Technical Report

TR-2010-007

Anonymizing User Profiles for Personalized Web Search

by

Yun Zhu, Li Xiong

Mathematics and Computer Science

EMORY UNIVERSITY

Anonymizing User Profiles for Personalized Web Search

Yun Zhu
Department of Math & CS
Emory University
Atlanta, GA
yzhu23@emory.edu

Li Xiong
Department of Math & CS
Emory University
Atlanta, GA
lxiong@emory.edu

ABSTRACT

We study the problem of anonymizing user profiles so that user privacy is sufficiently protected while the anonymized profiles are still effective in enabling personalized web search. We propose a Bayes-optimal privacy based principle to bound the prior and posterior probability of associating a user with an individual term in the anonymized user profile set. We also propose a novel bundling technique that clusters user profiles into groups by taking into account the semantic relationships between the terms while satisfying the privacy constraint. We evaluate our approach through a set of preliminary experiments using real data demonstrating its feasibility and effectiveness.

Categories and Subject Descriptors: H.2.7 [Database Administration]: Security, integrity, and protection; H.3.3. [Information Search and Retrieval]

General Terms: Design, Experimentation, Security

Keywords: Anonymization, privacy-preserving data publishing, personalized search

1. INTRODUCTION

Personalized web search is a promising technique to improve retrieval effectiveness. However, it often relies on personal user profiles which may reveal sensitive personal information if disclosed. For example, a particular person can be identified by analyzing the queries in the AOL query log data released in 2006 [3]. The query terms with sensitive information could then be revealed or associated to that user. In this paper, we study the problem of anonymizing user profiles (represented as a weighted term list) so that user privacy is sufficiently protected while the anonymized profiles are still effective in enabling personalized web search.

Research Challenges and Existing Techniques. There are several research challenges for anonymizing user profiles for personalized web search due to the unique characteristics of user profile data and the requirements of personalization. First, user profiles are represented as transactional data or set-valued data and are highly sparse. Second, user profiles consist of terms that are semantically related to each other. Third, unlike relational data, sensitive items or values are not clearly defined in the set-valued data or transactional data. Finally, the application of personalized web search requires some level of "personal" information which poses utility requirement on the anonymization.

Privacy preserving data publishing or anonymization has been extensively studied in recent years and the techniques can be potentially applied to anonymize user profiles [7]. Most existing anonymization techniques focus on relational data. There are several recent works taking the first step towards anonymizing transactional data. While they could be potentially applied to user profiles, one main limitation is that they either assume a predefined set of sensitive items that need to be protected, which are hard to define in the web context in practice, or only guarantee the anonymity of a user but do not prevent the linking attack between a user and a potentially sensitive item. [23] proposed a technique for building user profiles with configurable levels of details. A few recent works specifically studied anonymizing query logs. Notably, [13, 11] have demonstrated the ineffectiveness or privacy risks of naive anonymization schemes. [12] studied anonymization techniques with differential privacy, however, the utility of the data is limited to statistical information and it is not clear how it can be used for personalized web search.

Contributions. In this paper, we propose a new privacy notion and grouping technique for anonymizing user profiles for personalized web search. The paper makes several contributions. First, it defines a Bayes-optimal privacy based principle for user profiles represented as set-valued data. It does not require predefined quasi-identifying or sensitive terms nor does it require external knowledge database. Rather, it treats every term as potentially sensitive or identifying and bounds the difference between the prior and posterior probability of linking an individual to any term. Second, we propose a novel bundling technique that clusters user profiles into user groups by taking into account the semantic relationships between the terms while satisfying the privacy constraint. Finally, we evaluate our approach through a set of preliminary experiments using real data, showing that our approach effectively enables personalized search with assured privacy.

