the sensitivity of individual posts. However, the solution is largely based on explicit categories (rather than latent embeddings) and the "go anonymous" posting option that users may choose. In contrast, our work aims to understand the sensitivity of any latently represented topic, and provide assessment for risk understood as topical exposure in a community.

Query log sanitization. This line of work tackles the challenge of an adversary using session information to infer user identities from queries [2]. A variety of techniques have been proposed for anonymizing query logs, e.g., hashing tokens, removing identifiers, deleting infrequent queries, shortening sessions, and more [8, 11, 16, 20, 21]. [15] compared different methods for publishing frequent keywords, queries and clicks, and showed that most methods are vulnerable to information leakage.

User-centric privacy. Stochastic privacy [32] is one of the few works that focus on users rather than data. This model introduces a user-defined threshold for sharing data to be obeyed by the platform provider. Closest in spirit to our approach is [4], which uses probabilistic graphical models to infer sensitive user properties, but is very limited in scope.

Linkability and de-anonymization. Privacy research for social networks has demonstrated the feasibility of linking user profiles across different communities [14] and de-anonymizing users [26, 27, 37]. To prevent such attacks, a family of methods (e.g., [34]) eliminates joinable attributes from published datasets.

User behavior modeling. It has been shown that search queries can often be used to predict identity of users, as well as their gender, location, and other demographic attributes [18, 17, 35]. Such information can be harnessed for personalization but may also incur privacy threats. [31] analyzed Twitter profiles and network information to predict the political affiliation and race of users.

Expertise identification and trust analysis. Expert and trustworthy users can be identified based on their questions/answers contents and community votes [1] or by analyzing user interaction graphs [19, 38]. Unlike in these works, our aim is not to identify experts, but to push the users who have a broad interest in a domain down the privacy risk ranking.

7. CONCLUSION

This paper proposes a framework for quantifying privacy risks from textual contents of user profiles in online communities. By employing IR techniques such as ranking and latent topic models, it specifically addresses the risk of exposure with respect to sensitive topics and targeting by a rational adversary with rich background knowledge about topic vocabulary and word correlations.

Although more large scale studies of adversarial risk scoring strategies are needed, our experiments constitute a proof of concept that the approach is a viable basis for privacy risk assessment for users who want to post about sensitive topics but would like to be warned when the risk of being targeted becomes high.

In the future, R-Susceptibility can be extended to incorporate other forms of online activities, and be integrated in a framework for risk mitigation through appropriately guided user actions. Our vision is a trusted personal privacy advisor which assesses risks, alerts the user when critical situations arise, and guides her in appropriate countermeasures.

8. REFERENCES

[1] L. A. Adamic et al. Knowledge sharing and yahoo answers: Everyone knows something. *WWW'08.*

[2] E. Adar. User 4xxxxx9: anonymizing query logs. *Workshop on Query Log Analysis, WWW'07.*

[3] E. Bertino, D. Lin, W. Jiang: A Survey of Quantification of Privacy Preserving Data Mining Algorithms. In: *Privacy-Preserving Data Mining*, Springer 2008.

[4] J. Biega, I. Mele, G. Weikum. Probabilistic Prediction of Privacy Risks in User Search Histories. *Workshop on Privacy and Security of Big Data, CIKM'14.*

[5] D. Blei, A. Ng, M. Jordan. Latent Dirichlet Allocation. *JMLR'03.*

[6] C. Carpineto, G. Romano. Kθ-affinity privacy: Releasing infrequent query refinements safely. Information Processing & Management '15.

[7] Chicago Tribune Online: Social media activity might affect your credit score. `www.chicagotribune.com/lifestyles/sns-201511020000--tms--kidmoneyctnsr-cc20151102-20151102-story.html`, accessed on 2016/01/21.

[8] A. Cooper. A survey of query log privacy-enhancing techniques from a policy perspective. *TWeb'08.*

[9] W. Day, N. Li. Differentially Private Publishing of High-dimensional Data Using Sensitivity Control. *ASIACCS'15.*

[10] C. Dwork. Differential Privacy: A Survey of Results. *TAMC'08.*

[11] L. Fan et al. Monitoring Web Browsing Behavior with Differential Privacy. *WWW'14.*

[12] Forbes Online: How Social Media Can Help (or Hurt) Your Job Search. `www.forbes.com/sites/jacquelynsmith/2013/04/16/how-social-media-can-help-or-hurt-your-job-search/`, accessed on 2016/01/21.

[13] B. Fung et al. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv. 42(4), 2010.*

[14] O. Goga et al. Exploiting Innocuous Activity for Correlating Users Across Sites. *WWW'13.*

[15] M. Gotz et al. Publishing search logs – a comparative study of privacy guarantees. *TKDE'12.*

[16] Y. Hong et al. Differentially Private Search Log Sanitization with Optimal Output Utility. *EDBT'12.*

[17] J. Hu et al. Demographic prediction based on user's browsing behavior. *WWW'07.*

[18] R. Jones et al. "i know what you did last summer": Query logs and user privacy. *CIKM'07.*

[19] P. Jurczyk, E. Agichtein. Discovering authorities in question answer communities by using link analysis. *CIKM'07.*

[20] A. Korolova et al. Releasing search queries and clicks privately. *WWW'09.*

[21] R. Kumar et al. On anonymizing query logs via token-based hashing. *WWW'07.*

[22] N. Li, T. Li, S. Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. *ICDE'07.*

[23] N. Li et al. Membership privacy: a unifying framework for privacy definitions. *CCS'13.*

[24] A. Machanavajjhala et al. L-diversity: Privacy beyond k-anonymity. *TKDD'07.*

[25] T. Mikolov et al. Distributed Representations of Words and Phrases and their Compositionality. *NIPS'13.*

[26] A. Narayanan et al. On the Feasibility of Internet-Scale Author Identification. *SP'12.*

[27] A. Narayanan, V. Shmatikov. De-anonymizing Social Networks. *SP'09.*

[28] G. Navarro-Arribas et al. User k-anonymity for privacy preserving data mining of query logs. *Information Processing & Management '12.*

[29] H. Pang, X. Xiao, J. Shen. Obfuscating the Topical Intention in Enterprise Text Search. *ICDE'12.*

[30] S. T. Peddinti et al. Cloak and swagger: Understanding data sensitivity through the lens of user anonymity *IEEE S&P '14.*

[31] M. Pennacchiotti, A.-M. Popescu. A machine learning approach to twitter user classification. *ICWSM'11.*

[32] A. Singla et al. Stochastic Privacy. *AAAI'14.*

[33] L. Sweeney. k-Anonymity: A Model for Protecting Privacy. *J. of Uncertainty, Fuzziness and Knowledge-Based Systems '02.*

[34] D. Vatsalan, P. Christen, V. Verykios. A taxonomy of privacy-preserving record linkage techniques. *IS'13.*

[35] I. Weber, C. Castillo. The demographics of web search. *SIGIR'10.*

[36] S. Yekhanin. Private Information Retrieval. *CACM'10.*

[37] A. Zhang et al. Privacy Risk in Anonymized Heterogeneous Information Networks. *EDBT'14.*

[38] J. Zhang, M. S. Ackerman, L. Adamic. Expertise networks in online communities: Structure and algorithms. *WWW'07.*