Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

4

Data sharing and privacy issues arising with COVID-19 data and applications

Z. Müftüoğlu[1], M.A. Kızrak[1], T. Yıldırım[2]

[1]THE PRESIDENCY OF REPUBLIC OF TURKEY, THE DIGITAL TRANSFORMATION OFFICE, CANKAYA, ANKARA, TURKEY; [2]YILDIZ TECHNICAL UNIVERSITY, ESENLER, ISTANBUL, TURKEY

1. Introduction

As of February 11, 2020, the coronavirus disease 2019 (COVID-19) (2019-nCov) virus, which was declared as the new coronavirus by the World Health Organization (WHO), has spread to the whole world in a short time [1]. Coronaviruses are pleomorphic RNA viruses typically containing crown-shaped peplomers that are 80−160 nM in size and 27−32 kb positive polarity [2]. Coronaviruses are zoonotic pathogens that are present in humans and various animals, with their high mutation ratio. The coronavirus infection presents with a broad range of clinical features from asymptomatic course to requirement of hospitalization in the intensive-care unit [2]. This virus often causes serious complications with acute respiratory problems and causes death. According to WHO's Situation report, the COVID-19 virus continues to spread around the world and 2,719,897 confirmed cases were documented via case reporting forms received from 113 countries as of April 25 [1]. There is a large literature examining the adaptation of the virus to natural changes, its biological structure, its spreading and contagion structure, and prevention methods. Countries are developing various strategies against the epidemic threat. The main ones are to increase social distance, to provide support products to health units by using technology, and to develop applications to reduce the spread and economic damage of the epidemic by making use of data. Many countries aim to achieve fast and efficient results by collaborating on project calls for vaccine and drug development. There is a mobilization of physical measures taken to slow the spread of the outbreak as well. Masks are produced voluntarily using the 3D printer technology. There are also studies on the course of the epidemic and human behavior by following social media flows. By making use of radiologic images, early diagnostic studies supported by artificial intelligence feed the literature day by day. Using online tools more effectively has become widespread for patient assistance. The world is fighting against the epidemic

Data Science for COVID-19. https://doi.org/10.1016/B978-0-323-90769-9.00003-7

and its effects by using big data and related technology effectively. Many researchers are working on COVID-19 datasets, and also some specific web or mobile applications related to coronavirus all over the world has begun coming to the front.

At this point, privacy challenges arising with COVID-19 data and applications become very important. In an effort to manage the impact of the COVID-19 outbreak, technologic solutions may be collecting information from individuals that would not typically be collected. It is seen that the existing applications developed within the framework of these expectations contain absolute location information (direct), relative location information (indirect), and characteristic data defining people. Even if these data mean a lot to the world's struggle with COVID-19, it is necessary to foresee the risks that may occur after the epidemic when the relations of the information are considered. In order to measure the privacy risk of this kind of applications containing personal data, privacy metrics have been defined in the literature [27].

In this chapter, a perspective is given about the sharing and privacy of medical data within the scope of COVID-19. In this context, privacy models, metrics, and approaches for selecting the appropriate model are described, in particular for COVID-19 applications, and a new metric is proposed with the entropy approach to metrics defined in the literature and effective in determining the privacy score. Finally, the discussion and vision for the future are drawn at the level of the proposed privacy approach.

In this research, an approach is proposed that determines the level of privacy of personal data used in mobile and web-based applications developed in the scope of fighting against the lethal epidemic that emerged at the end of 2019 and spread around the world in the first half of 2020. In the second section, the accelerator studies conducted within the scope of COVID-19 are summarized. In the third part, a perspective is given about the sharing and privacy of medical data. In the fourth section, privacy models, metrics, and approaches for choosing the appropriate model are described, specifically for COVID-19 applications. In the fifth section, the discussion and vision for the future are drawn on the proposed level of privacy approach.

2. The process of accelerating COVID-19 research

To speed up the COVID-19 studies, data on diagnosis, recovery, and deceased are published day by day and competitions related to datasets are organized. Genome data for tracking pathogen evolution, chest radiographic imaging, maintaining social distance from geospatial data, follow-up of scientific research, and streams on twitter are the example of shared data types. Some of these data published in order to speed up the research within the scope of combating the COVID-19 outbreak are shown in Table 4.1.

The General Data Protection Regulation (GDPR) allows public health authorities and employers to process personal data in the condition of an epidemic, by national law, and within the conditions set therein [11]. In addition to vaccine development studies, it is aimed to use artificial intelligence solutions in order to estimate the spread and direction of the epidemic with medical data, social media data, and location data obtained from personal devices. However, the fact that mass surveillance and follow-up by using

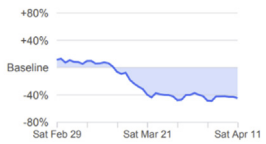
Table 4.1 Open datasets and their contents published for COVID-19.