2. RELATED WORK

Privacy preserving data publishing has received considerable attention in recent years. [7] provides an up-to-date survey. Most work on privacy preserving data publishing has been focused on structured or tabular data. One thread of work on data anonymization aims at devising privacy principles that serve as criteria for judging whether a published dataset provides sufficient privacy protection. Most practical principles consider specific types of attacks (attack specific) and assume the attacker has limited background knowledge (background knowledge sensitive). Notably, k-anonymity [16, 17] prevents identity disclosure (which usually leads to attribute disclosure). l-diversity [15] and t-closeness [14] prevent direct sensitive attribute disclosure. Contrary to most principles that protect against certain attacks and assume limited background knowledge of an attacker, differential privacy [6] is emerging as a strong notion for guaranteeing privacy with arbitrary background knowledge. A large body of work contributes to algorithms that transforms a dataset to meet one of the above privacy principles using techniques such as generalization, suppression (removal), permutation and perturbation [7].

Several works have been proposed recently for set-valued data or transactional data. Notably, Ghinita et al. [8] defined a privacy degree p which bounds the probability of associating any transaction with a particular sensitive item by $1/p$. Xu et al. [22] proposed a privacy notion called $(h, k, p) - coherence$ for transactional dataset that bounds the probability of linking an individual to a transaction by $1/k$ and the probability of linking an individual to a private item by h for an attacker with power p. ERASE [5] is another system proposed for sanitizing document (modeled as a set of terms). It requires an external database of knowledge which links terms to sensitive entities that need to be protected. [18] and the follow-up work [9] do not distinguish sensitive (private) and non-sensitive (public) items and proposed a notion called k^m-anonymity which bounds the probability of linking an individual to a transaction by $1/k$ for an attacker with power m. However, they do not prevent the linking attack between an individual and a potentially sensitive item. Compared to the above, our work has several important features. First, our privacy notion does not need to specify identifying, quasi-identifying or sensitive items nor does it need external knowledge database. Rather, it treats every item as potentially sensitive or identifying and bounds the probability of linking an individual to any additional item (aside what's been known to the attacker) by p. In essence, it generalizes the notion in [8] and [22] that bounds the probability of linking a transaction with a sensitive item. Second, it uses a microaggregation or bundling technique that takes into account the semantic relationships of the items and achieves high utility for personalized web search.

There have been also a few recent works that specifically study anonymizing query logs [1, 21, 12, 10, 20]. Notably, [13, 11] have demonstrated the ineffectiveness or privacy risks of naive anonymization schemes such as token based hashing and simple bundling. [12] provides an anonymization technique with rigorous differential privacy guarantee, however, the utility of the anonymized data is limited. [10] defines an interesting notion, k^δ-anonymity, that addresses the sparsity of the query terms but does not prevent the linking attack between an individual and potentially sensitive queries. [20] applies the interactive differential privacy querying framework PINQ on query logs. While demonstrating its effectiveness for certain queries, it is not clear how the interactive mechanism can be used for personalized web search.

3. APPROACH

Problem Definition. We consider a set of user profiles. Each user profile is represented as a vector of tuples: $UP = \{tw_1, tw_2, \ldots, tw_m\}$, where $tw_i = (term_i, weight_i)$, $term_i$ is a word or phrase representing a user's interest, and $weight_i$ quantifies the extent. Table 1 shows an example profile set with 2 users. Our goal is to cluster the user profiles into user groups so that the privacy of individual users is protected while the user groups are still useful for personalized web search.

Table 1: An example profile set with 4 users

UP_1:	(kitten, 1), (riding, 0.8)
UP_2:	(pup, 0.6), (equitation,1)
UP_3:	({mocha}, 0.7),({java}, 0.6),({coffee}, 0.8)
UP_4:	({programming language}, 0.6), ({java},1), ({C++},0.4)

3.1 Privacy Definitions

An adversary may link a user to a user group based on his background knowledge (e.g. certain terms that the user has searched for) and then identify additional terms in the user group that are contributed by the user. Our anonymization goal is to prevent such linking attacks that associate a user with an individual term in the anonymized user profile set. We adopt the Bayes-optimal privacy notion [15] to bound the difference between the prior and posterior beliefs (before and after access to the anonymized profiles) of linking a user to a term in the user groups. We propose an instantiation of the privacy notion for the grouping approach, called p-linkability.

