De-Health: All Your Online Health Information Are Belong to Us

Shouling Ji[†,‡], Qinchen Gu[§], Haiqin Weng[¶], Qianjun Liu[†], Pan Zhou[*], Jing Chen[♯], Zhao Li[£], Raheem Beyah[§], and Ting Wang[♮]

[†]Zhejiang University [‡]Alibaba-Zhejiang University Joint Institute of Frontier Technologies [§]Georgia Tech
[¶]Ant Financial Service Group [*]HUST [♯]Wuhan University [£]Alibaba Group [♮]Penn State University
{*sji, liuqj0522*}*@zju.edu.cn* {*guqinchen, rbeyah*}*@ece.gatech.edu* *haiqin.wenghaiqin@antfin.com*
panzhou@hust.edu.cn *chenjing@whu.edu.cn* *lizhao.lz@alibaba-inc.com* *inbox.ting@gmail.com*

***Abstract*—In this paper, we study the privacy of online health data. We present a novel online health data De-Anonymization (DA) framework, named *De-Health*. Leveraging two real world online health datasets WebMD and HealthBoards, we validate the DA efficacy of De-Health. We also present a linkage attack framework which can link online health/medical information to real world people. Through a proof-of-concept attack, we link 347 out of 2805 WebMD users to real world people, and find the full names, medical/health information, birthdates, phone numbers, and other sensitive information for most of the re-identified users. This clearly illustrates the fragility of the privacy of those who use online health forums.**

I. Introduction and Motivation

Status Quo. The advance of information technologies has greatly transformed the delivery means of healthcare services: from traditional hospitals/clinics to various online healthcare services. Ever since their introduction, online health services experienced rapid growth, and have had millions of users and accumulated billions of users' medical/health records [5]. According to several national surveys, $\sim 59\%$ of US adults have employed the Internet as a diagnostic tool in 2012 [1], and on average, the US consumers spend ~ 52 hours annually to search for and pursue online health information while only visiting doctors three times per year in 2013 [2]. Moreover, "on an average day, 6% of the US Internet users perform online medical searches to better prepare for doctors' appointments and to better digest information obtained from doctors afterwards" [3]. Therefore, online health services play a more important role in people's daily life.

When serving users (we use patients and users interchangeably in this paper), the online health services accumulate a huge amount of the users' health data. For instance, as one of the leading American corporations that provide health news, advice, and expertise [4], WebMD reached an average of approximately 183 million monthly unique visitors and delivered approximately 14.25 billion page views in 2014. Another leading health service provider, HealthBoards (HB), has over 10 million monthly visitors, 850,000 registered members, and over 4.5 million health-related/medical messages posted [5]. Due to the high value of enabling low-cost, large-scale data mining and analytics tasks, e.g., disease transmission and control research, disease inference, business, government applications, and other scenarios [12], those user-generated health data are increasingly shared, disseminated, and published.

Privacy Issues of Online Health Data. In addition to the high value for various applications, online health data carry numerous sensitive details of the users that generate them [23]. Therefore, before sharing, disseminating, and publishing the health data, proper privacy protection mechanisms should be applied and privacy policies should be followed. However, the reality is that it is still an open problem for protecting online health data's privacy with respect to both the *technical* perspective and the *policy* perspective.

From the technical perspective, most existing health data anonymization techniques (which are called *de-identification* techniques in the medical and policy literature [23]), e.g., the techniques in [13], if not all, focus on protecting the privacy of *structured medical/health data* that are usually generated from hospitals, clinics, and/or other official medical agencies (e.g., labs, government agencies). Nevertheless, putting aside their performance and effectiveness, existing privacy protection techniques for structured health data can hardly be applied to online health data due to the following reasons [8] [9] [12]. (i) *Structure and heterogeneity:* the structured health data are well organized with structured fields while online health data are usually heterogeneous and structurally complex. (ii) *Scale:* a structured health dataset usually consists of the records of tens of users to thousands of users [13], while an online health dataset can contain millions of users [5] [12]. (iii) *Threat:* Compared to online health data, the dissemination of structured health data is easier to control, and thus a privacy compromise is less likely. Due to its open-to-public nature, however, the dissemination of online health data is difficult to control, and adversaries may employ multiple kinds of means and auxiliary information to compromise the data's privacy.

From the policy making perspective, taking the US Health Insurance Portability and Accountability Act (HIPAA) [6] as an example, although HIPAA sets forth methodologies for anonymizing health data, once the data are anonymized, they are no longer subject to HIPAA regulations and can be used for any purpose. However, when anonymizing the data, HIPAA does not specify any concrete techniques. Therefore, the naive

2375-026X/20/$31.00 ©2020 IEEE
DOI 10.1109/ICDE48307.2020.00143

anonymization technique may be applied.

Our Work. Towards helping users, researchers, data owners, and policy makers comprehensively understand the privacy vulnerability of online health data, we study the privacy of online health data. Specifically, we focus on the health data generated on online health forums like WebMD [4] and HB [5]. These forums disseminate personalized health information and provide a community-based platform for connecting patients among each other as well as with doctors via interactive questions and answers, symptom analysis, medication advice and side effect warning, and other interactions [4] [5] [8].

As we mentioned earlier, a significant amount of medical records have been accumulated in the repositories of these health websites. According to the website privacy policies [4] [5], they explicitly state that they collect personal information of users (patients), including contact information, personal profile, medical information, transaction information, Cookies, and other sensitive information. To use these online health services, users have to accept their privacy policies. For instance, in HB's privacy policy, it is explicitly indicated that "if you do not agree to this privacy policy, please do not use our sites or services". Therefore, using the online health services requires the enabling of these service providers like WebMD and HB to collect users' personal information.

As stated, the collected personal information will be used for research and *various business purposes*, e.g., precise advertisements from pharmaceutical companies. Although these medical records are only affiliated with user-chosen pseudonyms or anonymized IDs, some natural questions arise: when those data are shared with commercial partners (one of the most typical business purposes)[1], or published for research, or collected by adversaries, can they be de-anonymized even if the patients who generated them are anonymized? And can those medical records be connected to real world people? In this paper, we answer these two questions by making the following contributions.

(1) We present a novel DA framework, named *De-Health*, for large-scale online health data. De-Health is a two-phase DA framework. In the first phase, De-Health performs *Top-K DA* and constructs a *Top-K candidate set* for each anonymized user. In the second phase, *refined DA* is performed, and De-Health maps an anonymized user to a candidate user.

(2) Leveraging two real world online health datasets WebMD (89,393 users, 506K posts) and HB (388,398 users, 4.7M posts), we conduct extensive evaluations to examine the performance of De-Health in both the closed-world setting and the open-world DA setting. The results show that De-Health is powerful in practice.

(3) We present a linkage attack framework, which can link online health service users to real world people. We validate the framework leveraging proof-of-concept attacks. For instance, it can successfully link 347 out of 2805 (i.e., 12.4%) target WebMD users to real world people, and find most of their full names, medical information, birthdates, addresses, and other sensitive information. Thus, those users' privacy can be compromised, and one can learn various aspects of information such as the sexual orientation and related infectious diseases, mental/psychological problems, and even suicidal tendency from some users' medial data.

[1]The business model (revenue model) of most online health forums is advertisement based [4] [5].

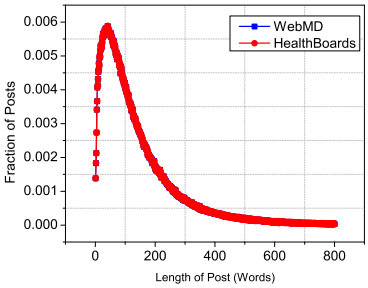


Fig. 1. Post length distribution.

II. Data Collection & Feature Extraction

A. Data Collection

We collect online medical postings from two leading US online health services providers WebMD [4] and HB [5]. The collection continued for approximately 4 months, from May to August 2015, and resulted in 540,183 webpages from WebMD and 2,596,433 webpages from HB. After careful analysis and processing, we extracted 506,370 disease/condition/medicine posts that were generated by 89,393 registered users from the WebMD dataset (5.66 posts/user on average) and 4,682,281 posts that were generated by 388,398 registered users from the HB dataset (12.06 posts/user on average).

We also conduct preliminary analysis on the collected data. The length distribution of the posts in WebMD and HB in terms of the number of words is shown in Fig.1. Most of the posts in the two datasets have a length less than 300 words. On average, the length of WebMD posts is 127.59 and the length of HB posts is 147.24.