Name of dataset	About properties of the data	Publisher
UNCOVER COVID-19 Challenge [3]	It consists of a compiled collection of more than 200 COVID-19 related datasets from public sources such as Johns Hopkins University, WHO, World Bank, *New York Times*, and others. It includes data on potentially strong statistics and indicators such as local and national infection rates, global social distance policies, geospatial data on people's movement, and more.	Roche Data Science Coalition, Kaggle
COVID19 Global Forecasting [4]	It contains data for estimating the number of cumulative confirmed COVID-19 cases around the world and the number of deaths occurring for future dates.	Kaggle
Novel CoronaVirus 2019 dataset [5]	Johns Hopkins University has an excellent dashboard using affected case data. The data has been removed from the associated Google pages and is available here.	SRK, H2O.ai, Data Science, Kaggle
COVID-19 Open Research Dataset Challenge [6]	A coalition of the White House and leading research groups has been prepared for the COVID-19 Open Research Dataset (CORD-19). It is the source of over 47,000 scientific articles containing more than 36,000 full texts on CORD-19, COVID-19, SARS-CoV-2, and related coronaviruses.	Allen Institute for AI, Kaggle
Coronavirus (COVID-19) data in the United States [7]	Over time, the *New York Times* publishes a series of data files in the United States containing a cumulative number of coronavirus cases at the state and county levels. This time series data was compiled from state and local governments and health departments to provide a complete record of the ongoing outbreak. Data begins with the first case of coronavirus reported in the Washington state on January 21, 2020. It is updated regularly for the data in this repository.	*New York Times*, GitHub
COVID chest X-ray dataset [8]	It is a database of COVID-19 cases with chest X-ray or CT images. It is published for COVID-19 cases, as well as MERS, SARS, and ARDS.	GitHub
COVID-19 CT segmentation dataset [9]	It is a dataset of 100 axial CT images from ~60 COVID-19 patients converted from open-access JPG images.	MedSeg
CoronaVirus (COVID-19) Tweets dataset [10]	Contains CSV files containing Tweet IDs. The tweets were collected by the LSTM model placed here on sentiment.live. The model tracks real-time Twitter broadcasts for tweets about coronavirus.	IEEE Data Port

ARDS, acute respiratory distress syndrome; *COVID-19*, coronavirus disease 2019; *CT*, computed tomography; *MERS*, Middle East respiratory syndrome; *SARS-CoV-2*, severe acute respiratory syndrome coronavirus-2.

relative and location data within the scope of this struggle is possible within the framework of the stretched rules may create an issue where social and national security is discussed. According to mobility changes, comparison between the countries where the epidemic is intense and Turkey is depicted in Table 4.2. Here are mobility trends for

Table 4.2 Mobility changes during the COVID-19 pandemic [12].[1]

The United States	Italy
Retail & recreation -45% compared to baseline 	Retail & recreation -86% compared to baseline 
Spain	**Turkey**
Retail & recreation -92% compared to baseline 	Retail & recreation -92% compared to baseline 

places like restaurants, cafes, shopping centers, theme parks, museums, libraries, and movie theaters.

3. Medical data and sharing

With the rapid advancement of technology, large amounts of collected data provide many advantages in different areas of our daily life, with the inclusion of finance, medicine, and industry [13,14]. Current technologic advances in biomedical and health research also dramatically increase digital data production, enabling large amounts of data in a variety of disciplines ranging from finance to medicine [13]. Big data in the field of health can improve the clinical decision-making process and patient care as well as increase the statistical power of clinical research studies by obtaining more accurate results and strong prediction models [13,15]. The stunning rising in the rapidity of the data gathering process and large amounts of data from distributed data sources enable scientific innovation, exclusively in healthcare. The kinds of data used in healthcare range from bio-signals to medical imaging and laboratory tests to omics data. Bio-signals are made by electric activity resulting from the biological function of organs in the human body. Medical images contain another type of medical data that is of great importance in clinical diagnosis and imaging procedures [4]. The domain of omics amount to a large number of areas, with each one having a specific individual clinical significance for the mentality of the underlying contraption behind the cellular

[1]Updated 11th April 2020.

interactions. Omics areas constitute a large area of medical data with many subfields such as genomics, lipidomics, proteomics, metabolomics, microbiomics, epigenomics, and transcriptomics domains [16]. Medical laboratory tests, on the other hand, can provide a strong base of mentality for the underlying contraption of a virus and to detect various pathologic conditions in tissue samples. Hematologic tests, serologic tests, skin tests, histopathologic tests, immunologic tests, urine tests, endocrine function tests, and coagulation tests are accepted as the most widespread laboratory tests [13,16].

