Investigating Similarity Between Privacy Policies of Social Networking Sites as a Precursor for Standardization

Emma Cradock
Electronics and Computer Science
University of Southampton
Southampton, UK
erc1e10@soton.ac.uk

Dr David Millard
Electronics and Computer Science
University of Southampton
Southampton, UK
dem@ecs.soton.ac.uk

Dr Sophie Stalla-Bourdillon
School of Law
University of Southampton
Southampton, UK
S.StallaBourdillon@soton.ac.uk

ABSTRACT
The current execution of privacy policies, as a mode of communicating information to users, is unsatisfactory. Social networking sites (SNS) exemplify this issue, attracting growing concerns regarding their use of personal data and its effect on user privacy. This demonstrates the need for more informative policies. However, SNS lack the incentives required to improve policies, which is exacerbated by the difficulties of creating a policy that is both concise *and* compliant. Standardization addresses many of these issues, providing benefits for users and SNS, although it is only possible if policies share attributes which can be standardized. This investigation used thematic analysis and cross- document structure theory, to assess the similarity of attributes between the privacy policies (as available in August 2014), of the six most frequently visited SNS globally. Using the Jaccard similarity coefficient, two types of attribute were measured; the clauses used by SNS and the coverage of forty recommendations made by the UK Information Commissioner's Office. Analysis showed that whilst similarity in the clauses used was low, similarity in the recommendations covered was high, indicating that SNS use different clauses, but to convey similar information. The analysis also showed that low similarity in the clauses was largely due to differences in semantics, elaboration and functionality between SNS. Therefore, this paper proposes that the policies of SNS already share attributes, indicating the feasibility of standardization and five recommendations are made to begin facilitating this, based on the findings of the investigation.

Categories and Subject Descriptors
H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *Standardization;* K.4.1 [**Computers and Society**]: Public Policy Issues – *Privacy*

General Terms
Human Factors, Standardization, Theory, Legal Aspects.

Keywords
Privacy Policies; Standardization; Web Science; Data Protection; Social Networking Sites.

WWW'15 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2743050

1. INTRODUCTION
Technological advancements, including the creation of the web, have dramatically increased the potential of personal data [24], which has lead to concerns over the effect of this on user privacy. To some extent, Data Protection laws have been used to strike the balance between the rights of individuals to privacy and the ability of organizations to use personal data [24]. One right given to data subjects under the EU Data Protection Directive [12] is the right to information about the nature of processing of their personal data, leading to the adoption of privacy policies. However, despite the many benefits of a well-executed privacy policy, their current role in informing users about the use of their data has been heavily criticized [19]. The growing concern SNS attract regarding their use of personal data and its effect on user privacy [3] highlights the need for more informative policies. However, SNS lack the incentive to improve policies, which is only exacerbated by the difficulties which creating a policy that is both concise *and* compliant entail. As a suggestion for improvement, the standardization of privacy policies of SNS, addresses many of these issues [10]. However, standardization is only possible if policies share attributes on which standards can be built. Therefore, using thematic analysis [8] and cross-document structure theory [1] our research investigates the similarity of the privacy policies of SNS and answers the following research questions: Firstly, what is the similarity, between the privacy policies of the top six SNS globally, in the clauses they use? Secondly, what is the similarity between the privacy policies of the top six SNS globally, in the coverage of forty recommendations, made by the UK Information Commissioners Office (ICO)? Thirdly, is standardization possible between the privacy policies of SNS?

2. RELATED WORK
2.1 Privacy Policies
Also called privacy notices, privacy policies are the explanations individuals are given when information is collected about them [16]. However, in reality, their role in informing users is unsatisfactory. They have been heavily criticized for being too long [19], legalistic, complex [16] and ineffective in helping users understand their rights [26], all of which result in them not being read, defeating their underlying purpose. Arguably, organizations are complicit in this, as with individuals still using their services, they lack strong incentives to improve privacy policies. Even with incentives, creating a policy that is both concise and compliant is not easy, given the supranational nature of the web, where data is processed in numerous jurisdictions, each with differing requirements [24]. However, if executed well, privacy policies can promote transparency and reduce information asymmetry [30]. As a risk trade-off decision [15], in a world increasingly

