Opening More Data

A New Privacy Risk Scoring Model for Open Data

Ali-Eldin, Amr; Zuiderwijk-van Eijk, Anneke; Janssen, Marijn

Publication date
2017

Document Version
Publisher's PDF, also known as Version of record

Published in
Proceedings of the 7th International Symposium on Business Modeling and Software Design 2017

Citation (APA)
Ali-Eldin, A., Zuiderwijk-van Eijk, A., & Janssen, M. (2017). Opening More Data: A New Privacy Risk Scoring Model for Open Data. In Proceedings of the 7th International Symposium on Business Modeling and Software Design 2017 (pp. 146-154)

Opening More Data
A New Privacy Risk Scoring Model for Open Data

Amr M. T. Ali- Eldin[1,2,3], Anneke Zuiderwijk[2] and Marijn Janssen[2]
[1]*Leiden Institute of Advanced Computer Science, Leiden University, Leiden, the Netherlands*
[2]*TBM - Delft University of Technology, Jaffalaan 5, 2628 BX Delft, The Netherlands*
[3]*Computer and Control Systems Dept., Faculty of Engineering, Mansoura University, Mansoura, Egypt*

Keywords: Open Data, Open Government Data, Privacy, Risk, Data Release, Data Mining, Scoring Systems.

Abstract: While the opening of data has become a common practice for both governments and companies, many datasets are still not published since they might violate privacy regulations. The risk on privacy violations is a factor that often blocks the publication of data and results in a reserved attitude of governments and companies. Additionally, even published data, which might seem privacy compliant, can violate user privacy due to the leakage of real user identities. This paper proposes a privacy risk scoring model for open data architectures to analyse and reduce the risks associated with the opening of data. The key elements consist of a new set of open data attributes reflecting privacy risks versus benefits trades-offs. Further, these attributes are evaluated using a decision engine and a scoring matrix intro a privacy risk indicator (PRI) and a privacy risk mitigation measure (PRMM). Privacy Risk Indicator (PRI) represents the predicted value of privacy risks associated with opening such data and privacy risk mitigation measures represent the measurements need to be applied on the data to avoid the expected privacy risks. The model is exemplified through five real use cases concerning open datasets.

1 INTRODUCTION

Governments and publicly funded research organizations are encouraged to disclose their data and to make this data accessible without restrictions and free of charge (B. 2009, Commision 2011, B. 2012). Opening public and private data is a complex activity that may result in benefits yet might also encounter risks (Conradie and Choenni 2014, Zuiderwijk and Janssen 2014, Zuiderwijk and Janssen 2015). An important risk that may block the publication of the data is that organizations might violate the privacy of citizens when opening data about them (Conradie and Choenni 2014). Moreover, when opening data, organizations lose control on who will be using this data and for what purpose. Once data is published, there is no control over who will download, use and adapt the data.

To avoid privacy violations, data publishers can remove sensitive information from datasets, however, this makes datasets less useful. In addition, even published data, which may seem privacy compliant, can violate user privacy due to leakage of real user identities when various datasets and other resources are linked to each other (Kalidien, Choenni et al. 2010). The possibility of mining the data afterwards to get meaningful conclusions can lead to leakage of private data or users real identities. Although organizations remove identifying information from the dataset before publishing the data, some studies demonstrate that anonymized data can be de-anonymized and hence real identities can be revoked (Kalidien, Choenni et al. 2010).

Various existing studies have pointed at the risks and challenges of privacy violations for publishing and using open data (Kalidien, Choenni et al. 2010, Conradie and Choenni 2014, Janssen and van den Hoven 2015, Perera, Ranjan et al. 2015). Some studies have identified privacy risks or policies for organizations in collecting and processing data (Drogkaris, Gritzalis et al. 2015, Kao 2015), some have provided decision support for opening data in general (Zuiderwijk and Janssen 2015), and some have focused on releasing information and data on the individual level (James, Warkentin et al. 2015). Nevertheless, there is still limited insight in how organizations can reduce privacy violation risks for open data in particular, and there is no uniform approach for privacy protection (Janssen and van den Hoven 2015). From existing studies it has not become

clear which open data model can be used to reduce the risk on open data privacy violations. An open data model is needed that helps making decisions on opening data and that provides insight in whether the data may violate users' privacy.