DEFINITION 1. *(**p-linkability**) A user profile grouping satisfies p-linkability if the probability of linking a user to an individual term in a user group does not exceed p.*

In essence, this protects a user from *attribute disclosure* by preventing the association of a user to potentially sensitive terms (attribute values). Formally, we consider the following attack and derive the bound on the change of prior and posterior beliefs given the requirement of p-linkability. Consider an attacker who possesses certain background knowledge about a victim user profile V and amounts an attack to identify additional terms or interests of the user: V contains certain term t. The background knowledge can be represented as a subset of V, e.g. a set of terms or interests of the user, and is denoted by V_b. Suppose the released anonymized user grouping G contains user groups G_i (i=1 to |G|). From the attacker's point of view, we derive the bound on the change of prior and posterior beliefs of V containing t below followed by explanations.

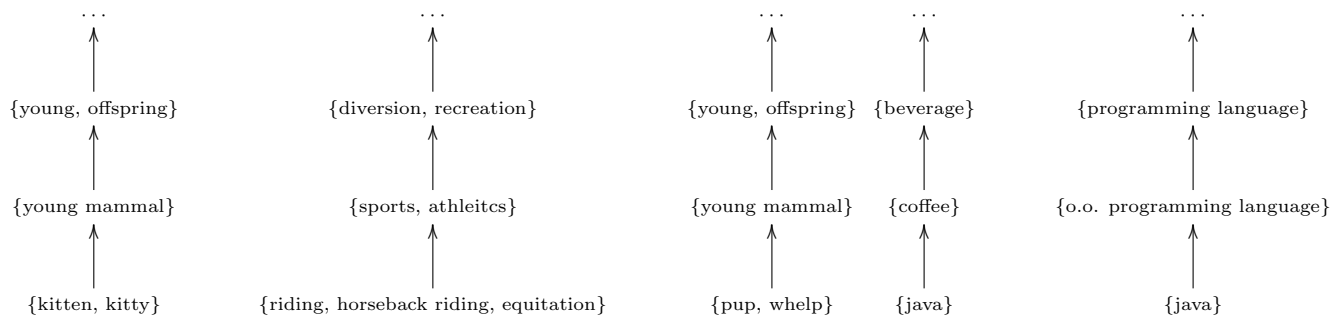

Figure 1: Example hypernym paths

$$\Delta = P(t \in V|V_b, G) - P(t \in V|V_b) \quad (1)$$
$$\leq P(t \in V|V_b, G) \quad (2)$$
$$\leq \sum_{i=1}^{|G|} P(V \subset G_i|V_b, G) * P(t \in V|V \subset G_i) \leq p \quad (3)$$

Given a large domain of terms, without any additional background knowledge besides V_b, we can assume $P(t \in V|V_b)$ is close to 0 which will result in an upper bound on the change of beliefs. The probability of linking a user to an individual group based on the background knowledge, $P(V \subset G_i|V_b, G)$ may vary depending on the overlap of V_b and G_i but they add up to 1. The probability of linking a user to an individual term in a user group, $P(t \in V|V \subset G_i)$, does not exceed p according to the p-linkability requirement. Hence, the change of beliefs is bounded by p.

3.2 User Profile Grouping

We propose a bundling technique that groups the user profiles and the user group representative will be used for personalized search. Since the grouped user profiles will be used for re-ranking the search results, we would like the users within each group to be similar to each other. Concretely, our goal is to perform similarity-based clustering while satisfying the privacy constraint.

Clustering constraint. Given the p-linkability requirement, we need to enforce that the probability of linking a user to an individual term in a user group G_i, $P(t \in V|V \subset G_i)$, does not exceed p. We suppose the average user profile size is $|U|_{avg}$. For simplicity, we assume the terms in each user profile are independent. The probability, $P(t \in V|V \subset G_i)$, can be derived as : $P(t \in V|V \subset G_i) = \frac{|U|_{avg}}{|G_i|}$. For grouping purposes, we have to enforce the constraint: $|G_i| \geq \frac{|U|_{avg}}{p}$.

Semantic similarity between user profiles. While traditional similarity metrics such as cosine similarity can be used to measure the similarity between user profiles, a main challenge is how to address the sparsity of the data and take into account the semantic similarity of two user profiles. For example, a user interested in *riding* and a user interested in *equitation* should be considered similar as the two terms are semantically equivalent. Moreover, a user interested in *riding* and a user interested in *sports* should be similar to some extent as riding is one type of sports.