3.1 Medical data acquiring

Medical data collection should generally be carried out considering international standards and protocols for each sort of data [16]. With increasing numbers of data produced daily with health sensors, medical imaging data, laboratory test data, electronic patient records, analog patient records (PR), and clinical and pharmaceutical invoices, the estimated data amount is expected to be in the range of yottabytes (10^{24} gigabytes) [17]. The increased amount of data also poses a threat to the control of privacy risk.

3.2 Medical data sharing

Medical data sharing includes mechanisms to protect the rights and privacy of patients. This kind of work contains the essence of a federal platform, as it provides the interconnection of health studies worldwide. Data abuse and fear of losing control of the data are a major obstacle to sharing data. Confident data management and data identification are imperative to ensure the sharing of sensitive data while respecting privacy [16].

A data-sharing framework consists of two main functions [18]:

I. The data source and acquisition processes should comply with the guidelines stipulated by the relevant data protection regulations.

II. The quality and accuracy of medical data should be assessed, considering current clinical field information and relevant public health policies. This function is often called as data governance and is associated with the following concepts:
   - **a.** the assessment of data quality,
   - **b.** the examination of the data organizational architecture,
   - **c.** the generic information management.

The data sharing framework is the phase before the improvement and application of federated data analysis services. An organization that desires to move its medical data to a federal platform must complete all necessary ethical and legal processes before any data operation. The data protection regulations (GDPR in Europe, HIPAA in the

United States, KVKK in Turkey) should be the basis for these documents and include the following points [16]:

a. the certain definition for legitimate interests,
b. full data protection impact assessments,
c. certain goals of the processing,
d. approved forms for the processing of personal data from the data owner,
e. purpose of transporting to third parties,
f. guarantees of data protection.

Different data sharing attempts have been started for the unity of clinical trial data [19,20]. These attempts intend to supply frameworks and guidelines for sharing health and related data. To encourage research in the medical field worldwide, the initiatives focalize on the transparency of data collection protocols and patient deidentification [20] processes [16]. However, central patient databases are usually tend to data delicts and are sometimes not in rapport with data protection regulations [19].

Fig. 4.1 presents the 10 elements necessary for trustworthy data sharing. The main data sharing initiatives whose aim is to provide out of border data sharing to encourage science in the clinical domain are given in Table 4.3.

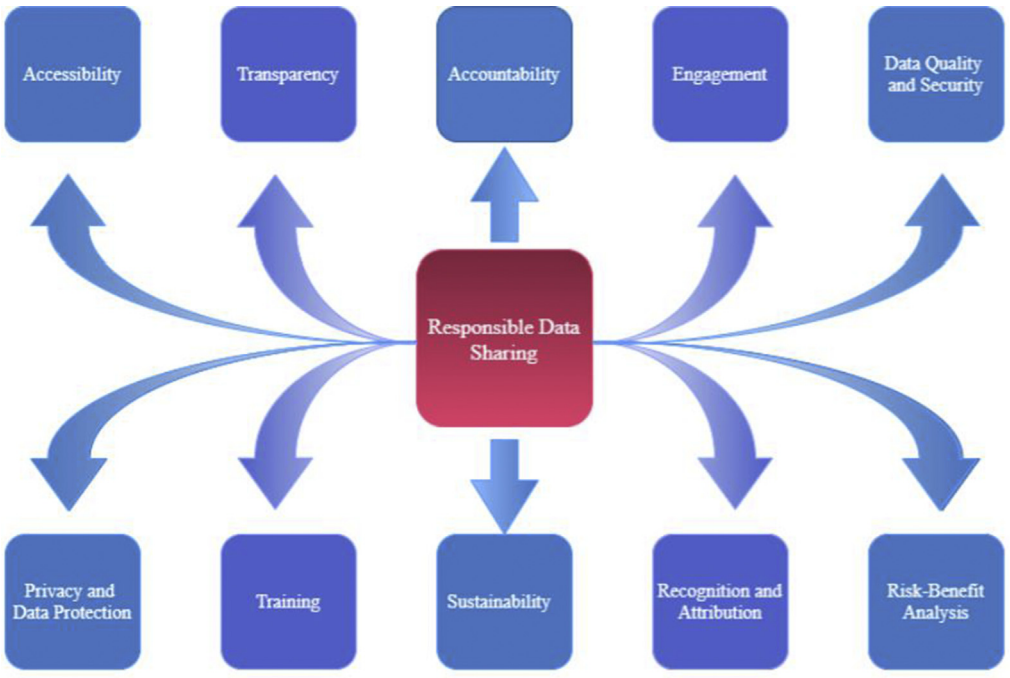


FIGURE 4.1 The10 essentials for reliable data sharing [22].