The objective of this paper is to propose a model to analyse privacy violation risks of publishing open data. To do so, a new set of what are called open data attributes is proposed. Open data attributes reflect privacy risks versus benefits trade-offs associated with the expected use scenarios of the data to be open. Further, these attributes are evaluated using a decision engine to a privacy risk indicator (PRI) and a privacy risk mitigation measure (PRMM). In particular this can help to determine whether to open data or keep it closed.

This paper is organized as follows. Section 2 discusses related work while section 3 presents privacy violation risks associated with open data, followed by section 4 which introduces the proposed model. The model helps identifying the risks and highlights possible alternatives to reduce these risks. Section 5 exemplifies the model by providing some use cases and preliminary results. Section 6 discusses the key findings and concludes the paper.

2 RELATED WORK

Public bodies are considered the biggest creators of data in the society in what is known as public data. Public data may range from data on procurement opportunities, weather, traffic, tourist, energy consumption, crime statistics, to data about policies and businesses (Janssen and van den Hoven 2015). Data can be classified into different levels of confidentiality, including confidential, restricted, internal use and public (ISO27001 2013). We consider public data that has no relation with data about citizens as outside the scope of this work.

Anonymized data about citizens can be shared to understand societal problems, such as crime or diseases. An example of citizen data is the sharing of patient data to initiate collaboration among health providers which is expected to be beneficial to the patient and researchers. The highly expected benefits behind this data sharing are the improved understanding of specific diseases and hence allowing for better treatments. It can also help practitioners to become more efficient. For example, a general practitioner can quickly diagnose and prescribe medicines. Nevertheless, this sharing of patients' information should be done according to data protection policies and privacy regulations.

A variety of Data Protection Directives has been created and implemented. Based on the Data Protection Directive of 1995 (European Parliament and the Council of the European Union 1995), a comprehensive reform of data protection rules in the European Union was proposed by the European Commission (2012). Also the Organization for Economic Co-operation and Development has developed Privacy Principles (OECD, 2008), including principles such as "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, the ISO/IEC 29100 standard has defined 11 privacy principles (ISO/IEC-29100 2011).

Nowadays a relatively new approach for privacy protection called privacy-by-design has received attention of much organization such as the European Network and Information Security Agency (ENISA). Privacy-by-Design suggests integrating privacy requirements into the design specifications of systems, business practices, and physical infrastructures (Hustinx 2010). In the ideal situation data is collected in such a way that privacy cannot be violated.

The Data Protection Directives are often defined on a high level of abstraction, and provide limited guidelines for translating the directives to practice. Despite the developed Data Protection Directives and other data protection policies, organizations still risk privacy violations when publishing open data. In the following sections we elaborate on the main risks of privacy violation associated with open data.

A number of information security standards were estalished to achieve effective information security governance, among which are ISO (2013), COBIT5 and NIST (2016). Most work on privacy risk assessment aim to conduct surveys or questionnaires that assess companies' ways of dealing with personal data according to regulatory frameworks and moral or ethical values. When it comes to open data, such frameworks to assess privacy risks cannot be used since the data to be published will contain no identifying information as a pre-requisite by the law. Having said that, normal ways of assessing privacy risks cannot be applied and new ways are needed that outweigh the benefits of sharing the data compared to expected privacy risks of the leakage of personally identifiable information.

3 PRIVACY THREATS FOR OPENING DATA

3.1 Real Identities Disclosure

Privacy can be defined as a person's desire to manage information and interaction about him or her (James, Warkentin et al. 2015). It appears that privacy threats are caused mainly by the risks associated with anonymizing the data and making it public for re-use. Privacy legislation and data protection policies force organizations and governments not to publish private information. In this context, organizations are asked to remove any identifying information from the data before making it available online. Nevertheless, some studies in anonymization techniques show that anonymized data can be de-anonymized and hence real identities can be revoked. For example, Narayanan and Shmatikov (2008) showed that an adversary with very little information about a user, could identify his or her record in the Netflix openly published datasets of 500,000 anonymized subscribers. In addition, removing real names, birth dates and other sensitive information from datasets may not always have the desired effect. For instance, the Dutch police has started to publish open data about cars and bicycle thefts after removing real names of people involved. Although removing these names might sound satisfactory for user privacy protection, more research is needed to analyse whether user identities are safe and whether this is a robust approach.

