Issues in Information Systems
Volume 16, Issue III, pp. 99-107, 2015

INFORMATION EXPOSURE MODEL: A FRAMEWORK FOR INDIVIDUAL ASSESSMENT OF RISKS AND BENEFITS FROM USING A WEBSITE

Robert L. Leitheiser, University of Wisconsin – Whitewater, leitheir@uww.edu
L. Roger Yin, University of Wisconsin – Whitewater, yinl@uww.edu

ABSTRACT

Popular websites are major contributors to Big Data. In the trade-off between risks and benefits, most of existing literature focuses on the organization. This paper focuses on the individual website user. How does the user trade-off the benefits they receive from using a website with the risks they incur from providing their personal information? The authors develop a model of Information Exposure that can be used as a framework for this assessment. The paper reviews the top 10 independent U.S. websites to see if they have the information necessary for individual users to make a rational decision about website usage.

Keywords: Big Data, Privacy, Website, Internet, Risk Assessment, Information Security, Cybersecurity

INTRODUCTION

Websites are important sources of "Big Data." When a user visits a website, he/she is often promptly identified and tracked. Data are collected from forms, navigation choices, time spent, cookies, and other technologies. This is often combined with other data to create a more complete profile of the user. These profiles provide significant benefits to businesses and advertisers. This collected information by the same token also presents risks to users from the negative consequences of intentional or accidental disclosure.

Privacy, the right of an individual to be left alone, is becoming an increasing concern for computer users connected to the Internet. A recent survey by the Pew Research Center [9] of 607 adults found that 91% of them felt they had lost control over how personal information was collected and used by companies. Of adults who used social media, 80% were concerned about advertisers and other businesses using the data they entered on the site. Eighty-eight percent of the sample felt that it would be very hard to remove inaccurate information that was captured online. Overall, the sample was very concerned about surveillance by both government and businesses alike.

Privacy can be thought of as having two forms: [7, p. 6]

- Physical privacy – the ability of a person to maintain their own physical space or solitude.
- Informational privacy – the ability of a person to control, edit, manage and delete information about themselves and to decide how and to what extent such information is communicated to others.

Concern for information privacy has led to the passage of laws and regulations in the United States (The Privacy Act of 1974), the United Kingdom (The Data Protection Act of 1984; 1998), Canada (Personal Information Protection and Electronic Documents Act of 2001), and the European Union (Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, 1995). In the U.S. the Federal Trade Commission (FTC) published the Fair Information Practice Principles in 1998. These regulations and frameworks have made privacy an issue for firms that collect user information on their websites.

From the firms' perspective, they are at risk when individuals disclose personal information to them. They are responsible and accountable for the security of the personal information they collect. It could also be to their benefit if they understand the high-stakes risk involved. Courtney [6] proposed an early model of security risk that included privacy. In his model, risk is the impact of disclosure times the probability of disclosure. Disclosure could be intentional or accidental. Risk increased with the time the information is held. Besides disclosure, his model also included the risks modification and destruction.

Bakerville [3] continues the focus on organizational risk in his survey of security methods. The intent is to reduce organizational risks rather than to focus on individuals. Boehm [4] presents an in-depth model for identifying information risks associated with system development. It is also organization and project based.

Risk assessment for information systems projects, has been formalized for the US Federal Government by the National Institute of Standards and Technologies (NIST).[8] It's risk model includes:

1. Threat Source, initiates a
2. Threat Event, exploits a
3. Vulnerability, causing an
4. Adverse Impact, producing
5. Organizational Risk

The focus continues to be on the risk to the organization. Privacy is only part of the equation and it is from the organization's perspective.

Privacy, however, is gaining in importance. Clarke [5] argues that there is an increasing demand for formal privacy assessment. The driving forces are:

1. A public reaction to privacy-invasive actions by governments and corporations, and
2. A desire by corporations to reduce resistance to the introduction of new systems.

Clarke then describes initiatives in several countries to support Privacy Impact Assessments (PIA).