Table 4.3 Global initiatives in the clinical domain.

Name of the initiative	Explanation
ClinicalTrials.gov	Its purpose is to make public and private clinical research and studies active for the scientific community to support multidisciplinary science [16–23].
The Data Sphere Project	It aims to make cancer clinical trials available to researchers [24].
The database of Genotypes and Phenotypes (dbGaP)	It was established to archive the scientific results of genotype and phenotype studies, in an attempt to demonstrate the relationships between them, as well as phenotype, exposure, genotype, and sequence data at the individual level.
Biogrid	It is an advanced data sharing platform that currently runs on different sites in Australia [25].
NEWMEDS	By aiming to explore new therapeutic treatments for schizophrenia and depression, it gathers clinical researchers [26].
The Query Health initiative	It is a new example of a safe, allocated system for data sharing and allocated inquiry processing intended at combining different heterogeneous clinical systems to a learning health system [20].

4. COVID-19 applications and privacy

Countries are developing various strategies against the epidemic threat. The main ones are to increase social distance, providing support products for health units by using technology and developing practices to reduce the spread and economic harm of the epidemic by utilizing data.

While individuals expect violence of the epidemic to be mapped on location basis from these applications, health institutions are also expecting solutions for the identification of people in the risk group and rapid identification of those who are in contact with COVID-19-positive individuals. It is very important that the technology solutions developed during this period comply with the regulations on data privacy. It has appeared that the applications developed in the scope of these expectations include absolute location, relative location, and personal data of users. Although these data provide great benefits in the fight against COVID-19, it is also necessary to foresee the risks that may occur after the epidemic when the relationship of metrics with each other is considered. When the applications developed in this context are examined, it is observed that all of them track ground position and some of them do not have any kind of conditional electronic document such as Terms and Conditions, which provides information to user about what his/her rights are, what kind of data are used, and whether his/her data are shared with third parties or not. Table 4.4 shows the most privacy-invasive COVID-19 applications.[2]

[2]https://surfshark.com/blog/privacy-invasive-covid-19-apps. Updated: 26.04.2020.

Table 4.4 The most privacy-invasive COVID-19 apps.[3]

Country	Name	Type of measure
Spain (Madrid)	Corona Madrid	Digital tracking
Singapore	Trace Together	Digital tracking
Poland	Home Quarantine	Digital tracking
China	Hangzhou Health Code app	Digital tracking
Hong Kong	Electronic wristbands synced with the app	Digital tracking
South Korea	Corona 100m	Digital tracking
Israel	Track Virus	Digital tracking
Thailand	IoT app with sim cards	Digital tracking
Colombia	CoronApp-Colombia	Digital tracking
The United Kingdom	TBC	Digital tracking
Belgium	TBC	Digital tracking
Iran	AC19	Digital tracking

4.1 Privacy metrics

The purpose of the privacy metrics is to compute the level of privacy of users in a system and the amount of guard presented by technologies that increase privacy. The technical privacy metric takes the features of a system as input and achieves a numerical value. This value allows both to achieve the level of privacy and to compare it with other privacy-protecting methods [27].

Privacy-enhancing technologies (PETs) protect privacy based on technology rather than policy, thereby providing much stronger protection. Privacy measures to evaluate the effectiveness of PETs are technical privacy criteria, which can compute the level of privacy in a system or the privacy maintained by a specific PET, take the features of a system as input (for instance, the quantity of sensitive information disclosed or the number of users that are indistinctive by some features), and give a numerical (or sometimes rule-based) value.

Although there are many metrics in the literature, a structured and comprehensive overview of privacy metrics is not yet available. Hence it is hard to take conscious decisions about which metrics to choose for the evaluation of PETs. Inefficient PETs can be chosen, considering the prevalence of systems that could violate privacy [33]. Here, the landscape of privacy metrics is structured by focusing on technical metrics that measure the degree of privacy in a system or the effectiveness of PETs.

Fig. 4.2 presents a summary of the outputs and the related metrics. Even if there are many possible classifications for metrics, an output-based classification is the most intuitive. The borders between categories might not be clear, and as with any classification, some metrics can be appointed to other categories. The metrics from the data similarity class in terms of metrics from the uncertainty and information gain/loss

[3] https://docs.google.com/spreadsheets/d/1shGO0YJKJPOf0jvTZPsmR_iLhO6-tUOOlfgG8Fg-q_I/edit#gid=0. Updated: 26.04.2020.

