A privacy risk assessment model for open data

Ali-Eldin, Amr; Zuiderwijk-van Eijk, A.M.G.; Janssen, Marijn

DOI
10.1007/978-3-319-78428-1_10

Publication date
2018

Document Version
Final published version

Published in
Business Modeling and Software Design - 7th International Symposium, BMSD 2017, Revised Selected Papers

A Privacy Risk Assessment Model for Open Data

Amr Ali-Eldin[1,2](✉), Anneke Zuiderwijk[3], and Marijn Janssen[3]

[1] Leiden Institute of Advanced Computer Science,
Leiden University, Leiden, The Netherlands
a.m.t.ali-eldin@liacs.leidenuniv.nl
[2] Computer and Control Systems Department, Faculty of Engineering,
Mansoura University, Mansoura, Egypt
[3] Faculty of Technology, Policy and Management,
Delft University of Technology, Delft, The Netherlands

Abstract. While the sharing of information has turned into a typical practice for governments and organizations, numerous datasets are as yet not openly published since they may violate users' privacy. The hazard on data protection infringement is a factor that regularly hinders the distribution of information and results in a push back from governments and organizations. Moreover, even published information, which may appear safe, can disregard client security because of the uncovering of users' personalities. This paper proposes a privacy risk assessment model for open data structures to break down and diminish the dangers related with the opening of data. The key components are privacy attributes of open data reflecting privacy risks versus benefits exchanges-off related with the utilization situations of the information to be open. Further, these attributes are assessed using a decision engine into a privacy risk indicator value and a privacy risk mitigation measure. Privacy risk indicator expresses the anticipated estimation of data protection dangers related with opening such information and privacy risk mitigation measure expresses the estimations that should be connected on the information to evade the expected security risks. The model is exemplified through five genuine scenarios concerning open datasets.

Keywords: Open data · Privacy risks
Personally identifiable information (PII) · Data mining · Scoring systems

1 Introduction

Governments and openly subsidized research associations are urged to unveil their information and to make this information available without limitations and for free [1]. Opening public and private information is a mind boggling movement that may bring about advantages yet may likewise experience risks [2]. An essential risk that may hinder the production of the information is that associations may abuse the privacy of citizens when opening data about them [3]. In addition, when opening data, associations lose control on who is utilizing this information and for what reason. When information are distributed, there is no power over who will download, utilize and adjust the information.

B. Shishkov (Ed.): BMSD 2017, LNBIP 309, pp. 186–201, 2018.
https://doi.org/10.1007/978-3-319-78428-1_10

To maintain a strategic distance from data protection infringement, information distributers and owners can remove delicate data from datasets, in any case, this makes datasets less helpful. Furthermore, even distributed information, which may appear security agreeable, can disregard user privacy because of leakage of genuine user personalities when different datasets and different assets are connected to each other [4]. The likelihood of mining the information subsequently to get important conclusions can prompt leakage of private information or users' real identities. In spite of the fact that organisations remove personally identifiable information (PII) from the dataset before distributing the information, a few investigations exhibit that anonymized information can be de-anonymized and thus real identities can be uncovered [4].

Different existing investigations have pointed at the dangers and difficulties of privacy infringement for distributing and utilizing open data [3–6]. A few investigations have distinguished privacy risks or approaches for organisations in gathering and preparing information [7, 8], some have given choice help to opening information as a rule [2], and some have concentrated on discharging data and information on the individual level [9]. All things considered, there is as yet constrained knowledge in how associations can lessen privacy infringement dangers for open data specifically, and there is no uniform approach for privacy assurance [5]. From existing studies, it has not turned out to be clear which open data frameworks can be utilized to lessen the hazard on open data privacy infringement. An open data design is required that helps settling on choices on opening data and that gives understanding in whether the information may abuse users privacy.

The goal of this paper is to propose a model to analyse privacy infringement risks of publishing open data. To do as such, a new arrangement of what are called open data attributes is proposed. Open data attributes reflect privacy risks versus benefits exchanges off related with the normal utilize situations of the information to be open. Further, these attributes are assessed utilizing a decision engine into a privacy risk indicator (PRI) and a privacy risk mitigation measure (PRMM). Specifically this can decide if to open data or keep it closed. This paper is organized as follows: Sect. 2 discusses related work while Sect. 3 presents privacy violation risks associated with open data. Section 4 introduces the proposed model. The model helps identifying the risks and highlights possible alternatives to reduce these risks. Section 5 highlights how the proposed model can be implemented in reality. Section 6 exemplifies the model by providing some scenarios and preliminary results. Section 7 discusses the key findings and concludes the paper.

