IDEAL: An Interactive De-Anonymization Learning System

Na Li, Rajkumar Murugesan, Lin Li, and Hao Zheng
Department of Computer Science
Prairie View A&M University
Prairie View, Texas 77446, USA

***Abstract*—In the era of digital communities, a massive volume of data is created from people's online activities on a daily basis. Such data is sometimes shared with third-parties for commercial benefits, which has caused people's concerns about privacy disclosure. Privacy preserving technologies have been developed to protect people's sensitive information in data publishing. However, due to the availability of data from other sources, e.g., blogging, it is still possible to de-anonymize users even from anonymized data sets. This paper presents the design and implementation of an Interactive De-Anonymization Learning system—IDEAL. The system can help students learn about de-anonymization through engaging hands-on activities, such as tuning different parameters to evaluate their impact on the accuracy of de-anonymization, and observing the affect of data anonymization on de-anonymization. A pilot lab session to evaluate the system was conducted among thirty-five students at Prairie View A&M University and the feedback was very positive.**

***Index Terms*—online social networks, anonymization, de-anonymization, target set, auxiliary set**

I. Introduction

With the fast growth of smart phone apps and online social media sites, a massive volume of data has been collected from users, from scalar data to relationship data and from healthcare data (e.g., DNA gen data) to mobile trace data. However, how the data is used has caused people's privacy concerns.

Researchers have been dedicated to designing advanced technologies for protecting people's privacy in digital communities. Completely ensuring data privacy is quite challenging due to the following reasons: (1) data utility needs to be ensured; (2) too much information has been shared online by people themselves, e.g. on Facebook, which can be used as background knowledge for attacks; and (3) many data sources are available to be used as auxiliary information for de-anonymization. All of these make it possible for malicious analyzers to build sufficient background knowledge so as to re-identify people even from anonymized data sets.

After a thorough investigation, we realized the lack of well-developed teaching materials for educating younger generations on de-anonymization. Hence, we were motivated to develop a system to engage students in learning such a critical topic in today's digital era. This paper introduces the design and development of our system—IDEAL (Interactively De-Anonymization Learning). IDEAL is a web application implemented using the latest technologies to support data storage and visualization, and processes a large Weibo data set for de-anonymization. A pilot lab session was conducted among thirty-five students at Prairie View A&M University to evaluate their learning outcomes and the effectiveness of the labware. Pre and post survey analyses showed that students' feedback was very positive and encouraging.

The road map of this paper is outlined as follows. Section II introduces the related work. Section III gives an overview of the IDEAL system, followed by the detailed description of system design and implementation. Section IV presents the experimental study. Section V discusses the pilot lab and survey results with a conclusion made in Section VI.

II. Related Work

There were several serious data breaches on the Internet in recent years, which pushed the development of advanced anonymization technologies [1]–[3]. These technologies aim to preserve people's privacy over various data types by adopting a group of privacy preservation models, including k-anonymity [4], ℓ-diversity [5], t-closeness [6], and differential privacy [7]. However, the easy accessibility of data on the Internet makes it challenging to fully prevent attackers from identifying people from anonymized data sets or disclosing people's private information, especially when attackers can stitch multiple data sources together to dig deeper.

De-anonymization attacks can be categorized in different ways. First, in terms of different data types, the attacks can be based on either descriptive information or structural information. The former utilizes all kinds of descriptive information, such as users' hobbies, membership groups, location information or behavioral patterns online [8]–[10] to re-identify users, while the latter relies on users' structural information, such as centrality and neighborhood topology [11]. Some recent work combines these two types of information for de-anonymization [12]. Second, the attacks can be categorized according to different attacking approaches, namely seed based and signature based attacks. The seed based attacks [13], [14] start with a small number of seed mappings, where seeds are defined as identifiable users, and try to identify the neighbors of the seeds, and then the neighbors' neighbors, and so forth.

Unlike the seed based attacks, the signature based attacks do not assume the availability of any seeds; instead, they depend on node signatures [12], [15], [16], which are unique and can be generated from the nodes' descriptive or/and structural information. This strategy is to first generate the signatures for nodes in both the anonymized data set and the data set

978-1-7281-7303-0/20/$31.00 ©2020 IEEE
DOI 10.1109/COMPSAC48688.2020.0-210

with background knowledge, and then calculate the similarity score between each pair of nodes across these two data sets, and find the best match for nodes.