worried about risk [4], privacy policies can communicate information enabling users to make effective choices regarding their personal data. Indeed, evidence suggests that users are privacy aware and active [7], just that they do not view privacy policies as a means of expressing consent [23]. Social Networking Sites (SNS) are a prime example of the need for informative privacy policies. A product of Web 2.0, SNS allow users to upload and share content, and an influential factor in their popularity is that they are free to use [27]. However, as businesses, the trade-off for free use is the data harvested from users. However, a 2011 survey [28] found that 72% of social network users worry that they are giving away *too* much data online. Indeed, it is not just the data users willingly share, or the data gathered without users knowing which SNS can access, but also the information they can infer about users. A recent study found that publicly available Facebook 'likes' could be used to predict a variety of attributes, including ethnic origin, religious beliefs and sexual orientation [18]. Therefore, although SNS rely on personal data, it is questionable how much data this entitles them to, especially given the criticisms they have received, including of the wide licenses they have imposed [3]. Thus, improving the privacy policies of SNS, as the second most frequently visited type of website [2], will benefit numerous users. Various suggestions have been made to improve privacy policies including using visualizations to aid communication [5] and taking an approach similar to the creative commons model [23]. Alternatively, a technical approach could be taken, such as making privacy policies machine-readable, as was the aim of the Platform for Privacy Preferences Project (P3P) [20]. Despite achieving limited adoption [20], some believe P3P-based techniques have considerable potential, the challenge being to design formalized privacy policy languages [20]. If achieved, an intelligent recommender system could be used to help users make decisions about their online privacy, combining user data and policies to provide recommendations for privacy management to the user [22].

2.2 Standardization

As a suggestion, standardization has the most potential here, offering benefits to various stakeholders as well as beginning the groundwork for other improvements. Benefits for users include facilitating comparisons between policies and allowing consumers to become familiar with terminology and the locations of particular types of information [10]. Benefits to organizations include allowing them to verify compliance with the law [10] and reducing the hassle of creating the policies completely on their own. Standardization also allows for large-scale analysis of privacy policies [11], allowing regulators and researchers to assess compliance and move away from human annotation, which is required to understand and compare privacy policies (as in this study). Standardizing elements of policies also begins the groundwork for other suggested improvements by beginning the process of information reduction and refinement required to develop formalized privacy policy languages [20], or standardized descriptions for a creative commons approach [23]. However, standardization requires policies to have shared attributes. Given the fragmented evolution of SNS privacy policies, in their creation by different organizations, combined with the differing legal requirements between jurisdictions, the shared attributes required may not be present in the policies in their current form. Therefore, prior to attempting standardization, it is important to assess similarity of the data in question, to ascertain whether standardization is possible.

3. METHODOLOGY

3.1 Selection of the Data

To assess the similarity between SNS, the most frequently visited SNS (as ranked by Alexa.com [2], a web analytics website that publishes a global traffic rank for major websites) were chosen, as these attract the most users. Alexa.com allows visitors to browse websites by category and their category 'Social Networking' was used for the purpose of this investigation. As the top five websites in 'Social Networking' were also ranked in the top thirty of *all* websites globally, they were included, in addition to the next ranked SNS, giving a total of six. The seventh ranked website in 'Social Networking' was ranked 175th of all websites, this and all others were excluded from the investigation. As a result, the six SNS selected for investigating were: Facebook, Twitter, LinkedIn, Google+, Pinterest and Flickr. Because online privacy policies are made of numerous layers and links to further information, the scope of the investigation was limited to the 'first layer'. This was because the familiarization stage of the investigation (at Stage 1) revealed a large amount of repetition and circularity in further layers. Neither Facebook nor Google provided a separate 'first layer'. When you clicked on their 'privacy' or 'privacy policy' tab, their policies were broken into sections, with an option to download complete versions. Therefore, the complete versions were chosen as the 'first layer' because looking at one section would have created a confounding variable. All other 'first layers' were treated as first screen shown when the 'privacy' or 'privacy policy' tabs were clicked.