2 Previous Work

Open bodies are viewed as the greatest makers of data in the general public in what is known as open data. Open data may extend from information on acquirement openings, climate, movement, traveller, energy utilization, crime statistics, to information about arrangements and organizations [1, 2]. Information can be arranged into various levels of secrecy, including exceedingly private, classified, confined and open. We consider open data that has no connection with information about citizens as outside the extent of this work.

Anonymized information about citizens can be shared to comprehend societal issues, for example, crimes or diseases. A case of subject information is the sharing of patient information to start joint effort among healthcare providers which is relied upon to be gainful to the patient and scientists. The profoundly expected advantages behind this information sharing is the enhanced comprehension of particular illness and subsequently considering better medications. It can likewise help professionals to become plainly more productive. For instance, a general specialist can rapidly analyse and recommend drug. However, this sharing of patients' data ought to be achieved by information security approaches and privacy controls.

An assortment of Data Protection Directives has been made and executed. In light of the Data Protection Directive [2], a thorough change of data protection rules in the European Union was proposed [3]. Additionally, the Organization for Economic Co-operation and Development (OECD) has created Privacy Principles [4], including standards, for example, "There should be limits to the collection of personal data" and "Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with Paragraph 9 except (a) with the consent of the data subject; or (b) by the authority of law." In addition, ISO/IEC 29100 standard has defined 11 principles for privacy [5].

It turns into a pattern these days that organizations put a greater amount of their consideration on the privacy issue, since information is assumed to be a central asset of any business. The Data Protection Directives are frequently characterized on high level of abstraction, and give restricted rules to making an interpretation of the directives to practice. In spite of the created Data Protection Directives and other information assurance arrangements, associations are still subject to privacy infringement when distributing open data. In the accompanying sections, we expound on the principle risks of privacy infringement related with open data.

There has been expanding enthusiasm for outlining privacy assurances into advancements from the beginning known as privacy by design (PbD). PbD is a proactive way to deal with privacy assurance that considers privacy ramifications of new advances amid the plan organize, as opposed to as a bit of hindsight [6].

Privacy awareness is increasingly being raised. A lot of privacy related investigations are really being directed, however they either concentrate on legal aspects like [7], or on conducting formal Privacy Impact Assessment (PIA) as [8]. Most work on privacy impact assessment plan to lead reviews or surveys that evaluate organizations methods for managing individual information as indicated by regulatory frameworks and moral or ethical esteems into what is known as PIA. According to [9], a PIA is a procedure which should start at the most punctual conceivable stages, when there are still chances to impact the result of a project. It is a procedure that should proceed until and even after the undertaking has been sent. A PIA has frequently been depicted as an early cautioning framework as it gives an approach to identify potential privacy issues [9].

The General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679), embraced on 27 April 2016 as a replacement of Directive 95/46/EC) [2], becomes enforceable from 25 May 2018 and does not require national governments to pass any empowering enactment [10]. The directive points fundamentally to give control back to subjects and EU occupants over their own information and to streamline the administrative condition for global business by bringing together the control inside the EU.

GDPR enforces organizations which deal with personal information of EU citizens to include privacy protection activities into the development lifecycle of software and business processes.

With regards to open data, such frameworks like to assess privacy risks cannot be utilized since the information to be distributed will contain no identifying data as a requirement by the law. Having said that, ordinary methods for assessing privacy risks cannot be applied and new ways are required that exceed the advantages of sharing the information contrasted with expected privacy risks of the leakage of personally identifiable information.

3 Privacy Threats and Opening Data

In this section, we elaborate on privacy threats associated with making data openly accessible.

3.1 Disclosure of Real Identities

Privacy can be characterized as a need to oversee data and associated interactions [11]. It is clear that privacy risks are caused mainly by the risks related with anonymizing the information and making it open for re-utilize. Privacy legislation and data protection policies oblige organisations and governments not to distribute private data. In this specific situation, associations are requested to dismiss any distinguishing data from the information before making it accessible on the web. In any case, a few researches in anonymization methods demonstrate that anonymized information can be de-anonymized and consequently identifying information can be disavowed. For instance, Narayanan and Shmatikov [12] demonstrated that adversary with very little information about a user, could recognize his or her record in the Netflix openly distributed datasets of 500,000 anonymized endorsers. Likewise, expelling genuine names, birth dates and other sensitive data from datasets may not generally have the coveted impact. For example, the police may distribute open information about autos and bike robberies after removing genuine names of individuals included. In spite of the fact that excluding these names may sound attractive for privacy and security insurance, more research is expected to examine whether user identities are safeguarded and whether this is a robust approach.