A major force in the development of Privacy Impact Assessment is the UK's Information Commissioner's Office (ICO). It produced an initial PIA Handbook in 2007. After 6 years the ICO commissioned a review [12] that caused it to make changes that were implemented in a revised handbook. [7]

The revised handbook was designed to help firms assess the risks associated with collecting personal information by performing a Privacy Impact Assessment (PIA). Such an assessment involves the following steps [7, p. 15]:

1. Identifying the need for a PIA,
2. Describing the information flows,
3. Identifying the privacy and related risks,
4. Identifying and evaluating privacy solutions,
5. Signing off and recording the PIA outcomes, and
6. Integrating the PIA outcomes back into the project plan.

It is recommended that firms preform this analysis for new projects and systems. The focus remains on the organization.

Polonetsky, Tene and Jerome [10] go further by adding a benefits component to the analysis. They called this a Data Benefit Analysis (DBA) and it consists of the following steps:

1. Assess the "raw value" of the benefit to be gained by sharing information including:
   - Nature of the benefit
   - Potential beneficiaries
   - Degree of benefit – size and scope
2. Discount the benefit by the probability that it can be achieved.

This creates a kind of "expected benefit value" that should be compared to the PIA risk. They propose a 2x2 comparison grid that has risk on one dimension and benefit on another. Low Benefit/High Risk projects should be canceled. High Benefit/Low Risk projects should proceed. Projects in the other 2 cells are "maybes".

While in their model, Polonetsky, Tene and Jerome [10] identify benefits for individuals, community, organizations and society; their focus is on the organization. It is the organization that is making the decision and their main concern is to manage their risk and maximize their benefit.

In another paper, Tene, and Polonetsky [11] argue that the benefits of Big Data overwhelm traditional conventions on data privacy and that new approaches to information transparency and individual data access should prevail. While seeming to provide benefits to the individual their approach also provides major benefits to Big Data entities. The authors believe that individuals whose privacy is on the line should have a framework that represents their perspective.

The purpose of this paper is to explore the risk/benefit analysis process that an individual user can do when he/she visits a website and is asked to provide personal information. More specifically, the authors address the following research question:

> Do users of popular websites have the information they need to make rational decisions about whether the risks of providing personal information are worth the benefits from using the website?

The question will be addressed in two steps. First, the authors propose an Information Exposure Model that can be used as a framework for rational decision-making in the web usage context. Second, the authors review the most popular US websites to see if the information needed to apply the Information Exposure model is readily available on the website. Conclusions and suggestions are made based on the findings.

THE INFORMATION EXPOSURE MODEL

The individual who visits a website does so to obtain benefits from the website. The individual exchanges personal information and perhaps money for these benefits. The personal information has value to the company that owns the website because the company can sell it or use it internally. There are risks to the individual that are created when the company shares this information. That risk occurs because of the individual's information exposure. This section will develop a model of information exposure that will be used to try to assess the risk/benefit tradeoff for individuals visiting websites that generate big data.

The model is presented in Figure 1. The individual has information about him/herself that is potentially of value to websites and 3rd Parties; e.g., advertisers. The individual can chose to disclose this information to a website through data entry, search terms, option selections, link clicks, etc. That information then becomes "exposed". There is no risk to the individual if they keep information to themselves. Once they disclose it to a website it becomes exposed to possible disclosure to unknown 3rd parties.