B. Feature Extraction

User Correlation Graph. Online health services provide a platform for connecting patients via interactive disease and symptom discussion, health question answering, medicine and possible side effect advice, etc. For instance, on WebMD and HB, when one disease topic is raised by some user, other users may join the discussion of this topic by providing suggestions, sharing experience, making comments, etc.

Therefore, if we take such user interactivity into consideration, there is some correlation, i.e., the co-disease/health/medicine discussion relation, among users. To characterize such interactivity, we construct a *user correlation graph* based on the relationships among the data (posts) of different users. Particularly, for each user in the WebMD/HB dataset, we represent him/her as a node in the correlation graph. For two users i and j, if they post under the same health/disease topic, i.e., they made posts on the same topic initialized by some user (could be i, j, or some other user), we consider that there is an undirected edge, denoted by e_{ij},

between i and j. Furthermore, we note that the number of interactive discussions between different pairs of users might be different. Therefore, we assign a *weight* for each edge to characterize the interactivity strength, which is defined as the number of times that the corresponding two users co-discussed under the same topic.

Now, we formally define the user correlation graph as $G = (V, E, W)$, where $V = \{1, 2, \cdots, n\}$ denotes the set of users, $E = \{e_{ij} | i, j \in V\}$ denotes the set of edges among users, and $W = \{w_{ij} |$ for $e_{ij} \in E, w_{ij}$ is the weight (interactivity strength) associated with edge $e_{ij}\}$. For $i \in V$, we define its *neighborhood* as $N_i = \{j | e_{ij} \in E\}$. Let $d_i = |N_i|$ be the number of neighbor users of i, i.e., the *degree* of user i. When taking the weight information into consideration, we define $wd_i = \sum_{j \in N_i} w_{ij}$ to be the *weighted degree* of i. For our following application, we also define a *Neighborhood Correlation Strength* (NCS) vector for each user. Specifically, for $i \in V$, its NCS vector is defined as $\mathbf{D}_i =< w'_{ij} | j \in N_i >$, where $< w'_{ij} | j \in N_i >$ is a decreasing order sequence of $\{w_{ij} | j \in N_i\}$. Given $i, j \in V$, we define the *distance* (resp., *weighted distance*) between i and j as the length of the shortest path from i to j in G when the weight information is overlooked (resp., considered), denoted by h_{ij} (resp., wh_{ij}).

We analyze the degree distributions of the WebMD graph and the HB graph, as well as the community structure of the WebMD graph in the technical report. Basically, the graph's connectivity is not strong.

Stylometric Features. The use of writing style for author attribution can be traced back to the 19th century [37]. Recently, stylometric approaches have been applied to broad security and privacy issues, from author attribution [26] to fraud and deception detection, underground cybercriminal analysis, and programmer DA [29]. Furthermore, it is difficult for users to intentionally obfuscate their writing style or attempt to imitate the writing styles of others in a long term. Even that happens, with a high probability, specific linguistic features can still be extracted from the long term written materials to identify the users. Therefore, for our purpose, we seek to employ the linguistic features of the health data (posts of users) to de-anonymize the associated users.

We extract various stylometric features from the WebMD and HB datasets as shown in Table I. Generally, the features in Table I can be classified into three groups: *lexical* features, *syntactic* features, and *idiosyncratic* features. The lexical features include length, word length, vocabulary richness, letter frequency, digit frequency, uppercase letter percentage, special characters, and word shape. They measure the writing style of users with respect to the characteristics of employed characters, words, and vocabularies. The syntactic features include punctuation frequency, function words, POS tags, and POS tag bigrams. They measure the writing style of users with respect to the arrangement of words and phrases to create well-formed sentences in posts. For idiosyncratic features, we consider misspelled words, which measure some peculiar writing style of users.

TABLE I
STYLOMETRIC FEATURES.

Category	Description	Count
Length	# of characters and paragraphs, average # of characters per word	3
Word Length	freq. of words of different lengths	20
Vocabulary richness	Yule's K, hapax /tris/dis/tetrakis legomena	5
Letter freq.	freq. of 'a/A' $\sim$ 'z/Z'	26
Digit freq.	freq. of '0' $\sim$ '9'	10
Uppercase letter percentage	% of uppercase letters in a post	1
Special characters	freq. of special characters	21
Word shape	freq. of all uppercase words, all lowercase words, first character uppercase words, camel case words	21
Punctuation freq.	freq. of punctuation, e.g., !,;?	10
Function words	freq. of function words	337
POS tags	freq. of POS tags, e.g., NP, JJ	< 2300
POS tag bigrams	freq. of POS tag bigrams	$< 2300^2$
Misspelled words	freq. of misspellings	248

Since the number of POS tags and POS tag bigrams could be variable, the number of total features is denoted by a variable M for convenience. According to the feature descriptions, all the features are real and positive valued. Without loss of generality, we organize the features as a vector, denoted by $\mathbf{F} =< F_1, F_2, \cdots, F_M >$. Then, given a post, we extract its features with respect to $\mathbf{F}$ and obtain a feature vector consisting of 0 and positive real values, where 0 implies that this post does not have the corresponding feature while a positive real value implies that this post has the corresponding feature.

Note that, it is possible to extract more stylometric features from the WebMD/HB dataset, e.g., content features. However, in this paper, we mainly focus on developing an effective online health data DA framework. For feature extraction, we mainly employ the existing techniques such as those in [26] [29], and we do not consider this part as the technical contribution of this paper.

User-Data-Attribute Graph and Structural Features. Previously, we constructed a correlation graph G for the users in a health dataset. Now, we extend G to a User-Data-Attribute (UDA) graph. As the stylometric features demonstrate the writing characteristics of users, logically, they can also be considered as the *attributes* of users, which are similar to the social attributes of users, e.g., career, gender, citizenship. Therefore, at the user level, we define an *attribute set/space*, denoted by A, based on $\mathbf{F}$, i.e., $A = \{A_i | A_i = F_i, i = 1, 2, \cdots, M\}$. Then, following this idea, for each feature $F_i \in \mathbf{F}$, if a user u has a post that has feature F_i (i.e., the F_i dimension is not 0 in the feature vector of that post), we say u has attribute A_i, denoted by $u \sim A_i$. Note that, each attribute is actually *binary* to a user, i.e., a user either has an attribute A_i or not, which is different from the feature, which could be either a continuous or a discrete real value. We define $A(u)$ as the set of all the attributes that user u has, i.e., $A(u) = \{A_i | A_i \in A, u \sim A_i\}$. Since u may have

multiple posts that have feature F_i, we assign a *weight* to the relation $u \sim A_i$, denoted by $l_u(A_i)$, which is defined as the number of posts authored by u that have the feature F_i.

Based on the attribute definition, we extend the correlation graph to the UDA graph, denoted by $G = (V, E, W, A, O, L)$, where V, E, and W are the same as defined before, A is the attribute set, $O = \{u \sim A_i | u \in V, A_i \in A\}$ denotes the set of all the user-attribute relationships, and $L = \{l_u(A_i) | u \sim A_i \in O\}$ denotes the set of the user-attribute relationship weights. Since the UDA graph is an extension of the correlation graph, we use the same notation G for these two concepts. In practice, one may consider more attributes of a user, e.g., their social attributes (user's social information) and behavioral attributes (user's activity pattern), when defining A.

From the definition of the UDA graph G, we can see that it takes into account the data's correlation as well as the data's linguistic features (by introducing the concept of attribute in a different way compared to the traditional manner [26] [29]). We will introduce how to use the UDA graph to conduct the user-level DA and analyze the benefits in the following section. Before that, we introduce more user-level features from the health data leveraging the UDA graph.

The features extracted from the UDA graph are classified as *structural features*, which can be partitioned into three categories: *local correlation features*, *global correlation features*, and *attribute features*. The local correlation features include user degree (i.e., d_u for $u \in V$), weighted degree (i.e., wd_u), and NCS vectors (i.e., $\mathbf{D}_i$). Basically, the local correlation features measure the *direct interactivity* of a health forum user.

Given $u \in V$ and a subset $S \subseteq V$, the global correlation features of u are defined as the distances and weighted distances from u to the users in S, denoted by vectors $\mathbf{H}_u(S) =< h_{uv} | v \in S >$ and $\mathbf{WH}_u(S) =< wh_{uv} | v \in S >$, respectively. Basically, the global correlation features measure the *indirect interactivity* of a user in a dataset.

Based on $A(u)$ of $u \in V$, we introduce a new notation to take into account the weight of each attribute of u. We define $WA(u) = \{(A_i, l_u(A_i)) | A_i \in A(u)\}$. Then, the attribute features of $u \in V$ are defined as $A(u)$ and $WA(u)$. The attribute features measure the linguistic features of users in the form of binary attributes and weighted binary attributes. The defined structural features are helpful in conducting user-level DA. We show this in detail in the De-Health framework as well as in the experiments.