3.2 Measuring Similarity

With the aim of the investigation being to measure similarity as a precursor to standardization, it was important to select the appropriate attributes to measure. In terms of granularity, whilst attributes such as words could be compared, the clauses used by the SNS proved an appropriate attribute, as they convey enough information to make comparison meaningful. A 'clause' is defined as '*a part of a treaty, law or contract*' [21]. Whilst it may be useful for other investigations to compare how many times the word 'privacy' appears, here it would not provide a meaningful measure of similarity on which the potential for standardization could be assessed. A second attribute was also measured for similarity, the coverage of forty recommendations from the UK Information Commissioner's Office 'Privacy notices code of practice' [16]. The code provides recommendations, which aim to help websites comply with the UK Data Protection Act 1998 (DPA) [14], the implementation of the EU Data Protection Directive in the UK. EU and UK law was chosen as the framework for which to compare the privacy policies against in this investigation primarily for two reasons. Firstly, that the DPD provides a single law, aimed at harmonizing data protection laws throughout the European Union. Therefore, although differing implementations between member states (who are left to decide the means to achieve the aims of the Directive) sometimes result in slight nuances between the individual implementations, compliance with the UK DPA can, to a certain extent, indicate findings about the rest of the EU. In comparison, the US has no single comprehensive federal (national) law so findings of similarity of the privacy policies with individual laws are not as widely applicable. Secondly, comparing the policies of US-based SNS with the EU and UK law provides for an interesting juxtaposition and strengthens conclusions about the possibility of a global standardized policy. A combination of thematic analysis and cross-document structure theory (CST) were used in this investigation. Thematic analysis [8] is used to pinpoint, examine and

record themes within data and occurs in the six stages outlined below. Cross-document structure theory [1] is a formal discourse theory for multi-document analysis, which establishes relationships among segments of different documents about the same topic. Similarity for both attributes was measured using Jaccard's similarity coefficient, a statistic used for comparing the similarity of sample sets. It is defined as the size of the intersection divided by the size of the union of the sample sets [17].

Stage 1: Familiarization with data: At this stage, researchers immerse themselves in the data, gaining familiarity with its depth and breadth [8]. Therefore, the policies were read multiple times, first passively then actively, recognizing meanings and patterns.

Stage 2: Generating Initial Codes: This phase involved the production of initial 'codes' from the data, defined as '*the most basic segment of the raw data that can be assessed in a meaningful way regarding the phenomenon*' [6]. Unlike some legal documents, privacy policies are only broken into sections, requiring the clauses to be identified. As the thematic analysis definition of 'code' and the (above) definition of 'clause' were compatible, this stage was used to identify the atomic clauses in the policies. The policies were divided into sentences and beginning with Facebook (as the longest policy) a table was created, initially treating each sentence as a clause. Sentences were then examined to see whether multiple sentences needed to be combined to form a clause, or whether multiple clauses were contained within one sentence. The clauses resulting from this formed the initial list of clauses. Here a technique from CST was introduced and sentence pairs were examined [25], similar to the thematic analysis 'compare and contrast approach' [13]. All other policies were also split into sentences and each sentence pair was compared individually asking each time; what is the sentence about? What question is it trying to answer? Is it equivalent to the examined clause in these respects? Would adding or subtracting information from the same privacy policy make the clause equivalent?

Once all policies had been worked through, the table was repeatedly checked until no more clauses were moved, known as achieving *theoretical saturation* [29]. As a result of breaking the policies down into atomic clauses, each clause could only be coded once (i.e. only be classed equal to one other clause) unlike other applications of thematic analysis, which code individual extracts of data into numerous codes. During this stage, some information from the policies was removed, such as duplicate clauses in the same policy, sub-headings mentioned in the body of the text and sentences preceding lists. For example, the subheading '*your information*' was removed from Facebook when the first line in the section began '*your information is*'. However, '*Information for users outside the United States and Canada*' was left in because following this, only contact information was provided, so the subheading was required for context. Removing these was to normalize the data, because including them could inflate the number of clauses some SNS had and skew the results.