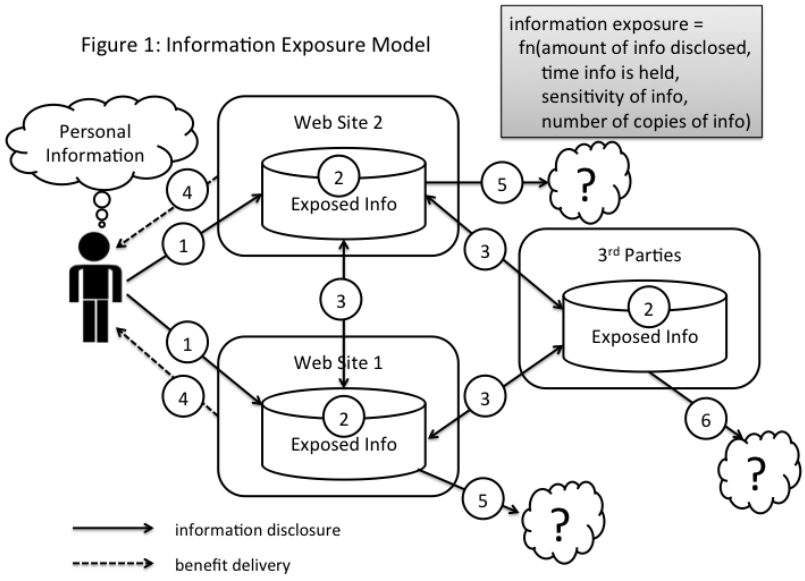

Figure 1. Information Exposure Model

The degree of exposure is a function of:

(1) the amount of information disclosed,
(2) the length of time the information is held,
(3) the number of copies that are made of the information, and
(4) the sensitivity of the information.

The sensitivity of the information is determined by the negative consequences that would result if certain 3rd parties knew the information. Third parties can obtain information about persons that does not come from the individual. Loan histories for example. That is not the focus of this model or paper. The decision of the individual about whether to disclose information to a website in return for perceived benefits is the subject here.

Individuals visit websites for many reasons. At each website they should assess the benefit they are receiving vs. the risk they are taking from information exposure. This can only be done if they know what information they are providing. If the site is collecting data that is not explicit then the individual cannot make a good assessment or a good decision.

Websites that share information increase the information exposure of the individual and undermine the individual's decision on the benefit/risk assessment for the one site. Perhaps the individual would not provide information to Site 2, which he/she provided to Site 1, because the benefits of Site 2 were not worth it.

Websites often share information with known 3rd parties. This may be part of the value they provide to individuals or it may be a way for them to generate revenue. In either case this increases the information exposure. These 3rd parties could intentionally or accidentally disclose copies of information. For example, the 3rd parties could be "friends" who share personally embarrassing information or health care organizations that leak health data to insurance companies. Criminal actors accessing any of the website or 3rd party information stores could also obtain information. The user may not know what 3rd parties are involved in information sharing, may not know what 3rd parties are doing with their information, and may not know how their information is connected to other information. Security and tight disclosure practices are the only ways to control information exposure risks in these circumstances.

To summarize, if the individual is to make a rational decision about whether the benefits of providing information to a website are worth the risks, he/she should know:

1. What information they are providing the website (this would help them assess exposure from amount of data and sensitivity of the data).
2. How long the site will store their data before deleting it (assess exposure from length of information retention).
3. Who will the site share the information with (help assess the risk of information exposure due to copies).
4. What benefits the individual will gain by providing the information. Note: this should not include benefits that individuals would receive if they did not disclose any information.
5. What security is in place on the original site organization to keep the exposed information from being disclosed to unknown 3rd parties.
6. What security is in place in the 3rd party organizations to keep the exposed information from being disclosed to unknown 3rd parties.

These knowledge requirements are keyed to the model in Figure 1 and are the basis for analyzing the online privacy statements of the top 10 U.S. websites.

RESEARCH METHODOLOGY

The top 10 U.S. websites were identified by using the ranking by the Alexa website [1] that was available on May 9, 2015. The No. 4 site on that day was YouTube.com, which uses the Google.com privacy statement. Site No. 11 (Reddit.com) was added to create a listing of ten sites with independent privacy statements. The Alexa listing is based on web traffic and is defined as follows:

> Alexa's Traffic Ranks are based on the traffic data provided by users in Alexa's global datapanel over a rolling 3 month period. Traffic Ranks are updated daily. A site's ranking is based on a combined measure of Unique Visitors and Pageviews. Unique Visitors are determined by the number of unique Alexa users who visit a site on a given day. Pageviews are the total number of Alexa user URL requests for a site. However, multiple requests for the same URL on the same day by the same user are counted as a single Pageview. The site with the highest combination of unique visitors and pageviews is ranked #1. Additionally, we employ data normalization to correct for biases that may occur in our data [2].

For each website, the authors entered the home page URL into a web browser. Then they searched on the home page for a link to the site's privacy statement. For all the top 10 sites the link was at the bottom of the page and was clearly labeled.