III. De-anonymization

In this section, we present De-Health. The considered anonymized data, denoted by Δ_1, are the data generated from current online health services, e.g., WebMD and HB. There are multiple applications of these anonymized online health data: (i) as indicated in the privacy policies of WebMD and HB, the health data of their users can be shared with researchers for multiple research and analytics tasks [4] [5]; (ii) again, according to their privacy policies, the data could be shared with commercial partners (e.g., insurance companies and pharmaceutical companies) for multiple business purposes [4] [5]; and (iii) the data might be publicly released for multiple government and societal applications [33]–[36]. Considering various applications of the online health data, our question is: *can those data be de-anonymized to the users of online health services and can they be linked to the users' real identities*? We answer the first part of this question in this section and discuss the second part in Section V.

To de-anonymize the anonymized data Δ_1, we assume that the adversary[2] can collect some auxiliary data, denoted by Δ_2, from the same or other online health service. According to our knowledge, this is possible in practice: from the adversary's perspective, for some online health services, e.g., HB, it is not difficult to collect data from them using some intelligent crawling techniques; for some other online health services with strict policies, e.g., PatientsLikeMe [7], an adversary can also collect their data by combining intelligent crawling techniques and anonymous communication techniques (e.g., Tor). In this paper, we assume both Δ_1 and Δ_2 are generated from online health services like WebMD and HB.

After obtaining Δ_1 and Δ_2, we extract the features of the data and transform them into an anonymized graph and an auxiliary graph, denoted by $G_1 = (V_1, E_1, W_1, A_1, O_1, L_1)$ and $G_2 = (V_2, E_2, W_2, A_2, O_2, L_2)$, respectively, using the techniques discussed in Section II. When it is necessary, we use the subscript '1' and '2' to distinguish the anonymized data/graph and the auxiliary data/graph. Now, the DA of Δ_1 leveraging Δ_2 can be approximately defined as: for an anonymized (unknown) user $u \in V_1$, seeking an auxiliary (known) user $v \in V_2$, such that u can be identified to v (i.e., they correspond to the same real world person), denoted by $u \rightarrow v$. However, in practice, it is unclear whether Δ_1 and Δ_2 are generated by the same group of users, i.e., it is unknown whether $V_1 \stackrel{?}{=} V_2$. Therefore, we define *closed-world DA* and *open-world DA*. When the users that generate Δ_1 are a subset of the users that generate Δ_2, i.e., $V_1 \subseteq V_2$, the DA problem is a *closed-world DA* problem. Then, a successful DA is defined as $u \in V_1, v \in V_2$, $u \rightarrow v$ and u and v correspond to the same user. When $V_1 \neq V_2$, the DA problem is an *open-world DA* problem. Let $V_o = V_1 \cap V_2$, the overlapping users between V_1 and V_2. Then, a successful DA is defined as $u \in V_1, v \in V_2$, $u \rightarrow v$, u and v are in V_o, and u and v correspond to the same user; or $u \rightarrow \perp$, if $u \notin V_o$, where $\perp$ represents *not-existence*. For $u \in V_1$ and $v \in V_2$, if u and v correspond to the same real world user, we call v the *true mapping* of u in V_2. In this paper, the presented De-Health framework works for both the closed-world and the open-world situations.

A. De-Health

Overview. In this subsection, we present the De-Health framework. We show the high level idea of De-Health in

[2]Here, the adversaries are defined as the ones who want to compromise the privacy of the users in the anonymized dataset. During the data sharing and publishing process (for research, business, and other purposes), every data recipient could be an adversary. In our paper, we focus on studying the potential privacy vulnerability of online health data.

Algorithm 1: De-Health

1 construct G_1 and G_2 from Δ_1 and Δ_2, respectively;
2 **for** *every* $u \in V_1$ **do**
3 **for** *every* $v \in V_2$ **do**
4 compute the *structural similarity* between u and v, denoted by s_{uv};
5 compute the Top-K candidate set for each user $u \in V_1$, denoted by $C_u = \{v_i | v_i \in V_2, i = 1, 2, \cdots, K\}$, based on the structural similarity scores;
6 filter C_u using a *threshold vector*;
7 **for** $u \in V_1$ **do**
8 leveraging the stylometric and structural features of the users in C_u, build a classifier, using benchmark machine learning techniques (e.g., SMO);
9 using the classifier to de-anonymize u;

Algorithm 1 and give the details later. At a high level, De-Health conducts user DA in two phases: *Top-K DA* (line 2-6) and *refined DA* (line 7-9). In the Top-K DA phase, we mainly focus on de-anonymizing each anonymized user $u \in V_1$ to a Top-K candidate set, denoted by $C_u = \{v_i | v_i \in V_2, i = 1, 2, \cdots, K\}$, that consists of the K most structurally similar auxiliary users with the anonymized user (line 2-5). Then, we optimize the Top-K candidate set using a threshold vector by eliminating some less likely candidates (line 6). In the *refined DA* phase, an anonymized user will be de-anonymized to some candidate user using a benchmark machine learning model trained leveraging both stylometric and structural features. Note that, we do not limit the DA scenario to closed-world or open-world. De-Health is designed to take both scenarios into consideration.

Top-K DA. Now, we discuss how to implement Top-K DA and optimization (filtering).

Structural Similarity. Before we compute the Top-K candidate set for each anonymized user, we compute the *structural similarity* between each anonymized user $u \in V_1$ and each auxiliary user $v \in V_2$, denoted by s_{uv}, from the graph perspective (line 2-3 in Algorithm 1). In De-Health, s_{uv} consists of three components: *degree similarity* s^d_{uv}, *distance similarity* s^s_{uv}, and *attribute similarity* s^a_{uv}. Specifically, s^d_{uv} is defined as $s^d_{uv} = \frac{\min\{d_u, d_v\}}{\max\{d_u, d_v\}} + \frac{\min\{wd_u, wd_v\}}{\max\{wd_u, wd_v\}} + \cos(\mathbf{D}_u, \mathbf{D}_v)$, where $\cos(\cdot, \cdot)$ is the *cosine similarity* between two vectors. Note that, it is possible that $\mathbf{D}_u$ and $\mathbf{D}_v$ have different lengths. In that case, we pad the short vector with zeros to ensure that both have the same length. From the definition, s^d_{uv} measures the degree similarity of u and v in G_1 and G_2, i.e., their local direct interactivity similarity in Δ_1 and Δ_2, respectively.

To define s^s_{uv}, we need to specify a set of *landmark users* from G_1 and G_2, respectively. Usually, the landmark users are some *pre-de-anonymized* users that serve as *seeds* for a DA [33]–[36]. There are many techniques to find landmark users, e.g., clique-based technique, community-based technique, and optimization-based technique [33]–[36]. In De-Health, we do not require accurate landmark users. In particular, we select $\hbar$ users with the largest degrees from V_1 and V_2 as the landmark users, denoted by S_1 and S_2, respectively. We also sort the users in S_1 and S_2 in the degree decreasing order. Then, we define s^s_{uv} as $s^s_{uv} = \cos(\mathbf{H}_u(S_1), \mathbf{H}_v(S_2)) + \cos(\mathbf{WH}_u(S_1), \mathbf{WH}_v(S_2))$. Basically, s^s_{uv} measures the relative global structural similarity, i.e., indirect interactivity similarity, of u and v.

For u and v, we define $WA(u) \cap WA(v) = \{(A_i, l_{u\cap v}(A_i)) | A_i \in A(u) \cap A(v), l_{u\cap v}(A_i) = \min\{l_u(A_i), l_v(A_i)\}\}$ and $WA(u) \cup WA(v) = \{(A_i, l_{u\cup v}(A_i)) | A_i \in A(u) \cup A(v), l_{u\cup v}(A_i) = \max\{l_u(A_i), l_v(A_i)\}\}$. Further, let $|\cdot|$ be the cardinality of a set and for the weighted set, we define $|WA(u) \cap WA(v)| = \sum_{A_i \in A(u)\cap A(v)} l_{u\cap v}(A_i)$ and $|WA(u) \cup WA(v)| = \sum_{A_i \in A(u)\cup A(v)} l_{u\cup v}(A_i)$. Then, s^a_{uv} is defined as $s^a_{uv} = \frac{|A(u)\cap A(v)|}{|A(u)\cup A(v)|} + \frac{|WA(u)\cap WA(v)|}{|WA(u)\cup WA(v)|}$, which measures the attribute similarity (i.e., linguistic similarity) between u and v.