To address this issue, we propose a user profile augmentation technique using term co-occurrence networks or term hierarchies when computing the similarity between users. The basic idea is to augment each user profile with semantically related terms and then compute similarity based on the augmented profiles using traditional measures such as cosine similarity.

DEFINITION 2. *(**Semantic similarity**) The semantic similarity of two user profiles are the cosine similarity of their augmented user profiles $UP = \{tw_1, tw_2, \ldots, tw_m\}$ and UP' $= \{tw'_1, tw'_2, \ldots, tw'_n\}$ and is computed as:*

$$Sim(UP, UP') = \frac{\sum_{t \in T} UP(t) \cdot UP'(t)}{\sqrt{\sum_{i=1}^{m} t_i^2} \sqrt{\sum_{j=1}^{n} t_j'^2}} \quad (4)$$

where

$$T = \{t_1, t_2, ..., t_m\} \cup \{t'_1, t'_2, ..., t'_n\},$$

$$UP(t) = \begin{cases} w_i & \text{if } \exists (t_i, w_i) \in up,\ t_i = t \\ 0 & \text{otherwise} \end{cases}$$

External term co-occurrence networks or taxonomy trees can be used for augmentation. For this study, we use the WordNet[1], a large online lexical database of English words. We introduce the following definitions before we introduce our augmentation steps.

Table 2: Example synonym sets

Kitten:	{kitten, kitty}
Riding:	{riding, horseback riding, equitation}
Pup:	{pup, whelp}
Equitation:	{riding, horseback riding, equitation}

DEFINITION 3. ***(Synonym set)** A synonym set of a term is a set of words and phrases including the term and all its synonyms.*

For example, {scarlet, vermilion, carmine, crimson} is the synonym set for any term within this set. Table 2 shows the synonym sets for the terms within the example user profiles.

[1] `http://wordnet.princeton.edu`

DEFINITION 4. ***(Hypernym path)*** *In linguistics, a hypernym is a word or phrase whose semantic range includes that of another word, its hyponym. Hypernym path for a synonym set is a list of synonym sets including the root synonym set and all its Hypernym sets.*

For example, scarlet, vermilion, carmine, and crimson are all hyponyms of red (their hypernym), which is, in turn, a hyponym of color. Figure 1 shows the hypernym paths generated from WordNet for all the example synonym sets.

We use the following two augmentation steps based on WordNet.

1. Synonym set replacement. This step will replace every single term in user profiles with its synonym set, a set of words and phrases including the term and all its synonyms.

2. Hypernym set augmentation. This step adds the hypernym sets for all existing synonym sets, a set of words or phrases whose semantic range includes that of another word and its hyponym. We use a parameter a to indicate that only hypernym sets that could be reached within a steps from the current synonym set will be added into the user profile.

An issue could rise in the hypernym set augmentation step if more than one hypernym paths exist for a synonym set. For example, Figure 1 shows two different hypernym paths for the synonym set {java}. Our approach is to compute the similarity between the terms on the candidate paths and the user profile and select the path that has the highest similarity. For example, to augment UP_3 and UP_4 in Table 1, we will select the first path for UP_3 and the other for UP_4.

Table 3: The augmented profiles after synonym replacement (similarity = 0.536)

UP_1:	({kitten, kitty}, 1),({riding, horseback riding, equitation}, 0.8)
UP_2:	({pup, whelp}, 0.6), ({riding, horseback riding, equitation},1)

Table 4: The augmented profiles after hypernym augmentation (similarity = 0.737)

UP_1:	({kitten, kitty}, 1),({riding, horseback riding, equitation}, 0.8)}, ({young mammal},1), ({sport, athletics}, 0.8)
UP_2:	({pup, whelp}, 0.6), ({riding, horseback riding, equitation}, 1), ({young mammal}, 0.6), ({sport, athletics}, 1)

In general, the synonym set replacement relates synonyms in different lexical forms and hypernym set augmentation introduces common hypernym sets for terms in the same semantic categories. Table 3 and Table 4 show the resulting user profile set after synonym replacement and hypernym augmentation. As we can see, both replacement and augmentation increase the similarity value.