Stage 3: Searching for Themes Among Codes: This phase re-focuses the analysis at broader themes and involved sorting the clauses into potential themes [8]. Rather than inductively producing the themes from the clauses, forty ICO Code [16] recommendations were used as themes into which the data was placed. The code states that it can be used as a list for organizations to check their privacy policies against, so it was parsed manually and forty-six recommendations were identified (also using the process of Stages 1 and 2 of this investigation). Although forty-six were identified, six were too broad or vague to assess e.g. '*Any further information necessary, in the specific circumstances, to enable the processing in respect of the individual to be fair*'. They were removed, leaving forty themes for analysis. Each one of the clauses from Stage 2 was then placed into at least one of the forty categories, or into a category of 'miscellaneous'. The focus here was to see whether the privacy policies contained information, which addressed the theme, not whether the privacy policy was legally compliant with the recommendation. For example, one of the forty ICO code recommendations was: '*Obtain assurances (in form of written agreements) from any organizations you share personal information with about what they will do with the information and what the effect on people is likely to be'*. Two clauses coded into this recommendation LinkedIn's privacy policy were:

- '*These third-party developers have either negotiated an agreement to use LinkedIn platform technology or have agreed to our self-service API and Plugin terms in order to build applications ("Platform Applications")*'.

- *'Both the negotiated agreements and our API and Plugin terms contain restrictions on how third parties may access, store, and use the personal information you provide to LinkedIn'*.

Although this meant that LinkedIn included information in its privacy policy, which addressed the recommendation, it would take further investigation (outside the scope of this paper) to assess whether the assurances obtained are legally compliant and therefore whether LinkedIn complies with the recommendation in full or just addresses an aspect of it.

Stage 4: Reviewing Themes: This stage involves two levels. Level one involves reading the collated extracts for each theme and considering whether they appear to form a coherent pattern [8]. If not, the researcher considers whether the theme is problematic or whether the data simply does not fit there, in which case, the theme can be re-worked. Level two involves a similar process but in relation to the whole data set [8]. Because the themes used here were pre-determined from the ICO Code, all this stage required was to check that each clause had been allocated to a recommendation correctly.

Stage 5: Defining and Naming Themes: In this stage, themes are named and content of the theme is paraphrased, clearly defining what themes are and are not [8]. As with the previous stage, because the themes were pre-determined and (as recommendations) defined already by the ICO code (thematic codes were essentially the recommendations themselves), this stage was not required. The definitions the code provided allowed me to state categorically whether a privacy policy had addressed a recommendation or not, giving me the binary classification required to use Jaccard's Coefficient. Table 1 shows three examples of code recommendations and their definitions, which formed the thematic codes.

Table 1: Examples of code recommendations

ICO Code Recommendation	Definition For Thematic Coding	Example of Clause
Tell people how long you or other organizations intend to keep the data.	The privacy policy refers to how long it (or organizations it shares the data with) intend to keep the data for.	*'Typically, information associated with your account will be kept until your account is deleted'.* **Facebook**
Tell people who their information will be shared with/ disclosed to.	The privacy policy advises who users information will be shared with/ disclosed to.	*'Secret boards are visible to you and other participants in the board, and any participant may choose to make the contents of the board available to anyone else'.* **Pinterest**
Tell people the purpose for using the information.	The privacy policy tells the user the purpose for using the information.	*(If you email us, we may keep your message, email address and contact information)* *to respond to your request.* **Twitter**

Stage 6: Producing the Final Report: Here the story of the data is told and this can be found in the next three sections.

4. RESULTS AND ANALYSIS

Table 2 displays the results from Stage 2 showing how many clauses were identified initially, how many were removed and the remaining number of clauses, which were assessed for similarity.

Table 2: Number of clauses identified, removed and remaining

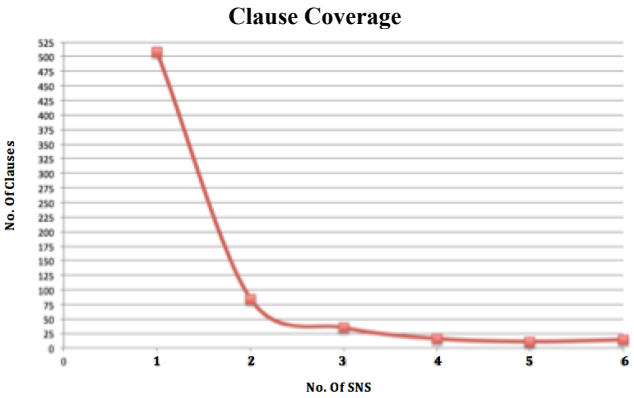
	FB	P	T	F	L	G	Total
No. Clauses	479	106	131	89	384	186	1375
No. Clauses Removed	141	19	23	23	150	33	389
% Clauses Removed	29.44	17.92	17.56	25.84	39.06	17.74	28.29
Remaining No. Of Clauses	338	87	108	66	234	153	986

From Table 2 it can be seen that Facebook and LinkedIn 's 'first layer' included significantly more clauses than the other SNS. Google (ranked third in descending order of number of clauses) had less than half the number of LinkedIn (ranked second). Interestingly, there is no direct relationship between the number of clauses identified and the number of clauses removed, indicating that the length of the policy did not necessitate repetition.