After specifying s^d_{uv}, s^s_{uv}, and s^a_{uv}, the structural similarity between u and v is defined as $s_{uv} = c_1 \cdot s^d_{uv} + c_2 \cdot s^s_{uv} + c_3 \cdot s^a_{uv}$, where c_1, c_2 and c_3 are positive constant values adjusting the weights of each similarity component.

Top-K Candidate Set. After obtaining the structural similarity scores, we compute the Top-K candidate set C_u for each $u \in V_1$ (line 5 in Algorithm 1)[3]. Here, we propose two approaches: *direct selection* and *graph matching based selection*. In *direct selection*, we directly select K auxiliary users from V_2 that have the Top-K similarity scores with u. In *graph matching based selection*: Step 1: we first construct a *weighted completely connected bipartite graph* $G(V_1, V_2)$ (anonymized users on one side while auxiliary users on the other side), where the weight on each edge is the structural similarity score between the two corresponding users; Step 2: we find a *maximum weighted bipartite graph matching* on $G(V_1, V_2)$, denoted by $\{(u_i, v_i) | u_i \in V_1, v_i, i = 1, 2, \cdots, |V_1|\}$; Step 3: for each (u_i, v_i) in the matching, we add v_i to the Top-K candidate set of u_i and remove the edge between u_i and v_i in the bipartite graph $G(V_1, V_2)$; Step 4: repeat Steps 2 and 3 until we find a Top-K candidate set for each user in V_1.

Optimization/Filtering. After determining the Top-K candidate set for each $u \in V_1$, we further optimize C_u using the *filtering procedure* shown in Algorithm 2 (to finish line 6 in Algorithm 1), where $\epsilon \in [0, s_u - \min\{s_{uv} | u \in V_1, v \in V_2\}]$ is a positive constant value, l is the length of the threshold vector $\mathbf{T}$ (defined later), and C'_u is a temporary candidate set. The main idea of the filtering process is to pre-eliminate some less likely candidates in terms of structural similarity using a threshold vector. Below, we explain Algorithm 2 in detail. First, the threshold interval $[s_l, s_u]$ is specified based on ϵ, and the maximum and minimum similarity scores between the users in V_1 and V_2 (line 1-2). Then, the threshold interval is partitioned into l segments with the threshold value $T_i = s_u - \frac{i}{l-1} \cdot (s_u - s_l)$ $(i = 0, 1, \cdots, l-1)$. We organize the threshold values as a *threshold vector* $\mathbf{T} =< T_i >$ (line 3).

[3]Here, we assume that K is far less than the number of auxiliary users. Otherwise, it is meaningless to seek Top-K candidate sets.

Algorithm 2: Filtering

1 $s_u \leftarrow \max\{s_{uv} | u \in V_1, v \in V_2\}$;
2 $s_l \leftarrow \min\{s_{uv} | u \in V_1, v \in V_2\} + \epsilon$;
3 construct a *threshold vector* $\mathbf{T} =< T_i >$, where for $i = 0, 1, \cdots, l-1 >$, $T_i = s_u - \frac{i}{l-1} \cdot (s_u - s_l)$;
4 **for** *every* $u \in V_1$ **do**
5 **for** $i = 0; i \leq l-1; i++$ **do**
6 $C'_u \leftarrow C_u$;
7 **for** $v \in C'_u$ **do**
8 **if** $s_{uv} < T_i$ **then**
9 $C'_u = C'_u \setminus \{v\}$;
10 **if** $C'_u \neq \emptyset$ **then**
11 $C_u \leftarrow C'_u$, **break**;
12 **if** $C'_u = \emptyset$ **then**
13 $u \rightarrow \perp$, $V_1 \leftarrow V_1 \setminus \{u\}$;

Third, we use $\mathbf{T}$ to filter each candidate set C_u starting from large thresholds to small thresholds (line 5-13). If one or more candidate users pass the filtering at a certain threshold level, we then break the filtering process and take those candidate users as the final C_u (line 7-10). If no candidate users are left even after being filtered by T_{l-1} (the smallest threshold), we conclude that u does not appear in the auxiliary data (i.e., $u \rightarrow \perp$) and remove u from V_1 for further consideration (line 12-13).

Note that, the filtering process is mainly used for reducing the size of the candidate set for each anonymized user, and thus to help obtain a better refined DA result and accelerate the DA process in the following stage. In practice, there is no guarantee for the filtering to improve the DA performance. Therefore, we set the filtering process as an *optional choice* for De-Health.

Refined DA. In the first phase of De-Health, we seek a Top-K candidate set for each anonymized user. In the second phase (line 7-9 of Algorithm 1), De-Health conducts refined DA for each $u \in V_1$ and either de-anonymizes u to some auxiliary user in C_u or concludes that $u \rightarrow \perp$. To fulfill this task, the high level idea is: leveraging the stylometric and correlation features of the users in C_u, train a classifier employing benchmark machine learning techniques, e.g., Support Vector Machine (SVM), Nearest Neighbor (NN), Regularized Least Squares Classification (RLSC), which is similar to that in existing stylometric approaches [26] [29][4]. Therefore, we do not go to further details to explain existing benchmark techniques.

By default, existing benchmark machine learning techniques are satisfiable at addressing the closed-world DA problem (e.g., [26]). However, their performance is far from expected in open-world DA [27]. To address this issue, we present two schemes: *false addition* and *mean-verification*, which are motivated by the open-world author attribution techniques proposed by Stolerman et al. in [27].

[4]In [26] [29], multiple benchmark machine learning based stylometric approaches are proposed to address the post/passage-level author attribution. Although we focus on user-level DA, those approaches could be extended to our refined DA phase.

In the *false addition* scheme, when de-anonymizing $u \in V_1$, we randomly select K' users from $V_2 \setminus C_u$, and add these K' users to C_u as *false users*. Then, if u is de-anonymized to a false user in C_u, we conclude that $u \rightarrow \perp$, i.e., u does not appear in the auxiliary data. Otherwise, u is de-anonymized to a non-false user.

In the *mean-verification* scheme, we first use the trained classifier to de-anonymize u to some user, say v, in C_u by assuming it is a closed-world DA problem. Later, we verify this DA: let $\lambda_u = (\sum_{w \in C_u} s_{uw})/|C_u|$ be the *mean similarity* between u and its candidate users; then, if $s_{uv} \geq (1+r) \cdot \lambda_u$, where $r \geq 0$ is some predefined constant value, the DA $u \rightarrow v$ is accepted; otherwise, it is rejected, i.e., $u \rightarrow \perp$. Note that, the verification process can also be implemented using other techniques, e.g., distractorless verification [38], Sigma verification [27].

Remark. The Top-K DA phase of De-Health can improve the DA performance from multiple perspectives. On one hand, it significantly reduces the possible mapping space for each anonymized user, and thus a more accurate classifier can be trained, followed by the improved DA performance. From the description of De-Health (Algorithm 1), it seems that the Top-K DA might degrade its DA performance if many true mappings of the anonymized users cannot be included into their Top-K candidate sets. However, we seek the candidate set for each anonymized user u based on structure similarities between u and the users in V_2, and the auxiliary users that have high structural similarities with u are preferred to be selected as candidates, e.g., in the direct selection approach. Furthermore, as shown in our experiments (Section IV), most anonymized users' true mappings can be selected into their candidate sets when a proper K is chosen. On the other hand, since the possible mapping space is significantly reduced by the Top-K DA, the computational cost for both constructing the classifiers and performing refined DA can be reduced. Most real world DA tasks are open-world problems. By using the false addition and mean-verification schemes, De-Health can address both closed-world and open-world DA issues.

IV. Experiments

In this section, we first evaluate De-Health's performance in the closed-world DA setting. Then, we extend our evaluation to the more practical open-world DA setting: the anonymized data and the auxiliary data only have partial overlapping users.

A. Closed-world DA

As described in Algorithm 1, De-Health consists of two phases, where the first phase is for Top-K DA, i.e., seeking a candidate set for each anonymized user, and the second phase is for refined DA, i.e., de-anonymizing an anonymized user either to some user in the corresponding candidate set or to $\perp$ (non-existence).