3.2 Personal Information Discovery Through Linking Data

The blend of factors from different datasets could bring about the distinguishing proof of people and uncover personalities [13]. Information characteristics, alluded to with the term ‘semi-identifier’, can be connected to outer information assets and consequently can prompt the arrival of concealed personalities [14]. Cases of semi identifiers are a man’s age, sex and address.

An attacker may recognize an individual John from a dataset. By linking this dataset to another on the sexual orientation, origin and the city where John lives, John’s records might be distinguished in the open dataset. Hence, these information types are critical to be covered up too. Furthermore, delicate information, for example, illness ought to be excluded from the datasets. In any case, information suppliers frequently cannot anticipate ahead of time which blend of factors will prompt privacy infringement [13], and along these lines this expectation is a perplexing action. Some domains like healthcare providers, utilize a linkage unit to interface diverse datasets together with a linkage id instead of personally identifiable information (PII) so as to lessen privacy risks. More often, the linkage unit depends on a blend of obscured PII data (for example first letters of names + birth date) [15]. However in the same way an attacker can mimic a linkage unit and retrieves actual records of users.

3.3 Personal Information Discovery Through Data Mining

Open information makes information accessible online for specialists and organizations. Organizations utilize data mining procedures to retrieve significant information from these datasets which help them in their organizations. While doing as such, they can damage user privacy since mining the information can derive private data. In order to help overcome these issues, privacy preserving data mining techniques should be used to reduce privacy risks [14].

3.4 Data Utilization Versus Privacy

Once a dataset has been exchanged from the information owner to the information distributer, the information proprietor is never again responsible for his or her information. Information control has exchanged to the information distributer who is legally responsible for the assurance of individuals' privacy. Before distributing the information on the web, the information distributer anonymizes the information and evacuates any delicate information that makes it conceivable to distinguish people.

A large portion of the circumstances, the information distributer does not know who will get the information and for what reason he or she will get to the information. Further, the information distributer does not really realize what mining systems will be utilized by the information beneficiary and how much sensitive data can be found from the anonymized information. On the off chance that the information distributer expels all identifying data, modifies related semi-identifiers, and evacuates delicate information, the distributed information can lose its esteem. Subsequently, there ought to be a harmony between what can be distributed, with the goal for clients to have the capacity to infer valuable data, and in the meantime guaranteeing privacy insurance. Complete privacy assurance may bring about no utilization of the information by any means, and consequently the distributed information can happen to have no esteem.

4 Proposed Privacy Risks Assessment Model

Uncertainty related with the exposure of information makes it hard to come up with a decent way to ensure user privacy. Whenever distributed, obscure outsider associations and different clients can access sensitive data. Sharing data under uncertainty conditions while having the capacity to ensure privacy protection is one of the difficulties in these situations [16]. Since evaluating privacy and security risks is basic for ventures [17], we expect the same is needed for open data environments for the sake of protection of user data. The key components of the proposed model are presented as follows (see Fig. 1):

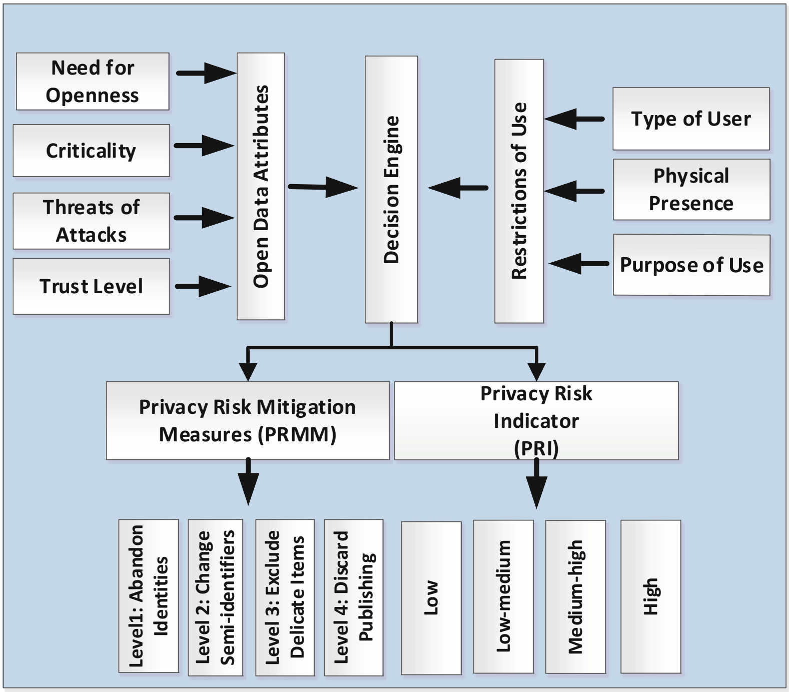
Need for Openness
Criticality
Threats of Attacks
Trust Level
Open Data Attributes
Decision Engine
Restrictions of Use
Type of User
Physical Presence
Purpose of Use
Privacy Risk Mitigation Measures (PRMM)
Privacy Risk Indicator (PRI)
Level1: Abandon Identities
Level 2: Change Semi-identifiers
Level 3: Exclude Delicate Items
Level 4: Discard Publishing
Low
Low-medium
Medium-high
High