III. System Design and Implementation

IDEAL is a web-based application. It was developed with several technologies and platforms, such as Angular JS, Node.js, and Spring. The back-end database was built upon MongoDB, a NoSQL database.

A data set [17] collected from Weibo (i.e., a Chinese Twitter) was used to generate the anonymized data set, which is called Target Set (TarS), as well as the data set used as background knowledge, which is called Auxiliary Set (AuxS). We discovered that the data has missing values for some attributes, so data cleaning was conducted prior to the generation of the two data sets. The de-anonymization process is to re-identify users from the TarS, using the AuxS as background knowledge. The de-anonymization algorithm implemented in the system is seed based. The web application consists of four primary components: data cleaning, data sets generation, de-anonymization, and experiment analysis, which are described in the following subsections.

A. Data Sets Preparation

1) Data Cleaning: The Weibo data set contains several files. We only used two of them, user profile and user relation. The user profile contains basic user information, such as year of birth (YoB) and gender. The user relation file has two user IDs separated with a space on each row, indicating a following relationship. In the profile data, some users' YoB values or gender values are missing or unknown. For the missing YoB, we filled it with a random year in the range [1940-2010]. For the unknown gender values, we assigned them with 1, where 1 indicates female and 2 indicates male. In the relation file, some relations are between users whose IDs do not exist in the profile, so we cleaned the data by removing those relations. Additionally, the original relation is unidirectional, which means if 1 and 2 are in one row in the relation file, user 1 follows user 2, but not vice versa. We converted the relation to bi-directional to simplify its impact on de-anonymization without distinguishing incoming and outgoing degrees, i.e., if user 1 follows user 2, then user 2 also follows user 1.

2) Data Sets Generation: The web interface for generating the TarS and the AuxS is shown in Figure 1. The relationship information in the data set forms a graph, where nodes and edges represent users and their relationship, respectively. A TarS of the graph is a connected subgraph which is composed of selected target nodes and their edges. To generate the TarS, we first randomly select a node from the original data set, and then perform a Breadth-First-Search (BFS) to visit a certain number of nodes. Moreover, the TarS needs to be anonymized before data publishing. In the current implementation, anonymization is explicitly applied to the profile attributes such as YoB and gender, while the topology of the TarS is not changed. For YoB, we generalize a specific year to a range. Specifically, our system user first sets a year

Fig. 1: The web page of data sets generation

range value (e.g., 10), and then generates two random integers between 0 and half of the range value (i.e., [0,5]), $randmin$ and $randmax$. The generalized YoB will be [YoB-$randmin$, YoB+$randmax$]. The gender anonymization allows the user to set a probability of keeping people's original gender.

Next, the user can set the percentage of nodes in the TarS which overlap with the nodes in the AuxS. It should be noted that making overlapping nodes connected is unnecessary. We assume the connectivity of the TarS and the AuxS in order to simplify the de-anonymization process. Otherwise, each of the data sets may consist of several disjoint components, and then any pairs of the components between these two sets need to be checked for the possibility of de-anonymization. Given the percentage value, the strategy to generate the overlap is to randomly pick a node from the TarS and then run Breadth-First-Search (BFS) in the TarS to get enough nodes into the overlap. To generate the AuxS, the process starts with the nodes in the overlap generated, then runs BFS in the original set to find their neighbors which are not in the TarS, and keeps running BFS until discovering enough nodes for the AuxS. Algorithm 1 has the pseudo code of generating the AuxS.

Algorithm 1: The Generation of the AuxS

Input: Original Data Set - OS, TarS - targetlist, the overlap size, and the AuxS size
Output: The list of nodes in the AuxS
$initOverlap$ = RandomlySelectNode($targetlist$);
$overlaplist$ = BreadthFirstSearch($initOverlap$, $TarS$);
▷ ensure $overlaplist.size$ = overlap size;
$initAuxiliary$ = RandomlySelectNode($overlaplist$);
$auxiliarylist$ = BreadthFristSearch($initAuxiliary$, $overlaplist$, $TarS$, OS);
▷ all nodes in the AuxS must be either in $overlaplist$ or OS;
▷ ensure $auxiliarylist.size$ = AuxS size;
return $auxiliarylist$

Figure 2 illustrates the process of generating the TarS and the AuxS, where the configuration is to have 9 nodes in both sets and set 60% as the overlap percentage. After calculation,