The authors only reviewed the main privacy policy page. Additional links to pages on cookies, children's privacy and other special privacy issues were not included. The authors believe that the crucial information for users to make the decision on whether to contribute information to a website should be delivered in a single webpage.

The primary privacy policy pages for the websites were downloaded on May 9, 2015, and reviewed for text that related to the six model dimensions introduced above. The two authors, independently coded privacy page statements into:

1. Statements about information collected by the website.
2. Statements about how long information will be retained.
3. Statements about how information will be shared.
4. Statements about how the information will be used to provide benefits to users.
5. Statements about how site will secure personal information.
6. Statements about how 3rd party partners will protect user personal information.

After coding statements in a site's privacy page, the researchers rated the policy statements in terms of coverage using the following scale:

N – no coverage: did not mention
M – minimal coverage: mentioned briefly but did not provide any details
S – significant: coverage including several sentences or paragraphs with some details

E – extensive: coverage including many paragraphs and extensive details.

The coders in this preliminary study were the two authors who are academics. One author specializes in networking and security while the other author's area is databases, analytics and system development. The authors believe that extensive coverage of all six areas of the Information Exposure model will provide users with the information they need to make reasoned decisions about providing personal information to a website.

RESULTS

The results from using the Information Exposure model as a guide to evaluating the top 10 websites are shown below. Table 1 shows the ratings for each of the top 10 independent websites by each rater. Table 2 shows the deltas between the ratings. A zero indicates a match. A one indicates that the ratings were within one of each other; etc. A calculation was made of the percentage of matches and the percentage of within-one ratings.

Overall, the two raters agreed on 62% of the ratings and were within 1 rating level 90% of the time. There were interesting differences across the rating dimensions and the rated websites. For example, the authors were in significant agreement about the levels of Information Collection and Information Sharing coverage but had serious differences in amount of coverage of benefits. There was perfect agreement on one website and for 70% of the websites the raters were within 1 on all dimensions. (See Tables 1 and 2; Figure 2). For Twitter, Linkedin and Wikipedia there were distinct differences in coding. These sites deserve additional investigation.

Table 1. Ratings (Rater 1/Rater 2)

Site	Collection	Retention	Sharing	Benefits	Security	3rd Party
01 Google.com	E/E	N/N	E/E	S/S	S/E	M/N
02 Facebook.com	E/E	N/N	E/E	E/S	S/S	M/N
03 Amazon.com	E/E	N/N	E/E	M/N	S/S	N/N
05 Yahoo	S/E	N/N	E/E	M/N	M/S	M/N
06 Wikipedia.com	E/E	S/S	E/E	E/N	S/S	S/N
07 Ebay.com	E/E	S/S	E/E	E/E	M/S	M/S
08 Twitter.com	E/E	S/N	E/E	E/M	N/S	M/N
09 Craigslist.org	E/E	M/M	E/E	M/M	M/M	M/M
10 Linkedin.com	E/E	S/E	E/E	E/N	S/S	M/S
11 Reddit.com*	E/E	E/S	E/E	S/M	S/E	M/M

* Note: the number #4 website is YouTube.com which has the same privacy page as #1 Google.com. The #11 website, Reddit.com, was added to provide the top 10 sites with unique privacy statements.

Table 2. Rating Deltas

Site	Collection	Retention	Sharing	Benefits	Security	3rd Party	Match%	+1%
01 Google.com	0	0	0	0	1	1	67%	100%
02 Facebook.com	0	0	0	1	0	1	67%	100%
03 Amazon.com	0	0	0	1	0	0	83%	100%
05 Yahoo	1	0	0	1	1	1	33%	100%
06 Wikipedia.com	0	0	0	3	0	2	67%	67%
07 Ebay.com	0	0	0	0	1	1	67%	100%
08 Twitter.com	0	2	0	2	2	1	33%	50%
09 Craigslist.org	0	0	0	0	0	0	100%	100%
10 Linkedin.com	0	1	0	3	0	1	50%	83%
11 Reddit.com	0	1	0	1	1	0	50%	100%
Match %	90%	70%	100%	30%	50%	30%		
+1 %	100%	90%	100%	70%	90%	90%		