3.2 Privacy Leakage through Linked - Data

The combination of variables from various datasets could result in the identification of persons and reveal identities (Zuiderwijk and Janssen 2015). Data attributes, referred to with the term 'quasi-identifier', can be linked to external data resources and hence can lead to the release of hidden identities (XU, JIANG et al. 2014). Examples of quasi-identifiers are a person's age, gender and address.

Figure 1 shows an example of privacy leakage through data linkage. In this example, an attacker may identify the person John from this dataset. By combining information about the gender, birth place and the city where John lives, John may be identified in the open dataset. Therefore, these data types are important to be hidden as well. In addition, sensitive data such as diseases should be removed from the datasets. However, data providers often cannot predict in advance which combination of variables will lead to privacy leakage (Zuiderwijk and Janssen 2015), and thus this prediction is a complex activity.

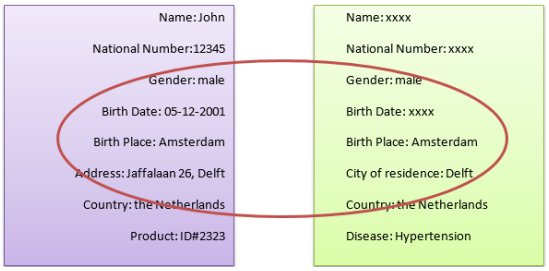

Figure 1: Privacy leakage through data linkage.

3.3 Data Mining

Open data makes data available online for researchers and companies. Companies use data mining techniques to conclude meaningful information from these datasets which help them in their businesses. When doing so, they can violate users' privacy because mining the data can deduce private information. In order to help overcome these issues, privacy preserving data mining techniques should be used to reduce privacy risks (XU, JIANG et al. 2014).

3.4 Data Utilization versus Privacy

Once a dataset has been transferred from the data owner to the data publisher, the data owner is no longer in control over his or her data. Data control has transferred to the data publisher who is responsible against the law for the protection of people's privacy. Before publishing the data online, the data publisher anonymizes the data and removes any sensitive data that makes it possible to identify persons.

Most of the times, the data publisher does not know who will receive the data and for what purpose he or she will access the data. Further, the data publisher does not necessarily know what mining techniques will be used by the data receiver and how much sensitive information can be deduced from the anonymized data. If the data publisher removes all identifying information, alters associated quasi-identifiers, and removes sensitive data, the published data can lose its value. Hence, there should be a balance between what can be published, in order for users to be able to derive useful information, and at the same time ensuring privacy protection. Complete privacy protection might result in no use of the data at all, and hence the published data can become of no value.

4 PRIVACY RISKS SCORING MODEL

Uncertainty associated with the disclosure of data makes it difficult to come up with a good approach to protect users' privacy. When published, unknown third-party organizations and other users can get access to sensitive information. Sharing information under uncertainty conditions while being able to guarantee user privacy represents one of the challenges in these environments (Ali Eldin and Wagenaar 2007). Since assessing privacy and security risks is critical for enterprises (Jones 2005), we expect the same is needed for open data environments for the sake of protection of user data.

The key elements of the proposed model are presented as follows (see Figure 2):

4.1 Open Data Attributes

Based on authors' observations from previous case studies on open data architectures, five open data attributes are assumed in this research. The first four are shown at the top left corner in Figure 2, whereas the fifth attribute is shown at the top right corner. The five open data attributes are as follows:

- Need for openness: referring to the need for publishing the data openly. If the data criticality level is high but the need of openness is high, then a trade-off exists and the need for openness can outweigh the high criticality level or vice versa.
- Criticality level: this attribute represents the importance of the data, analogous to the importance of the benefit of data publishing to the community.
- Security alarm/threat: this refers to what degree is the expected cyber security threat alert. If the security threat is set to high, then this can have impact on the nature of data being published and made available to others.
- Trust level: refers to how the data publisher is rated by others with respect to his or her trustworthiness. Reputation of the data publisher influences the quality of the data and the way privacy is dealt with. For example, whether the data publisher is trusted that he or she will comply with privacy regulations and whether the quality of data will be high.