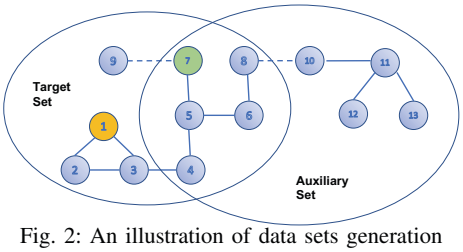


Fig. 2: An illustration of data sets generation

the number of overlap nodes is 5. If node 1 is picked to start BFS, the nodes selected will be: $1 \rightarrow 2$ and 3; $3 \rightarrow 4$; $4 \rightarrow 5$; $5 \rightarrow 6$ and 7; $6 \rightarrow 8$; and $7 \rightarrow 9$. So the TarS has nodes 1, 2, 3, 4, 5, 6, 7, 8, and 9. Suppose the overlap discovery starts with node 5. Then another BFS starts but only in the TarS: $5 \rightarrow 6$, 7 and 4; $6 \rightarrow 8$. The generation of the AuxS begins with the overlap nodes, say node 7, then $7 \rightarrow 5$; $5 \rightarrow 4$ and 6; $6 \rightarrow 8$; $8 \rightarrow 10$; $10 \rightarrow 11$; and $11 \rightarrow 12$ and 13. Although node 9 is node 7's neighbor, it is in the TarS; therefore, it cannot be included in the AuxS. Similarly, node 3 cannot be included in the AuxS either.

B. De-anonymization

The de-anonymization component contains two parts: configuration setting and graph visualization. In the configuration part, as a seed-based de-anonymization algorithm was developed in the system, the user needs to specify the number of initial seeds, the maximum value of which is the number of overlapping nodes. The seeds are the nodes which are identifiable; in other words, we know the original IDs of those nodes in the TarS. Additionally, the user can decide what information he wants to leverage to launch de-anonymization, either profile attributes (i.e., Year of Birth and Gender) or structural attributes (i.e., degree and centrality). Bonacich [18] proposed a family of centrality measurements that evaluate the importance or influence of a node in a graph. One of the measurements is called the eigenvector centrality which we chose to implement in the system. It is defined as the principal eigenvector of a graph's adjacency matrix. We adapted the code [19] in our implementation.

After setting the parameters, we can start the de-anonymization process. An example of the execution of the back-end de-anonymization algorithm is as follows: suppose the TarS and the AuxS are generated as shown in Figure 4. First, the process pairs the initially selected seeds, (T_7,A_7) and (T_8, A_8) and then visits each pair: (1) for the pair (T_7,A_7),