Table 2 also shows that although the descending order of SNS in terms of the number of clauses identified and number of clauses remaining stays the same (Facebook, LinkedIn, Google+, Twitter, Pinterest, Flickr), the order varies in terms of the number and percentage of clauses removed. Flickr in particular had just over a quarter of its clauses removed, which given only 89 were identified to begin with (the lowest), is a significant amount.

4.1 Similarity in Clause Coverage

Graph 1 shows evidence of a power-law relationship between the number of clauses and how many policies they appear in. Generally, as the number of clauses examined increases, the number of SNS they can be found in decreases. Although, there is an increase (rather than decrease) in the number of common clauses as the number of SNS increases from five to six, few empirical phenomena obey power laws for all values [9].



Graph 1: Shows number of clauses covered by SNS

In answering research question 1 (section 1), Table 3 shows that the similarity between the SNS in the clauses they use was low. The least similar were Flickr and Facebook with 8% similarity and the most similar were Pinterest and Twitter with 27% similarity. Average similarity was 15%.

Table 3: Jaccard Similarity of Clause Coverage

SNS	FB	P	T	F	L	G+
Facebook (FB)		0.09	0.14	0.08	0.15	0.10
Pinterest (P)			0.27	0.18	0.13	0.19
Twitter (T)				0.17	0.21	0.19
Flickr (F)					0.11	0.15
LinkedIn (L)						0.13
Google+ (G+)						

Interestingly, Table 2 shows that Facebook and Flickr were at separate ends of the continuum in terms of number of clauses, with Facebook having the most and Flickr the least. This may explain their dissimilarity. Whereas, Table 2 shows that Pinterest and Twitter sit next to each other on this continuum, with a similar number of clauses, which may be why they have a higher similarity. Three prominent reasons for differences between SNS in the clauses they used were:

- **Functionality:** Differences in functionality between SNS, such as LinkedIn 's use of Polls and Facebook's 'Instant Personalization', resulted in a number of

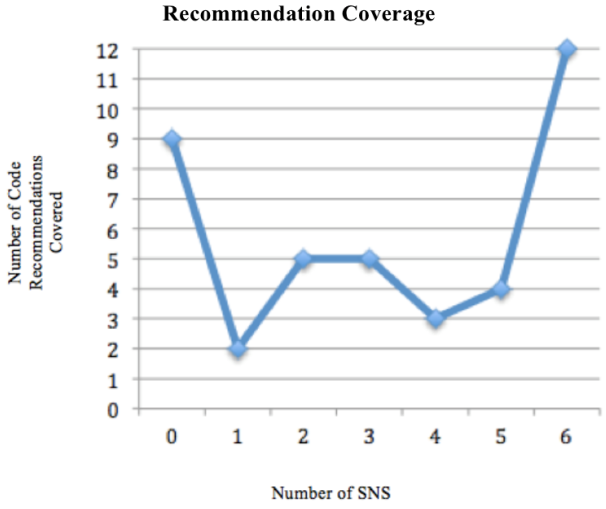
clauses communicating these, which other SNS would not include because they do not offer the functionality.

- **Semantics:** Different words were often used to discuss the same topics without being defined. When discussing account termination, the words 'close', 'delete' and 'deactivate' were used across the policies. Facebook confirmed 'delete' meant permanent deletion but Pinterest only stated users could 'close your account at any time'. Without defining 'close', it was difficult to ascertain whether the clauses were comparable, meaning they had to be treated as different.