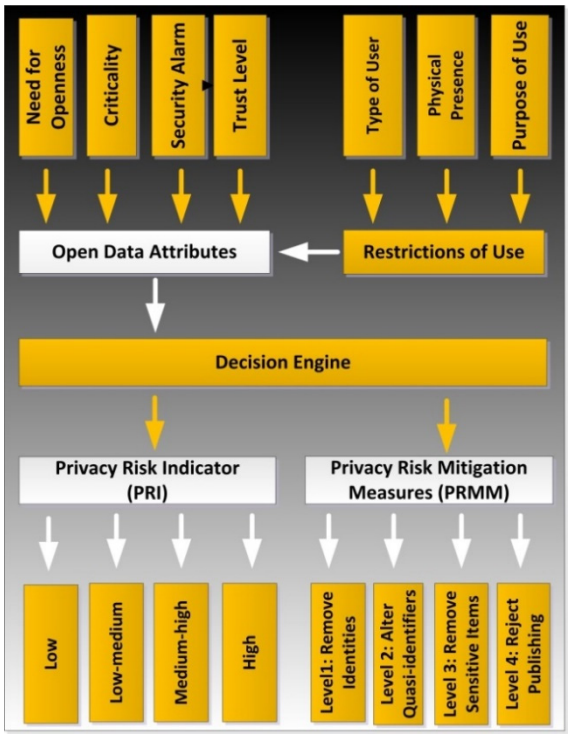

Figure 2: Proposed Privacy Risks Scoring (PRS) Model.

- Restrictions of Use: Restrictions of use represents access privileges allowed on the data. We distinguish three ways to describe this restriction:
  - Type of user. This means a restriction is applied on basis of the role the user plays.
  - Physical Presence. This means that data access depends on the physical location where it is accessed from.
  - Purpose of use. This means different types of restriction may apply depending on the purpose data is needed for.

4.2 Decision Engine

A box in the middle of Figure 2 depicts the decision engine. The decision engine component is responsible for deciding upon perceived privacy risks and recommends a suitable privacy risk mitigation measure. This is done based on a scoring matrix and a rule engine having scores of open data attributes as input. Rules are specified by subject matter experts and by analysis of the model associated data access records.

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, a 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: Remove identifiers. This is the least measure that needs to be taken by a data publisher when the risk indicator is classified as low risk. By doing that, they adhere to the European data directives and data protection laws. The use of database anonymization tools is mandatory in order to remove the identities and make the data anonymous. Examples of such tools are (Anonymizer , ARX , Camouflage's-CX-Mask).
- Level 2: Alter Quasi-identifiers. Changing quasi-identifiers' data values can help reduce identity leakage. Quasi-identifiers are data types which if linked with other datasets can reveal real identities. Examples are age, sex and zip code (Ali Eldin and Wagenaar 2007, Fung, Wang et al. 2010). Researchers in this area developed algorithms that can detect and find quasi-identifiers (Shadish, Cook et al. 2002, Motwani and Xu 2007, Shi, Xiong et al. 2010). To meet PRMM level 2, PRMM level 1 activities must be completed as well.
- Level 3: Remove Sensitive Items. For some cases, there are data items such as medical diseases which are considered sensitive and need to be protected when publishing the data if the risk indicator is high risk. The type of data that is considered sensitive varies from dataset to another which makes it complex to safely identify and remove it. Some commercial tools exist that could be used which can be adjusted for specific data type and specific operating systems. An example of such tools is Nessus (Nessus). To meet PRMM level 3, PRMM levels 1 and 2 activities must be completed as well.
- Level 4: Reject Publishing. 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 publishing.