1) Top-K DA.: First, we evaluate the Top-K DA performance of De-Health. In the Top-K DA phase, we seek a candidate set $C_u \subseteq V_2$ for each anonymized user u. We define that the Top-K DA of u is *successful/correct* if u's

true mapping is included in the C_u returned by De-Health. Note that, the Top-K DA is crucial to the success and overall performance of De-Health: given a relatively large auxiliary dataset and a small K, if there is a high success rate in this phase, the candidate space of finding the true mapping of an anonymized user can be significantly reduced (e.g., from millions or hundreds of thousands of candidates to several hundreds of candidates). Then, many benchmark machine learning techniques can be employed to conduct the second phase refined (precise) DA, since as shown in [24]- [30], benchmark machine learning techniques can achieve much better performance on a relatively small training dataset than on a large training dataset[5].

Methodology and Setting. We partition each user's data (posts) in WebMD and HB into two parts: *auxiliary data* denoted by Δ_2 and *anonymized data* denoted by Δ_1. Specifically, we consider three scenarios: randomly taking 50%, 70%, and 90% of each user's data as auxiliary data and the rest as anonymized data (by replacing each username with some random ID), respectively. Then, we run De-Health to identify a Top-K candidate set for each user in Δ_1 and examine the CDF of the successful Top-K DA with respect to the increase of K. For the parameters in De-Health, the default settings are: $c_1 = 0.05$, $c_2 = 0.05$, and $c_3 = 0.9$. We assign low weights to degree and distance similarities when computing the structural similarity. This is because, as shown in Section II, even in the UDA graph constructed based on the whole WebMD/HB dataset, (i) the degree of most of the users is low; and (ii) the size of most identified communities is small and the UDA graph is disconnected (consisting of tens of disconnected components). After partitioning the original dataset into auxiliary and anonymized data, the degree of most users gets lower and the connectivity of the UDA graph decreases further, especially in the scenario of 10%-anonymized data (the anonymized UDA graph consists of hundreds of disconnected components in our experiments). Thus, intuitively, the degree and distance (vector) do not provide much useful information in distinguishing different users for the two leveraged datasets here, and we assign low weights to degree and distance similarities. Furthermore, we set the number of landmark users as $\hbar = 50$ (the Top-50 users with respect to degree). For the structural similarity based Top-K candidate selection, we employ the *direct selection approach*. Since we conduct closed-world evaluation in this subsection, the filtering process is omitted. All the experiments are run 10 times. The results are the average of those 10 runs.

Results. We show the CDF of successful Top-K DA with respect to different K ranges ($K \in [1, 50]$, $K \in [1, 100]$, $K \in [1, 500]$, and $K \in [1, 1000]$) in Fig.2. We have the following observations.

First, with the increase of K, the CDF of successful Top-K DA increases. The reason is evident. When K increases, the probability of including the true mapping of an anonymized user to its Top-K candidate set also increases. Second, when comparing the Top-K DA performance of De-Health on WebMD and HB, De-Health has a better performance on WebMD than that on HB. This is due to the fact that the HB dataset (388,398 users) has many more users than the WebMD dataset (89,393 users), and thus with a higher probability, the correct Top-K candidate set can be found for a WebMD user under the same experimental setting. Third, the size of the available dataset (either the auxiliary data or the anonymized data) is important to constructing the UDA graph and thus has an explicit impact on the Top-K DA performance. This is because in the 90%-auxiliary data scenario, only 10% of the original dataset serves as the anonymized data. Then, only a very sparse anonymized UDA graph that consists of hundreds of disconnected components can be constructed. Thus, the Top-K DA performance has been clearly degraded.

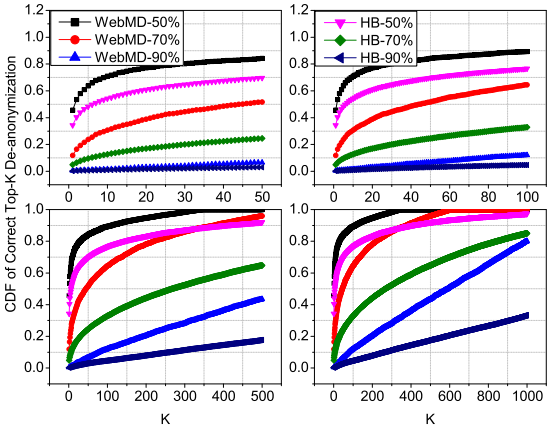


Fig. 2. CDF of correct Top-K DA.

Overall, De-Health is powerful in conducting Top-K DA on large-scale datasets (especially, when sufficient data appear in the auxiliary/anonymized data). By seeking each anonymized user Top-K candidate set, it decreases the DA space for a user from 100K-level to 100-level with high accuracy. This is further very meaningful for the following up refined DA, which enables the development of an effective machine learning based classifier.

2) Refined DA: We have demonstrated the effectiveness of the Top-K DA of De-Health on large-scale datasets. Now, we evaluate the refined DA phase of De-Health. As we indicated in Section III, the refined DA can be implemented by training a classifier employing existing benchmark machine learning techniques similar to those in [24]- [30]. In addition, more than 96.6% (resp., 98.2%) WebMD users and more than 92.2% (resp., 95.6%) HB users have less than 20 (resp., 40) posts, and the average length of those posts is short (the average lengths for WebMD posts and HB posts are 127.59 words and 147.24 words, respectively). Therefore, to enable the application of

[5]In the closed-world author attribution setting, state-of-the-art machine learning based stylometric approaches can achieve $\sim 80\%$ accuracy on 100-level of users [24], $\sim 30\%$ accuracy on 10K-level of users [25], and $\sim 20\%$ accuracy on 100K-level of users [26].

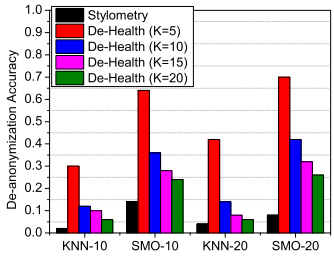


Fig. 3. DA accuracy (closed-world).

machine learning techniques to train a meaningful classifier[6] and to minimize the ethical concern, we conduct this group of evaluation on small-scale datasets extracted from the WebMD dataset, which is actually sufficient to show the performance of De-Health.

Methodology and Settings. We construct the auxiliary (training) and anonymized (testing) data for two evaluation settings. First setting: we randomly select 50 users each with 20 posts. Then, for the posts of each user, we take 10 for training and the other 10 for testing. Second setting: we randomly select 50 users each with 40 posts. Then, we take 20 posts from each user for training and the remaining for testing. Each setting is run 10 times. The results are the average of the 10 runs.

For the parameters in De-Health, the default settings are: $c_1 = 0.05$, $c_2 = 0.05$, $c_3 = 0.9$ (the reason is the same as before), $\hbar = 5$, $\epsilon = 0.01$, and $l = 10$; the employed Top-K candidate set selection approach is *direct selection*. In the refined DA phase, the employed machine learning techniques for training the classifier are the k-Nearest Neighbors (KNN) algorithm [26] and the Sequential Minimal Optimization (SMO) Support Vector Machine [27]. Note that, our settings and evaluations can be extended to other machine learning techniques directly. The features used to train the classifier are the stylometric features and structural features extracted from the auxiliary data (as defined in Section II).

We also compare De-Health with a DA method that is similar to traditional stylometric approaches [24]- [32]: leveraging the same feature set as in De-Health, training a classifier using KNN and SMO without of our Top-K DA phase, and employing the classifier for DA. We denote this comparison method as *Stylometry* (although we included correlation features in addition to stylometric features). Actually, Stylometry is equivalent to the second phase (refined DA) of De-Health.

Results. Let Y be the number of anonymized users that have true mappings in Δ_2 and Y_c be the number of anonymized users that have true mappings in Δ_2 and are successfully de-anonymized by algorithm $\mathcal{A}$. Then, the *accuracy* of $\mathcal{A}$ is defined as Y_c/Y.

We demonstrate the DA accuracy of De-Health and Stylometry in Fig.3, where $K = 5, 10, 15, 50$ indicate the setting of Top-K DA in De-Health, and '-10' (e.g., SMO-10) and '-20' (e.g., SMO-20) represent the evaluation settings with 10 and 20 posts of each user for training/testing, respectively. From the results, SMO has a better performance than KNN with respect to de-anonymizing the employed WebMD datasets.

De-Health significantly outperforms Stylometry, e.g., in the setting of SMO-20, De-Health ($K = 5$) successfully de-anonyimzes 70% users (with accuracy of 70%) while Stylometry only successfully de-anonymizes 8% users: (i) for Stylometry, given 20 (resp., 10) posts and the average length of WebMD posts is 127.59, the training data is 2551.8 (resp., 1275.9) words on average, which might be insufficient for training an effective classifier to de-anonymize an anonymized user; and (ii) as expected, this demonstrates that De-Health's Top-K DA phase is very effective, which can clearly reduce the DA space (from 50 to 5) with a satisfying successful Top-K DA rate (consistent with the results in Top-K DA).