- **Elaboration:** Some SNS elaborated on certain topics more than others. For example, although all SNS included a link to follow if users had any questions, comments or complaints, some also included their physical address and information regarding the procedure. Equally, SNS provided definitions and examples of varying length and content for technical terms. For example, only Pinterest and Twitter elaborated on the definition of cookies to mean 'persistent' and 'session' cookies, which resulted in additional clauses, not present in other policies.

4.2 Similarity in Recommendation Coverage

Interestingly, Graph 2 shows that the largest percentages of recommendations covered, were for those covered by either none of (22.5%), or all six SNS (30%). These percentages combined account for over half of the recommendations and show similarity between SNS in the recommendations of the code that they do and do not address.



Graph 2: Shows number of recommendations covered by SNS

Unlike Graph 1, there is no evidence of a power-law relationship in Graph 2 between the number of SNS and how many code recommendations they address, rather that the majority of recommendations were either addressed by all SNS or none.

Table 4: Jaccard Similarity in Covering Code Recommendations

	C	FB	P	T	F	L	G+
ICO Code (C)		0.58	0.45	0.48	0.48	0.65	0.65
Facebook (FB)			0.52	0.68	0.68	0.81	0.63
Pinterest (P)				0.76	0.61	0.52	0.69
Twitter (T)					0.65	0.61	0.66
Flickr (F)						0.66	0.61
LinkedIn (L)							0.68
Google+ (G+)							

Table 4 shows the similarity of SNS with the Code, ranges from 45-65% (of course the code includes 100% of recommendations). This may be because the recommendations are based on the EU and UK Data Protection Framework and none of the SNS are based within the UK. Whereas, in answer to research question 2 (section 1), Table 4 shows that similarity, between SNS in addressing the code recommendations, ranges from 52% - 81%. This evidences a higher percentage of similarity amongst SNS in the specific recommendations they addressed, than their overall similarity with the code. These percentages corroborate with Graph 2, in showing that there were certain recommendations that SNS collectively did or did not address.

However, failing to address a recommendation and a lower similarity with the code does not necessitate non-compliance with it. Failing to cover a recommendation could be for one of two reasons: because it was not applicable, or it was applicable, but the SNS failed to address it. For example, none of the SNS explicitly addressed the recommendation that '*Where individuals are required by law to provide personal details, be open and explain why information is being collected and what it will be used for'* i.e. none of the privacy policies explicitly stated that individuals were required by law to provide certain personal details. This may be because individuals are not required by law to provide SNS with personal details or equally, because individuals are, but SNS failed to explicitly address this in their policy. However, which scenario is correct cannot be ascertained without further investigation and without access to information which SNS often do not divulge in full, such as what data the SNS collects and whether this is collection is required by law.

Table 4 also shows that the least similar (with each other) were jointly Facebook and Pinterest (52%) and LinkedIn and Pinterest (52%) and the most similar were Facebook and LinkedIn (81%). Interestingly, Facebook and LinkedIn had the highest numbers of clauses (Table 2) and although Pinterest did not have the lowest, it did have the second lowest with only 21 more than Flickr who had 66 (Table 2). This indicates that the more clauses SNS have, the more ICO recommendations they are likely to share. However, as stated above, failing to address recommendations is not indicative of non-compliance and therefore a lesser length should not be assumed to mean a less compliant policy.

5. DISCUSSION

Based on the results of this investigation, in answering research question 3, this paper asserts that standardization is possible between the privacy policies of SNS, although standardization by

clause may not be appropriate initially. However, thematic similarity suggests standardization is feasible and based on the findings of this investigation five recommendations are made below to facilitate this.

1. **Begin with an as-exhaustive-as-possible list of themes, which a SNS should address rather than focusing on clauses initially**. Because the investigation showed that similarity was far higher between the policies in the recommendations they covered than the clauses they used, SNS policies are already in a position to begin to be standardized by theme. This could form a visually familiar table for users as a first step, consisting of two columns, with the list of standardized themes on the left and the SNS clauses allocated to those themes on the right. In addition to looking at the legal requirements and the advice of data protection authorities to create this list of themes, the privacy policies should also be examined as a source.

2. **General functionality and functionality specific to that website should appear as separate themes.** General functionality would include functionality common to all websites (such as log data) and website- specific functionality would include functionality only that SNS offer. Then users could easily identify differences between SNS by looking at the website- specific functionality theme, in addition to familiarizing themselves with standard collections of data by websites in the general functionality theme.