5 IMPLEMENTATION AND RESULTS

In this section, we describe five use cases to illustrate the proposed model. These cases are based on real scenarios. In each case we determine the Privacy Risk Indicator (PRI) and the Privacy Risk Mitigation Measure (PRMM). The cases involve different types of actors who conduct different activities. Some of the actors upload datasets, others use them or both upload and use them. The type of data provided varies between the cases, since some of the opened data are provided real-time, while others are static with or without updates.

The criticality of the data ranges from low to high, and the data use is restricted in various ways. The use of some datasets is not restricted, whereas for other datasets the restriction depends on the purpose of use, the type of data user, the physical presence of the data user at a certain location or the type of user. The level of trust in data quality is different for each of the cases, ranging from large issues (low trust level) to very limited issues (high trust level).

5.1 Decision Engine

Before we can assess the different cases, we need to specify the rules used in the decision engine. For the sake of simplicity, a scoring matrix is used where attributes are given scores on a scale from 0 to 1 according to their threat to privacy. Table 1 shows an example of the scoring approach based on the authors' experiences.

Each attribute is valuated with a score *s* such that $s \leq 1$. These scores are created based on assumptions on privacy risks associated with each attribute value. Each attribute category A_i has a weight $(0 < w_i \leq 10)$ associated with it such that when aggregating all scores they get weighted as follows:

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

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 advantage of using weights is to introduce some flexibility such that the influence of each attribute category can get updated over time according to lessons learned from gathered data and previously found privacy treats. Table 2 shows how PRI value is mapped to a

corresponding privacy mitigation risk measure level (PRMM).

5.2 Case 1: Use and Provision of Open Crime Data

A citizen of a large European city wants to know how many crimes occur in her neighbourhood compared to other neighbourhoods in the city.

She searches various open data infrastructures for the data that she is looking for. When she finds real-time open crime data, she downloads and analyses them. According to the license, the data can be used in various forms, both non-commercially and commercially. Data visualizations help the citizen to make sense of the data. Nevertheless, she has only limited information about the quality of the dataset and about the provider of the data, which decreases her trust in the data.

The open data infrastructure that the citizen uses does not only allow governmental organizations to open datasets, but offers this function to any user of the infrastructure. This citizen also wants to share some data herself. She has collected observation of theft in the shop that she owns, and publishes these data on the internet as open data. This means that the citizen both downloads and uploads open data. Using the proposed model, an overview of open data attributes for this case can be given (see Table 3).

From table 1, the *PRI* can be calculated using equation (1): $PRI = 0.61$. *PRI* can be seen to be medium-high meaning a relatively high privacy risk with associated PRMM set at level 3: *remove sensitive data*. The data publisher should filter the published data from identifying information, quasi-identifiers and sensitive data to avoid this expected relatively high privacy risk.

Table 1: Open Data Attributes Scoring Matrix.

Attribute (A)	Weight (w)	Attribute Value	Score (s)
Type of User	1	Government	0.2
		Researcher	0.4
		Citizen	0.6
		Student	0.8
		Company	1.0
Purpose of use	1	Information	0.2
		Research	0.4
		Commercial	0.6
		Sharing	0.8
		Unknown	1.0
	1	Static	0.33
Type of data		Updated	0.67
		Real-time	1.0
Data Criticality	1	Low	0.25
		Low-medium	0.50
		Medium –high	0.75
		High	1.00
Restrictions of use	1	None	0.25
		type of user / purpose of use	0.50
		Restricted by country	0.75
		Restricted by network	1.00
Need for Openness	1	Low	0.33
		Medium	0.67
		High	1.00
Trust in Data Quality	1	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: Remove identities
Low - Medium	0.25-0.50	Level 2: Remove Quasi-identifiers
Medium – High	0.50-0.75	Level 3: Remove Sensitive data
High	0.75-1.00	Level 4: Reject publishing

Table 3: Case 1 Overview.