Interestingly, De-Health has better accuracy for a smaller K than for a larger K. Although a large K implies a high successful Top-K DA rate, it cannot guarantee a better refined (precise) DA accuracy in the second phase, especially when the training data for the second phase (same to Stylometry) are insufficient. On the other hand, a smaller K is more likely to induce a better DA performance since it reduces more of the possible DA space. Therefore, *when less data are available for training, the Top-K DA phase is more likely to dominate the overall DA performance*.

B. Open-world DA

1) Top-K DA: We start the open-world evaluation from examining the effectiveness of the Top-K DA of De-Health.

Methodology and Settings. Leveraging WebMD and HB, we construct three open-world DA scenarios under which the anonymized data and the auxiliary data have the same number of users and their overlapping user ratios are 50%, 70%, and 90%, respectively[7]. Then, we employ De-Health to examine the Top-K DA performance in each scenario with the default setting: for each overlapping user, take half of its data (posts) for training and the other half for testing; $c_1 = 0.05$, $c_2 = 0.05$, and $c_3 = 0.9$ (for the same reason as explained before); $\hbar = 50$; and for the Top-K candidate selection approach, employ *direct selection*. All the evaluations are repeated 10 times. The results are the average of those 10 runs.

Results. We show the Top-K DA performance given different K ranges ($K \in [1, 50]$, $K \in [1, 100]$, $K \in [1, 500]$, and $K \in [1, 1000]$) in Fig.4. First, similar to that in the closed-world setting, the CDF of successful Top-K DA increases with the increase of K since the true mapping of an anonymized user (if it has) is more likely to be included in its Top-K candidate set for a large K. Second, De-Health has a better Top-K DA performance when more users are shared between the anonymized data (graph) and the auxiliary data (graph). This is

[6] As indicated in [24] [26] [28] [30], when applying machine learning based stylometric approaches for author attribution, there is a minimum requirement on the number of training words, e.g., 4500 words and 7500 words, for obtaining a meaningful classifier.

[7] Let n be the number of users in WebMD/HB, and x and y be the number of overlapping and non-overlapping users in the auxiliary/anonymized dataset. Then, it is straightforward to determine x and y by solving the equations: $x + 2y = n$ and $\frac{x}{x+y} = 50\%$ (resp., 70% and 90%).

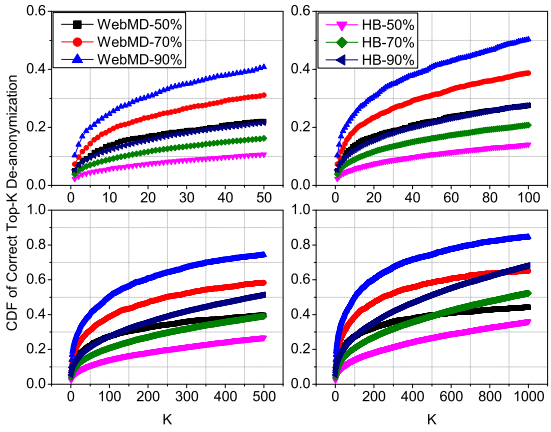


Fig. 4. CDF of correct Top-K DA (open-world).

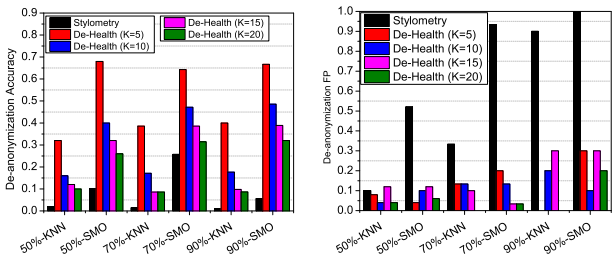
because a higher overlapping user ratio implies more common users between the anonymized and auxiliary data, followed by higher structural similarity between the anonymized and auxiliary UDA graphs. Thus, De-Health can find the correct Top-K candidate sets for more users (which are determined by the users' structural similarities). Third, when comparing closed-world (Fig.2) and open-world (Fig.4) Top-K DA, better performance can be achieved in the closed-world setting. The reason is the same as our analysis for the second observation. Finally, under the open-world setting, De-Health can still achieve a satisfying Top-K DA performance (compared to the closed-world setting, a larger K, e.g., $K = 1500$, might be necessary), and thus significantly reduces the possible DA space for an anonymized user.

2) Refined DA: Following the Top-K DA, we evaluate the refined DA performance of De-Health in the open-world setting. Due to the same reason as analyzed before and meanwhile to minimize the ethical concern, we conduct this group of evaluations on small WebMD datasets.

Methodology and Settings. We construct an anonymized dataset and an auxiliary dataset such that (i) each dataset has 100 users and each user has 40 posts; (ii) the overlapping user ratio between the two datasets is 50%; and (iii) for each overlapping user, half of its posts appear in the anonymized data while the others appear in the auxiliary data. Taking the same approach, we construct two other pairs of anonymized datasets and auxiliary datasets except for with overlapping user ratios of 70% and 90%, respectively.

For De-Health, its default settings are: $c_1 = 0.05$, $c_2 = 0.05$, and $c_3 = 0.9$; $\hbar = 5$; $\epsilon = 0.01$ and $l = 10$ for filtering; the Top-K candidate selection approach is *direct selection*; the leveraged features are the stylometric and structural features defined in Section II and the employed machine learning techniques are KNN and SMO; and after classification, we apply for the *mean-verification* scheme with $r = 0.25$. We also compare De-Health with *Stylometry* (which can be considered as equivalent to the second phase of De-Health). All the experiments are run 10 times and the results are the average of those 10 runs.

Results. We report the DA accuracy and False Positive



(a) DA accuracy (b) FP rate

Fig. 5. DA accuracy and FP rate (open-world).

(FP) rate in Fig.5, where 50%, 70%, and 90% indicate the overlapping user ratios. First, in the open-world setting, De-Health again significantly outperforms Stylometry with respect to both DA accuracy and the FP rate. For Stylometry, insufficient training data is one reason for its poor performance. In addition, in the open-world DA setting, non-overlapping users, which can be considered as noise, further degrade its performance. On the other hand, for De-Health, there are also two reasons responsible for its better performance: (i) the Top-K DA reduces the possible DA space while preserving a relatively high success rate, and thus high DA accuracy is achieved; and (ii) the *mean-verification* scheme eliminates FP DAs and thus reduces the FP rate. Second, similar to the closed-world scenario, De-Health with a smaller K has better DA accuracy (not necessary the FP rate) than that with a larger K. The reason is the same as discussed before: when less data are available for training, the Top-K DA is more likely to dominate the overall DA performance of De-Health. From the figure, we also observe that SMO-trained classifier induces better performance than KNN-trained classifier in most cases.

V. Real Identity Identification

Leveraging De-Health, an adversary can now have the health information of online health services users. On top of the DA results of De-Health, we present a *linkage attack* framework to *link those health information of the service users to real world people* in this section.

A. Linkage Attack Framework

In the linkage attack framework, we mainly conduct *username-based linkage* and *avatar-based linkage*.

Username-based Linkage. Users' usernames of most online health services are publicly available. In addition to that, there are many other social attributes that might be publicly available, e.g., gender and location of users are available on HB. In [39], Perito et al. empirically demonstrated that Internet users tend to choose a small number of correlated usernames and use them across many online services. They also developed a model to characterize the *entropy* of a given Internet username and demonstrated that a username with high (resp., low) entropy is very unlikely (resp., likely) picked by multiple users. Motivated by this fact, we implement a tool, named *NameLink*, to semi-automatically connect usernames on one online health service and other Internet services, e.g., Twitter.

NameLink works in the following manner: (i) collect the usernames of the users of an online health service; (ii) compute the entropy of the usernames using the technique in [39] and sort them in the entropy decreasing order; (iii) perform *general* and/or *targeting* online search using the sorted usernames (leveraging Selenium, which automates browsers and imitates user's mouse click, drag, scroll and many other input events). For general online searches, NameLink searches a username with/without other attributes (e.g., location) directly, e.g., "jwolf6589 + California"; for targeted searches, in addition to terms used in general search, NameLink adds a targeting Internet service, e.g., "jwolf6589 + Twitter"; and (iv) after obtaining the search results, NameLink filters unrelated results based on predefined heuristics. The main functionalities of NameLink include: (i) *information aggregation*; For instance, there is not too much information associated with WebMD users. However, there is rich information associated with HB users and BoneSmart users [42]. By linking the users on those three services, we may obtain richer information of WebMD users; (ii) *real people linkage*; For instance, for the WebMD users that have high entropy, e.g., "jwolf6589", we may try to link them to social network services, e.g., Twitter, and thus reveal their true identities; and (iii) *cross-validation*. For each user, we may link her to a real world person using multiple techniques, e.g., the username-based linkage and the following avatar-based linkage. Therefore, using the linkage results from different techniques can further enrich the obtained information as well as cross-validate the search results, and improve the linkage accuracy.