Figure 2 compares the Raters' scores on the different model dimensions. The dimension with the least agreement is Benefits. Only 30% of the Benefits ratings were matches and 70% were within-one. The differences appear to be due to these sites intermingling statements about benefits with statements about data collection and sharing. One coder split up the paragraphs into separate statement codes and the other applied the dominant code to the whole

section. Websites that clearly separated out and emphasized benefits were more consistently coded. A similar situation appeared to occur with the Retention dimension for Twitter.

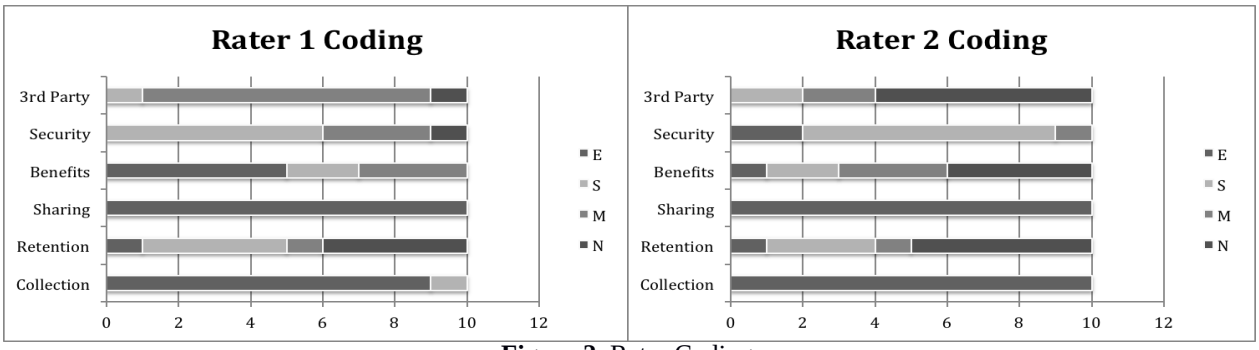

Figure 2. Rater Coding

The differences on the Security dimension for Twitter had to do with how the coders treated mention of the Safe

Harbor Framework. One coder did not feel this was directly related to most U.S. users' information exposure while the other coder found it to be significant.

While the coding differences and similarities are interesting, the main point of the study was to assess whether individual website users had enough information to evaluate their information exposure and risk. To visualize this for each of the top websites, the authors numerically coded the coverage levels; i.e., E=3, S=2, M=1, and N=0. The ratings for the two raters were combined by averaging their numerical scores. The results are shown in Figure 3.