Case Attributes	Case 1
Type of User	Citizen
Purpose of use	Use and upload open data about neighbourhood
Type of data	Real-time
Data Criticality	Low
Restrictions of use	None
Need for Openness	High
Trust Level	High

5.3 Case 2: Provision of Open Social Data

An archivist working for a governmental agency maintains the open data infrastructure of this agency. Datasets cannot be uploaded by anyone but only by an employee of the governmental organization. The archivist has the task to make various social datasets that are found appropriate for publication by the agency employees available to the public. The archivist uploads static datasets that are non-sensitive, so that the risk on privacy breaches is minimized. The datasets can be reused by anyone; there are no restrictions regarding the type of user or the purpose of use. Since the datasets are provided online with much metadata, including data about the quality of the dataset, this reduces the trust issues that data users may have. Using the proposed model, an overview of

this case's open data attributes is shown in Table 4. The PRI for case 2 is 0.39 with Low − medium privacy risks. PRMM is set at level 2: remove Quasi-identifiers. This implies removing identifying information as well.

Table 4: Case 2 Overview.

Case Attributes	Case 2
Type of User	Governmental Archivist
Purpose of use	Upload open social data
Type of data	Static
Data Criticality	Low
Restrictions of use	None
Need for Openness	High
Trust Level	Low-Medium

5.4 Case 3: Use of Restricted Archaeology Data

A student conducts a study in the area of archaeology. To obtain access to the data, the student needs to submit a request at the organization that owns the data. In his request, the student needs to provide information about himself, his study and about the purpose for which he wants to use the data. The data are not completely open, since access to the data is restricted by a data request procedure.

Since the data user needs to provide the governmental agency with information, the governmental organization can decide to provide more sensitive data than the data that they offer with open access.

For study purposes, more sensitive data can be disclosed to this single user, under the condition (contractually agreed) that he will not provide the information to others. Since the user can personally contact the data provider, trust issues are less common than they may be for other (open) datasets. Using the proposed model, an overview of this case is given in Table 5. $PRI = 0.54$, PRMM is at level 3: remove sensitive data. Again as mentioned earlier, special contractual agreement can be put in place with this particular student before sensitive data can be shared with him otherwise sensitive data has to be removed.

Table 5: Case 3 Overview.

Case Attributes	Case 3
Type of User	Student
Purpose of use	Use open data for study
Type of data	Static
Data Criticality	Low-medium
Restrictions of use	Purpose of use, type of user
Need for Openness	Medium
Trust Level	Medium-high

5.5 Case 4: Use of Physically Restricted Statistics Data

A researcher would like to use open statistical data that is provided by a governmental statistics organization. The statistics office has been opening data for many years and has a good reputation in this area, since it offers high-quality data. The researcher therefore trusts the data of the statistics office and believes that he can reuse these data for his own research. While the researcher can access various open datasets on the internet, some datasets are provided in a more restricted form. To access the more sensitive datasets, the researcher needs to physically go to the statistics office.

The statistics office does not open these sensitive data, since this may lead to privacy breaches. The researcher can analyse the data at the location of the statistics office, yet it is not allowed to take any data along with him and to publish these data as open data. Since the researcher physically needs to travel to the statistics office, the office can obtain insight in the purposes for which the researcher wants to use the data, and based on this purpose, they approve or disapprove the use of their data. Using the proposed model, an overview of this case is given in Table 6. Accordingly, $PRI = 0.51$, and the PRMM is at level 3: remove sensitive data. This means before sharing this data openly, all sensitive data has to be removed together with identifying information and quasi-identifiers.

Table 6: Case 4 Overview.

Case Attributes	Case 4
Type of User	Researcher
Purpose of use	Use open data for research
Type of data	Static, updated frequently
Data Criticality	Medium-high
Restrictions of use	Physical presence, type of user
Need for Openness	Low
Trust Level	Low

5.6 Case 5: Use of Physically Restricted Agency Data

A civil servant may not only be involved in opening datasets, but may also reuse datasets that are provided by her own organization. The agency's data can only be accessed internally by its employees who are present at the agency, and is therefore restricted by type of user and by physical barriers. The datasets are both real-time and static, yet they are updated frequently. The agency's data are highly sensitive; since they have not been anonymized and sensitive information has not been removed. The data cannot

be used by anyone and are not open. Trust of the data user is high, since the user is familiar with the context in which the data have been created and has access to colleagues who can answer questions about the data if necessary. Using the proposed model, an overview of this case is given in Table 7. Accordingly, $PRI = 0.68$, and the PRMM is at level 3: remove sensitive data.