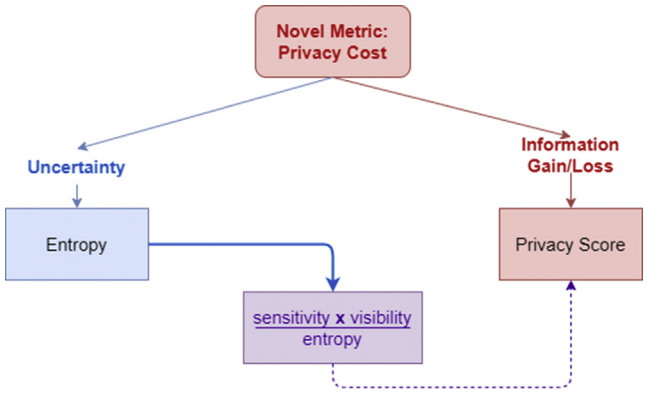


FIGURE 4.2 The proposed metric.

categories are defined in Ref. [31], and the data similarity metrics can be concerned with metrics from the indistinguishability category [32].

In the literature, domains that frequently use PETs are

a. communication systems,
b. databases,
c. location-based services,
d. smart metering,
e. social networks
f. genome privacy.

Besides, four basic characteristics are defined for these metrics, which are the factors that determine the level of privacy in the relevant domain in the literature:

a. Adversary Goals,
b. Adversary Capabilities,
c. Data Sources,
d. Inputs for Computation of Metrics.

Adversary Goals: The purpose of the adversary is achieving sensitive information and risking users' privacy. User identities, user features, or both can be called as sensitive information [30].

Adversary Capabilities: Knowledge and skill of the attacker in the field is an important parameter to damage the privacy more quickly and successfully.

Data Sources: Data sources define which data needs to be protected and how the adversary is supposed to access the data. Data is considered to be publicly available and/or misused.

Inputs for Calculation of Metrics: Privacy metrics depend on a dissimilar sort of input data to calculate privacy values. Usability of a metric in a specific scenario is assigned according to the efficacy of input data or convenient assumptions.

Output Measure: It is the most important characteristic in determining privacy metrics. Wagner and Eckhoff [27] introduce the taxonomy with eight output properties. Each of them is associated with related metrics. In Table 4.5, these output measures are given.

Table 4.5 Eight characteristic criteria for determining privacy metrics.

Uncertainty	As uncertainty increases, privacy increases. In this case, the attacker becomes unable to make accurate predictions.
Information gain or loss	It symbolizes the quantity of information acquired by the attacker or the quantity of privacy lost by users for the disclosure of information.
Indistinguishability	It is a classic concept used in security. It serves to analyze whether the attacker can distinguish two consequences of the privacy mechanism.
Data similarity	It serves to measure the similarity in the dataset.
Time	It is the time taken for the attacker to overcome the privacy mechanism or the duration of the attacker to hesitate. The shorter the hesitation period, the higher the privacy. The longer the time to overcome the mechanism, the higher the privacy.
Error	Error-based metrics help measure how accurate the attacker's estimation is, for instance, by using the span between the real result and the estimation. Elevated accuracy and small errors are associated with nominal privacy.
Adversary's success probability	It indicates the probability that an attacker would succeed in one or more attempts.
Accuracy/precision	It is used to determine how precise the attackers are, regardless of their accuracy. More precise estimates are associated with lower privacy.

4.2 Privacy metric selection method

Studies about the attacks against privacy head for using the metric attributes time, error, or the adversary's achievement probability, whereas studies about the new PETs incline toward the accuracy, similarity, and indistinguishability metrics. In both cases, this is an appropriate selection: while most metrics in the early group focus more on competitors, the second group highlights the PET activity presented. The "attack" and "defense" aspects can select metrics from the other. Finally, it is important to select the metrics that the enemy intends to reveal sensitive information, that is, to evaluate user identities or features, and to measure the relevant direction.

Considering the number and variety of privacy metrics, it can be difficult to select metrics for a specific scenario. To overcome this problem, Wagner et al. has prepared a set of nine questions that will help in choosing the metric. Thus suitable metrics are selected with this methodology. While determining the metrics, the following points are important:

I. against which types of attackers and which data sources are aimed to protect,
II. determining the requirements and input data to be used in metric calculations for these targets.

4.3 Proposed method: privacy cost

As can be seen in Table 4.6, there are many privacy metrics defined in the literature. In this study, the entropy approach will be used to calculate the privacy risk value. Assuming a database containing medical records of n items and N users, for each profile item, users set a privacy level that assigns their willingness to disclose information associated with this item [29].