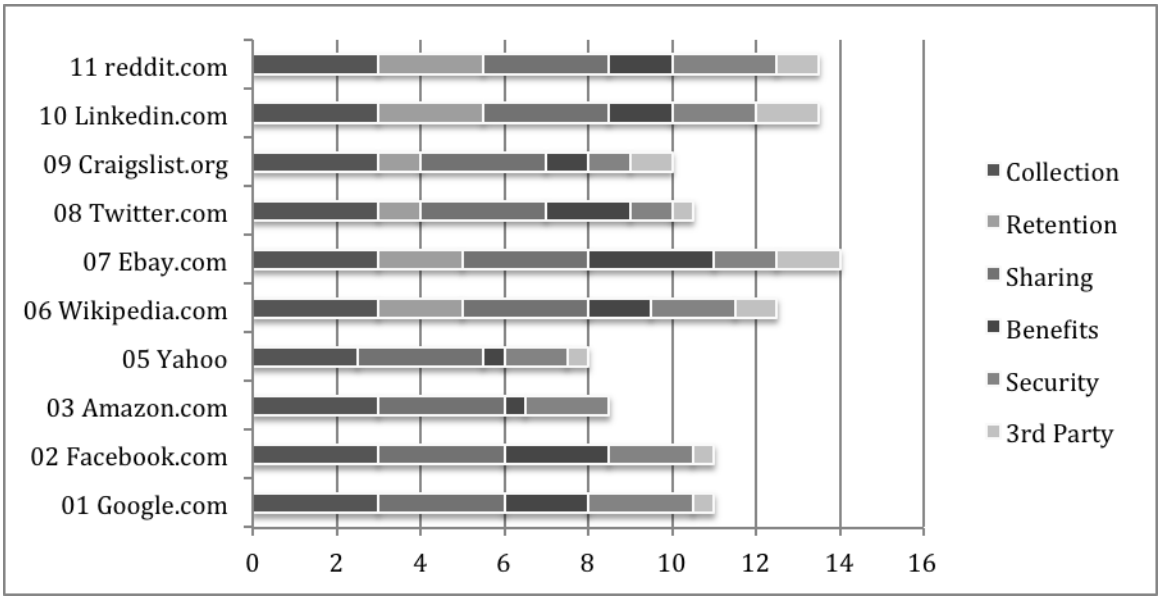

Figure 3. Combined Coverage Scores

The figure illustrates that most of the websites provided extensive coverage of the information they collect. This would be useful to the user to assess how much information they are disclosing to the website and how sensitive the information is. In our analysis we did not separate those two model dimensions in the coding. The figure also shows that websites provided extensive information about information sharing. Sharing leads to multiple copies of the disclosed information and an increased risk due to more information exposure. The figure indicates that all sites provided some information about their internal security but the amount varied significantly. This is important because it helps the user evaluate the possibility of unintentional disclosure of the exposed information.

The biggest differences among websites come from the Retention, Benefits and 3rd Party dimensions. The longer information is retained by an organization the greater the information exposure. Unfortunately some privacy statements didn't mention it at all. All sites documented their information sharing, but few provided information about how that shared information would be secured by 3rd parties. This is information exposure risk caused by the website but there is no information provided to the user on how they are protected. Where there was a mention of 3rd party security it usually was a suggestion for the user to "review the party's privacy statement". Clearly, this is a barrier to the user getting information critical to assessing their information exposure risk.

Interestingly, the biggest differences in coding, and a major difference among websites, is the provision of information about the benefits users will gain from disclosing their personal information. This is a critical part of the decision by the user to disclose personal information. Figure 3 suggests that all websites mention it but the coverage goes from a minimum by Amazon and Yahoo to extensive by Ebay and Facebook. The coding differences mentioned above suggest that users may have trouble getting the information they need from the privacy pages.

CONCLUSIONS

The purpose of this study is to determine if users visiting a popular website are given enough information to determine if they are receiving enough benefit to justify the risk they are incurring by providing personal information to the site. In order to help them assess the risk/benefit of that transaction an Information Exposure model was proposed. The goal of the proposed model was to be simple to apply using information readily available from the website. The model (Figure 1) was based on the concept of Information Exposure as the determinant of risk. The user's goal would be to limit information exposure while maximizing information disclosure benefit. As an initial test of the model, the authors reviewed popular website privacy statements to see if they had the information necessary to apply the model for individual decision-making.

Through their independent coding the authors found that it was easier to find important model information from some website privacy statements than from others. Websites tended to talk about information collection and sharing but not about information retention and 3rd party security. For some websites it was also hard to relate specific benefits to specific information disclosure.

If it is our goal to provide users with the information they need to make rational decisions about trading the risks of information exposure for the benefits of information disclosure, then the information they need should be made more prominent and it should be standardized. All of the tested websites placed a link at the bottom of their homepages that connected to their privacy statement. Most provided a complete privacy statement on clicking the link but a few required an additional link to get to the complete statement. The authors also downloaded privacy statements from the next 10 most popular websites and found a few that seemed to hide their link to a privacy statement. The websites reviewed here are commended for not doing that.

We believe that the privacy statement should be even more prominent since the risk/disclosure decision should be made before using the website. A standard for having the link to the privacy statement at the top left of every homepage would be preferable. The link should be clearly labeled and first time users should be directed to it. The statement itself should clearly address the model dimensions. Every data item collected should have a matching benefit for the user. If the benefit is only for the website or 3rd parties and not for the user, then that should be made clear. Data sharing and data retention should be clearly stated. When data is shared with 3rd parties the website should still be responsible for communicating about its security. The transaction the user is making is between himself/herself and the website. The website needs to take full responsibility for all information exposure including by 3rd parties.