Fig. 1. Proposed privacy risk assessment model

4.1 Open Data Attributes

Privacy attributes represent factors influencing publishing of the data openly which is inspired by the work of [18] who defined a number of factors influencing users willingness to share their private information and [19] who proposed a decision engine to evaluate privacy risks associated with sharing information on social networks. In this context, five attributes are distinguished as follows:

- Criticality level: this factor expresses the significance of the information, practically equivalent to the significance of the advantage of information distributing to the group. Criticality level can be measured by running a privacy impact assessment.
- Openness: alluding to the requirement for distributing the information straightforwardly. This factor expresses the advantage of data publishing from public and business perspectives. In the event that the information criticality level is high and the need for openness is high, at that point an exchange-off exists and the requirement for transparency can exceed the high criticality level or the other way around.
- Risk of Attacks: this alludes to what degree is the normal digital security danger alarm. In the event that the threat of an attack is set to high, at that point this can

have effect on the idea of information being distributed and made accessible to others. This is similar to the threat level of an attack which takes four levels in the Netherlands; minimal meaning attacks are unlikely, limited meaning an attack is less likely, substantial meaning an attack is most likely and critical meaning very strong indication an attack will happen [20]. Here, we use similar notation which reflects the status of security threats at the body publishing information, its neighbourhood or country.

- Trust: this corresponds to how the data publisher/owner is evaluated by others with respect to his or her dependability. Disrepute of the data publisher impacts the nature of the information and the way protection is managed.
- Restrictions of Use: restrictions of use represents access privileges allowed on the data. We distinguish four types to describe this restriction:
  - None. This means no restriction is applied.
  - Role of user. This means a restriction is applied on basis of the user role.
  - Purpose of use. This means different types of restriction may apply depending on the purpose data is needed for.
  - Physical Presence. This means that data access depends on the physical location where it is accessed from.

4.2 Decision Engine

The decision engine is in charge of settling on apparent privacy risks and suggests an appropriate privacy risk mitigation measure. This is done in light of a scoring value and a decision algorithm having scores of open data attributes as input. Practices are indicated by topic specialists and by investigation of related work. For effortlessness, a scoring mechanism is utilized where attributes are given scores on a scale from 1 to 5 as indicated by their risk to privacy. Each attribute is valuated with a score s such that $0 < s \leq 1$. These scores are created based on assumptions on privacy risks associated with each attribute value. Each attribute category A_i has a weight $(0 < w_i \leq 1)$ associated with it such that when aggregating all scores they get weighted as follows:

$$PRI = \frac{1}{n} * \sum_{i=1}^{n} w_i * Max(s_i), \quad PRI \leq 1 \tag{1}$$

Max(Si) means that if more than one score is possible within one attribute category because of the existence of more than one attribute value like for example two types of use, then the maximum score is selected to reflect the one with the highest risk. The upside of utilizing weights is to present some adaptability with the end goal that the influence of each characteristic class can get refreshed after some time as indicated by lessons gained from assembled information and already discovered privacy threats.

4.3 Privacy Risk Indicator (PRI)

The PRI represents the predicted value of privacy risks associated with opening such data. PRI can have four values; low, low-medium, medium-high and high. A high PRI means the threat to privacy violation is expected to be high. PRI is determined by the decision engine based on the scoring matrix and the rules associated with the decision engine.

4.4 Privacy Risk Mitigation Measures (PRMM)

Based on the decision engine, privacy risk indicator score is predicted together with a privacy risk mitigation measure. For example, what should be done if there is a risk that the identity of an owner of a stolen bike can be tracked down if we publish stolen bike records online? The following measures are used in our framework:

- Level 1: Abandon identifiers. This is the slightest measure that should be taken by a data publisher when the risk indicator shows low risk. By doing that, they adhere to the European directives and the law. The utilization of database anonymization software is compulsory with a specific end goal to anonymize the data [21–23].
- Level 2: Change Semi-identifiers. Modifying semi identifiers' information can help diminish identity capturing. Semi identifiers are information constructs which if connected with other datasets can uncover user identities. Illustrations are age, sex and postal district [16, 24]. Researchers around there created mechanisms that can identify and find semi-identifiers [25–27]. To meet PRMM level 2, PRMM level 1 activities must be completed as well.
- Level 3: Exclude Delicate Data. For a few cases, there are data items, for example, restorative infections which are considered delicate and should be safeguarded when publishing the data if the privacy risk is high. The sort of information that is viewed as delicate or sensitive shifts from dataset to another which makes it complex to securely recognize and evacuate. Some commercial tools exist that could be utilized. Examples of such tools are Nessus [28]. To meet PRMM level 3, PRMM levels 1&2 activities must be completed as well.
- Level 4: Discard Publication. In case that the threat is high, it is advised not to publish the data at all, and therefore the recommended measure would be to reject publication.

5 PRI Model Implementation

In this section, we elaborate on the implementation features of the PRI model. Functional requirements needed to implement the proposed model in an open data platform are introduced.

5.1 Functional Components

It is obvious that for such a platform privacy represents a critical requirement which must be met. The proposed functionalities are inspired by the proposed ones in [29]. These functionalities are (see Fig. 2):

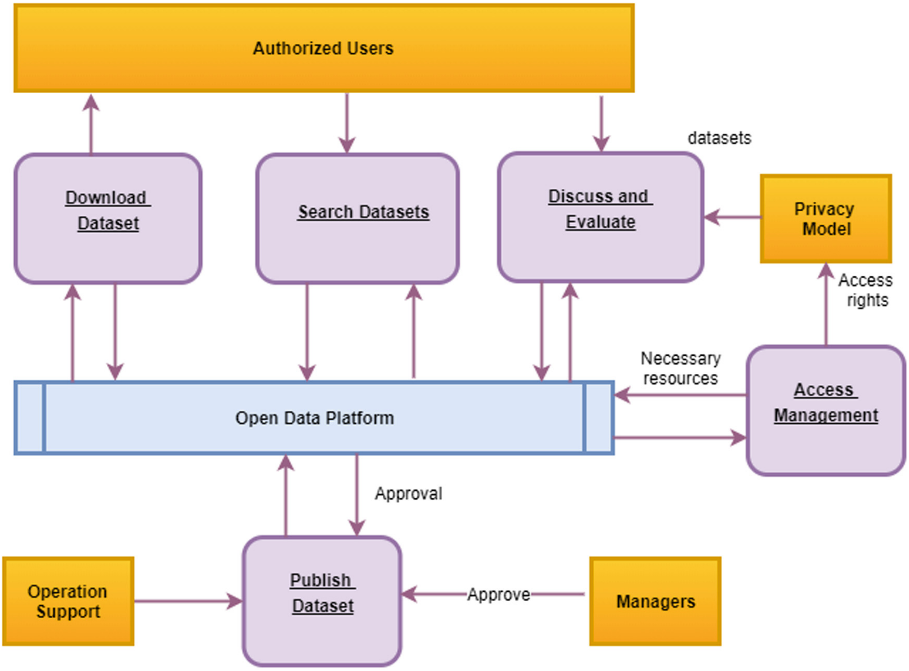

Fig. 2. PRI model functional components

- Search Datasets: Via this functionality, users can search for available open data by specifying specific keywords.
- Publish Datasets: data can be published to be used by other researchers or practitioners as well. Before publishing the data, the dataset will be reviewed by the management. A decision need to be take on the privacy measurements by the privacy model first to guarantee users' privacy without impacting the quality of the data. The privacy model is where the proposed privacy assessment model will be implemented.
- Discuss an Evaluate: Through this functionality researchers can discuss together certain data elements with other experts who have had similar experiences. Further this component implements datamining tools to evaluate data properties.
- Access Management and Privacy Model: Access control is needed in order to restrict access to this platform for authorized persons only. These include: registered professionals, technicians and decision makers etc. Further, sensitive data and privacy identifying information are removed before publishing.
- Evaluation Engine: This is responsible for providing an evaluation for the privacy mitigation decision using the PRI model.
- Request for Download of Dataset: Via this functionality, professionals will be able to download the data they want after taking the necessary approvals.

- User Roles: Three types of users are distinguished; management, authorized users and operators. The management is responsible for granting access to users. Further, they approve datasets being uploaded by the operators to be published on the platform.

5.2 Technical Implementation

Using a common platform for sharing data could only exist due to the recent developments in services technology and computer networks [29]. Such specialized arrangements have been there since over two decades with the development of web technology innovation by Microsoft [30]. The early usage of this approach relied upon the high adaptability offered by the World Wide Web. Later, and with the improvement of web 2.0 technology, ventures put resources into enormous IT change ventures towards achieving business and public strategic goals [31]. This prompt presentation of web 2.0 innovation assets examples are Representational State Transfer (REST) protocol and RESTFUL web services [32]. In this work, RESTFUL web services are prescribed on the grounds that REST gives preferable execution over SOAP web services. In addition, its usage is simpler than that of SOAP [32].