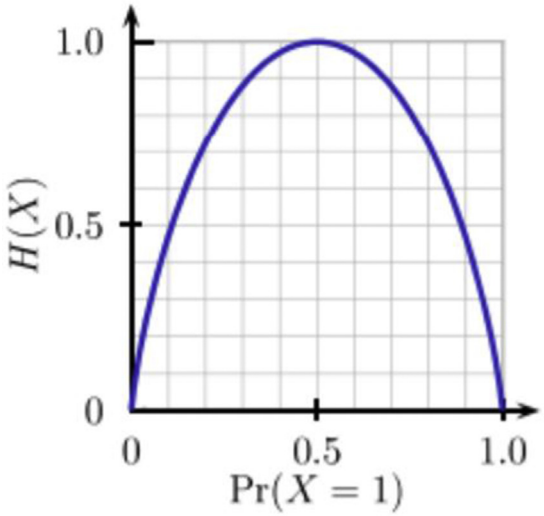


FIGURE 4.3 Shannon entropy for coin toss status in case the probability Pr $[X = 1]$ ranges from 0 to 0.1.

Many other metrics rely on the Shannon entropy. As can be seen from Fig. 4.2, many uncertainty metrics are built on entropy to measure uncertainty. In Table 4.6, it is seen that most metrics in the uncertainty category arise from the communication domain, where, for instance, they can be utilized to rate an adversary's uncertainty regarding diverse users and messages.

Table 4.6 Metrics and original domains placed in the uncertainty category [27].

Metric	Original domain
Anonymity set size	Communication
Entropy	Communication
Rényi entropy	Communication
Max-entropy (Hartley)	Communication
Min-entropy	Communication
Normalized entropy	Communication
Degree of unlikability	Communication
Quantiles on entropy	Communication
Conditional entropy	Communication
Conditional privacy	Database
Inherent privacy	Database
Cross-entropy	Database
Cumulative entropy	Location
Protection level	Location
Asymmetric entropy	Genome privacy
Genomic privacy	Genome privacy
User-centric privacy	Location

The privacy levels selected by all N users for the n profile items are stored in a response matrix R. The rows of R correspond to profile items, and the columns of R correspond to users. $R(i,u)$ refers to the privacy setting of user u for item i.

There are two definitions for the R matrix:

- If the entries of R are limited to take values in $\{0,1\}$, then R is called a dichotomous response matrix.
- If $R(i,u)$ takes any nonnegative integer values in $\{0,\ldots,l\}$, then R is called a polytomous response matrix.

In a dichotomous response matrix R, $R(i,u)=1$ means that user u has made the information about the profile item i publicly available. If $R(i,u)=0$, it means that user u has kept the item i private.

The commentary of values appearing in the polytomous response matrix is alike. $R(i,u)=0$ means that user u does not share item i with anyone, whereas $R(i,u)=k$, with $k \in \{0,1\}$, means that u disclose item i to other users that are at most k-hops away in the social graph [29].

4.3.1 Privacy algorithm

A user's privacy score [27] indicates the potential privacy risk of the user. The risk of privacy increases as long as the user discloses his/her sensitive information to the public. Maximilien et al. [29] defines the privacy risk of user u as an increasing function of two parameters monotonically: the sensitivity of the user's profile items and the visibility these items get.

The sensitivity of an item: It is a parameter that determines the damage caused by the disclosure of information. For example, in a medical record, the diagnostic information of a patient is considered much more sensitive than the patient's address. The sensitivity of the i item will be depicted as Δ_i.

Visibility of an item: The visibility of item i due to user u captures how widely known the value of i becomes in the social network; the more it spreads, the higher the item's visibility. The visibility is denoted by $V(i,u)$. The visibility $V(i,u)$ is denoted in the following for $I_{\text{condition}}$ indicator variable that becomes 1 when "condition" is true:

$$V(i,u) = I_{(R(i,u)=1)} \tag{4.1}$$

In statistics, one can assume that R is a sample from a probability distribution over all possible response matrices. For $P_{iu} = Prob\{R(i,u)=1\}$,

$$V(i,u) = P_{iu} \times 1 + (1 - P_{iu}) \times 0 = P_{iu} \tag{4.2}$$

4.3.2 Privacy cost

Wagner et al. defined the privacy score as a privacy metric. According to this definition, a privacy score has a relationship directly proportional to sensitivity and visibility. On the other hand, Longpre et al. [28] defined entropy as the mean number of binary ("yes"-"no") questions that we require to inquire to unambiguously determine the alternative. Privacy means that we do not know everything about a person, so we need to ask extra questions

to receive the full information about the person. It, therefore, appears reasonable to measure the degree of privacy in each situation by the mean number of binary ("yes"-"no") questions that we need to inquire to determine the full information, which is accurately Shannon entropy.