Hardness of constrained clustering. Given the privacy constraint, our problem is a constrained clustering problem. With an exhaustive search for an optimal solution involved, most clustering problems are potentially exponential. An optimal k-anonymity by suppression has been proven NP-hard by Aggarwal et al. in [2] and k-member clustering problem NP-complete by Byun et al. in [4].

Greedy algorithm. Based on the hardness of the problem, we use a greedy algorithm. In the beginning, a user profile is randomly selected as the seed of a new cluster. The closest user profile is continuously selected and combined with the seed until the cluster satisfies p-linkability or the size of the cluster $|G_i|$ satisfies the constraint $|G_i| \geq \frac{|U|_{avg}}{p}$. At next step, a user profile with the longest distance to the previous seed is selected as the seed of the new cluster. The process repeats until every user profile is clustered. The last step checks the last cluster, which may not have sufficient user profiles to satisfy the constraint, and assign each user profile to the closest existing clusters. Algorithm 1 presents a sketch of the greedy clustering algorithm.

Algorithm 1 Greedy constrained clustering

```
result ← ∅
C ← ∅
seed ← a randomly picked user profile from S
while |S| > 0 do
  seed ← the furthest user profile(with the min similarity
  value) to seed
  while C does NOT satisfy p-linkability AND |S|>0 do
    add the closest user profile (with the max similarity
    value) to C
  end while
  if C does satisfy p-linkability then
    result ← result ∪ C;
    C ← ∅
  end if
end while
for each user profile in C do
  assign it to the closest cluster
end for
result result
```

Complexity of the algorithm. The algorithm spends most of its time selecting the most similar profile for the current cluster in each iteration. Assume the original user profile set has n profiles and the algorithm generates l clusters which means it has l iterations. Since it searches at most n profiles at each iteration, the overall time complexity is $O(ln)$.

Table 5: Example anonymized user groups

G_1:	(kitten, 1), (riding, 0.8),(pup, 0.6), (equitation,1)
G_2:	({mocha}, 0.7),({java}, 0.6),({coffee}, 0.8), ({programming language}, 0.6), ({java},1), ({C++},0.4)

Cluster representative. The final step of our approach is to compute a representative for all member profiles in each profile cluster. The cluster centroid or union (based on the original user profiles not the augmented ones) is computed and used as the group representative. Table 5 shows an example user grouping result containing the user profiles from Table 1.

4. PRELIMINARY RESULTS

We performed a set of preliminary experiments using real-world data and present the results in this section. Our main goal is to answer the following questions: 1) can the anonymized profiles achieve personalized web search? 2) what is the impact of semantic similarity on search results? 3) what is the impact of privacy level on the search result? 4) what is the actual change of beliefs given a privacy level?

Data and experiment setup. We generated a set of user profiles from the AOL search query log[2]. The log was collected over three months. Each entry of the query log contains user ID, query terms, query time, item rank, clicked URL's. We only extracted the query terms for each user id and they are assigned with the same initial weight value 1.0. In the pre-processing, terms that do not exist in WordNet dictionary are considered as typos and invalid terms and removed. We also discard stop words which are language-words that have no significance meaning in a keyword based search system. We use the stop words set from Lucene[3] which includes words like "a", "an", "and", "are" etc.

We implemented our grouping algorithm for grouping the user profiles. We also implemented a personalized search engine on top of Lucene. We used the TIPSTER Information-Retrieval Text Research Collection[4] as our search corpus. When a result list is returned from Lucene, we re-rank them according to their similarity to the user's profile (both the original and anonymized profiles). Since our focus is to evaluate the effectiveness of anonymization, rather than the personalized search, we use the search result based on the original user profiles as the gold standard and measured the precision of the search results based on anonymized profiles. We use Average Precision [19] as our search quality metric, which is widely used as a measure in web search quality evaluation.

The settings of parameters are as follows. For the experiment comparing personalized search using anonymized profiles and non-personalized search, we used $p = 0.4$ and $a = 1$. For the experiments evaluating the impact of semantic similarity and privacy level, a varies from 0 to 3, and p varies from 0.1 to 0.4. The user profile size is fixed at 150 and number of users is fixed at 720. All experiments assume that the top 30 search results in the re-ranked result list by original user profiles (gold standard) are relevant. All the implementations are in Java and all experiments are run on a PC with 3G CPU and 4G RAM.

