Literature Review

Personal data privacy protection using knowledge graph

By: Harshith Shankar Tarikere Ravikumar - M.Sc. Computer Science (Data Analytics)

19230323

Supervisor: Prof. Mathieu D'Aquin

1 Introduction

This research thesis aims to provide a technical solution to evaluate and warns users of the potential data breach by an organization they are interacting with. The software uses semantic web technologies and knowledge graphs to analyse the user relationship with the organization whose website is visited by the user. Based on the user relationship with the organization, the software generates a privacy score metric which gives a data sensitivity level to the particular user with respect to the organization. The user will be warned based on data sensitivity level. A browser extension is used to collect website information, to get user's feedback to improve the privacy score metric and warn the user of possible data breach.

2 Privacy metrics

One of the major components in this software which generates data sensitivity scores used to warn the user of the potential privacy consequences of visiting and entering data to a particular website. The data sensitivity score is based on the relationship between user and organisation, and the organisation's privacy practice.

PrivacyMeter, is a browser extension described in Starov and Nikiforakis (2018). PrivacyMeter computes a relative privacy score for any website that a user is visiting. This score is computed based on practices of a website's privacy such as the reputation of trackers, the amount of third-party content or presence of insecure leaky web forms. PrivacyMeter gives information to the user about the discovered privacy issues and recommends action to take.

The severity of a privacy issue is evaluated by comparing the value of the current page to the mean value over other websites. If the value is greater than the mean plus one standard deviation then the current page is marked as "potentially dangerous", and if the value is greater than the mean plus two standard deviations then the current page is marked as "dangerous" issue. Customisation of score calculation can be done by the user by removing or adding some privacy preference factor, and default scoring thresholds can be modified by the user. PrivacyMeter also provides an option to compare relative factors to the median and the third quartile of distribution and it also allows user to change base threat level. (Starov and Nikiforakis, 2018)

PrivacyMeter tests a webpage against the four groups of privacy risks, such as Third-party trackers, Fingerprinting activity, Third-party content, and Leaky web forms. First, three groups mentioned above measures the potential privacy risk by calculating comparative metrics on how the number of trackers, an application program interface (APIs), or third-party content differs from the average number for similar websites respectively. If the difference is high then the current website is potentially dangerous. In the last group, privacy risks are evaluated by identifying web forms that may leak entered information such as web with the only Hypertext Transfer Protocol (HTTP) GET method. (Starov and Nikiforakis, 2018)

Therefore, Starov and Nikiforakis (2018) provides an approach to calculate the privacy score to evaluate privacy risks. From this paper, calculating privacy risks by identifying web forms that may leak entered information can be used in our thesis. This paper fails to consider the user relationship with the organisation whose website is visited by the user to calculate the privacy score.

Another approach of defining privacy metrics for online users is proposed in Blauw and Solms (2017). Privacy metric is calculated by assigning different points or weights to different aspects such as the scope of personal Identity visibility, Identity exposure and Identity elements.

Privacy metrics are calculated by measuring the personal information of an individual stored at different levels (the scope of personal Identity visibility) such as the World Wide Web (WWW), the Deep web, and the Dark Web. If the information of an individual is present on the Dark web then the more weight is added to the privacy metric. Similarly, if the information is on the deep web then less weight is added than the Dark web. If it is on the World Wide Web (WWW) even less weight is added than the deep web. (Blauw and Solms, 2017)

Privacy metrics are measured by considering how easily identity elements of an individual can be accessed (Identity exposure). If the identity elements/personal information are easily available then the privacy weights will be high. If the identity elements are public or public with authentication will have weights less than the easily available. If the identity elements are private or private with public authentication or private with full/part authentication will have less weights than the publicly available. (Blauw and Solms, 2017)

Privacy metrics also considers the different weights for different identity elements for an individual. Identity elements like a credit card, passport number, Personal identification number will have high weights. Identity elements like Financial history, medical history, criminal history will have comparatively less weights. Identity elements like full name, gender, and the race will have low weights. All these weights are added to calculate the privacy metric of an individual. (Blauw and Solms, 2017)

These weighting schemes can be used in our thesis to calculate the privacy metric. Blauw and Solms (2017) intelligently assigns the weights to the individual but these weights are generic. This approach fails to assign different weights to a different group of users/different individuals.

In Hamed and Ayed (2015), authors proposed one more scoring model which quantifies the privacy risk. The score is calculated based on the assumptions that the tracking component presence is a risk by itself (Hamed and Ayed, 2015). In addition to that presence of more than one component including JavaScript increases risk as those component's interaction might permit to collect more data about a user. Tracking component, such as cookies, JavaScripts (JS), Local shared objects (LSOs) and Iframes.

The privacy score of a website is computed as the sum of component existence score and interaction score. Component existence score is a sum of all the component scores. Interaction score is a combination of JS and any component score present on the same webpage. If the privacy score is high then the webpage is not secure. (Hamed and Ayed, 2015)

Hamed and Ayed (2015) provides a different approach to calculate privacy score by using web trackers and it also considers browsing time of the user to calculate the privacy score. It is limited to only trackers and it's interactions and therefore fails to consider the relationship between the user and the website.