6 Illustration Scenarios

In this section, we describe five scenarios to illustrate the proposed model. These scenarios are based on the authors' experience with recent projects in open government. In each case we evaluate PRI and PRMM. Table 1 demonstrates the scoring approach in light of the authors' experiences. Table 2 demonstrates how PRI is mapped to a privacy risks mitigation measure level (PRMM). The situations include distinctive sorts of actors who conduct different activities. A portion of the actors upload datasets, others utilize them or both transfer and utilize them. The kind of information provided fluctuates between the scenarios, since a portion of the opened data is updated regularly, while others are static with or without refreshes.

The criticality of the data ranges from low to high, and the information utilize is confined in different ways. The utilization of some datasets is not limited, while for different datasets the limitation relies upon the reason for utilize, and the type of user. The level of trust in information quality is diverse for each of the scenarios, extending from exceptionally constrained issues (low) to no issues (high).

6.1 Scenario S1: Open Crime Data Usage and Provisioning

A resident of a European city needs to know what number of violations happen in her neighbourhood contrasted with different neighbourhoods in the city. She scans different open information frameworks for the information that she is searching for. When she discovers ongoing open crime information, she downloads and examines them. As per the permit, the information can be utilized as a part of different structures, both non-economically and monetarily. Data perceptions help the citizen to understand the

Table 1. Open data attributes matrix

Attribute	Attribute value	Score (s)
Type of user	Government	0.2
	Researcher	0.4
	Citizen	0.6
	Student	0.8
	Company	1.0
Purpose of use	Information	0.2
	Research	0.4
	Commercial	0.6
	Sharing	0.8
	Unknown	1.0
Type of data	Static	0.33
	Updated	0.67
	Real-time	1.0
Data criticality	Low	0.25
	Low-medium	0.50
	Medium-high	0.75
	High	1.00
Restrictions of use	None	0.25
	Type of user/purpose of use	0.50
	Restricted by country	0.75
	Restricted by network	1.00
Need for openness	Low	0.33
	Medium	0.67
	High	1.00
Trust	Low	0.25
	Low-medium	0.50
	Medium-high	0.75
	High	1.00

Table 2. Mapping PRI to PRMM

PRI	Score	PRMM
Low	0.00–0.25	Level 1: Abandon identities
Low-medium	0.25–0.50	Level 2: Remove Semi-identifiers
Medium-high	0.50–0.75	Level 3: Exclude Delicate data
High	0.75–1.00	Level 4: Discard publishing

information. In any case, she has just constrained data about the nature of the dataset and about the supplier of the data, which diminishes her trust in the information.

The open data framework that the subject uses does not just permit governmental associations to open datasets, yet offers this capacity to any client of the infra-structure.

This subject likewise needs to share a few information herself. She has gathered observation of robbery in the shop that she possesses, and submits these information on the web as open data. This implies the citizen both downloads and transfers open data. An overview of open data characteristics for the scenarios is given in Table 3. Utilizing the proposed model, From Table 1, PRI can be calculated using Eq. (1): $PRI = 0.61$. From Table 2, PRI can be seen to be medium-high meaning a relatively high privacy risk with associated PRMM set at level 3: remove delicate data. The data publisher should filter the published data from identifying information, semi-identifiers and sensitive data to avoid this expected relatively high privacy risk.

Table 3. Scenarios overview

Attributes	S1	S2	S3	S4	S5
Type of user	Citizen	Governmental archivist	Student	Researcher	Civil servant
Purpose of use	Use and upload open data about neighbourhood	Upload open social data	Use open data for study	Use open data for research	Use data provided by own organization
Type of data	Real-time	Static	Static	Static, updated frequently	Real-time and static, updated frequently
Data criticality	Low	Low	Low-medium	Medium-high	High
Restrictions of use	None	None	Purpose of use, type of user	Physical presence, type of user	Physical presence, type of user
Openness	High	High	Medium	Low	Low
Trust level	High	Low-medium	Medium-high	Low	Low

6.2 Scenario S2: Open Social Data Provisioning

An archivist working for an administrative organization keeps up the open data framework of this office. Datasets cannot be transferred by anybody yet just by a representative of the administrative association. The analyst has the undertaking to make different social datasets that are discovered fitting for production by the office representatives accessible to general society. The analyst transfers static datasets that are non-sensitive, with the goal that the risks on security ruptures is limited. The datasets can be reused by anybody; there are no limitations in regards to the kind of user or the reason for utilize. Since the datasets are given online much metadata, including information about the nature of the dataset, this lessens the trust issues that clients may have needed to utilize the dataset. Utilizing the proposed model, a review of this scenario open data qualities is appeared in Table 3. Like scenario S1, PRI = 0.39 with Low-medium privacy risk. PRMM is set at level 2: expel semi-identifiers. This infers expelling identifying information too.