3. **Definitions, explanations and examples of technical terms should be standardized so that each policy uses the same ones.** Given that it is almost impossible to avoid using technical terms in relation to SNS, at least by doing this, the amount or type of information a user gets in this context will not vary with the SNS they use, lessening confusion and possibly supporting familiarity with definitions and examples.

4. **Certain words should also be standardized.** For example, close, delete and deactivate should not be used interchangeably, but either one word is used or their individual, but separate, definitions (in relation to terminating an account on a SNS) should be standardized i.e. close account always means one thing as does delete account.

5. **Make sure that when standardizing, there is a way for users to easily ascertain when a theme is not addressed and why.** As mentioned, if a theme was not covered it was unclear whether this was because it was not applicable or because the SNS simply failed to do so. Fulfilling this recommendation would solve that issue, making SNS justifications clear to users, regulators and researchers.

6. CONCLUSION AND FUTURE WORK

Therefore, with the aim here being to assess the similarity between the privacy policies of the top six SNS, as a precursor for standardization, this paper proposes that the privacy policies of SNS demonstrate homogeneity and promising potential for standardization.

In answering research question one, analysis initially showed that similarity between the policies in the clauses they used was low. However, in answering research question two, analysis of a second attribute showed that similarity between SNS was far higher in the themes of information addressed. This showed that SNS express similar themes of information, but in different ways. This analysis enabled the answer to research question three, which was that SNS evidence the shared attributes required for standardization to be possible and that by following the recommendations in the previous section, SNS could begin by standardizing their policies by theme and then begin looking at the possibility of standardizing clauses within themes, beginning with certain definitions and the theme of 'general functionality'. Although standardization by clause is not currently feasible, because of the low similarity in clauses, the analysis showed that the differences in clause coverage were largely due to differences in functionality, semantics and elaboration between SNS. Overcoming these, with the recommendations in the previous section, would allow for another assessment of similarity to investigate the potential for standardization by clause.

Future works in this area could also extend the analysis to further layers of the policies, or conduct inter-rater reliability or intra-rater reliability on the findings. An extension of the work could also measure similarity against thematic codes which are a superset of legal requirements from multiple jurisdictions or compare the similarity with a different type of website such as those operating under a chargeable business model. Further investigations could (where the information is available) also look to clarify the ambiguous clauses discovered in the analysis to enable meaningful conclusions on them. Finally, whilst the aim of this paper was to assess the possibility of standardization, the recommendations made as an outcome of this investigation could be followed up with further work into how to put these into practice. Standardizing the privacy policies of SNS would be a major success for individual management of personal data online. Our work demonstrates that this is possible and we hope it will result in a step closer to the standardization of privacy policies of SNS in the future.

7. ACKNOWLEDGEMENTS

This research was funded by the Research Councils UK Digital Economy Programme, Web Science Doctoral Training Centre, University of Southampton. EP/G036926/1.

8. REFERENCES

[1] Aleixo, P. and Pardo, T.A.S. 2008. Finding Related Sentences in Multiple Documents for Multidocument Discourse Parsing of Brazilian Portuguese Texts. In *Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana* – TIL, pp. 298-303. Vila Velha, Espírito Santo. October, 26-28.

[2] Alexa. 2014. *Actionable Analytics for the Web.*

[Online]Available: http://www.alexa.com [Accessed: 21st August 2014].

[3] Anderson, H. 2009. A privacy wake-up call for social networking sites. *Ent. L.R. 20(7), 245-248*

[4] Beck, U. 1992. *Risk society: Towards a new modernity* (Vol. 17). London: Sage Publications

[5] Becker, J., Heddier, M., Oksuz, A. and Knackstedt, R. (2014). The Effect of Providing Visualizations in Privacy Policies on Trust in Data Privacy and Security. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on* (pp. 3224-3233). IEEE

[6] Boyatzis, R. E. 1998 Transforming qualitative information: Thematic analysis and code development. London: Sage Publications.

[7] Boyd, D. and Hargittai, E. 2010. Facebook privacy settings: Who cares? *First Monday* 15(8).

[8] Braun, V., and Clarke, V. 2006 Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77- 101.