3 Browser Extensions

Browser extensions play a major role in this thesis so efficient development of browser extension is crucial. Implementing better functionality and user interface of an extension is essential.

Carmi and Bouhnik (2016) analyses the functions of four common data security and private applications such as Ghostery, Privacy Badger, Disconnect, and Privatedog. The author analyses the three aspects of these applications: functional activity, user interface design and economic-financial sources.

Each application works similarly by providing information regarding tracking measures to users by clicking on the extension icon but takes a different approach to block the tracking measure. (Carmi and Bouhnik, 2016)

Ghostery application blocks the user activity tracking by scanning code sections of the advertisement bodies which are found on the webpage visited by the user. By clicking on the extension icon, a new window will open which has the tracking information and the customisation options for blocking. It also allows the user to create a white list of websites.(Carmi and Bouhnik, 2016)

Disconnect Private Application detects that the browser is trying to communicate with a different server than the user requested. This request is categorized into seven groups including a group called content. If the request is other than content then it blocks. Clicking on the icon opens a new window which has blocked tracking measures and the option is also available for the user to unblock those measures. (Carmi and Bouhnik, 2016)

Privacy Badger analyses the cookies to block the tracking measures. This application displays a number on the icon in the browser which represents the number of tracking measures found. To know the information regarding the tracking measures, the user has to click on the icon. Privacy badger categorises each tracking measure into three groups with the colour red representing a completely blocked measure, yellow with icon marked by a red X representing allowed tracking but blocks creating cookies, and green which represents that it is not a tracking measure. (Carmi and Bouhnik, 2016)

In our thesis, colour-changing icons, detailed information about blocked/privacy risk website in the separate window, and user feedback interface in the separate window approaches can be used. Carmi and Bouhnik (2016) doesn't mention the performance of the extensions related to time, space and privacy risk by the extension.

Schaub et al. (2016) analyses the user interface and the performance of the privacy extensions such as Ghostery, DoNotTrackMe and Disconnect. These extensions provide clarity to the user regarding who is collecting and what is collected by the trackers. Performance in tracking is best for Ghostery extension and DoNotTrackMe and Disconnect have similar performance. Authors of this paper argue that a displayed number on the icon which represents the number of blocked trackers is very small to be noticed by a user. Detailed information of blocked trackers in the new window will help the user to gain more clarity. Colour used in these extensions will lead to confusion to the user.

Displaying numbers on the icon will lead to ambiguity and colour selection should be appropriate. These observations in paper are solely based on the experiment with 24 participants and the number of websites and search topics were also limited.

In paper Starov and Nikiforakis (2018), the user interface of the extension(PrivacyMeter) has an icon which shows the number of privacy issues. By clicking on the icon a pop up will appear. This interface consists of the box plot to display the relative privacy risks and text warning with clearly stated privacy risk. Three colour scheme is also used to show the level of privacy risks, where green represents safe, yellow represents potentially dangerous and red represents dangerous.

The invisibility of a browser extension is crucial to prevent the privacy risk of the user from trackers and website. Trackers and website can detect the browser extension if an extension tries to modify the document.(Starov and Nikiforakis, 2018)

Statistical analysis such as box plot can be used to visualise the privacy risk level of the webpage. PrivacyMeter doesn't explicitly allow the user to white list/blacklist websites. (Starov and Nikiforakis, 2018)

4 Research Gaps

From the literature review, it can be inferred that privacy-protecting extensions only consider third-party trackers present on a website but ignore other signals such as the presence of privacy policy and the presence of HTTPS. By analysing the privacy policy and transfer protocol of a website, and adding these analysed factors to calculate privacy metric will solve this issue.

Extensions only block the third-party trackers but don't provide detailed information about the tracker to the user. All these extensions fail to consider the relationship between user and organisation whose website the user visited. Building a knowledge graph around the user and organisation will address this issue. This knowledge graph will also provide information about the organisation to the user.

References

Oleksii Starov and Nick Nikiforakis. Privacymeter: Designing and developing a privacy-preserving browser extension. In *ESSoS*, 2018. 1, 2, 5

Frans Blauw and Sebastiaan Solms. Towards quantifying and defining privacy metrics for online users. pages 1–9, 05 2017. doi: 10.23919/ISTAFRICA.2017.8102366. 2, 3

A. Hamed and H. K. Ayed. Privacy scoring and users' awareness for web tracking. In *2015 6th International Conference on Information and Communication Systems (ICICS)*, pages 100–105, 2015. 3

Golan Carmi and Dan Bouhnik. Functional analysis of applications for data security and for surfing privacy protection in the internet. volume 4, pages 201–208, 07 2016. 3, 4

Florian Schaub, Aditya Marella, Pranshu Kalvani, Blase Ur, Chao Pan, Emily Forney, and Lorrie Cranor. Watching them watching me: Browser extensions impact on user privacy awareness and concern. 01 2016. doi: 10.14722/usec.2016.23017. 4