Benefit of personalized web search using anonymized profiles. Figure 2 compares the average precision of personalized search using anonymized profiles with the non-personalized search. We observe that the average precision at top 10 relevant search results is 78.86% which indicates that the search using anonymized user groups achieves good precision and provides significant improvement over non-personalized search.

Impact of semantic similarity on search precision.

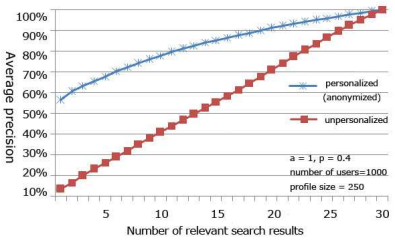

Figure 2: Search precision using anonymized profiles

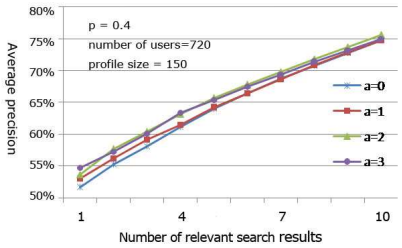

Figure 3: Impact of augmentation level on search precision

Figure 3 shows the impact of the hypernym augmentation level, a, on the search precision. It is verified, to some extent, that the higher a, the more semantic similarity between user profiles are taken into account and thus the better the search quality.

Tradeoff between privacy and search precision. Figure 4 presents the tradeoff between the privacy level, p, and the search precision. A lower value of p means a lower bound of the probability that a specific term could be associated with a user profile and a higher privacy protection. As expected, a lower value p provides stronger privacy guarantee at the cost of search precision.

To illustrate the reason of the impact on search precision, Figure 5 shows the average group size (the average number of profiles in each group) with respect to different p values. The group size decreases with increasing value of p because a lower p requires more profiles grouped together to satisfy the privacy constraint. On other other hand, a larger number of profiles in one group will introduce more noise for every member profile in that group and hence lower the search precision as shown in Figure 4.

Actual privacy. Figure 6 presents an estimate of the change of beliefs of the linking attack by the actual probability of linking a term in a user group to a user based on the grouping result compared to the given bound p. It shows that the actual linkability is indeed close to but lower than the given bound. This is due to the fact that in certain cases it is necessary to group user profiles into groups that over-qualify for the privacy constraint.

5. CONCLUSION AND FUTURE WORK

Personalized web search customizes the search results to improve the search quality for web users. However, user's personal information might be exposed in the user profile

[2] http://gregsadetsky.com/aol-data/
[3] http://lucene.apache.org
[4] http://www.ldc.upenn.edu/Catalog

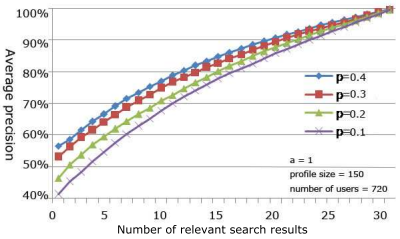

Figure 4: Impact of privacy level on search precision

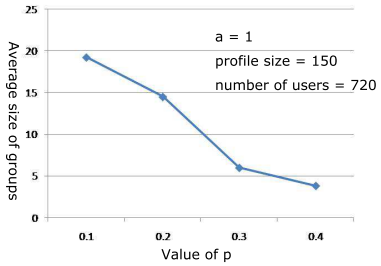

Figure 5: Impact of privacy level on average group size

which is the basis in personalized web search. In this paper, we proposed a grouping approach for anonymizing user profiles with a $p - linkability$ notion to bound the probability of linking a potentially sensitive term to a user by p. We presented a greedy clustering technique with novel semantic similarity metric based on augmented user profiles in order to address the sparsity of user profiles and take into account semantic relationships between user profiles. The experiment results showed that the search precision was raised by using anonymized user profile set compared to the non-personalized search results. The tradeoff between search quality and privacy protection were also presented in our experiment.