[9] Clauset, A., Shalizi, C. R., & Newman, M. E. 2009. Power-law distributions in empirical data. *SIAM review*, *51*(4), 661-703.

[10] Cranor, L. F. 2012. Necessary But Not Sufficient: Standardized Mechanisms for Privacy Notice and Choice Journal on Telecommunications and High Technology Law 10(2), 273-307.

[11] Cranor, L.F., Idouchi, K., Leon, P.G., Sleeper, M. & Ur, B. 2013 Are They Actually Any Different? Comparing Thousands of Financial Institutions' Privacy Practices. In *Proceedings of the 12th Workshop on the Economics of Information Security* (WEIS 2013), Jun 11-12, Washington, DC.

[12] European Parliament and of the Council. 1995. DIRECTIVE 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data [Online] Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML [Accessed 21st December 2014].

[13] Glaser, B. G. 1978. Theoretical sensitivity: Advances in the methodology of grounded theory (Vol. 2). Mill Valley, CA: Sociology Press.

[14] Great Britain. *Data Protection Act 1998: Elizabeth II (1998)* London: The Stationary Office

[15] Grossklags, J., and Acquisti, A. 2007. When 25 Cents is Too Much: An Experiment on Willingness-To-Sell and Willingness-To-Protect Personal Information. In *WEIS*. [Online] Available at: http://weis2007.econinfosec.org/papers/66.pdf [Accessed: 21st December 2014].

[16] Information Commissioner's Office. 2010. *Privacy notices code of practice* [Online] Available at: http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Detailed_specialist_guides/PRIV ACY_NOTICES_COP_FINAL.ashx [Accessed 21st December 2014].

[17] Jaccard, P. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, *11*(2), 37-50.

[18] Kosinski, M., Stillwell, D., and Graepel, T. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802-5805.

[19] McDonald, A. M. and Cranor, L. F. 2008. The Cost of reading privacy policies *ISJLP*, *4*, 543.

[20] Olurin, M., Adams, C., and Logrippo, L. 2012. Platform for privacy preferences (P3P): Current status and future directions. In *Privacy, Security and Trust (PST), 2012 Tenth Annual International Conference on* (pp. 217-220). IEEE.

[21] Oxford University Press Oxford English Mini Dictionary. 2011. New York: Oxford University Press

[22] Rasmussen, C., and Dara, R. 2014. Recommender Systems for Privacy Management: A Framework. In *High-Assurance Systems Engineering (HASE), 2014 IEEE 15th International Symposium on* (pp. 243-244). IEEE.

[23] Robinson, N., Grauz, H., Botterman, M. and Valeri, L. 2009. *Review of the European Data Protection Directive* [Online] Available at: http://ico.org.uk/~/media/documents/library/data_protection/detailed_specialist_guides/review_of_eu_dp_directive.ashx [Accessed 4th September 2014]

[24] Rowland, D., Kohl, U. and Charlesworth, A. 2012. *Information Technology Law. 4th Ed.* Oxon: Routledge Publishing

[25] Ryan, G. W., and Bernard, H. R. 2003. Techniques to identify themes. *Field methods*, *15*(1), 85-109.

[26] Scribbins, K. 2001. *Privacy@ net: an international comparative study of consumer privacy on the internet*. Consumers International [Online] Available: http://www.consumersinternational.org/media/304817/privacy@net%20an%20international%20comparative%20study%20of%20consumer%20privacy%20on%20the%20internet.pdf [Accessed 21st December 2014].

[27] Sellars, S. 2011. Online privacy: do we have it and do we want it? A review of the risks and UK case law. *European Intellectual Property Review*, *33*(1), 9-17.

[28] Special Eurobarometer 359. 2011. *Attitudes on Data Protection and Electronic Identity in the European Union* [Online] Available at: http://ec.europa.eu/public_opinion/archives/ebs/ebs_359_en.pdf [Accessed 4th September 2014]

[29] Strauss, A., and Corbin, J. M. 1990. Basics of qualitative research: Grounded theory procedures and techniques. London: Sage

[30] Tsai, J. Y., Egelman, S., Cranor, L. Acquisti, A. 2011. The Effect of Online Privacy Information on Purchasing Behaviour: An Experimental Study, Information Systems Research, 1047-7047, Vol. 22(2) 254-268.