Fig. 3: The web page of de-anonymization configuration

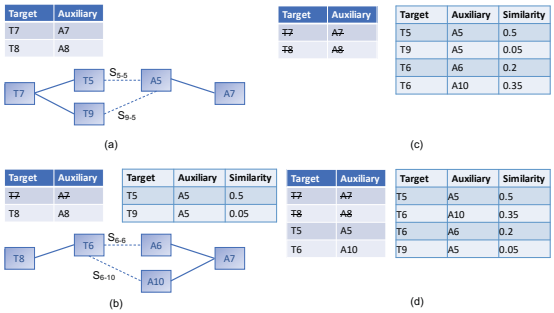


Fig. 4: An illustration of de-anonymization

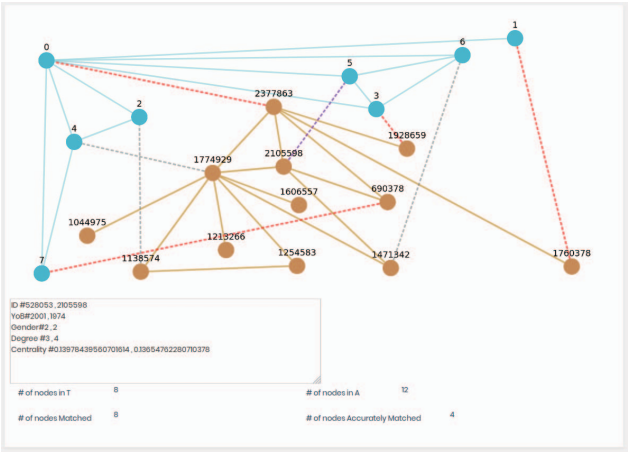


Fig. 5: An example of de-anonymization visualization

finding all of neighbors of T_7 which include T_5 and T_9 and finding all of the neighbors of A_7 which contains only A_5; (2) pairing all neighbors across these two sets and calculating their similarity scores, $S_{5-5} = 0.5$ and $S_{9-5} = 0.05$; (3) for the pair (T_8, A_8), finding their neighbors, pairing them and calculating their similarity scores; (4) sorting all pairs (T_5, A_5), (T_9, A_5), (T_6, A_6), (T_6, A_10) in the descending order of the similarity score; (5) adding (T_5, A_5) to the matched pairs as it has the highest score, and then deleting all other pairs which end with either T_5 or A_5 as they are matched already; and (6) picking the pair with the 2nd highest similarity score, (T_6, A_{10}), adding it to the matched pairs and deleting (T_6, A_6). Now two more non-visited but matched pairs are discovered, so the process continues with these pairs to check their neighbors. Algorithm 2 has the pseudo code of de-anonymization.

The node similarity score is calculated based on the user's configuration, using profile attributes or structural attributes. Two vectors are generated according to the attributes selected. Then cosine similarity is applied to the calculation.

The visualization of the de-anonymization result is implemented with Cytoscape.js [20]. Particularly, the graphs of the TarS and the AuxS are colored differently, as depicted in Figure 5. The TarS nodes only have their alternate IDs displayed since they are anonymized, while the AuxS nodes display their real IDs. The solid lines are drawn among nodes in their own groups, and the dotted lines are used to connect the matched nodes between the TarS and the AuxS in the de-anonymization. Red dotted lines indicate the correct

matches while the gray dotted lines signify incorrect ones. While hovering the cursor over a dotted line, it pops up an information box with the nodes' real IDs, attribute values, and their similarity score. The line being hovered over is purple in color. The result also includes statistical data such as the number of nodes in the TarS and the AuxS, the number of pairs of nodes matched, and the number of correct matches.

Algorithm 2: Seed-Based De-anonymization

Input: The Pairs of seeds $pairednodes$, $TarS$ and $AuxS$
Output: The Pairs of Nodes Matched $pairednodes$
$curindex$ = 0;
$tmpPairs$ = list(); **while** $pairednodes$ *has non-visited pair* **do**
  curpair = GetNodesPair(curindex);
  if $curpairisnotvisited$ **then**
    tn = GetNeighbors($curpair.target$, $TarS$);
    an = GetNeighbors($curpair.auxiliary$, $AuxS$);
    foreach t *in* tn **do**
      foreach a *in* an **do**
        if t *is not in* $pairednodes.targetNodes$ *and* a *is not in* $pairednodes.auxiliaryNodes$ **then**
          ▷ They are not paired between each other or with any other nodes.
          p=CreateNodePair(t, a);
          $p.simi$=CalculateSimilarityScore(t,a);
          $tmpPairs$.add(p);
SortTmpPairsBasedOnSimilarity($tmpPairs$);
foreach tp *in* $tmpPairs$ **do**
  $pairednodes$.add(tp);
  Remove pairs from $tmpPairs$ ending with $tp.target$ or $tp.auxiliary$;
Seed-Based De-anonymization($pairednodes$, $TarS$, $AuxS$);

C. Experiment Analysis

The analysis component provides a user-friendly interface to analyze the experiment results. As shown in Figure 6, the top section of this page has a date filter, allowing users to focus on the experiments executed during a specific time period. The filtering result is presented in a table of four columns: time stamp of experiment execution, number of seeds configured initially, configuration code (profile or structural attributes), and the de-anonymization accuracy. The de-anonymization accuracy is calculated as follows:

$$\frac{\#\ of\ correctly\ identified\ nodes - \#\ of\ seeds}{|TarS \cap AuxS| - \#\ of\ seeds} \quad (1)$$

The bottom section on this analysis page displays bar charts, as shown in Figure 7. The charts can help understand the impacts of the two parameters on the de-anonymization