Avatar-based Linkage. Many online health services, e.g., WebMD, allow users to choose their own avatars. Thus, many users take this option by uploading an avatar without awareness of the privacy implications of their actions. However, as shown in [40], those photos may also cause serious privacy leakage. The reason behind is that a significant amount of users upload the same photo/avatar across different Internet services (websites). Similar to NameLink, we develop another semi-automatic tool, named *AvatarLink*, to link the users of one online health service to other Internet services, e.g., Facebook. AvatarLink generally follows the same working procedure as NameLink except for the search engine, which takes either an image URL or user uploaded image as a search key. AvatarLink can also fulfill the same functionalities as NameLink, i.e., information aggregation, real people linkage, and cross-validation.

B. Evaluation

We validate the linkage attack framework using the collected WebMD dataset since all its users have publicly available usernames and many of them have publicly available avatars. Note that, *the employed WebMD dataset is collected from a real world online health service* (and thus *generated by real people*). Therefore, *it might be illegal, at least improper, to employ NameLink and AvatarLink to conduct a large-scale linkage attack although we can do that. When linking the medical/health information to real world people, we only show a proof-of-concept attack and results*.

Objectives and Settings. Considering that there is not too much information associated with WebMD users, we have two objectives for our evaluation: (i) information aggregation, i.e., enrich the information of WebMD users; and (ii) link WebMD users to real world people, reveal their identities, and thus compromise their medical/health privacy.

To achieve the first objective, we employ NameLink for targeting linkage and the targeting service is HB, which has rich user information. Since we have both a WebMD dataset and a HB dataset, we limit our linkage to the users within the available datasets and thus we can do the linkage offline. Note that, this is a proof-of-concept attack and it can be extended to large-scale directly.

To achieve the second objective, we employ AvatarLink to link WebMD users to some well known social network services, e.g., Facebook, Twitter, and LinkedIn. There are 89,393 users in the WebMD dataset, which are too many for a proof-of-concept linkage attack. Thus, we filter avatars (i.e., users) according to four conditions: (i) exclude default avatars; (ii) exclude avatars depicting non-human objects, such as animals, natural scenes, and logos; (iii) exclude avatars depicting fictitious persons; and (iv) exclude avatars with only kids in the picture. Consequently, we have 2805 avatars left. When using AvatarLink to perform the linkage attack, the employed search engine is Google Reverse Image Search. In order to avoid the violation of Google's privacy and security policies, we spread the searching task of the 2805 avatars in five days (561 avatars/day) and the time interval between two continuous searches is at least 1 minute.

Results and Findings. For understanding and analyzing the results returned by NameLink and AvatarLink, a challenging task is to validate their accuracy. To guarantee the preciseness as much as possible, we **manually** validate all the results and only preserve the ones with high confidence. Specifically, for the results returned by NameLink, in addition to using the technique in [39] to filter out results with low entropy usernames, we manually compare the users' posts on two websites with respect to writing style and semantics, as well as the users' activity pattern, e.g., post written time. Interestingly, many linked users post the same description of their medical conditions on both websites. For the results returned by AvatarLink, we **manually** compare the person in the avatar and the person in the found picture, and only results in which we are confident are preserved.

Finally, using NameLink, we successfully link 1676 WebMD users to HB users and thus, those users' medical records and other associated information can be combined to provide us (or adversaries) more complete knowledge about them. Using AvatarLink, we successfully link 347 WebMD users to real world people through well known social network services (e.g., Facebook, Twitter, LinkedIn, and Google+), which consists 12.4% of the 2805 target users. Among the 347 WebMD users, more than 33.4% can be linked to two or more social network services, and leveraging the Whitepage service

[41], detailed social profiles of most users can be obtained. More interestingly, the WebMD users linked to HB and the WebMD users linked to real people have 137 overlapping users. This implies that information aggregation and linkage attacks are powerful in compromising online health service users' privacy. Overall, we can acquire most of the 347 users' full name, medical/health information, birthdate, phone numbers, addresses, jobs, relatives, friends, co-workers, etc. Thus, those users' privacy suffers from a serious threat. For example, after observing the medical/health records of some users, we can find their sexual orientation, relationships, and related infectious diseases. More concerning, some of the users even have serious mental/psychological problems and show suicidal tendency.

VI. Discussion

De-Health: Novelty versus Limitation. As shown in the experiments (Section IV), the Top-K DA of De-Health is effective in reducing the DA space (from 100K-order of possible space to 100-order of possible space) while preserving a satisfying precision (having the true mapping of an anonymized user included into the candidate set). Further, when the training data for constructing a powerful classifier are insufficient, such DA space reduction is more helpful for De-Health to achieve a promising DA accuracy. Therefore, the Top-K DA is stable and robust. For the refined DA phase, technically, it can be implemented by existing benchmark machine learning techniques. Nevertheless, due to the benefit of the Top-K DA phase, the possible DA space is reduced by several orders of magnitude, which enables us to build an effective classifier even with insufficient training data. Therefore, the Top-K DA together with the refined DA lead to promising performance of De-Health in both closed-world and open-world scenarios.

It is important to note that we do not apply advanced anonymization techniques to the health data when evaluating the performance of De-Health. This is mainly because no feasible or dedicated anonymization technique is available for large-scale online health data, to the best of our knowledge. Actually, developing proper anonymization techniques for large-scale online health data is a challenging open problem. The challenges come from (i) the data volume is very big, e.g., WebMD has millions of users that generate millions to billions of health/medical posts every month; (ii) unlike well-structured traditional medical records, the online health data are generated by millions of different users. It is a challenging task to organize those unstructured (complex) data; and (iii) different from other kinds of data, health/medical data have sensitive and important information. A proper health data anonymization scheme should appropriately preserve the data's utility (e.g., preserve the accurate description of a disease). We take developing effective online health data anonymization techniques as a future work.

Online Health Data Privacy and Policies. Based on our analysis and experimental results (especially the results of the linkage attack), online health data privacy suffers from serious threats. Unfortunately, there is no effective solution for protecting the privacy of online health service users from either the technical perspective or the policy perspective. Therefore, our results in this paper are expected to shed light in two areas: (i) for our De-Health and linkage attack frameworks and evaluation results, they are expected to show users, data owners, researchers, and policy makers the concrete attacks and the corresponding serious privacy leakage; and (ii) for our theoretical analysis, it is expected to provide researchers and policy makers a clear understanding of the impacts that different features have on the data anonymity, and thus help facilitate them to develop effective online health data anonymization techniques and proper privacy policies.

VII. Related Work

Hospital/Structured Data Anonymization and DA. To anonymize the claims data, Emam proposed several anonymization methods based on a risk threshold [13]. Fernandes et al. developed an anonymous psychiatric case register [14]. For the scenario of statistical health information release, Gardner et al. developed SHARE [15]. To defend against the re-assembly attack, Sharma et al. proposed DAPriv [16]. In [17], Emam et al. systematically evaluated existing DA attacks to structured health data. A comprehensive survey on existing privacy-preserving structured health data publishing techniques (45+) was given in [18].

Online Health Data. In [8], Nie et al. sought to bridge the vocabulary gap between health seekers and online healthcare knowledge. Another similar effort is [9], where Luo and Tang developed iMed, an intelligent medical Web search engine. Along the line of analyzing users' behavior in searching, Cartright et al. studied the intentions and attention in exploratory health search [10] and White and Horvitz studied the onset and persistence of medical concerns in search logs [11]. Nie et al. studied automatic disease inference in [12].

Health Data Policy. In [19], Barth-Jones re-examined the 're-identification' attack of Governor William Weld's medical information. In [20], Señor et al. conducted a review of free web-accessible Personal Health Record (PHR) privacy policies. In [21], McGraw summarized concerns with the anonymization standard and methodologies under the HIPAA regulations. In [22], Hripcsak et al. summarized the ongoing gaps and challenges of health data use, stewardship, and governance, along with policy suggestions. In [23], Emam et al. analyzed the key concepts and principles for anonymizing helath data.