While our preliminary results demonstrated the feasibility of the approach, it certainly warrants further research. The current AOL dataset places many limitations for extracting users' specific interests. We plan to explore other options to collect or extract user profiles to further verify our approach. Moreover, we are interested in extending the work with similarity constraint in each group to provide certain utility guarantee. Finally, we are also exploring mechanisms for anonymizing user profiles with differential privacy.

Acknowledgement

The work is partially supported by a Career Enhancement Fellowship by Woodrow Wilson Foundation.

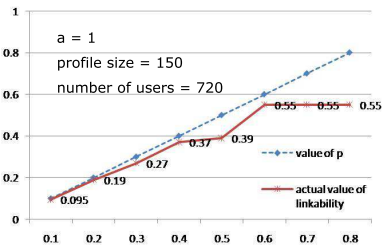

Figure 6: Actual linking probability vs. given privacy level

6. REFERENCES

[1] E. Adar. User 4xxxxx9: Anonymizing query logs. In *Query Log Analysis Workshop at WWW Conference*, 2007.

[2] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu. Achieving anonymity via clustering. In *PODS*, pages 153–162, 2006.

[3] M. Barbaro and T. Zeller. A face is exposed for aol searcher no. 4417749. *New York Times*, August 9 2006.

[4] J.-W. Byun, A. Kamra, E. Bertino, and N. Li. Efficient k-anonymity using clustering technique. Technical report, Purdue University.

[5] V. Chakaravarthy, H. Gupta, P. Roy, and M. Mohania. Efficient techniques for document sanitization. *ACM 17th Conference on Information and Knowledge Management (CIKM)*, 2008.

[6] C. Dwork. Differential privacy: A survey of results. In M. Agrawal, D.-Z. Du, Z. Duan, and A. Li, editors, *TAMC*, volume 4978 of *Lecture Notes in Computer Science*, pages 1–19. Springer, 2008.

[7] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey on recent developments. *ACM Com Surveys*, 42(4), 2010.

[8] G. Ghinita, Y. Tao, and P. Kalnis. On the anonymization of sparse high-dimensional data. In *ICDE '08: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pages 715–724, Washington, DC, USA, 2008. IEEE Computer Society.

[9] Y. He and J. F. Naughton. Anonymization of set-valued data via top-down, local generalization. *PVLDB*, 2(1):934–945, 2009.

[10] Y. Hong, X. He, J. Vaidya, N. R. Adam, and V. Atluri. Effective anonymization of query logs. In *CIKM*, pages 1465–1468, 2009.

[11] R. Jones, R. Kumar, B. Pang, and A. Tomkins. Vanity fair: privacy in querylog bundles. In *CIKM '08*, 2008.

[12] A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas. Releasing search queries and clicks privately. In *WWW '09*, 2009.

[13] R. Kumar, J. Novak, B. Pang, and A. Tomkins. On anonymizing query logs via token-based hashing. In *WWW '07*, 2007.

[14] N. Li and T. Li. t-closeness: Privacy beyond k-anonymity and l-diversity. In *To appear in International Conference on Data Engineering (ICDE)*, 2007.

[15] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*, page 24, 2006.

[16] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Trans. Knowl. Data Eng.*, 13(6):1010–1027, 2001.

[17] L. Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, 2002.

[18] M. Terrovitis, N. Mamoulis, and P. Kalnis. Privacy preserving anonymization of set-valued data. *Proc. of the 34st Very Large DataBases (VLDB)*, 2008.

[19] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *SIGIR '06*, 2006.

[20] E. Viegas and P. A. Kodeswaran. Applying Differential Privacy to Search Queries in a Policy Based Interactive Framework. *ACM Workshop on Privacy and Annonymity for Very Large Datasets*, 2009.

[21] L. Xiong and E. Agichtein. Towards privacy preserving query log publishing. In *Query Log Analysis Workshop at Internatioanl Conference on World Wide Web (WWW)*, 2007.

[22] Y. Xu, K. Wang, A. W.-C. Fu, and P. S. Yu. Anonymizing transaction databases for publication. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2008.

[23] Y. Xu, K. Wang, B. Zhang, and Z. Chen. Privacy-enhancing personalized web search. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, 2007.