Fig. 6: The result table with filter

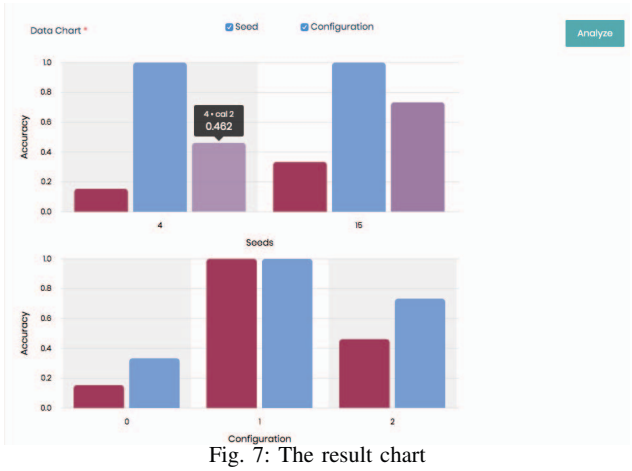


Fig. 7: The result chart

accuracy, namely the number of initial seeds and the attribute configuration for de-anonymization. These parameters can be configured on the de-anonymization page, as mentioned in Section III-B. In the seed chart, the bars are grouped in terms of the seed number. Each bar group has the same number of seeds but different attribute configuration (0, 1 or 2), which makes it easy to see the impact of configuration on de-anonymization accuracy. When a user hovers the cursor over a bar, he can see the corresponding average accuracy and the configuration code. Note that the average accuracy is calculated from the experiments executed with the same number of seeds and the same configuration setting. In the configuration chart, the bars are grouped in terms of attribute configuration. Each bar group has the same configuration but different seed numbers, making it clear to see the impact of profile/structural attributes on de-anonymization accuracy.

IV. Experimental Study

We conducted experiments to assess the impact of different factors on the de-anonymization accuracy, including the randomness injected by anonymization, the number of initial seeds, the attribute configuration for de-anonymization, and the overlap of the TarS and the AuxS. All these impacts are what we expect students to learn and observe in the lab activities.

A. Anonymization and De-anonymization

In the current implementation of the system, the anonymization can be applied only to the profile attributes. In the future,

we will add more options, such as anonymizing structural attributes. We ran two experiments upon the following setting: 30 nodes for the TarS and 30 nodes for the AuxS with 100% overlap; 5 and 20 for the generalization range; 90% and 20% for the probability of retaining gender; and using profile attributes for de-anonymization. We randomly picked 12 initial seeds and kept them the same for both experiments. Each experiment was repeated three times. The averaged de-anonymization accuracy values are 0.78 and 0.39.

B. The Impact of Seeds

The setting for this group of experiments is: 30 nodes for the TarS and 30 nodes for the AuxS with 100% overlap; five for the generalization range; 90% for the probability of retaining gender; and using profile attributes for de-anonymization. We picked seeds of 2, 12, and 22 for different experiments and repeated each experiment three times. The averaged accuracy values are 0.39, 0.48, and 0.58, respectively.

C. Profile and Structural Attributes

The setting for this group of experiments is: 30 nodes for the TarS and 30 nodes for the AuxS with 100% overlap; five for the generalization range; 90% for the probability of retaining gender; five initial seeds. We ran three experiments using profile attribute only, structural attribute only, and both for de-anonymization, which kept the same five seeds at the beginning. Again, each experiment was repeated three times. The averaged accuracy values are 0.67, 1.0 and 0.83, respectively. It should be noted that since only node profile is manipulated for the TarS anonymization, the structural attributes provide more accurate information for de-anonymization, which is even better than using both types of attributes.

D. The Overlap of TarS and AuxS

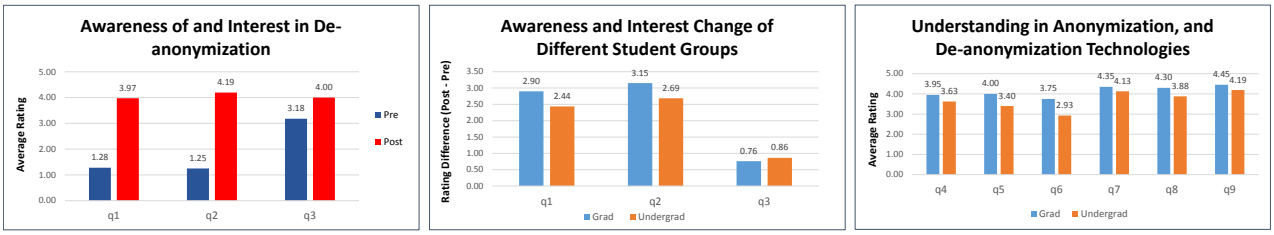
In practice, there may be differences between the TarS and the AuxS. Specifically, the AuxS that the attacker holds is incomplete or comes from a totally different online social network. Therefore, the information in the overlap of the two sets impacts the de-anonymization accuracy. A meaningful de-anonymization should be conducted only among the nodes in the overlap; however, the attacker does not know the size of the overlap and the nodes involved. Therefore, in the de-anonymization, an overlap node from the TarS may be linked to a non-overlap node in the AuxS, which reduces the de-anonymization accuracy. Moreover, the TarS and the AuxS not being fully overlapped causes some difference in nodes' structural attributes (i.e., degree and centrality). For example, the neighbors of an overlap node x in the TarS may be different from those in the AuxS. This can be regarded as the randomness of structural anonymization. So one can see the overlap between the TarS and the AuxS may also impact the de-anonymization accuracy.