6.3 Scenario S3: Use of Restricted Archaeology Data

An understudy directs an examination in the region of archaeology studies. To acquire access to the information, the understudy needs to present a demand at the association that claims the information. In his demand, the understudy needs to give data about himself, his examination and about the reason for which he needs to utilize the administrative organization with data, the legislative organization can choose to give more delicate information than the information that they offer with open access.

More delicate data can be unveiled to this single client, under the condition that he will not give the data to others. Since the client can directly contact the information supplier, trust issues are much lower than they might be for other (open) datasets. Utilizing the proposed model, an outline of this situation is given in Table 3. Similarly, PRI = 0.54, PRMM is at level 3: remove delicate data. Special contractual agreement can be put in place with this particular student before delicate data can be shared with him otherwise this data has to be unveiled.

6.4 Scenario S4: Use of Physically Restricted Statistics Data

A scientist might want to utilize some statistics that is given by a governmental measurements association. The measurements office has been opening information for a long time and has a good fame around there, since it offers great information. The researcher in this manner puts stock in the information of the measurements office and trusts that he can reuse these information for his own particular research. While the investigator can get to different open datasets on the web, some datasets are given in a more confined frame. To get to the more delicate datasets, the specialist needs to physically go to the measurements office. The measurement office does not open these delicate information, since this may prompt privacy breaches. The scientist can investigate the information at the area of the measurements office, yet it is not permitted to take any information alongside him and to distribute these information as open information. Since the specialist physically needs to move to the measurements office, the workplace can acquire knowledge in the reasons for which the analyst needs to utilize the information, and in light of this reason, they support or object the utilization of their information. Utilizing the proposed model, an illustration of this case is given in Table 3. Similarly, PRI = 0.51, PRMM is at level 3: remove delicate data. This means before sharing this data with the researcher, all sensitive data has to be removed together with identifying information and semi-identifiers.

6.5 Scenario S5: Use of Physically Restricted Agency Data

A government worker may be engaged with opening datasets, as well as reuse datasets that are given by her own association. The organization's information must be gotten to inside by its workers who are available at the office, and is in this manner confined by type of user and by physical boundaries. The datasets are both run-time and static, yet they are refreshed much of the time. The office's information are exceptionally sensitive; since they have not been anonymized and delicate data has not been removed. The information cannot be utilized by anybody and are not open. Trust of the information client is high, since the client knows about the setting in which the information have been made and approaches partners who can answer inquiries concerning the information if vital. Utilizing the proposed model, an outline of this case is given in Table 3. Similarly, PRI = 0.68, PRMM is at level 3: remove delicate data.

In the past cases, risk of attack is thought to be low in this way it was excluded in the evaluations. From the above, we see that for the diverse situations of the same dataset, we could have distinctive privacy risks and in this manner need to consider applying measures for mitigation of these risks (see Table 4). The use of the proposed model has given knowledge into this relationship between the datasets and the situations in view of privacy risks scores related with these cases. This knowledge will help in applying the appropriate privacy risk mitigation measure (PRMM) before distributing the information straightforwardly.

Table 4. Overview of scenarios evaluation

Scenario	PRI	PRMM
S1	Medium-high	Level 3: exclude delicate data
S2	Low-medium	Level 2: expel semi-identifiers
S3	Medium-high	Level 3: exclude delicate data
S4	Medium-high	Level 3: exclude delicate data
S5	Medium-high	Level 3: exclude delicate data

7 Conclusions

The opening and sharing of information is regularly hindered by security and privacy observations. Most work on privacy assesses privacy breaches in view of evaluation of organizations' methods and taken procedures for managing individual information and their development in doing as such according to benchmarks and normal practices. These systems cannot be effectively utilized in open data platforms in light of the fact that the information does not contain personally identifiable data (PII) as a matter of course if distributed out in the open. In any case, in this paper, we demonstrated that PII can in any case be uncovered even when evacuated through various ways. We additionally contended for the need of assessing the diverse scenarios related with the use of the dataset before a decision to be made on whether to open the data.