While self-information measures the uncertainty of a particular event, Shannon entropy measures uncertainty in a probability distribution as can be seen from Fig. 4.3. When one does not know the result of the random experiment, it can also be seen as a measure of the average information content that the person is missing. Shortly, Shannon entropy can be seen as a measure of unpredictability or disorder [31].

In our study, the current privacy score metric is associated with another privacy metric, *Entropy*. The Shannon entropy, which forms the basis for many different metrics, measures the uncertainty of a random variable's value estimation.

Normally, we have n different alternatives, $(a_1, a_2, a_3, \ldots, a_n)$ to identify the user u. Each alternative has probability values as $p(a_1)$, $p(a_2)$, …, $p(a_n)$. All these probabilities will be equal to 1.

$$\sum_{i=1}^{n} p(a_i) = 1 \tag{4.3}$$

In this case, the amount of privacy is defined by entropy as follows:

$$H_0 = \sum_{i=1}^{n} p(a_i) \log_2(p(a_i)) \tag{4.4}$$

For this study, entropy can be described as the average number of binary ("yes"-"no") questions to disclosure the i-th item of u user.

The new approach, which we call as privacy cost, is expressed as follows:

$$\text{priv}_{PC} = \frac{\Delta_i \times V(i,u)}{H(u)} \tag{4.5}$$

From the expression, it is observed that, as an acceptable way of defining the loss of privacy, the low entropy value increases the risk of privacy.

5. Discussion and suggestions for further research

In this study, we propose a new privacy metric approach by presenting a perspective on the sharing and privacy medical data within the scope of COVID-19. Many researchers have been working on COVID-19 datasets and developing web or mobile applications aiming to control the spread of coronavirus all over the world. The developed products need to keep absolute position, relative position, and some personal data in order to meet the expectation. This process, which operates fast due to a vital situation, brings with it some risks of privacy. Developers use various methods to protect privacy in their products, which are called as PETs. To evaluate the effectiveness of a PET, privacy criteria are needed that can measure the level of privacy. A privacy metric is calculated by taking features of a system as input, such as the number of users who have indistinguishable characteristics or the quantity of sensitive information. At the end of

calculation a numeric value is obtained that provides the privacy level. This value also can be used as a comparative parameter for other PETs. This study proposes a new privacy metric with the Shannon entropy approach called privacy cost using metrics defined in the literature [27]. Entropy is also a privacy metric, which is the base for many other metrics, that measures uncertainty. Therefore the expression of privacy cost demonstrates that the low entropy value means high privacy risk.

As a further work, the relationship between privacy metrics can be tackled. This is important to develop more effective PETs. On the other hand, although there are many metrics in the literature, a structured and exhaustive overview of privacy metrics is not yet available. This makes it hard to take conscious decisions about which metrics to choose for the evaluation of PETs. Given the prevalence of systems that could violate privacy, developing methods for the selection of suitable metrics provide an efficient PET choice.

References

[1] S. Salehi, A. Abedi, et al., Coronavirus disease 2019 (COVID-19): a systematic review of imaging findings in 919 patients, Am. J. Roentgenol. (2020). Available from: https://www.ajronline.org/doi/full/10.2214/AJR.20.23034.

[2] S.N. Mali, A.P. Pratap, B.R. Thorat, The rise of new coronavirus infection-(COVID-19): a recent update, Eurasian J. Med. Oncol. (EJMO) 4 (1) (2020) 35–41.

[3] UNCOVER COVID-19 Challenge, United Network for COVID Data Exploration and Research, Kaggle, 2020. Available from: https://www.kaggle.com/roche-data-science-coalition/uncover.

[4] COVID19 Global Forecasting (Week1) Forecast Daily COVID-19 Spread in Regions Around World, Kaggle, 2020. Available from: https://www.kaggle.com/c/covid19-global-forecasting-week-1.

[5] Novel Coronavirus 2019 Dataset Day Level Information on Covid-19 Affected Cases, Kaggle, 2020. Available from: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset.

[6] COVID-19 Open Research Dataset Challenge (CORD-19) an AI Challenge With AI2, CZI, MSR, Georgetown, NIH & the White House, Kaggle, 2020. Available from: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge.

[7] Coronavirus (Covid-19) Data in the United States, NYTimes, GitHub, 2020. Available from: https://github.com/nytimes/covid-19-data.

[8] COVID Chest X Ray Dataset, IEEE8023, GitHub, 2020. Available from: https://github.com/ieee8023/covid-chestxray-dataset.