This was only a preliminary study to test out the Information Exposure model. Further work is required. The model and assessment should be done for other websites and for other countries. More raters should be used to get different perspectives on coding privacy statements. Perhaps an experiment could be done with user subjects who try to evaluate the privacy statements of popular websites. Different prototype privacy statements could be tested to see which is most effective at communicating the necessary risk/benefit information to users. The authors ruminated about a possible web application that would use the Information Exposure model to structure the decision-making about using a website. On visiting the website for the first time, the application would execute and lead the user through a short, simple dialog to determine if they should expose their personal information for that site.

There are numerous other limitations to this study. One obvious limitation is that most users do not even consider looking at the privacy statement. The author's believe that this is the case because (1) the statement is linked at the bottom of the page and (2) that it is cumbersome to read and evaluate. When printed out the privacy statements for this study ranged from a little over 1 page (Craigslist) to 22 pages (Wikipedia). No one will read 22 pages of privacy statement. The challenge is to communicate what is necessary in a straightforward and effective way. We believe the Information Exposure model provides a framework for doing that.

Another limitation of the study is that it is based on the number of statements related to the model dimensions. It does not evaluate the content of those statements. The assumption is that more statements about information content collected or information shared means more information about the amount and sensitivity of information involved. The user needs these details in order to assess the risk. It may be that the statements are vague and redundant. A more concise communication of details would be preferred. That would require a much more detailed and time-consuming analysis.

This paper has advanced the understanding of the personal risks and benefits involved when individuals visit websites and contribute their personal information to a Big Data entity. It is hoped that by utilizing the Information Exposure model that users will be able to make better decisions about whether to trade their personal information for website benefits.

ACKNOWLEDGEMENT

The authors wish to thank the anonymous reviewers for their helpful and constructive comments.

REFERENCES

1. Alexa. Top Sites in the United States. http://www.alexa.com/topsites/countries/US, accessed May 9, 2015.
2. Alexa. How are Alexa's Traffic Rankings Determined? https://support.alexa.com/hc/en-us/articles/200449744-How-are-Alexa-s-traffic-rankings-determined-, accessed May 13, 2015.
3. Bakerville, R. (1993). Information Systems Security Design Methods: Implications for Information System Design. *ACM Computing Surveys*, Vol. 25, No. 4, December 1993.
4. Boehm, B. (1991). Software Risk Management: Principles and Practices. IEEE Software, Jan 1991, 32-41.
5. Clarke, R. (2009) Privacy Impact Assessment: Its Origins and Development. *Computer Law and Security Review*, 25, 2009, pp. 123-135, Elsevier, ScienceDirect.
6. Courtney, R. (1977). Security Risk Assessment in Electronic Data Processing. In the AFIPS Proceedings of the National Computer Conference 46. AFIPS, Arlington, Va., 97-104.
7. ICO. (2014) *Conducting Privacy Impact Assessments Code of Practice*, Information Commissioner's Office (UK), Version 1.0, February 2014.
8. NIST (2012). Guide to Conducting Risk Assessments, Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology, U.S. Department of Commerce, NIST Special Publication 800-30 Revision 1.
9. Pew Research Center (2014). *Public Perceptions of Privacy and Security in the Post-Snowden Era*. November , 2014. (http://www.pewinternet.org/2014/11/12/public-privacy-perceptions/)
10. Polonetsky, J., Tene, O., and Jerome, J. (2014). *Benefit-Risk Analysis for Big Data Projects*. Future of Privacy Forum (FPF). Washington, D.C.
11. Tene, O. and J. Polonetsky, (2013). Big Data for All: Privacy and User Control in the Age of Analytics, *Northwestern Journal of Technology and Intellectual Property*, Vol. 11, Issue 5. 273.
12. Trilateral Research & Consulting, (2013). *Privacy Impact Assessment and Risk Management*. Report prepared for the Information Commissioner's Office, May 4, 2013.