We ran two experiments with the following setting: 30 nodes for the TarS and 30 nodes for the AuxS; 50% and 100% overlap, respectively; five for the generalization range; 90% for the probability of retaining gender; eight initial seeds; and using profile attributes for de-anonymization. We repeated the experiments for three times. The averaged accuracy values are 0.91 and 1.0, respectively.

V. Pilot Lab and Student Feedback

In order to verify the effectiveness of IDEAL in teaching de-anonymization, we pilot tested it at the beginning of Fall 2019. Teaching slides and lab instructions were also developed. A total of thirty-five Computer Science students volunteered to participate in this security lab. Eighteen participants were graduate students and the rest were undergraduate students. The learning and evaluation activities fell into five categories: (1) lecture to introduce de-anonymization; (2) class presentation to introduce the lab tool; (3) lab environment setup; (4) hands-on activities using the tool to examine the de-anonymization strategy and evaluate the performance in terms of user re-identification accuracy; and (5) pre and post surveys to evaluate the system and analyze the learning outcomes.

Table I presents the survey questions. All questions except for the last use a rating scale of 1 to 5 with 5 being the greatest deal or the most positive. Survey results showed that students' feedback was positive and encouraging. For questions 1-3, we analyzed the average ratings with regard to the discrepancy of the pre and post surveys. The result showed a significant increase of students' awareness of and interest in anonymization and de-anonymization technologies after the lab, as depicted in Figure 8 (a). Figure 8 (b) compares the rating change of the three questions between the graduate students and the undergraduate students. Since the participants were volunteers who already held high interest in cyber security before the lab, their interest was not increased significantly. An interesting finding is that although both cohorts had similar increase in their ratings on the awareness and interest, the graduate students showed stronger increase in their ratings on the concept understanding, while the undergraduate students showed stronger increase in the rating on their interest. This may be because the older students can grasp the concepts quickly due to their richer experience and the younger students tend to be more enthusiastic to hands-on activities. For questions 4-6, as depicted in Figure 8 (c), the students' average ratings are high. The percentage of participants who said that they gained a lot or a great deal in understanding of anonymization, de-anonymization, and de-anonymization technologies is 62.9%, 58.8%, and 42.4%, respectively. For questions 7-9, over ninety percent of the participants said that they understood the possibility of re-identifying users through the lab and knew different de-anonymization technologies. Almost all the participants felt that the lab should be taught in a security course. The percentages of students who rated either "agree" or "strongly agree" in the questions are 94.3%, 82.9%, and 91.4%, respectively. Besides, students gave very positive comments with regard to question 10. Many said that the hands-on activities enhanced their learning. Most participants felt that more exciting learning materials in privacy protection and labware like this should be developed in the future.



(a) (b) (c)

Fig. 8: (a) Students' awareness of and interest in de-anonymization (b) Comparison of graduate and undergraduate students' awareness and interest change pre and post lab (c) Feedback of different student groups on the effectiveness of the learning tool

TABLE I: Pre and post survey questions

#	Survey Questions	Type
1	How do you rate your awareness about de-anonymization in data sharing?	Pre & Post
2	How do you rate your awareness about de-anonymization technologies?	Pre & Post
3	How do you rate your interest in de-anonymization technologies?	Pre & Post
4	How do you rate your gains in understanding of anonymization?	Post
5	How do you rate your gains in understanding of de-anonymization?	Post
6	How do you rate your gains in understanding of de-anonymization technologies?	Post
7	Understand the possibility to re-identify users from anonymized data sets.	Post
8	Knowing there are different de-anonymization technologies.	Post
9	I would like this lab and de-anonymization to be taught in a computer security course.	Post
10	How could students' learning about de-anonymization be improved in this lab?	Post

VI. Conclusion

This paper discussed the importance of educating students on de-anonymization and introduced the design and implementation of the IDEAL system that was developed for teaching students about de-anonymization concepts and technologies, as well as its relevance to anonymization. The system was pilot tested among thirty-five students. The very positive feedback from the students proved the system's effectiveness in education and encouraged us to continue to develop more tools to teach different topics relevant to information privacy.