In the previous cases, the security threat is assumed to be low and thus it was not included in the computations. From the above, we see that for the different use cases of the same dataset, we can have different privacy risks and thus we need to consider applying measures for privacy risks mitigation. The application of the proposed model has given insight into this association between the datasets and the use cases based on privacy risks scores associated with these cases (see Figure 3). This insight will help in applying the suitable privacy risk mitigation measure (PRMM) before publishing the data openly.

Table 7: Case 5 Overview.

Case	Case 5
Type of User	Civil servant
Purpose of use	Use data provided by own organization
Type of data	Real-time and static, updated frequently
Data Criticality	High
Restrictions of use	Physical presence, type of user
Need for Openness	Low
Trust Level	Low

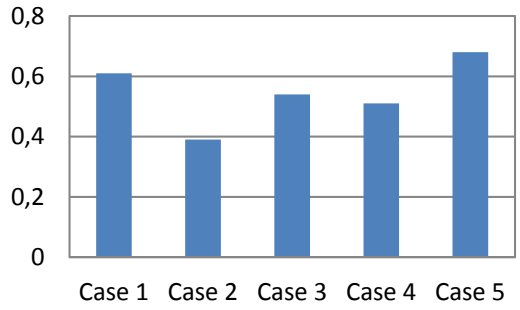

Figure 3: PRI scores for five different open data cases with equally weighted Open Data Attributes.

6 DISCUSSION AND CONCLUSION

The opening and sharing of data is often blocked by privacy considerations. Most work on privacy risk assessment evaluates privacy risks based on assessment of companies' ways of dealing with personal data and their maturity in doing so according to standards and common practices. These frameworks cannot be applied in open data architectures because the data does not contain personally identifiable information (PII) by default if published in public. However, in this paper, we showed that PII can still be disclosed even after being removed through different ways. We also argued for the need of evaluating the different use cases associated with the dataset before a decision to be made on whether to open the data.

In this paper, a new model for privacy risk scoring in open data architectures was proposed. The model is based on defining a new set of, what is called, open data attributes and privacy risk mitigation measures. Each open data attribute is given a score according to a predefined scoring matrix. From the implemented cases, it was clear that different privacy risk mitigation measures are considered depending on risks associated with these attributes. Each defined privacy risk mitigation measure should be applied before making this dataset available online openly.

Further research is needed to define a common basis for the scoring matrix and all possible open data attributes. Statistical analysis should be conducted to validate the possible generalizability of the proposed model. In addition, details of the realization architecture should be discussed together with the implementation details of the privacy risk mitigation measures.

REFERENCES

Ali Eldin, A. and R. Wagenaar (2007). " Towards Autonomous User Privacy Control." International Journal of Information Security and Privacy 1(4): 24-46.

Anonymizer. "http://www.eyedea.cz/image-data-anonymization/ - Last visited on 1-3-2017."

ARX "ARX Data Anonymization Tool - available at http://arx.deidentifier.org/ - Last visited on 1-3-2017."

B., O. (2009). "Memorandum for the Heads of executive Departments and Agencies: Transparency and Open Government. Available at: http://www.whitehouse.gov/the_press_office/Transparency_and_Open_Government - lat visited on 10-3-2017."

B., O. (2012). "Digital Government. Building a 21st Century Platform to Better Serve the American People. Available at: http://www.whitehouse.gov/sites/default/files/omb/egov/digital-government/digital-government.html - last visisted on 10-3-2017."

Camouflage's-CX-Mask "https://datamasking.com/products/static-masking/ - Last visited on 1-3-2017."

Commision, E. (2011). Communication From The Commision To The European Parliment, The Council, The European Economic and Social Committee and The Committe of The Regions: Open data An engine for innovation, growth and transparent governance E. Commision. Brussels.