Stylometric Approaches. In [24], Abbasi and Chen proposed the use of stylometric analysis techniques to identify authors based on writing style. In [25], Koppel et al. studied the authorship attribution problem in the wild. Later, in [26], Narayanan et al. studied the feasibility of Internet-scale author identification. Stolerman et al. presented a Classify-Verify framework for open-world author identification [27]. In [28], Afroz et al. studied the performance of stylometric techniques when faced with authors who intentionally obfuscate their writing style or attempt to imitate that of other authors. In [29], Afroz et al. investigated stylometry-based

adapting authorship attribution. In [30], Caliskan-Islam et al. de-anonymized programmers via code stylometry. To defend against stylometry-based author attribution, McDonald et al. presented Anonymouth [31]. In [32], Brennan et al. proposed a framework for creating adversarial passages.

VIII. Conclusion

In this paper, we study the privacy of online health data. Our main conclusions are three-fold. First, we present a novel two-phase online health data DA attack, named De-Health, which can be applied to both closed-world and open-world DA settings. We also conduct the first theoretical analysis on the soundness and effectiveness of online health data DA. Second, leveraging two large real world online health datasets, we validate the performance of De-Health. Finally, we present a linkage attack framework that can link online health data to real world people and thus clearly demonstrate the vulnerability of existing online health data. Our findings have meaningful implications to researchers and policy makers in helping them understand the privacy vulnerability of online health data and develop effective anonymization techniques and proper privacy policies.

Acknowledgment

This work was partly supported by the National Key Research and Development Program of China under No. 2018YFB0804102, NSFC under No. 61772466, U1936215, and U1836202, the Zhejiang Provincial Natural Science Foundation for Distinguished Young Scholars under No. LR19F020003, the Provincial Key Research and Development Program of Zhejiang, China under No. 2019C01055, the Ant Financial Research Funding, and the Alibaba-ZJU Joint Research Institute of Frontier Technologies.

References

[1] S. Fox and M. Duggan, "Health Online 2013", *Pew Research Center, Survey*, 2013.
[2] "Online Health Research Eclipsing Patient-Doctor Conversations", *Makovsky Health and Kelton, Survey*, 2013.
[3] C. Sherman,, "Curing Medical Information Disorder", *http://searchenginewatch.com/showPage.html?page =3556491*, 2005.
[4] WebMD, *http://www.webmd.com/*.
[5] HealthBoards, *http://www.healthboards.com/*.
[6] US HIPPA, *http://www.hhs.gov/ocr/privacy/*.
[7] PatientsLikeMe, *https://www.patientslikeme.com/*.
[8] L. Nie, Y.-L. Zhao, M. Akbari, J. Shen, and T.-S. Chua, "Bridging the Vocabulary Gap between Health Seekers and Healthcare Knowledge", TKDE 2014.
[9] G. Luo and C. Tang, "On Iterative Intelligent Medical Search", *ACM SIGIR*, 2008.
[10] M.-A. Cartright, R. W. White, and E. Horvitz, "Intentions and Attention in Exploratory Health Search", *ACM SIGIR*, 2011.
[11] R. W. White and E. Horvitz, "Studies of the Onset and Persistence of Medical Concerns in Search Logs", *ACM SIGIR*, 2012.
[12] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T.-S. Chua, "Disease Inference from Health-Related Questions via Sparse Deep Learning", TKDE 2015.
[13] K. E. Emam, L. Arbuckle, G. Koru, B. Eze, L. Gaudette, E. Neri, S. Rose, and J. Howard, "De-identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset", *JMIR*, 2012.
[14] A. C. Fernandes, D. Cloete, et al., "Development and Evaluation of a De-identification Procdure for a Case Register Sourced from Mental Health Electronic Records", BMC MIDM 2013.
[15] J. Gardner, L. Xiong, Y. Xiao, J. Gao, A. R. Post, X. Jiang, and L. Ohno-Machado, "SHARE: System Design and Case Studies for Statistical Health Information Release", JAMIA 2013.
[16] R. Sharma, D. Subramanian, S. N. Srirama, "DAPriv: Decentralized Architecture for Preserving the Privacy of Medical Data", *arXiv:1410.5696*.
[17] K. E. Emam, E. Jonker, L. Arbuckle, and B. Malin, "A Systematic Review of Re-Identification Attacks on Health Data", *PloS ONE*, 2011.
[18] A. Gkoulalas-Divanis, G. Loukides, and J. Sun, "Publishing Data from Electronic Health Records while Preserving Privacy: A Survey of Algorithms", *Journal of Biomedical Informatics*, No. 50, pp. 4-19, 2014.
[19] D. C. Barth-Jones, "The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now", *http://dx.doi.org/10.2139/ssrn.2076397*, 2012.
[20] I. C. Señor, J. L. Fernández-Alemán, and A. Toval, "Are Personal Health Records Safe? A Review of Free Web-Accessible Personal Health Record Privacy Policies", *J Med Internet Res*, 2012.
[21] D. McGraw, "Building Public Trust in Uses of Health Insurance Portability and Accountability Act De-identified Data", *J Am Med Inform Assoc*, 2013.
[22] G. Hripcsak, M. Bloomrosen, P. FlatelyBrennan, et al., "Health Data Use, Stewardship, and Governance: Ongoing Gaps and Challenges: JAMIA 2014.
[23] K. E. Emam, S. Rodgers, and B. Malin, "Anonymising and Sharing Individual Patient Data", *BMJ*, 2015.
[24] A. Abbasi and H. Chen, "Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace", TIS 2008.
[25] M. Koppel, J. Schler, and E. Bonchek-Dokow, "Authorship Attribution in the Wild", LRE 2011.
[26] A. Narayanan, H. Paskov, N. Gong, J. Bethencourt, E. Stefanov, E. C. R. Shin, and D. Song, "On the Feasibility of Internet-Scale Author Identification", *IEEE S&P*, 2012.
[27] A. Stolerman, R. Overdorf, S. Afroz, and R. Greenstadt, "Classify, but Verify: Breaking the Closed-World Assumption in Stylometric Authorship Attribution", IFIP WG 11.9 ICDF 2014.
[28] S. Afroz, M Brennan, and R. Greenstadt, "Detecting Hoaxes, Frauds, and Deception in Writing Style Online", *IEEE S&P*, 2012.
[29] S. Afroz, A. Caliskan-Islam, A. Stolerman, R. Greenstadt, and D. McCoy, "Doppelgänger Finder: Taking Stylometry To the Underground", *IEEE S&P*, 2014.
[30] A. Caliskan-Islam, R. Harang, A. Liu, A. Narayanan, C. Voss, F. Yamaguchi, and R. Greenstadt, "De-anonymizing Programmers via Code Stylometry", *USENIX Security*, 2015.
[31] A. W. E. McDonald, S. Afroz, A. Cliskan, A. Stolerman, and R. Greenstadt, "Use Fewer Instances of the Letter "i": Toward Writing Style Anonymization", PETS 2012.
[32] M. Brennan, S. Afroz, and R. Greenstadt, "Adversarial Stylometry: Circumventing Authorship Recognition to Preserve Privacy and Anonymity", TISS 2012.
[33] S. Ji, W. Li, M. Srivatsa, and R. Beyah, "Structural Data De-anonymization: Quantification, Practice, and Implications", *ACM CCS*, 2014.
[34] S. Ji, W. Li, P. Mittal, X. Hu, and R. Beyah, "SecGraph: A Uniform and Open-source Evaluation System for Graph Data Anonymization and De-anonymization", *USENIX Security*, 2015.
[35] S. Ji, W. Li, N. Gong, P. Mittal, and R. Beyah, "On Your Social Network De-anonymizablity: Quantification and Large Scale Evaluation with Seed Knowledge", *NDSS*, 2015.
[36] S. Ji, P. Mittal, and R. Beyah, "Graph Data Anonymization, De-anonymization Attacks, and De-anonymizability Quantification: A Survey", *COMST*, 2016.
[37] T. C. Mendenhall, "The Characteristic Curves of Composition",, *Science*, 1887.
[38] J. Noecker Jr and M. Ryan, "Distractorless Authorship Verification", LREC, 2012.
[39] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils, "How Unique and Traceable are Usernames?", PETS 2011.
[40] P. Ilia, I. Polakis, E. Athanasopoulos, F. Maggi, and S. Loannidis, "Face/Off: Preventing Privacy Leakage From Photos in Social Networks", *ACM CCS*, 2015.
[41] *http://www.whitepages.com/*.
[42] BoneSmart, *http://bonesmart.org/*.