References

1. Janssen, M., van den Hoven, J.: Big and Open Linked Data (BOLD) in government: a challenge to transparency and privacy? Gov. Inf. Q. **32**(4), 363–368 (2015)
2. European Parliament and the Council of the European Union: Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (1995)

3. European_Commission: Communication from the commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Towards better access to scientific information: Boosting the benefits of public investments in research (2012). Accessed 6 Oct 2013
4. OECD: OECD recommendation of the council for enhanced access and more effective use of on Public Sector Information (2008). http://www.oecd.org/dataoecd/41/52/44384673.pdf. Accessed 8 Nov 2011
5. ISO/IEC-29100: INTERNATIONAL STANDARD ISO/IEC Information technology - Security techniques - Privacy framework (2011)
6. Kroener, I., Wright, D.: A strategy for operationalizing privacy by design. Inf. Soc. **30**(5), 355–365 (2014)
7. ISACA AICPA/CICA: Privacy Maturity Model (2011)
8. Revoredo, M., et al.: A privacy maturity model for cloud storage services. In: Proceedings of the 7th International Conference on Cloud Computing (2014)
9. Wright, D.: The state of the art in privacy impact assessment. Comput. Law Secur. Rev. **28** (1), 54–61 (2012)
10. Blackmer, W.S.: GDPR: getting ready for the new EU general data protection regulation. In: Information Law Group (2016)
11. James, T.L., Warkentin, M., Collignon, S.E.: A dual privacy decision model for online social networks. Inf. Manag. **52**, 893–908 (2015)
12. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: Proceedings of the IEEE Symposium on Security and Privacy, pp. 111–125 (2008)
13. Zuiderwijk, A., Janssen, M.: Towards decision support for disclosing data: closed or open data? Inf. Polit. **20**(2), 103–117 (2015)
14. Xu, L., et al.: Information security in big data: privacy and data mining. IEEE Access **2**, 1149–1176 (2014)
15. Randall, S.M., et al.: Privacy-preserving record linkage on large real world datasets. J. Biomed. Inform. **50**, 205–212 (2014)
16. Eldin, A., Wagenaar, R.: Towards autonomous user privacy control. Int. J. Inf. Sec. Priv. **1** (4), 24–46 (2007)
17. Jones, J.A.: An Introduction to Factor Analysis of Information Risk (Fair) (2005). http://www.fairinstitute.org/. Accessed 13 Dec 2016
18. Ali-Eldin, A., Wagenaar, R.: A fuzzy logic based approach to support users self control of their private contextual data retrieval, In: European Conference on Information Systems (ECIS). Association for Information Systems (AISeL), Turku (2004)
19. Ali-Eldin, A., van den Berg, J., Ali, H.: A risk evaluation approach for authorization decisions in social pervasive applications. Computer and Electrical Engineering **55**, 59–72 (2016)
20. Government_of_the_Netherlands: Risk of an attack (threat level). https://www.government.nl/topics/counterterrorism-and-national-security/risk-of-an-attack-threat-level. Accessed 28 Jan 2018
21. Anonymizer. http://www.eyedea.cz/image-data-anonymization/. Accesed 1 Mar 2017
22. ARX: Data Anonymization Tool. http://arx.deidentifier.org/. Accessed 1 Mar 2017
23. Camouflage's-CX-Mask: https://datamasking.com/products/static-masking/. Accessed 1 Mar 2017
24. Fung, B.C., et al.: Privacy preserving data publishing: a survey of recent developments. ACM Comput. Surv. **42**(4) (2010)
25. Shi, P., Xiong, L., Fung, B.: Anonymizing data with quasi-sensitive attribute value. In: Proceedings of the 19th ACM International Conference (2010)

26. Motwani, R., Xu, Y.: Efficient algorithms for masking and finding quasi-identifiers (PDF). In: Proceedings of the Conference on Very Large Data Bases (VLDB) (2007)
27. Shadish, W.R., Cook, T.D., Campbell, D.T.: Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton-Mifflin, Boston (2002)
28. Nessus: Nessus Vulnerability Scanner. https://www.tenable.com/products/nessus-vulnerability-scanner. Accessed 1 Mar 2017
29. Ali-Eldin, A.M.T., Hafez, E.A.: Towards a universal architecture for disease data models sharing and evaluation. In: 2017 International Symposium on Networks, Computers and Communications (ISNCC) (2017)
30. Josuttis, N.M.: SOA in Practice: The Art of Distributed System Design. O'Reilly, Sebastopol (2007)
31. Ali-Eldin, A.M.T.: Towards a shared public electronic services framework. Int. J. Comput. Appl. **93**(14), 48–52 (2014)
32. Abeysinghe, S.: Restful PHP Web Services. PACKT Publishing, Birmingham (2008)