Conradie, P. and S. Choenni (2014). "On the barriers for local government releasing open data." Government Information Quarterly 31(supplement 1): S10–S17.

Drogkaris, P., S. Gritzalis, C. Kalloniatis and C. Lambrinoudakis (2015). "A hierarchical multitier approach for privacy policies in e-government environments." Future Internet 7: 500-515.

European Parliament and the Council of the European Union (1995). 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

European_Commission. (2012). "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." Retrieved October 6, 2013.

Fung, b., C., K. Wang, R. Chen and P. Yu (2010). "Privacy Preserving Data Publishing: A Survey of Recent Developments." ACM Computing Surveys 42(4).

Hustinx, P. (2010). "Privacy by design: delivering the promises." Identity in the Information Society 3(2): 253-255.

ISO27001 (2013). ISO/IEC 27001:2013 - Information technology -- Security techniques -- Information security management systems -- Requirements - Last visited on 10-3-2017., .International Organization for Standardization.

ISO. (2013). "ISO/IEC 27002:2013 Information technology - Security techniques - Code of practice for information security controls. ."

ISO/IEC-29100 (2011). "INTERNATIONAL STANDARD ISO / IEC Information technology - Security techniques - Privacy framework."

James, T. L., M. Warkentin and S. E. Collignon (2015). "A dual privacy decision model for online social networks." Information & Management 52: 893–908.

Janssen, M. and J. van den Hoven (2015). "Big and Open Linked Data (BOLD) in government: A challenge to transparency and privacy?" Government Information Quarterly 32(4): 363–368.

Jones, J. A. (2005). "An Introduction to Factor Analysis of Information Risk (Fair) Available from: http://www.fairinstitute.org/ - Last visited on 13-12-2016."

Kalidien, S., S. Choenni and R. F. Meijer (2010). Crime Statistics Online: Potentials and Challenges. 11th Annual International Digital Government Research Conference on Public Administration Online: Challenges and Opportunities. S. A. Chun, R. Sandoval and A. Philpot. Puebla, Mexico, Digital Government Society of North America: 131-137.

Kao, D.-Y. (2015). "Exploring Privacy Requirements and Their Online Managements." Journal of Computers 26(2): 34-45.

Motwani, R. and Y. Xu (2007). Efficient Algorithms for Masking and Finding Quasi-Identifiers (PDF). Proceedings of the Conference on Very Large Data Bases (VLDB).

Narayanan, A. and V. shmatikov (2008). Robust De-anonymization of Large Sparse Datasets. IEEE symposium on Security and Privacy: 111-125.

Nessus "Nessus Vulnerability Scanner, https://www.tenable.com/products/nessus-vulnerability-scanner - Last visited on 1-3-2017."

NIST. (2016). "Cybersecurity Framework version 1.1. Retrieved from National Institute of Standards and Technology (NIST): https://www.nist.gov/topics/cybersecurity."

OECD. (2008). "OECD recommendation of the council for enhanced access and more effective use of on Public Sector Information." Retrieved November 8, 2011, from http://www.oecd.org/dataoecd/41/52/44384673.pdf.

Perera, C., R. Ranjan and L. Wang (2015). "End-to-End Privacy for Open Big Data Markets." IEEE Cloud Computing 2(4): 44-53.

Shadish, W. R., T. D. Cook and D. T. Campbell (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, Houghton-Mifflin.

Shi, P., L. Xiong and B. Fung (2010). Anonymizing data with quasi-sensitive attribute value. Proceedings of the 19th ACM International Conference.

XU, L., C. JIANG, J. WANG and J. YUAN (2014). "Information Security in Big Data: Privay and Data Mining." IEEE Access 2.

Zuiderwijk, A. and M. Janssen (2014). The negative effects of open government data: investigating the dark side of open data. the 15th Annual International Conference on Digital Government Research. Aguascalientes, Mexico.

Zuiderwijk, A. and M. Janssen (2015). "Towards decision support for disclosing data: Closed or open data?" Information Polity 20(2, 3): 103-117.