Acknowledgment

This project is supported in part by the National Science Foundation (NSF) under grant DUE-1712496. Any opinions, findings, and conclusions expressed in this paper are those of the authors, and do not necessarily reflect the views of NSF.

References

[1] L. Zou, L. Chen, and M. T. zsu, "K-automorphism: A general framework for privacy preserving network publication," in *Proceedings of Very Large Database Endowment*, 2009, pp. 946–957.

[2] B. Zhou and J. Pei, "Preserving privacy in social networks against neighborhood attacks," in *Proceedings of the IEEE 24th International Conference on Data Engineering*, 2008, pp. 506–515.

[3] J. Cheng, A. W. chee Fu, and J. Liu, "K-isomorphism: Privacy preservation in network publication against structural attack," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, 2010, pp. 459–470.

[4] L. Sweeney, "k-anonymity: a model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, 2002.

[5] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 3, 2007.

[6] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proceedings of the IEEE 23rd International Conference on Data Engineering*, 2007, pp. 106–115.

[7] C. Dwork, "Differential privacy," in *International Colloquium on Automata, Languages, and Programming. Lecture Notes in Computer Science*, vol. 4052. Springer, 2007, pp. 106–115.

[8] R. Zafarani and H. Liu, "Connecting users across social media sites: a behavioral-modeling approach," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. SIGKDD'08, 2013, pp. 41–49.

[9] T. Okuno, M. Ichino, T. Kuboyama, and H. Yoshiura, "Content-based de-anonymization of tweets," in *Proceedings of the Seventh International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, oct 2011, pp. 53–56.

[10] J. Qian, X.-Y. Li, C. Zhang, L. Chen, T. Jung, and J. Han, "Social network de-anonymization and privacy inference with knowledge graph model," *IEEE Transactions on Dependable and Secure Computing*, 2007.

[11] S. Ji, W. Li, M. Srivatsa, J. S. He, and R. Beyah, "General graph data de-anonymization: From mobility traces to social networks," *ACM Transactions on Information and System Security*, vol. 18, no. 4, 2016.

[12] H. Fu, A. Zhang, and X. Xie, "Effective social graph deanonymization based on graph structure and descriptive information," *ACM Transactions on Intelligent Systems and Technology*, vol. 6, no. 4, 2015.

[13] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *Proceedings of the 2009 IEEE Symposium on Security and Privacy*, 2009, pp. 173–187.

[14] A. Narayanan, E. Shi, and B. I. P. Rubinstein, "Link prediction by de-anonymization: How we won the kaggle social network challenge," in *Proceedings of the 2011 International Joint Conference on Neural Networks*, 2011, pp. 1825–1834.

[15] K. Henderson, B. Gallagher, L. Li, L. Akoglu, T. Eliassi-Rad, H. Tong, and C. Faloutsos, "It's who you know: Graph mining using recursive structural features," in *Proceedings of the 17th ACM International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 663–671.

[16] M. Korayem and D. J. Crandall, "De-anonymizing users across heterogeneous social computing platforms," in *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, 2013.

[17] "Kdd cup 2012, track 1, predict which users (or information sources) one user might follow in tencent weibo." 2012. [Online]. Available: https://www.kaggle.com/c/kddcup2012-track1

[18] P. Bonacich, "Power and centrality: A family of measures," *American Journal of Sociology*, 1987.

[19] M. Needham, "Java/jblas: Calculating eigenvector centrality of an adjacency matrix," 2013. [Online]. Available: https://markhneedham.com/blog/2013/08/05/javajblas-calculating-eigenvector-centrality-of-an-adjacency-matrix/

[20] M. Franz, C. T. Lopes, G. Huck, Y. Dong, O. Sumer, and G. D. Bader, "Cytoscape.js: a graph theory library for visualisation and analysis," *Bioinformatics*, vol. 32, no. 2, 2016. [Online]. Available: http://js.cytoscape.org/