[9] COVID-19 CT Segmentation Dataset, MedSeg, 2020. Available from: http://medicalsegmentation.com/covid19/.

[10] R. Lamsal, Coronavirus (COVID-19) Tweets Dataset, IEE Data-Port, 2020. Available from: https://ieee-dataport.org/open-access/corona-virus-covid-19-tweets-dataset.

[11] European Data Protection Board (EDPB), Statement on the Processing of Personal Data in the Context of the COVID-19 Outbreak. Adopted on 19 March 2020, Available from: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_statement_2020_processingpersonaldataandcovid-19_en.pdf.

[12] Google COVID-19 Community Mobility Reports, 2020. Available from: https://www.google.com/covid19/mobility/.

[13] J.W. Song, K.C. Chung, Observational studies: cohort and case-control studies, Plast. Reconstr. Surg. 126 (6) (2010) 2234e42.

[14] H.M. Krumholz, Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system, Health Aff. 33 (7) (2014) 1163e70.

[15] D.W. Bates, S. Saria, et al., Big data in health care: using analytics to identify and manage high-risk and high-cost patients, Health Aff. 33 (7) (2014) 1123e31.

[16] V.C. Pezoulas, T.P. Exarchos, D.I. Fotiadis, Medical Data Sharing, Harmonization and Analytics, Academic Press, 2020, ISBN 9780128165591 eBook.

[17] C.S. Greene, J. Tan, et al., Big data bioinformatics, J. Cell. Physiol. 229 (12) (2014) 1896e900.

[18] V. Khatri, C.V. Brown, Designing data governance, Commun. ACM 53 (1) (2010) 148e52.

[19] A.S. Downey, S. Olson, Sharing Clinical Research Data: Workshop Summary, National Academies Press (US), Washington (DC), 2013.

[20] J.G. Klann, M.D. Buck, et al., Query health: standards-based, cross-platform population health surveillance, J. Am. Med. Inf. Assoc. 21 (4) (2014) 650e6.

[21] B.M. Knoppers, Framework for responsible sharing of genomic and health-related data, HUGO J. 8 (1) (2014) 3.

[22] US National Institutes of Health, ClinicalTrials.Gov., 2012. Available from: https://clinicaltrials.gov/.

[23] A.K. Green, K.E. Reeder-Hayes, et al., The project data sphere initiative: accelerating cancer research by sharing data, Oncologist 20 (5) (2015). 464-e20.

[24] M.D. Mailman, M. Feolo, et al., The NCBI dbGaP database of genotypes and phenotypes, Nat. Genet. 39 (10) (2007) 1181.

[25] R.B. Merriel, P. Gibbs, et al., Australia facilitates collaborative medical and bioinformatics research across hospitals and medical research institutes by linking data from disease and data types, Hum. Mutat. 32 (5) (2011) 517e25.

[26] K.E. Tansey, M. Guipponi, et al., Genetic predictors of response to serotonergic and noradrenergic antidepressants in major depressive disorder: a genome-wide analysis of individual-level data and a meta-analysis, PLoS Med. 9 (10) (2012) e1001326.

[27] I. Wagner, D. Eckhoff, Technical privacy metrics: a systematic survey, technical privacy metrics: a systematic survey, ACM Comput. Surv. 51 (3) (2018) 45. https://doi.org/10.1145/3168389. Article 57.

[28] L. Longpre, V. Kreinovich, T. Dumrongpokaphan, Entropy as a Measure of Average Loss of Privacy, University of Texas at El Paso Digital Commons@UTEP, 2017. Available from: https://pdfs.semanticscholar.org/10bd/406c549a159a73e7554f026e7c998ad874cf.pdf.

[29] E.M. Maximilien, T. Grandison, et al., Privacy-as-a-Service: Models, Algorithms, and Results on the Facebook Platform, 2017. Available from: https://pdfs.semanticscholar.org/a230/5ab835252e5bb8c050ea4a0dedc561122326.pdf.

[30] J. Heurix, P. Zimmermann, et al., A taxonomy for privacy enhancing technologies, Comput. Secur. 53 (2015) 1−17.

[31] M. Bezzi, An information theoretic approach for privacy metrics, Trans. Data Privacy 3 (3) (2010) 199−215.

[32] J. Soria-Comas, J. Domingo-Ferrer, Differential privacy via t-closeness in data publishing, in: Proc. 11th Annu. Conf. on Privacy, Security and Trust, IEEE, Tarragona, Spain, 2013, pp. 27−35.

[33] D. Eckhoff, I. Wagner, Privacy in the smart city − applications, technologies, challenges and solutions, IEEE Commun. Surv. Tutor. 20 (1) (2018) 489−516.